RECENT ADVANCES in MATHEMATICAL METHODS in APPLIED SCIENCES

Proceedings of the 2014 International Conference on Mathematical Models and Methods in Applied Sciences (MMAS '14)

Proceedings of the 2014 International Conference on Economics and Applied Statistics (EAS '14)

Saint Petersburg State Polytechnic University
Saint Petersburg, Russia
September 23-25, 2014

RECENT ADVANCES in MATHEMATICAL METHODS in APPLIED SCIENCES

Proceedings of the 2014 International Conference on Mathematical Models and Methods in Applied Sciences (MMAS '14)

Proceedings of the 2014 International Conference on Economics and Applied Statistics (EAS '14)

Saint Petersburg State Polytechnic University
Saint Petersburg, Russia
September 23-25, 2014

All papers of the present volume were peer reviewed by no less than two independent reviewers. Acceptance was granted when both reviewers' recommendations were positive.

Series: Mathematics and Computers in Science and Engineering Series | 32

ISSN: 2227-4588
ISBN: 978-1-61804-251-4

RECENT ADVANCES in MATHEMATICAL METHODS in APPLIED SCIENCES

Proceedings of the 2014 International Conference on Mathematical Models and Methods in Applied Sciences (MMAS '14)

Proceedings of the 2014 International Conference on Economics and Applied Statistics (EAS '14)

Saint Petersburg State Polytechnic University
Saint Petersburg, Russia
September 23-25, 2014

Organizing Committee

Editors:
Prof. Yuri B. Senichenkov, St. Petersburg State Politechnical University, Saint Petersburg, Russia
Prof. V. Korablev, St. Petersburg State Politechnical University, Saint Petersburg, Russia
Prof. Igor Chernorytski, St. Petersburg State Politechnical University, Saint Petersburg, Russia
Prof. N. Korovkin, St. Petersburg State Politechnical University, Saint Petersburg, Russia
Prof. S. Pozdnjkov, St. Petersburg State Politechnical University, Saint Petersburg, Russia
Prof. Klimis Ntalianis, Technological Educational Institute of Athens, Greece

Associate Editor:
Prof. Massimo Ceraolo, University of Pisa, Italy.

Senior Program Chair

Prof. Ljubiša Kočinac, University of Nis,
Nis, Serbia

Program Chairs

Prof. Yuri B. Senichenkov,
St. Petersburg State Politechnical University,
Saint Petersburg, Russia
Prof. Constantin Udriste,
University Politehnica of Bucharest,
Bucharest
Romania
Prof. Marcia Cristina A. B. Federson,
Universidade de São Paulo,
São Paulo, Brazil

Tutorials Chair

Prof. Yury A. Rossikhin,
Voronezh State University of
Architecture and Civil Engineering,
Voronezh, Russia

Special Session Chair

Prof. Yuri B. Senichenkov,
St. Petersburg State Politechnical
University,
Saint Petersburg, Russia

Workshops Chair

Prof. Sehie Park,
The National Academy of Sciences,
Republic of Korea

Local Organizing Chair

Prof. Vadim Korablev,
St. Petersburg State Politechnical University,
Saint Petersburg, Russia

Publication Chair

Prof. Marina Shitikova,
Voronezh State University of Architecture and
Civil Engineering, Voronezh, Russia

Publicity Committee

Prof. Vjacheslav Yurko,
Saratov State University,
Astrakhanskaya, Russia
Prof. Myriam Lazard
Institut Superieur d' Ingenierie de la Conception
Saint Die, France

International Liaisons

Professor Jinhu Lu, IEEE Fellow
Institute of Systems Science
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing 100190, P. R. China
Prof. Olga Martin
Applied Sciences Faculty
Politehnica University of Bucharest
Romania
Prof. Vincenzo Niola
Departement of Mechanical Engineering for Energetics
University of Naples "Federico II"
Naples, Italy
Prof. Eduardo Mario Dias
Electrical Energy and Automation
Engineering Department
Escola Politecnica da Universidade de Sao Paulo
Brazil

Steering Committee

Prof. Dr. H. M. Srivastava, University of Victoria, Canada
Prof. Stefan Siegmund, Technische Universitaet Dresden, Germany
Prof. Natig M. Atakishiyev, National Autonomous University of Mexico, Mexico
Prof. Narcisa C. Apreutesei, Technical University of Iasi, Iasi, Romania
Prof. Imre Rudas, Obuda University, Budapest, Hungary

Program Committee

Prof. Nasser-Eddine Mohamed Ali Tatar, King Fahd University of Petroleum and Mineral, Dhahran, Saudi Arabia

Prof. Jianqing Chen, Fujian Normal University, Cangshan, Fuzhou, Fujian, China
Prof. Josef Diblik, Brno University of Technology, Brno, Czech Republic
Prof. Stanislaw Migorski, Jagiellonian University in Krakow, Krakow, Poland
Prof. Qing-Wen Wang, Shanghai University, Shanghai, China
Prof. Luis Castro, University of Aveiro, Aveiro, Portugal
Prof. Alberto Fiorenza, Universita' di Napoli "Federico II", Napoli (Naples), Italy
Prof. Patricia J. Y. Wong, Nanyang Technological University, Singapore
Prof. Salvatore A. Marano, Universita degli Studi di Catania, Catania, Italy
Prof. Sung Guen Kim, Kyungpook National University, Daegu, South Korea
Prof. Maria Alessandra Ragusa, Universita di Catania, Catania, Italy
Prof. Gerassimos Barbatis, University of Athens, Athens, Greece
Prof. Jinde Cao, Distinguished Prof., Southeast University, Nanjing 210096, China
Prof. Kailash C. Patidar, University of the Western Cape, 7535 Bellville, South Africa
Prof. Mitsuharu Otani, Waseda University, Japan
Prof. Luigi Rodino, University of Torino, Torino, Italy
Prof. Carlos Lizama, Universidad de Santiago de Chile, Santiago, Chile
Prof. Jinhu Lu, Chinese Academy of Sciences, Beijing, China
Prof. Narcisa C. Apreutesei, Technical University of Iasi, Iasi, Romania
Prof. Sining Zheng, Dalian University of Technology, Dalian, China
Prof. Daoyi Xu, Sichuan University, Chengdu, China
Prof. Ferhan M. Atici, Western KentuckyUniversity, Bowling Green, KY 42101, USA
Prof. Ravi P. Agarwal, Texas A&M University - Kingsville, Kingsville, TX, USA
Prof. Martin Bohner, Missouri University of Science and Technology, Rolla, Missouri, USA
Prof. Dashan Fan, University of Wisconsin-Milwaukee, Milwaukee, WI, USA
Prof. Paolo Marcellini. University of Firenze, Firenze, Italy
Prof. Xiaodong Yan, University of Connecticut, Connecticut, USA
Prof. Ming Mei, McGill University, Montreal, Quebec, Canada
Prof. Enrique Llorens, University of Valencia, Valencia, Spain
Prof. Yuriy V. Rogovchenko, University of Agder, Kristiansand, Norway
Prof. Yong Hong Wu, Curtin University of Technology, Perth, WA, Australia
Prof. Angelo Favini, University of Bologna, Bologna, Italy
Prof. Andrew Pickering, Universidad Rey Juan Carlos, Mostoles, Madrid, Spain
Prof. Guozhen Lu, Wayne state university, Detroit, MI 48202, USA
Prof. Gerd Teschke, Hochschule Neubrandenburg - University of Applied Sciences, Germany
Prof. Michel Chipot, University of Zurich, Switzerland
Prof. Juan Carlos Cortes Lopez, Universidad Politecnica de Valencia, Spain
Prof. Julian Lopez-Gomez, Universitad Complutense de Madrid, Madrid, Spain
Prof. Jozef Banas, Rzeszow University of Technology, Rzeszow, Poland
Prof. Ivan G. Avramidi, New Mexico Tech, Socorro, New Mexico, USA
Prof. Kevin R. Payne, Universita' degli Studi di Milano, Milan, Italy
Prof. Juan Pablo Rincon-Zapatero, Universidad Carlos III De Madrid, Madrid, Spain
Prof. Valery Y. Glizer, ORT Braude College, Karmiel, Israel
Prof. Norio Yoshida, University of Toyama, Toyama, Japan
Prof. Feliz Minhos, Universidade de Evora, Evora, Portugal
Prof. Mihai Mihailescu, University of Craiova, Craiova, Romania
Prof. Lucas Jodar, Universitat Politecnica de Valencia, Valencia, Spain
Prof. Dumitru Baleanu, Cankaya University, Ankara, Turkey
Prof. Jianming Zhan, Hubei University for Nationalities, Enshi, Hubei Province, China
Prof. Zhenya Yan, Institute of Systems Science, AMSS, Chinese Academy of Sciences, Beijing, China
Prof. Zili Wu, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, China
Prof. Wei-Shih Du, National Kaohsiung Normal University, Kaohsiung City, Taiwan
Prof. Khalil Ezzinbi, Universite Cadi Ayyad, Marrakesh, Morocco
Prof. Youyu Wang, Tianjin University of Finance and Economics, Tianjin, China
Prof. Satit Saejung, Khon Kaen University, Thailand

Additional Reviewers

Bazil Taha Ahmed	Universidad Autonoma de Madrid, Spain
M. Javed Khan	Tuskegee University, AL, USA
James Vance	The University of Virginia's College at Wise, VA, USA
Minhui Yan	Shanghai Maritime University, China
Angel F. Tenorio	Universidad Pablo de Olavide, Spain
Jose Flores	The University of South Dakota, SD, USA
Francesco Zirilli	Sapienza Universita di Roma, Italy
Kei Eguchi	Fukuoka Institute of Technology, Japan
Jon Burley	Michigan State University, MI, USA
Imre Rudas	Obuda University, Budapest, Hungary
Philippe Dondon	Institut polytechnique de Bordeaux, France
Zhong-Jie Han	Tianjin University, China
George Barreto	Pontificia Universidad Javeriana, Colombia
Frederic Kuznik	National Institute of Applied Sciences, Lyon, France
Stavros Ponis	National Technical University of Athens, Greece
Lesley Farmer	California State University Long Beach, CA, USA
Francesco Rotondo	Polytechnic of Bari University, Italy
Genqi Xu	Tianjin University, China
Manoj K. Jha	Morgan State University in Baltimore, USA
Hessam Ghasemnejad	Kingston University London, UK
Ole Christian Boe	Norwegian Military Academy, Norway
Deolinda Rasteiro	Coimbra Institute of Engineering, Portugal
Masaji Tanaka	Okayama University of Science, Japan
Takuya Yamano	Kanagawa University, Japan
Konstantin Volkov	Kingston University London, UK
José Carlos Metrôlho	Instituto Politecnico de Castelo Branco, Portugal
Moran Wang	Tsinghua University, China
Santoso Wibowo	CQ University, Australia
Yamagishi Hiromitsu	Ehime University, Japan
Kazuhiko Natori	Toho University, Japan
Abelha Antonio	Universidade do Minho, Portugal
Matthias Buyle	Artesis Hogeschool Antwerpen, Belgium
Tetsuya Yoshida	Hokkaido University, Japan
Miguel Carriegos	Universidad de Leon, Spain
Andrey Dmitriev	Russian Academy of Sciences, Russia
Dmitrijs Serdjuks	Riga Technical University, Latvia
Shinji Osada	Gifu University School of Medicine, Japan
Tetsuya Shimamura	Saitama University, Japan
Valeri Mladenov	Technical University of Sofia, Bulgaria
João Bastos	Instituto Superior de Engenharia do Porto, Portugal
Sorinel Oprisan	College of Charleston, CA, USA
Alejandro Fuentes-Penna	Universidad Autónoma del Estado de Hidalgo, Mexico
Xiang Bai	Huazhong University of Science and Technology, China
Eleazar Jimenez Serrano	Kyushu University, Japan

Table of Contents

ISBN: 978-1-61804-251-4

ISBN: 978-1-61804-251-4

ISBN: 978-1-61804-251-4

ISBN: 978-1-61804-251-4

ISBN: 978-1-61804-251-4

ISBN: 978-1-61804-251-4

Plenary Lecture 1

Stiff Models and Gradient Methods with the Exponential Relaxation

Professor Igor G. Chernorutskiy
Saint Petersburg State Polytechnical University
Russia
E-mail: igcher@spbstu.ru

Abstract: 1. For a class of matrix gradient methods a new concept of the relaxation function is suggested. This concept allows to evaluate the effectiveness of each gradient optimization procedure, and to synthesize new methods for special classes of ill conditioned (stiff) non-convex optimization problems. According to the suggested formula , it is possible to build relevant search procedures for any given relaxation function.
2. The theorem about the relaxation conditions of each matrix gradient method is proven. Based on the concept of the relaxation functions it is given the geometric interpretation of relaxation properties of gradient methods. According to this interpretation it is possible to build a relaxation area, and to evaluate the speed of the objective function values decreasing.
3. The analysis of classical matrix gradient schemes such as simple gradient method, Newton's methods, Marquardt method is given. It is shown that the relaxation function and its geometric interpretation gives almost full information about the properties and capabilities of relevant gradient optimization methods.
4. A new class of matrix gradient methods with the exponential relaxation function (ERF) is suggested. It is shown that ERF-method summarizes the classical gradient methods including Newton methods, and Marquardt method. In contrast to these methods, ERF-methods have the relaxation functions, entirely located in the relaxation area, which significantly increases the computational efficiency of gradient methods.
5. The ERF-methods convergence for a wide class of non-convex objective functions is established.

Brief Biography of the Speaker: Dr. Chernorutskiy currently is a Professor of Saint-Petersburg State Polytechnical University (SPbSPU). Degrees (SPbSPU): Professor, 1990; Doctor of Technical Science, 1987; Associate Professor, 1982; Ph.D., 1978; M.S., 1970.
Professor Chernorutskiy is the Chair of Information & Control Systems Division of Computer Science and Engineering School (CSES).
Research Interests
Applied Software Engineering, Optimization Tools, Real - Time Systems Modeling and Simulation, Parameter Estimation, and Adaptive Optimization, Decision Support Systems, Artificial Intelligence and Expert Systems.

ISBN: 978-1-61804-251-4

Plenary Lecture 2

Enhancing Efficiency and Performance of Electric Power Systems by Using Smart Grid Technology

Professor Nikolay V. Korovkin
Chef of Theoretical Electrical Engineering Department
Saint Petersburg State Polytechnical University
Russia
E-mail: nikolay.korovkin@gmail.com

Abstract: A new approach for optimization of power system states with Smart grid utilities will be proposed.
The development of electric power systems (EPS) goes to the construction of power plants, connection of new consumers to networks, introduction into service of new power transmission lines. The complication of electric power system structure and configuration results in reduction of their flexibility and has an adverse effect on the main indices of EPS performance: power distribution losses, power quality and power supply security. Actual conditions of operation and development of large EPS call for new control techniques to be introduced, that is why the elaboration of methods to control the power system operation and to optimize its states with respect to various criteria is now the trend of scientific researches of current concern.

Brief Biography of the Speaker: Education (degrees, dates, universities):
1978, Leningrad Polytechnic Institute, research engineer
1984, Leningrad State University, candidate of science (Phd)
1997, Saint Petersburg Polytechnic university, doctor of science
Career/Employment (employers, positions and dates):
1978, Leningrad State University, assistance professor
1984, Leningrad State University, docent
1997, Saint Petersburg Polytechnical university, professor
2010, Saint Petersburg Polytechnical university, head of Theoretical Electrical Engineering department

ISBN: 978-1-61804-251-4

Plenary Lecture 3

Modeling of Mechanism of State and Private Partnership Development of the Social Infrastructure in the Regions

Professor, Doctor of Science, Vice Rector Vladimir V. Gluhov
Saint Petersburg State Polytechnical University
Polytechnicheskaja str., 29, 195251, St. Petersburg
Russia
E-mail: vicerector.me@spbstu.ru

Abstract: 1. There are identified and analyzed the problems of development of social infrastructure in the regions of Russia. It is developed the mechanism and proposed the forms of cooperation for their solution on the basis of private and state partnership.
2. It is developed institutional framework for interaction between city administrations and business communities, aimed at creating an environment for effective development of the social infrastructure in the regions.
3. It is developed the game theory approach for modeling the interaction of city administrations and businesses considering the possible development of the institutional environment.
4. It is described a class of cooperative games simulating the interaction of businesses and city administration.
5. It is proposed a mechanism for solving the problems of social infrastructure development based on the analysis of game interaction models of city administrations and businesses.

Brief Biography of the Speaker: Vice-Rector for administrative and economic activity of St. Petersburg State Polytechnic University, Professor of Russian-German Center of Management and Marketing “Progress”, laureate of state prize “President of Russian Federation Prize in Higher Education”, laureate of St. Petersburg governor prize for excellence in higher education, laureate of V.V. Novozhilov prize (the Russian Academy of Sciences).
Member of following Academies:
• International Academy of Technological Cybernetics
• International Academy of Informational Support
• Baltic Academy of Informational Support
• International Academy of Ecology and Security Sciences
• Academy of Humanities
• International Academy of Higher School Science
• Academy of Municipal Sciences
The scholarly works of Vladimir V. Gloukhov develop the “effective management” research area.

ISBN: 978-1-61804-251-4

Vladimir V. Gloukhov developed the full system of optimization mathematical models for iron and steel enterprises, which found their places in engineering practice and were described in “Mathematical methods and models in manufacturing planning and management” scientific work. These models formed a basis of new school of thought and applied research area – optimization models of iron and steel production.

Vladimir V. Gloukhov has also developed some methods of economic analysis of newest technological processes (in the fields of powder metallurgy, laser processing, ferrous and non-ferrous industry), which have later been implemented in many production enterprises of Russia. The theory of economic analysis of newest technological processes allowed to form the “economics and management of innovation technologies” educational direction.

ISBN: 978-1-61804-251-4

Plenary Lecture 4

On Complete Monotonicity of Some Functions of the Mittag-Leffler Type in Non-Debye Relaxation Processes

Professor Francesco Mainardi
Department of Physics, University of Bologna, and INFN
Via Irnerio 46, I-40126 Bologna, Italy
E-mail: francesco.mainardi@bo.infn.it.it

Abstract: In this talk we discuss some interesting examples of relaxation occurring in viscoelastic and dielectric materials, which are described by special completely monotone functions of the Mittag-Leffler type. This means that these response functions are represented by continuous distributions of elementary (i.e. exponential) relaxation processes via non-negative spectra of relaxation in frequency or time. In addition to the well known functions of Mittag-Leffler type in one and two parameters, we revisit two more general kinds of Mittag-Leffler functions in three parameters, that is the Prabhakar and the Kilbas-Saigo functions. For all these functions we prove the conditions on the parameters to ensure the complete monotonicity and compute the corresponding frequency spectra. For some study-cases we present numerical results with illustrative plots for the field variable and for the corresponding spectral distribution. We hope that our results can be adopted when the field variable is the response function associated with non-Debye relaxation processes found e.g. in dielectrics. In particular we have derived as noteworthy particular cases the classical models of non-Debye relaxation phenomena referred to as Cole-Cole, Davidson-Cole, Havriliak-Negami along with the so-called Kohlrausch-Williams-Watts (KWW) law based on the stretched exponential function.

Brief Biography of the Speaker: For a full biography, list of references on author's papers and books see:
Home Page: http://www.fracalmo.org/mainardi/index.htm
and http://scholar.google.com/citations?user=UYxWyEEAAAAJ&hl=en&oi=ao

ISBN: 978-1-61804-251-4

Plenary Lecture 5

From Physical to Mathematical Circuits: Theoretical and Practical Issues

Professor Massimo Ceraolo
University of Pisa
Italy
E-mail: massimo.ceraolo@unipi.it

Abstract: Electrical engineers typically talk about “circuits”, without first defining what a circuit really is. If we mean circuits to be sets of elements containing insulating and conducting material, as well as magnetic material, nearly everything is a circuit.
If, instead, we mean circuits as “sets of elements in which some wires that connect components to each other are clearly distinguishable”, they constitute a set (the set of all possible circuits) that is a bit more limited, and maybe clear enough.
When talking about circuits, typically electrical engineers think of this latter definition. In addition, they typically assume that Kirchhoff’s equations are valid for all circuits.
This creates theoretical and practical issues that are normally underestimated. In particular:
• Kirchhoff’s laws are not valid in general. In the speech examples of “circuits” (according to the above definition) for which they are not valid are reported;
• the very concept of “potential” of points of the circuits is vague if not totally wrong.
The speech will discuss this inconsistence thoroughly and proposes a solution to the issues the fol-lowing approach:
• Systems in which electric and magnetic phenomena occur are simply called electromagnetic systems; for them Maxwell’s equations are valid, where Kirchhoff’s laws not only are not valid, but even loose meaning
• Systems in which electric magnetic phenomena occur and have a circuital shape, i.e. are composted by lumped components connected to each other by means of insulated wires, are called physical circuits. For them Maxwell’s equations are still valid; they are susceptible to be abstracted in such a way that, under given conditions, mathematical circuits can be inferred from them
• Mathematical circuits, or simply circuits, are abstracted structures, that constitute under given conditions, approximations of actual physical circuits, for which Kirchhoff’s equations are valid, or better, are postulated to be valid. As such, Kirchhoff’s equations are just the version of the continuity (charge conservation) equation and energy conservation for mathematical circuits. Instead of the Maxwell’s equations, for circuits Kirchhoff’s and constitutive equations are valid.
Once circuits (the short name of mathematical circuits) are defined, not all problems are solved.

ISBN: 978-1-61804-251-4

In the speech, the author shows that to obtain circuits from physical circuits containing transmission lines, for which Kirchhoff's laws are valid, is not always possible; however, a special version of them, that will be called metacircuit, will be introduced.
Again, it will be discussed that in circuits with ideal transformers do not allow Kirchhoff's laws to be written in their more common form, and special treatment is needed.
Circuits are lumped component systems: i.e., systems composed by components that are connected to each other through interfaces. Therefore their behavior over time can be computer-simulated using object-oriented tools and languages. The final part of the speech will show that the modern si-mulation language Modelica has an approach that is one perfectly in line with the analysis of this speech, and even the graphical tricks used to evidence lumped components and connections are in total agreement with the Modelica approach.
This gives additional usefulness to the approach proposed in the speech, and in its companion paper.

Brief Biography of the Speaker: Born in 1960, he took his Ms Degree in Electrical Engineering from the University of Pisa, with honours, in 1985. For some years he has worked in an Italian private research centre. Since 1992 he has been working in Electric Power Systems first as a researcher, then as a professor.
He is full professor of Electric Power Systems since 2002, and teaches Electric and Hybrid Vehicles at the University of Pisa and on-board Electrical Systems at the Naval Academy of Livorno.
He is author or co-author of more than one hundred National and International scientific papers, mainly regarding power systems, electrochemical energy storage, and electric and hybrid vehicles.
He is the chairman of the School of Engineering of the University of Pisa, that coordinates teaching activities of around 250 researchers and professors.
He is the main author of the IEEE-Wiley book "Fundamentals of Electric Power Engineering – from Electromagnetics to Power Systems".

ISBN: 978-1-61804-251-4

ISBN: 978-1-61804-251-4

Dynamic response of a hereditarily elastic beam with Rabotnov's kernel impacted by an elastic rod

Yury A. Rossikhin[1], Marina V. Shitikova[1] and Ivan I. Popov[1,2]
[1]Research Center on Dynamics of Solids and Structures
Voronezh State University of Architecture and Civil Engineering
Voronezh 394006, Russian Federation
[2] National Taiwan University of Science and Technology, Taipei, R.O.C.
Email: yar@vgasu.vrn.ru mvs@vgasu.vrn.ru

Dedicated to the 100th Birthday of Russian Academician Yury N. Rabotnov

Abstract – The problem on low-velocity impact of a long thin elastic rod with a flat end upon an infinite Timoshenko-type beam, the viscoelastic features of which are exhibited only within the contact domain and are governed by the fractional derivative standard linear solid model, is formulated. The part of the beam being out of the contact region is considered to be elastic, and its behavior is described by a set of equations taking the rotary inertia and transverse shear deformation into account. At the moment of impact, shock waves are generated both in the impactor and target, the influence of which on the contact domain is considered via the theory of discontinuities. The contact zone moves like a rigid whole under the action of the contact force and transverse forces applied to the boundary of the contact region, which are obtained on the basis of one-term ray expansions. The contact force has been determined analytically via the Laplace transform technique.

Keywords–Impact response, hereditarily elastic Timoshenko-like beam, fractional derivative standard linear solid model, ray method, dynamic conditions of compatibility, Laplace transform.

I. Introduction

The problems connected with the analysis of the shock interaction of thin bodies (rods, beams, plates, and shells) with other bodies have widespread application in various fields of science and technology. The physical phenomena involved in the impact event include structural responses, contact effects and wave propagation. These problems are topical not only from the point of view of fundamental research in applied mechanics, but also with respect to their applications. Because these problems belong to the problems of dynamic contact interaction, their solution is connected with severe mathematical and calculation difficulties. To overcome this impediment, a rich variety of approaches and methods have been suggested, and the overview of current results in the field can be found in recent state-of-the-art articles [1]–[4].

In recent decades fractional calculus (integral and differential operators of noninteger order), which has a long history [5], has been the object of ever increasing interest in many branches of natural science, and of engineering interest as well. Thus, Rossikhin and Shitikova [3] have reviewed the application of fractional calculus to dynamic problems of linear and nonlinear hereditary mechanics of solids, among them, the problems of dynamic contact interaction. Two approaches have been discussed for studying the impact response of fractionally damped systems subjected to falling impactors. The first one is based on the assumption that viscoelastic properties of the target manifest themselves only in the contact domain, while the other part of the target remains elastic one. This approach results in defining the contact force and the local penetration of target by an impactor from the set of linear fractional differential equations. The second approach is the immediate generalization of the Timoshenko approach utilizing the viscoelastic analog of Hertz's contact law by substituting elastic constants by the corresponding viscoelastic operators. This approach results in the nonlinear functional equation for determining the contact force or the impactor's relative displacement.

In the present paper, the analytical approach proposed in [2], [6] for the analysis of the dynamic response of the elastic isotropic Timoshenko beam subjected to the impact by elastic long rod has been extended to the problem of the dynamic response of a hereditarily elastic Timoshenko-like beam impacted by an elastic prismatic long rod of a rectangular cross-section. As this takes place, the impact response of thin isotropic beams is investigated under the assumption that the viscosity of the target exhibits only within the contact domain, while out of the contact region the beam remains to be elastic with a non-relaxed elastic modulus, in so doing viscous features are described by the fractional derivative standard linear solid model.

II. Problem Formulation

Let a long prismatic elastic rod of a rectangular cross-section with the dimensions $2\tau_{\rm im}$ and a move along the z-normal with the velocity V_0 towards an isotropic rectangular Timoshenko beam of infinite extent (this assumption is introduced due to the short duration of contact interaction in order to ignore reflected waves) with width a and thickness h, in so doing the normal z is erected at the middle of the beam.

The beam out of the contact zone is considered to be elastic, while within the contact domain its microstructure changes and it gains viscoelastic properties, which are described by the generalized fractional-derivative standard linear solid model. For the projectile with a flat end, such a scheme could be

ISBN: 978-1-61804-251-4

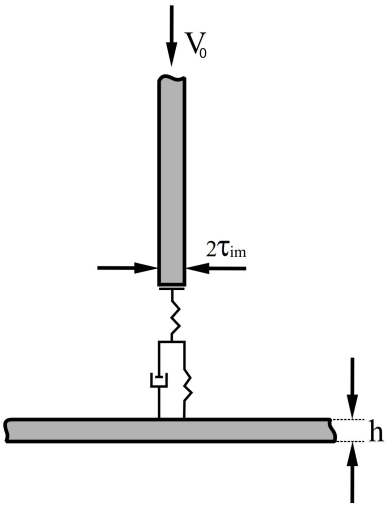


Fig. 1. Scheme of the shock interaction of a plain-end impactor with a target

realized if a viscoelastic buffer involving two springs and a viscous damper is embedded by its low end in the target (Figure 1).

Thus, the rod falls vertically upon the target. Impact occurs at $t = 0$ at the origin of the coordinate system x, y, z. At the moment of impact, shock waves (surfaces of strong discontinuity) are generated in the beam and in the rod, which then propagate along the projectile and the target with the velocities of the transient waves.

Further we will assume that during the process of impact the transverse forces and transverse shear deformations dominate in the stress-strain state of the beam within the vicinity of the contact zone. Besides, the elastic rod and the beam are considered to be somewhat extended, so that the waves reflected from rod's free edge and beam's boundary have had no time to return to the contact region to terminate the collision.

III. Governing Equations

The dynamic behavior of an elastic homogeneous prismatic beam with due account for the rotary inertia and transverse shear deformations is described by the following set of equations [2], [6]:

$$\frac{\partial Q_r}{\partial z} = \rho A \dot{W}, \tag{1}$$

$$\frac{\partial M}{\partial z} - Q = -\rho I \dot{\beta}, \tag{2}$$

$$\dot{Q}_r = K\mu_\infty A\left(\partial W/\partial z - \beta\right), \tag{3}$$

$$\dot{M} = -E_\infty I \partial\beta/\partial z, \tag{4}$$

where M is the bending moment, Q is the shear force, $W = \dot{w}$ is the transverse displacement velocity of a beam central axis (velocity of deflection), β is the angular velocity of a cross-section about the z-axis which is perpendicular to the plane of flexure $y - z$ (the axes z and y are directed along the beam axis and vertically down, respectively), E_∞ and μ_∞ are the nonrelaxed magnitudes of the elastic and shear moduli corresponding to the elastic beam, respectively, ρ is the density, K is the shear coefficient, A and I are the cross-sectional area and the moment of inertial with respect to the z-axis, respectively, and an overdot denotes the time derivative.

To equations (1) to (4), one should add equations describing the dynamic behavior of the elastic rod (impactor)

$$\frac{\partial \sigma}{\partial z} = \rho_{\rm im}\dot{v}, \tag{5}$$

$$\dot{\sigma} = E_{\rm im}\,\frac{\partial v}{\partial z}, \tag{6}$$

where σ is the stress, v is the velocity, $\rho_{\rm im}$ and $E_{\rm im}$ are the density and Young's modulus of impactor's material, respectively, as well as the equation of motion of the contact domain of the length $2\tau_{\rm im}$ (Figure 1)

$$2\tau_{\rm im} A\rho\ddot{w} = 2Q|_{z=\tau_{\rm im}} + F_{\rm cont}, \tag{7}$$

and the equation for the contact force which could be written as the fractional derivative standard linear solid constitutive relationship

$$F_{\rm cont} + \tau_\varepsilon^\gamma D^\gamma F_{\rm cont} = E_0\left[(\alpha - w) + \tau_\sigma^\gamma D^\gamma(\alpha - w)\right], \tag{8}$$

wherein α and w are the displacements of the upper and lower ends of the buffer, respectively, in so doing the displacement w is equal to the displacement of the beam in the place of contact (Figure 1), γ $(0 < \gamma \le 1)$ is the fractional parameter, τ_ε and τ_σ are the relaxation and retardation (creep) times, respectively, in so doing

$$\tau_\varepsilon^\gamma \tau_\sigma^{-\gamma} = E_0 E_\infty^{-1}, \tag{9}$$

E_∞ and E_0 are the non-relaxed (instantaneous modulus of elasticity, or the glassy modulus) and relaxed elastic (prolonged modulus of elasticity, or the rubbery modulus) moduli, respectively,

$$D^\gamma F_{\rm cont} = \frac{\rm d}{{\rm d}t}\int\limits_0^t \frac{F_{\rm cont}(t-t')}{\Gamma(1-\gamma)t'^\gamma}\,{\rm d}t' \tag{10}$$

is the Riemann-Liouville fractional derivative, and $\Gamma(1-\gamma)$ is the Gamma-function.

The above equations are subjected to the initial conditions

$$\alpha|_{t=0} = w|_{t=0} = \dot{w}|_{t=0} = 0, \quad \dot{\alpha}|_{t=0} = V_0, \tag{11}$$

as well as the boundary condition

$$\partial W/\partial z|_{z=\pm\tau_{\rm im}} = 0. \tag{12}$$

IV. Methods of Solution

To find the solution of the stated problem, two methods are used, namely: the ray method and Laplace transform technique. The ray method is applied for constructing an approximate solution within the elastic part of the beam from the surface of strong discontinuity upto the boundary of the contact region, as well as for finding the exact solution within the disturbed domain of the elastic rod. Within the contact domain, the Laplace transformation method is utilized to determine the contact force.

ISBN: 978-1-61804-251-4

A. A Ray Method for the Elastic Part of the Beam

To find the solution outward the contact region, i.e., for the elastic part of the target, we shall interpret a shock wave in the beam (surface of strong discontinuity) as a layer of the thickness δ, within which the desired function Z changes from the magnitude Z^- to the magnitude Z^+ but remaining a continuous function [2]. Then integrating equations from (1) to (4) over the layer's thickness from $-\delta/2$ to $\delta/2$, with δ tending to zero, and considering that inside the layer the condition of compatibility [7] is fulfilled in the form of

$$\dot{Z} = -\,G\,\frac{\partial Z}{\partial r} + \frac{\delta Z}{\delta t}, \tag{13}$$

where G is the normal velocity of the wave surface, and $\delta/\delta t$ is the δ-derivative with respect to time [8], we find the dynamic conditions of compatibility

$$[Q] = -\,\rho A G[W], \qquad -\,G[Q] = K\mu_\infty A[W], \tag{14}$$

$$[M] = -\,\rho I\; G[\beta], \qquad -\,G[M] = E_\infty I[\beta], \tag{15}$$

where $[Z] = Z^+ - Z^-$.

Eliminating the values $[Q]$ and $[M]$ from equations (14) and (15), respectively, we define the velocities of the quasi-transverse $G_\infty^{(2)}$ and quasi-longitudinal $G_\infty^{(1)}$ waves as follows

$$G_\infty^{(2)} = \left(\frac{K\mu_\infty}{\rho}\right)^{1/2}, \qquad G_\infty^{(1)} = \left(\frac{E_\infty}{\rho}\right)^{1/2}. \tag{16}$$

If the contact spot is considered to be a rigid body, then the values $Q \approx [Q]$ and $W \approx [W]$, which are connected by the relationship

$$Q = -\,\rho A G_\infty^{(2)} W, \tag{17}$$

are the dominating values in the vicinity of the contact spot and on its boundary [2].

B. Ray Method for the Elastic Rod

At the moment of impact of a projectile (rod) against a target (beam), the shock waves are generated not only in the beam but in the rod (a longitudinal shock wave) as well, which propagates along the rod with the velocity $G_{\rm im}$.

Using the same reasoning for determining the dynamic conditions of compatibility as we have adopted above for the elastic beam, we find

$$[\sigma] = -\,\rho_{\rm im} G_{\rm im}[v], \qquad -\,G_{\rm im}[\sigma] = E_{\rm im}[v], \tag{18}$$

whence it follows that

$$G_{\rm im} = \sqrt{\frac{E_{\rm im}}{\rho_{\rm im}}}. \tag{19}$$

Behind the front of this wave (a surface of the strong discontinuity), the relationships for the stress σ^- and velocity v^- could be obtained using the ray series [2]

$$\sigma^- = -\sum_{k=0}^{\infty} \frac{1}{k!}\;\left[\sigma_{,(k)}\right]\left(t - \frac{z}{G_{\rm im}}\right)^k, \tag{20}$$

$$v^- = V_0 - \sum_{k=0}^{\infty} \frac{1}{k!}\;\left[v_{,(k)}\right]\left(t - \frac{z}{G_{\rm im}}\right)^k, \tag{21}$$

where $\sigma_{,(k)} = \partial^k/\partial t^k$ and $v_{,(k)} = \partial^k/\partial t^k$.

Considering that the discontinuities in the elastic rod remain constant during the process of the wave propagation, and using the condition of compatibility

$$G_{\rm im}\left[\frac{\partial Z_{,(k-1)}}{\partial z}\right] = -[Z_{,(k)}],$$

which is obtained from equation (13) by substitution of the function Z with $Z_{,(k)} = \partial^k Z/\partial t^k$, we have

$$\left[\frac{\partial \sigma_{,(k-1)}}{\partial z}\right] = -G_{\rm im}^{-1}[\sigma_{,(k)}]. \tag{22}$$

With due account for (22), the equation of motion on the wave surface is written in the form

$$[\sigma_{,(k)}] = -\rho_{\rm im} G_{\rm im}[v_{,(k)}]. \tag{23}$$

Substituting (23) in (20) yields

$$\sigma^- = \rho_{\rm im} G_{\rm im} \sum_{k=0}^{\infty} \frac{1}{k!}\;\left[v_{,(k)}\right]\left(t - \frac{z}{G_{\rm im}}\right)^k. \tag{24}$$

Comparing relationships (24) and (21), we obtain

$$\sigma^- = \rho_{\rm im} G_{\rm im}(V_0 - v^-). \tag{25}$$

At $z = 0$, expression (25) takes the form

$$\sigma_{\rm cont} = \rho_{\rm im} G_{\rm im}(V_0 - W - \dot{\alpha}), \tag{26}$$

where $\sigma_{\rm cont} = \sigma^-|_{z=0}$ is the contact stress.

Using formula (26), it is possible to find the contact force

$$F_{\rm cont} = b(V_0 - \dot{w} - \dot{\alpha}), \tag{27}$$

where $b = 2\tau_{\rm im} a \rho_{\rm im} G_{\rm im}$.

C. Determination of the Contact Force by the Laplace Transform Technique

The contact force can be determined not only by formula (27) but according to the following equation [9] as well:

$$F_{\rm cont} = E_\infty(\alpha - w) - \triangle E \int_0^t \ni_\gamma\left(-\frac{t-t'}{\tau_\varepsilon}\right)[\alpha(t') - w(t')]\mathrm{d}t', \tag{28}$$

where $\triangle E = E_\infty - E_0$ is the defect of the modulus, i.e., the value characterizing the decrease in the elastic modulus from its nonrelaxed value to its relaxed value, and

$$\ni_\gamma\left(-\frac{t}{\tau_\varepsilon}\right) = \frac{t^{\gamma-1}}{\tau_\varepsilon^\gamma} \sum_{n=0}^{\infty} \frac{(-1)^n (t/\tau_\varepsilon)^{\gamma n}}{\Gamma[\gamma(n+1)]} \tag{29}$$

is the fractional exponential function suggested by Rabotnov [10].

Really, we could rewrite equation (8) in the form

$$F_{\rm cont} = E_0\,\frac{1 + \tau_\sigma^\gamma D^\gamma}{1 + \tau_\varepsilon^\gamma D^\gamma}(\alpha - w), \tag{30}$$

or with due account for formula (9) in the form

$$F_{\rm cont} = E_\infty \, \frac{E_\infty^{-1} E_0 + \tau_\varepsilon^\gamma D^\gamma}{1 + \tau_\varepsilon^\gamma D^\gamma} (\alpha - w). \tag{31}$$

Adding and subtracting the unit in the numerator of equation (31) yields

$$F_{\rm cont} = E_\infty (\alpha - w) - \triangle E \, \ni_\gamma^* \, (\tau_\varepsilon^\gamma) \, (\alpha - w), \tag{32}$$

where

$$\ni_\gamma^* \, (\tau_\varepsilon^\gamma) = \frac{1}{1 + \tau_\varepsilon^\gamma D^\gamma}$$

is the dimensionless Rabotnov operator [11].

Considering that $D^\gamma I^\gamma = 1$, we could represent the operator $\ni_\gamma^* \, (\tau_\varepsilon^\gamma)$ as

$$\ni_\gamma^* \, (\tau_\varepsilon^\gamma) = \frac{I^\gamma \tau_\varepsilon^{-\gamma}}{1 - \left(-I^\gamma \tau_\varepsilon^{-\gamma}\right)}, \tag{33}$$

where

$$I^\gamma x(t) = \int\limits_0^t \frac{(t-t')^{\gamma-1}}{\Gamma(\gamma)} \, x(t') {\rm d}t' \tag{34}$$

is the fractional integral.

If we suppose that the right part of formula (33) is the sum of an infinite decreasing geometrical progression, the denominator of which is equal to $d = -I^\gamma \tau_\varepsilon^{-\gamma}$, then $\ni_\gamma^* \, (\tau_\varepsilon^\gamma)$ could be represented as

$$\ni_\gamma^* \, (\tau_\varepsilon^\gamma) = \sum_{n=0}^{\infty} (-1)^n \tau_\varepsilon^{-\gamma(n+1)} I^{\gamma(n+1)}, \tag{35}$$

or with due account for equation (34), we find

$$\ni_\gamma^* \, (\tau_\varepsilon^\gamma) \, x(t) = \int\limits_0^t \ni_\gamma \left(-\frac{t'}{\tau_\varepsilon} \right) x(t-t') {\rm d}t'. \tag{36}$$

If we change the subtrahend in equation (32) by formula (36) with $x(t) = \alpha(t) - w(t)$, then we are led to relationship (28).

Equations (27), (28) and (7) rewritten with due account for formula (17)

$$M\ddot{w} + MB\dot{w} = F_{\rm cont}, \tag{37}$$

where $B = \tau_{\rm im}^{-1} G_\infty^{(2)}$ and $M = 2\tau_{\rm im}\rho A$ is the mass of the contact region, provide a closed set of three equations in terms of three unknowns: $F_{\rm cont}$, w, and α.

Now applying Laplace transformation to equations (37), (30), and (27), we have

$$Mp\bar{w}(p+B) = \bar{F}_{\rm cont}, \tag{38}$$

$$\bar{F}_{\rm cont} = E_0 \, \frac{1 + (p\tau_\sigma)^\gamma}{1 + (p\tau_\varepsilon)^\gamma} \, (\bar{\alpha} - \bar{w}), \tag{39}$$

$$\bar{F}_{\rm cont} = b \left(\frac{V_0}{p} - p\bar{\alpha} - p\bar{w} \right), \tag{40}$$

where a bar over a value denotes the Laplace transform the given value, and p is the Laplace variable.

Eliminating $\bar{F}_{\rm cont}$ from equations (38) and (40), we find

$$\bar{\alpha}(p) = \frac{V_0}{p^2} - \left[\frac{M}{b}(p+B) + 1 \right] \bar{w}. \tag{41}$$

Now eliminating $\bar{F}_{\rm cont}$ from equations (38) and (39) and considering (41), we obtain

$$\bar{w}(p) = \frac{V_0 \Omega_\infty^2 (\tau_\sigma^{-\gamma} + p^\gamma)}{p^2 f_\gamma(p)}, \tag{42}$$

where $\Omega_\infty^2 = E_\infty M^{-1}$, and

$$\begin{aligned} f_\gamma(p) &= p^{2+\gamma} + \tau_\varepsilon^{-\gamma} p^2 + (B + E_\infty b^{-1}) p^{1+\gamma} \\ &+ (B + E_0 b^{-1}) \tau_\varepsilon^{-\gamma} p + E_\infty (Bb^{-1} + M^{-1}) p^\gamma \\ &+ E_0 (Bb^{-1} + M^{-1}) \tau_\varepsilon^{-\gamma}. \end{aligned} \tag{43}$$

Substituting formulas (41) and (42) in (40) yields

$$\bar{F}_{\rm cont}(p) = \frac{V_0 E_\infty (p+B)(\tau_\sigma^{-\gamma} + p^\gamma)}{p f_\gamma(p)}. \tag{44}$$

Besides, it is possible to find the value $\bar{\alpha}(p)$, if we exclude the value $\bar{w}(p)$ defined by (42) from equation (41). As a result, we obtain

$$\bar{\alpha}(p) = \frac{V_0}{p^2} \left\{ 1 - \left[\frac{M}{b}(p+B) + 1 \right] \frac{\Omega_\infty^2 (\tau_\sigma^{-\gamma} + p^\gamma)}{f_\gamma(p)} \right\}. \tag{45}$$

Now we will carry out the inverse transformation of formula (44). For this purpose, first we will investigate the roots of the characteristic equation

$$f_\gamma(p) = 0. \tag{46}$$

Let us multiply equation (46) by τ_ε^γ, represent p in the geometrical form

$$p = re^{{\rm i}\psi} \tag{47}$$

and introduce a new variable $x = (r\tau_\varepsilon)^\gamma$. As a result, equation (46) could be rewritten in the form

$$\begin{aligned} & r^2 \left[x e^{i(2+\gamma)\psi} + e^{2i\psi} \right] \\ + \; & r \left[(B + E_\infty b^{-1}) x e^{i(1+\gamma)\psi} + (B + E_0 b^{-1}) e^{i\psi} \right] \\ + \; & \left(Bb^{-1} + 2M^{-1} \right) \left(E_\infty x e^{i\gamma\psi} + E_0 \right) = 0. \end{aligned} \tag{48}$$

Separating the real and imaginary parts in (48), we have

$$r^2 a_1 + r a_2 + a_3 = 0, \tag{49}$$

$$r^2 b_1 + r b_2 + b_3 = 0, \tag{50}$$

where

$$\begin{aligned} a_1 &= \cos 2\psi + x \cos(2+\gamma)\psi, \\ b_1 &= \sin 2\psi + x \sin(2+\gamma)\psi, \\ a_2 &= (B + E_0 b^{-1})[\cos\psi + x(B + E_\infty b^{-1}) \cos(1+\gamma)\psi], \\ b_2 &= (B + E_0 b^{-1})[\sin\psi + x(B + E_\infty b^{-1}) \sin(1+\gamma)\psi], \\ a_3 &= \left(Bb^{-1} + 2M^{-1} \right) (E_0 + x E_\infty \cos\gamma\psi), \\ b_3 &= \left(Bb^{-1} + 2M^{-1} \right) x E_\infty \sin\gamma\psi. \end{aligned}$$

First we fix the angle $\frac{\pi}{2} \le \psi \le \pi$ in equations (49) and (50), and then eliminate r^2. As result, we obtain

$$r = \frac{a_1 b_3 - a_3 b_1}{a_2 b_1 - a_1 b_2}. \quad (51)$$

Substituting (51) in (49) yields

$$\begin{aligned}(a_3 b_1 - a_1 b_3)^2 a_1 &- (a_2 b_1 - a_1 b_2)(a_3 b_1 - a_1 b_3) a_2 \\ &+ (a_2 b_1 - a_1 b_2)^2 a_3 = 0. \quad (52)\end{aligned}$$

From equation (52) at each fixed angle ψ from the segment $\frac{\pi}{2} \le \psi \le \pi$, we could find the values x_i $(i = 1, 2, ...)$, and then we substitute the chosen ψ with the found magnitude of x_i in equation (51), what allows us to find the corresponding module r_i $(i = 1, 2, ...)$. Knowing the values of x_i and r_i, it is possible to determine $(\tau_\varepsilon^\gamma)_i = x_i r_i^{-\gamma}$. The set of values involving the angle ψ, radii r_i, and parameters $(\tau_\varepsilon^\gamma)_i$ completely defines the roots of the characteristic equation (46).

In order to clarify the number of characteristic equation roots, we consider their asymptotic behavior.

1) The case $\tau_\varepsilon^\gamma \to 0$: Suppose that $\tau_\varepsilon^\gamma \to 0$ $(\tau_\varepsilon^{-\gamma} \to \infty)$. In this case, the characteristic equation (46) takes the form

$$f_{\gamma 0}(p_0) = p_0^2 + (B + E_0 b^{-1}) p_0 + E_0 (B b^{-1} + M^{-1}) = 0, \quad (53)$$

whence it follows that

$$\begin{aligned}p_{0i} = (p_0)_{1,2} = &- \frac{1}{2}(B + E_0 b^{-1}) \\ &\pm \frac{1}{2}\sqrt{(B - E_0 b^{-1})^2 - 8 E_0 M^{-1}}. \quad (54)\end{aligned}$$

2) The case $\tau_\varepsilon^\gamma = \varepsilon$: Now we suppose that the relaxation time of the system is a small value, i.e. $\tau_\varepsilon^\gamma = \varepsilon$, where ε is a small value. We will seek the solution of the characteristic equation (46) in the form:

$$p_i = p_{0i} + \varepsilon \chi_i, \quad (55)$$

where χ_i is yet unknown function.

Substituting (55) in (46) and ignoring the values of the order higher than ε, we find

$$\chi_i = - \frac{f_{\gamma\infty}(p_{0i})}{f'_{\gamma 0}(p_{0i})}, \quad (56)$$

where $f'_\gamma(p)$ denotes the derivative of the function $f_\gamma(p)$ with respect to p,

$$f'_\gamma(p_{0i}) = 2 p_{0i} + B + E_0 b^{-1},$$

$$f_{\gamma\infty} = p_{0i}^{2+\gamma} + (B + E_\infty b^{-1}) p_{0i}^{1+\gamma} + E_\infty (B b^{-1} + 2 M^{-1}) p_{0i}^\gamma.$$

3) The case $\tau_\varepsilon^\gamma \to \infty$: Suppose that $\tau_\varepsilon^\gamma \to \infty$ $(\tau_\varepsilon^{-\gamma} \to 0)$. In this case, the characteristic equation (46) takes the form

$$\begin{aligned}f_{\gamma\infty}(p_\infty) &= p_\infty^{2+\gamma} + (B + E_\infty b^{-1}) p_\infty^{1+\gamma} \\ &+ E_\infty (B b^{-1} + 2 M^{-1}) p_\infty^\gamma = 0. \quad (57)\end{aligned}$$

From equation (57) we find

$$\begin{aligned}p_{\infty i} = (p_\infty)_{1,2} = &- \frac{1}{2}(B + E_\infty b^{-1}) \\ &\pm \frac{1}{2}\sqrt{(B - E_\infty b^{-1})^2 - 8 E_\infty M^{-1}}. \quad (58)\end{aligned}$$

4) The case $\tau_\varepsilon^{-\gamma} = \varepsilon$: Now we suppose that the relaxation time of the system is a large value, i.e. $\tau_\varepsilon^{-\gamma} = \varepsilon$, where ε is a small value. We will seek the solution of the characteristic equation (46) in the form:

$$p_i = p_{\infty i} + \varepsilon \eta_i. \quad (59)$$

Substituting (59) in equation (46) and ignoring the values of the order higher than ε, we find

$$\eta_i = - \frac{f_{\gamma 0}(p_{\infty i})}{f'_{\gamma\infty}(p_{\infty i})}, \quad (60)$$

where

$$f_{\gamma 0}(p_{\infty i}) = p_{\infty i}^2 + (B + E_0 b^{-1}) p_{\infty i} + E_0 (B b^{-1} + 2 M^{-1}),$$

$$\begin{aligned}f'_{\gamma\infty}(p_{\infty i}) = &(2 + \gamma) p_{\infty i}^{1+\gamma} + (B + E_\infty b^{-1})(1 + \gamma) p_{\infty i}^\gamma \\ &+ (B b^{-1} + 2 M^{-1}) \gamma p_{\infty i}^{\gamma - 1}.\end{aligned}$$

On the ground of the above asymptotic formulas, it could be assumed that the characteristic equation (46) possesses two complex conjugate roots, which we will represent in the following form:

$$p_{1,2} = r e^{\pm \mathrm{i}\psi} = -\alpha \pm \mathrm{i}\omega. \quad (61)$$

Further it is convenient to rewrite $\bar{F}_{\mathrm{cont}}(p)$ defined by (44) in the form

$$\bar{F}_{\mathrm{cont}}(p) = \frac{1}{p}\, \bar{F}_0(p), \quad (62)$$

where

$$\bar{F}_0(p) = V_0\, \frac{g_\gamma(p)}{f_\gamma(p)}, \quad (63)$$

$$g_\gamma(p) = E_\infty p^{1+\gamma} + E_0 \tau_\varepsilon^{-\gamma} p + E_\infty B p^\gamma + E_0 \tau_\varepsilon^{-\gamma} B.$$

The function $F_0(t)$ in the time domain is governed by the Mellin-Fourier inversion formula

$$F_0(t) = \frac{1}{2\pi\mathrm{i}} \int\limits_{c - \mathrm{i}\infty}^{c + \mathrm{i}\infty} \bar{F}_0(p) e^{pt} \mathrm{d}p. \quad (64)$$

To calculate the integral (64), it is necessary to define all singular points of the complex function $\bar{F}_{\mathrm{cont}}(p)$. This multi-valued function possesses the branch points at $p = 0$ and $p = \infty$ and the simple poles at the same magnitudes of $p = p_k$ which vanish to zero the denominator in equation (63), i.e. they are the roots of the characteristic equation (46).

The inversion theorem is applicable to multi-valued functions possessing branch points only on the first sheet of the Riemann surface, i.e. when $0 < |\arg p| < \pi$. Thus a closed contour of integration should be chosen in the form presented in Figure 2. Considering Jordan lemma and applying the main theorem of the theory of residues, we rewrite the integral (64) in the following form:

$$\begin{aligned}F_0(t) &= \frac{1}{2\pi\mathrm{i}} \int\limits_0^\infty \left[\bar{F}_0(s e^{-\mathrm{i}\pi}) - \bar{F}_0(s e^{\mathrm{i}\pi})\right] e^{-st} \mathrm{d}s \\ &+ \sum_k \mathrm{res}\left[\bar{F}_0(p_k) e^{p_k t}\right], \quad (65)\end{aligned}$$

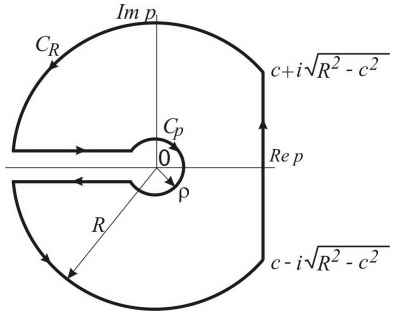


Fig. 2. Closed contour of integration

where the summation is carried out over all isolated singular points (poles).

Knowing the function $F_0(t)$, it is possible to determine the contact force $F_{\rm cont}(t)$ via the following formula:

$$F_{\rm cont}(t) = \int_0^t F_0(t')\mathrm{d}t'. \tag{66}$$

Since the roots of the characteristic equation (46) are complex conjugate ones and are defined by formula (61), then equation (65) is reduced to

$$F_0(t) = A_0(t) + A\exp(-\alpha t)\cos(\omega t + \varphi), \tag{67}$$

where

$$A_0(t) = \int_0^\infty B(s)e^{-st}\mathrm{d}s,$$

$$B(s) = \frac{(s-B)\left[Y\mathrm{Re}f_\gamma(se^{\mathrm{i}\pi}) - X\mathrm{Im}f_\gamma(se^{\mathrm{i}\pi})\right]V_0\pi^{-1}}{\left[\mathrm{Re}f_\gamma(se^{\mathrm{i}\pi})\right]^2 + \left[\mathrm{Im}f_\gamma(se^{\mathrm{i}\pi})\right]^2},$$

$$A_j = \frac{2V_0\sqrt{\left[N_1(p_j)\right]^2 + \left[N_2(p_j)\right]^2}}{\left[\mathrm{Re}f'_\gamma(p_j)\right]^2 + \left[\mathrm{Im}f'_\gamma(p_j)\right]^2}, \quad A_1 = A_2 = A,$$

$$\tan\varphi_j = \frac{\mathrm{Re}f'_\gamma(p_j)\mathrm{Re}\,g_\gamma(p_j) + \mathrm{Im}f'_\gamma(p_j)\mathrm{Im}\,g_\gamma(p_j)}{\mathrm{Im}f'_\gamma(p_j)\mathrm{Re}\,g_\gamma(p_j) - \mathrm{Re}f'_\gamma(p_j)\mathrm{Im}\,g_\gamma(p_j)},$$
$$\tan\varphi_2 = -\tan\varphi_2 = \tan\varphi,$$

$$\begin{aligned}&\mathrm{Re}f_\gamma(se^{\mathrm{i}\pi}) = \tau_\varepsilon^{-\gamma}\left\{s^2\left[(s\tau_\varepsilon)^\gamma\cos(2+\gamma)\pi + 1\right]\right.\\&+s\left[(s\tau_\varepsilon)^\gamma(B+E_\infty b^{-1})\cos(1+\gamma)\pi - (B+E_0b^{-1})\right]\\&\left.+\,(Bb^{-1}+2M^{-1})\left[E_\infty(s\tau_\varepsilon)^\gamma\cos\gamma\pi + E_0\right]\right\}\\&= \mathrm{Re}f_\gamma(se^{-\mathrm{i}\pi}),\end{aligned}$$

$$\begin{aligned}&\mathrm{Im}f_\gamma(se^{\mathrm{i}\pi}) = \tau_\varepsilon^{-\gamma}\left[s^2(s\tau_\varepsilon)^\gamma\sin(2+\gamma)\pi\right.\\&+s(s\tau_\varepsilon)^\gamma(B+E_\infty b^{-1})\sin(1+\gamma)\pi\\&\left.+\,(Bb^{-1}+2M^{-1})E_\infty(s\tau_\varepsilon)^\gamma\sin\gamma\pi\right] = -\mathrm{Im}f_\gamma(se^{-\mathrm{i}\pi}),\end{aligned}$$

$$\begin{aligned}&\mathrm{Re}f'_\gamma(p_1) = (2+\gamma)r^{1+\gamma}\cos(1+\gamma)\psi + 2r\tau_\varepsilon^{-\gamma}\cos\psi\\&+(1+\gamma)(B+E_\infty b^{-1})r\cos\gamma\psi\\&+\gamma E_\infty(Bb^{-1}+2M^{-1})r^{\gamma-1}\cos(\gamma-1)\psi\\&+(B+E_0b^{-1})\tau_\varepsilon^{-\gamma} = \mathrm{Re}f'_\gamma(p_2),\end{aligned}$$

$$\begin{aligned}&\mathrm{Im}f'_\gamma(p_1) = (2+\gamma)r^{1+\gamma}\sin(1+\gamma)\psi + 2r\tau_\varepsilon^{-\gamma}\sin\psi\\&+(1+\gamma)(B+E_\infty b^{-1})r\sin\gamma\psi\\&+\gamma E_\infty(Bb^{-1}+2M^{-1})r^{\gamma-1}\sin(\gamma-1)\psi = -\mathrm{Im}f'_\gamma(p_2),\end{aligned}$$

$$X = E_0\tau_\varepsilon^{-\gamma} + E_\infty s^\gamma\cos\pi\gamma, \qquad Y = E_\infty s^\gamma\sin\pi\gamma,$$

$$N_1(p_j) = \mathrm{Re}f'_\gamma(p_j)\mathrm{Re}\,g_\gamma(p_j) + \mathrm{Im}f'_\gamma(p_j)\mathrm{Im}\,g_\gamma(p_j),$$

$$N_1(p_1) = N_1(p_2),$$

$$N_2(p_j) = \mathrm{Im}f'_\gamma(p_j)\mathrm{Re}\,g_\gamma(p_j) - \mathrm{Re}f'_\gamma(p_j)\mathrm{Im}\,g_\gamma(p_j),$$

$$N_2(p_1) = -N_2(p_2),$$

$$\begin{aligned}&\mathrm{Re}\,g_\gamma(p_1) = E_\infty r^{1+\gamma}\cos(1+\gamma)\psi + E_0\tau_\varepsilon^{-\gamma}r\cos\psi\\&+BE_\infty r^\gamma\cos\gamma\psi + BE_0\tau_\varepsilon^{-\gamma} = \mathrm{Re}\,g_\gamma(p_2),\end{aligned}$$

$$\begin{aligned}&\mathrm{Im}\,g_\gamma(p_1) = E_\infty r^{1+\gamma}\sin(1+\gamma)\psi + E_0\tau_\varepsilon^{-\gamma}r\sin\psi\\&+BE_\infty r^\gamma\sin\gamma\psi = -\mathrm{Im}\,g_\gamma(p_2).\end{aligned}$$

The first term in equation (67) defines the drift of the equilibrium position, while the second term governs damping vibrations around the drifting equilibrium position.

According to equation (66), for determining the function $F_{\rm cont}(t)$, it is a need to integrate the function (67) over t from o to t. As a result we obtain

$$\begin{aligned}F_{\rm cont}(t) = &\int_0^\infty B(s)\left(1-e^{-st}\right)\mathrm{d}s\\&+\frac{A}{\alpha^2+\omega^2}\Big\{\alpha\sin\varphi + \omega\cos\varphi\\&-e^{-\alpha t}\left[\alpha\sin(\omega t+\varphi) + \omega\cos(\omega t+\varphi)\right]\Big\}.\end{aligned} \tag{68}$$

V. Conclusion

The impact of a long thin cylindrical plain-ended elastic rod upon an infinite isotropic rectangular prismatic beam is investigated for the case when the viscoelastic features of the beam represent themselves only in the place of contact as a result of changes of target's microstructure during the process of contact interaction and are governed by the standard linear solid model with fractional derivatives. Out of the contact domain the target remains elastic with the non-relaxed magnitude of the elastic modulus. Due to the short duration of contact interaction, the reflected waves are not taken into account. In other words, it is assumed that the impactor will bounce from the target before the reflected waves have a time to reach the place of contact. The problem of determining the contact force is a quasi-linear one, and the Laplace transform technique has been used for its analytical solution.

Acknowledgment

The research described in this publication has been supported by the international project from the Russian Foundation for Basic Research No.14-08-92008-HHC-a and Taiwan National Science Council No. NSC 103-2923-E-011-002-MY3.

REFERENCES

[1] S. Abrate, "Modeling of impacts on composite structures," *Composite Structures*, vol. 51, pp. 129–138, 2001.

[2] Yu. A. Rossikhin and M. V. Shitikova, "Transient response of thin bodies subjected to impact: Wave approach," *Shock and Vibration Digest*, vol. 39, pp. 273–309, 2007.

[3] Yu. A. Rossikhin and M. V. Shitikova, "Application of fractional calculus for dynamic problems of solid mechanics: Novel trends and recent results," *Applied Mechanics Reviews*, vol. 63(1), pp. 010801-1–52, 2010.

[4] Yu. A. Rossikhin and M. V. Shitikova, "Two approaches for studying the impact response of viscoelastic engineering systems: An overview," *Computers and Mathematics with Applications*, vol. 66, pp. 755–773, 2013.

[5] D. Valério, J. T. Machado and V. Kiryakova, "Some pioneers of the applications of fractional calculus," *Fractional Calculus and Applied Analysis*, vol. 17(2), pp. 552–578, 2014.

[6] Yu. A. Rossikhin and M. V. Shitikova, "The impact of an elastic bodies upon a Timoshenko beam," in *Proceedings of the IFIP W67.2 on Modelling and Optimization of Distributed Parameter Systems with Application to Engineering* (K. Malanowski, Z. Nahorski and M. Peszynska, eds.), Warsaw, Poland, June 1995, London: Chapman & Hall, pp. 370-374, 1996.

[7] Yu. A. Rossikhin and M. V. Shitikova, "The ray method for solving boundary problems of wave dynamics for bodies having curvilinear anisotropy," *Acta Mechanica*, vol. 109, 49–64, 1995.

[8] T. Y. Thomas, *Plastic Flow and Fracture in Solids*. New York: Academic Press, 1961.

[9] Yu. A. Rossikhin and M. V. Shitikova, "The analysis of the impact response of a thin plate via fractional derivative standard linear solid model," *Journal of Sound and Vibration*, vol. 330, pp. 1985–2003, 2011.

[10] Yu. N. Rabotnov, "Equilibrium of an elastic medium with after-effect" (in Russian), *Prikladnaya Matematika i Mekhanika*, vol. 12(1), 53–62, 1948 (English translation of this paper could be found in *Fractional Calculus and Applied Analysis*, vol. 17, no. 3, pp. 684–696, 2014; DOI: 10.2478/s13540-014-0193-1).

[11] Yu. A. Rossikhin and M. V. Shitikova, "Centennial jubilee of Academician Rabotnov and contemporary handling of his fractional operator," *Fractional Calculus and Applied Analysis*, vol. 17(3), pp. 675–683, 2014.

ISBN: 978-1-61804-251-4

Robust normal two-armed bandit and parallel data processing

Alexander V. Kolnogorov

Abstract – According to the main theorem of the theory of games, we search minimax strategy and minimax risk for the two-armed bandit problem as Bayes' ones corresponding to the worst prior distribution. Incomes are assumed to be normally distributed with unit variances and mathematical expectations depending on currently chosen actions only. In this case, asymptotically the worst prior distribution is symmetric and asymptotically uniform one. We obtain invariant integro-difference equation for recurrent calculation of minimax strategy and minimax risk. In the limiting case, when the horizon of the control goes to infinity we obtain the second order partial differential equation. Results can be applied to systems with parallel data processing. The usual approach is to process data sequentially, one by one. However, if the problem is considered in minimax setting, it turned out that the control may be implemented in parallel almost without the lack of its quality, i.e. under mild conditions minimax risks in both cases of parallel and sequential controls have close values.

Keywords – two-armed bandit problem, control in random environment, stochastic robust control, minimax and bayesian approaches, parallel processing.

I. Introduction

We consider the two-armed bandit problem (see, e.g. [1]). The name originates from the slot machine with two arms, corresponding model is considered in Section II. This is a sequential design problem. It is often considered as a control problem as well. The general setting assumes that there are two alternative actions. Each choice of any action generates a random income which distribution is fixed during control process but unknown to the person choosing arms. The goal is to maximize the total expected income by identification the most profitable action and its preferable application.

Different models for the problem have been proposed depending on their possible applications. For example, the finite automata and stochastic automata with variable structure (see, e.g. [2, 3]) were investigated in order to describe expedient behavior of biological systems. More effective control procedures for the problems of adaptive control were propose by [4, 5]. Models for optimization in economics were considered by [6] .

In above-mentioned approaches, one assumes that incomes occur and are processed sequentially, one by one. The parallel processing of incomes was first proposed in medicine. Suppose, that there are two alternative treatments with different but unknown in advance probabilities of success and, say, 1000 patients. The result of the treatment is available in a week after its assignment. Treatments of the patients sequentially would take about twenty years, but another way is to use parallel treatments. One can assign different treatments to two test groups of, say, 100 patients. A week later, one should compare the results and most effective treatment assign to the rest 800 patients. Then the process would tale two weeks. This approach is considered, e.g., by [7, 8]. The disadvantage of this approach is the small number of stages in which parallel treatments are implemented. So, the losses may be significant.

We consider the two-armed bandit problem in application to processing of a large number items of data. Two universal alternative methods of data processing are available, each method has a fixed but a priori unknown probability of successful processing. Since the best method is not known in advance, it should be estimated during the control process. The usual approach to the control is to process data sequentially, one by one. However, if the problem is considered in minimax setting, it turned out that the control may be implemented in parallel almost without the lack of its quality, i.e. under mild conditions minimax risks in both cases of parallel and sequential controls have close values. For example, $T = 30000$ items of data may be partitioned into $N = 30$ groups each containing $M = 1000$ items of data so that data in each group are processed in parallel and the results of processing are summarized. Calculations show that $N = 30$ provides a high quality of the control.

The usage of summarized results of data processing leads to the following setup of the problem. Distributions of incomes of the two-armed bandit are assumed to be normal with unknown mathematical expectations and unit variances. According to the main theorem of the theory of games minimax strategy and minimax risk are searched as Bayes ones corresponding to the worst prior distribution. In considered case, the worst prior distribution is symmetric and asymptotically uniform. This allows to use numerical methods.

The structure of the paper is the following. In section II we give the model. In section III we consider minimax and Bayesian approaches and their relation. In section IV we show that asymptotically the worst prior distribution can be chosen a symmetric and asymptotically uniform one. Recurrent equation for Bayes risk calculation over this prior distribution is given. In section V we give the invariant equation with the unit time horizon. More convenient form of this equation is obtained after extracting a singularity in its solution. Then its limiting description by the second order partial differential

[1]This work was supported in part by Project Part of the State Assignment in Field of Scientific Activity by the Ministry of Education and Science of Russian Federation, project no. 1.949.2014/K, and by Russian Foundation for Basic Research, project no. 13-01-00334-a.

ISBN: 978-1-61804-251-4

equation is given. The results of numerical experiments and Monte Carlo simulations are presented in Section VI. Some presented in the article results were published earlier in [9–11] .

II. The Model

In this section, we describe the normal two-armed bandit, give formal setup of the problem and discuss the application to parallel data processing.

A. Normal Two-Armed Bandit

Normal two-armed bandit is a slot machine with two arms. If the gambler chooses the ℓ-th arm he gets normally distributed random income with unit variance and mathematical expectation m_ℓ ($\ell = 1, 2$). The gambler has to play against the two-armed bandit N times totally (he knows this value) and his goal is to maximize (in some sense) his total expected income. Expectations m_1, m_2 are fixed at play but unknown to the gambler. In the sequel the arms are also called actions.

This problem is closely connected with a dilemma "Information vs Control". It states that the best control policy of the gambler is always to choose the arm corresponding to the largest value of m_1, m_2. However, due to the lack of the information on this arm the gambler should try them both and this diminishes his total expected income.

B. Formal Setup

Formally, let ξ_n, $n = 1, \dots, N$, be a controlled random process which values are interpreted as incomes, depend on currently chosen actions η_n only and are normally distributed with probability densities

$$f(x|m_\ell) = (2\pi)^{-1/2} \exp\left\{-(x - m_\ell)^2/2\right\},$$

if $\eta_n = \ell$ ($\ell = 1, 2$). Such two-armed bandit is described by a vector parameter $\theta = (m_1, m_2)$. Control strategy σ at a point of time n is a measurable function of the current history of the process, i.e. incomes $x^{n-1} = x_1, \dots, x_{n-1}$ and chosen actions $y^{n-1} = y_1, \dots, y_{n-1}$. Hence,

$$\Pr(\eta_n = \ell | \eta^{n-1} = y^{n-1}, \xi^{n-1} = x^{n-1}) = \sigma_\ell(y^{n-1}, x^{n-1}),$$

$\ell = 1, 2$. There is no history at $n = 1$ and, hence, it can be omitted in expressions below. The set of strategies is denoted by Σ.

Let's describe the loss function. We assume that the goal is to maximize (in some sense) the total expected income. Hence, if parameter θ is known, the optimal strategy should always prescribe to choose the action corresponding to the largest value of m_1, m_2. The total expected income would thus be equal to $N(m_1 \vee m_2)$. If parameter θ is unknown, then the function

$$L_N(\sigma, \theta) = N(m_1 \vee m_2) - \mathbb{E}_{\sigma,\theta} \left(\sum_{n=1}^{N} \xi_n \right) \qquad (1)$$

describes losses of total expected income due to incomplete information. Here $\mathbb{E}_{\sigma,\theta}$ denotes the mathematical expectation over the measure generated by a strategy σ and a parameter θ. The set of parameters is assumed to be the following

$$\Theta = \{\theta : |m_1 - m_2| \le 2c\},$$

where $0 < c < \infty$. Here condition $c < \infty$ ensures the boundedness of the loss function on Θ.

C. Why Normal Two-Armed Bandit?

We consider the problem in application to processing of large numbers of data. Assume that there are $T = NM$ items of data and two alternative methods of their processing. Probabilities of successful and unsuccessful processing depend on applied methods (actions) only, i.e. $\Pr(\xi'_t = 1 | \eta_t = \ell) = p_\ell$, $\Pr(\xi'_t = 0 | \eta_t = \ell) = 1 - p_\ell$ ($\ell = 1, 2$; $t = 1, \dots, T$). Assume that it is known that p_1, p_2 are close to p. We partition all data into N blocks each containing sufficiently large M items of data and define a process $\xi_n = (DM)^{-1/2} \sum_{t=(n-1)M+1}^{nM} \xi'_t$, $n = 1, \dots, N$, with $D = p(1 - p)$. One can see that according to the central limit theorem distributions of $\{\xi_n\}$, $n = 1, \dots, N$, are close to normal and their variances are close to 1 just like in considered setting.

Note, that data in the same block can be processed in parallel. It means that actual duration of data processing depends on the number of blocks and not on the total number of data. For example, 30000 items of data can be partitioned into 30 blocks each containing 1000 items of data and then processed in 30 stages.

Of course, the following question arises. How much do maximal losses of the total expected income grow due to such group processing? And the answer is that the maximal losses are approximately the same as if the data were optimally processed sequentially one by one, if the goal of the control is stated in minimax terms.

III. Goal of the Control

Two approaches, minimax and Bayesian, are mostly used in order to define the quality of the control. First, we'll briefly describe both of them and then state the goal of the control.

A. Minimax Approach

According to the minimax approach the minimax risk is defined as follows

$$R_N^M(\Theta) = \inf_{\Sigma} \sup_{\Theta} L_N(\sigma, \theta), \qquad (2)$$

the corresponding strategy σ^M is called the minimax strategy. Note, that minimax approach to the problem for the first time was proposed by [13]. This article caused a significant interest to considered problem. The classic object of the most of arisen articles was Bernoulli two-armed bandit which can be described by distribution

$$\Pr(\xi_n = 1 | \eta_n = \ell) = p_\ell, \quad \Pr(\xi_n = 0 | \eta_n = \ell) = 1 - p_\ell,$$

$\ell = 1, 2$. Such two-armed bandit is described by a parameter $\theta = (p_1, p_2)$ with the set of values $\Theta = \{\theta : 0 \le p_\ell \le 1; \ell = 1, 2\}$.

Let's discuss some properties of minimax approach.

ISBN: 978-1-61804-251-4

Robustness. A very good property of the minimax approach is its robustness. Namely, if the minimax strategy σ^M is applied then the following inequality holds

$$L_N(\sigma^M, \theta) \le R_N^M(\Theta),$$

for all θ and this ensures boundedness of maximal losses.

Impossibility of Direct Determination. A direct determination of the minimax risk is impossible. For Bernoulli two-armed bandit, it was shown by [14] that exact determination of the minimax risk is practically impossible already for $N > 4$. As they write, "the algebra involved becomes progressively more complicated with increasing N and seems to remain prohibitive already for N as small as 5". For the normal two-armed bandit the problem is not easier.

An Asymptotic Minimax Theorem. However, for Bernoulli two-armed bandit an asymptotic minimax theorem by [15] holds:

$$0.530 \le (DN)^{-1/2} R_N^M(\Theta) \le 0.752, \tag{3}$$

as $N \to \infty$. Here $D = 0.25$ is the maximal variance of one-step income. The estimates (3) can be easily generalized to considered normal two-armed bandit with a glance that $D = 1$ in this case.

B. Bayesian Approach

Let's define a prior distribution $\Lambda(d\theta)$ on Θ. According to the Bayesian approach the Bayes risk is defined as follows

$$R_N^B(\Lambda) = \inf_{\Sigma} \int_{\Theta} L_N(\sigma, \theta) \Lambda(d\theta), \tag{4}$$

the corresponding strategy σ^B is called the Bayes strategy.

Let's discuss some properties of Bayesian approach.

A Simple Algorithm of Finding Bayes Risk and Bayes Strategy. Bayesian approach is very popular one because it allows to write recurrent equations for calculation of both Bayes strategy and Bayes risk by a dynamic programming technique. As Berry and Fristedt write,"...it is not that researchers in bandit problems tend to "Bayesians"; rather Bayes's theorem provides a convenient mathematical formalism that allows for adaptive learning and so is an ideal tool in sequential decision problems".

No Clear Criteria for Prior Distribution Assignment. The prior distribution is often assigned for reasons of convenience of calculations. In addition, some settings allow to take into consideration expert opinions and this may help to assign the prior distribution.

C. Main Theorem of the Theory of Games

Minimax and Bayesian approaches are related by the main theorem of the theory of games. Denote by μ and Λ distributions on the sets Σ and Θ. The corresponding loss function is equal to

$$L_N(\mu, \Lambda) = \int_{\Theta} \int_{\Sigma} L_N(\sigma, \theta) \mu(d\sigma) \Lambda(d\theta).$$

According the main theorem of theory o games under mild conditions the equality holds

$$\inf_{\mu} \sup_{\Lambda} L_N(\mu, \Lambda) = \sup_{\Lambda} \inf_{\mu} L_N(\mu, \Lambda). \tag{5}$$

Since the sets $\{\mu\}$ and Σ are the same in considered case, and $\inf_\mu L_N(\mu, \Lambda) = R_N^B(\Lambda)$, $\sup_\Lambda L_N(\mu, \Lambda) = \sup_\Theta L_N(\mu, \theta)$, it follows from (5) that

$$R_N^M(\Theta) = \inf_{\Sigma} \sup_{\Theta} L_N(\sigma, \theta) = \sup_{\Lambda} R_N^B(\Lambda) = R_N^B(\Lambda^0), \tag{6}$$

where Λ^0 denotes the worst prior distribution corresponding to the maximum of Bayes risk. Equality (6) means that minimax risk is equal to Bayes risk corresponding to the worst prior distribution. And the minimax strategy is equal to corresponding Bayes strategy as well.

D. The Goal and the Method

In the sequel the problem is considered in minimax (robust) setting. According to the main theorem of the theory of games minimax risk and minimax strategy are searched as Bayes' ones calculated over the worst prior distribution.

IV. Equation for Minimax Risk and Minimax Strategy Calculation

In this section we provide standard recurrent equation for Bayes risk and Bayes strategy calculation. Then we analyze the properties of the worst prior distribution and obtain recurrent equation for Bayes risk and Bayes strategy calculation over the worst prior distribution. One can find more detail proofs of the most results in [9, 10].

A. Standard Equation for Bayes Risk and Bayes Strategy Calculation

In the sequel, we consider strategies which apply the same actions to the groups of M incomes. If incomes arise sequentially, one-by-one, these strategies allow to switch actions more rarely. If incomes arise by groups, these strategies allow their parallel processing.

Let's write standard equations for Bayes risk and Bayes strategy calculation. Denote by $f_D(x|M) = (2\pi D)^{-1/2} \exp\left\{-(x-M)^2/(2D)\right\}$ the probability density of normal distribution with mathematical expectation M and variance D. Let $\lambda(m_1, m_2)$ denote the prior distribution density on the set Θ. Let (X_1, n_1, X_2, n_2) denote the history of the process by the point of time n with n_1, n_2 total numbers of both actions application ($n_1 + n_2 = n$) and X_1, X_2 corresponding total incomes. We assume that $X_\ell = 0$ if $n_\ell = 0$. Hence, the posterior distribution density is as follows

$$\lambda(m_1, m_2 | X_1, n_1, X_2, n_2) \tag{7}$$

$$= \frac{f_{n_1}(X_1|n_1 m_1) f_{n_2}(X_2|n_2 m_2) \lambda(m_1, m_2)}{\iint\limits_{\Theta} f_{n_1}(X_1|n_1 m_1) f_{n_2}(X_2|n_2 m_2) \lambda(m_1, m_2) dm_1, dm_2}.$$

Let's put $f_n(X|nm) = 1$ if $n = 0$. In this case, (7) remains correct if $n_1 = 0$ and/or $n_2 = 0$.

ISBN: 978-1-61804-251-4

Denote by $R^B_{N-n}(\lambda; X_1, n_1, X_2, n_2)$ the Bayes risk at the final $N-n$ steps calculated over the posterior distribution density (7) (here $n = n_1 + n_2$). Denote $x^+ = \max(x, 0)$. Then

$$R^B_{N-n}(\cdot) = \min(R^{(1)}_{N-n}(\cdot), R^{(2)}_{N-n}(\cdot)), \tag{8}$$

with $R^{(1)}_0(\cdot) = R^{(2)}_0(\cdot) = 0$ and

$$R^{(1)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = \iint\limits_{\Theta} \big(M(m_2 - m_1)^+ + \mathbb{E}^{(1)} R^B_{N-n-1}(\lambda; X_1 + x, n_1 + M, X_2, n_2)\big) \times \lambda(m_1, m_2 | X_1, n_1, X_2, n_2) dm_1 dm_2, \tag{9}$$

$$R^{(2)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = \iint\limits_{\Theta} \big(M(m_1 - m_2)^+ + \mathbb{E}^{(2)} R^B_{N-n-1}(\lambda; X_1, n_1, X_2 + x, n_2 + M)\big) \times \lambda(m_1, m_2 | X_1, n_1, X_2, n_2) dm_1 dm_2, \tag{10}$$

if $n < N$. Here

$$\mathbb{E}^{(\ell)} R(x) = \int\limits_{-\infty}^{+\infty} R(x) f_M(x | M m_\ell) dx, \quad \ell = 1, 2.$$

Note, that $R^{(\ell)}_{N-n}(\cdot)$ are equal to expected losses on the final $N-n$ steps if at first the ℓ-th action is chosen M times and then the optimal control is implemented ($\ell = 1, 2$). The Bayes strategy prescribes to choose at first M steps the action corresponding to the smallest value of $R^{(1)}_{N-n}(\cdot)$, $R^{(2)}_{N-n}(\cdot)$, the choice may be arbitrary if these values are equal to each other.

B. Properties of the Worst Prior Distribution

Recall that the worst prior distribution corresponds to the maximum of Bayes risk (4). This allows to describe its properties. The arguments use the following lemmas.

Lemma 1. Bayes risk is a continuous concave function of distribution density, i.e. for any densities λ_1, λ_2 and positive real numbers α_1, α_2, such that $\alpha_1 + \alpha_2 = 1$, inequality holds

$$R^B_N(\alpha_1 \lambda_1 + \alpha_2 \lambda_2) \geq \alpha_1 R^B_N(\lambda_1) + \alpha_2 R^B_N(\lambda_2). \tag{11}$$

Proof. The property follows from inequalities

$$\begin{aligned} &R^B_N(\alpha_1 \lambda_1 + \alpha_2 \lambda_2) \\ &= \inf_{\Sigma} \int\limits_{\Theta} \left(\alpha_1 L_N(\sigma, \lambda_1) + \alpha_2 L_N(\sigma, \lambda_2)\right) d\theta \\ &\geq \alpha_1 \inf_{\Sigma} \int\limits_{\Theta} L_N(\sigma, \lambda_1) d\theta + \alpha_2 \inf_{\Sigma} \int\limits_{\Theta} L_N(\sigma, \lambda_2) d\theta \\ &= \alpha_1 R^B_N(\lambda_1) + \alpha_2 R^B_N(\lambda_2). \end{aligned}$$

Lemma 2. The following transformations $\tilde{\lambda}$ of the prior distribution density λ do not change the Bayes risk, i.e. $R^B_N(\tilde{\lambda}) = R^B_N(\lambda)$:

1) $\tilde{\lambda}^{(1)}(m_1, m_2) = \lambda(m_2, m_1)$ (for all m_1, m_2),
2) $\tilde{\lambda}^{(2)}(m_1, m_2) = \lambda(m_1 + m, m_2 + m)$ (for all m_1, m_2 and any fixed m).

Let's discuss these properties without proof. Property 1) means that Bayes risk does not change if the actions are swaped and this information is given to the gambler. Property 2) means that Bayes risk does not change if all incomes are shifted on the same value m and this information is given to the gambler.

Lemma 1 and lemma 2 allow to describe the worst prior distribution. First, this distribution density can be chosen a symmetric one, i.e. $\lambda(m_1, m_2) = \lambda(m_2, m_1)$ for all m_1, m_2. If it is not the case, let's put $\lambda^{(1)}(m_1, m_2) = (\lambda(m_1, m_2) + \lambda(m_2, m_1))/2$. Using (11) and property 1), one can see that $R^B_N(\lambda^{(1)}(m_1, m_2)) \geq R^B_N(\lambda(m_1, m_2))$ and, hence, $\lambda^{(1)}(m_1, m_2)$ can be chosen as the worst distribution density.

Similarly, if $\lambda(m_1, m_2)$ is the worst one, let's put for sufficiently large a

$$\lambda^{(2)}(m_1, m_2) = (2a)^{-1} \int\limits_{-a}^{a} \lambda(m_1 + m, m_2 + m) dm,$$

Using (11) and property 2), one can see that $R^B_N(\lambda^{(2)}(m_1, m_2)) \geq R^B_N(\lambda(m_1, m_2))$ and, hence, $\lambda^{(2)}(m_1, m_2)$ can be chosen as the worst distribution density. However, $\lambda^{(2)}(m_1, m_2)$ is more uniform than $\lambda(m_1, m_2)$ with respect to the shift of expectations (m_1, m_2).

In the sequel it is convenient to modify parameterization. Let $m_1 = u + v$, $m_2 = u - v$, then $\theta = (u + v, u - v)$ and $\Theta = \{\theta : |v| \leq c\}$. Taking into account that $|\partial(m_1, m_2)/\partial(u, v)| = 2$, one obtains that a prior distribution density is equal to $\nu(u, v) = 2\lambda(u + v, u - v)$. The prior distribution density which asymptotically does not decrease the Bayes risk can be chosen as follows

$$\nu_a(u, v) = \kappa_a(u) \rho(v), \tag{12}$$

where $\kappa_a(u)$ is the uniform density on the interval $|u| \leq a$, $\rho(v)$ is a symmetric density (i.e. $\rho(-v) = \rho(v)$) on the interval $|v| \leq c$ and $a \to \infty$.

C. Equations for Bayes Risk and Bayes Strategy Calculation over the Worst Prior Distribution

Let's consider a strategy which at the initial $2M_0$ steps applies both actions by turn receiving incomes X_1 X_2. At the rest horizon the strategy applies the same actions to groups of M incomes. The values $X = (X_1 + X_2)/2$, $Y = (X_1 - X_2)/2$ can be used for the posterior probability density calculation:

$$\nu_a(u, v | X, Y) = \kappa_a(u | X) \rho(v | Y) \tag{13}$$

with

$$\kappa_a(u | X) = \frac{f_{M_0/2}(X | M_0 u) \kappa_a(u)}{p(X)},$$

$$p(X) = \int\limits_{-\infty}^{\infty} f_{M_0/2}(X | M_0 u) \kappa_a(u) du$$

ISBN: 978-1-61804-251-4

and

$$\rho(v|Y) = \frac{f_{M_0/2}(Y|M_0 v)\rho(v)}{q(Y)},$$

$$q(Y) = \int\limits_{-\infty}^{\infty} f_{M_0/2}(Y|M_0 v)\rho(v)dv.$$

One can see that if X and u are fixed, then $\kappa_a(u)/p(X) \to 1$ as $a \to \infty$. Hence, $\nu_a(u,v|X,Y) \to \mu(u,v|X,Y)$ with

$$\mu(u,v|X,Y) = f_{M_0/2}(X|M_0 u)\rho(v|Y). \tag{14}$$

The following lemma is given without proof.

Lemma 3. Let $\nu_a(u,v)$ be chosen satisfying (12). Corresponding Bayes risk satisfies the equality

$$\lim_{a\to\infty} R_N^B(\nu_a(u,v)) = 4M_0 \int\limits_0^c v\rho(v)dv \tag{15}$$

$$+ \int\limits_{-\infty}^{\infty} R^B_{N-2M_0}(\mu(u,v|X,Y))q(Y)dY$$

with distribution density $\mu(u,v|X,Y)$ chosen from (14). Risks $\{R^B_{N-2M_0}(\mu(u,v|X,Y))\}$ do not depend on X.

Now let's write the dynamic programming equation for Bayes risk calculation according to (15). This equation follows from (8)–(10) if the prior distribution density is formally assumed to be constant with respect to u and this gives the true expressions for the posterior densities provided $n_1 \geq M_0$, $n_2 \geq M_0$. Note, that equations are more simple for risks

$$R_{n_1,n_2}(X_1,X_2) = R^B_{N-n}(X_1,n_1,X_2,n_2)p_{n_1,n_2}(X_1,X_2)$$

with

$$p_{n_1,n_2}(X_1,X_2) = \iint\limits_{\Theta} f_{n_1}(X_1|n_1 m_1)f_{n_2}(X_2|n_2 m_2)$$
$$\times\lambda(m_1,m_2)dm_1 dm_2.$$

Denote by $R_{n_1,n_2}(Z) = R_{n_1,n_2}(X_1,X_2)$ with $Z = X_1 n_2 - X_2 n_1$.

Theorem 1. Let $\nu_a(u,v)$ be chosen from (12) and $a \to \infty$. Then

$$R_{n_1,n_2}(\cdot) = \min(R^{(1)}_{n_1,n_2}(\cdot), R^{(2)}_{n_1,n_2}(\cdot)), \tag{16}$$

with $R^{(1)}_{n_1,n_2}(Z) = R^{(2)}_{n_1,n_2}(Z) = 0$ at $n_1 + n_2 = N$,

$$R^{(1)}_{n_1,n_2}(Z) = 2M\int\limits_0^c v g_{n_1,n_2}(Z,v)\rho(v)dv$$
$$+\frac{1}{n_2}\int\limits_{-\infty}^{+\infty} R_{n_1+M,n_2}(Z+z)h_{n_1,M}\left(\frac{MZ-n_1 z}{n_2}\right)dz, \tag{17}$$

$$R^{(2)}_{n_1,n_2}(Z) = 2M\int\limits_0^c v g_{n_1,n_2}(Z,-v)\rho(v)dv$$
$$+\frac{1}{n_1}\int\limits_{-\infty}^{+\infty} R_{n_1,n_2+M}(Z+z)h_{n_2,M}\left(\frac{MZ-n_2 z}{n_1}\right)dz \tag{18}$$

at $n_1 + n_2 < N$, $n_1 \geq M_0$, $n_2 \geq M_0$. Here

$$g_{n_1,n_2}(Z,v) = \frac{1}{(2\pi n_1 n_2(n_1+n_2))^{1/2}}$$
$$\times\exp\left(-\frac{(Z+2vn_1n_2)^2}{2n_1n_2(n_1+n_2)}\right), \tag{19}$$

$$h_{n,M}(z) = \left(\frac{n+M}{2\pi Mn}\right)^{1/2} \times \exp\left(-\frac{z^2}{2Mn(n+M)}\right). \tag{20}$$

Bayes risk (15) is calculated according to the formula

$$\lim_{a\to\infty} R_N^B(\nu_a(u,v)) = 4M_0\int\limits_0^c v\rho(v)dv + \int\limits_{-\infty}^{\infty} R_{M_0,M_0}(z)dz. \tag{21}$$

Proof. Let's check the correctness of (17). Multiplying left-hand side and right-hand side of (9) by $p_{n_1,n_2}(X_1,X_2)$ one obtains

$$R^{(1)}_{n_1,n_2}(X_1,X_2) = M\int\limits_0^c 2v g_{n_1,n_2}(X_1,X_2,v)\rho(v)dv$$
$$+\int\limits_{-\infty}^{+\infty} R_{n_1+M,n_2}(X_1+x,X_2)$$
$$\times h_{n_1,M}(MX_1 - n_1 x)dx. \tag{22}$$

Here $g_{n_1,n_2}(X_1,X_2,v)$ should be calculated under assumption that the prior distribution density is constant for all $u \in (-\infty,+\infty)$, i.e.

$$g_{n_1,n_2}(X_1,X_2,v)$$
$$= \int\limits_{-\infty}^{+\infty} f_{n_1}(X_1|n_1(u+v))f_{n_2}(X_2|n_2(u-v))du$$
$$= \frac{\exp\left(-\frac{(\overline{X}_1 - \overline{X}_2 + 2v)^2}{2(n_1^{-1}+n_2^{-1})}\right)}{(2\pi n_1 n_2(n_1+n_2))^{1/2}},$$

where $\overline{X}_\ell = X_\ell/n_\ell$, $\ell = 1,2$. Function $h_{n_\ell}(X_\ell - n_\ell x)$ does not depend on the prior distribution:

$$h_{n_\ell,M}(MX_\ell - n_\ell x)$$
$$= \Big(\iint\limits_{\Theta} f_M(x|Mm_\ell)f_{n_\ell}(X_\ell|n_\ell m_\ell)$$
$$\times f_{n_{\overline{\ell}}}(X_{\overline{\ell}}|n_{\overline{\ell}}m_{\overline{\ell}})\lambda(m_1,m_2)dm_1dm_2\Big)$$
$$\Big/\Big(\iint\limits_{\Theta} f_{n_\ell+M}(X_\ell+x|(n_\ell+M)m_\ell)$$
$$\times f_{n_{\overline{\ell}}}(X_{\overline{\ell}}|n_{\overline{\ell}}m_{\overline{\ell}})\lambda(m_1,m_2)dm_1dm_2\Big)$$
$$= \frac{f_M(x|Mm_\ell)f_{n_\ell}(X_\ell|n_\ell m_\ell)}{f_{n_\ell+M}(X_\ell+x|(n_\ell+M)m_\ell)}$$
$$= \left(\frac{n_\ell+M}{2\pi Mn_\ell}\right)^{1/2}\exp\left(-\frac{(MX_\ell - n_\ell x)^2}{2Mn_\ell(n_\ell+M)}\right).$$

Obviously, $g_{n_1,n_2}(X_1,X_2)$, $h_{n_\ell,M}(MX_\ell - n_\ell x)$ correspond to those given in (19), (20). Then, note that at $n_1 + M$, n_2 the value Z is recalculated by expression $Z \leftarrow (X_1 + x)n_2 - X_2(n_1+M) = Z + z$ with $z = xn_2 - X_2 M$. Noting that $MX_1 - n_1 x = n_2^{-1}(MZ - n_1 z)$ and changing the integration variable in (22) from x to z one obtains (17).

Equation (18) is proved in the similar way. Equality (21) follows from (15) if one takes into account that $R_{M_0,M_0}(z) = R^B_{N-2M_0}(\mu(u,v|X,Y))q(Y)$.

V. Invariant Equation and Limiting Description

In this section, we give invariant form of recurrent equation with unit time horizon. The disadvantage of this equation is that its solution is singular at $t = 0$. However, this singularity can be easily extracted. Then we give the limiting description of recurrent integro-difference equation by the second order partial differential equation. Some of results are presented in [11, 12].

A. Invariant Equation with Unit Time Horizon

Now let's write equation in invariant form with unit time horizon. Denote $S = ZN^{-3/2}$, $s = zN^{-3/2}$, $t_1 = n_1N^{-1}$, $t_2 = n_2N^{-1}$, $t = nN^{-1}$, $w = vN^{1/2}$, $c = CN^{-1/2}$, $\varepsilon = MN^{-1}$, $\varepsilon_0 = M_0N^{-1}$, and $r_\varepsilon(S,t_1,t_2) = NR_{n_1,n_2}(Z)$, $r^{(\ell)}_\varepsilon(S,t_1,t_2) = NR^{(\ell)}_{n_1,n_2}(Z)$, $\varrho(w) = N^{-1/2}\rho(v)$.

Let's consider the set of parameters $\Theta' : \{(m_1,m_2) : |m_1 - m_2| \le 2cN^{-1/2}\}$ which describes the set of close distributions.

Theorem 2. Let $\nu_a(u,v)$ be chosen from (12) and $a \to \infty$. Then the following equation holds

$$r_\varepsilon(S,t_1,t_2) = \min(r^{(1)}_\varepsilon(S,t_1,t_2), r^{(2)}_\varepsilon(S,t_1,t_2)), \tag{23}$$

where $r^{(1)}_\varepsilon(S,t_1,t_2) = r^{(2)}_\varepsilon(S,t_1,t_2) = 0$ if $t_1 + t_2 = 1$,

$$\begin{gathered} r^{(1)}_\varepsilon(S,t_1,t_2) = \varepsilon g^{(1)}(S,t_1,t_2) \\ +t_2^{-1}\int\limits_{-\infty}^{+\infty} r_\varepsilon(S+s,t_1+\varepsilon,t_2)h_\varepsilon\left(\frac{S\varepsilon - t_1 s}{t_2},t_1\right)ds, \\ r^{(2)}_\varepsilon(S,t_1,t_2) = \varepsilon g^{(2)}(S,t_1,t_2) \\ +t_1^{-1}\int\limits_{-\infty}^{+\infty} r_\varepsilon(S+s,t_1,t_2+\varepsilon)h_\varepsilon\left(\frac{S\varepsilon - t_2 s}{t_1},t_2\right)ds \end{gathered} \tag{24}$$

if $t_1 + t_2 < 1$, $t_1 \ge \varepsilon_0 \;\; t_2 \ge \varepsilon_0$. Here

$$\begin{gathered} g^{(\ell)}(S,t_1,t_2) = \int\limits_0^C 2wg(S,(-1)^{\ell+1}w,t_1,t_2)\varrho(w)dw, \\ g(S,w,t_1,t_2) = (2\pi t_1t_2(t_1+t_2))^{-1/2} \\ \times\exp\left(-\frac{(S+2wt_1t_2)^2}{2t_1t_2(t_1+t_2)}\right), \\ h_\varepsilon(s,t) = \left(\frac{t+\varepsilon}{2\pi t\varepsilon}\right)^{1/2}\exp\left(-\frac{s^2}{2t\varepsilon(t+\varepsilon)}\right). \end{gathered}$$

The Bayes risk (4) corresponding to asymptotically the worst prior distribution $\nu_a(u,v)$ is equal to

$$\begin{gathered} \lim_{a\to\infty} R^B_N(\nu_a(u,v)) = N^{1/2} \\ \times\left(4\varepsilon_0\int\limits_0^C w\varrho(w)dw + \int\limits_{-\infty}^{\infty} r_\varepsilon(s,\varepsilon_0,\varepsilon_0)ds\right). \end{gathered} \tag{25}$$

Proof. Making above-mentioned substitution of variables one obtains $g(S,w,t_1,t_2) = N^{3/2}g_{n_1,n_2}(Z,v)$, $g^{(\ell)}(S,t_1,t_2) = N^2g^{(\ell)}_{n_1,n_2}(Z)$, $h_\varepsilon(s,t) = N^{1/2}h_{n,M}(z)$, $\varepsilon g^{(\ell)}(S,t_1,t_2) = NMg^{(\ell)}_{n_1,n_2}(Z)$, $t_\ell^{-1}\int_{-\infty}^{\infty} r_\varepsilon(\cdot)h_\varepsilon(\cdot)\,ds = Nn_\ell^{-1}\int_{-\infty}^{\infty} R(\cdot)h(\cdot)dz$ and, hence, (23), (24) follow from (17), (18). Since $4\varepsilon_0\int_0^C w\varrho(w)dw + \int_{-\infty}^{\infty} r_\varepsilon(s,\varepsilon_0,\varepsilon_0)ds = N^{-1/2}\left(4M_0\int_0^c v\rho(v)dv + \int_{-\infty}^{\infty} R_{M_0,M_0}(z)dz\right)$, then (25) follows from (21).

B. Extracting the Singularity from the Solution

A disadvantage of the solution to equation (23), (24) is that it is a singular one if $t_1 = 0$ or $t_2 = 0$. However, this singularity is easily extracted in considered case. Denote by $f_D(x) := (2\pi D)^{-1/2}\exp(-x^2/(2D))$ a probability density of normal distribution. Let's also denote $t := t_1 + t_2$. The following theorem holds.

Theorem 3. Let $r_\varepsilon(s,t_1,t_2)$ be a solution to integro-difference equation (23), (24). Then $r_\varepsilon(s,t_1,t_2)$ can be represented as

$$\begin{gathered} r_\varepsilon(s,t_1,t_2) = f_{t_1t_2t}(s)\mathrm{r}_\varepsilon(st^{-1},t_1,t_2), \\ \mathrm{r}_\varepsilon(st^{-1},t_1,t_2) = \min_{\ell=1,\,2}\mathrm{r}^{(\ell)}_\varepsilon(st^{-1},t_1,t_2), \end{gathered} \tag{26}$$

where $\mathrm{r}^{(1)}_\varepsilon(u,t_1,t_2) = \mathrm{r}^{(2)}_\varepsilon(u,t_1,t_2) = 0$ if $t_1 + t_2 = 1$ and then

$$\begin{gathered} \mathrm{r}^{(1)}_\varepsilon(u,t_1,t_2) = \varepsilon\mathrm{g}^{(1)}(u,t_1,t_2) \\ +\int\limits_{-\infty}^{\infty}\mathrm{r}_\varepsilon(x,t_1+\varepsilon,t_2)f_{\varepsilon t_2^2t^{-1}(t+\varepsilon)^{-1}}(u-x)dx, \\ \mathrm{r}^{(2)}_\varepsilon(u,t_1,t_2) = \varepsilon\mathrm{g}^{(2)}(u,t_1,t_2) \\ +\int\limits_{-\infty}^{\infty}\mathrm{r}_\varepsilon(x,t_1,t_2+\varepsilon)f_{\varepsilon t_1^2t^{-1}(t+\varepsilon)^{-1}}(u-x)dx \end{gathered} \tag{27}$$

if $t_1 + t_2 < 1$. Here

$$\begin{gathered} \mathrm{g}^{(\ell)}(u,t_1,t_2) \\ = \int\limits_0^c 2w\mathrm{g}(u,(-1)^{\ell+1}w,t_1,t_2)\varrho(w)dw, \\ \tilde{\mathrm{g}}^{(\ell)}(s,t_1,t_2) \\ = \int\limits_0^c 2w\tilde{\mathrm{g}}(s,(-1)^{\ell+1}w,t_1,t_2)\varrho(w)dw, \\ \mathrm{g}(u,w,t_1,t_2) = \tilde{\mathrm{g}}(ut,w,t_1,t_2), \\ \tilde{\mathrm{g}}(s,w,t_1,t_2) = g(s,w,t_1,t_2)/f_{t_1t_2t}(s),\;\ell = 1,2. \end{gathered} \tag{28}$$

Proof. First, let's check up the following equality

$$\begin{gathered} r_\varepsilon(s,t_1,t_2) = f_{t_1t_2t}(s)\tilde{\mathrm{r}}_\varepsilon(s,t_1,t_2), \\ \tilde{\mathrm{r}}_\varepsilon(s,t_1,t_2) = \min_{\ell=1,\,2}\tilde{\mathrm{r}}^{(\ell)}_\varepsilon(s,t_1,t_2), \end{gathered} \tag{29}$$

where $\tilde{\mathrm{r}}^{(1)}_\varepsilon(s,t_1,t_2) = \tilde{\mathrm{r}}^{(2)}_\varepsilon(s,t_1,t_2) = 0$ if $t_1 + t_2 = 1$ and

$$\begin{gathered} \tilde{\mathrm{r}}^{(1)}_\varepsilon(s,t_1,t_2) = \varepsilon\tilde{\mathrm{g}}^{(1)}(s,t_1,t_2) + \int\limits_{-\infty}^{\infty}\tilde{\mathrm{r}}_\varepsilon(y,t_1+\varepsilon,t_2) \\ \times f_{\varepsilon t_2^2(1+\varepsilon t^{-1})}(y - s(1+\varepsilon t^{-1}))dy, \\ \tilde{\mathrm{r}}^{(2)}_\varepsilon(s,t_1,t_2) = \varepsilon\tilde{\mathrm{g}}^{(2)}(s,t_1,t_2) + \int\limits_{-\infty}^{\infty}\tilde{\mathrm{r}}_\varepsilon(y,t_1,t_2+\varepsilon) \\ \times f_{\varepsilon t_1^2(1+\varepsilon t^{-1})}(y - s(1+\varepsilon t^{-1}))dy, \end{gathered} \tag{30}$$

if $t_1 + t_2 < 1$.

We restrict our consideration to the first equation (30). The check is made by induction. Obviously, (29) holds if $t = 1$. Assume that it holds if $t + \varepsilon = t_1 + t_2 + \varepsilon \le 1$. Making the substitution of variables $s + x = y$ in (24) one obtains

$$r_\varepsilon^{(1)}(s,t_1,t_2) = \varepsilon g^{(1)}(s,t_1,t_2) + t_2^{-1} \times \int_{-\infty}^{\infty} r_\varepsilon(y,t_1+\varepsilon,t_2) h_\varepsilon\left(\frac{s(t_1+\varepsilon)-t_1 y}{t_2}, t_1\right) dy. \quad (31)$$

By assumption, (29) holds if $t_1 + t_2 + \varepsilon$. Note, that all functions in (31) are exponents. Since the following equalities hold for the powers of these exponents

$$\frac{y^2}{(t_1+\varepsilon)t_2(t_1+\varepsilon+t_2)} + \frac{(s(t_1+\varepsilon)-t_1 y)^2}{t_1\varepsilon(t_1+\varepsilon)t_2^2} = \frac{s^2}{t_1 t_2(t_1+t_2)} + \frac{(y - s(1+\varepsilon(t_1+t_2)^{-1}))^2}{\varepsilon t_2^2(1+\varepsilon(t_1+t_2)^{-1})}$$

and for their factors

$$\frac{1}{(t_1 t_2(t_1+t_2))^{1/2}} = t_2^{-1}\frac{1}{((t_1+\varepsilon)t_2(t_1+\varepsilon+t_2))^{1/2}} \times \left(\frac{t_1+\varepsilon}{t_1\varepsilon}\right)^{1/2} (\varepsilon t_2^2(1+\varepsilon t^{-1}))^{1/2},$$

then (29) and (30) follow from (23), (31).

Now let's put $u = st^{-1}$, $\mathrm{r}_\varepsilon(u,t_1,t_2) = \tilde{\mathrm{r}}_\varepsilon(ut,t_1,t_2)$ and $\mathrm{r}_\varepsilon^{(\ell)}(u,t_1,t_2) = \tilde{\mathrm{r}}_\varepsilon^{(\ell)}(ut,t_1,t_2)$, $\ell = 1,2$, and let's check up that (26) and (27) follow from (29) and (30). Let's check up the first equality (27). In the first equality (30), let's put $u = st^{-1}$, $x = y(t+\varepsilon)^{-1}$, $\mathrm{r}_\varepsilon(u,t_1,t_2) = \tilde{\mathrm{r}}_\varepsilon(s,t_1,t_2)$ and $\mathrm{r}_\varepsilon(x,t_1+\varepsilon,t_2) = \tilde{\mathrm{r}}_\varepsilon(y,t_1+\varepsilon,t_2)$. Then

$$\mathrm{r}_\varepsilon^{(1)}(u,t_1,t_2) = \varepsilon \mathrm{g}^{(1)}(u,t_1,t_2) + (t+\varepsilon) \times \int_{-\infty}^{\infty} \mathrm{r}_\varepsilon(x,t_1+\varepsilon,t_2) f_{\varepsilon t_2^2(1+\varepsilon t^{-1})}((t+\varepsilon)(x-u))dx = \varepsilon \mathrm{g}^{(1)}(u,t_1,t_2) + \int_{-\infty}^{\infty} \mathrm{r}_\varepsilon(x,t_1+\varepsilon,t_2) f_{\varepsilon t_2^2 t^{-1}(t+\varepsilon)^{-1}}(x-u)dx.$$

One can check up the second equality (27) in the similar way. Theorem is proved.

C. Limiting Description of Invariant Equation

Let's assume that $\mathrm{r}_\varepsilon(u,t_1,t_2)$ has continuous partial derivatives of proper orders and show that equations (23) may be reduced to the form

$$\begin{aligned} \mathrm{r}_\varepsilon^{(1)}(u,t_1,t_2) &= \mathrm{r}_\varepsilon(u,t_1+\varepsilon,t_2) \\ &+\frac{\varepsilon t_2^2}{2t(t+\varepsilon)}\cdot\frac{\partial^2 \mathrm{r}_\varepsilon(u,t_1+\varepsilon,t_2)}{\partial u^2} \\ &+\varepsilon \mathrm{g}^{(1)}(u,t_1,t_2) + o(\varepsilon), \\ \mathrm{r}_\varepsilon^{(2)}(u,t_1,t_2) &= \mathrm{r}_\varepsilon(u,t_1,t_2+\varepsilon) \\ &+\frac{\varepsilon t_1^2}{2t(t+\varepsilon)}\cdot\frac{\partial^2 \mathrm{r}_\varepsilon(u,t_1,t_2+\varepsilon)}{\partial u^2} \\ &+\varepsilon \mathrm{g}^{(2)}(u,t_1,t_2) + o(\varepsilon). \end{aligned} \quad (32)$$

Let's check up the first equation (32). For this purpose we present $\mathrm{r}_\varepsilon(u-x,t_1+\varepsilon,t_2)$ as follows:

$$\mathrm{r}_\varepsilon(u-x,\cdot) = \mathrm{r}_\varepsilon(u,\cdot) - x\cdot\frac{\partial \mathrm{r}_\varepsilon(u,\cdot)}{\partial u} + \frac{x^2}{2}\cdot\frac{\partial^2 \mathrm{r}_\varepsilon(u,\cdot)}{\partial u^2} + o(x^2). \quad (33)$$

Noting that

$$\int_{-\infty}^{\infty} f_\varepsilon(x)dx = 1, \quad \int_{-\infty}^{\infty} x f_\varepsilon(x)dx = 0, \quad \int_{-\infty}^{\infty} x^2 f_\varepsilon(x)dx = \varepsilon,$$

and substituting (33) into the first equation (27), one obtains

$$\begin{aligned} &\mathrm{r}_\varepsilon^{(1)}(u,t_1,t_2) = \varepsilon \mathrm{g}^{(1)}(u,t_1,t_2) \\ &+\int_{-\infty}^{\infty} \mathrm{r}_\varepsilon(u-x,t_1+\varepsilon,t_2) f_{\varepsilon t_2^2 t^{-1}(t+\varepsilon)^{-1}}(x)dx \\ &= \varepsilon \mathrm{g}^{(1)}(u,t_1,t_2) + \mathrm{r}_\varepsilon(u,t_1+\varepsilon,t_2) \\ &+\frac{\varepsilon t_2^2}{2t(t+\varepsilon)}\cdot\frac{\partial^2 \mathrm{r}_\varepsilon(u,t_1+\varepsilon,t_2)}{\partial u^2} + o(\varepsilon), \end{aligned}$$

i.e. the first equation (32) is valid. The validity of the second equation (32) is checked up in a similar way.

In the limiting case as $\varepsilon \downarrow 0$, one obtains from (32) two differential equations for $\mathrm{r} = \mathrm{r}(u,t_1,t_2)$:

$$\frac{\partial \mathrm{r}}{\partial t_1} + \frac{t_2^2}{2t^2}\cdot\frac{\partial^2 \mathrm{r}}{\partial u^2} + \mathrm{g}^{(1)}(u,t_1,t_2) = 0 \quad \text{if } (u,t_1,t_2) \in D_1,$$

$$\frac{\partial \mathrm{r}}{\partial t_2} + \frac{t_1^2}{2t^2}\cdot\frac{\partial^2 \mathrm{r}}{\partial u^2} + \mathrm{g}^{(2)}(u,t_1,t_2) = 0 \quad \text{if } (u,t_1,t_2) \in D_2,$$

where D_1, D_2 correspond to choices of the first and the second actions respectively. The usual approach assumes that it is necessary to describe the boundary between D_1, D_2 and conditions on $\mathrm{r}(u,t_1,t_2)$ at this boundary. However, recall that equations (32) should be added by the second equation (26) which can be written in the form

$$\min_{\ell=1,2}(\mathrm{r}_\varepsilon^{(\ell)}(u,t_1,t_2) - \mathrm{r}_\varepsilon(u,t_1,t_2)) = 0,$$

and then the differential equation becomes as follows

$$\min_{\ell=1,2}\left(\frac{\partial \mathrm{r}}{\partial t_\ell} + \frac{t_\ell^2}{2t^2}\cdot\frac{\partial^2 \mathrm{r}}{\partial u^2} + \mathrm{g}^{(\ell)}(u,t_1,t_2)\right) = 0 \quad (34)$$

with initial and boundary conditions

$$\lim_{t_1+t_2\to 1-0} \mathrm{r}(u,t_1,t_2) = 0, \quad \lim_{u\to+\infty} \mathrm{r}(u,t_1,t_2) = \lim_{u\to-\infty} \mathrm{r}(u,t_1,t_2) = 0.$$

Equation (34) simultaneously describes the function $\mathrm{r}(u,t_1,t_2)$ and the sets D_1, D_2, because D_ℓ corresponds to minimum of ℓ-th member in the left-hand side of (34).

VI. Numerical Results

In this section, we give numerical results and Monte-Carlo simulations. First, we calculate minimax risk and minimax strategy as Bayes' ones corresponding to the worst prior distributions. Then we apply the strategy to parallel data processing and present Monte-Carlo simulations. Finally, we compare solutions to integro-difference and partial differential equations.

ISBN: 978-1-61804-251-4

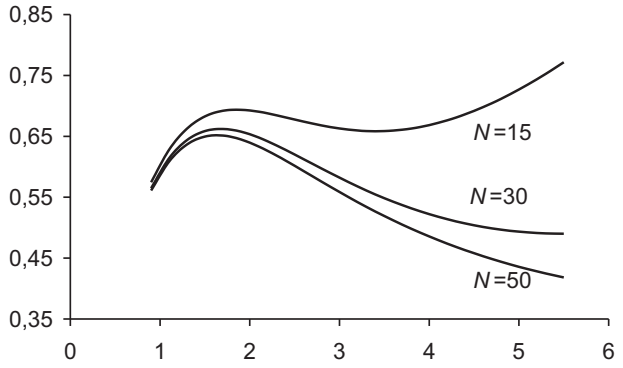


Fig. 1. Bayes risks: Determination of parameters of the worst prior distributions

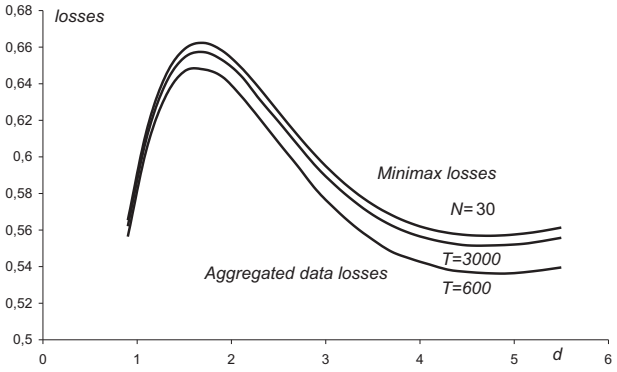


Fig. 2. Application to aggregated data control

A. Finding Minimax Risk and Minimax Strategy

Numerical optimization was made using (16)–(21) with $M_0 = M = 1$ under assumption that probability density $\rho(v)$ is concentrated at two points $v = \pm dN^{-1/2}$ with probabilities 0.5. Hence, the worst prior distribution should correspond to the maximum of the normalized Bayes risk $r_N(d) = N^{-1/2} R_N^B(\cdot)$. Calculations were implemented for $d = 0.9, 1.1, \ldots, 5.5$. Then maxima of normalized risks $r_N(d)$ and corresponding values of the argument d_N were determined. The results are presented in Table I and on Fig. 1. Two latter maxima $\max r_N(d)$ are internal ones and are in agreement with (3). The first maximum $\max r_N(d)$ is achieved at the right bound of d and does not suit to (3). This can be explained by the fact that $r_N(d)$ is a growing function of d if d is large enough. And the less value of N, the less value of d demonstrates this behavior of $r_N(d)$.

Computed Bayes strategy at $n > 2$ was the following: apply the first action if $Z > T(n_1, n_2)$ and the second action if $Z < T(n_1, n_2)$, where $\{T(n_1, n_2)\}$ is the set of determined threshold values. Then for determined Bayes strategy normalized losses $l_N(d) = N^{-1/2} L_N(\cdot)$ were computed as well. They had the same values $\max l_N(d)$ as $\max r_N(d)$ at the same points d_N. It confirms the trueness of the original assumption of probability density $\rho(v)$ structure.

TABLE I
OPTIMIZATION RESULTS

N	d_N	$\max r_N(d)$	$\max l_N(d)$
15	5.5	0.77	0.77
30	1.7	0.66	0.66
50	1.7	0.65	0.65

B. Monte-Carlo Simulations of Parallel Data Processing

Consider the following example. Let $T = 600$ packages of data be given such that two possible actions can be used for their processing. Processing is successful ($\xi_t = 1$) or unsuccessful ($\xi_t = 0$). Probabilities of successful and unsuccessful processing depend on applied actions only, i.e. $\Pr(\xi_t = 1 | \eta_t = \ell) = p_\ell$, $\Pr(\xi_t = 0 | \eta_t = \ell) = 1 - p_\ell$ ($\ell = 1, 2$). Assume that it is known that p_1, p_2 are close to $p = 0.6$. We partition all packages into $N = 30$ blocks each containing $M = 20$ packages and define a process $\xi'_n = (DM)^{-1/2} \sum_{t=(n-1)NM+1}^{nM} \xi_t$, $n = 1, \ldots, N$ with $D = p(1-p)$. Distributions of $\{\xi'_n\}$ are close to normal and their variances are close to 1.

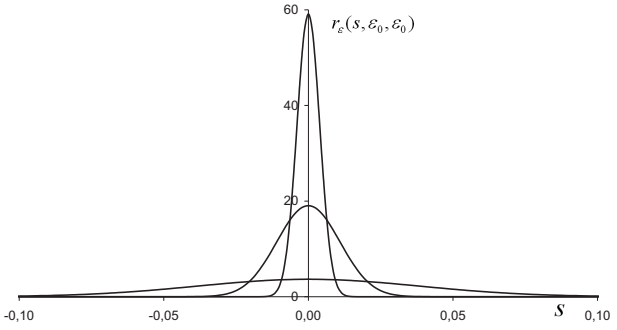


Fig. 3. Singular solutions to integro-difference equation

We use invariant equation (23)–(25) with $\varepsilon_0 = \varepsilon = 1/30$ and apply corresponding minimax strategy to $\{\xi'_n\}$. On Fig. 2 Monte-Carlo simulations of normalized losses

$$l_T(d) = (DT)^{-1/2} E_{\sigma,\theta} \left(\sum_{t=1}^{T} ((p_1 \vee p_2) - \xi_t) \right)$$

in comparison with $l_N(d)$ calculated by (23)–(25) with $\varepsilon_0 = \varepsilon = 1/30$ are presented. Parameters d and $\theta = (p_1, p_2)$ are related as $p_1 - p_2 = 2d(D/T)^{1/2}$ and p_1, p_2 are close to $p = 0.6$. One can see on Fig. 2 that $l_T(d)$ follows $l_N(d)$ and is even less than the latter. The possible explanation is that distributions of aggregated data are not sufficiently close to normal. This explanation is confirmed by the curve obtained for $T = 3000$ data which were partitioned into $N = 30$ groups each containing $M = 100$ data. This curve is closer to theoretical one.

C. Comparison of Numerical Solutions to Integro-Difference and Partial Differential Equations

Numerical solutions $r_\varepsilon(s, \varepsilon_0, \varepsilon_0)$ and $\mathrm{r}_\varepsilon(u, \varepsilon_0, \varepsilon_0)$ with $\varepsilon_0 =$ 0.02; 0.04; 0.1 are presented on Fig. 3 and Fig. 4 respectively.

ISBN: 978-1-61804-251-4

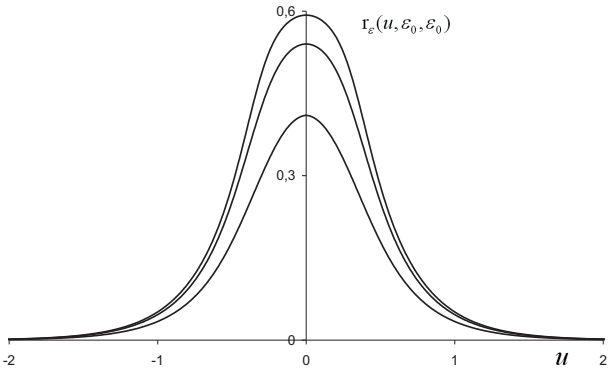


Fig. 4. Nonsingular solutions to integro-difference equation

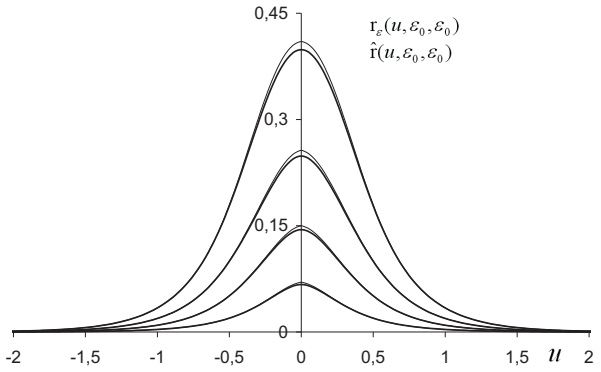


Fig. 5. Comparison of solutions to integro-difference and to differential equations

Here probability density $\varrho(w)$ is concentrated in two points $w = \pm 1.7$ with probabilities 0.5. In all cases $\varepsilon = 0.02$.

Solutions $r_\varepsilon(s, \varepsilon_0, \varepsilon_0)$ are singular if $\varepsilon_0 \downarrow 0$. On Fig. 3 singularity becomes more expressed with diminishing of ε_0. Solutions $\mathrm{r}_\varepsilon(u, \varepsilon_0, \varepsilon_0)$ are not singular. On Fig. 4 the less values of ε_0 correspond to larger values of $\mathrm{r}_\varepsilon(u, \varepsilon_0, \varepsilon_0)$.

Finally, note that Bayes risk at $\varepsilon_0 = 0.02$ calculated by (25) (using (26) in the second case) in both cases was approximately equal to 0.651. However, the calculations in the second case took time twice less than in the first case.

Numerical experiments also show the proximity of solutions to equations (26)–(27) and (26)–(32) (the second case corresponds to differential equation (34)). Partial derivatives in (32) were calculated as

$$\frac{\partial^2 \mathrm{r}(u,\cdot)}{\partial u^2} \leftarrow \frac{\mathrm{r}(u+\Delta,\cdot) - 2\mathrm{r}(u,\cdot) + \mathrm{r}(u-\Delta,\cdot)}{\Delta^2},$$

and it was always assumed that $\varepsilon < \Delta^2$ for stability of calculations.

All curves on Fig. 5 are obtained under assumption that $\varrho(w)$ is concentrated at $w = \pm 1, 7$ with probabilities 0.5. Numerical solutions to integro-difference $\mathrm{r}_\varepsilon(u, \varepsilon_0, \varepsilon_0)$ and to differential $\hat{\mathrm{r}}(u, \varepsilon_0, \varepsilon_0)$ equations for $\varepsilon_0 = 0.1$; 0.2; 0.3; 0.4, are presented by thin and thick lines respectively on Fig. 5. Less values of ε_0 correspond to larger values of functions.

REFERENCES

[1] D. A. Berry and B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London, New York, 1985.

[2] M. L. Tsetlin, *Automation Theory and Modeling of Biological Systems*. Academic Press, New York, 1973.

[3] V. I. Varshavsky, *Collective Behavior of Automata*. Nauka, Moscow, 1973. (In Russian)

[4] V. G. Sragovich, *Mathematical Theory of Adaptive Control*. Interdisciplinary Mathematical Sciences, Vol. 4. World Scientific. New Jersey, London, ..., 2006.

[5] A. V. Nazin and A. S. Poznyak, *Adaptive Choice of Alternatives*. Nauka, Moscow, 1986. (In Russian)

[6] E. L. Presman and I. M. Sonin, *Sequential Control with Incomplete Information*. Academic Press, New York, 1990.

[7] T. L. Lai, B. Levin, H. Robbins, et al., *Sequential medical trials (stopping rules/asymptotic optimality).* Proc. Nati. Acad. Sci. USA. Vol. 77, No. 6, pp. 3135–3138, 1980.

[8] J. A. Witmer, *Bayesian Multistage Decision Problems.* Ann. Statist. Vol. 14, pp. 283–297, 1986.

[9] A. V. Kolnogorov, *Determination of the Minimax Risk for the Normal Two-Armed Bandit*. In *Proceedings of the IFAC Workshop "Adaptation and Learning in Control and Signal Processing ALCOSP 2010"*, Antalya, Turkey, August 26–28, 2010. DOI 10.3182/20100826-3-TR-4015.00044. http://www.ifac-papersonline.net.

[10] A. V. Kolnogorov, *Finding minimax strategy and minimax risk in a random environment (the two-armed bandit problem).* Automation and Remote Control, Vol. 72, pp. 1017–1027, 2011.

[11] A. V. Kolnogorov, *Parallel design of robust control in the stochastic environment (the two-armed bandit problem).* Automation and Remote Control, Vol. 73, pp. 689–701, 2012.

[12] A. V. Kolnogorov, *Contributions to the limiting description of parallel design of robust control in the stochastic environment* . Automation and Remote Control, submitted for publication.

[13] H. Robbins, *Some aspects of the sequential design of experiments.* Bulletin AMS., Vol. 58(5), pp. 527–535, 1952.

[14] J. Fabius and W. R. van Zwet, *Some remarks on the two-armed bandit.* Ann. Math. Statist., Vol. 41, pp. 1906–1916, 1970.

[15] W. Vogel, *An asymptotic minimax theorem for the two-armed bandit problem.* Ann. Math. Stat., Vol. 31, pp. 444–451, 1960.

Nonlinear heat conduction problem in doubly periodic 2D composite materials

Marina Dubatovskaya[2], Gennady Mishuris[1], Sergei Rogosin[1,2]
[1]IMPACS, Aberystwyth University
Penglais, Physical Building, SY23 3BZ Aberystwyth, UK
[2]Department of Economics, Belarusian State University
4, Nezavisimosti, BY–220030 Minsk, Belarus
Email: ser14@aber.ac.uk

Abstract—An analytic solution to heat conduction problem in 2D unbounded doubly periodic composite materials with temperature dependent conductivities of its components (matrix and inclusions) is given. Linear boundary value problem for a quasi-linear differential equation is reduced to a non-linear boundary value problem for Laplace equation. By introducing complex potentials, the later is reduced to a nonlinear boundary value problem for analytic functions. This problem is investigated via application of a combination of the method of functional equations and the method of the successive approximation. Detailed description of a new algorithm for the construction of any level approximate solution to the starting problem is given.

I. Introduction

The study of the macroscopic properties of the composite materials (heterogeneous media) is of the high importance for the modern material science and technology. Different approaches for study linear inhomogeneous material are presented in well-known monographs [1], [2], [3], [4], [5]. One of the leading methods in the study of inhomogeneous media is so called homogenization method (see [6], [7]). Mathematical aspects of the higher order homogenization have been intensively developed (see, e.g. [8]). The limiting case for large (close to the maximal value) rectangular cross-section cylindrical cavities by means of an asymptotic procedure were studied in [6], simple analytical expressions for effective parameters are also found there. Non-local phenomena resulting from a high contrast (or anisotropy) of composite structures were studied in [1], [9]. In two- and three-dimensional cases the Rayleigh multipole expansions method and its generalizations is effectively used (see, e.g. [5]). Analytic type approaches were developed in [10], [11].

The main idea in linear case is to find a homogeneous structure (with so called effective properties), which represent in some sense a similar overall response as the original inhomogeneous one. to achieve this goal two main approaches can be applied. The first one consists in solving a periodic boundary value problem within separate unit (elementary) cell (see, e.g., [6], [12], [3]). The second is the direct solution of a fully periodic (doubly-periodic) boundary value problem in the whole infinite domain. The second approach has been developed only last decade (see, e.g., [13]) after progress in the areas following from the work of [14].

In the case of linear problem, one can prove (see, e.g., [10]) that the approaches lead to the same final results. However, the second approach is much more informative as it allows to construct the full solution in the composite with the compound structure, not only to evaluate its average properties.

Nonlinear composites (in this case - those with nonlinear dependence of their conductivities on the temperature) are of virtual importance not only in the study of engineering materials and structures, but also for modelling the behaviour of biological tissues. In this case, in fact, the only first approach has been used to evaluate the average properties. Note also, that in the case of linear composites, the solutions are defined with an accuracy to additive constant which is not influencing the average properties. for nonlinear problems, this should be specially addressed.

The situation in the nonlinear case is partially clarified by using asymptotic homogenization method developed in [15], [16], [17]. In particular, in [16], [17] it is proposed a Padé approximation approach for estimating of the effective behaviour of nonlinear temperature-depending composites. The key idea is to restate the homogenization problem for nonlinear composites in terms of corresponding problems for equivalent linear heterogeneous media. The quasi-linear transport equation is studied in [15]. For periodically micro-heterogeneous media, asymptotic homogenization has been performed with the local problem formulated as a minimization problem. The problem of estimation of the effective properties of special 2D composites is discussed in [18] on the base of analytic functions approach (see also [13] and [10]). Anyway the problem of understanding the macroscopic behaviour of nonlinear composite is far from completeness.

We consider the case of steady state thermal conductivity of unbounded 2D composite materials with circular inclusions geometrically formed doubly periodic structure. We suppose that each component of the composite is filled in by materials of different conductivity. Inside the components the conductivity can be changed with changing of the temperature. The

ISBN: 978-1-61804-251-4

steady state (external) flux of a given intensity is directed at certain angle and is, in general, non-parallel to the vectors of the periodic cell. The components are coupled together into unique structure due to so called ideal contact conditions.

In this case the temperature is the solution of boundary value problems for a quasi-linear differential equation corresponding to ideal contact condition. Our main problem is to determine the temperature distribution subject of this problem in an analytic form and to derive certain conclusions concerning macroscopic behaviour of the composites followed from the analysis of the obtained formulas.

The paper is organized as follows. In Section 2 we describe the geometry of the considered composites and formulate the mathematical problem basing on proper physical assumptions.

Section 3 is devoted to the reduction of the considered linear (real) problem for quasilinear equation to nonlinear (complex) problem for linear differential equation.

A novel algorithm for the solution to the nonlinear boundary value problem is proposed in Section 4. At each step the solution is reduced to the solution of linear boundary value problem by using the method of functional equations (see, e.g. [10]). The numerical analysis of this solution and related features of the nonlinear composites will be presented in forthcoming papers.

II. Statement of the problem

Let us first describe the geometry of the composites. We consider a lattice L which is defined by the two fundamental translation vectors 1 and $\imath$ (where $\imath^2$ = -1) in the complex plane $\mathbb{C} \cong \mathbb{R}^2$ of the complex variable $z = x + \imath y$). Here, the representative cell is the square

$$Q_{(0,0)} := \left\{ z = t_1 + \imath t_2 \in \mathbb{C} : -\frac{1}{2} < t_p < \frac{1}{2},\ p = 1, 2 \right\}.$$

Let $\mathcal{E} := \bigcup\limits_{m_1,m_2} \{m_1 + \imath m_2\}$ be the set of the lattice points, where $m_1, m_2 \in \mathbf{Z}$. The cells corresponding to the points of the lattice $\mathcal{E}$ are denoted by

$$Q_{(m_1,m_2)} = Q_{(0,0)} + m_1 + \imath m_2 := \\ := \left\{ z \in \mathbb{C} : z - m_1 - \imath m_2 \in Q_{(0,0)} \right\}.$$

It is considered the situation when mutually disjoint disks (inclusions) of different radii $D_k := \{z \in \mathbb{C} : |z - a_k| < r_k\}$ with boundaries $\partial D_k := \{z \in \mathbb{C} : |z - a_k| = r_k\}$ $(k = 1, 2, \ldots, N)$ are located inside the cell $Q_{(0,0)}$ and periodically repeated in all cells $Q_{(m_1,m_2)}$. We denote by

$$D_0 := Q_{(0,0)} \setminus \left(\bigcup_{k=1}^{N} D_k \cup \partial D_k \right)$$

the connected domain obtained by removing of the inclusions from the cell $Q_{(0,0)}$.

We investigate the steady state heat conduction problem for nonlinear composite materials modeling by the above described geometry, i.e. determination of a distribution of the temperature T (and/or heat flux q) in such composites. Physical assumptions are presented below.

We consider a doubly periodic composite material with matrix

$$D_{matrix} = \bigcup_{m_1,m_2} ((D_0 \cup \partial Q_{(0,0)}) + m_1 + \imath m_2)$$

and inclusions

$$D_{inc} = \bigcup_{m_1,m_2} \bigcup_{k=1}^{N} (D_k + m_1 + \imath m_2)$$

occupied by materials of conductivities $\lambda(T)$ and $\lambda_k(T)$, respectively.

The thermal loading for the composite is described by the flux given at infinity and its intensity A. We assume, that the flux is directed θ which does not coincide, in general, with the orientation of the periodic cell. According to the conservation law and the ideal (perfect) contact condition between the different materials the flux is continuous in the entire structure. Moreover, as a result of such formulation, the temperature possesses jumps across any cell. This effectively means that there are continuous isotherms crossing the composite. We choose the system of coordinate in such a way that the zero-isotherm has an internal point in $Q_{(0,0)}$ cell. It is clear that the choice is not unique, but it is not essential for the analysis.

Note, that on the base of the solution of linear problem one can assume that there exists a solution with the flux represented by a doubly periodic function within the matrix.

We assume that the conductivities $\lambda(T), \lambda_k(T) \in \mathcal{C}^\infty(\mathbb{R}), k = 1, \ldots, N$, are bounded continuous positive functions on $\mathbb{R}$ such that[1]

$$0 < \lambda^- \le \lambda(t) \le \lambda^+ < +\infty, \tag{1}$$

$$0 < \lambda_k^- \le \lambda_k(t) \le \lambda_k^+ < +\infty, \quad k = 1, \ldots, N. \tag{2}$$

Our attention will be paid to the physically substantive model case, when $\lambda(T)$ and all $\lambda_k(T)$ are continuous functions of T with emphasis on different types of behaviour. It makes sense to assume that the functions $\lambda(T)$, $\lambda_k(T)$, $k = 1, \ldots, N$, have bounded derivatives, i.e.[2]

$$|\lambda'(t)| \le \mu^+ < +\infty, \tag{3}$$

$$|\lambda_k'(t)| \le \mu_k^+ < +\infty, \quad k = 1, \ldots, N. \tag{4}$$

We search for steady-state distribution of the temperature and heat flux within the above described composite. The problem is equivalent to determination of the function $T = T(x, y) \in \mathcal{C}^2(D_{matrix} \cup D_{inc}) \cap \mathcal{C}^1(cl(D_{matrix} \cup D_{inc}))$ satisfying the quasi-linear differential equation

$$\nabla(\lambda(T)\nabla T) = 0,\ z \in D_{matrix}, \tag{5}$$

$$\nabla(\lambda_k(T)\nabla T) = 0,\ z \in \bigcup_{m_1,m_2} (D_k + m_1 + \imath m_2). \tag{6}$$

[1]Without loss of generality we can suppose that the estimates in (1), (2) are sharp, i.e. $\lambda^- = \inf_{t \in \mathbb{R}} \lambda(t)$, $\lambda^+ = \sup_{t \in \mathbb{R}} \lambda(t)$, $\lambda_k^- = \inf_{t \in \mathbb{R}} \lambda_k(t)$, $\lambda_k^+ = \sup_{t \in \mathbb{R}} \lambda_k(t)$.

[2]For determines we suppose that constants μ^+, μ_k^+ are sharp in these estimates.

ISBN: 978-1-61804-251-4

As it was already assumed, the steady state flux is directed at an angle θ to axis Ox. Besides, the heat flux is periodic in y. Thus,

$$\lambda\Big(T\Big(x,\frac{1}{2}\Big)\Big)T_y\Big(x,\frac{1}{2}\Big)=\lambda\Big(T\Big(x,-\frac{1}{2}\Big)\Big)T_y\Big(x,-\frac{1}{2}\Big)= \quad (7)$$
$$=-A\sin\theta+q_1(x),$$

where A is the intensity of an external flux. The heat flux is periodic on x too, consequently,

$$\lambda\Big(T\Big(-\frac{1}{2},y\Big)\Big)T_x\Big(-\frac{1}{2},y\Big)=\lambda\Big(T\Big(\frac{1}{2},y\Big)\Big)T_x\Big(\frac{1}{2},y\Big)= \quad (8)$$
$$=-A\cos\theta+q_2(y).$$

To complement to the flux conditions at infinity, the later immediately proves that the equalities

$$\int_{-1/2}^{1/2} q_j(\xi)d\xi=0. \quad (9)$$

are valid for the unknown functions q_j, $(j=1,2)$. As a result of (7) and (8), the heat flux has a zero mean value along the cell

$$\int\limits_{\partial\, Q_{(m_1,m_2)}} \lambda(T(s))\frac{\partial T(s)}{\partial n}ds=0. \quad (10)$$

From the physics point of view, condition (10) is the consequence of the fact that no source (sink) exists in the cells.

Finally, we assume that the ideal contact conditions on the boundaries between the matrix and the inclusions hold for all $t\in\bigcup\limits_{m_1,m_2\in\mathbf{Z}}(\partial D_k+m_1+\imath m_2)$:

$$T(t)=T_k(t), \quad (11)$$

$$\lambda(T(t))\frac{\partial T(t)}{\partial n}=\lambda_k(T_k(t))\frac{\partial T_k(t)}{\partial n}. \quad (12)$$

Here, the vector $n=(n_1,n_2)$ is the outward unit normal vector to ∂D_k; $\frac{\partial}{\partial n}=n_1\frac{\partial}{\partial x}+n_2\frac{\partial}{\partial y}$; and $T(t):=\lim\limits_{z\to t,z\in D_0}T(z)$, $T_k(t):=\lim\limits_{z\to t,z\in D_k}T(z)$.

III. Transition from quasilinear to linear equation

We use the so called Baiocchi transformation (see [19]) in order to introduce continuous increasing functions $f:I\!R\to I\!R$, $f_k:I\!R\to I\!R$, $k=1,\dots,N$,

$$f(T)=\int_0^T\lambda(\xi)\,d\xi,\quad f_k(T)=\int_0^T\lambda_k(\xi)\,d\xi,\quad k=1,\dots,N. \quad (13)$$

and to perform the following change of unknown functions:

$$u(z)=f(T(z)),\quad u_k(z)=f_k(T_k(z)),\quad k=1,\dots,N. \quad (14)$$

Remark *3.1:* Note that an assumption that the zero-isotherm has an internal point in $Q_{(0,0)}$ (and thus in D_0) can be equivalently rewritten as

$$\exists\, z_0\in D_0:\quad u(z_0)=0. \quad (15)$$

Determination of the point z_0 will be discussed below.

By using representations (13) equations (5), (6) are transformed to the Laplace equations (see, e.g., [19])

$$\Delta u(z)=0,\quad z\in D_{matrix}, \quad (16)$$

$$\Delta u_k(z)=0,\quad z\in\bigcup_{m_1,m_2\in\mathbf{Z}}D_k+m_1+\imath m_2. \quad (17)$$

The boundary conditions (7)–(8) take the form

$$u_y\Big(x,\frac{1}{2}\Big)=u_y\Big(x,-\frac{1}{2}\Big)=-A\sin\theta+q_1(x), \quad (18)$$

$$u_x\Big(\frac{1}{2},y\Big)=u_x\Big(-\frac{1}{2},y\Big)=-A\cos\theta+q_2(y). \quad (19)$$

The transmission conditions (11) and (12) along the inclusions interfaces can be rewritten as follows:

$$u(t)=f(f_k^{-1}(u_k(t))), \quad (20)$$

$$\frac{\partial u(t)}{\partial n}=\frac{\partial u_k(t)}{\partial n},\quad t\in\bigcup_{m_1,m_2\in\mathbf{Z}}(\partial D_k+m_1+\imath m_2). \quad (21)$$

Note that generally speaking the newly introduced functions u and u_k are not continuous across the interface.

Zero mean value condition for the flux (10) can be rewritten in the form:

$$\int\limits_{\partial\, Q_{(m_1,m_2)}}\frac{\partial u}{\partial n}ds=0. \quad (22)$$

Remark *3.2:* One can consider (21) as the Neumann boundary value problem for determination of unknown harmonic functions u_k in a precise domain $D_{k+m_1+\imath m_2}$ in terms of the given function u, which is harmonic in the infinitely connected domain $D_{matrix}=\mathbb{C}\setminus\bigcup\limits_{k=1}^{N}\bigcup\limits_{m_1,m_2\in\mathbf{Z}}cl(D_k+m_1+\imath m_2)$. It is known that each these problems is solvable (up to a certain constant, which we specify later) if and only if

$$\int\limits_{\partial D_k+m_1+\imath m_2}\frac{\partial u_k}{\partial n}ds=0. \quad (23)$$

Therefore, the same condition is valid for the function u along the boundaries of each inclusion

$$\int\limits_{\partial D_k+m_1+\imath m_2}\frac{\partial u}{\partial n}ds=0. \quad (24)$$

The conditions (22) – (24) have simple physical sense: they confirm that there is no source (sink) inside the composite, i.e. neither in the matrix of the composite, nor in any inclusion (the total heat flux through any closed simply connected curve is equal to zero).

Let u be a harmonic function in the infinitely connected doubly periodic domain D_{matrix}, satisfying conditions (18), (19) and (24). We introduce (new unknown auxiliary) harmonic function v in D_{matrix} which is the harmonic conjugate to u. For this pair of functions the Cauchy-Riemann equations $\frac{\partial u}{\partial x}=\frac{\partial v}{\partial y}$, $\frac{\partial u}{\partial y}=-\frac{\partial v}{\partial x}$ (or the so called normal-tangent Cauchy-Riemann equations $\frac{\partial u}{\partial n}=\frac{\partial v}{\partial s}$, $\frac{\partial u}{\partial s}=-\frac{\partial v}{\partial n}$) have to be valid.

ISBN: 978-1-61804-251-4

One of possible ways do determine the function v is the following formula

$$v(z) = \int_{z_0}^{z} -\frac{\partial u}{\partial y}dx + \frac{\partial u}{\partial x}dy + v_0. \tag{25}$$

The integration here is performed along each simple arc in the domain D_{matrix}. Since v is an auxiliary function, then we can fix in (25) the constant $v_0 = 0$. Note that, as it follows from (22) – (24) we can write

$$\int_{\partial Q_{m_1,m_2}} \frac{\partial v}{\partial s}ds = 0, \qquad \int_{\partial D_k + m_1 + \imath m_2} \frac{\partial v}{\partial s}ds = 0. \tag{26}$$

These relations yield that the harmonic function v is single-valued in the domain D_{matrix}.

Conditions (18) and (19) can be written in the following form:

$$v_x\Big(x, \frac{1}{2}\Big) - v_x\Big(x, -\frac{1}{2}\Big) = 0, \tag{27}$$

$$v_y\Big(\frac{1}{2}, y\Big) - v_y\Big(-\frac{1}{2}, y\Big) = 0. \tag{28}$$

$$v_x\Big(x, \frac{1}{2}\Big) = A\sin\,\theta - q_1(x), \tag{29}$$

$$v_y\Big(\frac{1}{2}, y\Big) = -A\cos\,\theta + q_2(y). \tag{30}$$

Integrating (27) and (28), we have, in particular, for any point on the edges of a cell

$$v\Big(x, \frac{1}{2}\Big) - v\Big(x, -\frac{1}{2}\Big) = C_1, \tag{31}$$

$$v\Big(\frac{1}{2}, y\Big) - v\Big(-\frac{1}{2}, y\Big) = C_2. \tag{32}$$

In particular, we have the following equalities at the corner points:

$$v\Big(\frac{1}{2}, \frac{1}{2}\Big) - v\Big(\frac{1}{2}, -\frac{1}{2}\Big) = C_1, \quad v\Big(-\frac{1}{2}, \frac{1}{2}\Big) - v\Big(-\frac{1}{2}, -\frac{1}{2}\Big) = C_1, \tag{33}$$

$$v\Big(\frac{1}{2}, \frac{1}{2}\Big) - v\Big(-\frac{1}{2}, \frac{1}{2}\Big) = C_2, \quad v\Big(\frac{1}{2}, -\frac{1}{2}\Big) - v\Big(-\frac{1}{2}, -\frac{1}{2}\Big) = C_2. \tag{34}$$

Integrating the equalities (29) – (30) in the interval on $\left[-\frac{1}{2}, \frac{1}{2}\right]$, with respect to variables x and y, respectively, we have:

$$v\Big(\frac{1}{2}, \frac{1}{2}\Big) - v\Big(-\frac{1}{2}, \frac{1}{2}\Big) = A\sin\,\theta, \tag{35}$$

$$v\Big(\frac{1}{2}, \frac{1}{2}\Big) - v\Big(\frac{1}{2}, -\frac{1}{2}\Big) = -A\cos\,\theta. \tag{36}$$

Therefore, the constants C_1, C_2 from (31) – (32) can be computed

$$C_1 = -A\cos\,\theta, \quad C_2 = A\sin\,\theta. \tag{37}$$

Therefore, we introduced a single-valued harmonic (harmonic conjugate to u) having constant jumps along the edges of cells. It is convenient to eliminate these jumps and to introduce (single-valued) complex potential $\varphi(z)$ in D_{matrix} by using the following formula

$$v(z) = \mathrm{Im}\,\{\varphi(z) + \alpha z\}. \tag{38}$$

with constant α defined as follows:

$$\alpha = -Ae^{-i\theta}. \tag{39}$$

Since the harmonic function u in (14) is the harmonic conjugate to v, then the following relation holds:

$$u(z) = \mathrm{Re}\,\{\varphi(z) + \alpha z\} + u_0, \tag{40}$$

with the constant u_0 satisfying the condition (15).

We note first that by construction it follows from (18) and (19), (7) and (8) that the function $\varphi(z)$ has zero-jump along the edges of cells. Moreover, integrating (7) and (8) and using connection between harmonic conjugates we can conclude that

$$\varphi(x+\imath/2) - \varphi(x-\imath/2) = 0, \qquad \varphi(1/2+\imath y) - \varphi(-1/2+\imath y) = 0.$$

Hence $\varphi(z)$ has zero-jump along each interval of unit length parallel to edges of a cell and connected points on the opposite sides of the cell. Thus, the function $\varphi(z)$ introduced in (38) is a doubly periodic one, i.e.

$$\varphi(z+1) - \varphi(z) = 0, \qquad \varphi(z+\imath) - \varphi(z) = 0.$$

The condition (40) has different form for each cell. In fact, in the central cell $Q_{(0,0)}$ we have

$$\mathrm{Re}\,\{\varphi(z_0)\} + \mathrm{Re}\,\{\alpha z_0\} + u_0 = 0. \tag{41}$$

Hence,

$$u_0 = -\mathrm{Re}\,\{\alpha z_0\} - \mathrm{Re}\,\{\varphi(z_0)\}. \tag{42}$$

In the cell $Q_{(m_1,m_2)}$

$$u(z_0 + m_1 + im_2) = \mathrm{Re}\,\{\varphi(z_0 + m_1 + im_2)\} + \tag{43}$$
$$+\mathrm{Re}\,\{\alpha\,(z_0 + m_1 + im_2)\} + u_0.$$

Since $\varphi(z)$ is doubly periodic, the values of the unknown function in the periodic copies of z_0 are equal to

$$u(z_0 + m_1 + im_2) = \mathrm{Re}\,\{\alpha\,(m_1 + im_2)\}. \tag{44}$$

The condition (43) determines an interrelation between two unknown constants, namely, between z_0 and u_0. The relation (44) shows, in particular, that an unknown harmonic function u is not periodic in the whole plane.

By using normal-tangent Cauchy-Riemann equations we can rewrite boundary condition (21) in equivalent form

$$\frac{\partial v_k(t)}{\partial s} = \frac{\partial v(t)}{\partial s}, \quad t \in \partial D_k + m_1 + \imath m_2,\ k = 1, \ldots, N, \tag{45}$$

where $v_k(z),\ z \in D_k + m_1 + \imath m_2$, are unknown auxiliary harmonic functions. Integrating (45) on s, we have for each $k = 1, \ldots, N,\ m_1, m_2 \in \mathbf{Z}$,

$$v_k(t) = v(t) + d_{(k,m_1,m_2)},\ t \in \partial D_k + m_1 + \imath m_2,\ k = 1, \ldots, N, \tag{46}$$

ISBN: 978-1-61804-251-4

where $d_{(k,m_1,m_2)}$ are undetermined constants. We can specify these constants by calculating integrals

$$\frac{1}{2\pi}\int\limits_0^{2\pi} v_k(a_k+m_1+\imath m_2+r_k e^{\imath\eta})d\eta =$$

$$=\frac{1}{2\pi}\int\limits_0^{2\pi}\left[v(a_k+m_1+\imath m_2+r_k e^{\imath\eta})\right]d\eta + d_{(k,m_1,m_2)},$$

where a_k is the center of k-th inclusion, and r_k is its radius. We get from the mean value theorem for harmonic functions (see, e.g. [20]) that

$$\frac{1}{2\pi}\int\limits_0^{2\pi} v_k(a_k+m_1+\imath m_2+r_k e^{\imath\eta})d\eta = v_k(a_k+m_1+\imath m_2).$$

Besides, it follows from the Cauchy integral formula (see, e.g. [20]), that

$$\frac{1}{2\pi}\int\limits_0^{2\pi} v(a_k+m_1+\imath m_2+r_k e^{\imath\eta})d\eta = 0.$$

As a result, we have the relation between the constants $d_{(k,m_1,m_2)}$ and the values of unknown harmonic functions v_k in centers of inclusions

$$d_{(k,m_1,m_2)} = v_k(a_k+m_1+\imath m_2). \tag{47}$$

Remark *3.3:* It follows from (47) that unknown harmonic functions v_k are different in the domains $D_k+m_1+\imath m_2$ for different m_1, m_2.

Anyway, for shortness, we keep below the notation v_k (instead of $v_{(k,m_1,m_2)}$) in all cases when it does not lead to a contradiction.

Analogously to the function $\varphi(z)$, we introduce the (single-valued) complex analytic functions $\varphi_k(z) = u_k(z)+\imath v_k(z)$ in the domains $D_k+m_1+\imath m_2$ for each $k=1,\dots,N,\ m_1,m_2\in\mathbf{Z}$. Namely, having the functions v_k known, we can than determine the analytic functions $\varphi_k(z)$ via the following relations

$$v_k(z) = \mathrm{Im}\ \{\varphi_k(z)\} + v_0^{(k,m_1,m_2)}, \tag{48}$$

and

$$u_k(z) = \mathrm{Re}\ \{\varphi_k(z)\}. \tag{49}$$

It follows from the boundary condition (20) that

$$\mathrm{Re}\,\varphi(t) = f(f_k^{-1}(\mathrm{Re}\,\varphi_k(t))) - \mathrm{Re}\,(\alpha t) - u_0.$$

From (46) and (48), we have

$$\mathrm{Im}\,\varphi(t) = \mathrm{Im}\,\varphi_k(t) - \mathrm{Im}\,(\alpha t) - d_{(k,m_1,m_2)} + v_0^{(k,m_1,m_2)} m_2)).$$

Since the auxiliary harmonic functions v_k have no influence on the solution of a starting problem (16) – (21) then it makes sense to assume

$$d_{(k,m_1,m_2)} = v_0^{(k,m_1,m_2)},\quad \forall k=1,\dots,N,\ m_1,m_2\in\mathbf{Z}. \tag{50}$$

Thus, we arrive at the following system of boundary relations ($k=1,\dots,N$):

$$\varphi(t) = f\left(f_k^{-1}\left(\frac{\varphi_k(t)+\overline{\varphi_k(t)}}{2}\right)\right) + \frac{\varphi_k(t)-\overline{\varphi_k(t)}}{2} - \alpha t - u_0. \tag{51}$$

IV. Algorithm for the solution to the nonlinear boundary value problem

In order to formulate an algorithmic procedure for the solution of the considered nonlinear boundary value problem we return to the starting form of the relation (51).

Step 1. Let us put $\lambda = \lambda^{(0)} = \lambda(0)$, $\lambda_k = \lambda_k^{(0)} = \lambda_k(0)$, $\rho_k^{(0)} = \frac{\lambda^{(0)}-\lambda_k^{(0)}}{\lambda^{(0)}+\lambda_k^{(0)}}$. Fix also the point $z_0 = z_0^{(0)}$ and corresponding value of the constant $u_0 = u_0^{(0)} = \mathrm{Re}\{\alpha z_0^{(0)}\}$.

Let us introduce new unknown functions

$$\widetilde{\varphi}(z) = \varphi(z) - \varphi_\alpha(z),\quad z\in D, \tag{52}$$

$$\widetilde{\varphi}_k(z) = \frac{\lambda^{(0)}+\lambda_k^{(0)}}{2\lambda_k^{(0)}}\cdot\left[\varphi_k(z) - \frac{2\lambda_k^{(0)}}{\lambda^{(0)}+\lambda_k^{(0)}}\varphi_{\alpha,k}(z)\right], \tag{53}$$

$$z\in D_k,\ k=1,\dots,N,$$

where $\varphi_\alpha(z)$, $\varphi_{\alpha,k}(z)$ is a solution in the space $\mathbf{X}$ to the system of linear problems ($k=1,\dots,N$)

$$\varphi_\alpha(t) = \varphi_{\alpha,k}(t) + \rho_k^{(0)}\overline{\varphi_{\alpha,k}(t)} - \alpha t - u_0^0,\ t\in\partial D_k, \tag{54}$$

$$\rho_k^{(0)} = \frac{\lambda^{(0)}-\lambda_k^{(0)}}{\lambda^{(0)}+\lambda_k^{(0)}}.$$

An analytic solution of the equation (54) is done below.

We represent boundary conditions (51) by using the form of the Taylor formula near the point $\sigma_{0,k} = f_k(T_k(a_k))$. For this we introduce N unknown constants

$$\tau_k = T_k(a_k), \tag{55}$$

which are equal to (an unknown) temperature in the centers of inclusions. These constants are subject to further determination.

In terms of such Taylor formula conditions (51) can be rewritten as

$$\widetilde{\varphi}(t) = \widetilde{\varphi}_k(t) + \rho_k(\tau_k)\overline{\widetilde{\varphi}_k(t)} + G_k\left(\widetilde{\varphi}_k(t), \overline{\widetilde{\varphi}_k(t)}, \lambda, \lambda_k, \tau_k\right)(t), \tag{56}$$

where

$$G_k\left(\widetilde{\varphi}_k(t), \overline{\widetilde{\varphi}_k(t)}, \lambda, \lambda_k, \tau_k\right)(t) =$$

$$= f(T_k(t)) - \frac{2\lambda(\tau_k)}{\lambda(\tau_k)+\lambda_k(\tau_k)}\mathrm{Re}\,[\widetilde{\varphi}_k(t)+\varphi_{\alpha,k}(t)] \tag{57}$$

$$= f(T_k(t)) - \frac{\lambda(\tau_k)}{\lambda_k(\tau_k)} f_k(T_k(t)).$$

Step 2. Substitute

$$\widetilde{\varphi}_k(z) = \widetilde{\varphi}_k^{(0)}(z) \equiv 0, \tag{58}$$

ISBN: 978-1-61804-251-4

$\tau_k = \tau_k^{(0)} = 0$, $\lambda(\tau_k) = \lambda(0) =: \lambda^{(0)}$, and $\lambda_k(\tau_k) = \lambda_k(0) =: \lambda_k^{(0)}$ into the nonlinear term G_k in (56) and put also $\rho_k = \rho_k(0) = \rho_k^{(0)}$ in (54) and (56).

Then $G_k(t)$ becomes a constant, namely

$$G_k^{(0)} = G_k(0, 0, \lambda(0), \lambda_k(0), 0),$$

and boundary condition (56) can be rewritten as

$$\widetilde{\varphi}(t) = \widetilde{\varphi}_k(t) + \rho_k(\tau_k)\overline{\widetilde{\varphi}_k(t)} + G_k^{(0)}, \quad t \in \partial\, D_k. \qquad (59)$$

In this case, the problem (56) has an analytic solution which is described in the section below.

Step 3. Therefore we obtain the first approximation of our solution $\widetilde{\varphi}^{(1)}(z), \widetilde{\varphi}_k^{(1)}(z)$. By using this approximation we calculate an approximate values $\tau_k^{(1)}$ of the temperature at the centers of inclusions and corresponding approximate values of conductivities $\lambda = \lambda(\tau_k^{(1)}) = \lambda^{(1)}$, and $\lambda_k = \lambda_k(\tau_k^{(1)}) = \lambda_k^{(1)}$, as well as $\rho_k = \rho_k(\tau_k^{(1)}) = \rho_k^{(1)}$. We also calculate $\widetilde{\varphi}_k^{(1)}(a_k), \overline{\widetilde{\varphi}_k^{(1)}(a_k)}$. Besides, we fix the constant

$$u_0 = u_0^{(1)} = -\mathrm{Re}\,\{\alpha\, z_0^{(0)}\} - \mathrm{Re}\,\{\varphi(z_0^{(0)})\}. \qquad (60)$$

Corresponding point $z_0(1)$ is chosen in this case from the relation:

$$u_0^{(1)} = -\mathrm{Re}\,\{\alpha\, z_0^{(1)}\}.$$

The point $z_0(1)$ will be used at the next stage of approximation.

Step 4. Solve in the space $\mathbf{X}$ auxiliary problem (54) with the values of parameters obtained at the step 2, namely with $\lambda = \lambda(\tau_k^{(1)}) = \lambda^{(1)}$, and $\lambda_k = \lambda_k(\tau_k^{(1)}) = \lambda_k^{(1)}$, as well as $\rho_k = \rho_k(\tau_k^{(1)}) = \rho_k^{(1)}$, and $u_0^{(1)}$.

Step 4. Next we calculate the values of constants $G_k^{(1)} = G_k(\widetilde{\varphi}_k^{(1)}(a_k), \overline{\widetilde{\varphi}_k^{(1)}(a_k)}, \lambda^{(1)}, \lambda_k^{(1)}, \tau_k^{(1)})$ and solve the problem (59) with the constant $G_k^{(1)}$ replacing $G_k^{(0)}$. This problem is of the same type but corresponds to another values of parameters and another inhomogeneous term. Hence it can be solved by using the same approach.

ACKNOWLEDGMENT

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement PIRSES-GA-2013-610547 - TAMER.

REFERENCES

[1] G. Allaire, *Shape Optimization by the Homogenization Method*. Berlin: Springer Verlag, 2002.

[2] A. Cherkaev, *Variational Methods for Structural Optimization*. New York: Springer Verlag, 2000.

[3] L. Manevitch, I. Andrianov, and V. Oshmyan, *Mechanics of periodically heteregeneous structures*, ser. Foundations of Engineering Mechanics. Berlin: Springer, 2002.

[4] G. Milton, *The Theory of Composites*, ser. Cambridge Monographs on Applied and Computational Mathematics. Cambridge: Cambridge University Press, 2002, vol. 6.

[5] A. Movchan, N. Movchan, and C. Poulton, *Asymptotic Models of Fields in Delute and Densely Packed Composites*. London: Imperial College Press, 2002.

[6] N. Bakhvalov and G. Panasenko, *Homogenisation: Averaging Processes in Periodic Media*, ser. Soviet Series. Dordrecht: Kluwer Academic Publishers, 1989, vol. 36, Mathematical Problems in the Mechanics of Composite Materials, Mathematics and Its Applications.

[7] V. Jikov, S. Kozlov, and O. Oleinik, *Homogenization of differential operators and integral functionals*. Berlin: Springer, 1994.

[8] K. Cherednichenko and V. Smyshlyaev, "On full two-scale expansion of the solutions of nonlinear periodic rapidly oscillating problems and higher-order homogenised variational problems," *Arch. Ration. Mech. Anal.*, vol. 174, pp. 385–442, 2004.

[9] K. Cherednichenko, V. Smyshlyaev, and V. Zhikov, "Non-local homogenized limits for composite media with highly anisotropic periodic fibres," *Proc. R. Soc. Edin.*, vol. A 136, pp. 87–114, 2006.

[10] V. Mityushev and S. Rogosin, *Constructive Methods for Linear and Nonlinear Boundary Value Problems for Analytic Functions. Theory and Applications*, ser. Monographs and Surveys in Pure and Applied Mathematics. Boca Raton - London: Chapman & Hall / CRC, 1999, vol. 108.

[11] Y. Obnosov, *Boundary value problems of the theory of heterogeneous media: multiphase media, separated by second order curves*. Kazan: Kazan State University, 2009, (in Russian).

[12] E. Grigolyuk and L. Filshtinskii, *Periodic piece-wise homogeneous elastic structures*. Moscow: Nauka, 1992, (in Russian).

[13] V. Mityushev, E. Pesetskaya, and S. Rogosin, "Analytical methods for heat conduction in composites and porous media," in *Thermal Properties of Cellular and Porous Materials*, A. Öchsner, G. Murch, and M. de Lemos, Eds. WILEY-VCH, 2007, pp. 124–167.

[14] V. Mityushev, "Convergence of the Poincaré series for classical Schottky groups," *Proc. Amert. Math. Soc.*, vol. 126, pp. 2399–2406, 1998.

[15] A. Galka, J. Telega, and S. Tokarzewski, "Heat equation with temperature-dependent conductivity coefficients and macroscopic properties of microheterogeneous media," *Math. and Comp. Modelling*, vol. 33, pp. 927–942, 2001.

[16] S. Tokarzewski and I. Andrianov, "Effective coefficients for real nonlinear and fictitious linear temperature-dependent periodic composites," *J. Non-Linear Mech.*, vol. 36, no. 1, pp. 187–195, 2001.

[17] S. Tokarzewski, I. Andrianov, V. Danishevsky, and G. Starushenko, "Analytical continuation of asymptotic expansion of effective transport coefficients by Padé approximants," *Nonlinear Analysis*, vol. 47, pp. 2283–2292, 2001.

[18] V. Mityushev, "Sready heat conduction of a material with an array of cylindrical holes in the nonlinear case," *J. Applied Mathematics*, vol. 61, pp. 91–102, 1998.

[19] C. Baiocchi and A. Capelo, *Variational and quasivariational inequalities. Applications to free boundary problems*. Chichester etc.: John Wiley and Sons, 1984, a Wiley-Interscience Publication.

[20] M. Ablowitz and A. Fokas, *Complex Variables*. Cambridge: Cambridge University Press, 1997.

Variable Structure Algorithm Using Explicit and L-Stable Methods

Eugeny A. Novikov
Institute of Computational Modeling SB RAS
ICM SB RAS
Krasnoyarsk, Russia
novikov@icm.krasn.ru

Anton E. Novikov
Siberian Federal University
SFU
Krasnoyarsk, Russia
aenovikov@bk.ru

***Abstract* —** An explicit two-stage Runge-Kutta scheme and a L-stable (2,1)-method are constricted, both scheme of order two. A numerical formula of order one is developed, which is based on stages of the explicit method and its stability interval is extended to 8. An integration algorithm of alternating order and step is constructed, choice of the most efficient numerical scheme is performed on each step with applying an inequality of stability control. Numerical results, confirming efficiency of the algorithm are given.

Keywords — stiff system; accuracy and stability control; variable structure algorithms.

I. INTRODUCTION

In modeling kinetics of chemical reactions, calculation of electronic circuits, and other important applications, there is a problem of numerical solution of the Cauchy problem for stiff systems of ordinary differential equations. Main trends in construction of numerical methods are associated with expansion of their possibilities in solving problems of more and more high dimension. Mathematical formulation of practical problems become more and more accurate, which leads to increase of dimension and complexity of a right part of a system of differential equations. Despite improvement in computer performance, complexity of problems arising in practice outgrows development of computer technology, which in turn leads to increasing demands for computational algorithms.

In many cases, calculations are required to be conducted within limits of so called engineering accuracy about - 1% and lower. This is due to the fact, that measurement of constants in a right part of a system of differential equations is often quite rough. Sometimes, such accuracy of calculations is satisfactory in terms of a goal. It is well-known (see, e.g., [1]), that order of approximation of a numerical scheme should be associated with required accuracy of calculations. Therefore, below we shall consider only those numerical formulas, that have order of accuracy less or equal to two.

Modern methods for solving stiff problems usually use calculation and inversion of the Jacobi matrix of a system of differential equations. In case of a sufficiently large dimension, efficiency of numerical methods is almost completely determined by inversion (decomposition) of the matrix. To increase efficiency of calculations in a number of algorithms, the freezing of the Jacobi matrix is used, that means using same matrix on several integration steps [2]. This approach is most successful in algorithms, based on multistep methods and, in particular, in backward differentiation formulas [3]. This problem does not cause any particular difficulties in constructing integration algorithms, based on other numerical schemes, if their stages are computed with the Jacobian matrix in some iterative process. This is due to the fact, that in this case the Jacobian matrix does not affect accuracy order of a numerical scheme, but only determines rate of convergence of iterations. So, it needs to be recomputed, when there is a significant slowdown in convergence rate of the iterative process.

The situation is worse in an integration algorithm, based on the known noniterative methods, which include methods of the Rosenbrock type [4] and their various modifications [2]. It should be noted that the noniteration method is much simpler in terms of computer implementation than algorithms based on numerical formulas, which are evaluated with using iterations. However, in methods of form [4], the Jacobi matrix affect accuracy order of a numerical scheme and, therefore, difficulties with its freezing arise. If a problem of using same matrix on several steps of integration is left unsolved, then, obviously one is limited to solve only problems of low dimensions. In [5, 6], this problem is considered in relation to the Rosenbrock methods. It is proved, that maximum accuracy order of the Rosenbrock methods is equal to two, if in an integration algorithm the same Jacobi matrix is applied on several steps of integration. There is an algorithm with freezing Jacobi matrix, based on *L*-stable numerical formulas of second accuracy order and results of calculations, confirming its high efficiency.

Another important requirement for modern integration algorithms is numerical approximation of the Jacobi matrix. This is due to the fact that a right part of a system of differential equations often has large dimension and quite complex form. A typical example is provided by problems of chemical kinetics, where complexity of a right part side increases with number of elementary stages in a chemical reaction. Nowadays, simulation involves reactions, which contain dozens of reagents and hundreds of elementary stages.

ISBN: 978-1-61804-251-4

Therefore, in some cases, less effective numerical methods is more preferable, if their implementation does not require analytical calculation of elements of the Jacobi matrix. This barrier can be removed if an integration algorithm includes possibility of numerical approximation of the Jacobi matrix. Note, that the problem of freezing and numerical approximation are in some sense close to each other and, therefore, can be solved simultaneously.

Some analog of freezing the Jacobi matrix is using in calculations integration algorithms, based on explicit and L-stable methods with automatic selection of a numerical scheme. In this case, efficiency of the algorithm can be improved by calculating transitive regions corresponding to a maximum eigenvalue of the Jacobi matrix by an explicit method. It is natural to apply an inequality for stability control [7] as a criterion for choosing an efficient numerical formula. Note, that using such hybrid algorithms does not fully eliminate the problem of freezing the Jacobi matrix, because the explicit method can applied, generally speaking, only for a boundary layer solution, corresponding to a maximum eigenvalue of the Jacobi matrix.

Here, based on the explicit methods of the Runge-Kutta type of the first and second orders, as well as the L-stable (2,1)-method of second-order accuracy, an algorithm of variable structure is constructed, which allows freezing both a numerical and an analytical Jacobi matrix. Numerical results confirm efficiency of the integration algorithm.

II. L-STABLE (2,1)- METHOD

In [8] for numerical solution of the Cauchy problem for stiff systems of ordinary differential equations

$$y' = f(t,y),\ y(t_0) = y_0,\ t_0 \le t \le t_k, \qquad (1)$$

where y and f are real N-dimensional vector functions, and t is an independent variable, the class of (m, k)-methods is proposed. From the standpoint of computer implementation, (m,k)-methods are as simple as the Rosenbrock schemes. However, in contrast to the Rosenbrock methods, in this class it is much easier to solve a problem of freezing the Jacobi matrix and its numerical approximation. In addition, (m,k)-methods have more good properties of accuracy and stability with slight increase of computational cost. In traditional methods, number of stages m completely describes a numerical formula. In (m,k)-methods two constants are required to describe numerical schemes: number of stages m and number of calculations of a right part of the system (1) on an integration step k.

To solve the problem (1), we consider a (2,1)-scheme

$$y_{n+1} = y_n + p_1 k_1 + p_2 k_2,$$
$$D_n k_1 = hf\left(t_n + \beta h, y_n\right),\ D_n k_2 = k_1, \qquad (2)$$

where k_1 and k_2 are stages of the method; $D_n = E - ahA_n$, E is an identity matrix, h is an integration step, A_n is a some matrix, which can be represented in a following form

$$A_n = f_n' + hB_n + O\left(h^2\right), \qquad (3)$$

$f'_n = \partial f(t_n,y_n)/\partial y$ is the Jacobi matrix of the system (1), B_n is an independent of an integration step arbitrary matrix, and a, β, p_1, p_2 are numerical coefficients. Using the matrix A_n represented in form (3) allows us to apply (2) with freezing both an analytical and a numerical Jacobi matrix [9]. In case of using the Jacobi matrix f'_{n-k}, calculated k steps back, we have

$$B_n = -kf_n'' f_n,\ f_n'' f_n = \partial^2 f(y_n)/\partial y^2.$$

If the Jacobi matrix is computed numerically with a step $r_j = c_j h$, then elements $b_{n,ij}$ of the matrix B_n have a form

$$b_{n,ij} = 0.5 c_j \partial^2 f_i\left(t_n, y_n\right)/\partial y_j^2.$$

In calculations the step r_j is chosen according to a formula $r_j = \max(10^{-14}, 10^{-7}|y_j|)$.

Let's obtain the coefficients for the L-stable numerical scheme (2) of second order and an inequality for accuracy control. An expansion of an exact solution $y(t_{n+1})$ in the Taylor series in a vicinity of a point t_n to terms with h^3 inclusive has a form

$$y\left(t_{n+1}\right) = y\left(t_n\right) + hf + 0.5h^2\left[f_t' + f_y' f\right] +$$
$$+h^3\left[f_{tt}'' + f_y' f_t' + 2f_{yt}'' f + f_y'^2 f + f_{yy}'' f^2\right]/6 + O\left(h^4\right), \qquad (4)$$

where the elementary differentials are calculated on an exact solution $y(t_n)$. To find the coefficients a, β, p_1 and p_2 of the scheme (2), we write an expansion of stages k_1 and k_2 in a Taylor series in a vicinity of a point y_n to terms with h^3 inclusive, and substitute it in (2). We obtain

$$y_{n+1} = y_n + \left(p_1 + p_2\right)hf_n + \beta\left(p_1 + p_2\right)h^2 f'_{t,n} +$$
$$+a\left(p_1 + 2p_2\right)h^2 f'_{y,n} f_n + 0.5\beta^2\left(p_1 + p_2\right)h^3 f''_{tt,n} +$$
$$+a\beta\left(p_1 + 2p_2\right)h^3 f'_{y,n} f'_{t,n} + a^2\left(p_1 + 3p_2\right)h^3 f'^2_{y,n} f_n +$$
$$+a\left(p_1 + 2p_2\right)h^3 B_n f_n + O\left(h^4\right), \qquad (5)$$

where the elementary differentials are calculated on an approximate solution y_n. Assuming, that $y_n = y(t_n)$ and comparing (4) and (5) to terms with h^2 inclusive, we obtain conditions of second-order accuracy of scheme (2), i.e.,

$$p_1 + p_2 = 1,\ a(p_1 + 2p_2) = 0.5,\ \beta = 0.5. \qquad (6)$$

Let's investigate stability of numerical formula (2). Applying it to a problem

$$y' = \lambda y,\ y(0) = y_0,\ \mathrm{Re}(\lambda) < 0, \qquad (7)$$

we obtain $y_{n+1} = Q(x)y_n$, and $x = h\lambda$, where a function of stability $Q(x)$ has the form

$$Q(x) = \frac{1 + \left(p_1 + p_2 - 2a\right)x + a\left(a - p_1\right)x^2}{\left(1 - ax\right)^2}.$$

Then, the scheme (2) is L-stable, if $p_1 = a$. Substituting this relation in (6), we obtain a set of coefficients

$$p_1 = a,\ p_2 = 1 - a,\ \beta = 0.5, \qquad (8)$$

where a is determined from a L-stability condition

$$a^2 - 2a + 0.5 = 0. \qquad (9)$$

Comparing (4) and (5) to terms with h^3 inclusive, we find, that a local error δ_n of the numerical scheme (2) with the coefficients (8) has a form

$$\delta_n = \left(a - 1/3\right)h^3 f_y'^2 f + h^3 f_{tt}''/24 + h^3 f_{yy}'' f^2/6 +$$
$$+h^3 f_{yt}'' f/3 - h^3 f_y' f_t'/2 - h^3 B_n f/2 + O\left(h^4\right). \qquad (10)$$

The equation (9) has two roots a_1=1–0.5√2 and a_2=1+0.5√2. We choose $a= a_1$, as in this case the coefficient in the leading term $(a – 1/3)h3f^2f$ of error (10) is less.

Let's consider simultaneously the numerical Rosenbrock formula with two calculations of the function f on an each step

$$y_{n+1} = y_n + p_1k_1 + p_2k_2,$$
$$D_nk_1 = hf\left(y_n\right),\ D_nk_2 = hf\left(y_n + \gamma k_1\right), \qquad (11)$$

According to [5], for $\gamma = a$, a set of coefficients (8) provides a second accuracy order of (11), and condition (9) provides its L-stability. It follows from [5], that the numerical formula (11) with coefficients (8) is one of the most efficient among the methods of the Rosenbrock type, with two computations of a right part of a differential problem on an integration step. A local error of the numerical formula (11) has a form

$$\delta_n^{roz} = h^3\left(a - \frac{1}{3}\right)f'^2f + \left(\frac{1}{6} + \frac{1-\sqrt{2}}{2}a\right)h^3f''f^2 -$$
$$-ah^3B_nf + O\left(h^4\right). \qquad (12)$$

The scheme (2) with coefficients (8) as well as scheme (11) with coefficients (8) has second accuracy order and L-stability, and their local errors (10) and (12) differ slightly. At the same time, the scheme (2) requires one less calculation of function f than (11) on an each step, with other costs being equal, which makes it preferable.

We construct accuracy control of the numerical scheme (2) by analogy with [5]. For this purpose, we denote

$$v\left(j_n\right) = D_n^{1-j_n}\left(k_2 - k_1\right), \qquad (13)$$

where k_1 and k_2 calculated by the formulas (2). Then according to [5], in order to control the accuracy on an each step, one has to control an inequality

$$\left\|v(j_n)\right\| \le \varepsilon,\ 1 \le j_n \le 2, \qquad (14)$$

where ε is required accuracy of calculations, $\|\cdot\|$ is some norm in R^N, and the integer variable j_n is selected as the lowest, for which inequality (14) holds.

Note one important feature of error estimation (13). The scheme (2) is L-stable, that is, for its stability function $Q(x)$, the relation $Q(x) \to 0$ for $x \to -\infty$ holds. Since for an exact solution $y(t_{n+1}) = \exp(x)y(t_n)$ of the problem (7), a similar property holds, it is natural to require convergence to zero of the error estimation for $x \to -\infty$. However, for $v(1)$, this property is not satisfied — this estimation has an A-stable manner. To correct asymptotic behavior of the estimated error we introduced an estimation $v(j_n)$, $1 \le j_n \le 2$ instead of $v(1)$. In this case behavior of error estimations for $j_n = 2$ will be coordinated with behavior of the exact solution of the test problem for $x \to -\infty$. We emphasize, that in sense of a general member, estimations $v(1)$ and $v(2)$ coincide. Using $v(j_n)$ actually does not lead to increase of computational costs. This is due to the fact, that $v(j_n)$ for $j_n = 2$ is checked only if it is violated for $j_n = 1$. This situation appears rarely, mainly when an integration step grows rapidly. However, this allows us to choose the step more accurately and thereby reduce the number of unnecessary recomputing solutions (returns).

An estimation of a maximum eigenvalue $\omega_{n,0} = h\lambda_{n,\max}$ of the Jacobi matrix of the system (1), necessary to switch to an explicit formula, is estimated through its norm $w_{n,\,0} = h\|\partial f(t_n,y_n)/\partial y\|$. Below, this estimation will be used for automatic selection of a numerical scheme.

III. RUNGE–KUTTA METHOD OF SECOND ORDER

For solution the problem (1), we consider an explicit two-stage Runge–Kutta method [11]:

$$y_{n+1} = y_n + p_1k_1 + p_2k_2,$$
$$k_1 = hf\left(y_n\right),\ k_2 = hf\left(y_n + \beta k_1\right). \qquad (15)$$

Let's consider the autonomous problem (1) to simplify formulas. In case of a nonautonomous system $y' = f(t, y)$, scheme (15) is written as

$$y_{n+1} = y_n + p_1k_1 + p_2k_2,$$
$$k_1 = hf\left(t_n, y_n\right),\ k_2 = hf\left(t_n + \beta h, y_n + \beta k_1\right).$$

We obtain relations for the coefficients of the method (15) of second accuracy order. For this purpose, we expand the stages k_1 and k_2 in Taylor series in powers of h up to terms with h^3 inclusive, and substitute them in the first formula (15). The result

$$y_{n+1} = y_n + \left(p_1 + p_2\right)hf_n + \beta p_2h^2f_n'f_n +$$
$$+0.5\beta^2h^3p_2f_n''f_n^2 + O\left(h^4\right),$$

where the elementary differentials are calculated on the approximate solution y_n. Comparing this expression with (4) to terms with h^2 inclusive, assuming, that $y_n = y(t_n)$, we write conditions $p_1 + p_2 = 1$ and $\beta p_2 = 0.5$ of second accuracy order of the scheme (15). In these relations, the local error δ_n of the scheme (15) can be written as follows:

$$\delta_n = h^3\left[\frac{1}{6}f'^2f + \frac{2-3\beta}{12}f''f^2\right] + O\left(h^4\right).$$

We construct an inequality to check accuracy. For this purpose we consider an auxiliary scheme $y_{n+1,\,1} = y_n + k_1$ of first accuracy order. Using an idea of nested methods, estimation of error $\varepsilon_{n,\,2}$ of the second order method can be calculated by formula [10]:

$$\varepsilon_{n,2} = y_{n+1} - y_{n+1,1} = p_2\left(k_2 - k_1\right).$$

To improve the reliability of this estimation, we choose $\beta = 1$. Then, stage k_1 is computed at the point t_n, and k_2 is computed at the point t_{n+1}. Calculations show, that using information in extreme points of a step leads to more reliability. For $\beta = 1$, coefficients of the method of second order are uniquely determined $p_1 = p_2 = 0.5$ and local error and an inequality for accuracy control are, respectively, given below:

$$\delta_n = \frac{h^3}{12}\left[2f'^2f - f''f\right] + O\left(h^4\right),\ 0.5\left\|k_2 - k_1\right\| \le \varepsilon.$$

Now, we construct an inequality for stability control of (15) by the method proposed in [7]. For this purpose, we consider an auxiliary stage $k_3 = hf(y_{n+1})$. Note, that k_3 coincides with the stage k_1, which is used on a next integration step and, therefore, its applying does not lead to additional computing of the right part of (1). We write stage k_1, k_2, k_3,

applied to a problem $y' = Ay$, where A is a matrix with constant coefficients. A result is given below

$$k_1 = Xy_n,\ k_2 = \left(X + X^2\right)y_n,\ k_3 = \left(X + X^2 + 0.5X^3\right)y_n,$$

where $X = hA$. It is easy to see, that

$$k_2 - k_1 = X^2 y_n,\ 2\left(k_3 - k_2\right) = X^3 y_n.$$

Then, according to [7], an estimation of a maximum eigenvalue $w_{n,2} = h\lambda_n$, max of the Jacobi matrix of the system (1) can be calculated by a following formula

$$w_{n,2} = 2\max_{1\le i\le N}\left\{\left|k_3^i - k_2^i\right| / \left|k_2^i - k_1^i\right|\right\}. \qquad (16)$$

A stability region of the scheme (15) is showed on a Fig. 1

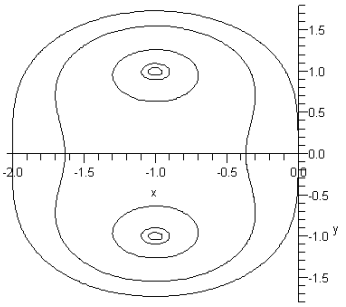


Fig. 1. Stability region of the scheme (15)

A stability interval of (15) of second accuracy order is approximately equal to two. Therefore, an inequality $w_{n,2} \le 2$ can be applied for stability control. In case of using this inequality for step selection, roughness of estimation (16) should be considered, because a maximum eigenvalue is strongly separated from rest, in a power method few iterations are applied and additional distortions are occur because of nonlinearity of the problem (1). Therefore, stability control is used to limit size of an integration step. As a result, we will calculate the projected step h_{n+1} as follows. We define a new step h^{ac} by criterion of accuracy according to the formula $h^{ac} = qh_n$, where h_n is the last successful step of the integration, and q, taking into account a relation $k_2 - k_1 = O(h^2)$, is given by equation $q^2\|k_2 - k_1\| = \varepsilon$. Step h^{st} by criterion of stability is given by a formula $h^{st} = dh_n$, where d, taking into account a relation $w_{n,2} = O(h)$ is determined from an equation $dw_{n,2} = 2$. Then, the projected step h_{n+1} is calculated by a formula

$$h_{n+1} = \max\left[h_n, \min\left(h^{ac}, h^{st}\right)\right]. \qquad (17)$$

Note, that the formula (17) is used to predict a value of the integration step h_{n+1} after successful computation of solution with the previous step h_n and, therefore, does not actually lead to increase of computational cost. If the step by criterion of stability is less than the last successful one, it will not be reduced, because it may be caused by roughness of estimation of a maximum eigenvalue. However, the step will not be increased, because there is a possibility of instability of the numerical scheme. If the step should be reduced by criterion of stability, then the last successful step h_n is applied again. As a result, the formula (17) is proposed to select a step. This formula allows to stabilize step behavior on a settling region of solution, where stability has a defining role. Indeed, existence of this region limits possibilities of applying explicit methods for solving stiff problems.

IV. RUNGE–KUTTA METHOD OF FIRST ORDER

For numerical solution of the problem (1), we consider a scheme

$$y_{n+1} = y_n + r_1 k_1 + r_2 k_2,$$
$$k_1 = hf\left(y_n\right),\ k_2 = hf\left(y_n + k_1\right). \qquad (18)$$

Note, that when $r_1 = r_2 = 0.5$, the numerical formula (18) has second order of accuracy, and coincides with (15) with coefficients $p_1 = p_2 = 0.5$. We construct a less accurate scheme with a maximum interval of stability. For this purpose, we use (18) for solution the scalar test equation (7). We obtain $y_{n+1} = Q(x)y_n$, where the function of stability $Q(x)$ has a form

$$Q\left(x\right) = 1 + \left(r_1 + r_2\right)x + r_2 x^2,\ x = h\lambda.$$

Requirement of first order of accuracy leads to the relation $r_1 + r_2 = 1$, which, below, we will assume to be satisfied. Now, we choose r_2 so that the method (18) has a maximum stability interval. For this purpose, we consider the Chebyshev polynomial $T_2(z) = (2z^2 - 1)$ on an interval $[-1,1]$. We carry out change of variables, setting $z = 1 - 2x/\gamma$. We obtain $T_2(x) = 1 - 8x/\gamma + 8x^2/\gamma^2$, and the interval $[\gamma, 0]$ passes to $[-1,1]$. It is easy to show, that among all polynomials of a form $P_2(x) = 1 + x + c_2x^2$ for $T_2(x)$, the inequality $|T_2(x)| \le 1$ is satisfied at a maximum interval $[\gamma, 0]$, $\gamma = -8$. We require, that the coefficients of $Q(x)$ and $T_2(x)$ coincide at $\gamma = -8$. This leads to relations $r_1 + r_2 = 1$ and $r_2 = 1/8$. As a result, we have coefficients $r_1 = 7/8$ and $r_2 = 1/8$ of the first accuracy order method with a maximum stability interval, with a local error

$$\delta_n = 3h^2 f f' / 8 + O\left(h^3\right).$$

We will use estimation of the local error to control accuracy of the numerical formulas of the first order,. Taking into account, that

$$k_2 - k_1 = h^2 f_n' f_n + O\left(h^3\right)$$

and a form of the local error, an inequality for accuracy control may be written as

$$\left\|k_2 - k_1\right\| \le 8\varepsilon/3,$$

where $\|\cdot\|$ is a some norm in R^N, and ε is required accuracy of calculations.

We construct an inequality for stability control for a first order method. For this purpose, we consider an auxiliary stage $k_3 = hf(y_{n+1})$. We write k_1, k_2 and k_3, applied to the problem $y' = Ay$, where A is the matrix with constant coefficients. As a result, we obtain

$$k_1 = Xy_n,\ k_2 = \left(X + X^2\right)y_n,\ k_3 = \left(X + X^2 + 0.125X^3\right)y_n,$$

where $X = hA$. It is easy to see, that

$$k_2 - k_1 = X^2 y_n,\ 8\left(k_3 - k_2\right) = X^3 y_n.$$

Then, according to [7], estimation of a maximum eigenvalue $w_{n,1} = h\lambda_{n,\max}$ of the Jacobi matrix of the system (1) can be calculated by a formula

$$w_{n,1} = 8\max_{1\le i\le N}\left\{\left|k_3^i - k_2^i\right| / \left|k_2^i - k_1^i\right|\right\}.$$

A stability region of the scheme (18) is showed on a Fig. 2.

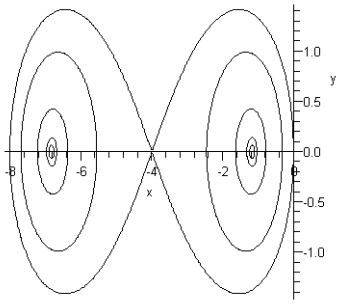


Fig. 2. Stability region of the scheme (18)

An interval of stability of the numerical scheme (18) is equal to eight. Therefore, an inequality $w_{n,1} \le 8$ can be applied to control stability.

V. INTEGRATION ALGORITHM WITH AUTOMATIC SELECTION OF NUMERICAL SCHEME

An algorithm of alternating order and step may be easily formulated on a base of constructed explicit methods of first and second orders of accuracy. Calculations are always begun with the second order method as it is more accurate. Transition to first order scheme is carried out in case of violation an inequality $w_{n,2} \le 2$. Reverse transition to the second order method is carried out if an inequality $w_{n,1} \le 2$ holds. On calculations by the first order method in addition to accuracy control there is stability control, and choice of a predicted step is carried out in the same manner as in the second order method applying a formula of type (17).

In case of using the scheme (2) formulation of an integration algorithm also does not present difficulties. Violation of an inequality $w_{n,1} \le 8$ causes a transition to the scheme (2). Transfer to explicit methods is carried out if an inequality $w_{n,0} \le 8$ holds.

The numerical formula (2), without the loss of the accuracy order, can be applied to freezing a matrix D_n. Note, that during the freezing the Jacobi matrix, size of an integration step remains constant. An attempt to freeze the matrix D_n, is carried out after each successful step. The matrix thaws in following cases: (1) violation of accuracy of calculation; (2) if number of steps with a frozen matrix reaches a defined maximum number i_h; and (3) if a projected step is greater than the last one by q_h times. By the numbers i_h and q_h, we can affect the redistribution of computational cost. When $i_h = 0$ and $q_h = 0$, freezing does not occur; with increasing i_h and q_h, the number of calculations of a right part increases, while number of inversions of the Jacobi matrix decreases.

The norm in a left part of inequality for accuracy control is calculated by a formula

$$\left\| k_2 - k_1 \right\| = \max_{1 \le i \le N} \left\{ \frac{| k_2^i - k_1^i |}{| y_n^i | + r} \right\},$$

where i is number of components, and r is a positive parameter. If, for the i-th component of solution, an inequality $|y_n^i| < r$ holds, then an absolute error $r\varepsilon$ is controlled, otherwise, the relative error ε is controlled. Below the algorithm of variable order and step with automatic selection of an explicit or a L-stable numerical scheme is called RKMK2.

VI. NUMERICAL RESULTS

Calculations were carried out on PC Intel(R) Core(TM) i7-3770S CPU@3.10GHz with double precision. In the calculations, the parameter r was chosen so that practical accuracy of all components of solution was not worse than defined accuracy. The calculations were performed with accuracy $\varepsilon = 10^{-2}$. This is due to the fact, that the algorithm, based on low accuracy order schemes, and, therefore, it is impractical to carry out calculations with higher accuracy with this method. A comparison of its efficiency was carried out with the well-known Gear method in the implementation of A. Hindmarsh named DLSODE from the ODEPACK collection [12].

Below *ifu* and *ija* denote, respectively, total number of calculations of a right part and number of inversions (decompositions) of the Jacobi matrix of the problem (1), which allow us to evaluate objectively the efficiency of the integration algorithm.

As the first test, the simplest model of the Belousov–Zhabotinsky reaction [13] was chosen:

$$\begin{gathered} y_1' = 77.27\left(y_2 - y_1 y_2 + y_1 - 8.375 \cdot 10^{-6} y_1^2\right), \\ y_2' = \left(-y_2 - y_1 y_2 + y_3\right)/77.27, \\ y_3' = 0.161\left(y_1 - y_3\right), \\ t \in [0, 300],\ y_1(0) = y_3(0) = 4,\ y_2(0) = 1.1,\ h_0 = 2 \cdot 10^{-3}. \end{gathered} \tag{19}$$

Calculations were made with numerical Jacobi matrix. A solution of this problem by the algorithm RKMK2 was calculated with costs ifu = 1 214 and ija = 65. Calculations only by the L-stable scheme (2) give ifu = 926 and ija = 88. Practical accuracy of the calculations at the end of the interval of integration is not worse than defined accuracy. Solution of (19) was calculated by explicit methods of variable order and step with cost ifu = 2 112 678. This problem is too stiff for explicit methods. However, the results of calculations are shown here to demonstrate principal possibility of application explicit methods for solving stiff enough examples, those in solving some high-dimensional problems may be more efficient than L-stable methods. For calculation by the DLSODE program, the required accuracy of 10^{-2} is achieved, when defined accuracy is 10^{-4} with costs ifu = 1129 and ija = 107. On calculations with higher accuracy, DLSODE is more efficient than the constructed algorithm. This is a sequence of low accuracy order of the constructed numerical formulas. On defined accuracy equal 10^{-2}, the algorithm RKMK2 is more efficient than the well-known method DLSODE in 1.5 times in number of inversions of the Jacobi matrix, while number of computations of the right part of (19) for RKMK2 and DLSODE vary slightly. In case of the large-

scale problem (1), the constructed integration algorithm may be more efficient than DLSODE in calculating time. Time dependence of y_1 is showed on a Fig. 3.

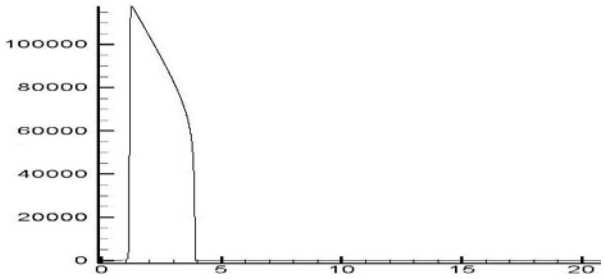


Fig. 3. Time dependence of y_1 (fragment)

A second example describes the modified oregonator exhibiting complicated limit cycle. A reaction includes following six stages [14]

$$\text{A+Y} \to \text{X+P}, \quad k_1 = 0.084, \; k_{-1} = 10^4,$$
$$\text{X+Y} \to 2\text{P}, \; k_2 = 4\cdot 10^8, \; k_{-2} = 5\cdot 10^{-5},$$
$$\text{A+X} \to 2\text{W}, \; k_3 = 2\cdot 10^3, \; k_{-3} = 2\cdot 10^7,$$
$$\text{C+W} \to \text{X+Z}, \; k_4 = 1.3\cdot 10^5, \; k_{-4} = 2.4\cdot 10^7,$$
$$2\text{X} \to \text{A+P}, \; k_5 = 4\cdot 10^7, \; k_{-5} = 4\cdot 10^{-11},$$
$$\text{Z} \to \text{C+0.462Y}, \; k_6 = 0.65,$$

where $k_i, 5 \le i \le 6$, – constants of velocities forward (with positive indices) and counter (with negative indices) stages. There are 7 entities in this reaction, denoted by

$$\text{A}=\text{BrO}_3^-, \; \text{C}=\text{M(n)}, \; \text{P}=\text{HOBr}, \; \text{W}=\text{BrO}_2,$$
$$\text{X}=\text{HBrO}_2, \; \text{Y}=\text{Br}^-, \; \text{Z}=\text{M(n+1)}.$$

In this notations M(n) is an ion of a metal accelerant, M(n+1) is an oxygenated form of the ion. Let's denote concentrations of reagents by

$$c_1=[\text{BrO}_3^-], \; c_2=[\text{Br}^-], \; c_3=[\text{M(n)}],$$
$$c_4=[\text{HBrO}_2], \; c_5=[\text{HOBr}], \; c_6=[\text{BrO}_2], \; c_7=[\text{M(n+1)}].$$

This reaction proceeds in an isothermal reactor with constant capacity with substance exchange. A Corresponding system of equations is given below

$$\begin{aligned}
c_1' &= -v_1 - v_3 + v_5 + (c_{p1} - c_1)/\theta, \\
c_2' &= -v_1 - v_2 + 0.462v_6 + (c_{p2} - c_2)/\theta, \\
c_3' &= -v_4 + v_6 + (c_{p3} - c_3)/\theta, \\
c_4' &= v_1 - v_2 - v_3 + v_4 - 2v_5 + (c_{p4} - c_4)/\theta, \\
c_5' &= v_1 + 2v_2 + v_5 + (c_{p5} - c_5)/\theta, \\
c_6' &= 2v_3 - v_4 + (c_{p6} - c_6)/\theta, \\
c_7' &= v_4 - v_6 + (c_{p7} - c_7)/\theta,
\end{aligned} \qquad (20)$$

where velocities $v_1, v_2, \ldots, v_6$ of stages are defined by formulas

$$v_1 = k_1c_1c_2 - k_{-1}c_4c_5, \; v_2 = k_2c_2c_4 - k_{-2}c_5^2,$$
$$v_3 = k_3c_1c_4 - k_{-3}c_6^2, \; v_4 = k_4c_3c_6 - k_{-4}c_4c_7,$$
$$v_5 = k_5c_4^2 - k_{-5}c_1c_5, \; v_6 = k_6c_7.$$

Integration of the system (20) was made on an interval [0,1000] with an initial step equal to 10^{-5}. Inlet concentrations of reagents are given below

$$c_{p1} = 0.14, \; c_{p2} = 0.151\cdot 10^{-5}, \; c_{p3} = 0.125\cdot 10^{-3},$$
$$c_{p4} = c_{p5} = c_{p6} = c_{p7} = 0$$

and Θ=125.5. Initial values of reagent concentrations are equal

$$c_1 = 0.1387, \; c_2 = 0.1534\cdot 10^{-6}, \; c_3 = 0.1176\cdot 10^{-3},$$
$$c_4 = 0.3165\cdot 10^{-7}, \; c_5 = 0.1956\cdot 10^{-3},$$
$$c_6 = 0.5814\cdot 10^{-6}, \; c_7 = 0.631\cdot 10^{-5}.$$

Time dependence of $[BrO_2]$ is showed on a Fig. 4.

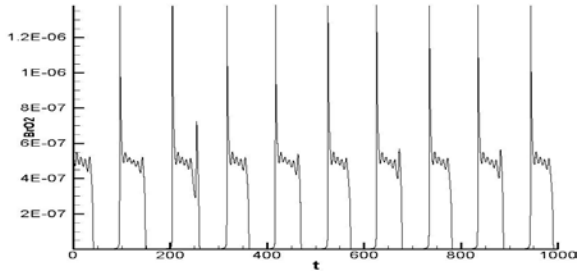


Fig. 4. Time dependence of $[BrO_2]$ concentration

Calculations was performed with a numerical Jacobian matrix. Solution of the problem was calculated by the algorithm RKMR2 with ifu = 5 623 and ija = 533. Calculations with the L-stable scheme only give ifu = 5 371 and ija = 591. Practical accuracy in the end of the integration interval as good as defined one. In calculations by DLSODE required accuracy 10^{-2} is achieved for defined accuracy equal to 10^{-4} with computational costs ifu = 7 806 and ija = 542. On higher accuracy of calculations DLSODE is more efficient than the constructed algorithm. It is a result of low order of accuracy of constructed numerical formulas. In case of high dimension of the problem (1) the constructed algorithm may be more efficient in time than DLSODE.

VII. CONCLUSION

The constructed algorithm RKMK2 is designed for low precision calculations — about 1% and lower. In this case, its maximum efficiency is reached.

In RKMK2, with its parametrs, one can specify different modes of calculation:

(1) explicit methods of first or second order of accuracy with or without stability control;

(2) explicit methods with variable order and step;

(3) L-stable method with or without freezing, both an analytical and a numerical Jacobi matrix;

(4) with automatic selection of the numerical scheme.

This allows us to apply this algorithm to solving both stiff and nonstiff problems. In calculations with automatic selection of a numerical scheme, the integration algorithm makes a decision whether a problem is stiff or not by itself.

Using the inequality for stability control does not actually lead to increase of computational cost, because estimation of a maximum eigenvalue of the Jacobi matrix of (1) is carried out

through a previously computed stages and does not lead to increase of number of computed values of function *f*. This estimation is rough. However, using stability control for limiting step growth allows us to avoid negative effects of roughness of estimation. Moreover, in some cases this leads to an exceptionally high growth of efficiency of the algorithm. In a settling region, the old errors tend to zero due to stability control, and the new ones are low, due to small values of derivatives of a solution. In some cases, following eigenvalue is estimated instead of a maximum one. An integration step becomes greater than limit, and with such a step the integration step is carried out as long as it does not disturb the inequality for checking the accuracy. Typically, the number of such steps is small. However, the step may be an order of magnitude greater than the maximum step for stability. After violation of the inequality for accuracy control, the step is reduced to a maximum possible. This effect can be repeated many times, depending on length of a region of settling. As a result, an average integration step cannot exceed the maximum allowable.

Application the explicit first order method with an extended interval of stability on a settling region allows increase an integration step in 4 times in comparison with the explicit second order method without increasing of computational cost. On transition regions, where accuracy of calculations has a defining role, the second accuracy order method with a small area of stability is more efficient.

Combining methods of low and high order applying an inequality for stability control, improves efficiency of calculations.

ACKNOWLEDGMENT

This work was supported by the Russian Foundation for Basic Research (project 14-01-00047).

REFERENCES

[1] S. S. Artem'ev, G. V. Demidov, E. A. Novikov, and L. A. Yumatova, "The Way to Optimize the Numerical Solution of Cauchy Problem for Ordinary Differential Equations," *Chislennye Metody Mekh. Sploshn. Sredy*, no. 2, pp. 5–14, 1984.

[2] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations. Stiff and Differential-Algebraic Problems*. Cambridge Univ. Press, Cambridge, 1987.

[3] G. D. Byrne and A. C. Hindmarsh, "ODE Solvers: a Review of Current and Coming Attractions," *J. Comput. Phys.*, no. 70, pp. 1–62, 1987.

[4] H. H. Rosenbrock, "Some General Implicit Processes for the Numerical Solution of Differential Equations," *Computer*, no. 5, pp. 329–330, 1963.

[5] E. A. Novikov, V. A. Novikov, and L. A. Yumatova, "The Way to Increase the Efficiency of an Integration Algorithm That is Based on the Rosenbrock_Type Formula with the Second Order of Accuracy by Freezing the Jacoby Matrix," Novosibirsk, *preprint* no. 592, Institute of Computational Mathematics and Mathematical Geophysics SB RAS, 1985.

[6] E. A. Novikov and A. L. Dvinskii, "Jacobi Matrix Freezing for Rosenbrock-Type Methods," *Vychislitelnye Tekhnologii*, vol. 10, pp. 108–114, 2005.

[7] E. A. Novikov, *Explicit Methods for the Stiff Systems*. Novosibirsk, Nauka, 1997.

[8] E. A. Novikov, Yu. A. Shitov, and Yu. I. Shokin, "Single-Step Methods for Solving the Stiff Systems," *Dokl. Akad. Nauk SSSR*, vol. 301, no. 6, pp. 1310–1314, 1988.

[9] E. A. Novikov and Yu. A. Shitov, "The Algorithm for Stiff Systems Integrating on the Base of (m,k)-Method with Second Order of Accuracy with Numerical Calculation of Jacoby Matrix," Krasnoyarsk , *preprint* no. 20, Institute of computational modeling SB RAS, 1988.

[10] G. V. Demidov and E. A. Novikov, "Error Estimation for One_step Methods for Integrating the Ordinary Differential Equations," *Chisl. Metody Mekhan. Splosh. Sredy*, vol. 16, no. 1, pp. 27–42, 1985.

[11] L. V. Knaub, Yu. M. Laevskii, and E. A. Novikov, "Integration Algorithm of Variable Order and Step that is Based on the Explicit Two_Stage Runge_Kutta Method," *Sib. Zh. Vychisl. Mat.* , vol. 10, no. 2, pp. 177–185, 2007.

[12] http://www.netlib.org/odepack/index.html

[13] W. H. Enright and T. E. Hull, "Comparing Numerical Methods for the Solutions of Systems of ODE's," *BIT*, vol. 15, pp. 10–48, 1975.

[14] Showalter K., Noyes R.M., Bar-Eli K. "A Modified Oregonator Model Exhibiting Complicated Limit Cycle Behavior in a Flow System," *J. Chem. Phys*., vol. 69, pp. 2514–2524, 1978.

Necessary Conditions of Optimality for Stochastic Switching Systems With Delay

Charkaz Aghayeva

***Abstract*—** This paper provides necessary condition of optimality, in the form of maximum principle for optimal control problems of switching systems with constraints. Dynamics of the processes are defined by the stochastic differential equations with delay in the drift and diffusion coefficients. The restrictions on switches between operating mode are described by collection of functional constraints.

***Keywords*—** Optimal control problem, Stochastic differential equation with delay, Stochastic switching system, Switching law.

I. Introduction

THE stochastic differential equations with delay find much exhibits in description of the real systems, which in one or another degree are subjected to the influence of the random noises. Systems with stochastic uncertainties have provided a lot of interest for problems of nuclear fission, communication systems , self-oscillating systems and etc., where the influences of random disturbances can not be ignored [1]-[3]. Switching systems consist of several subsystems and a switching law indicating the active subsystem at each time instantly. Optimization problems for switching systems have attracted a lot of interest. Theoretical results and applications were developed in [4]-[6]. For general theory of stochastic switching systems it is referred to [7] .

Therefore problems of optimal control for switching systems, described by deterministic and stochastic differential equations with delay, are actual at present [8],[9]. Earlier the problems of stochastic optimal control of switching systems without delay were considered in [10]-[12]. Stochastic optimal control problem of unrestricted switching systems with delay is investigated in [13].

The present work is devoted to the optimal control problem of delayed stochastic switching system with uncontrolled diffusion coefficient. It is obtain maximum principle in the case when endpoint constraints are imposed. Using Ekeland's variational principle [14], given problem is convert into the sequence of unconstrained problems . Due to the result from [13 it is established maximum principle and transversality conditions. Finally, taking the limit, we achieve the necessary condition of optimality in the case when endpoint constraints are imposed.

Ch.Aghayeva is with the Industrial Engineering Department, University of Anadolu, Eskisehir, Turkey, on leave from the Institute of Cybernetics of ANAS, Baku, Azerbaijan (corresponding author to provide phone: 90-222-335 05 80; fax: 90-222-335 36 16; e-mail: c_aghayeva@ anadolu.edu.tr; cherkez.agayeva@gmail.com).

II. Preliminaries and Statement of Problem

Let $(\Omega, F^l, P), l = 1,...,r$ be a probability spaces with filtration $\left\{ F_t^l, t \in [t_{l-1}, t_l], l = 1,...,r \right\}$, $0 = t_0 < t_1 < ... < t_r = T$. Assume that $w_t^1, w_t^2, ..., w_t^r$ are independent Wiener processes, which generate $F_t^l = \overline{\sigma}(w_q^l, t_{l-1} \le t \le t_l),\ l = 1,...,r$. Let R^n denotes the n - dimensional real vector space and $|.|$ denotes the Euclidean norm in R^n . E represents the expectation. $L^2_{F^l}(a,b;R^n)$ denotes the space of all predictable processes $x_t(\omega)$ such that: $E\int_a^b |x_t(\omega)|^2 dt < +\infty$. $R^{m\times n}$ is the space of linear transformations from R^m to R^n . Let, $O_l \subset R^{n_l}, Q_l \subset R^{m_l}$ be open sets and $\mathrm{T} = [0,T]$ be a finite interval.

Consider the following stochastic control system with delay:

$$dx_t^l = g^l(x_t^l, x_{t-h}^l, u_t^l, t)dt + f^l(x_t, x_{t-h}, t)dw_t^l \ t \in (t_{l-1}, t_l] \quad (1)$$

$$x_t^{l+1} = \mathrm{K}^{l+1}(t), t \in [t_l - h, t_l),\ l = 0,1,...,r-1, \quad (2)$$

$$x_{t_l}^{l+1} = \Phi^l(x_{t_l}^l, t_l),\ l = 1,...,r-1,\ x_{t_0}^1 = x_0, \quad (3)$$

$$u_t^l \in U_{\partial}^l \equiv \left\{ u^l(\cdot,\cdot) \in L_F^2(t_{l-1}, t_l; R^m) \,|\, u^l(t,\cdot) \in U^l \subset R^m \right\} \quad (4)$$

where $U^l,\ l = 1,...,r$ are non-empty bounded sets.

Let $\Lambda_l, l = 1,...,r$ be the set of piecewise continuous functions $\mathrm{K}^l(\cdot),\ l = 1,...,r : [t_{l-1} - h, t_{l-1}) \to N_l \subset O_l$ and $h \ge 0$.

The problem is concluded to find the control $u^1, u^2, ..., u^r$ and the switching law $t_1, t_2, ..., t_r$ which minimize the cost functional :

$$J(u) = \sum_{l=1}^{r} E\left[\varphi^l(x_{t_l}^l) + \int_{t_{l-1}}^{t_l} p^l(x_t^l, u_t^l, t)dt \right] \quad (5)$$

which is determined on the decisions of the system (1)- (3), which are generated by all admissible controls $U = U^1 \times U^2 \times ... \times U^r$ at conditions:

$$Eq^r(x_{t_r}^r) \in G \quad (6)$$

G is a closed convex set in R^k .

ISBN: 978-1-61804-251-4

Consider the sets: $A_i = \mathrm{T}^{i+1} \times \prod_{j=1}^{i} O_j \times \prod_{j=1}^{i} \Lambda_j \times \prod_{j=1}^{i} U^j$ with the elements

$\pi^i = (t_0, t_1, t_i, x^1_{t_0}, x^2_{t_2}, ..., x^i_{t_i}, K_1, ..., K_i, u^1, u^2, ..., u^i)$.

Definition 1: The set of functions $\left\{x_t^l = x^l\left(t, \pi^l\right),\ t \in [t_{l-1} - h, t_l], l = 1, ...r\right\}$ is said to be a solution of the equation with variable structure which corresponds to an element $\pi^r \in \mathrm{A}_r$, if the function $x_t^l \in O_l$ on the interval $[t_l - h, t_l]$ satisfies the conditions (2),(3), while on the interval $[t_{l-1}, t_l]$ it is absolutely continuous with probability 1 and satisfies the equation (1) almost everywhere.

Definition 2: The element $\pi^r \in A_r$ is said to be admissible if the pairs $\left(x_t^l,\ u_t^l\right), t \in [t_{l-1} - h, t_l], l = 1, ..., r$ are the solutions of system (1)-(4) and satisfied the conditions (6) .

A_r^0 denotes the set of admissible elements.

Definition 3: The element $\tilde{\pi}^r \in A_r^0$, is said to be an optimal solution of problem (1)-(6) if there exist admissible controls $\tilde{u}_t^l, t \in [t_{l-1}, t_l], l = 1, ...r$ and corresponding solutions $\left\{\tilde{x}_t^l,\ t \in [t_{l-1} - h, t_l], l = 1, ..., r\right\}$ of system (1)-(4) with constraints (6), and pairs $\left(\tilde{x}_t^l, \tilde{u}_t^l\right), l = 1, ..., r$ minimize the functional (5).

Assume that the following requirements are satisfied:

I. Functions g^l, f^l, p^l, $l = 1, ..., r$ and their derivatives are continuous in (x, y, u, t):

$g^l(x, y, u, t) : O_l \times O_l \times Q_l \times T \to R^{n_l}$

$f^l(x, y, t) : O_l \times O_l \times T \to R^{n_l \times n_l}$ $p^l(x, u, t) : O_l \times Q_l \times T \to R^1$.

II. When (t, u) are fixed , functions g^l, f^l, p^l, $l = \overline{1, r}$ hold the conditions:

$$\left(1 + |x| + |y|\right)^{-1}\left(\left|g^l(x, y, u, t)\right| + \left|g_x^l(x, y, u, t)\right| + \left|g_y^l(x, y, u, t)\right| + \left|f^l(x, y, t)\right| + \left|f_x^l(x, y, t)\right| + \left|f_y^l(x, y, t)\right| + \left|p^l(x, u, t)\right| + \left|p_x^l(x, u, t)\right|\right) \le N.$$

III. Functions $\varphi^l(x) : R^{n_l} \to R^1, l = 1, ..., r$ are continuously differentiable and satisfies the following:

$$\left|\varphi^l(x)\right| + \left|\varphi_x^l(x)\right| \le N(1 + |x|).$$

IV Functions $\Phi^l(x, t_l) : R^{n_l} \times \mathrm{T} \to R^1, l = 1, ..., r-1$ are continuously differentiable in respect to (x, t):

$$\left|\Phi^l(x, t_l)\right| + \left|\Phi_x^l(x, t_l)\right| \le N(1 + |x|).$$

V. Functions $q^r(x) : R^{n_r} \to R$ are continuously differentiable in respect to (x, t) :

$$\left|q^l(x)\right| + \left|q_x^l(x)\right| \le N(1 + |x|).$$

III. Methods of Solution

The following result that is a necessary condition of optimality for problem (1)-(5) has been obtained in [6].

Theorem 1 . Suppose that assumptions I-IV hold, $\pi^r = (t_0, .., t_r, x_t^1, ..., x_t^r, K_1, ..., K_r, u^1, ..., u^r)$ is an optimal solution of problem (1)-(5) and random processes $(\psi_t^l, \beta_t^l) \in L^2_{F^l}(t_{l-1}, t_l; R^{n_l}) \times L^2_{F^l}(t_{l-1}, t_l; R^{n_l x n_l})$ are the solutions of the following adjoint equations:

$$\begin{cases} d\psi_t^l = -\left[H_x^l(\psi_t^l, x_t^l, y_t^l, u_t^l, t) + H_y^l(\psi_{t+h}^l, x_{t+h}^l x_t^l, u_t^l, t)\right]dt \\ \qquad + \beta_t^l dw_t^l, \ \ t_{l-1} \le t < t_l - h, \\ d\psi_t^l = -H_x^l(\psi_t^l, x_t^l, y_t^l, u_t^l, t)dt + \beta_t^l dw_t^l, \ \ t_{l-1} - h_l \le t < t_l, \\ \psi_{t_l}^l = -\varphi_x^l(x_{t_l}^l) + \psi_{t_{l+1}}^l \Phi_x^l(x_{t_l}^l, t_l), \ l = 1, ..., r-1, \\ \psi_{t_r}^r = -\varphi^r(x_{t_r}^r). \end{cases}$$

Then ,

a) almost certainly for $\forall \tilde{u}^l \in U^l, l = 1, ...r,$ a.e. in $[t_{l-1}, t_l]$ the maximum principle hold:

$H^l(\psi_\theta^l, x_\theta^l, y_\theta^l, \tilde{u}^l, \theta) - H^l(\psi_\theta^l, x_\theta^l, y_\theta^l, u_\theta^l, \theta) \le 0$

b) Following transversality conditions hold:

$$-a_l \psi_{t_l}^l g^l\left(x_{t_l}^l, y_{t_l}^l, u_{t_l}^l, t_l\right) + b_l \psi_{t_l}^{l+1} g^{l+1}\left(x_{t_l}^l, y_{t_l}^l, u_{t_l}^l, t_l\right) + b_l \psi_{t_l+h}^{l+1} g^{l+1}\left(x_{t_l}^l, K^l(t_l), u_{t_l}^l, t_l\right) - b_l \Phi_t^l\left(x_{t_l}^l, t_l\right) = 0, \ l = 0, 1, ..., r$$

Here

$$H^l(\psi_t, x_t, y_t, u_t, t) = \psi_t g^l(x_t, y_t, u_t, t) + \beta_t f^l(x_t, y_t, t) - p^l(x_t, u_t, t), \ t \in [t_{l-1}, t_l],$$

$y_t^l = x_{t-t}^l$, $a_0 = 0, a_1 = ... = a_r = 1$ and

$b_0 = ... = b_{r-1} = 1, b_r = 0.$

Further, by applying Theorem 1 and Ekeland's Variational Principle it is obtained the necessary condition of optimality for stochastic control problem of switching systems with delay (1)-(6).

Theorem 2 . Suppose that, assumptions I-V hold, $\pi^r = (t_0, ..., t_r, x_t^1, x_t^2, ..., x_t^r, K_1, ..., K_r, u^1, u^2, ..., u^r)$ is a optimal solution of problem (1)-(6) and random processes $(\psi_t^l, \beta_t^l) \in L^2_{F^l}(t_{l-1}, t_l; R^{n_l}) \times L^2_{F^l}(t_{l-1}, t_l; R^{n_l x n_l})$ are the solutions of the following adjoint equations:

$$\begin{cases} d\psi_t^l = -\left[H_x^l(\psi_t^l, x_t^l, y_t^l, u_t^l, t) + H_y^l(\psi_{t+h}^l, x_{t+h}^l x_t^l, u_t^l, t)\right]dt \\ \qquad + \beta_t^l dw_t^l, \ \ t_{l-1} \le t < t_l - h, \\ d\psi_t^l = -H_x^l(\psi_t^l, x_t^l, y_t^l, u_t^l, t)dt + \beta_t^l dw_t^l, \ \ t_{l-1} - h \le t < t_l \\ \psi_{t_l}^l = -\lambda_l \varphi_x^l(x_{t_l}^l) + \psi_{t_{l+1}}^l \Phi_x^l(x_{t_l}^l, t_l), \ l = 1, ..., r-1 \\ \psi_{t_r}^r = -\lambda_0 \varphi_x^r(x_{t_r}^r) - \lambda_r q_x^r(x_{t_r}^r). \end{cases} \quad (7)$$

Then ,

a) almost certainly for $\forall \tilde{u}^l \in U^l, l = 1, ...r,$ a.e. in $[t_{l-1}, t_l]$ the maximum principle holds:

$$H^l(\psi_\theta^l, x_\theta^l, y_\theta^l, \tilde{u}^l, \theta) - H^l(\psi_\theta^l, x_\theta^l, y_\theta^l, u_\theta^l, \theta) \le 0 \quad (8)$$

b) following transversality condition holds a.c.:

b) Following transversality conditions hold:

ISBN: 978-1-61804-251-4

$$-a_l\psi^l_{t_l}g^l\left(x^l_{t_l},y^l_{t_l},u^l_{t_l},t_l\right)+b_l\psi^{l+1}_{t_l}g^{l+1}\left(x^l_{t_l},y^l_{t_l},u^l_{t_l},t_l\right)+$$
$$b_l\psi^{l+1}_{t_l+h}g^{l+1}\left(x^l_{t_l},K^l(t_l),u^l_{t_l},t_l\right)-b_l\Phi^l_t\left(x^l_{t_l},t_l\right)=0,\ l=0,1,...,r \quad (9)$$

where $a_0=0, a_1=....=a_r=1$ and $b_0=....=b_{r-1}=1, b_r=0.$

Proof. For any natural j let's introduce the approximating functional:

$$I_j(\mathbf{u})=S^l_j(\sum_{l=1}^{r}\left[E\varphi^l\left(x^l_{t_l}\right)+E\int_{t_{l-1}}^{t_l}p^l\left(x^l_t,u^l_t,t\right)dt\right],Eq^r(x^r_{t_r}))=$$
$$\min_{c^l_j\in\varepsilon}\sqrt{\sum_{l=1}^{r}\left|c^l_j-1/j-E\left[\varphi^l(x^l_{t_l})+\int_{t_{l-1}}^{t_l}p(x^l_t,u^l_t,t)dt\right]\right|^2+\left|y-Eq^r(x^r_{t_r})\right|^2}$$

Where $\varepsilon=\left\{c: c\le J^0, y\in G\right\}$ and J^0 is minimal value of the functional in the problem (1)-(5). Let $\mathbf{V}\equiv(V^1,...,V^r)$, here $V^k\equiv(U^k,d)$ be space of controls obtained by means of the following metric:

$d(u^k,v^k)=(l\otimes P)\left\{(t,\omega)\in[t_{k-1},t_k]\times\Omega: v^k_t\ne u^k_t\right\}.$

It is easy to prove the following fact:

Lemma 1. Assume that conditions I-IV hold, $u^{l,n}_t, l=1,...,r$ be the sequence of admissible controls from V^l, and $x^{l,n}_t$ be the sequence of corresponding trajectories of the system (1)-(3). If the following condition is met: $d(u^{l,n}_t,u^l_t)\to 0$.

Then

$$\lim_{n\to\infty}\left\{\sup_{t_{l-1}\le t\le t_l}E\left|x^{l,n}_t-x^l_t\right|^2\right\}=0$$

where x^l_t is a trajectory corresponding to an admissible controls $u^l_t, l=1,...,r$.

According to Ekeland's variational principle, there are controls such as; $u^{l,j}_t: d(u^{l,j}_t,u^l_t)\le\sqrt{\varepsilon^l_j}$ and for $\forall u^l_t\in V^l$ the following is achieved:

$$I_j(\mathbf{u}^j)\le I_j(\mathbf{u})+\sum_{l=1}^{r}\sqrt{\varepsilon^l_j}d(u^{l,j},u^l),\ \varepsilon^l_j=\frac{1}{j}.$$

This inequality means that $(t_0,t_1,...,t_r,x^{1,j}_t,...,x^{r,j}_t,K_1,...,K_r,u^{1,j}_t,...,u^{r,j}_t)$ is a solution of the following problem:

$$\begin{cases}J_j(\mathbf{u})=I_j(\mathbf{u}^j)+\sum_{l=1}^{r}\sqrt{\varepsilon^l_j}E\int_{t_{l-1}}^{t_l}\delta(u^l_t,u^{l,j}_t)dt\to\min\\ dx^l_t=g^l(x^l_t,y^l_t,u^l_t,t)dt+f^l(x^l_t,y^l_t,t)dw_t,\ t\in(t_{l-1},t_l]\\ x^{l+1}_t=\mathrm{K}^{l+1}(t), t\in[t_l-h,t_l),\ l=0,1,...,r-1,\\ x^{l+1}_{t_l}=\Phi^l\left(x^l_{t_l},t_l\right)\ \ l=1,...,r\\ x^1_{t_0}=x_0,\\ u^l_t\in U^l_\partial\end{cases}\quad(10)$$

Function $\delta(u,v)$ is determined in the following way:

$$\delta(u,v)=\begin{cases}0, u=v\\ 1, u\ne v.\end{cases}$$

Then according to the Theorem 1, it is obtained as follows:

1) there exist the random processes $\psi^{l,j}_t\in L^2_{F^l}(t_{l-1},t_l;R^{n_l})$,

$\beta^{l,j}_t\in L^2_{F^l}(t_{l-1},t_l;R^{n_l\times n_l})$, which are solutions of the following system

$$\begin{cases}d\psi^{l,j}_t=-H^l_x\left(\psi^{l,j}_t,x^{l,j}_t,y^{l,j}_t,u^{l,j}_t,t\right)dt-H^l_y\left(\psi^{l,j}_{t+h},x^{l,j}_{t+h},y^{l,j}_{t+h},u^{l,j}_{t+h},t+h\right)dt\\ +\beta^{l,j}_t dw_t,\ t\in\left[t_{l-1},t_l-h\right),\ l=1,...,r\\ d\psi^{l,j}_t=-H^l_x\left(\psi^{l,j}_t,x^{l,j}_t,y^{l,j}_t,u^{l,j}_t,t\right)dt+\beta^{l,j}_t dw_t,\ t\in\left[t_{l-1}-h,t_l\right),\\ \psi^{l,j}_{t_l}=-\lambda^j_0\varphi^l_x\left(x^{l,j}_{t_l}\right)+\psi^l_{t_{l+1}}\Phi^l_x(x^{l,j}_{t_l},t_l),\ l=1,...,r-1\\ \psi^r_{t_r}=-\lambda^j_0\varphi^r_x\left(x^{r,j}_{t_r}\right)-\lambda^j_r q^r_x\left(x^{r,j}_{t_r}\right).\end{cases}$$

(11)

where non-zero $(\lambda^j_0,\lambda^j_1,...,\lambda^j_r)\in R^{r+1}$ meet the following requirement:

$$\lambda^j_l=\left(-c_l+1/j+E\varphi^l\left(x^{l,j}_{t_l}\right)+E\int_{t_{l-1}}^{t_l}p^l(x^{l,j}_t,u^{l,j}_t,t)dt\right)/J^0_j$$

$$\lambda^j_r=Eq^r\left(x^{r,j}_{t_r}\right)/J^0_j$$

$$J^0_j=\sqrt{\sum_{l=1}^{r}\left|c_l-1/j-E\left[\varphi^l(x^l_{t_l})+\int_{t_{l-1}}^{t_l}p(x^l_t,u^l_t,t)dt\right]\right|^2+\left|y-Eq^r(x^r_{t_r})\right|^2}$$

2) almost certainly for any $\tilde{u}^l\in U^l$ and a.e. $t\in[t_{l-1},t_l]$ is satisfied:

$$H^l\left(\psi^{l,j}_t,x^{l,j}_t,y^{l,j}_t,\tilde{u}^l_t,t\right)-H^l\left(\psi^{l,j}_t,x^{l,j}_t,y^{l,j}_t,u^{l,j}_t,t\right)\le 0 \quad (12)$$

3) the following transversality conditions hold:

$$-a_l\psi^{l,j}_{t_l}g^l\left(x^{l,j}_{t_l},y^{l,j}_{t_l},u^{l,j}_{t_l},t_l\right)+b_l\psi^{l+1,j}_{t_l}g^{l+1}\left(x^{l,j}_{t_l},y^{l,j}_{t_l},u^{l,j}_{t_l},t_l\right)+$$
$$b_l\psi^{l+1,j}_{t_l+h}g^{l+1}\left(x^{l,j}_{t_l},K^{l,j}(t_l),u^{l,j}_{t_l},t_l\right)-b_l\Phi^l_t\left(x^{l,j}_{t_l},t_l\right)=0,\ l=0,1,...,r \quad (13).$$

Since the following exists $\left|(\lambda^j_0,\lambda^j_1,...,\lambda^j_r)\right|=1$, then according to conditions I-IV it is implied that

$$(\lambda^j_0,\lambda^j_1,...,\lambda^j_r)\to(\lambda_0,\lambda_1,...,\lambda_r)\ \text{ if }\ j\to\infty.$$

Let us introduce the following result which will be needed in the future.

Lemma 2. Let $\psi^l_{t_l}$ be a solution of system (7), and $\psi^{l,j}_{t_l}$ be a solution of system (11). Then

$$E\int_{t_{l-1}}^{t_l}|\psi_t^{l,j}-\psi_t^l|^2\,dt+E\int_{t_{l-1}}^{t_l}|\beta_t^{l,j}-\beta_t^l|^2\,dt\to 0,\ \text{if}$$

$$d(u_t^{l,j},u_t^l)\to 0,\ \ j\to\infty.$$

Proof: It is clear that $\forall t\in[t_{l-1},t_l], l=1,...,r-1$:

$$d\left(\psi_t^{l,j}-\psi_t^l\right)=-\left[H_x^l\left(\psi_t^{l,j},x_t^{l,j},y_t^{l,j},u_t^{l,j},t\right)-H_x^l\left(\psi_t^l,x_t^l,y_t^l,u_t^l,t\right)\right]dt$$
$$+\left(\beta_t^{l,j}-\beta_t^l\right)dw_t$$

According to Ito formula, for $\forall s\in[t_l-h,t_l]$ it is satisfied:

$$E|\psi_{t_l}^{l,j}-\psi_{t_l}^l|^2-E|\psi_s^{l,j}-\psi_s^l|^2=$$
$$2E\int_s^{t_r}[\psi_t^{l,j}-\psi_t^l][(g_x^{l*}(x_t^{l,j},y_t^{l,j},u_t^{l,j},t)-g_x^{l*}(x_t^l,y_t^l,u_t^l,t))\psi_t^{l,j}+$$
$$g_x^{l*}(x_t^l,y_t^l,u_t^l,t)(\psi_t^{l,j}-\psi_t^l)+(f_x^{l*}(x_t^{l,j},y_t^{l,j},t)-f_x^{l*}(x_t^l,y_t^l,t))\beta_t^{l,j}$$
$$-p^l(x_t^{l,j},u_t^{l,j},t)+p_x^l(x_t^l,u_t^l,t)]dt+E\int_s^{t_r}|\beta_t^{l,j}-\beta_t^l|^2\,dt.$$

Due to assumptions I-IV and using simple transformations, the following is obtained:

$$E\int_s^{t_l}|\beta_t^{l,j}-\beta_t^l|^2\,dt+E|\psi_s^{l,j}-\psi_s^l|^2\le EN\int_s^{t_l}|\psi_t^{l,j}-\psi_t^l|^2\,dt+$$
$$+EN\varepsilon\int_s^{t_r}|\beta_t^{l,j}-\beta_t^l|^2\,dt+E\left|\psi_{t_l}^{l,j}-\psi_{t_l}^l\right|^2.$$

Hence, according to Gronwall inequality [3] it suggests that:

$$E|\psi_s^{l,j}-\psi_s^l|^2\le De^{N(t_r-s)}\ \ \text{a.e. in}\ \ [t_l-h,t_l]$$

(14)

where constant D is determined in the way below:

$$D=E|\psi_{t_l}^{l,j}-\psi_{t_l}^l|^2.$$

According to (7) and (11), it is obtained that: $\psi_{t_l}^{l,j}\to\psi_{t_l}^l$,

which leads to $D\to 0$. Consequently, from (14) it follows:

$\psi_s^{l,j}\to\psi_s^l$ in $L^2_{F^l}(t_l-h,t_l;R^{n_l})$ and $\beta_s^{l,j}\to\beta_s^l$

in $L^2_{F^l}(t_l-h,t_l;R^{n_l\times n_l})$.

Then, $\forall t\in[t_{l-1},t_l-h), l=1,...,r$ from the expression:

$$d\left(\psi_t^{l,j}-\psi_t^l\right)=-\left[H_x^l\left(\psi_t^{l,j},x_t^{l,j},y_t^{l,j},u_t^{l,j},t\right)-H_x^l\left(\psi_t^l,x_t^l,y_t^l,u_t^l,t\right)\right]dt$$
$$-\left[H_y^l\left(\psi_{t+h}^{l,j},x_{t+h}^{l,j},y_{t+h}^{l,j},u_{t+h}^{l,j},t+h\right)-H_y^l\left(\psi_{t+h}^l,x_{t+h}^l,y_{t+h}^l,u_{t+h}^l,t+h\right)\right]dt$$
$$+\left(\beta_t^{l,j}-\beta_t^l\right)dw_t$$

using simple transformations , in view of assumptions I-IV the following is obtained:

$$E\int_s^{t_l-h}|\beta_t^{l,j}-\beta_t^l|^2\,dt+E|\psi_s^{l,j}-\psi_s^l|^2\le EN\int_s^{t_l-h}\left|\psi_t^{l,j}-\psi_t^l\right|^2dt+$$
$$+EN\varepsilon\int_s^{t_l-h}|\beta_t^{l,j}-\beta_t^l|^2\,dt+E|\psi_{t_l-h}^{l,j}-\psi_{t_l-h}^l|^2.$$

Hence, according to Gronwall inequality, the following result is achieved:

$$E|\psi_s^{l,j}-\psi_s^l|^2\le De^{N(t_l-s)}\ \ \text{a.e. in}\ \ [t_{l-1},t_l-h)$$

where constant D is determined as follows:

$D=E|\psi_{t_l-h}^{l,j}-\psi_{t_l-h}^l|^2$, which leads to $D\to 0$.

It is inferred that $\psi_s^{l,j}\to\psi_s^l$ in $L^2_{F^l}(t_{l-1},t_l;R^{n_l})$ and $\beta_s^{l,j}\to\beta_s^l$ in

$L^2_{F^l}(t_{l-1},t_l;R^{n_l\times n_l})$.

Lemma 2 is proved.

It follows from Lemma 2 that it can be proceeded to the limit in system (11) and the fulfilments of (7) are obtained. Following the similar scheme by taking limit in (12) and (13) it is proved that (8), (9) are true. Theorem 2 is proved.

IV. Conclusion

It is obtained a necessary condition of optimality for stochastic control problem of switching systems with delay on state. Necessary conditions satisfied by an optimal solution , play an important role for investigation of optimal control problems. The result can be used in various optimal control problems of biological, technical and economic systems. The necessary conditions developed in this study can be viewed as a stochastic analogues of the problems formulated in ([4]-[6]). However, Theorem 2 is a natural evolution of the results given in [10]-[13].

References

[1] V.B. Kolmanovsky, A.D. Myshkis, *Applied Theory of Functional Differential Equations*. Kluwer Academic Publishers, 1992.

[2] I.I. Gikhman, A.V. Skorokhod, *Stochastic Differential Equations.* Springer ,1972, (Translated from Russian).

[3] W.H. Fleming, R.W. Rishel, *Deterministic and Stochastic Optimal Control*. Springer, 1975.

[4] D.I. Capuzzo, L.C. Evans, "Optimal Switching for ordinary differential equations" , *SIAM J. on Control and Optimization*, vol.22, no 1, pp. 143-161, 1984.

[5] S.C. Bengea, A.C. Raymond, "Optimal Control of Switching systems", *Automatica* , vol.41, pp.11-27,2005.

[6] T. I. Seidmann,, "Optimal control for switching systems", in *Proc. 21st Annu. conference on informations science and systems*, 1987, pp.485-489.

[7] E.-K. Boukas, *Stochastic Switching Systems. Analysis and Design*, Birkhauer, 2006.

[8] G. Kharatatishvili, T.Tadumadze, "The problem of optimal control for nonlinear systems with variable structure, delays and piecewise continuous prehistory", *Memorirs on Differential Equations and Mathematical Physics*, Moskow,vol.11, pp. 67-88,1997,.

[9] H. Shen, Sh. Xu, X. Song, J. Luo, "Delay-dependent robust stabilization for uncertain stochastic switching systems with distributed delays", *Asian Journal of Control*, vol. 5, no 11, pp. 527-535,2009.

[10] Ch. Agayeva, Q.Abushov "Necessary condition of optimality for stochastic control systems with variable structure", in *Proc. 20th International Conference :Continuous, Optimization and Knowledge – based technologies"*, Lithuania, 2008, pp. 77-81.

[11] Ch. Agayeva, Q.Abushov "The maximum principle for some stochastic control problem of switching systems", *Selected Papers of International Conference, "MiniEURO 24rd Conference on Continuous Optimization @ Information-Based Technologies in Financial Sector*, Izmir, 2010, pp.100-105.

[12] Ch. Agayeva, Q. U.Abushov . "The maximum principle for some nonlinear stochastic control system with variable structure", *Theory of Stochastic Processes*, vol 16,no.1, pp.1-11, 2010.

[13] Ch.A Aghayeva, "Stochastic optimal control problem of switching systems with lag", *Transactions ANAS, mathematics and mechanics series of physical-technical & mathematical science*, vol.31, no.3, pp. 68-73, 2011.

[14] I. Ekeland, "On the variational principle", *Journal of Math. Anal. Appl.*, vol.47, pp.324-353, 1974.

Finding minimax strategy and minimax risk for Bernoulli multi-armed bandit

Alexander V. Kolnogorov

Abstract – Bernoulli multi-armed bandit with arbitrary set of unknown parameters is considered. Using the strategy variation, it is shown that minimax risk and strategy on the whole set of parameters are equal to those on some finite subset of its closure. According to the main theorem of the theory of games minimax strategy and minimax risk on the finite set of parameters are searched as Bayes' ones corresponding to the worst prior distribution. These properties allow to reduce the problem to finding the global maximum of the function depending on finite number of variables. On the other hand, they provide a convenient method to represent and to keep determined strategy. Some numerical examples are presented.

Keywords – multi-armed bandit problem, behavior in random environment, stochastic robust control, minimax and bayesian approaches, strategy variation, linear uniform inequalities.

I. Introduction

We consider Bernoulli multi-armed bandit problem (see, e.g. [1]). The name originates from the slot machine with K ($K \geq 2$) arms, corresponding model is considered in Section II. This is a sequential design problem. It is often considered as a control problem as well. The general setting assumes that there are K ($K \geq 2$) alternative actions. Each choice of any action generates a random binary income with possible values $\{1, 0\}$ which distribution depends on currently chosen arm only, is fixed during control process but unknown to the person choosing arms. The goal is to maximize the total expected income by identification the most profitable action and its preferable application.

Different models for the problem have been proposed depending on their possible applications. For example, the finite automata and stochastic automata with variable structure (see, e.g. [2 ,3]) were investigated in order to describe expedient behavior of biological systems. More effective control procedures for the problems of adaptive control were propose in [4, 5]. Models for optimization in economics were considered in [6].

We consider the minimax approach to the problem. The importance of minimax approach is due to its robustness. This approach was proposed in [7] which caused a significant interest to the problem. However, it turned out that there is no a direct method of finding minimax risk and minimax strategy.

[1]This work was supported in part by Project Part of the State Assignment in Field of Scientific Activity by the Ministry of Education and Science of Russian Federation, project no. 1.949.2014/K, and by Russian Foundation for Basic Research, project no. 13-01-00334-a.

On the other hand, Bayesian approach was successfully used by many researchers. The popularity of Bayesian approach is due to the possibility to calculate Bayes risk and Bayes strategy by dynamic programming technique. Minimax and Bayesian approaches are related by the main theorem of the theory of games. According to this theorem minimax risk is equal to Bayes risk calculated over the worst prior distribution corresponding to the maximum of Bayes risk. This theorem gives indirect method to find minimax risk and minimax strategy.

However, the worst prior distribution might have a complex structure and, hence, this approach would not make the problem easy to solve. Fortunately, it is not the case. Below we show by strategy variation that minimax risk and minimax strategy on the whole set of parameters are equal to those on some *finite subset* belonging to its closure. This property allows to reduce the problem to finding the global maximum of function depending on finite number of variables. On the other hand, it provides a convenient method to represent and to keep determined strategy. This approach was considered in [8, 9].

The structure of the paper is the following. In section II, we describe the model. In section III, we consider minimax and Bayesian approaches to the problem and their relation. Then we state the goal of the control. In section IV, we investigate some properties of the strategy and its variation. In section V, the reduction of the problem to finding minimax risk and minimax strategy on the finite subset of parameters is considered. In section VI, calculation and representation of the minimax risk and strategy are given.

II. The Model

In this section, we describe Bernoulli two-armed bandit, give formal setup of the problem and consider some examples of the sets of parameters.

A. Bernoulli Two-Armed Bandit

Bernoulli two-armed bandit is a slot machine with two arms. If the gambler chooses the ℓ-th arm he gets unit reward with probability p_ℓ or nothing with probability $q_\ell = 1 - p_\ell$ ($\ell = 1, 2$). The gambler has to play against the two-armed bandit N times totally (he knows this value) and his goal is to maximize (in some sense) his total expected income. Probabilities p_1, p_2 are fixed at play but unknown to the gambler. In the sequel, the arms are also called actions.

This problem is closely connected with a dilemma "Information vs Control". It states that the best control policy of the gambler is always to choose the arm corresponding to

ISBN: 978-1-61804-251-4

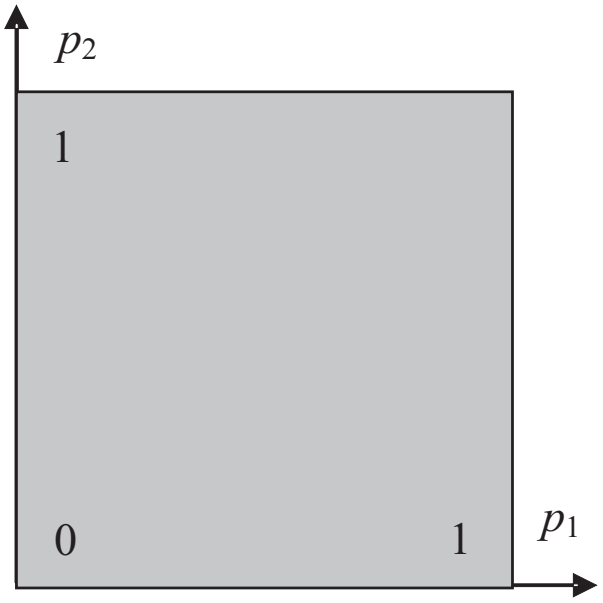


Fig. 1. The set of all Bernoulli two-armed bandits

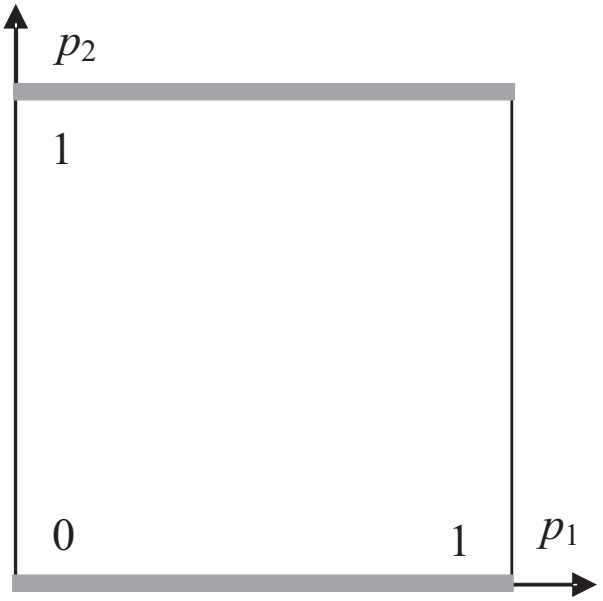


Fig. 2. p_1 is arbitrary, p_1 may take two possible values

the largest value of p_1, p_2. However, due to the lack of the information on this arm the gambler should try them both and this diminishes his total expected income.

B. Formal Setup

Formally, let ξ_n, $n = 1, \dots, N$, be a controlled random process with two possible values 1, 0, which are interpreted as incomes, depend on currently chosen actions η_n only and are described by the distribution

$$\Pr(\xi_n = 1|\eta_n = \ell) = p_\ell, \quad \Pr(\xi_n = 0|\eta_n = \ell) = q_\ell,$$

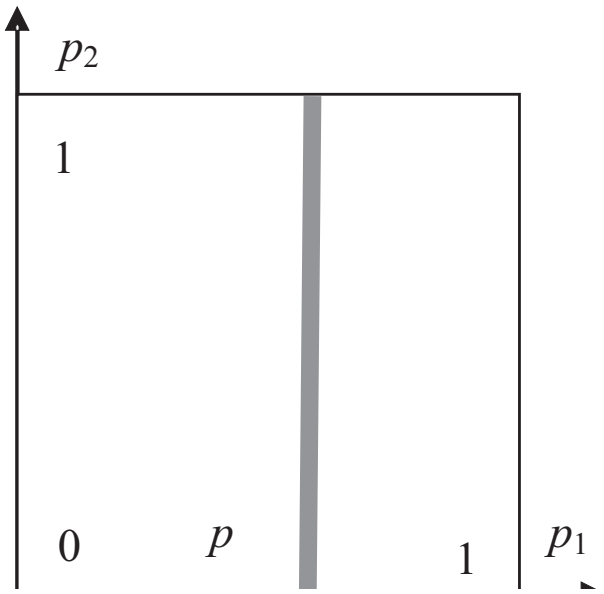


Fig. 3. The first arm is known and the second is arbitrary

where $p_\ell + q_\ell = 1$, $\ell = 1, \dots, K$, $K \geq 2$. Such multi-armed bandit can be described by a vector parameter $\theta = (p_1, \dots, p_K)$. Control strategy σ at the instant of time n is a function of the current history of the process $\xi^{n-1} = \xi_1, \dots, \xi_{n-1}$, $\eta^{n-1} = \eta_1, \dots, \eta_{n-1}$. Thus

$$\sigma_\ell(\eta^{n-1}, \xi^{n-1}) = \Pr(\eta_n = \ell|\eta^{n-1}, \xi^{n-1}),$$

$\ell = 1, \dots, K$. There is no history at $n = 1$ and, hence, it can be omitted in expressions below. Obviously, equalities hold

$$\sum_{\ell=1}^{K} \sigma_\ell(\eta^{n-1}, \xi^{n-1}) = 1 \tag{1}$$

for all histories (η^{n-1}, ξ^{n-1}). The set of strategies is denoted by Σ.

Let's describe the loss function. We assume that the goal of the control is to maximize (in some sense) the total expected income. Hence, if parameter θ is known, the optimal strategy prescribes always to apply the action corresponding to the largest value of $p_1, \dots, p_K$. The total expected income would thus be equal to $N \max\limits_{\ell=1\dots,K} p_\ell$. If parameter θ is unknown, then the function

$$L_N(\sigma, \theta) = N \max_{\ell=1\dots,K} p_\ell - \mathbb{E}_{\sigma,\theta} \left(\sum_{n=1}^{N} \xi_n) \right) \tag{2}$$

describes expected losses of total income due to the incomplete information. Here $\mathbb{E}_{\sigma,\theta}$ denotes the mathematical expectation over the measure generated by a strategy σ and a parameter θ. The set of possible values of parameter Θ is known. It can be any subset of K-dimensional cube $\{\theta : 0 \leq p_\ell \leq 1, \ \ell = 1, \dots, K\}$ which itself corresponds to the set of all possible multi-armed bandits.

ISBN: 978-1-61804-251-4

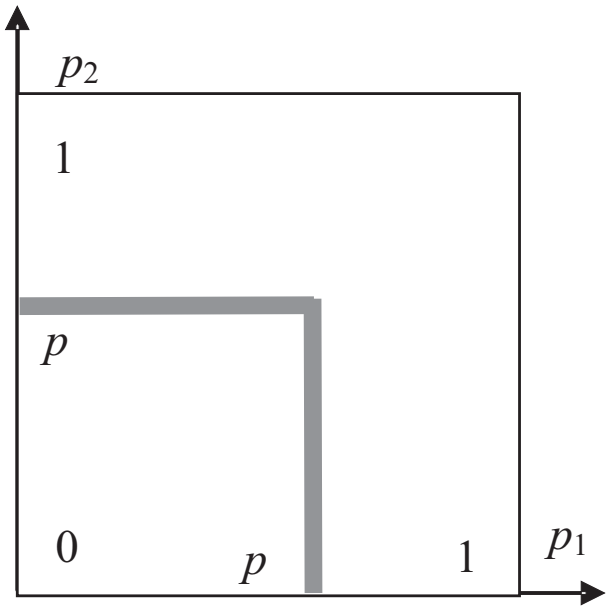


Fig. 4. Maximum of probabilities p_1, p_2 is known

C. Examples of Sets of Parameters

Different structures of the sets of parameters depend on prior information. Let's consider some examples with $K = 2$.

The set of all Bernoulli two-armed bandits is described by $\Theta = \{(p_1, p_2) : 0 \le p_1 \le 1, 0 \le p_2 \le 1\}$. This set is presented on Fig. 1.

On Fig. 2 we present the set of parameters considered by [1]. It can be formally described as $\Theta = \{(x, 0) : 0 \le x \le 1\} \bigcup \{(y, 1) : 0 \le y \le 1\}$. In this case, there is a priori information on the value p_2.

On Fig. 3 the so-called one-armed bandit is presented. In this case, the first arm is known and the second arm is unknown. The set of parameters can be formally described as $\Theta = \{(p, x) : 0 \le x \le 1\}$.

On Fig. 4 the set of Bernoulli two-armed bandits with maximal known probability p_1, p_2 is presented. The set of parameters can be formally described as $\Theta = \{(p, x) : 0 \le x \le p\} \cup \{(y, p) : 0 \le y \le p\}$.

III. The Goal of the Control

In this section, we consider minimax and Bayesian settings of the problem and relation between them.

A. Minimax Approach

According to the minimax approach the maximal total expected losses on the set of parameters Θ should be minimized over the set of strategies Σ. The value

$$R_N^M(\Theta) = \inf_{\Sigma} \sup_{\Theta} L_N(\sigma, \theta) \tag{3}$$

is called the minimax risk and corresponding strategy σ^M (if it exists) is called the minimax strategy.

Minimax setting of the problem was first proposed in [7]. The importance of minimax approach is due to its robustness. It means that if minimax strategy σ^M is applied then the following inequality holds

$$L_N(\sigma^M, \theta) \le R_N^M(\Theta),$$

for all $\theta \in \Theta$. The article [7] initiated a significant interest to considered problem. However, minimax strategy and minimax risk for Bernoulli two-armed bandit were not found in general case. For example, Fabius and van Zwet [10] calculated minimax strategy and minimax risk for $N = 1, 2, 3, 4$, and then wrote: "the algebra involved becomes progressively more complicated with increasing N and seems to remain prohibitive already for N as small as 5". On the other hand, an asymptotic minimax theorem [11] was proved:

$$0.265 \le N^{-1/2} R_N^M(\Theta) \le 0.376,$$

as $N \to \infty$. However, this theorem was proved by indirect method.

B. Bayesian Approach

Much more popular approach to the problem is Bayesian one. Denote by Λ a prior distribution of the parameter on the set Θ. The value

$$R_N^B(\Lambda) = \inf_{\Sigma} \int_{\Theta} L_N(\sigma, \theta) \Lambda(d\theta) \tag{4}$$

is called the Bayes risk and corresponding strategy σ^B is called the Bayes strategy. The popularity of Bayesian approach is because it allows to calculate Bayes strategy and Bayes risk by dynamic programming technique. As Berry and Fristedt [1] write: "it is not that researchers in bandit problems tend to "Bayesians"; rather Bayes's theorem provides a convenient mathematical formalism that allows for adaptive learning and so is an ideal tool in sequential decision problems".

On the other hand, a disadvantage of Bayesian approach is that there are no clear criteria for assignment a prior distribution. Very often researchers assign the prior distribution so that to simplify calculations. Sometimes it is possible to assign the prior distribution on the base of expert guidelines.

C. Relation between Minimax and Bayesian Approaches

Minimax and Bayesian approaches are related by the main theorem of the theory of games. According to this theorem minimax risk (3) is equal to Bayes risk (4) over the worst prior distribution.

However, the worst prior distribution might have a complex structure and, hence, this approach would not make the problem easy to solve. Fortunately, it is not the case. Below we show by the strategy variation that minimax risk and strategy on the whole set of parameters are equal to those on some *finite subset* belonging to its closure. This property allows to reduce the problem to finding global maximum of function depending on finite number of variables. On the other hand, it provides a convenient method to represent and to keep determined strategy.

ISBN: 978-1-61804-251-4

IV. Loss Function, Strategy and Strategy Variation

In this section, we give the formula for calculation the loss function (2) and show that it depends on sufficient statistics only. Then we modify presentation of the strategy, define strategy variation and obtain formula (12). Note that in case of the two-armed bandit these properties were considered in [10]. Finally, we partition all variations into permissible and prohibited ones and give a criterion which allows to distinguish between them.

A. Loss Function

Denote $r_\ell(0) = p_\ell$, $r_\ell(1) = q_\ell$, $\Delta_\ell = (\max\limits_{i=1,\dots,K} p_i - p_\ell)$, $\ell = 1, \dots, K$. Defined above loss function (2) is equal to

$$L_N(\sigma, \theta) = \sum_{n=1}^{N} M_n(\sigma, \theta) \tag{5}$$

with

$$\begin{aligned} M_n(\sigma,\theta) &= \sum_{\{\eta^{n-1};\xi^{n-1}\}} \Big(f(\eta^{n-1};\xi^{n-1};\theta) \\ &\times \sum_{\ell=1}^{K} \pi_\ell(\eta^{n-1};\xi^{n-1})\Delta_\ell \Big), \\ \pi_{\eta_n}(\eta^{n-1};\xi^{n-1}) &= \prod_{\nu=1}^{n} \sigma_{\eta_\nu}(\eta^{\nu-1};\xi^{\nu-1}), \\ f(\eta^n;\xi^n;\theta) &= \prod_{\nu=1}^{n} r_{\eta_\nu}(\xi_\nu). \end{aligned} \tag{6}$$

Here $\pi_\ell(\eta^{n-1};\xi^{n-1}) f(\eta^{n-1};\xi^{n-1};\theta)$ is the probability of the history $(\eta^{n-1}, \ell; \xi^{n-1})$. Probabilities $\{\pi_\ell(\eta^{n-1};\xi^{n-1})\}$ describe the strategy as well as $\{\sigma_\ell(\eta^{n-1};\xi^{n-1})\}$. Given probabilities $\{\pi_\ell(\eta^{n-1};\xi^{n-1})\}$, it follows from (6) that probabilities $\{\sigma_\ell(\eta^{n-1};\xi^{n-1})\}$ can be determined as

$$\sigma_\ell(\eta^{n-1};\xi^{n-1}) = \pi_\ell(\eta^{n-1};\xi^{n-1}) / \pi_{\eta_{n-1}}(\eta^{n-2};\xi^{n-2})$$

if $\pi_{\eta_{n-1}}(\eta^{n-2};\xi^{n-2}) \neq 0$; otherwise they can have arbitrary permissible values.

B. Sufficient Statistics

Note, that sufficient statistics of the history $(\eta^{n-1};\xi^{n-1})$ is $\zeta_{n-1} = ((n_1, m_1), \dots, (n_K, m_K))$ where (n_ℓ, m_ℓ) are equal to total number of choices and total number of nonzero incomes respectively for ℓ-th action ($\ell = 1, \dots, K$). Let's denote by S the operator $S : \eta^{n-1} \times \xi^{n-1} \to \zeta_{n-1}$ which calculates sufficient statistics ζ_{n-1} of the history (η^{n-1}, ξ^{n-1}) and by

$$\pi_\ell(\zeta_{n-1}) = \sum_{(\eta^{n-1};\xi^{n-1}) \in S^{-1}(\zeta_{n-1})} \pi_\ell(\eta^{n-1};\xi^{n-1})$$

the sum of probabilities $\pi_\ell(\cdot)$ corresponding to all histories having sufficient statistics ζ_{n-1}. Note, that in general it can not be interpreted as probability because its value can exceed 1. Then (6) can be rewritten as

$$\begin{aligned} M_n(\sigma,\theta) &= \sum_{\{\zeta_{n-1}\}} \left(F(\zeta_{n-1};\theta) \sum_{\ell=1}^{K} \pi_\ell(\zeta_{n-1})\Delta_\ell \right), \\ \pi_\ell(\zeta_{n-1}) &= \sum_{(\eta^{n-1};\xi^{n-1}) \in S^{-1}(\zeta_{n-1})} \pi_\ell(\eta^{n-1};\xi^{n-1}), \\ F(\zeta_{n-1};\theta) &= f(\eta^{n-1};\xi^{n-1};\theta) = \prod_{\ell=1}^{K} p_\ell^{m_\ell} q_\ell^{n_\ell - m_\ell}, \end{aligned} \tag{7}$$

The set of values $\{\pi_\ell(\zeta_{n-1})\}$ in (7) can be interpreted as the strategy π depending on sufficient statistics only. Let us determine corresponding strategy σ. It follows from (6) that

$$\begin{aligned} &\pi_\ell = \sigma_\ell, \quad \text{if } n = 1, \\ &\pi_\ell(\eta^{n-1};\xi^{n-1}) \\ &= \sigma_\ell(\eta^{n-1};\xi^{n-1}) \pi_{\eta_{n-1}}(\eta^{n-2};\xi^{n-2}), \quad \text{if } n > 1, \end{aligned} \tag{8}$$

$\ell = 1, \dots, K$. Denote by S_1 operator $S_1 : \zeta_{n-2} \times (\eta_{n-1}, \xi_{n-1}) \to \zeta_{n-1}$ which calculates sufficient statistics at the time point $(n-1)$ on the base of that one at the time point $(n-2)$ and the pair (η_{n-1}, ξ_{n-1}). Assigning in (8) equal values $\sigma_\ell(\zeta_{n-1})$ to all $\sigma_\ell(\eta^{n-1};\xi^{n-1})$ such that $S(\eta^{n-1};\xi^{n-1}) = \zeta_{n-1}$ we obtain at $n > 1$

$$\begin{aligned} &\pi_\ell(\zeta_{n-1}) = \sigma_\ell(\zeta_{n-1}) \\ &\times \sum_{k=1}^{K} \sum_{x=0}^{1} \sum_{\substack{\zeta_{n-2}: \\ S_1(\zeta_{n-2},k,x)=\zeta_{n-1}}} \pi_k(\zeta_{n-2}). \end{aligned} \tag{9}$$

Obviously, (9) allows to determine strategy σ as a function of sufficient statistics at $n > 1$. Since $\{\sigma_\ell(\zeta_{n-1})\}$ satisfy equalities (1), we obtain using (8) and (9)

$$\begin{aligned} &\sum_{\ell=1}^{K} \pi_\ell = 1, \quad \text{if} \quad n = 1, \\ &\sum_{\ell=1}^{K} \pi_\ell(\zeta_{n-1}) \\ &= \sum_{k=1}^{K} \sum_{x=0}^{1} \sum_{\substack{\zeta_{n-2}: \\ S_1(\zeta_{n-2},k,x)=\zeta_{n-1}}} \pi_k(\zeta_{n-2}), \quad \text{if} \quad n > 1 \end{aligned} \tag{10}$$

for all $\{\zeta_{n-1}\}$. Note that (10) completely describes both $\{\pi_\ell(\zeta_{n-1})\}$ and $\{\sigma_\ell(\zeta_{n-1})\}$. So, probabilities $\{\sigma_\ell(\zeta_{n-1})\}$ can be determined as

$$\begin{aligned} &\sigma_\ell = \pi_\ell, \quad \text{if } n = 1, \\ &\sigma_\ell(\zeta_{n-1}) = \pi_\ell(\zeta_{n-1}) \\ &\times \left(\sum_{k=1}^{K} \sum_{x=0}^{1} \sum_{\substack{\zeta_{n-2}: \\ S_1(\zeta_{n-2},k,x)=\zeta_{n-1}}} \pi_k(\zeta_{n-2}) \right)^{-1}, \quad \text{if } n > 1, \end{aligned}$$

unless appropriate expressions in brackets are equal to 0. Otherwise, probabilities $\{\sigma_\ell(\zeta_{n-1})\}$ can have arbitrary non-negative values satisfying (1).

In the sequel we consider strategies depending on sufficient statistics and use the following notations

$$\mathrm{L}_N(\pi, \theta) := L_N(\sigma, \theta), \ \mathrm{M}_N(\pi, \theta) := M_N(\sigma, \theta).$$

Note that values $\{\pi_\ell(\zeta_{n-1}); \ell = 1, \dots, K-1\}$ may be arbitrary positive values satisfying (10). Their total number is denoted by $S(K.N)$. Then $\{\pi_K(\zeta_{n-1})\}$ can be determined using (10).

ISBN: 978-1-61804-251-4

C. Strategy Variation

Let's define strategy variation δ by the set of numbers $\{\delta_\ell(\zeta_{n-1})\}$ satisfying conditions

$$\begin{aligned}&\sum_{\ell=1}^{K}\delta_\ell=0,\quad \text{if } n=1,\\&\sum_{\ell=1}^{K}\delta_\ell(\zeta_{n-1})\\&=\sum_{k=1}^{K}\sum_{x=0}^{1}\sum_{\substack{\zeta_{n-2}:\\ S_1(\zeta_{n-2},k,x)=\zeta_{n-1}}}\delta_k(\zeta_{n-2}),\quad \text{if } n>1\end{aligned}\tag{11}$$

for all $\{\zeta_{n-1}\}$.

Given strategy π and variation δ, the set of values $\pi+\delta$ defined as $\{\pi_\ell(\zeta_{n-1})+\delta_\ell(\zeta_{n-1})\}$ satisfy (10) and is therefore also a strategy if all these values are nonnegative numbers. In the sequel, we distinguish between permissible and forbidden variations. Given permissible variation, $\pi+\delta$ is also a strategy. Given forbidden variation, $\pi+\delta$ is not a strategy because of the negative values of some $\pi_\ell(\zeta_{n-1})+\delta_\ell(\zeta_{n-1})$. Note that any permissible variation δ can be represented as $\delta{=}\pi^{(2)}-\pi^{(1)}$ where $\pi^{(1)}$ and $\pi^{(2)}$ are both strategies.

Obviously, given strategy π, its permissible variation δ and $0<\varepsilon<1$, all variations $\varepsilon\delta$ are permissible variations as well. Using (7), one obtains

$$\mathrm{L}_N(\pi+\varepsilon\delta,\theta)=\mathrm{L}_N(\pi,\theta)+\varepsilon\mathrm{L}_N(\delta,\theta).\tag{12}$$

The set of variations satisfying conditions (11) generates a linear space. All variations can be described by their arbitrary values $\{\delta_\ell(\zeta_{n-1});\ell=1,\dots,K-1\}$ and $\{\delta_K(\zeta_{n-1})\}$ can then be determined using (11). Hence, variations δ having a single nonzero value $\delta_\ell(\cdot)=1$ ($\ell\in\{1,\dots,K-1\}$) and $\delta_\ell(\cdot)=0$ for the rest ones generate the basis in this linear space of all strategy variations.

The following lemma allows to distinguish permissible and forbidden variations.

Lemma 1. Given strategy π and variation δ, the variation $\varepsilon_0\delta$ is permissible for the strategy π for some $\varepsilon_0>0$ if and only if

$$\delta_\ell(\zeta_{n-1})\ge 0\quad\text{whenever}\quad \pi_\ell(\zeta_{n-1})=0.\tag{13}$$

Proof. The necessity of (13) is the consequence of the fact that otherwise for any $\varepsilon>0$ the negative values $\pi_\ell(\zeta_{n-1})+\varepsilon\delta_\ell(\zeta_{n-1})$ occur. Let's check the sufficiency of (13). It follows from (10), (11) that there is at least one $\pi_\ell(\zeta_{n-1})>0$ and at least one $\delta_\ell(\zeta_{n-1})<0$ if $\delta\ne 0$. Denote by

$$m=\min_{\{\pi_\ell(\zeta_{n-1})>0\}}\pi_\ell(\zeta_{n-1}),\ M=\max_{\{\delta_\ell(\zeta_{n-1})<0\}}|\delta_\ell(\zeta_{n-1})|.$$

Obviously, $m>0$, $M>0$ and for $\varepsilon_0=m/M>0$ all inequalities $\pi_\ell(\zeta_{n-1})+\varepsilon_0\delta_\ell(\zeta_{n-1})\ge 0$ hold. Hence, variation $\varepsilon_0\delta$ is permissible. All variations $\varepsilon\delta$, $0<\varepsilon<\varepsilon_0$ are permissible as well.

V. An Existence, a Criterion and Some Properties of the Minimax Strategy

In this section, we prove the existence of the minimax strategy. In less general case, it was done earlier in [10]. Then we give a necessary and sufficient criterion of the minimax strategy. It is written in terms of system of linear uniform inequalities. Given any nonzero permissible variation, this system contains at least one true inequality. It was proved in [12] that every such system contains some finite subsystem possessing the same property. This allows to reduce the problem to finding minimax risk and minimax strategy on some finite subset of original set of parameters.

A. An Existence of the Minimax Strategy

Given any strategy π and permissible variation δ, both $\mathrm{L}_N(\pi,\theta)$ and $\mathrm{L}_N(\delta,\theta)$ are continuous and bounded functions, so as $\mathrm{L}_N(\pi,\theta)\le N$, $|\mathrm{L}_N(\delta,\theta)|\le N$. Obviously, sets $\{\pi\}$ and Θ are bounded. The set $\{\pi\}$ is closed as well. Hence, $\{\pi\}$ is a compact set and Θ is a subset of a compact set. The following theorem holds.

Theorem 1. Given an arbitrary set of parameters Θ, there exists a minimax strategy. This strategy is minimax one on the closure of Θ as well.

Proof. The existence of the minimax risk (3) follows from the uniform boundedness of the loss function $\mathrm{L}_N(\pi,\theta)$. It follows from (3) that there exists a sequence of positive numbers $\{\varepsilon_k\}$ and strategies $\{\pi^{(k)}\}$ such that inequalities hold

$$\sup_\Theta \mathrm{L}_N(\pi^{(k)},\theta)<R_N(\Theta)+\varepsilon_k,$$

where $\varepsilon_k\downarrow 0$ as $k\to\infty$. Since the set of strategies is a compact set, one can choose $\{\pi^{(k)}\}$ to be a convergent sequence and let π^0 denote its limit. Since $\mathrm{L}_N(\pi,\theta)$ is a continuous function of π, one obtains as $k\to\infty$ the limiting inequality $\mathrm{L}_N(\pi^0,\theta)\le R_N(\Theta)$, $\theta\in\Theta$. Obviously, this inequality holds on the closure of Θ as well. In the sequel Θ is supposed to be a compact set.

Note that theorem does not state the uniqueness of the minimax strategy.

B. A Criterion of the Minimax Strategy

Given strategy π, let's consider a set of parameters

$$\Theta(\pi)=\mathrm{Argmax}_\theta\ \mathrm{L}_N(\pi,\theta).$$

Theorem 2. Strategy π is minimax one on Θ if and only if there exists, for any permissible variation δ, a parameter $\theta(\delta)\in\Theta(\pi)$ such that inequality holds

$$\mathrm{L}_N(\delta,\theta(\delta))\ge 0.\tag{14}$$

The set $\Theta^0=\bigcup_\delta\theta(\delta)$ is closed.

Proof. Let π be a minimax strategy and δ its permissible variation. Then $\varepsilon\delta$ is also permissible for any $0<\varepsilon<1$. Consider a sequence $\{\varepsilon_k\}$, such that $0<\varepsilon_k<1$ and $\varepsilon_k\downarrow 0$ as $k\to\infty$. Then for some sequence $\{\theta_k\}$ inequalities hold

$$\mathrm{L}_N(\pi+\varepsilon_k\delta,\theta_k)\ge R_N(\Theta)\ge\mathrm{L}_N(\pi,\theta_k),\ k=1,2,\dots$$

Since Θ is a compact set, one can choose $\{\theta_k\}$ to be a convergent sequence which limiting value is $\theta(\delta)$. Obviously, $\theta(\delta)\in\Theta(\pi)$. Then using (12) and subtracting right hand part of inequality from its left hand part, one obtains $\mathrm{L}_N(\delta,\theta_k)\ge 0$, $k=1,2,\dots$ and this results to (14) as $k\to\infty$.

ISBN: 978-1-61804-251-4

On the other hand, suppose that π is not a minimax strategy and consider a minimax strategy $\pi(\Theta)$. Then inequality holds

$$\max_{\Theta} \mathrm{L}_N(\pi(\Theta),\theta) < M = \max_{\Theta} \mathrm{L}_N(\pi,\theta)$$

and hence for each $\theta \in \Theta(\pi)$ the following inequality holds as well $\mathrm{L}_N(\pi(\Theta),\theta) - \mathrm{L}_N(\pi,\theta) = \mathrm{L}_N(\pi(\Theta) - \pi,\theta) < 0$. Hence, variation $\delta = \pi(\Theta) - \pi$ is permissible for strategy π and condition (14) does not hold.

Example 1. Let $N = 1$, $K = 2$, then

$$\mathrm{L}_1(\pi,\theta) = \begin{cases} \pi_2(p_1 - p_2) & \text{if } \ p_1 > p_2, \\ \pi_1(p_2 - p_1) & \text{if } \ p_2 > p_1. \end{cases}$$

Let Θ be the set of all possible parameters. Consider a strategy $\pi = (\pi_1 = 0.5, \pi_2 = 0.5)$. Then $\Theta(\pi) = \{\theta_1, \theta_2\}$, where $\theta_1 = (1,0)$, $\theta_2 = (0,1)$ and $\mathrm{L}_1(\pi,\theta_1) = \mathrm{L}_1(\pi,\theta_2) = 0.5$. Permissible variations are the following $\delta = \{\delta_1 = z, \delta_2 = -z\}, |z| \le 0.5$. Since $\mathrm{L}_1(\delta,\theta_1) + \mathrm{L}_1(\delta,\theta_2) = \delta_1 + \delta_2 = 0$, then at least one of inequalities $L_1(\delta,\theta_1) > 0$ or $L_1(\delta,\theta_2) > 0$ holds if $\delta \neq 0$. So, π is a minimax strategy for considered set Θ.

On the other hand, consider a strategy $\pi' = (\pi'_1 = 0.75, \pi'_2 = 0.25)$. Then $\Theta(\pi') = \{\theta_2\}$, $L_1(\pi',\theta_2) = 0.75$. Since $L_1(\delta,\theta_2) = \delta_2$, then for permissible variation $\delta = \{\delta_1 = z, \delta_2 = -z, 0 < z \le 0.25\}$ inequality $L_1(\delta,\theta_2) \ge 0$ does not hold. So, the strategy π' is not a minimax strategy for the set of parameters Θ.

C. Reduction of the Problem to Finding Minimax Risk and Minimax Strategy for the Finite Subset of Parameters

Given finite set of parameters $\{\theta_1,\dots,\theta_r\}$, minimax risk is equal to

$$R_N(\theta_1,\dots,\theta_r) = \min_{\{\pi\}} \max_{i=1,\dots,r} \mathrm{L}_N(\pi,\theta_i)$$

and minimax strategy, which is denoted by $\pi(\theta_1,\dots,\theta_r)$, satisfies inequalities

$$\mathrm{L}_N(\pi(\theta_1,\dots,\theta_r),\theta_i) \le R_N(\theta_1,\dots,\theta_r), \ i = 1,\dots,r.$$

We need the following well known result which is one of the corollaries of Helly's theorem concerning the coverings by convex sets and follows from results of [12]. Let's consider a system, possibly infinite, of linear uniform inequalities

$$\sum_{j=1}^{s} a_j(\theta)\delta_j \ge 0, \ \theta \in \Theta, \tag{15}$$

where Θ is the set of parameters and factors $a_j(\theta)$, $j = 1,\dots,s$, are arbitrary real numbers not all equal to zero. Let's put

$$c_j(\theta) = a_j(\theta)\left(\sum_{i=1}^{s} a_i(\theta)\right)^{-1/2}$$

and define vector $c(\theta) = (c_1(\theta),\dots,c_n(\theta))$.

Lemma 2. Let the set $C = \bigcup\limits_{\theta\in\Theta} c(\theta)$ be closed. Given any real numbers $\{\delta_j\}$ which are not all equal to zero, let system

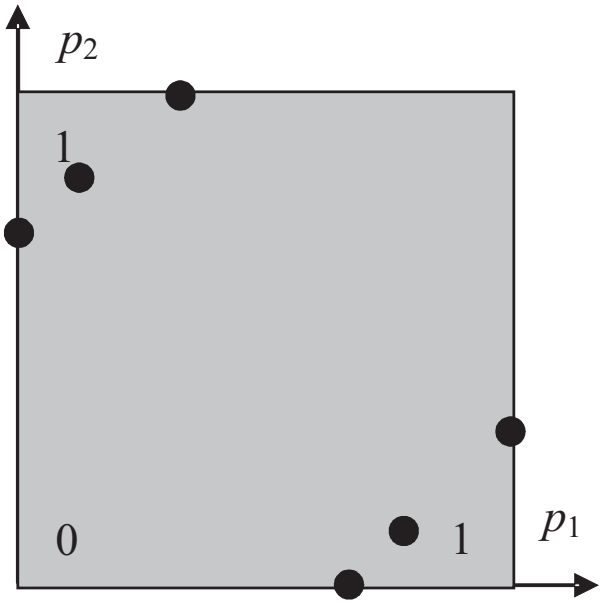


Fig. 5. The structure of the set of parameters, $N = 4$

(15) contain a true inequality for some $\theta \in \Theta$. Then there exists a finite subsystem

$$\sum_{j=1}^{s} a_j(\theta_i)\delta_j \ge 0, \ \theta_i \in \Theta, \ i = 1,\dots,S, \tag{16}$$

which belongs to system (15) and possesses the same property. That is, given any real numbers $\{\delta_j\}$ which are not all equal to zero, system (16) contains a true inequality for some θ_i, $i = 1,\dots,S$. The unimprovable estimate for S is the following $S \le s + 1$.

Theorem 3. Minimax strategy $\pi(\Theta)$ is equal to some minimax strategy $\pi(\theta_1^0,\dots,\theta_s^0)$ on the finite set of parameters. The minimax risk on this set of parameters achieves its maximal value, i.e.

$$R_N(\theta_1^0,\dots,\theta_s^0) = \max_{\{\theta_1,\dots,\theta_r\}} R_N(\theta_1,\dots,\theta_r), \tag{17}$$

and the estimate holds

$$s \le S(K,N) + 1. \tag{18}$$

Here $S(K,N)$ is a total number of arbitrary values $\{\delta_\ell(\zeta_{n-1}); \ell = 1,\dots,K-1\}$.

Proof. Suppose that variations are described by arbitrary values $\{\delta_\ell(\zeta_{n-1}); \ell = 1,\dots,K-1\}$ and $\{\delta_K(\zeta_{n-1})\}$ are expressed as their linear combinations according to (11). Hence, the set of variations generates $S(K,N)$-dimensional linear space. By lemma 1 all variations δ can be partitioned into two types, those ones that $\varepsilon\delta$ are forbidden for strategy π for any $\varepsilon > 0$ and those ones that $\varepsilon_0\delta$ are permissible for π for some $\varepsilon_0 > 0$.

Let $\pi(\Theta)$ be a minimax strategy and variation δ be such that $\varepsilon_0\delta$ is permissible for $\pi(\Theta)$ for some $\varepsilon_0 > 0$. It follows from theorem 2 that inequality holds

$$\mathrm{L}_N(\delta,\theta) \ge 0 \quad \text{for some } \theta(\delta) \in \Theta^0, \tag{19}$$

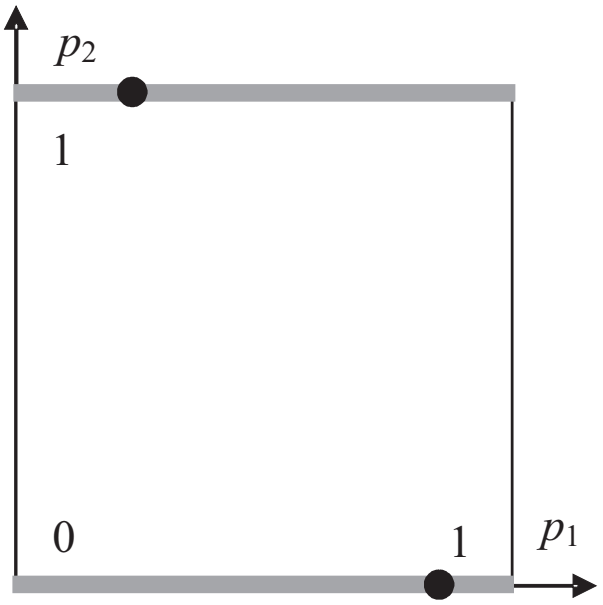


Fig. 6. The structure of the set of parameters $N = 3$

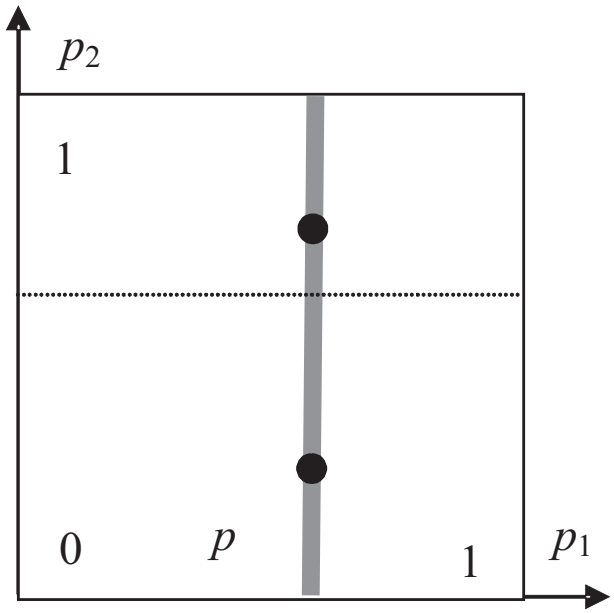


Fig. 7. The structure of the set of parameters $N = 10$

where Θ^0 is closed set. Obviously, corresponding set C_1 is closed.

Each forbidden variation satisfies to some inequality $-\delta_\ell(\zeta_{n-1}) > 0$, where ℓ is such that $\pi_\ell(\zeta_{n-1}) = 0$. We substitute this set by its closure

$$-\delta_\ell(\zeta_{n-1}) \geq 0 \quad \text{whenever} \quad \pi_\ell(\zeta_{n-1}) = 0. \tag{20}$$

According to previous remark all inequalities $-\delta_K(\zeta_{n-1}) > 0$ in the system (20) should be replaced by linear uniform inequalities of $\{\delta_\ell(\zeta_{n-1})\}$, $\ell = 1, \ldots, K-1$. Since (19) is a finite system, corresponding set C_2 is closed as well. Hence, the set $C = C_1 \bigcup C_2$ of the joint system (19), (20) is closed, too.

Since any nonzero variation δ satisfies either (20) or (19), conditions of lemma hold. Therefore a finite subsystem can be extracted from the system of inequalities (20), (19) which possesses the same property

$$\mathrm{L}_N(\delta, \theta_i^0) \geq 0, \ i = 1, \ldots s, \quad \theta_i \in \Theta^0, \tag{21}$$

$$-\delta_{\ell_i}(\zeta_{n-1}) \geq 0, \quad i = s+1, \ldots, S, \tag{22}$$

and contains $S \leq S(K, N) + 1$ inequalities. Since (21) covers the interior of permissible variations, it covers all permissible variations. Hence, any permissible strategy variation satisfies some inequality in (21). According to theorem 2 the strategy $\pi(\Theta)$ is minimax on the set of parameters θ_i^0, $i = 1, \ldots, r$ and the estimate (18) holds.

VI. Calculation and Representation of the Strategy

In this section, we give the algorithm of calculation minimax risk as Bayes one corresponding to the worst prior distribution. Then we give some examples.

Given a prior distribution $\lambda_1, \ldots, \lambda_r$ on the finite set of parameters $\theta_1, \ldots, \theta_r$, denote by

$$R_N^B(\lambda_1, \ldots, \lambda_r; \theta_1, \ldots, \theta_r) = \inf_{\{\pi\}} \sum_{i=1}^{r} \lambda_i L_N(\pi, \theta_i)$$

the corresponding Bayes risk. Note, that Bayes risk is a concave function of $\lambda_1, \ldots, \lambda_r$. According to the main theorem of the theory of games

$$R_N^M(\theta_1, \ldots, \theta_r) = \sup_{\{\lambda_1, \ldots, \lambda_r\}} R_N^B(\lambda_1, \ldots, \lambda_r; \theta_1, \ldots, \theta_r).$$

So, the problem is to find such set of parameters $\theta_1^0, \ldots, \theta_s^0 \in \Theta$ and corresponding prior distribution $\lambda_1^0, \ldots, \lambda_s^0$ which satisfy the equality

$$\begin{aligned} &R_N^B(\lambda_1^0, \ldots, \lambda_s^0; \theta_1^0, \ldots, \theta_s^0) \\ &= \max_{\{\theta_1, \ldots, \theta_r\}} \max_{\{\lambda_1, \ldots, \lambda_r\}} R_N^B(\lambda_1, \ldots, \lambda_r; \theta_1, \ldots, \theta_r), \end{aligned} \tag{23}$$

where all $R_N^B(\lambda_1, \ldots, \lambda_r; \theta_1, \ldots, \theta_r)$ are continuous functions of $\lambda_1, \ldots, \lambda_r$, $\theta_1, \ldots, \theta_r$ and, at fixed values $\theta_1, \ldots, \theta_r$, concave functions of $\lambda_1, \ldots, \lambda_r$. The number s is estimated as $s \leq S(K, N) + 1$.

Note that $\lambda_1^0, \ldots, \lambda_s^0$, $\theta_1^0, \ldots, \theta_s^0$ determined in (23) provide a convenient method to represent the strategy because it can be easily found by an appropriate computer program.

For example, in [10] the set $\Theta = \{(p_1, p_2) : 0 \leq p_1 \leq 1, 0 \leq p_2 \leq 1\}$ is considered and $N = 1, 2, 3, 4$. Then $\lambda_1^0 = \lambda_2^0 = 0.5$ on $\{\theta_1^0 = (0, 1), \theta_2^0 = (1, 0)\}$, $R_N^M(\Theta) = 0.5$ if $N = 1, 2$ and $\{\theta_1^0 = (0, 0.75), \theta_2^0 = (0.75, 0)\}$, $R_N^M(\Theta) = 9/16$ if $N = 3$. If $N = 4$ then $\{\theta_1^0 = (0, a), \theta_2^0 = (a, 0), \theta_3^0 = (1 - a, 1), \theta_4^0 = (1, 1 - a), \theta_5^0 = (0.7 - a, 0.7), \theta_6^0 = (0.7, 0.7 - a)\}$, where $a \approx 0.654$, and $\lambda_1^0 = \lambda_2^0 \approx 0.290$, $\lambda_3^0 = \lambda_4^0 \approx 0.160$, $\lambda_5^0 = \lambda_6^0 \approx 0.050$, $R_N^M(\Theta) \approx 0.617$. See Fig. 5.

In [1] the set $\Theta = \{(p_1, p_2) : 0 \leq p_1 \leq 1, p_2 = 0 \text{ or } p_2 = 1\}$ is considered and $N = 3$. In this case, $\{\theta_1^0 = (u, 0), \theta_2^0 =$

ISBN: 978-1-61804-251-4

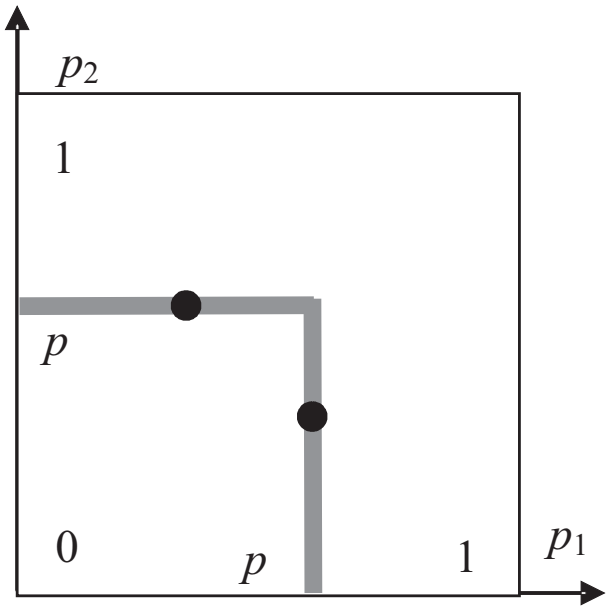


Fig. 8. The structure of the set of parameters $N = 10$

$(v, 1)\}$, where $u \approx 0.881$, $v \approx 0.218$ and $\lambda_1^0 \approx 0.591$, $\lambda_2^0 \approx 0.409$, $R_N^M(\Theta) \approx 0.521$. See Fig. 6.

Consider $\Theta = \{\theta = (p, x), 0 \leq x \leq 1\}$, i.e. $p_1 = p$ is known and p_2 can be any. If $p = 0.6$ and $N = 10$ then $\theta_1^0 = (p, p + b)$, $\theta_2^0 = (p, p - a)$, $\lambda_1^0 \approx 0.276$, $\lambda_2^0 \approx 0.724$, $a \approx 0.451$, $b \approx 0.209$, $R_N^M(\Theta) \approx 0.576$. See Fig. 7.

Finally, consider $\Theta = \{\theta = (p, x), 0 \leq x \leq p\} \bigcup \{\theta = (y, p), 0 \leq y \leq p\}$, i.e. maximum of p_1, p_2 is known. If $p = 0.6$ and $N = 10$ then $\theta_1^0 = (p, a)$, $\theta_2^0 = (a, p)$, $a \approx 0.210$, $\lambda_1^0 = \lambda_2^0 = 0.5$, $R_N^M(\Theta) \approx 0.880$. See Fig. 8.

REFERENCES

[1] D. A. Berry and B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London, New York, 1985.

[2] M. L. Tsetlin, *Automation Theory and Modeling of Biological Systems*. Academic Press, New York, 1973.

[3] V. I. Varshavsky, *Collective Behavior of Automata*. Nauka, Moscow, 1973. (In Russian)

[4] V. G. Sragovich, *Mathematical Theory of Adaptive Control*. Interdisciplinary Mathematical Sciences, Vol. 4. World Scientific. New Jersey, London, ..., 2006.

[5] A. V. Nazin and A. S. Poznyak , *Adaptive Choice of Alternatives*. Nauka, Moscow, 1986. (In Russian)

[6] E. L. Presman and I. M. Sonin, *Sequential Control with Incomplete Information*. Academic Press, New York, 1990.

[7] H. Robbins, Some aspects of the sequential design of experiments. *Bulletin AMS.*, Vol. 58(5), 1952, pp. 527–535.

[8] A. V. Kolnogorov, A minimax approach to optimal expedient behavior in stationary environments over finite time. *Sov. J. Comput. Syst. Sci.*, Vol. 27, No.4, 1989, pp. 33–35. (Translation from Russian)

[9] A. V. Kolnogorov, Determination of the minimax risk for Bernoulli multi-armed bandit. *IFAC Workshop "Adaptation and Learning in Control and Signal Processing ALCOSP'2010"*, Antalya, Turkey, August 26–28, 2010. pp. 237–242. DOI 10.3182/20100826-3-TR-4015.00045 Available online at http://www.ifac-papersonline.net.

[10] J. Fabius and W. R. van Zwet, Some remarks on the two-armed bandit. *Ann. Math. Statist.*, Vol. 41, 1970, pp. 1906–1916.

[11] W. Vogel, An asymptotic minimax theorem for the two-armed bandit problem. *Ann. Math. Stat.*, Vol. 31, 1960, pp. 444–451.

[12] L. M. Blumental, Metric methods in linear inequalities. *Duke Math. J.*, Vol. 15, 1948, pp. 955–966.

Thermochemical Non-Equilibrium Reentry Flows in Three-Dimensions: Seven Species Model – Part I – Structured Solutions

Edisson S. G. Maciel, Amilcar P. Pimenta and Nikos E. Mastorakis

***Abstract*—**This work presents a numerical tool implemented to simulate inviscid and viscous flows employing the reactive gas formulation of thermochemical non-equilibrium. The Euler and Navier-Stokes equations, employing a finite volume formulation, on the context of structured and unstructured spatial discretizations, are solved. These variants allow an effective comparison between the two types of spatial discretization aiming verify their potentialities: solution quality, convergence speed, computational cost, etc. The aerospace problem involving the hypersonic flow around a blunt body, in three-dimensions, is simulated. The reactive simulations will involve an air chemical model of seven species: N, O, N_2, O_2, NO, NO^+ and e^-. Eighteen chemical reactions, involving dissociation, recombination and ionization, will be simulated by the proposed model. This model was suggested by Blottner. The Arrhenius formula will be employed to determine the reaction rates and the law of mass action will be used to determine the source terms of each gas species equation. In this work is only presented the structured formulation and solutions. The unstructured formulation and solutions are presented in the second part of this study, which treats exclusively the unstructured context.**

***Keywords*—**Thermochemical non-equilibrium, Reentry flow, Seven species chemical model, Arrhenius formula, Structured and unstructured solutions, Euler and Navier-Stokes equations, Three-Dimensions**.**

I. Introduction

A HYPERSONIC flight vehicle has many applications for both military and civilian purposes including reentry vehicles such as the Space Shuttle and the Automated Transfer Vehicle (ATV) of the European Space Agency (ESA). The extreme environment of a hypersonic flow has a major impact on the design and analysis of the aerodynamic and thermal loading of a reentry or hypersonic cruise vehicle. During a hypersonic flight, the species of the flow field are vibrationally excited, dissociated, and ionized because of the very strong shock wave which is created around a vehicle. Because of these phenomena, it is necessary to consider the flow to be in thermal and chemical non-equilibrium.

In high speed flows, any adjustment of chemical composition or thermodynamic equilibrium to a change in local environment requires certain time. This is because the redistribution of chemical species and internal energies require certain number of molecular collisions, and hence a certain characteristic time. Chemical non-equilibrium occurs when the characteristic time for the chemical reactions to reach local equilibrium is of the same order as the characteristic time of the fluid flow. Similarly, thermal non-equilibrium occurs when the characteristic time for translation and various internal energy modes to reach local equilibrium is of the same order as the characteristic time of the fluid flow. Since chemical and thermal changes are the results of collisions between the constituent particles, non-equilibrium effects prevail in high-speed flows in low-density air.

In chemical non-equilibrium flows the mass conservation equation is applied to each of the constituent species in the gas mixture. Therefore, the overall mass conservation equation is replaced by as many species conservation equations as the number of chemical species considered. The assumption of thermal non-equilibrium introduces additional energy conservation equations – one for every additional energy mode. Thus, the number of governing equations for non-equilibrium flow is much bigger compared to those for perfect gas flow. A complete set of governing equations for non-equilibrium flow may be found in [1-2].

Analysis of non-equilibrium flow is rather complex because (1) the number of equations to be solved is much larger than the Navier-Stokes equations, and (2) there are additional terms like the species production, mass diffusion, and vibrational energy relaxation, etc., that appear in the governing equations. In a typical flight of the NASP (National AeroSpace Plane) flying at Mach 15, ionization is not expected to occur, and a 5-species air is adequate for the analysis (see [3]). Since the rotational characteristic temperatures for the constituent species (namely N, O, N_2, O_2 and NO) are small, the translational and rotational energy modes are assumed to be in equilibrium, whereas the vibrational energy mode is assumed to be in non-equilibrium.

Edisson S. G. Maciel works as a post-doctorate researcher at ITA (Aeronautical Technological Institute), Aeronautical Engineering Division – Praça Marechal do Ar Eduardo Gomes, 50 – Vila das Acácias – São José dos Campos – SP – Brazil – 12228-900 (corresponding author, phone number: +55 012 99165-3565; e-mail: edisavio@edissonsavio.eng.br).

Amilcar P. Pimenta teaches at ITA (Aeronautical Technological Institute), Aeronautical Engineering Division – Praça Marechal do Ar Eduardo Gomes, 50 – Vila das Acácias – São José dos Campos – SP – Brazil – 12228-900 (e-mail: amilcar@ita.br)

Nikos E. Mastorakis is with WSEAS (World Scientific and Engineering Academy and Society), A. I. Theologou 17-23, 15773 Zografou, Athens, Greece, E-mail: mastor@wseas.org as well as with the Technical University of Sofia, Industrial Engineering Department, Sofia, 1000, Bulgaria mailto:mastor@tu-sofia.bg

ISBN: 978-1-61804-251-4

[4] has simplified the thermodynamic model by assuming a harmonic oscillator to describe the vibrational energy. Ionic species and electrons are not considered. This simplifies the set of governing equations by eliminating the equation governing electron and electronic excitation energy. [4] has taken the complete set of governing equations from [1], and simplified them for a five-species two-temperature air model.

The problems of chemical non-equilibrium in the shock layers over vehicles flying at high speeds and high altitudes in the Earth's atmosphere have been discussed by several investigators ([5-8]). Most of the existing computer codes for calculating the non-equilibrium reacting flow use the one-temperature model, which assumes that all of the internal energy modes of the gaseous species are in equilibrium with the translational mode ([7-8]). It has been pointed out that such a one-temperature description of the flow leads to a substantial overestimation of the rate of equilibrium because of the elevated vibrational temperature [6]. A three-temperature chemical-kinetic model has been proposed by [9] to describe the relaxation phenomena correctly in such a flight regime. However, the model is quite complex and requires many chemical rate parameters which are not yet known. As a compromise between the three-temperature and the conventional one-temperature model, a two-temperature chemical-kinetic model has been developed ([10-11]), which is designated herein as the TT_v model. The TT_v model uses one temperature T to characterize both the translational energy of the atoms and molecules and the rotational energy of the molecules, and another temperature T_v to characterize the vibrational energy of the molecules, translational energy of the electrons, and electronic excitation energy of atoms and molecules. The model has been applied to compute the thermodynamic properties behind a normal shock wave in a flow through a constant-area duct ([10-11]). Radiation emission from the non-equilibrium flow has been calculated using the Non-equilibrium Air Radiation (NEQAIR) program ([12-13]). The flow and the radiation computations have been packaged into a single computer program, the Shock-Tube Radiation Program (STRAP) ([11]).

A first-step assessment of the TT_v model was made in [11] where it was used in computing the fAlow properties and radiation emission from the flow in a shock tube for pure nitrogen undergoing dissociation and weak ionization (ionization fraction less than 0.1%). Generally good agreement was found between the calculated radiation emission and those obtained experimentally in shock tubes ([14-16]). The only exception involved the vibrational temperature. The theoretical treatment of the vibrational temperature could not be validated because the existing data on the vibrational temperature behind a normal shock wave ([16]) are those for an electronically excited state of the molecular nitrogen ion N_2^+ instead of the ground electronic state of the neutral nitrogen molecule N_2 which is calculated in the theoretical model. The measured vibrational temperature of N_2^+ was much smaller than the calculated vibrational temperature for N_2.

This work, first of this study, describes a numerical tool to perform thermochemical non-equilibrium simulations of reactive flow in three-dimensions. The [17] scheme, in its first- and second-order versions, is implemented to accomplish the numerical simulations. The Euler and Navier-Stokes equations, on a finite volume context and employing structured and unstructured spatial discretizations, are applied to solve the "hot gas" hypersonic flow around a blunt body in two-dimensions. The second-order version of the [17] scheme is obtained from a "MUSCL" extrapolation procedure in a context of structured spatial discretization. In the unstructured context, only first-order solutions are obtained. The convergence process is accelerated to the steady state condition through a spatially variable time step procedure, which has proved effective gains in terms of computational acceleration (see [18-19]). In this paper only the structured formulation and results are presented.

The reactive simulations involve an air chemical model of seven species: N, O, N_2, O_2, NO, NO^+ and e^-. Eighteen chemical reactions, involving dissociation, recombination and ionization, are simulated by the proposed model. This model was suggested by [46]. The Arrhenius formula is employed to determine the reaction rates and the law of mass action is used to determine the source terms of each gas species equation.

The results have demonstrated that the shock position is closer to the geometry as using the reactive formulation, the stagnation pressure is better estimated by the [17] scheme, in its first-order, viscous, structured formulation, and the standoff distance is better predicted by its second-order, viscous, structured formulation.

II. Formulation to Reactive Flow in Thermochemical Non-Equilibrium

A. Reactive Equations in Three-Dimensions

The reactive Navier-Stokes equations in thermal and chemical non-equilibrium were implemented on a finite volume context, in the three-dimensional space. In this case, these equations in integral and conservative forms can be expressed by:

$$\frac{\partial}{\partial t}\int_V Q dV + \int_S \vec{F} \bullet \vec{n} dS = \int_V S_{CV} dV \text{ , with}$$
$$\vec{F} = (E_e - E_v)\vec{i} + (F_e - F_v)\vec{j} + (G_e - G_v)\vec{k} , \qquad (1)$$

where: Q is the vector of conserved variables, V is the volume of a computational cell, $\vec{F}$ is the complete flux vector, $\vec{n}$ is the unity vector normal to the flux face, S is the flux area, S_{CV} is the chemical and vibrational source term, E_e, F_e and G_e are the convective flux vectors or the Euler flux vectors in the x, y and z directions, respectively, E_v, F_v and G_v are the viscous flux vectors in the x, y and z directions, respectively. The $\vec{i}$, $\vec{j}$ and $\vec{k}$ unity vectors define the Cartesian coordinate system. Twelve (12) conservation equations are solved: one of general mass conservation, three of linear momentum conservation, one of total energy, six of species mass conservation and one of the vibrational internal energy of the molecules. Therefore, one of the species is absent of the iterative process. The CFD ("Computational Fluid Dynamics") literature recommends that

the species of biggest mass fraction of the gaseous mixture should be omitted, aiming to result in a minor numerical accumulation error, corresponding to the biggest mixture constituent (in the case, the air). To the present study, in which is chosen a chemical model to the air composed of seven (7) chemical species (N, O, N_2, O_2, NO, NO^+ and e^-) and eighteen (18) chemical reactions, being fifteen (15) dissociation reactions (endothermic reactions), two (2) of exchange or recombination, and one (1) of ionization, this species can be either the N_2 or the O_2. To this work, it was chosen the N_2. The vectors Q, E_e, F_e, G_e, E_v, F_v, G_v and S_{CV} can, hence, be defined as follows ([4]):

$$Q=\begin{Bmatrix}\rho\\ \rho u\\ \rho v\\ \rho w\\ e\\ \rho_1\\ \rho_2\\ \rho_4\\ \rho_5\\ \rho_6\\ \rho_7\\ \rho e_V\end{Bmatrix},\ E_e=\begin{Bmatrix}\rho u\\ \rho u^2+p\\ \rho uv\\ \rho uw\\ \rho Hu\\ \rho_1 u\\ \rho_2 u\\ \rho_4 u\\ \rho_5 u\\ \rho_6 u\\ \rho_7 u\\ \rho e_V u\end{Bmatrix},\ F_e=\begin{Bmatrix}\rho v\\ \rho vu\\ \rho v^2+p\\ \rho vw\\ \rho Hv\\ \rho_1 v\\ \rho_2 v\\ \rho_4 v\\ \rho_5 v\\ \rho_6 v\\ \rho_7 v\\ \rho e_V v\end{Bmatrix},\ G_e=\begin{Bmatrix}\rho w\\ \rho wu\\ \rho wv\\ \rho w^2+p\\ \rho Hw\\ \rho_1 w\\ \rho_2 w\\ \rho_4 w\\ \rho_5 w\\ \rho_6 w\\ \rho_7 w\\ \rho e_V w\end{Bmatrix};\quad (2)$$

$$E_v=\frac{1}{Re}\begin{Bmatrix}0\\ \tau_{xx}\\ \tau_{xy}\\ \tau_{xz}\\ \tau_{xx}u+\tau_{xy}v+\tau_{xz}w-q_{f,x}-q_{v,x}-\phi_x\\ -\rho_1 v_{1x}\\ -\rho_2 v_{2x}\\ -\rho_4 v_{4x}\\ -\rho_5 v_{5x}\\ -\rho_6 v_{6x}\\ -\rho_7 v_{7x}\\ -q_{v,x}-\phi_{v,x}\end{Bmatrix};\quad (3a)$$

in which: ρ is the mixture density; u, v and w are Cartesian components of the velocity vector in the x, y and z directions, respectively; p is the fluid static pressure; e is the fluid total energy; ρ_1, ρ_2, ρ_4, ρ_5, ρ_6, ρ_7 are densities of the N, O, O_2, NO, NO^+ and e^-, respectively; H is the mixture total enthalpy; e_v is the sum of the vibrational energy of the molecules; the τ's are the components of the viscous stress tensor; $q_{f,x}$, $q_{f,y}$ and $q_{f,z}$ are the frozen components of the Fourier-heat-flux vector in the x, y and z directions, respectively; $q_{v,x}$, $q_{v,y}$ and $q_{v,z}$ are the components of the Fourier-heat-flux vector calculated with the vibrational thermal conductivity and vibrational temperature; $\rho_s v_{sx}$, $\rho_s v_{sy}$ and $\rho_s v_{sz}$ represent the species diffusion flux, defined by the Fick law; ϕ_x, ϕ_y and ϕ_z are the terms of mixture diffusion; $\phi_{v,x}$, $\phi_{v,y}$ and $\phi v,z$ are the terms of molecular diffusion calculated at the vibrational temperature; $\dot{\omega}_s$ is the chemical source term of each species equation, defined by the law of mass action; e_v^* is the molecular-vibrational-internal energy calculated with the translational/rotational temperature; and τs is the translational-vibrational characteristic relaxation time of each molecule.

$$F_v=\frac{1}{Re}\begin{Bmatrix}0\\ \tau_{xy}\\ \tau_{yy}\\ \tau_{yz}\\ \tau_{xy}u+\tau_{yy}v+\tau_{yz}w-q_{f,y}-q_{v,y}-\phi_y\\ -\rho_1 v_{1y}\\ -\rho_2 v_{2y}\\ -\rho_4 v_{4y}\\ -\rho_5 v_{5y}\\ -\rho_6 v_{6y}\\ -\rho_7 v_{7y}\\ -q_{v,y}-\phi_{v,y}\end{Bmatrix};\quad (3b)$$

$$G_v=\frac{1}{Re}\begin{Bmatrix}0\\ \tau_{xz}\\ \tau_{yz}\\ \tau_{zz}\\ \tau_{xz}u+\tau_{yz}v+\tau_{zz}w-q_{f,z}-q_{v,z}-\phi_z\\ -\rho_1 v_{1z}\\ -\rho_2 v_{2z}\\ -\rho_4 v_{4z}\\ -\rho_5 v_{5z}\\ -\rho_6 v_{6z}\\ -\rho_7 v_{7z}\\ -q_{v,z}-\phi_{v,z}\end{Bmatrix};\quad (3c)$$

$$S_{CV}=\begin{Bmatrix}0\\ 0\\ 0\\ 0\\ 0\\ \dot{\omega}_1\\ \dot{\omega}_2\\ \dot{\omega}_4\\ \dot{\omega}_5\\ \dot{\omega}_6\\ \dot{\omega}_7\\ \sum_{s=mol}\rho_s\left(e_{v,s}^*-e_{v,s}\right)/\tau_s+\sum_{s=mol}\dot{\omega}_s e_{v,s}\end{Bmatrix},\quad (4)$$

ISBN: 978-1-61804-251-4

The viscous stresses, in N/m^2, are determined, according to a Newtonian fluid model, by:

$$\tau_{xx} = 2\mu\frac{\partial u}{\partial x} - \frac{2}{3}\mu\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}\right);$$
$$\tau_{yy} = 2\mu\frac{\partial v}{\partial y} - \frac{2}{3}\mu\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}\right);$$
$$\tau_{zz} = 2\mu\frac{\partial w}{\partial z} - \frac{2}{3}\mu\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}+\frac{\partial w}{\partial z}\right); \quad (5)$$
$$\tau_{xy} = \tau_{yx} = \mu\left(\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}\right);$$
$$\tau_{xz} = \tau_{zx} = \mu\left(\frac{\partial u}{\partial z}+\frac{\partial w}{\partial x}\right);$$
$$\tau_{yz} = \tau_{zy} = \mu\left(\frac{\partial w}{\partial y}+\frac{\partial v}{\partial z}\right), \quad (6)$$

in which μ is the fluid molecular viscosity.

The frozen components of the Fourier-heat-flux vector, which considers only thermal conduction, are defined by:

$$q_{f,x} = -k_f\frac{\partial T}{\partial x},\ q_{f,y} = -k_f\frac{\partial T}{\partial y},\ q_{f,z} = -k_f\frac{\partial T}{\partial z}, \quad (7)$$

where k_f is the mixture frozen thermal conductivity. The vibrational components of the Fourier-heat-flux vector are calculated as follows:

$$q_{v,x} = -k_v\frac{\partial T_v}{\partial x},\ q_{v,y} = -k_v\frac{\partial T_v}{\partial y},\ q_{v,z} = -k_v\frac{\partial T_v}{\partial z}, \quad (8)$$

in which k_v is the vibrational thermal conductivity and T_v is the vibrational temperature, what characterizes this model as of two temperatures: translational/rotational and vibrational.

The terms of species diffusion, defined by the Fick law, to a condition of thermal non-equilibrium, are determined by ([4]):

$$\rho_s v_{sx} = -\rho D_s\frac{\partial Y_{MF,s}}{\partial x},\ \rho_s v_{sy} = -\rho D_s\frac{\partial Y_{MF,s}}{\partial y};$$
$$\rho_s v_{sz} = -\rho D_s\frac{\partial Y_{MF,s}}{\partial z}, \quad (9)$$

with "s" referent to a given species, $Y_{MF,s}$ being the molar fraction of the species, defined as:

$$Y_{MF,s} = \frac{\rho_s/M_s}{\sum_{k=1}^{ns}\rho_k/M_k} \quad (10)$$

and Ds is the species-effective-diffusion coefficient.

The diffusion terms ϕ_x, ϕ_y and ϕ_z which appear in the energy equation are defined by ([20]):

$$\phi_x = \sum_{s=1}^{ns}\rho_s v_{sx} h_s,\ \phi_y = \sum_{s=1}^{ns}\rho_s v_{sy} h_s,\ \phi_z = \sum_{s=1}^{ns}\rho_s v_{sz} h_s, \quad (11)$$

being hs the specific enthalpy (sensible) of the chemical species "s". Details of the calculation of the specific enthalpy, see [21-22]. The molecular diffusion terms calculated at the vibrational temperature, $\phi_{v,x}$, $\phi_{v,y}$ and $\phi_{v,z}$ which appear in the vibrational-internal-energy equation are defined by ([4]):

$$\phi_{v,x} = \sum_{s=mol}\rho_s v_{sx} h_{v,s},\ \phi_{v,y} = \sum_{s=mol}\rho_s v_{sy} h_{v,s};\ \phi_{v,z} = \sum_{s=mol}\rho_s v_{sz} h_{v,s}, \quad (12)$$

with $h_{v,s}$ being the specific enthalpy (sensible) of the chemical species "s" calculated at the vibrational temperature Tv. The sum of Eq. (12), as also those present in Eq. (5), considers only the molecules of the system, namely: N_2, O_2, NO, and NO^+.

B. Thermodynamic Model/Thermodynamic Properties

Definition of general parameters.

$$p = RT\sum_{s=1}^{ns}\rho_s/M_s = \rho\sigma RT \therefore \rho\sigma = \sum_{s=1}^{ns}\rho_s/M_s = \rho\sum_{s=1}^{ns}c_s/M_s \Rightarrow$$
$$\sigma = \sum_{s=1}^{ns}c_s/M_s, \quad (13)$$

in which: σ is the mixture number in kg-mol/kg and cs is the mass fraction (non-dimensional), defined by $c_s = \rho_s/\rho$.

$$\sigma = \sum_{s=1}^{ns}\sigma_s \Rightarrow \sigma_s = c_s/M_s;$$
$$M_{mixt} = 1/\sigma \therefore M_{mixt} = 1\Big/\sum_{s=1}^{ns}c_s/M_s;$$
$$e^*_{v,s} = e_{v,s}(T_v = T), \quad (14)$$

with: σ_s being the number of kg-mol/kg of species "s" and M_{mixt} is the mixture molecular mass, in kg/kg-mol.

Thermodynamic model.

(a) Mixture translational internal energy:

$$e_T = \sum_{s=1}^{ns}e_{T,s}\sigma_s = \sum_{s=1}^{ns}\left[\int_0^T C_{v,T,s}(T')dT' + h^0\right]\sigma_s, \quad (15)$$

where: $e_{T,s}$ is the translational internal energy per kg-mol of species "s", in J/kg-mol. The specific heat at constant volume per kg-mol of species "s" due to translation, in J/(kg-mol.K), is

defined by:

$$C_{v,T,s}(T) = 1.5R. \quad (16)$$

Hence,

$$e_{T,s}(T) = 1.5RT + h^0 \Rightarrow e_T(T) = \sum_{s=1}^{ns} \sigma_s \left(1.5RT + h^0\right), \quad (17)$$

with: e_T being the translational internal energy per unity of the gaseous mixture mass, in J/kg, and h_0 being the formation enthalpy of the species "s" per kg-mol of species, J/kg-mol. It is important to note that:

$$e_T(T) = \sum_{s=1}^{ns} \sigma_s \left(1.5RT + h^0\right) = \sum_{s=1}^{ns} c_s \left(1.5\frac{R}{M_s}T + \frac{h^0}{M_s}\right) =$$

$$\sum_{s=1}^{ns} c_s \left(1.5R_s T + h_s^0\right) \Rightarrow e_T(T) = \sum_{s=1}^{ns} c_s \left(1.5R_s T + h_s^0\right), \quad (18)$$

with: R_s being the gas constant of species "s" and h_s^0 being the formation enthalpy of species "s" in J/kg. The species formation enthalpy per g-mol of species is specified in Tab. 1.

Table 1 Species formation enthalpy.

Species	h_0 (J/g-mol)
N	470,816.0
O	246,783.0
N_2	0.0
O_2	0.0
NO	90,671.0
NO^+	992,963.2
e^-	0.0

As can be noted, dividing each above term by the species molecular mass and multiplying by 10^3, it is possible to obtain the formation enthalpy in J/kg.

(b) Mixture rotational internal energy:

$$e_R = \sum_{s=1}^{ns} e_{R,s}\sigma_s = \sum_{s=mol} \left[\int_o^T C_{v,R,s}(T')dT'\right]\sigma_s =$$

$$\sum_{s=mol} \sigma_s \int_o^T C_{v,R,s}(T')dT', \quad (19)$$

where: $e_{R,s}$ is the rotational internal energy per kg-mol of species "s", in J/kg-mol. The specific heat at constant volume per kg-mol of species "s" due to rotation, in J/(kg-mol.K), is defined by:

$$C_{v,R,s} = R \Rightarrow e_{R,s}(T) = RT \therefore e_R(T) = \sum_{s=mol} \sigma_s RT$$

$$\text{or} \quad e_R(T) = \sum_{s=mol} c_s R_s T, \quad (20)$$

with e_R being the rotational internal energy per unity of gaseous mixture mass, in J/kg.

(c) Mixture vibrational internal energy:

$$e_V = \sum_{s=mol} e_{v,s}\sigma_s = \sum_{s=mol} \sigma_s \int_o^{T_V} C_{v,V,s}(T')dT'; \quad \text{with}$$

$$C_{v,V,s} = C_{v,V,s}(T_v) = R\frac{e^{\theta_{V,s}/T_V}}{\left(e^{\theta_{V,s}/T_V} - 1\right)^2}\left(\frac{\theta_{v,s}}{T_v}\right)^2, \quad (21)$$

in which: e_V is the vibrational internal energy per unity of gaseous mixture mass, in J/kg; $e_{v,s}$ is the vibrational internal energy per kg-mol of species "s", in J/kg-mol; $C_{v,V,s}$ is the specific heat at constant volume per kg-mol of species "s" due to vibration, in J/(kg-mol.K); $\theta_{v,s}$ is the characteristic vibrational+ temperature of species "s", in K; and T_v is the vibrational temperature, in K. The characteristic vibrational temperature to each molecule is specified in Tab. 2, obtained from [4]. It is important to note that eV is also directly obtained from the vector of conserved variables.

Table 2 Characteristic vibrational temperature of the molecular species.

Species	**N_2**	**O_2**	**NO**	**NO^+**
$\theta_{v,s}$ (K)	3,390.0	2,270.0	2,740.0	2,740.0

It is important to note that the modes of translational and rotational internal energy are assumed completely excited and, hence, the specific heats at constant volume to these modes are temperature independent. The vibrational-internal-energy mode is admitted not be completely excited, and, hence, the vibrational specific heat at constant volume is function of the vibrational temperature. The expression above to Cv,V,s is due to [23] and is the result of the hypothesis that the molecules can be considered as harmonic oscillators. Note that when the mode of vibrational internal energy is completely excited, i.e., when $T_v >> \theta_{v,s}$, $C_{v,V,s} = R$.

(c) Mixture internal energy:

$$e_{int} = e_T + e_R + e_V, \quad (22)$$

which is the internal energy per unity of mixture mass, in J/kg.

(d) Frozen speed of sound:

$$C_{v,TR} = \sum_{s=1}^{ns} \sigma_s C_{v,TR,s} = \sum_{s=1}^{ns} \sigma_s \begin{pmatrix} 2.5R & \text{molecules} \\ 1.5R & \text{atoms and } e^- \end{pmatrix};$$

$$\beta = R\sigma/C_{v,TR} \quad \therefore \quad a_f = \sqrt{(\beta+1)p/\rho}. \quad (23)$$

The frozen speed of sound, in a thermochemical non-equilibrium model, should be employed in the calculation of the convective flux of the [17] scheme. $C_{v,TR,s}$ is the specific heat at constant volume due to translation and rotation; in other words, $C_{v,TR,s}$ is the sum of $C_{v,T,s}$ with $C_{v,R,s}$.

ISBN: 978-1-61804-251-4

(e) Determination of the translational/rotational temperature:

$$\frac{e}{\rho}=\sum_{s=1}^{ns}c_s C_{v,TR,s}T+\sum_{s=1}^{ns}c_s h_s^0+e_V+\frac{1}{2}\left(u^2+v^2+w^2\right), \quad (24)$$

to the three-dimensional case. Hence, noting that T is constant at the right hand side of Eq. (24), it is possible to write:

$$T=\frac{1}{\sum_{s=1}^{ns}c_s C_{v,TR,s}}\left[\frac{e}{\rho}-\sum_{s=1}^{ns}c_s h_s^0-e_V-\frac{1}{2}\left(u^2+v^2+w^2\right)\right], \quad (25)$$

to the three-dimensional case;

(f) Determination of the vibrational temperature:

The vibrational temperature is calculated through an interactive process employing the Newton-Raphson method (a version to the five species model is found in [24]).

(g) Species pressure:

Applying the equation of a thermally perfect gas to each species:

$$p_s=\rho_s R_s T, \quad (26)$$

where: $\rho_s=c_s\rho$ is the density of species "s", Rs is the gas constant to species "s" and T is the translational/rotational temperature.

C. Transport Model/Transport Physical Properties

Collision integrals to species i and j. In Table 3 are presented values of $\mathrm{Log}_{10}\left[\pi\Omega_{i,j}^{(1,1)}\right]$ and $\mathrm{Log}_{10}\left[\pi\Omega_{i,j}^{(2,2)}\right]$ to temperature values of 2,000 K and 4,000 K. The indexes i and j indicate, in the present case, the collision partners; in other words, the pair formed by one atom and one atom, one atom and one molecule, etc. These data obtained from [1].

The data aforementioned define a linear interpolation to values of $\mathrm{Log}_{10}\left[\pi\Omega_{i,j}^{(k,k)}\right]$ as function of Ln(T), with k = 1, 2, through the linear equation:

$$\mathrm{Log}_{10}\left[\pi\Omega_{i,j}^{(k,k)}\right](T)=\mathrm{Log}_{10}\left[\pi\Omega_{i,j}^{(k,k)}\right](T=2{,}000\mathrm{K})+\mathrm{slope}\times\mathrm{Ln}\left(T/2{,}000\right), \quad (27)$$

in which:

$$\mathrm{slope}=\left\{\mathrm{Log}_{10}\left[\pi\Omega_{i,j}^{(k,k)}\right](T=4{,}000\mathrm{K})-\mathrm{Log}_{10}\left[\pi\Omega_{i,j}^{(k,k)}\right](T=2{,}000\mathrm{K})\right\}/\mathrm{Ln}2\,. \quad (28)$$

The value of $\pi\Omega_{i,j}^{(k,k)}$ is obtained from:

$$\pi\Omega_{i,j}^{(k,k)}(T)=e^{\left\{\mathrm{Log}_{10}\left[\pi\Omega_{i,j}^{(k,k)}\right](T=2{,}000\mathrm{K})+\mathrm{slope}\times\mathrm{Ln}\left(T/2{,}000\right)\right\}\times\mathrm{Ln}10}, \quad (29)$$

with the value of $\Omega_{i,j}^{(k,k)}$ in m^2.

Table 3 Collision integrals to five chemical species: N, O, N_2, O_2, NO, NO^+ and e^-.

Pairs		$\mathrm{Log}_{10}\left[\pi\Omega_{i,j}^{(1,1)}\right]$		$\mathrm{Log}_{10}\left[\pi\Omega_{i,j}^{(2,2)}\right]$	
i	j	2,000 K	4,000 K	2,000 K	4,000 K
N	N	-14.08	-14.11	-14.74	-14.82
N	O	-14.76	-14.86	-14.69	-14.80
N	N_2	-14.67	-14.75	-14.59	-14.66
N	O_2	-14.66	-14.74	-14.59	-14.66
N	NO	-14.66	-14.75	-14.67	-14.66
N	NO^+	-14.34	-14.46	-14.38	-14.50
N	e^-	-15.30	-15.30	-15.30	-15.30
O	N	-14.76	-14.86	-14.69	-14.80
O	O	-14.11	-14.14	-14.71	-14.79
O	N_2	-14.63	-14.72	-14.55	-14.64
O	O_2	-14.69	-14.76	-14.62	-14.69
O	NO	-14.66	-14.74	-14.59	-14.66
O	NO^+	-14.34	-14.46	-14.38	-14.50
O	e^-	-15.94	-15.82	-15.94	-15.82
N_2	N	-14.67	-14.75	-14.59	-14.66
N_2	O	-14.63	-14.72	-14.55	-14.64
N_2	N_2	-14.56	-14.65	-14.50	-14.58
N_2	O_2	-14.58	-14.63	-14.51	-14.54
N_2	NO	-14.57	-14.64	-14.51	-14.56
N_2	NO^+	-14.34	-14.46	-14.38	-14.50
N_2	e^-	-15.11	-15.02	-15.11	-15.02
O_2	N	-14.66	-14.74	-14.59	-14.66
O_2	O	-14.69	-14.76	-14.62	-14.69
O_2	N_2	-14.58	-14.63	-14.51	-14.54
O_2	O_2	-14.60	-14.64	-14.54	-14.57
O_2	NO	-14.59	-14.63	-14.52	-14.56
O_2	NO^+	-14.34	-14.46	-14.38	-14.50
O_2	e^-	-15.52	-15.39	-15.52	-15.39
NO	N	-14.66	-14.75	-14.67	-14.66
NO	O	-14.66	-14.74	-14.59	-14.66
NO	N_2	-14.57	-14.64	-14.51	-14.56
NO	O_2	-14.59	-14.63	-14.52	-14.56
NO	NO	-14.58	-14.64	-14.52	-14.56
NO	NO^+	-14.18	-14.22	-14.38	-14.50
NO	e^-	-15.30	-15.08	-15.30	-15.08
NO^+	N	-14.34	-14.46	-14.38	-14.50
NO^+	O	-14.34	-14.46	-14.38	-14.50
NO^+	N_2	-14.34	-14.46	-14.38	-14.50
NO^+	O_2	-14.34	-14.46	-14.38	-14.50
NO^+	NO	-14.18	-14.22	-14.38	-14.50
NO^+	NO^+	-11.70	-12.19	-11.49	-11.98
NO^+	e^-	-11.70	-12.19	-11.49	-11.98

ISBN: 978-1-61804-251-4

Table 3 Collision integrals to five chemical species: N, O, N2, O2, NO, NO+ and e-. (Continuation)

Pairs		$\text{Log}_{10}\left[\pi\Omega_{i,j}^{(1,1)}\right]$		$\text{Log}_{10}\left[\pi\Omega_{i,j}^{(2,2)}\right]$	
i	j	2,000 K	4,000 K	2,000 K	4,000 K
e^-	N	-15.30	-15.30	-15.30	-15.30
e^-	O	-15.94	-15.82	-15.94	-15.82
e^-	N_2	-15.11	-15.02	-15.11	-15.02
e^-	O_2	-15.52	-15.39	-15.52	-15.39
e^-	NO	-15.30	-15.08	-15.30	-15.08
e^-	NO^+	-11.70	-12.19	-11.49	-11.98
e^-	e^-	-11.70	-12.19	-11.49	-11.98

Modified collision integrals to the species i and j. [1] and [4] define the modified collision integrals to the species i and j as:

$$\Delta_{i,j}^{(1)}(T)=\frac{8}{3}\sqrt{\frac{2m_{i,j}}{\pi RT}}\pi\Omega_{i,j}^{(1,1)} \quad \text{and} \quad \Delta_{i,j}^{(2)}(T)=\frac{16}{5}\sqrt{\frac{2m_{i,j}}{\pi RT}}\pi\Omega_{i,j}^{(2,2)}, \tag{30}$$

with:

$$m_{i,j}=M_iM_j/\left(M_i+M_j\right), \tag{31}$$

being the reduced molecular mass. These integrals are given in m.s. With the definition of the modified collision integrals to species i and j, it is possible to define the mixture transport properties (viscosity and thermal conductivities) and the species diffusion property (diffusion coefficient).

Mixture molecular viscosity. [4] define the mixture molecular viscosity as:

$$\mu_{mixt}=\sum_{i=1}^{ns}\frac{m_i\sigma_i}{\sum_{j=1}^{ns}\sigma_j\Delta_{i,j}^{(2)}(T)}, \tag{32}$$

where:

$$m_i=M_i/N_{AV}, \tag{33}$$

being the mass of a species particle under study. N_{AV} = 6.022045x10^{23} particles/g-mol, Avogadro number. This mixture molecular viscosity is given in kg/(m.s).

Vibrational, frozen, rotational and translational thermal conductivities. All thermal conductivities are expressed in J/(m.s.K). [4] defines the mixture vibrational, rotational and translational thermal conductivities, as also the species diffusion coefficient, as follows.

(a) Translational thermal conductivity:
The mode of translational internal energy is admitted completely excited; hence, the thermal conductivity of the translational internal energy is determined by:

$$k_T=\frac{15}{4}k_{Boltzmann}\sum_{i=1}^{ns}\frac{\sigma_i}{\sum_{j=1}^{ns}\bar{a}_{i,j}\sigma_j\Delta_{i,j}^{(2)}(T)}, \tag{34}$$

in which:

$$k_{Boltzmann}=\text{Boltzmann constant}=1.380622\text{x}10^{-23}\text{J/K};$$
$$\bar{a}_{i,j}=1+\frac{(1-M_i/M_j)\left[0.45-2.54\left(M_i/M_j\right)\right]}{\left(1+M_i/M_j\right)^2}. \tag{35}$$

(b) Rotational thermal conductivity:
The mode of rotational internal energy is also considered fully excited; hence, the thermal conductivity due to rotational internal energy is defined by:

$$k_R=k_{Boltzmann}\sum_{i=mol}\frac{\sigma_i}{\sum_{j=1}^{ns}\sigma_j\Delta_{i,j}^{(1)}(T)}. \tag{36}$$

(c) Frozen thermal conductivity:

$$k_f=k_T+k_R. \tag{37}$$

(d) Thermal conductivity due to molecular vibration:
The mode of vibrational internal energy, however, is assumed be partially excited; hence, the vibrational thermal conductivity is calculated according to [3] by:

$$k_V=k_{Boltzmann}\sum_{i=mol}\frac{\left(C_{v,V,i}/R\right)\sigma_i}{\sum_{j=1}^{ns}\sigma_j\Delta_{i,j}^{(1)}(T)}, \tag{38}$$

with $C_{v,V,i}$ obtained from Eq. (21).

Species diffusion coefficient. The mass-diffusion-effective coefficient, Di, of the species "i" in the gaseous mixture is defined by:

$$D_i=\frac{\sigma^2M_i\left(1-\sigma_iM_i\right)}{\sum_{j=1}^{ns}\sigma_j/D_{i,j}} \quad \text{and} \quad D_{i,j}=\frac{k_{Boltzmann}T}{p\Delta_{i,j}^{(1)}(T)}, \tag{39}$$

where: Di,j is the binary diffusion coefficient to a pair of particles of the species "i" and "j" and is related with the modified collision integral conform described above, in Eq. (39). This coefficient is measured in m2/s.

D. Chemical Model

The chemical model employed to this case of thermochemical non-equilibrium is the seven species model of [46], using the

N, O, N_2, O_2, NO, NO^+ and e^- species. This formulation uses, in the calculation of the species production rates, a temperature of reaction rate control, introduced in the place of the translational/rotational temperature, which is employed in the calculation of such rates. This procedure aims a couple between vibration and dissociation. This temperature is defined as: $T_{rrc} = \sqrt{T \times T_v}$, where T is the translational/rotational temperature and T_v is the vibrational temperature. This temperature T_{rrc} replaces the translational/rotational temperature in the calculation of the species production rates, according to [25].

Law of Mass Action. The symbolic representation of a given reaction in the present work follows the [26] formulation and is represented by:

$$\sum_{s=1}^{ns} \upsilon'_{sr} A_s \leftrightarrow \sum_{s=1}^{ns} \upsilon''_{sr} A_s \ , \ r = 1,..., nr. \qquad (40)$$

The law of mass action applied to this system of chemical reactions is defined by:

$$\dot{\omega}_s = M_s \sum_{r=1}^{nr} \left(\upsilon''_{sr} - \upsilon'_{sr}\right) \left\{ k_{fr} \prod_{s=1}^{ns} \left(\frac{\rho_s}{M_s}\right)^{\upsilon'_{sr}} - k_{br} \prod_{s=1}^{ns} \left(\frac{\rho_s}{M_s}\right)^{\upsilon''_{sr}} \right\}, \qquad (41)$$

where A_s represents the chemical symbol of species "s", "ns" is the number of species of the present study (reactants and products) involved in the considered reaction; "nr" is the number of reactions considered in the chemical model; υ'_{sr} e υ''_{sr} are the stoichiometric coefficients to reactants and products, respectively; $k_{fr} = AT^B e^{-C/T}$ and $k_{br} = DT^{-E}$, with A, B, C, D and E being constants of a specific chemical reaction under study ["fr" = forward reaction and "br" = backward reaction].

Table 4. Chemical reactions and forward coefficients.

Reaction	Forward reaction rate coefficients, kfr, cm3/(mol.s)	Third body
O_2+M↔2O+M	3.61x1018T-1.0e(-59,400/T)	O, N, O_2, N_2, NO
N_2+M↔2N+M	1.92x1017T-0.5e(-113,100/T)	O, O_2, N_2, NO
N_2+N↔2N+N	4.15x1022T-0.5e(-113,100/T)	-
NO+M↔N+O+M	3.97x1020T-1.5e(-75,600/T)	O, N, O_2, N_2, NO
NO+O↔O_2+N	3.18x109T1.0e(-19,700/T)	-
N_2+O↔NO+N	6.75x1013e(-37,500/T)	-
N+O↔NO++e-	9.03x109e(-32,400/T)	-

Table 5. Chemical reactions and backward coefficients.

Reaction	Backward reaction rate coefficients, kbr, cm3/(mol.s) or cm6/(mol2.s)	Third body
O_2+M↔2O+M	3.01x1015T-0.5	O, N, O_2, N_2, NO
N_2+M↔2N+M	1.09x1016T-0.5	O, O_2, N_2, NO
N_2+N↔2N+N	2.32x1021T-0.5	-
NO+M↔N+O+M	1.01x1020T-1.5	O, N, O_2, N_2, NO
NO+O↔O2+N	9.63x1011T0.5e(-3,600/T)	-
N_2+O↔NO+N	1.5x1013	-
N+O↔NO++e-	1.80x1019Tv-1.0	-

It is important to note that $k_{br} = k_{fr} / k_{er}$, with ker being the equilibrium constant which depends only of the thermodynamic quantities. In this work, ns = 7 and nr = 18. Table 4 presents the values to A, B, C, D and E for the forward reaction rates of the 18 chemical reactions. Table 5 presents the values to A, B, C, D and E for the backward reaction rates. The eighth equation takes into account the formation of an electron from the ionization of the NO. For this case, the backward reaction rate depends only of the vibrational temperature.

E. Vibrational Model

The vibrational internal energy of a molecule, in J/kg, is defined by:

$$e_{v,s} = \frac{R_s \theta_{v,s}}{e^{\theta_{V,s}/T_V} - 1}, \qquad (42)$$

obtained by the integration of Eq. (21), and the vibrational internal energy of all molecules is given by:

$$e_V = \sum_{s=mol} c_s e_{v,s} \ . \qquad (43)$$

The heat flux due to translational-vibrational relaxation, according to [27], is given by:

$$q_{T-V,s} = \rho_s \frac{e^*_{v,s}(T) - e_{v,s}(T_v)}{\tau_s}, \qquad (44)$$

where: $e^*_{v,s}$ is the vibrational internal energy calculated at the translational temperature to the species "s"; and τ_s is the translational-vibrational relaxation time to the molecular species, in s. The relaxation time is the time of energy exchange between the translational and vibrational molecular modes.

ISBN: 978-1-61804-251-4

Vibrational characteristic time of [28]. According to [28], the relaxation time of molar average of [29] is described by:

$$\tau_s = \tau_s^{M-W} = \sum_{l=1}^{ns} X_l \Big/ \sum_{l=1}^{ns} X_l / \tau_{s,l}^{M-W} , \qquad (45)$$

with:

$\tau_{s,l}^{M-W}$ is the relaxation time between species of [29];

τ_s^{M-W} is the vibrational characteristic time of [29];

$X_l = c_l / (N_{AV} m_l)$ and $m_l = M_l / N_{AV}$. (46)

Definition of $\tau_{s,l}^{M-W}$**.** For temperatures inferior to or equal to 8,000 K, [29] give the following semi-empirical correlation to the vibrational relaxation time due to inelastic collisions:

$$\tau_{s,l}^{M-W} = \left(\frac{B}{p_l}\right) e^{\left[A_{s,l}\left(T^{-1/3} - 0.015\mu_{s,l}^{1/4}\right) - 18.42\right]}, \qquad (47)$$

where:

$B = 1.013 \times 10^5$ Ns/m^2 ([30]);

p_l is the partial pressure of species "l" in N/m^2;

$A_{s,l} = 1.16 \times 10^{-3} \mu_{s,l}^{1/2} \theta_{v,s}^{4/3}$ ([30]); (48)

$$\mu_{s,l} = \frac{M_s M_l}{M_s + M_l}, \qquad (49)$$

being the reduced molecular mass of the collision partners: kg/kg-mol;

T and $\theta_{v,s}$ in Kelvin.

[25] correction time. For temperatures superiors to 8,000 K, the Eq. (43) gives relaxation times less than those observed in experiments. To temperatures above 8,000 K, [25] suggests the following relation to the vibrational relaxation time:

$$\tau_s^P = \frac{1}{\xi_s \sigma_v n_s}, \qquad (50)$$

where:

$$\xi_s = \sqrt{\frac{8 R_s T}{\pi}}, \qquad (51)$$

being the molecular average velocity in m/s;

$$\sigma_v = 10^{-20} \left(\frac{50{,}000}{T}\right)^2, \qquad (52)$$

being the effective collision cross-section to vibrational relaxation in m^2; and

$$n_s = \rho_s / m_s , \qquad (53)$$

being the density of the number of collision particles of species "s". ρ_s in kg/m^3 and m_s in kg/particle, defined by Eq. (33).

Combining the two relations, the following expression to the vibrational relaxation time is obtained:

$$\tau_s = \tau_s^{M-W} + \tau_s^P . \qquad (54)$$

[25] emphasizes that this expression [Eq. (54)] to the vibrational relaxation time is applicable to a range of temperatures much more vast.

III. Structured [17] Algorithm to Thermochemical Non-Equilibrium

Considering the three-dimensional and structured case, the algorithm follows that described in [21], considering, however, the vibrational contribution ([31]) and the version of the two-temperature model to the frozen speed of sound [Eq. (23)]. Hence, the discrete-dynamic-convective flux is defined by:

$$R_{i+1/2,j,k} = |S|_{i+1/2,j,k} \left\{ \frac{1}{2} M_{i+1/2,j,k} \left[\begin{pmatrix} \rho a \\ \rho a u \\ \rho a v \\ \rho a w \\ \rho a H \end{pmatrix}_L + \begin{pmatrix} \rho a \\ \rho a u \\ \rho a v \\ \rho a w \\ \rho a H \end{pmatrix}_R \right] - \frac{1}{2} \phi_{i+1/2,j,k} \left[\begin{pmatrix} \rho a \\ \rho a u \\ \rho a v \\ \rho a w \\ \rho a H \end{pmatrix}_R - \begin{pmatrix} \rho a \\ \rho a u \\ \rho a v \\ \rho a w \\ \rho a H \end{pmatrix}_L \right] \right\} + \begin{pmatrix} 0 \\ S_x p \\ S_y p \\ S_z p \\ 0 \end{pmatrix}_{i+1/2,j,k} , \qquad (55)$$

the discrete-chemical-convective flux is defined by:

$$R_{i+1/2,j,k} = |S|_{i+1/2,j,k} \left\{ \frac{1}{2} M_{i+1/2,j,k} \left[\begin{pmatrix} \rho_1 a \\ \rho_2 a \\ \rho_4 a \\ \rho_5 a \\ \rho_6 a \\ \rho_7 a \end{pmatrix}_L + \begin{pmatrix} \rho_1 a \\ \rho_2 a \\ \rho_4 a \\ \rho_5 a \\ \rho_6 a \\ \rho_7 a \end{pmatrix}_R \right] - \frac{1}{2} \phi_{i+1/2,j,k} \left[\begin{pmatrix} \rho_1 a \\ \rho_2 a \\ \rho_4 a \\ \rho_5 a \\ \rho_6 a \\ \rho_7 a \end{pmatrix}_R - \begin{pmatrix} \rho_1 a \\ \rho_2 a \\ \rho_4 a \\ \rho_5 a \\ \rho_6 a \\ \rho_7 a \end{pmatrix}_L \right] \right\}, \qquad (56)$$

and the discrete-vibrational-convective flux is determined by:

$$R_{i+1/2,j,k} = |S|_{i+1/2,j,k} \left\{ \frac{1}{2} M_{i+1/2,j,k} \left[(\rho e_v a)_L + (\rho e_v a)_R \right] - \frac{1}{2} \phi_{i+1/2,j,k} \left[(\rho e_v a)_R - (\rho e_v a)_L \right] \right\}. \qquad (57)$$

The same definitions presented in [21-22] are valid to this algorithm. The time integration is performed employing the Runge-Kutta explicit method of five stages, second-order accurate, to the three types of convective flux. To the dynamic part, this method can be represented in general form by:

$$\begin{aligned} Q_{i,j,k}^{(0)} &= Q_{i,j,k}^{(n)} \\ Q_{i,j,k}^{(m)} &= Q_{i,j,k}^{(0)} - \alpha_m \Delta t_{i,j,k} \, R\left(Q_{i,j,k}^{(m-1)}\right) \Big/ V_{i,j,k} \ , \\ Q_{i,j,k}^{(n+1)} &= Q_{i,j,k}^{(m)} \end{aligned} \qquad (58)$$

to the chemical part, it can be represented in general form by:

$$\begin{aligned} Q_{i,j,k}^{(0)} &= Q_{i,j,k}^{(n)} \\ Q_{i,j,k}^{(m)} &= Q_{i,j,k}^{(0)} - \alpha_m \Delta t_{i,j,k} \left[R\left(Q_{i,j,k}^{(m-1)}\right) \Big/ V_{i,j,k} - S_C\left(Q_{i,j,k}^{(m-1)}\right) \right], \\ Q_{i,j,k}^{(n+1)} &= Q_{i,j,k}^{(m)} \end{aligned} \qquad (59)$$

where the chemical source term S_C is calculated with the temperature T_{rrc}. Finally, to the vibrational part:

$$\begin{aligned} Q_{i,j,k}^{(0)} &= Q_{i,j,k}^{(n)} \\ Q_{i,j,k}^{(m)} &= Q_{i,j,k}^{(0)} - \alpha_m \Delta t_{i,j,k} \left[R\left(Q_{i,j,k}^{(m-1)}\right) \Big/ V_{i,j,k} - S_v\left(Q_{i,j,k}^{(m-1)}\right) \right], \\ Q_{i,j,k}^{(n+1)} &= Q_{i,j,k}^{(m)} \end{aligned} \qquad (60)$$

in which:

$$S_v = \sum_{s=mol} q_{T-V,s} + \sum_{s=mol} S_{C,s} e_{v,s} \ ; \qquad (61)$$

where: m = 1,...,5; $\alpha_1 = 1/4$, $\alpha_2 = 1/6$, $\alpha_3 = 3/8$, $\alpha_4 = 1/2$ and $\alpha_5 = 1$. This scheme is first-order accurate in space and second-order accurate in time. The second-order of spatial accuracy is obtained by the "MUSCL" procedure (details in [32]).

The [17] scheme in its first-order two-dimensional unstructured version to an ideal gas formulation is presented in [33]. The extension to reactive flow in thermochemical non-equilibrium can be deduced from the present code.

The viscous formulation follows that of [34], which adopts the Green theorem to calculate primitive variable gradients. The viscous vectors are obtained by arithmetical average between cell (i,j,k) and its neighbours. As was done with the convective terms, there is a need to separate the viscous flux in three parts: dynamical viscous flux, chemical viscous flux and vibrational viscous flux. The dynamical part corresponds to the first four equations of the Navier-Stokes ones, the chemical part corresponds to the following six equations and the vibrational part corresponds to the last equation.

The spatially variable time step technique has provided excellent convergence gains as demonstrated in [18-19] and is implemented in the code presented in this work. Details in [18-19; 22].

IV. Results

Tests were performed in one personal computer Notebook with Dual Core Intel Pentium processor of 2.30 GHz of "clock" and 2.0 GBytes of RAM. As the interest of this work is steady state problems, it is necessary to define a criterion which guarantees the convergence of the numerical results. The criterion adopted was to consider a reduction of no minimal four (4) orders of magnitude in the value of the maximum residual in the calculation domain, a typical CFD-community criterion. The residual of each cell was defined as the numerical value obtained from the discretized conservation equations. As there are twelve (12) conservation equations to each cell, the maximum value obtained from these equations is defined as the residual of this cell. Hence, this residual is compared with the residual of the other cells, calculated of the same way, to define the maximum residual in the calculation domain. In the simulations, the attack angle was set equal to zero.

A. Initial and Boundary Conditions to the Studied Problem

The initial conditions are presented in Tab. 6. The Reynolds number is obtained from data of [35]. The boundary conditions to this problem of reactive flow are detailed in [24], as well the geometry in study, the meshes employed in the simulations and the description of the computational configuration.

Table 6 Initial conditions to the problem of the blunt body.

Property	Value
M_∞	8.78
ρ_∞	0.00326 kg/m^3
p_∞	687 Pa
U_∞	4,776 m/s
T_∞	694 K
$T_{v,\infty}$	694 K
altitude	40,000 m
c_N	10^{-9}
c_O	0.07955
c_{O_2}	0.13400
c_{NO}	0.05090
c_{NO+}	0.0
c_{e-}	0.0
L	2.0 m
Re_∞	2.3885x10^6

The geometry is a blunt body with 1.0 m of nose ratio and parallel rectilinear walls. The far field is located at 20.0 times the nose ratio in relation to the configuration nose. The dimensionless employed in the Euler and Navier-Stokes equations in this study are also described in [24].

B. Studied Cases

Table 7 presents the studied cases in this work, the mesh characteristics and the order of accuracy of the [17] scheme.

Table 7 Studied cases, mesh characteristics and accuracy order.

Case	Mesh	Accuracy Order
Inviscid – 3D	63x60x10	First
Viscous – 3D	63x60x10 (7.5%)[a]	First
Inviscid – 3D	63x60x10	Second
Viscous – 3D	63x60x10 (7.5%)	Second

[a] Exponential stretching.

C. *Results in Thermochemical Non-Equilibrium*

Inviscid, structured and first-order accurate case. Figure 1 exhibits the pressure contours around the blunt body geometry calculated at the computational domain by the [17] scheme, in its first-order version, in thermochemical non-equilibrium. The non-dimensional pressure peak is equal to 148.46 unities and is located at the configuration nose. The solution presents good symmetry characteristics. Figure 2 shows the Mach number contours calculated at the computational domain. A region of subsonic flow is formed behind the normal shock wave, at the geometry nose. The shock wave develops normally: normal shock wave at the configuration nose, decaying to oblique shock waves and finally reaching, far from the blunt body, the Mach wave.

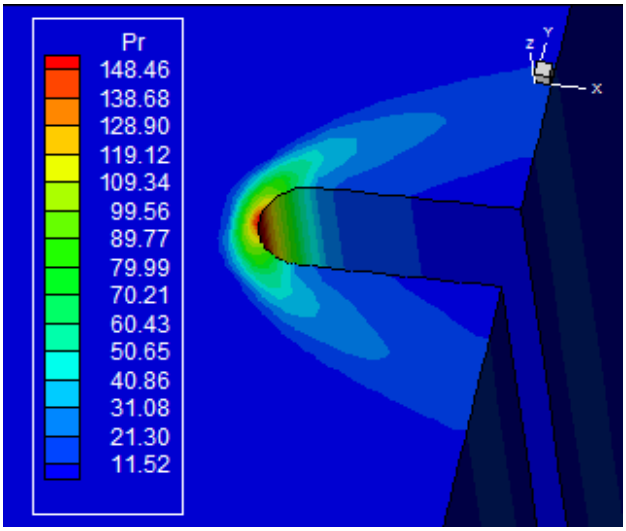


Figure 1. Pressure contours.

Figure 3 presents the contours of the translational/rotational temperature distribution calculated at the computational domain. The translational/rotational temperature reaches a peak of 8,103 K at the configuration nose and determines an appropriated region to dissociation of N_2 and O_2. Along the blunt body, the translational/rotational temperature assumes an approximated value of 6,000 K, what also represents a good value to the dissociation firstly of O_2 and, in second place, of the N_2.

Figure 4 exhibits the contours of the vibrational temperature calculated at the two-dimensional computational domain. Its peak reaches a value of 5,415 K and also contributes to the dissociation of N_2 and O_2, since the employed temperature to the calculation of the forward and backward reaction rates (reaction-rate-control temperature, T_{rrc}) in the thermochemical non-equilibrium is equal to $\sqrt{T.T_V}$, the square root of the product between the translational/rotational temperature and the vibrational temperature.

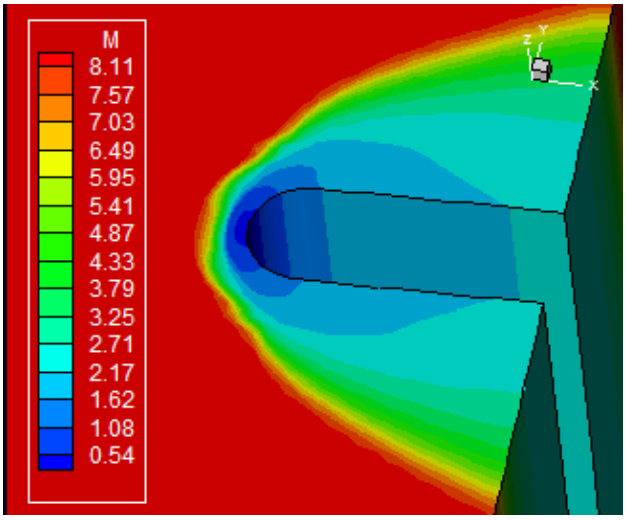


Figure 2. Mach number contours.

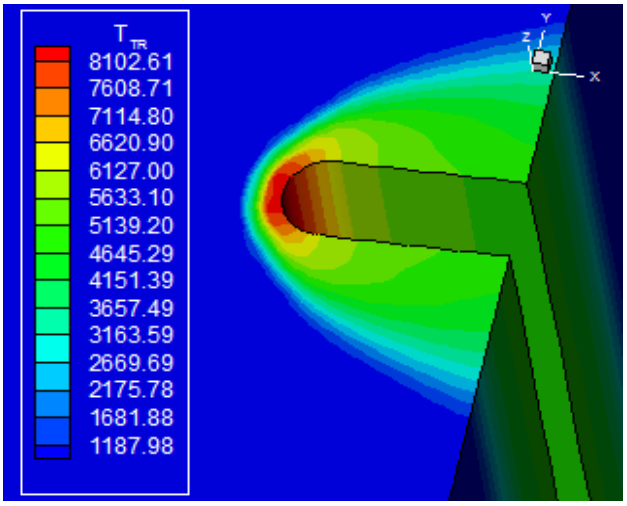


Figure 3. T/R temperature contours.

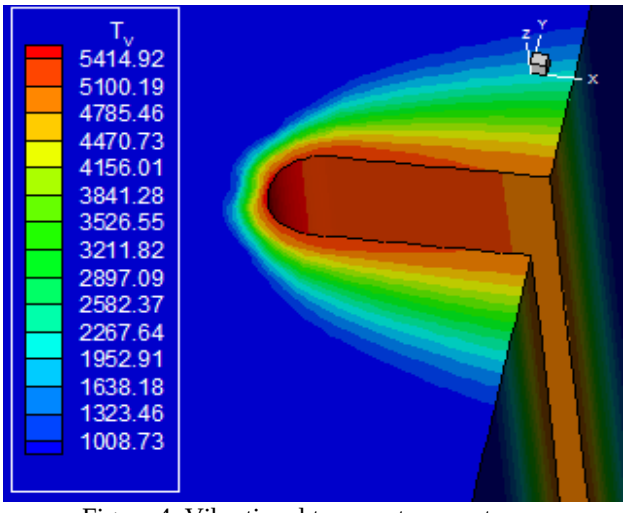


Figure 4. Vibrational temperature contours.

ISBN: 978-1-61804-251-4

Hence, the effective temperature to the calculation of the chemical phenomena guarantees the couple between the vibrational mode and the dissociation reactions. In this configuration nose region, the temperature T_{rrc} reaches, in the steady state condition, the approximated value of 6,624 K, assuring that the dissociation phenomena described above occurs. Good symmetry characteristics are observed.

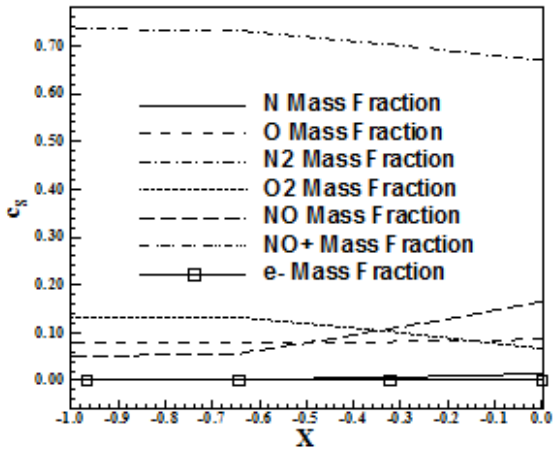


Figure 5. Mass fraction distribution at the blunt body stagnation line.

Figure 5 shows the mass fraction distribution of the seven chemical species under study, namely: N, O, N_2, O_2, NO, NO^+ and e^-, along the geometry stagnation line or geometry symmetry line. As can be observed from this figure, enough dissociation of N_2 and O_2 occur, with the consequent meaningful increase of N and of NO in the gaseous mixture. As mentioned early, this behaviour is expected due to the effective peak temperature reached at the calculation domain. The NO presented the biggest absolute increase in its formation, whereas the N presented the biggest relative increase. The O has not a meaningful increase due to the formation of the NO^+. The formation of e- is also discrete.

Viscous, structured and first-order accurate case. Figure 7 shows the Mach number contours calculated at the computational domain. The subsonic flow region, which is formed behind the normal shock, is well captured and propagates by the lower and upper geometry walls, due to the transport phenomena considered in the viscous simulations. The shock wave presents the expected behaviour: normal shock wave at the configuration nose, oblique shock waves and a Mach wave far from de blunt body.

Figure 8 exhibits the distribution of the +translational/rotational temperature calculated at the computational domain. The peak of translational/rotational temperature reaches the approximated value of 8,797 K at the configuration nose and this value is observed along the lower and upper surfaces of the geometry.

Figure 9 presents the vibrational temperature distribution calculated at the computational domain. Its peak, at the configuration nose, reaches an approximated value of 5,401 K. The effective temperature to the calculation of the dissociation and recombination reactions, T_{rrc}, is equal approximately to 6,893 K, which guarantees that processes of dissociation of O_2 and N_2 can be captured by the employed formulation.

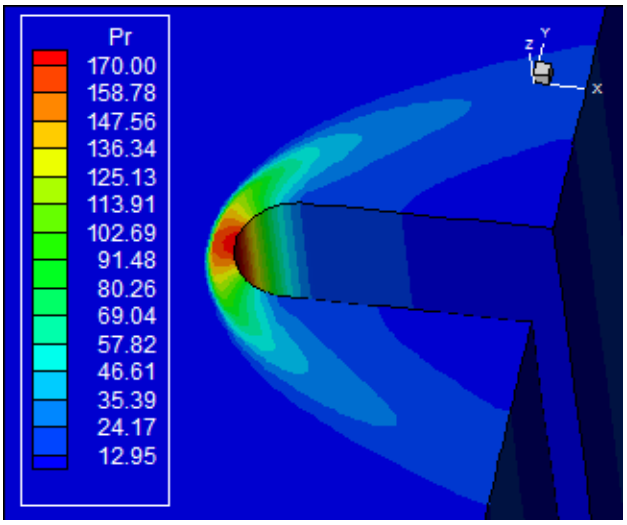


Figure 6. Pressure contours.

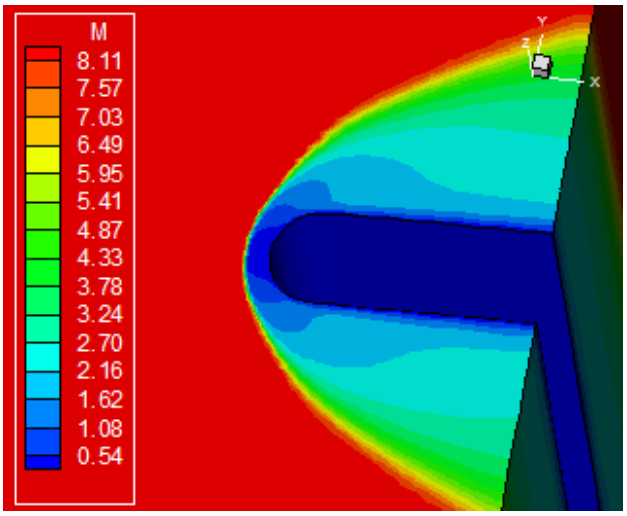


Figure 7. Mach number contours.

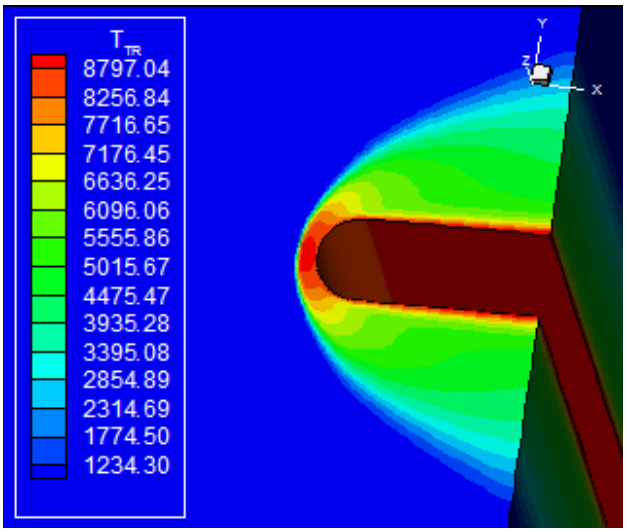


Figure 8. T/R temperature contours.

This value of effective temperature to the viscous reactive simulations is superior to that obtained in the inviscid case. Good symmetry characteristics are observed in these figures.

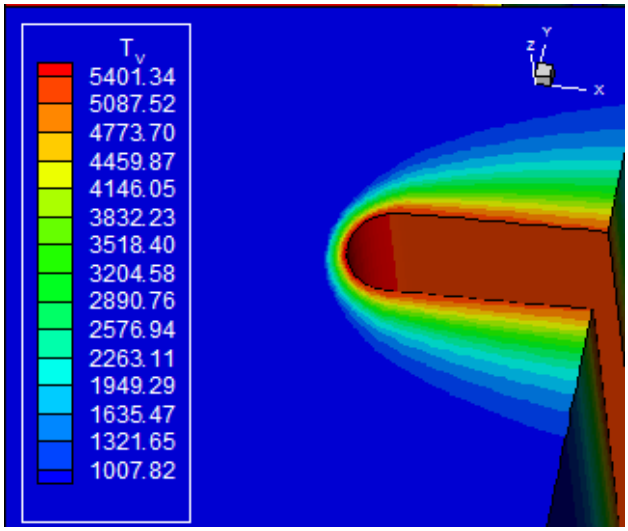


Figure 9. Vibrational temperature contours.

Figure 10 exhibits the mass fraction distribution of the seven chemical species *under* study along the geometry stagnation line. As can be observed, enough dissociation of the N_2 and O_2 occurs, with the consequent meaningful increase of the N and of the NO, with reduction of the mass fraction of the O, in the gaseous mixture. The behaviour of the N and of the NO is expected due to the temperature peak reached in the calculation domain. The O reduction is also expected due to the formation of the NO^+. The biggest absolute increase in the formation of a species was due to the NO, while, in relative terms, was due to the N.

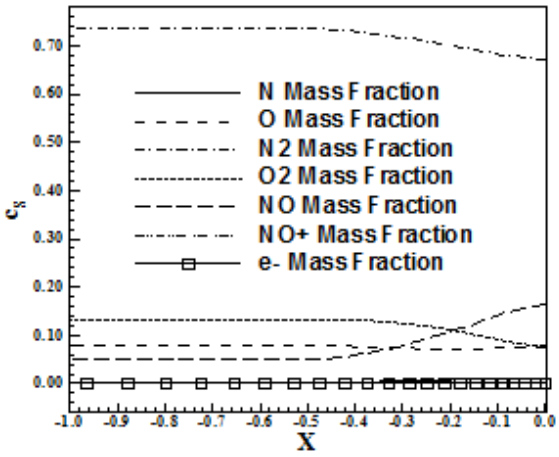


Figure 10. Mass fraction distribution at the blunt body stagnation line.

Inviscid, structured and second-order accurate case. Figure 11 shows the pressure contours obtained by the inviscid simulation performed by the second-order [17] scheme employing a minmod non-linear flux limiter. The non-dimensional pressure peak is approximately equal to 146 unities, slightly inferior to the respective peak obtained by the first-order solution. This pressure peak occurs at the configuration nose. The solution presents good symmetry characteristics. Figure 12 presents the Mach number contours obtained at the computational domain. The subsonic region which is formed behind the normal shock wave is well characterized at the configuration nose. Good symmetry characteristics are observed. The shock wave presents the expected behavior, passing from a normal shock at the configuration stagnation line to a Mach wave far from the blunt body.

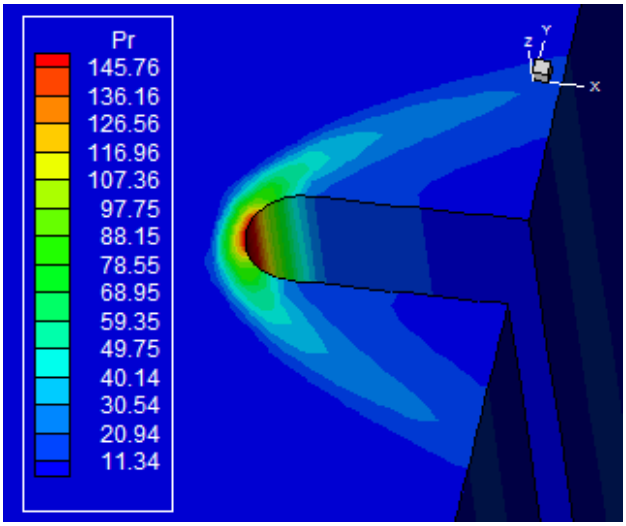


Figure 11. Pressure contours.

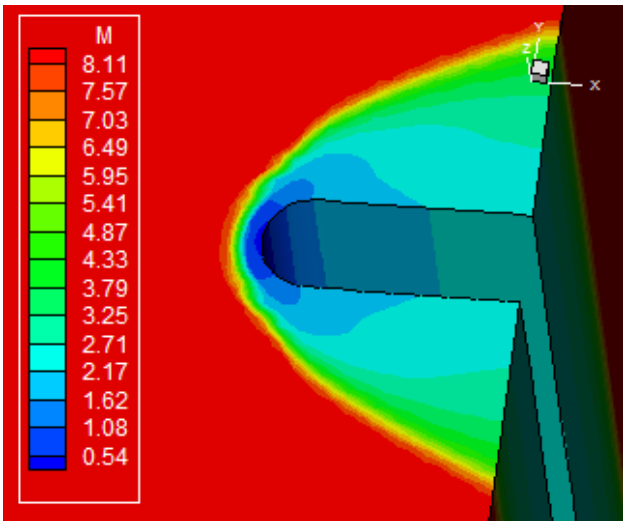


Figure 12. Mach number contours.

Figure 13 exhibits the contours of the translational/rotational temperature distribution calculated at the computational domain. The translational/rotational temperature peak occurs at the configuration nose and is approximately equal to 8,278 K. Figure 14 presents the contours of the vibrational temperature distribution calculated at the computational domain. The vibrational temperature peak is approximately equal to 2,365 K and is observed at the configuration nose. The effective temperature to calculation of the reaction rates (reaction rate control temperature, T_{rrc}) is approximately equal to 4,425 K, which represents a

temperature capable to capture the dissociation phenomena of N_2 and O_2. Good symmetry characteristics are observed in both figures.

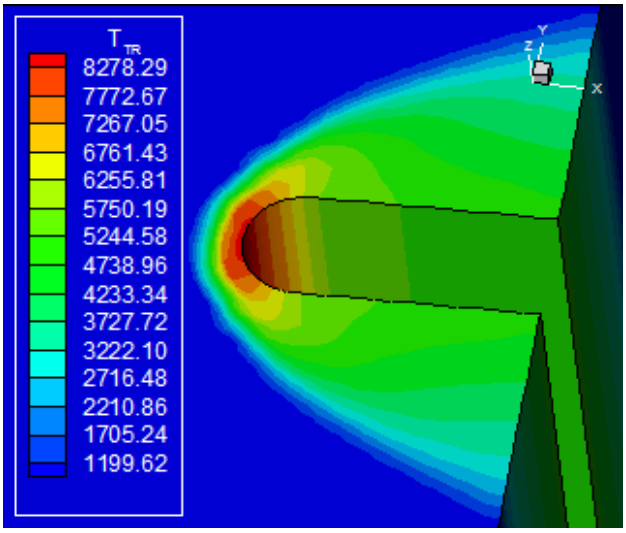


Figure 13. T/R temperature contours.

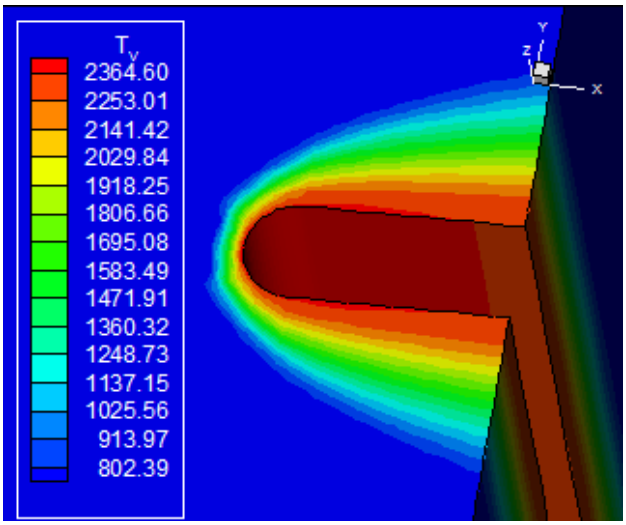


Figure 14. Vibrational temperature contours.

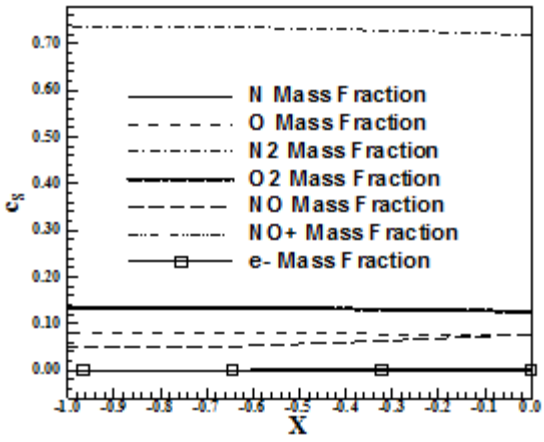


Figure 15. Mass fraction distribution at the blunt body stagnation line.

Figure 15 exhibits the mass fraction distribution of the seven chemical species under study, namely: N, O, N_2, O_2, NO, NO^+ and e^-, along the geometry stagnation line. As can be observed, discrete dissociation of N_2 and O_2 occur, with consequent discrete increase of the N and of the NO, with subsequent reduction of the O, in the gaseous mixture. This behaviour is expected due to the effective temperature peak reached at the computational domain to the calculation of thermochemical non-equilibrium and to a second-order numerical formulation, which behaves in a more conservative way (see [22]), providing minor dissociation of N_2 and O_2.

Viscous, structured and second-order accurate case. Figure 16 exhibits the pressure contours calculated at the computational domain to the studied configuration of blunt body. The non-dimensional pressure peak is approximately equal to 164 unities, less than the respective value obtained by the first-order solution. The shock is positioned closer to the blunt body due to the mesh stretching and the employed-viscous-reactive formulation. Good symmetry characteristics are observed.

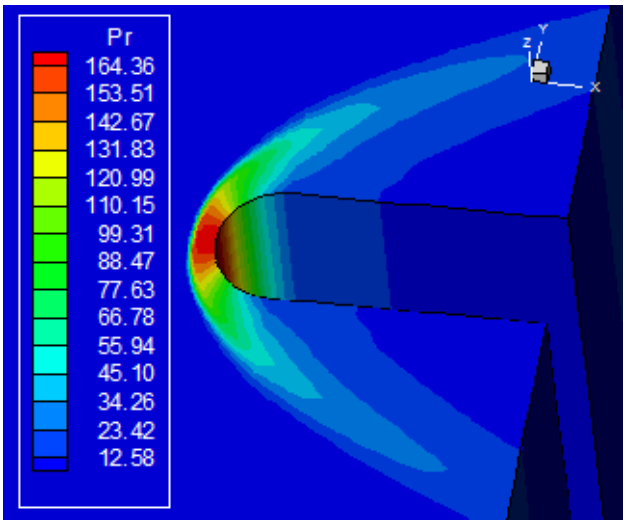


Figure 16. Pressure contours.

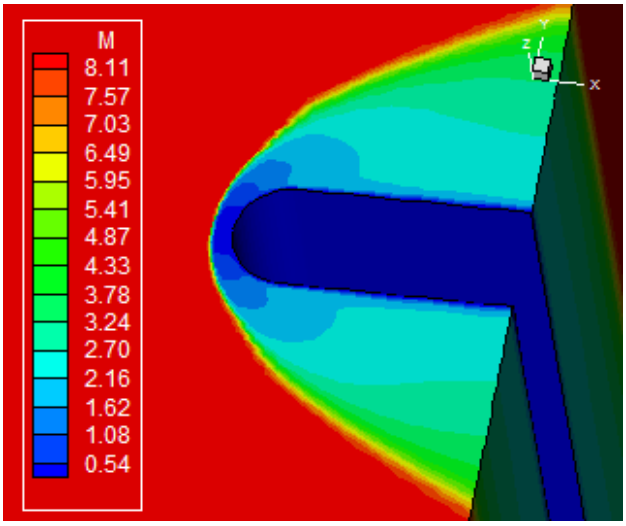


Figure 17. Mach number contour.

ISBN: 978-1-61804-251-4

Figure 17 shows the Mach number contours obtained at the computational domain. The subsonic region behind the normal shock wave, at the stagnation line, is well captured by the solution. This region propagates along the lower and upper surfaces of the geometry, due to the transport phenomena (viscosity, thermal conductivity and species diffusion). The shock wave behaviour is also the expected: normal shock at the geometry nose, oblique shock waves close to the configuration and Mach wave far from the geometry.

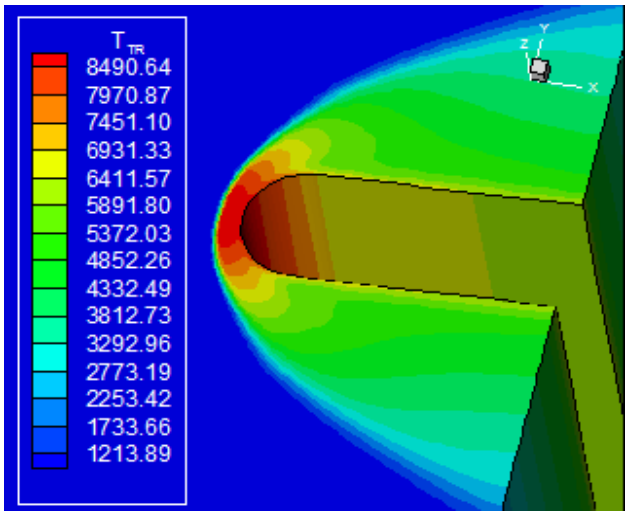


Figure 18. T/R temperature contours.

Figure 18 exhibits the translational/rotational temperature distribution calculated at the computational domain. The temperature peak at the configuration nose reaches approximately 8,491 K. Figure 19 shows the vibrational temperature distribution calculated at the computational domain. The temperature peak at the nose and along the lower and upper surfaces of the geometry is equal to 5,901 K. The effective temperature to the calculation of the reaction rates, T_{rrc}, was of 7,079 K, superior to that obtained with the first-order solution, which is representative to the calculation of the N_2 and O_2 dissociations. Both Figs. 18 and 19 exhibit good symmetry characteristics.

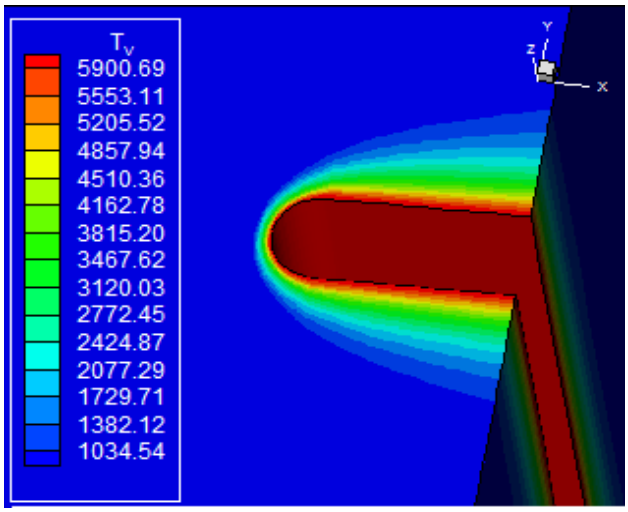


Figure 19. Vibrational temperature contours.

Figure 20 presents the mass fraction distribution of the seven chemical species under study, namely: N, O, N_2, O_2, NO, NO^+ and e^-, along the geometry stagnation line. As can be observed, good dissociation of N_2 and O_2 occur, with consequent good increase of N and NO in the gaseous mixture. This behavior is expected due to the effective temperature peak reached at the computational domain to the calculation of thermochemical non-equilibrium and to a second-order numerical formulation, which behaves in a more conservative way ([22]), providing major dissociation of N_2 and O_2. In other words, this solution provided by the second-order [17] scheme, as seen in other cases, tends to provide bigger dissociation of N_2 and O_2. As this solution is more precise (second-order), it should be considered as standard to comparison with other schemes. The NO^+ is formed with the subsequent reduction of the O species.

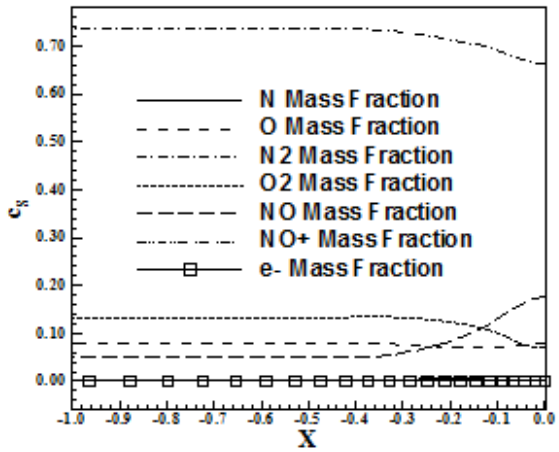


Figure 20. Mass fraction distribution at the blunt body stagnation line.

Shock Position. In this section is presented the behaviour of the shock position in thermochemical non-equilibrium conditions for the five and seven species models. Both first- and second-order solutions are compared between them.

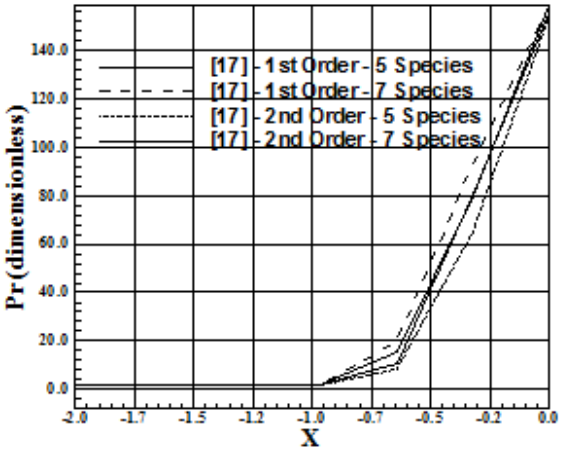


Figure 31. Shock position (inviscid).

The detached shock position in terms of pressure distribution, in the inviscid case, and first- and second-order

accurate solutions, is exhibited in Fig. 31. It is shown the thermochemical non-equilibrium shock position for the five and seven species models. As can be observed, the second-order results yield closer shock positions in relation to the blunt body nose. Particularly, the second-order, five species model, is the closest solution to the inviscid case.

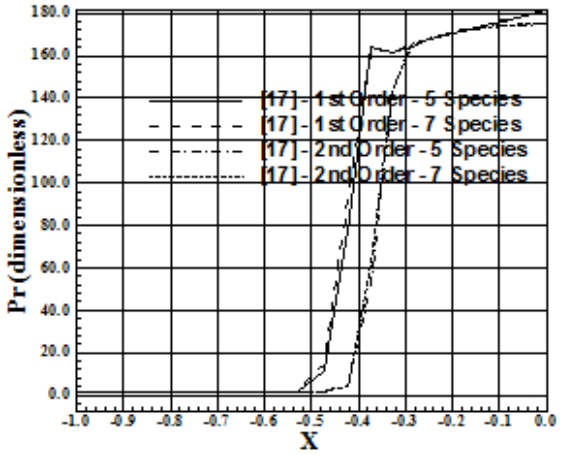


Figure 32. Shock position (viscous).

The detached shock position in terms of pressure distribution, in the viscous case, first- and second-order accurate solutions, is exhibited in Fig. 32. It is shown the thermochemical non-equilibrium shock position to the five and seven species models. As can be observed, the second-order positions are located at 0.48 m, whereas the first-order solutions are located at 0.54m, ratifying the best behaviour of the second-order results.

Quantitative Analysis. In terms of quantitative results, the present authors compared the reactive results with the perfect gas solutions. The stagnation pressure at the blunt body nose and the shock standoff distance were evaluated assuming the perfect gas formulation. Such parameters calculated at this way are not the best comparisons, but in the absence of practical reactive results, these constitute the best available results.

To calculate the stagnation pressure ahead of the blunt body, [36] presents in its B Appendix values of the normal shock wave properties ahead of the configuration. The ratio pr0/pr∞ is estimated as function of the normal Mach number and the stagnation pressure pr0 can be determined from this parameter. Hence, to a freestream Mach number of 9.0 (close to 8.78), the ratio pr_0/pr_∞ assumes the value 104.8. The value of pr_∞ is determined by the following expression:

$$pr_\infty = \frac{pr_{initial}}{\rho_{initial} \times a^2_{initial}} \quad (62)$$

In the present study, prinitial = 687 N/m^2, ρinitial = 0.004kg/m^3 and ainitial = 317.024m/s. Considering these values, one concludes that pr_∞ = 1.709 (non-dimensional). Using the ratio obtained from [36], the stagnation pressure ahead of the configuration nose is estimated as 179.10 unities. Table 9 compares the values obtained from the simulations with this theoretical parameter and presents the numerical percentage errors. As can be observed, all solutions present percentage errors less than 20%, which is a reasonable estimation of the stagnation pressure.

Table 9 Comparisons between theoretical and numerical results.

Case	pr_0	Error (%)
Inviscid/Structured/1st Order	148.46	17.11
Viscous/Structured/1st Order	170.00	5.08
Inviscid/Structured/2nd Order	145.76	18.62
Viscous/Structured/2nd Order	164.36	8.23

Another possibility to quantify the results is the determination of the shock standoff distance. [37] presents a graphic in which is plotted the shock standoff distance of a pre-determined configuration versus the Mach number. Considering the blunt body nose approximately as a cylinder and using the value 8.78 to the Mach number, it is possible to obtain the value 0.19 to the ratio δ/d, where δ is the position of the normal shock wave in relation to the body nose and d is a characteristic length of the configuration. In the present study, d = 2.0m (diameter of the body nose) and δ = 0.38m. Table 10 presents the values obtained by δ for the different cases and the percentage errors. This table shows that the best result is obtained with the structured, viscous, second order version of [17]. As the shock standoff distance presented in [37] is more realistic, presenting smaller dependence of the perfect gas hypothesis, improved results were expected to obtain in this study. Hence, the best solution is obtained by the [17] scheme in its second order version.

Table 10 Shock standoff distance obtained from numerical schemes.

Case	δ_{NUM} (m)	Error (%)
Inviscid/Structured/1st Order	0.80	110.53
Viscous/Structured/1st Order	0.48	26.32
Inviscid/Structured/2nd Order	0.60	57.89
Viscous/Structured/2nd Order	0.40	5.26

Computational performance of the studied algorithms. Table 11 presents the computational data of the reactive simulations performed with the [17] scheme to the problem of the blunt body in three-dimensions. The reactive simulations involved the thermochemical non-equilibrium solutions obtained from five [47] and seven chemical species.

In this table are exhibited the studied case, the maximum number of CFL employed in the simulation, the number of iterations to convergence and the number of orders of reduction in the magnitude of the maximum residual in relation to its initial value.

Table 11 Computational data of the reactive simulations with the 2D blunt body.

Studied case	CFL	Iterations	Orders of Residual Reduction
First-Order / Structured / Inviscid / FS[(a)]	0.9	373	4
First-Order / Structured / Viscous / FS	0.7	1,005	4
Second-Order / Structured / Inviscid / FS	0.3	982	4
Second-Order / Structured / Viscous / FS	0.3	2,412	4
First-Order / Structured / Inviscid / SS[(b)]	0.9	372	4
First-Order / Structured / Viscous / SS	0.7	997	4
Second-Order / Structured / Inviscid / SS	0.1	2,908	4
Second-Order / Structured / Viscous / SS	0.7	1,173	4

[(a)]: Five Species; [(b)]: Seven Species.

As can be observed, all test-cases converged with no minimal four orders of reduction in the value of the maximum residual. The maximum numbers of CFL presented the following distribution: 0.9 in two (2) cases (25.00%), 0.7 in three (3) cases (37.50%), 0.3 in two (2) cases (25.00%) and 0.1 in one (1) case (12.50%). The convergence iterations did not overtake 3,000, in all studied cases. However, the time wasted in the simulations was much raised, taking until days to convergence (to four orders of reduction in the maximum residual). It is important to emphasize that all two-dimensional viscous simulations were considered laminar, without the introduction of a turbulence model, although high Reynolds number were employed in the simulations.

V. Conclusion

This work, the first part of this study, presents a numerical tool implemented to simulate inviscid and viscous flows employing the reactive gas formulation of thermochemical non-equilibrium flow in three-dimensions. The Euler and Navier-Stokes equations, employing a finite volume formulation, on the context of structured and unstructured spatial discretizations, are solved. These variants allow an effective comparison between the two types of spatial discretization aiming verify their potentialities: solution quality, convergence speed, computational cost, etc. The aerospace problem involving the "hot gas" hypersonic flow around a blunt body, in three-dimensions, is simulated.

To the simulations with unstructured spatial discretization, a structured mesh generator developed by the first author ([38]), which creates meshes of hexahedrons (3D), was employed. After that, as a pre-processing stage ([39]), such meshes were transformed in meshes of tetrahedrons. Such procedure aimed to avoid the time which would be waste with the implementation of an unstructured generator, which was not the objective of the present work, and to obtain a generalized algorithm to the solution of the reactive equations.

In this work, first part of this study, the structured formulation of the three-dimensional Euler and Navier-Stokes reactive equations is presented. In [40], the second part of this study, it will be presented the unstructured version of the calculation algorithm in three-dimensions to complete the formulation in structured and in unstructured contexts.

The reactive simulations involved an air chemical model of seven species: N, O, N_2, O_2, NO, NO^+ and e^-. Eighteen chemical reactions, involving dissociation, recombination and ionization, were simulated by the proposed model. This model was suggested by [46]. The Arrhenius formula was employed to determine the reaction rates and the law of mass action was used to determine the source terms of each gas specie equation.

The results have demonstrated that the most correct aerodynamic coefficient of lift is obtained by the [17] scheme with first-order accuracy, in an inviscid formulation, to a five species model. The cheapest algorithm was due to [17], inviscid, first-order accurate, structured, and five species model. Moreover, the shock position is closer to the geometry as using the reactive formulation than the ideal gas formulation. It was verified in [22]. Comparing the five species model and the seven species model, the second order solution of both models present the best behaviour. Errors less than 20% were obtained with this version of the [17] algorithm in the determination of the stagnation pressure at the body nose and an error of 7.89% was found in the determination of the shock standoff distance, highlighting the correct implementation and good results obtained from the reactive formulation. Values of these parameters were evaluated and proved the significant potential of the present numerical tool.

This work, as also [40], is the continuation of the study started at [41], based on the work of [42]. Other references on the non-equilibrium reactive flows area are: [43], [44] and [45].

References

[1] P. A. Gnoffo, R. N. Gupta, and J. L. Shinn, Conservation Equations and Physical Models for Hypersonic Flows in Thermal and Chemical Nonequilibrium, *NASA TP 2867*, 1989.

[2] M. Liu and M. Vinokur, Upwind Algorithms for General Thermo-Chemical Nonequilibrium Flows, *AIAA Paper 89-0201*, 1989.

[3] R. N. Gupta, J. M. Yos, R. A. Thompson, and K. –P. Lee, A Review of Reaction Rates and Thermodynamic and Transport Properties for an 11-Species Air Model for Chemical and Thermal Nonequilibrium Calculations to 30000 K, *NASA RP-1232*, 1990.

[4] R. K. Prabhu, An Implementation of a Chemical and Thermal Nonequilibrium Flow Solver on Unstructured Meshes and Application to Blunt Bodies, *NASA CR-194967*, 1994.

[5] C. Park, Radiation Enhancement by Nonequilibrium in Earth's Atmosphere, *Journal of Spacecraft and Rockets*, Vol. 22, No. 1, 1985, pp. 27-36.

[6] C. Park, Problem of Rate Chemistry in the Flight Regimes of Aeroassissted Orbital Transfer Vehicles, *Thermal Design of Aeroassissted Orbital Transfer Vehicles*, Progress in Astronautics and Aeronautics, edited by H. F. Nelson, AIAA, NY, Vol. 96, 1985, pp. 511-537.

[7] P. A. Gnoffo, Three-Dimensional AOTV Flowfields in Chemical Nonequilibrium, *AIAA Paper 86-0230*, 1986.

[8] C. P. Li, Implicit Methods for Computing Chemically Reacting Flow, *NASA TM-58274*, 1986.

[9] J. H. Lee, Basic Governing Equations for the Flight Regimes of Aeroassisted Orbital Transfer Vehicles, *Thermal Design of Aeroassisted Transfer Vehicles*, Progress in Astronautics and Aeronautics, AIAA, Vol. 96, 1985, pp. 3-53.

[10] C. Park, Convergence of Computation of Chemically Reacting Flows, *Thermophysical Aspects of Re-entry Flows*, Progress in Astronautics and Aeronautics, edited by J. N. Moss and C. D. Scott, AIAA, NY, Vol. 103, pp. 478-513.

[11] C. Park, Assessment of Two-Temperature Kinetic Model for Dissociating and Weakly-Ionizing Nitrogen, *AIAA Paper 86-1347*, 1986.

[12] C. Park, Calculation of Nonequilibrium Radiation in the Flight Regimes of Aeroassissted Orbital Transfer Vehicles, *Thermal Design of Aeroassissted Orbital Transfer Vehicles*, Progress in Astronautics and Aeronautics, edited by H. F. Nelson, AIAA, NY, Vol. 96, 1985, pp. 395-418.

[13] C. Park, Nonequilibrium Air Radiation (NEQAIR) Program: User's Manual, *NASA TM-86707*, 1985.

[14] R. A. Allen, J. C. Camm, and J. C. Keck, Radiation from Hot Nitrogen, Research Report 102, AVCO-Everett Research Laboratory, Everett, MA, 1961.

[15] R. A. Allen, J. C. Keck, and J. C. Camm, Nonequilibrium Radiation from Shock Heated Nitrogen and a Determination of the Recombination Rate, Research Report 110, AVCO-Everett Research Laboratory, Everett, MA, 1961.

[16] R. A. Allen, Nonequilibrium Shock Front Rotational, Vibrational, and Electronic Temperature Measurements, Research Report 186, AVCO-Everett Research Laboratory, Everett, MA, 1964.

[17] B. Van Leer, Flux-Vector Splitting for the Euler Equations, *Lecture Notes in Physics*, Springer Verlag, Berlin, Vol. 170, pp. 507-512, 1982.

[18] E. S. G. Maciel, Analysis of Convergence Acceleration Techniques Used in Unstructured Algorithms in the Solution of Aeronautical Problems – Part I, *Proceedings of the XVIII International Congress of Mechanical Engineering (XVIII COBEM)*, Ouro Preto, MG, Brazil, 2005.

[19] E. S. G. Maciel, Analysis of Convergence Acceleration Techniques Used in Unstructured Algorithms in the Solution of Aerospace Problems – Part II, *Proceedings of the XII Brazilian Congress of Thermal Engineering and Sciences (XII ENCIT)*. Belo Horizonte, MG, Brazil, 2008.

[20] S. K. Saxena and M. T. Nair, An Improved Roe Scheme for Real Gas Flow, *AIAA Paper 2005-587*, 2005.

[21] E. S. G. Maciel, Relatório ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) sobre as atividades de pesquisa realizadas no período de 01/07/2008 até 30/06/2009 com relação ao projeto PDJ número 150143/2008-7, *Report to the National Council of Scientific and Technological Development (CNPq)*, São José dos Campos, SP, Brasil, 102p, 2009. [available in the website www.edissonsavio.eng.br]

[22] E. S. G. Maciel, and A. P. Pimenta, Thermochemical Non-Equilibrium Reentry Flows in Two-Dimensions – Part I, *WSEAS TRANSACTIONS ON MATHEMATICS*, Vol. 11, June, Issue 6, pp. 520-545.

[23] W. G. Vincent and C. H. Kruger Jr., *Introduction to Physical Gas Dynamics*, John Wiley & Sons, Ltd, New York, 1965.

[24] E. S. G., Maciel, Relatório ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) sobre as atividades de pesquisa realizadas no período de 01/07/2009 até 31/12/2009 com relação ao projeto PDJ número 150143/2008-7, *Report to the National Council of Scientific and Technological Development (CNPq)*, São José dos Campos, SP, Brasil, 102p, 2009. [available in the website www.edissonsavio.eng.br]

[25] C. Park, Assessment of Two-Temperature Kinetic Model for Ionizing Air, *Journal of Thermophysics and Heat Transfer*, Vol. 3, No. 13, pp. 233-244, 1989.

[26] G. Degrez, and E. Van Der Weide, Upwind Residual Distribution Schemes for Chemical Non-Equilibrium Flows, *AIAA Paper 99-3366*, 1999.

[27] L. Landau, and E. Teller, Theory of Sound Dispersion, *Physikalische Zeitschrift Der Sowjetunion*, Vol. 10, 1936, pp. 34-43.

[28] R. Monti, D. Paterna, R. Savino, and A. Esposito, Experimental and Numerical Investigation on Martian Atmosphere Entry, *AIAA Paper 2001-0751*, 2001.

[29] R. C. Millikan and D. R. White, Systematics of Vibrational Relaxation, *The Journal of Chemical Physics*, Vol. 39, No. 12, 1963, pp. 3209-3213.

[30] A. F. P. Houwing, S. Nonaka, H. Mizuno, and K. Takayama, Effects of Vibrational Relaxation on Bow Shock Stand-off Distance for Nonequilibrium Flows, *AIAA Journal*, Vol. 38, No. 9, 2000, pp. 1760-1763.

[31] D. Ait-Ali-Yahia, and W. G. Habashi, Finite Element Adaptive Method for Hypersonic Thermochemical Nonequilibrium Flows, *AIAA Journal* Vol. 35, No. 8, 1997, 1294-1302.

[32] C. Hirsch, *Numerical Computation of Internal and External Flows – Computational Methods for Inviscid and Viscous Flows*, John Wiley & Sons Ltd, 691p, 1990.

[33] E. S. G. Maciel, Comparison Between the First Order Upwind Unstructured Algorithms of Steger and Warming and of Van Leer in the Solution of the Euler Equations in Two-Dimensions, *Proceedings of the XIX International Congress of Mechanical Engineering (XIX COBEM)*, Brasília, DF, Brazil, 2007.

[34] L. N. Long, M. M. S. Khan, and H. T. Sharp, Massively Parallel Three-Dimensional Euler / Navier-Stokes Method, *AIAA Journal*, Vol. 29, No. 5, 1991, pp. 657-666.

[35] R. W. Fox, and A. T. McDonald, *Introdução à Mecânica dos Fluidos*, Guanabara Editor, 1988.

[36] J. D. Anderson Jr., Fundamentals of Aerodynamics, McGraw-Hill, Inc., 2nd Edition, 772p, 1991.

[37] H. W. Liepmann, and A. Roshko, Elements of Gasdynamics, John Wiley & Sons, Inc., 1st Edition, 439p, 1957.

[38] E. S. G. Maciel, Relatório ao Conselho Nacional de Pesquisa e Desenvolvimento Tecnológico (CNPq) sobre as Atividades de Pesquisa Desenvolvidas no Primeiro Ano de Vigência da Bolsa de Estudos para Nível DCR-IF Referente ao Processo No. 304318/2003-5, *Report to the National Council of Scientific and Technological Development (CNPq)*, Recife, PE, Brazil, 37p, 2004. [available in the website www.edissonsavio.eng.br]

[39] E. S. G. Maciel, Relatório ao Conselho Nacional de Pesquisa e Desenvolvimento Tecnológico (CNPq) sobre as Atividades de Pesquisa Desenvolvidas no Segundo Ano de Vigência da Bolsa de Estudos para Nível DCR-IF Referente ao Processo No. 304318/2003-5, *Report to the National Council of Scientific and Technological Development (CNPq)*, Recife, PE, Brazil, 54p, 2005. [available in the website www.edissonsavio.eng.br]

[40] E. S. G. Maciel, and A. P. Pimenta, Thermochemical Non-Equilibrium Reentry Flows in Three-Dimensions: Five and Seven Species Model – Part II – Unstructured Solutions, to be submitted to *WSEAS TRANSACTIONS ON MATHEMATICS*, 2012.

[41] E. S. G. Maciel, and A. P. Pimenta, Reentry Flows in Chemical Non-Equilibrium in Two-Dimensions, *Proceedings of the 10th International Symposium on Combustion and Energy Utilisation (ICCEU 2010)*, Mugla, Turkey, 2010.

[42] S. K. Saxena, and M. T. Nair, An Improved Roe Scheme for Real Gas Flows, *AIAA Paper 2005-587*, 2005.

[43] Y. Liu, M. Vinokur, M. Panesi, and T. Magin, A Multi-Group Maximum Entropy Model for Thermo-Chemical Nonequilibrium, *AIAA Paper 2010-4332*.

[44] M. L. da Silva, V. Guerra, and J. Loureiro, State-Resolved Dissociation Rates for Extremely Nonequilibrium Atmospheric Entries, *J. Thermo. Phys.*, Vol. 21, No.1, 2007, pp. 40-49.

[45] S. C. Spiegel, D. L. Stefanski, H. Luo, and J. R. Edwards, A Cell-Centered Finite Volume Method for Chemically Reacting Flows on Hybrid Grids, *AIAA Paper 2010-1083*, 2010.

[46] F. G. Blottner, Viscous Shock Layer at the Stagnation Point With Nonequilibrium Air Chemistry, *AIAA Journal*, Vol. 7, No. 12, 1969, pp. 2281-2288.

[47] E. S. G. Maciel and A. P. Pimenta, Thermochemical Non-Equilibrium Reentry Flows in Three-Dimensions – Part I – Structured Solutions, submitted to *WSEAS TRANSACTIONS ON APPLIED AND THEORETICAL MECHANICS* (under review).

Edisson S. G. Maciel (F'14), born in 1969, february, 25, in Recife, Pernambuco. He is a Mechanical Engineering undergraduated by UFPE in 1992, in Recife, PE, Brazil; Mester degree in Thermal Engineering by UFPE in 1995, in Recife, PE, Brazil; Doctor degree in Aeronautical Engineering by ITA in 2002, in São José dos Campos, SP, Brazil; and Post-Doctor degree in Aeronautical Engineering by ITA in 2009, in São José dos Campos, SP, Brazil.

Actually, he is working in a new post-doctorate project in Aerospace Engineering at ITA. The last researches are based on thermochemical non-equilibrium reentry simulations in Earth and thermochemical non-equilibrium entry simulations in Mars. They are: Maciel, E. S. G., and Pimenta, A. P., "Thermochemical Non-Equilibrium Reentry Flows in Two-Dimensions – Part I', WSEAS Transactions on Mathematics, Vol. 11, Issue 6, June, pp. 520-545, 2012; Maciel, E. S. G., and Pimenta, A. P., "Thermochemical Non-Equilibrium Entry Flows in Mars in Two-Dimensions – Part I', WSEAS Transactions on Applied and Theoretical Mechanics, Vol. 8, Issue 1, January, pp. 26-54, 2013; and he has three published books, the first one being: Maciel, E. S. G., "Aplicações de Algoritmos Preditor-Corretor e TVD na Solução das Equações de Euler e de Navier-Stokes em Duas Dimensões", Recife, PE, Editor UFPE, 2013. He is interested in the Magnetogasdynamic field with applications to fluid dynamics and in the use of ENO algorithms.

Amilcar P. Pimenta (F'14), born in 1954, january, 12, in Minas Gerais, MG. He is a Mechanical Engineering undergraduated by UFMG in 1977, in Minas Gerais, MG, Brazil; Mester degree in Aerodynamics, Propulsion and Energy by ITA in 1986, in São José dos Campos, SP, Brazil; Doctor degree in Energetique by Universite de Poitiers in 1993, in Paris, France.

Actually, he is professor of the ITA and is expert in Aerospatial Engineering, with focus in combustion and chemical reaction flows.

ISBN: 978-1-61804-251-4

Higher Symmetries and Inverse Problems for Ordinary Differential Equations

Valentin Zaitsev
Herzen State
Pedagogical University
of Russia
Saint-Petersburg, Russia
Email: valentin_zaitsev@mail.ru

Lidiya Linchuk
Saint-Petersburg
State Polytechnic University
Saint-Petersburg, Russia

Alexander Flegontov
Herzen State
Pedagogical University
of Russia
Saint-Petersburg, Russia
Email: aflegontoff@herzen.spb.ru

***Abstract*—Regular search algorithms of higher symmetries (tangential, Lie – Bäcklund [Bäcklund] and nonlocal – exponential and non-exponential) for ordinary differential equations of 2–nd and 3–rd order are considered. The inverse problem for search of all equations of the form $y''' = F(x, y, y')$, admitting certain non–exponential nonlocal operator is solved.**

***Index Terms*—ordinary differential equations, tangential operators, Lie–Bäcklund and nonlocal operators, inverse problem.**

I. Introduction

It is well known that the group analysis originated in the end of XIX century (Sophus Lie) and initially it considered only point transformation leaving ordinary differential equation (ODE) invariant. It is possible to equally describe almost all classic methods of integration known at that period on this basis. However, there didn't exist any new methods of solving practically important model equations and further development of group analysis became possible only after Ovsiannikov L. V. (1950 – 1960), who cogently demonstrated that usage of the group methods could give the huge number of physically significant partial solutions of nonlinear model equations in mathematical physics.

Since the search of particular solutions of partial equations (for example, self–model solutions) in partial derivatives is often reduced to the solution of nonlinear ODE, there appeared an urgent need to elaborate the new methods for ODE solving in closed form. Leaving meanwhile the occurrence of discrete–group analysis, it's possible to note that classic group analysis was developed in two directions: nonpoint transformations (local and nonlocal) research and inverse problem solution. Wherein the first direction allows significantly expand the number of "solvable" ODE and the second – to describe the multiplicity of **all** ODE of the given class coinciding with several a priori conditions , for example, those having a first integral (conservation law) of given structure or having symmetry corresponding to the symmetry of certain application. This in its turn gives the possibility to purposefully build the model equations regarding the requirement of maximum adequacy to describing phenomenon.

II. Local transformations

Point transformations do not describe all possible symmetries of ODE, even local. Natural generalization of point transformations in this case are **tangential** or **contact** transformations. The well-known example of such transformations – is the Legendre transformation. Let's consider the G group of point transformations in the space of independent variables (x, y, y')

$$\tilde{x} = \varphi(x, y, y', a), \quad \tilde{y} = \psi(x, y, y', a), \quad \tilde{y}' = \chi(x, y, y', a), \tag{1}$$

where

$$\varphi\Big|_{a=0} = x, \qquad \psi\Big|_{a=0} = y, \qquad \chi\Big|_{a=0} = y'. \tag{2}$$

Transformations (1) are called **contact**, if the G group preserves the following equation $\omega = dy - p\,dx$, i.e. the equation $dy - p\,dx = \rho(d\tilde{y} - \tilde{p}\,d\tilde{x})$ completes. This equation expresses the tangency condition of the first order.

Similarly it's possible to try to enter the tangential transformation of the highest order, but, apparently S. Lie already knew that there could **not exist any tangential transformation of the highest order** because transformations corresponding to them in form and properties turn out in prolongation of point transformations or contact transformations. Moreover, for functions which depend on more than one variable such contact transformations are indeed prolongation of point transformation.

***Theorem* 1** [1]. Operator

$$X = \xi(x, y, y')\frac{\partial}{\partial x} + \eta(x, y, y')\frac{\partial}{\partial y} + \zeta(x, y, y')\frac{\partial}{\partial y'} \tag{3}$$

is an infinitesimal operator of the group of contact transformations if and only if

$$\xi = -\frac{\partial W}{\partial y'}, \qquad \eta = W - y'\frac{\partial W}{\partial y'}, \qquad \zeta = \frac{\partial W}{\partial x} + y'\frac{\partial W}{\partial y} \tag{4}$$

with several function $W = W(x, y, y')$.

From the form (3) it is obvious that for equations of the second order the regular algorithm of search for contact transformations there do not exist, as the defining equation

ISBN: 978-1-61804-251-4

(invariance condition) does not split on the system due to the lack of an independent variable – all the required functions depend on all variables included. But even in the case when it is possible to find an assumed operator which coincides in its shape with the from (3) it is not guaranteed that it could be contact. For example, in article [2] there is an example of equation

$$y'' + \frac{1}{3}xy^{-5/3} = 0, \tag{5}$$

which permits the existence of the second operator (along with point $X_1 = 8x\,\partial_x + 9y\,\partial_y$)

$$X_2 = \left[(y')^2 - ty^{-2/3}\right]\partial_x - \frac{3}{2}y^{1/3}\,\partial_y + \frac{1}{2}y^{-2/3}y'\,\partial_{y'}. \tag{6}$$

(this operator was found using a discrete group of transformations for the equation of Emden–Fowler [3]). However, it's impossible to find the function W – the system of equations (4) for coordinates (6) is incompatible. Therefore the operator (6) is tangent **only on the manifold of solutions** for equation (5), i.e. essentially it's Lie–Bäcklund's operator.

For equations of the third and higher order contact transformations could be received using the algorithm of Lie. However, despite of the significant increase of possible dimension of permissible operators algebra (from 7 to 10 for equations of the third order), the class of equations which admit local transformations is indeed rather narrow. Let's illustrate this statement considering an inverse problem for class of contact transformations allowed by such equation:

$$y''' = F(x, y, y'). \tag{7}$$

This problem means the solving of system

$$\begin{cases} W_{y'y'y'} = 0, \\ W_{yy'y'}y' + W_{xy'y'} + W_{yy'} = 0, \\ W_{yyy'}(y')^2 + 2W_{xyy'}y' + W_{xxy'} + W_{yy}y' + \\ +W_{xy} + W_{y'y'}F = 0, \\ W_{y'}F_x + (y'W_{y'} - W)F_y - (y'W_y - W_x)F_{y'} + \\ +(3y'W_{yy'} + W_y + 3W_{xy'})F + \\ +W_{xxx} + 3W_{xxy}y' + 3W_{xyy}(y')^2 + W_{yyy}(y')^3 = 0. \end{cases} \tag{8}$$

The solution of first three equations (8) gives two solutions: 1) if $W_{y'y'} = 0$, so F is arbitrary and operator converts to point operator. This problem is already solved (see, f. ex. [4]) and therefore is not interesting for us; 2) $W_{y'y'} \neq 0$, and then

$$W = \frac{1}{2}f(x)(y')^2 - f'(x)yy' + g(x)y' + H(x, y),$$

$$F = \frac{1}{f(x)}\Big[\big(2f''(x) - H_{yy}\big)y' + f'''(x)y - g''(x) - H_{xy}\Big]. \tag{9}$$

The substitution of expressions (9) in the last equation of system (8) gives (after splitting on exponents y') a new system, from the first two equations of which follows

$$H(x, y) = \frac{[f'(x)]^2 + C}{2f(x)}y^2 + a_1(x)y + a_0(x),$$

i.e. the equation (7) is linear.

III. Nonlocal operators

Let's find out which structure should be for infinitesimal operator to describe all ODE of the second order, allowing reduction of the number of order by moving to new variables – invariants for admitted operator. Let the operator be written in canonical form, then a universal invariant $I_0 = x$, and let the first differential equations be $I_1 = H(x, y, y')$. Then the function H satisfies the equation

$$\Phi\frac{\partial H}{\partial y} + D_x\Phi\frac{\partial H}{\partial y'} = 0, \tag{10}$$

where Φ – is a coordinate of formal canonical operator. The equation (10) is an equation of the total derivatives of the first order regarding unknown coordinate for Φ, and it could be considered as an equation with ramifying variables. Its solution has the following form:

$$\Phi = \exp\left(-\int \frac{\partial H/\partial y}{\partial H/\partial y'}\,dx\right). \tag{11}$$

It's obvious that the formal operator defined by this coordinate is an **exponential nonlocal operator (ENO)**. Thus, to describe all equations of the second order, which allow a reduction to the equation of the first order by submitting invariants of possible operator, an exponential nonlocal operator is enough. For equations of higher order this is not true, but it's possible to prove more general statements playing an important role in the general theory of nonlocal operators.

Theorem 2 ***(second factorization theorem)*** [5, 6]. Random differential equation of n–th order $y^{(n)} = F\big(x, y, y', \ldots, y^{(n-1)}\big)$, could be factorized till the system of special form:

$$\begin{cases} z^{(n-k)} = G\big(x, z, z', \ldots, z^{(n-k-1)}\big), \\ z = H\big(x, y, y', \ldots, y^{(k)}\big), \qquad \dfrac{\partial z}{\partial y^{(k)}} \neq 0, \end{cases} \tag{12}$$

if it admits some formal operator for which $H\big(x, y,\ y', \ldots, y^{(k)}\big)$ is the ***youngest*** differential invariant on the manifold given by the equation. If the equation is factorized into the system (12), then it admits some formal operator for which $H\big(x, y, y', \ldots, y^{(k)}\big)$ is a differential invariant of k –order on the manifold.

Remark. If $k = n - 1$ and the first equation (12) has the form $z' = 0$, so the function H is the first integral of the original equation.

The search of the formal operator having a predetermined invariant in general consists in the solution of differential equation with full derivatives

$$\Phi\frac{\partial H}{\partial y} + D_x[\Phi]\frac{\partial H}{\partial y'} + \ldots + D_x^k[\Phi]\frac{\partial H}{\partial y^{(k)}} = 0.$$

ISBN: 978-1-61804-251-4

A perspective approach of this problem solution (alternative generalized symmetries) is proposed by one of the authors of the present article [7,8], but here we will consider only one special case leading to a closed form of nonlocal operator different from ENO.

IV. Nonexponential nonlocal operators

Let's consider the task of searching the classes of the third order equations admitting nonlocal nonexponential operator (NNO) of such form

$$X = \eta(x, y, y') \left(\int \zeta(x, y, y')\, dx \right) \partial_y. \tag{13}$$

Theorem 3. Any equation allowing the operator (13) admits also the local operator $\bar{X} = \eta(x, y, y')\, \partial_y$.

Consequence. To solve this problem we could consider the class of autonomous equations , i.e. put $\eta \equiv y'$ and look for the operator as

$$X = y' \left(\int \zeta(x, y, y')\, dx \right) \partial_y, \tag{14}$$

and then find all classes of equations using the principle of similarity of oneparametric point groups on the plane.

The following statement is fair.

Theorem 4 [9]. There is no nontrivial equation of the form $y''' = F(y)$, admitting the NNO of the form (13).

Therefore, let's consider an autonomous equation of the third order without an "elder" derivative admitting nonlocal nonexponential operator (14).

Theorem 5 [9]. The equation (7) with $F_x = 0$ admits NNO (14) if and only if the right part has the form

$$F(y, y') = y' \Big(C\big(y'\big)^2 + G(y) \Big) H(y) - \frac{1}{2C} G''(y) y', \tag{15}$$

wherein

$$\zeta(x, y, y') = C + \frac{G(y)}{\big(y'\big)^2}, \tag{16}$$

where $G(y)$ and $H(y)$ – are an arbitrary functions, $C \neq 0$ – is an arbitrary constant.

Remark. The value $C = 0$ is possible only if $G''(y) \equiv 0$. But in this case the original equation is trivial and can be easily integrated.

It's easy to prove [4] that the operator (14) does not have the first differential invariant (more precisely – an invariant, depending only on the first derivative). To calculate the second differential invariant of found operator it's useful to solve the equation

$$\tilde{\eta} \frac{\partial \Phi}{\partial y} + \tilde{\eta}_1 \frac{\partial \Phi}{\partial y'} + \tilde{\eta}_2 \frac{\partial \Phi}{\partial y''} = 0.$$

Substituting the received coordinates of the operator and splitting the equation by the nonlocal invariable I, we obtain a system of two equations

$$\begin{cases} y' \left[C + \dfrac{G(y)}{\big(y'\big)^2} \right] \dfrac{\partial \Phi}{\partial y'} + \big(2Cy'' + G'(y) \big) \dfrac{\partial \Phi}{\partial y''} = 0, \\ y' \dfrac{\partial \Phi}{\partial y} + y'' \dfrac{\partial \Phi}{\partial y'} + \\ + \left[y' \Big(C\big(y'\big)^2 + G(y) \Big) H(y) - \frac{1}{2C} G''(y) y' \right] \frac{\partial \Phi}{\partial y''} = 0. \end{cases} \tag{17}$$

It's necessary to note that in the second equation instead of y''' is used the right part of the equation (15), i.e. an invariant is placed **on the manifold** of solutions of the original equation. The solution of the first equation of the system (17) is a function

$$\Omega \left(y, \frac{2Cy'' + G'(y)}{C\big(y'\big)^2 + G(y)} \right), \tag{18}$$

the substitution of (18) in the second equation of the system leads to a linear equation of the first order with partial derivatives regarding the function Ω

$$\frac{\partial \Omega}{\partial y} + \big[H(y) - 2C\omega^2 \big] \frac{\partial \Omega}{\partial \omega} = 0, \tag{19}$$

where ω – is the second argument of the function Ω. The equation in characteristics of (19) is a Ricatti's equation in its canonical form, consequences, it is always could be solved as an linear equation of the second order. In a big amount of cases the solution of equation (19) could be expressed in a closed form – through an elementary or special functions. The type of submission subsequently depends on the function $H(y)$. For example, if $H(y) = y^k$ or $H(y) = e^y$, the second differential invariant is expressed through the Bessel functions, while in the case of degree function we obtain a special Ricatti's equation – if the expression $\frac{k+3}{k+2}$ is a half–integer, then the second differential invariant is an elementary function. For example, when $k = 0$

$$\Omega = \sqrt{2C} y - \operatorname{arth} \left(\frac{2Cy'' + G'(y)}{\sqrt{2C} \Big(C\big(y'\big)^2 + G(y) \Big)} \right).$$

Direct verification shows that because of the original equation is $\Omega' = 0$, i.e. there exists a factorization

$$\begin{cases} \Omega' = 0, \\ \Omega = \sqrt{2C} y - \operatorname{arth} \left(\dfrac{2Cy'' + G'(y)}{\sqrt{2C} \Big(C\big(y'\big)^2 + G(y) \Big)} \right). \end{cases}$$

Thus, the function Ω is an **autonomous first integral** of the original equation and the found symmetry is an analogue of the variation symmetry.

References

[1] Ibragimov N. H. *Transformation groups in mathematical physics*. M.: Nauka, 1983.

[2] Ibragimov N. H. *Group analysis experience*. M.: Znanie, "Math and cybernetics", N 7. – 1991.

[3] Zaitsev V. F. *About discrete–group analysis of ordinary differential equations*, // DAN SSSR, **299**, N 3, 1988. – P. 542–545.

[4] Avrashkov P. P., Zaitsev V. F. *Lee symmetries and first integrals of similar class of differential equations*, // Intercollegiate scientific works, vol. 8. Orel: OSTU, 1996. – P. 44–49.

[5] Zaitsev V. F. *Formal operators and factorization theorem for ordinary differential equations*, // " Symmetry and differential equations", Proceedings of the III International Conference. Krasnoyarsk, 2002. – P. 101–105.

[6] Zaitsev V. F. *About comprehensive description of symmetries of ordinary differential equations*, // Proceedings of Lobachevsky Center, v. 11 "Problems of modern mathematics." Kazan: ed. "UNIPRESS", 2001 – S. 93–96.

[7] Linchuk L. V. *Alternative generalized symmetries of ordinary differential equations of the second order*, // Some topical problems of modern mathematics and mathematical education. LXIII scientific conference "Herzenovskie Readings – 2010". St. Petersburg.: BAN, 2010. – P. 46–53.

[8] Linchuk L. V. *"Nonclassical" factorization of ordinary differential equations of the third order*, // Some topical issues of modern mathematics and mathematics education. Materials of LXVII scientific conference "Herzenovskie Readings – 2014". St. Petersburg.: Univ. RSPU. Herzen, 2014. – P. 90–98.

[9] Zaitsev V. F., Mkhitaryan M. G. *An inverse problem for nonexponential nonlocal operators admitted for ODE of the third order*, // Some topical issues of modern mathematics and mathematics education. Materials of LXVII scientific conference "Herzenovskie Readings – 2014". St. Petersburg.: Univ. RSPU. Herzen, 2014. – P. 58–62.

ISBN: 978-1-61804-251-4

Tangency-saddle singularities of Planar Bimodal Linear Systems

Josep Ferrer
Applied Mathematics Dept.
UPC- BarcelonaTech
08028-Barcelona, Av. Diagonal 647
Email: josep.ferrer@upc.edu

Marta Peña
Applied Mathematics Dept.
UPC- BarcelonaTech
08028-Barcelona, Av. Diagonal 647
Email: marta.penya@upc.edu

Antonio Susin
Applied Mathematics Dept.
UPC- BarcelonaTech
08028-Barcelona, Av. Diagonal 647
Email: toni.susin@upc.edu

***Abstract*—We continue the study of the bifurcations and the structural stability of planar bimodal linear dynamical systems (that is, systems consisting of two linear dynamics acting on each side of a straight line, assuming continuity along the separating line). Here we determine the tangency-saddle singularities in the saddle/spiral case, the only where they can appear.**

I. Introduction

Piecewise linear systems constitute a class of non-linear systems which have recently attracted the interest of researchers because of their interesting properties and the wide range of applications from which they arise. In [4] and [5] one studies the controllability of BLDS (bimodal linear dynamical systems). In [6] one begins the study of its structural stability. The structural stability of a system warrants that its qualitative behavior is preserved under small perturbations of their parameters. One focuses in planar BLDS, that is, two planar linear subsystems acting in complementary halfplanes, assuming continuity in the separating straight line. They have interesting theoretical properties as well as applications (see, for example, [1], [2], [4] and [7]).

In Section 3 we recall the results in [6]: by adapting the necessary conditions in [8], one obtains the list of possible structurally stable planar BLDS and one concludes that structural stability holds when (real) spirals do not appear; in addition, one studies the finite periodic orbits for the saddle/spiral case.

In Section 4 we enlarge this study to the tangency-saddle singularities. They can appear only in the saddle/spiral case. We prove that it does for a sequence of values of the trace of the spiral subsystem, whereas for the remainder values the BLDS is structurally stable.

Throughout the paper, $\mathbf{R}$ will denote the set of real numbers, $M_{n\times m}(\mathbf{R})$ the set of matrices having n rows and m columns and entries in $\mathbf{R}$ (in the case where $n = m$, we will simply write $M_n(\mathbf{R})$) and $Gl_n(\mathbf{R})$ the group of non-singular matrices in $M_n(\mathbf{R})$. Finally, we will denote by $e_1, \ldots, e_n$ the natural basis of the Euclidean space $\mathbf{R}^n$.

II. Structurally stable planar bimodal linear systems

Let us consider a bimodal linear dynamical system given by

$$\begin{cases} \dot{x}(t) = A_1 x(t) + B_1 & \text{if } Cx(t) \le 0 \\ \dot{x}(t) = A_2 x(t) + B_2 & \text{if } Cx(t) \ge 0 \end{cases}$$

where $A_1, A_2 \in M_n(\mathbf{R})$; $B_1, B_2 \in M_{n\times 1}(\mathbf{R})$; $C \in M_{1\times n}(\mathbf{R})$. We assume that the dynamics is continuous along the separating hyperplane $H = \{x \in \mathbf{R}^n : Cx = 0\}$; that is to say, that both subsystems coincide for $Cx(t) = 0$.
By means of a linear change in the state variable $x(t)$, we can consider $C = (1\ 0 \ldots 0) \in M_{1\times n}(\mathbf{R})$. Hence $H = \{x \in \mathbf{R}^n : x_1 = 0\}$ and continuity along H is equivalent to:

$$B_2 = B_1, \qquad A_2 e_i = A_1 e_i, \quad 2 \le i \le n.$$

We will write from now on $B = B_1 = B_2$.

Definition 1: In the above conditions, we say that the triple of matrices (A_1, A_2, B) defines a bimodal linear dynamical system. (BLDS.)

The placement of the equilibrium points will play a significative role in the dynamics of a BLDS. So, we define:

Definition 2: Let us assume that a subsystem of a BLDS has a unique equilibrium point, not lying in the separating hyperplane. We say that this equilibrium point is real *if it is located in the halfspace corresponding to the considered subsystem. Otherwise, we say that the equilibrium point is* virtual.

Our goal is to characterize the planar BLDS which are structurally stable in the sense of [8].

Definition 3: A triple of matrices (A_1, A_2, B) defining a BLDS is said to be (regularly) structurally stable if it has a neighborhood $V(A_1, A_2, B)$ such that for every $(A_1', A_2', B') \in V(A_1, A_2, B)$ there is a homeomorphism of $\mathbf{R}^2$ preserving the hyperplane H which maps the oriented orbits of (A_1', A_2', B') into those of (A_1, A_2, B) and it is differentiable when restricted to finite periodic orbits.

A natural tool in the study of BLDS is simplifying the matrices A_1, A_2, B by means of changes in the variables $x(t)$ which preserve the qualitative behavior of the system (in particular, the condition of structurally stability). So, we consider linear changes in the state variables space preserving the hyperplanes $x_1(t) = k$, which will be called *admissible basis changes*. Thus, they are basis changes given by a matrix $S \in Gl_n(\mathbf{R})$,

$$S = \begin{pmatrix} 1 & 0 \\ U & T \end{pmatrix}, \quad T \in Gl_{n-1}(\mathbf{R}), \quad U \in M_{n-1\times 1}(\mathbf{R}).$$

ISBN: 978-1-61804-251-4

See [3] for the resulting reduced forms.

Also, translations parallel to the hyperplane H are allowed.

III. Preliminaries

In [6] one proves the following results. Firstly,

Theorem 1: 1. The triples of matrices representing a structurally stable BLDS can be reduced (by means of an admissible basis change and a translation parallel to the separating line) to the form:

$$A_1 = \begin{pmatrix} T & 1 \\ -D & 0 \end{pmatrix}, A_2 = \begin{pmatrix} \tau & 1 \\ -\Delta & 0 \end{pmatrix}, B = \begin{pmatrix} 0 \\ b \end{pmatrix} \quad (*)$$

In particular, the only tangency point is $(0, 0)$.

2. The only possible structurally stable BLDS are those in Table 1.
3. A sufficient condition in order to be structurally stable is that none subsystem is a real spiral (cases 1, 2, 4, 5, 6, 8, 9, 10, 12, 13, 14 and 16).
4. In the case 3, it is structurally stable if and only if
   - 4.a the finite periodic orbits are hyperbolic and disjoint from the tangency points
   - 4.b there are not finite orbits connecting two saddles
   - 4.c there are not finite orbits connecting a saddle and a tangency point
5. In the cases 7, 11 and 15, it is structurally stable if and only if condition (a) holds.

Subsystem 1 \ 2	Virtual saddle	Real node	Real spiral	Real imp. node
Real saddle	1 ($b > 0$)	2 ($b > 0$)	3 ($b > 0$)	4 ($b > 0$)
Virtual node	5 ($b < 0$)	6 ($b > 0$)	7 ($b > 0$)	8 ($b > 0$)
Virtual spiral	9 ($b < 0$)	10 ($b < 0$)	11 ($b > 0$)	12 ($b > 0$)
Virtual imp. node	13 ($b < 0$)	14 ($b < 0$)	15 ($b < 0$)	16 ($b > 0$)

TABLE I.

Secondly, one focuses on conditions (a), (b) of case 3 for divergent spirals. Thus, let us assume a BLDS as in (*), verifying:

- The left subsystem is a (real) saddle, i.e.: $D < 0, b > 0$. In particular, its equilibrium point is $(\frac{b}{D}, -T\frac{b}{D})$, and the invariant manifold cut the separating line at $(0, -\frac{b}{\lambda_2})$ and $(0, -\frac{b}{\lambda_1})$, where $\lambda_2 < 0 < \lambda_1$ are the eigenvalues of A_1. ($\lambda_1 + \lambda_2 = T$, $\lambda_1\lambda_2 = D$.)
- The right subsystem is a (real) divergent spiral, i.e.: $\tau > 0, \tau^2 < 4\Delta, b > 0$. In particular, its equilibrium point is $(\frac{b}{\Delta}, -\tau\frac{b}{\Delta})$. We write $\alpha \pm i\beta$, $\beta > 0$ the eigenvalues of A_2. ($2\alpha = \tau$, $\alpha^2 + \beta^2 = \Delta$.)

Theorem 2: In the above conditions:

1. 1.a If $T > 0$, then there is not homoclinic orbit.
   - 1.b If $T = 0$, then there is a homoclinic orbit only for $\tau = 0$, which is a not considered case.
   - 1.c If $T < 0$, the only homoclinic (i.e., saddle-loop) orbit appears for $\tau = \tau_H > 0$ verifying

$$\exp(\alpha t)\sin(\beta t - \varphi) + \frac{\beta}{M} = 0, \quad \pi + \varphi \leq \beta t \leq \frac{3\pi}{2} + \varphi$$

being

$$t = \frac{1}{\tau}\ln\left(\frac{\lambda_2^2}{\lambda_1^2}\frac{\lambda_1^2 - \tau\lambda_1 + \Delta}{\lambda_2^2 - \tau\lambda_2 + \Delta}\right)$$

where $M\cos(\varphi) = \alpha - \frac{\alpha^2+\beta^2}{\lambda_2}$, $M\sin(\varphi) = \beta$.
Moreover, $\tau_H > \frac{T\Delta}{D}$.

2. 2.a If $T > 0$, then there are not finite periodic orbits.
   - 2.b If $T = 0$, then there are finite periodic orbits (all of them) only for $\tau = 0$, which is a not considered case.
   - 2.c If $T < 0$, a finite periodic orbit appears for $0 < \tau < \tau_H$, which is hyperbolic (indeed, attractive) and disjoint from the tangency points, and no saddle-tangency orbits appear.
3. In particular, the systems in case 3 with $T < 0$ and $0 < \tau < \tau_H$ are structurally stable.

IV. The tangency-saddle singularities

Here we tackle the case 3 for $T < 0$, $\tau > \tau_H$. We will see that there is a decreasing sequence $\tau_1, \tau_2, ... \to \tau_H$ of values of τ where tangency-saddle singularities appear. For the remainder values, the BLDS is structurally stable.

Theorem 3: In the conditions of case 3 and $T < 0$:

1. There exists a maximal value of τ, τ_1 (see Figure 1), for which a tangency-saddle orbit appears. This is for $\tau = \tau_1 > 0$ verifying

$$\exp(\alpha t)\sin(\beta t - \varphi) + \frac{\beta}{M} = 0, \quad \pi + \varphi \leq \beta t \leq \frac{3\pi}{2} + \varphi$$

being

$$t = \frac{1}{\tau}\ln\left(\frac{1 + \tau + \Delta}{\lambda_1^2}\right)$$

where $M\cos(\varphi) = \alpha$, $M\sin(\varphi) = \beta$.
Moreover, $\tau_1 > \lambda_1^2 - \Delta - 1$.
It is the only value of τ for which the tangent orbit at $(0, 0)$ has its first intersection with the separating hyperplane just at $(0, -b/\lambda_1)$.

2. There exists a decreasing sequence $(\tau_1, \tau_2, ..., \tau_k, ...) \to \tau_H$, $k \geq 1$ (see Figures 2 and 3), for which tangency-saddle orbits appear.
For the value $\tau = \tau_k$ the orbit starting in the tangency point $(0, 0)$ has its $(2k-1)$th intersection with the separating hyperplane just at $(0, -b/\lambda_1)$.
3. For the remainder values of $\tau > \tau_H$, the BLDS is structurally stable.

Proof:

1. From [6], the first intersection of a spiral passing through $(0, 0)$ with the hyperplane must verify

$$\exp(\alpha t)\sin(\beta t - \varphi) + \frac{\beta}{M} = 0, \quad \pi + \varphi \leq \beta t \leq \frac{3\pi}{2} + \varphi$$

where $M\cos(\varphi) = \alpha$, $M\sin(\varphi) = \beta$.
Moreover, again from [6], a spiral cuts $x_1 = 0$ in x_{21} and x_{22} if and only if

$$\exp(\mu t) = \frac{b + \mu x_{22}}{b + \mu x_{21}}$$

ISBN: 978-1-61804-251-4

where $\mu = \alpha + i\beta$. Imposing that $x_{21} = 0$ and $x_{22} = -b/\lambda_1$ we get

$$t = \frac{1}{\tau} \ln(\frac{1+\tau+\Delta}{\lambda_1^2})$$

and from it the bound for τ_1.

2. For $\tau = \tau_1$ the tangent orbit at $(0,0)$ intersects $x_1 = 0$ just at $(0, -b/\lambda_1)$ (see Figure 1). When τ decreases, this (first) intersection point ascends, so that the orbit completes a full turn and intersects the axe $x_1 = 0$ twice between $(0, -b/\lambda_2)$ and $(0, -b/\lambda_1)$, and again under $(0, -b/\lambda_1)$ if $\tau - \tau_1$ is small enough. It is clear that this third intersection ascends when τ increases, so that for a certain (unique) value $\tau = \tau_2$ this third intersection point is just $(0, -b/\lambda_1)$ (see Figure 2).
   Additional degrowth of τ gives a second turn (with two additional intersections between $(0, -b/\lambda_2)$ and $(0, -b/\lambda_1)$ and a fifth intersection with $x_1 = 0$ under $(0, -b/\lambda_1)$. As above, for a certain (unique) value $\tau = \tau_3$ this fifth intersection point is just $(0, -b/\lambda_1)$ (see Figure 3).
   By recurrence, one obtains a sequence of decreasing values $\tau_1, \tau_2, ..., \tau_k, ...$ for which the tangent orbit at $(0,0)$ intersects $x_1 = 0$ in $0, 1, ..., 2k-2, ...$ points between $(0, -b/\lambda_2)$ and $(0, -b/\lambda_1)$, and another one just at $(0, -b/\lambda_1)$.
   An analogous reasoning shows that $\lim \tau_k = \tau_H$: for $\tau = \tau_H$ the saddle orbit through $(0, -b/\lambda_2)$ intersects again $x_1 = 0$ at $(0, -b/\lambda_1)$, whereas the tangent orbit at $(0,0)$ turns over the spiral toward this homoclinic orbit; for any slightly greater value $\tau_H + \epsilon$ the above saddle orbit intersects $x_1 = 0$ under $(0, -b/\lambda_1)$, so that the tangent orbit passes between this new intersection point and $(0, -b/\lambda_1)$; therefore, the reasoning in the above paragraph shows that there is some $\tau_k < \tau_H + \epsilon$.
3. By construction, for $\tau_k < \tau < \tau_{k+1}$ there are not tangency-saddle orbits. Moreover, the orbits for τ run between the ones for τ_k and τ_{k+1} so that neither finite periodic orbits nor saddle-tangency orbits can occur.

■

Exemp 1: For $T = -1, D = -1, \Delta = 5, b = 1$, we plot the tangency-saddle orbits: $\tau_1 = 1.145, \tau_2 = 0.782, \tau_3 = 0.745; \tau_H = 0.742$.

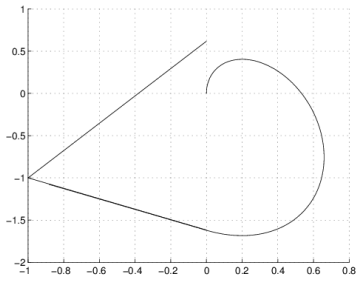


Fig. 1. Appearance of a tangency-saddle orbit: $T = -1, D = -1, \tau = \tau_1 = 1.145, \Delta = 5, b = 1$

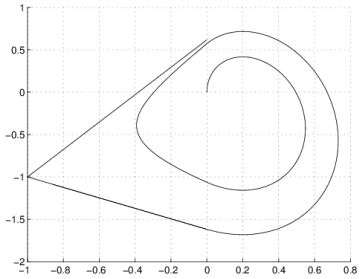


Fig. 2. Appearance of a tangency-saddle orbit: $T = -1, D = -1, \tau = \tau_2 = 0.782, \Delta = 5, b = 1$

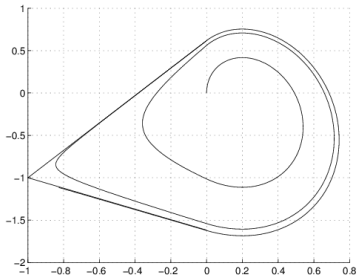


Fig. 3. Appearance of a tangency-saddle orbit: $T = -1, D = -1, \tau = \tau_3 = 0.745, \Delta = 5, b = 1$

ACKNOWLEDGMENT

We thank Prof. Rafael Ramirez for many helpful discussions during the preparation of the manuscript. This article is supported by DGICYTMTM2011-23892 (first and second author) and TIN2010-20590-C02-C1 (third author).

REFERENCES

[1] CAMLIBEL, K., HEEMELS, M. & SCHUMACHER, H., *Stability and controllability of planar bimodal linear complementarity systems,* Proceedings of the 42nd IEEE Conference on Decision and Control, 1651–1656, 2003.

[2] DI BERNARDO, M., PAGANO, D. J. & PONCE, E., *Nonhyperbolic boundary equilibrium bifurcations in planar Filippov systems: a case study approach*, J. Bifur. Chaos Appl. Sci. Engin., **18**, 1377–1392, 2008.

[3] FERRER, J., MAGRET, M. & PEÑA, M., *Bimodal piecewise linear systems. Reduced forms*, International Journal of Bifurcation and Chaos, **20**, 2795–2808, 2010.

[4] FERRER, J., MAGRET, M. & PEÑA, M., *Differentiable Families of Planar Bimodal Linear Control Systems*, Mathematical Problems in Engineering, Article ID 292813, 9 pages, http://dx.doi.org/10.1155/2014/292813, 2014.

[5] FERRER, J., PACHA, J. R. & PEÑA, M., *Controllability of continuous bimodal linear systems,* Mathematical Problems in Engineering (special issue Mathematical Modeling, Analysis, and Control of Hybrid Dynamical Systems), Volume 2013 (2013), Article ID 342548, 14 pages.

[6] FERRER, J., PEÑA, M. & SUSIN, A., *Structural stability of planar bimodal linear systems,* Conference Proceedings of ICNAAM 2013, 2013.

[7] LLIBRE, J., ORDONEZ, M. & PONCE, E., *On the existence and uniqueness of limit cycles in planar continuous piecewise linear systems without symmetry,* Nonlinear Analysis: Real World Applications, 2013.

[8] SOTOMAYOR, J. & GARCIA, R., *Structural stability of piecewise-linear vector fields*, Journal of Differential Equations, **192**, 553–565, 2003.

Lower Bounds on the Convergence Rate of the Markov Symmetric Random Search

Alexey Tikhomirov

Abstract—The convergence rate of the Markov random search algorithms designed for finding the extremizer of a function is investigated. It is shown that, for a wide class of random search methods that possess a natural symmetry property, the number of evaluations of the objective function needed to find the extremizer accurate to ε cannot grow slower than $|\ln \varepsilon|$.

Keywords—random search, stochastic optimization, global optimization, estimate of convergence rate.

I. Introduction

LET the objective function $f: X \mapsto \mathbf{R}$ (where, for instance, $X = \mathbf{R}^d$) take its minimal value at a single point x_*. Consider the problem of finding the global minimizer x_* of this function up to a given accuracy ε (approximation by argument). One way of solving this problem is to use random search algorithms (see [1]–[12]). Such algorithms have long been used for solving difficult optimization problems. However, theoretical results for the convergence rate of those algorithms are scarce (see [1], [2], [4]). This paper is devoted to the theoretical analysis of the convergence rate of the Markov random search algorithms designed for finding the global extremizer. Note that the simulated annealing algorithm, which is a well-known stochastic global optimization algorithm, belongs to this class.

We use the number of evaluations of the objective function required to achieve the prescribed accuracy ε of the solution as the characteristic of the convergence rate. The main reason for the choice of this characteristic is that the objective function evaluations require the most part of the computational effort involved in the algorithm execution. In addition, this characteristic is convenient for the comparison of different random search algorithms.

It turns out that Markov random search algorithms cannot be very fast. It is shown that, for a wide class of random search methods that possess a natural symmetry property, the number of evaluations of the objective function needed to find the extremizer accurate to ε cannot grow slower than $|\ln \varepsilon|$.

The results presented in this paper make it possible to estimate the potential capability of Markov algorithms and draw the conclusion that the convergence rate of some algorithms is close to the optimal one, at least, in the order of dependence on ε.

It is worth noting that the theoretical lower bounds on the computational effort of numerical methods for solving optimization problems for the standard classes (smooth, non-smooth convex, strongly convex, smooth convex, and convex stochastic) are discussed in [9]. The results obtained in the present paper can be considered as a small extension of the results presented in [9]. These results are some extension of the results presented in [12].

A. Tikhomirov is with the Applied Mathematics and Informatics Department, Novgorod State University, Velikiy Novgorod, Russia e-mail: Tikhomirov.AS@mail.ru.

Manuscript received July 10, 2014.

II. Statement of the Problem

A. Optimization Space

The *optimization space* is defined as a set X equipped with a metric . In this paper, we consider the case of the Euclidean space $\mathbf{R}^d$. We use the following metrics :

$$(x,y) = {}_2(x,y) = \left(\sum_{n=1}^{d}(x_n - y_n)^2\right)^{1/2},$$
$$(x,y) = {}_\infty(x,y) = \max_{1\le i\le d}|x_i - y_i|,$$

where $x = (x_1,\dots,x_d)$ and $y = (y_1,\dots,y_d)$.

The closed ball of radius r centered at the point x is denoted by

$$B_r(x) = \{y \in \mathbf{R}^d, \text{ such that } (x,y) \le r\}.$$

B. Objective Function

Throughout this paper, we assume that the objective function $f: \mathbf{R}^d \mapsto \mathbf{R}$ is measurable, and takes its minimum value at a unique point $x_* = \arg\min\{f(x), \text{ such that } x \in \mathbf{R}^d\}$.

C. Markov Random Search

A *random search* is defined as an arbitrary sequence of random variables $\{ {}_n\}_{n\ge 0}$ taking values in $\mathbf{R}^d$. Following [1], we give a general scheme of Markov random search algorithms.

The following algorithm simulates the Markov random search $\{ {}_n\}_{n\ge 0}$ in $\mathbf{R}^d$ with the initial point ${}_0 = x \in \mathbf{R}^d$.

Algorithm 1 (A general scheme of Markov algorithms)

Step 1. Set ${}_0 = x$ and the iteration number $n = 1$.

Step 2. Obtain a point ${}_n$ in $\mathbf{R}^d$ by sampling from the distribution $P_n(\ {}_{n\ 1}, \cdot\ )$. Here $P_n(\ {}_{n\ 1}, \cdot\ )$ is the *transition probability*; this probability may depend on n and ${}_{n\ 1}$.

Step 3. Set

$${}_n = \begin{cases} {}_n, & \text{with probability } Q_n, \\ {}_{n\ 1}, & \text{with probability } 1 \quad Q_n. \end{cases}$$

Here Q_n is the *acceptance probability*; this probability may depend on ${}_n$, ${}_{n\ 1}$, $f(\ {}_n)$ and $f(\ {}_{n\ 1})$.

ISBN: 978-1-61804-251-4

Step 4. Check a stopping criterion. If the algorithm does not stop, substitute $n+1$ for n and return to Step 2.

Upon calculating the next trial point $\ _n$ (at the second step of the algorithm), the search either goes to this point with the probability Q_n at the third step or remains at the preceding point $\ _{n\ 1}$.

Particular choices of transition probabilities $P_n(x,\cdot\,)$ and acceptance probabilities Q_n lead to specific Markov global random search algorithms. The most well-known among them is the celebrated 'simulated annealing' algorithm.

D. Simulated Annealing

The name of the algorithm originated from its similarity to the physical procedure called annealing used to remove defects from metals and crystals by heating them locally near the defect to dissolve the impurity and then slowly re-cooling them so that they could find a basic state with a lower energy configuration.

A general simulated annealing algorithm is Algorithm 1 with acceptance probabilities

$$Q_n = \begin{cases} 1, & \text{if } \Delta_n \le 0, \\ \exp(\ \ \beta_n \Delta_n), & \text{if } \Delta_n > 0, \end{cases} \tag{1}$$

where $\beta_n \ge 0$ $(n = 1, 2, \ldots)$ are search parameters and $\Delta_n = f(\ _n) \quad f(\ _{n\ 1})$.

The choice (1) for the acceptance probability Q_n means that any 'promising' new point $\ _n$ (for which $f(\ _n) \le f(\ _{n\ 1})$) is accepted unconditionally; a 'non-promising' point (for which $f(\ _n) > f(\ _{n\ 1})$) is accepted with probability $Q_n = \exp(\ \ \beta_n \Delta_n)$.

E. Markov Monotone Search

If the acceptance probabilities Q_n are determined by

$$Q_n = \begin{cases} 1, & \text{if } f(\ _n) \le f(\ _{n\ 1}), \\ 0, & \text{if } f(\ _n) > f(\ _{n\ 1}), \end{cases}$$

then we obtain the Markov monotone random search (see [1, p. 122]), which plays an important role in the further reasoning. This search is monotone in the sense that the inequalities $f(\ _n) \le f(\ _{n\ 1})$ hold true with the unit probability for all $n > 0$. The Markov monotone search can be interpreted as the limit case of the simulated annealing algorithm with $\beta_n = +\infty$ for all n.

F. Markov Symmetric Random Search

In accordance with the structure of Algorithm 1, we call the distributions $P_n(\ _{n\ 1}, \cdot\,)$ *trial transition probabilities*. We shall consider the case where the trial transition probabilities $P_n(x, \cdot\,)$ have *symmetric densities* $p_n(x,y)$ of the form

$$p_n(x,y) = g_{n,x} \quad (x,y)\ , \tag{2}$$

where $\ $ is a metric and $g_{n,x}$ are non-increasing functions of a positive argument.

The Markov search defined by Algorithm 1 with the trial transition functions with densities (2) is called *the Markov symmetric random search.*

G. Characteristics of Random Search

We use a random search to find the minimizer x_* with a given accuracy ε (approximation with respect to the argument). In this case, we want the search to hit the ball $B_\varepsilon(x_*)$. Denote by

$$\tau_\varepsilon = \min\{n \ge 0,\ \text{such that}\ \ _n \in B_\varepsilon(x_*)\}$$

the time when the search first hits the ε-neighborhood of the global minimizer.

The distribution of the random variable τ_ε provides sufficient information about the quality of the random search. Indeed, usually we assume that there is no need to evaluate the function f for the simulation of the distributions P_n. Therefore, in the process of performing τ_ε iterations of Algorithm 1, the function f is evaluated $\tau_\varepsilon + 1$ times.

We use one characteristic of the convergence rate of the random search. The *computational complexity* of the random search is defined as $\mathsf{E}\tau_\varepsilon$. It is interpreted as the average number of search steps needed to hit the set $B_\varepsilon(x_*)$.

III. Lower Bounds on the Convergence Rate

A. Estimate of the Computational Complexity

The main result of this paper is Theorem 1. It is proved in this theorem that the computational effort of the Markov symmetric random search needed to guarantee the required accuracy ε of the solution cannot grow slower than $|\ln \varepsilon|$.

Theorem 1. Let the function $f\colon \mathbf{R}^d \mapsto \mathbf{R}$ take its minimum value at a unique point x_*. Consider the Markov symmetric random search $\{\ _n\}_{n\ge 0}$ defined by Algorithm 1 whose transition probabilities have densities of form (2). Let x be the starting point of the search, $0 < \varepsilon < \delta = \ (x, x_*)$, and $n = 1, 2, \ldots$ Then, it holds that

$$\mathsf{E}\tau_\varepsilon \ge \ln(\delta/\varepsilon) + 1, \quad \mathsf{P}(\tau_\varepsilon \le n) \le \frac{\varepsilon}{\delta} \sum_{i=0}^{n} \frac{\ln^i(\delta/\varepsilon)}{i!}. \tag{3}$$

The proof of the Theorem 1 is similar to the proof of results in [12].

B. The Accuracy of the Estimates of Theorem 1

We restrict ourselves to the consideration of the one-dimensional optimization space $\mathbf{R}$, the simple objective function $f(x) = |x|$, and the Markov monotone random search in which the trial transition functions have form

$$P(x,\cdot\,) = U_{2|x|}(x,\cdot\,), \tag{4}$$

where $U_r(x,\cdot\,)$ is the uniform distribution on the ball $B_r(x)$ of radius r with center at x.

We can calculate the computational complexity of this random search and compare it with estimate (3) of Theorem 1.

Lemma 1. Let the optimization space $X = \mathbf{R}$ and $f(x) = |x|$. Let $\{\ _n\}_{n\ge 0}$ be the Markov monotone search with the trial transition functions given by (4). Let x be the starting point of the search, and $0 < \varepsilon < |x|$. Then, it holds that

$$\mathsf{E}\tau_\varepsilon = 2\ln(|x|/\varepsilon) + 2.$$

This result shows that estimates (3) of Theorem 1 are accurate estimates of the convergence rate.

ISBN: 978-1-61804-251-4

REFERENCES

[1] A. Zhigljavsky and A. Žilinskas, *Stochastic Global Optimization*, Springer, Berlin, 2008.

[2] J.C. Spall, *Introduction to stochastic search and optimization: estimation, simulation, and control*, Wiley, New Jersey, 2003.

[3] S.M. Ermakov and A.A. Zhigljavsky, *On the Random Search of Global Extremum*, Teor. Veroyatn. Ee Primen., 1983, No. 1, pp. 129–136.

[4] J.C. Spall, S.D. Hill and D.R. Stark, *Theoretical framework for comparing several stochastic optimization approaches*, Probabilistic and randomized methods for design under uncertainty. London, Springer, 2006, pp. 99–117.

[5] G. Yin, *Rates of convergence for a class of global stochastic optimization algorithms*, SIAM Journal on Optimization, 1999, vol. 10, no. 1, pp. 99–120.

[6] L. Ingber, *Very fast simulated re-annealing*, Mathl. Comput. Modelling, 1989, v. 12. pp. 967–973.

[7] A.S. Lopatin, *Simulation Annealing Method*, Stokhasticheskaya Optimizatsiya v Informatike, No. 1, 133–149, 2005.

[8] O.N. Granichin and B.T. Polyak, *Randomized Estimation and Optimization Algorithms under Almost Arbitrary Noise*, Nauka, Moscow, 2003.

[9] A.S. Nemirovskii and D.B. Yudin, *Complexity of Problems and Efficiency of Optimization Methods*, Nauka, Moscow, 1979.

[10] A. Tikhomirov, T. Stojunina and V. Nekrutkin, *Monotonous Random Search on a Torus: Integral Upper Bounds for the Complexity*, Journal of Statistical Planning and Inference, 2007, 137, pp. 4031–4047.

[11] A.S. Tikhomirov, *On the Convergence Rate of the Simulated Annealing Algorithm*, Computational Mathematics and Mathematical Physics, 2010, vol. 50, no. 1, pp. 19–31.

[12] A.S. Tikhomirov, *Lower Bounds on the Convergence Rate of the Markov Symmetric Random Search*, Computational Mathematics and Mathematical Physics, 2011, Vol. 51, No. 9, pp. 1524–1538.

ISBN: 978-1-61804-251-4

Simulation of emission spectra for LH2 ring: Fluctuations in radial positions of molecules

Pavel Heřman, David Zapletal and Pavel Kabrhel

Abstract—Absorption and steady state fluorescence spectra of exciton states for ring molecular systems are presented. The B850 ring from peripheral cyclic antenna unit LH2 of the bacterial photosystem from purple bacteria can be modeled by such system. The cumulant-expansion method of Mukamel et al. is used for the calculation of spectral responses of the system with exciton-phonon coupling. Dynamic disorder, interaction with a bath, in Markovian approximation simultaneously with uncorrelated static disorder in radial positions of molecules on the ring is taking into account in our simulations. We compare calculated absorption and steady state fluorescence spectra for LH2 ring obtained within the full Hamiltonian model with our previous results within the nearest neighbour approximation model.

Keywords—Absorption and fluorescence spectrum, fluctuations in radial positions of molecules, LH2, static and dynamic disorder.

I. Introduction

In the process of photosynthesis (in plants, bacteria, and blue-green algae), solar energy is used to split water and produce oxygen molecules, protons and electrons. Our interest is mainly focused on the first (light) stage of photosynthesis in purple bacteria. Solar photon is absorbed by a complex system of membrane-associated pigment-proteins (light-harvesting (LH) antenna) and absorbed energy is efficiently transferred to a reaction center (RC), where it is converted into a chemical energy [1].

The antenna systems of photosynthetic units from purple bacteria are formed by ring units LH1, LH2, LH3, and LH4. Their geometric structures are known in great detail from X-ray crystallography. The general organization of above mentioned light-harvesting complexes is the same: identical subunits are repeated cyclically in such a way that a ring-shaped structure is formed. However the symmetries of these rings are different.

Crystal structure of LH2 complex contained in purple bacterium Rhodopseudomonas acidophila was first described in high resolution by McDermott et al. [2] in 1995, then further e.g. by Papiz et al. [3] in 2003. The bacteriochlorophyll (BChl) molecules are organized in two concentric rings. One ring features a group of nine well-separated BChl molecules (B800) with absorption band at about 800 nm. The other ring consists of eighteen closely packed BChl molecules (B850) absorbing around 850 nm. LH2 complexes from other purple bacteria have analogous ring structure.

Some bacteria express also other types of complexes such as the B800-820 LH3 complex (*Rhodopseudomonas acidophila* strain 7050) or the LH4 complex (*Rhodopseudomonas palustris*). LH3 complex like LH2 one is usually nonameric but LH4 one is octameric. While the B850 dipole moments in LH2 ring have tangential arrangement, in the LH4 ring they are oriented more radially. Mutual interactions of the nearest neighbour BChls in LH4 are approximately two times smaller in comparison with LH2 and have opposite sign. At this article we focus on LH2 complex.

Despite intensive study of bacterial antenna systems, e.g. [2]–[5], the precise role of the protein moiety for governing the dynamics of the excited states is still under debate. At room temperature the solvent and protein environment fluctuates with characteristic time scales ranging from femtoseconds to nanoseconds. The simplest approach is to substitute fast fluctuations by dynamic disorder and slow fluctuation by static disorder.

In our previous papers we presented results of simulations doing within the nearest neighbour approximation model. In several steps we extended the former investigations of static disorder effect on the anisotropy of fluorescence made by Kumble and Hochstrasser [6] and Nagarajan et al. [7]–[9] for LH2 ring. After studying the influence of diagonal dynamic disorder for simple systems (dimer, trimer) [10]–[12], we added this effect into our model of LH2 ring by using a quantum master equation in Markovian and non-Markovian limits [13]–[16]. We also studied influence of four types of uncorrelated static disorder [17]–[19] (Gaussian disorder in local excitation energies, Gaussian disorder in transfer integrals, Gaussian disorder in radial positions of BChls on the ring and Gaussian disorder in angular positions of BChls on the ring). Influence of correlated static disorder, namely an elliptical deformation of the ring, was also taken into account [13]. We also investigated the time dependence of fluorescence anisotropy for the LH4 ring with different types of uncorrelated static disorder [15], [20].

Recently we have focused on the modeling of absorption and steady state fluorescence spectra. Our results for LH2 and LH4 rings within the nearest neighbour approximation model have been presented in [21]–[26]. The results for LH2 ring within full Hamiltonian model have been published in [27]–[29].

Main goal of our present paper is the comparison of the results for B850 ring from LH2 complex calculated within full Hamiltonian model with our previous results calculated within

This work was supported in part by the Faculty of Science, University of Hradec Králové - specific research project No. 2106/2014.

P. Heřman and P. Kabrhel are with Department of Physics, Faculty of Science, University of Hradec Králové, Rokitanského 62, 50003 Hradec Králové, Czech Republic, e-mail: pavel.herman@uhk.cz

D. Zapletal is with Institute of Mathematics and Quantitative Methods, Faculty of Economics and Administration, University of Pardubice, Studentská 95, 53210 Pardubice, Czech Republic, e-mail: david.zapletal@upce.cz

ISBN: 978-1-61804-251-4

the nearest neighbour approximation model [22], [23]. The rest of the paper is organized as follows. Section II introduces the ring model with the static disorder and dynamic disorder (interaction with phonon bath) and the cumulant expansion method, which is used for the calculation of spectral responses of the system with exciton-phonon coupling. Computational point of view is mentioned in Section III. The presented results of our simulations and used units and parameters could be found in Section IV, in Section V some conclusions are drawn.

II. Physical Model

We assume that only one excitation is present on the ring after an impulsive excitation. The Hamiltonian of an exciton in the ideal ring coupled to a bath of harmonic oscillators reads

$$H^0 = H^0_{\rm ex} + H_{\rm ph} + H_{\rm ex-ph}. \quad (1)$$

Here the first term,

$$H^0_{\rm ex} = \sum_{m,n(m\neq n)} J_{mn} a^\dagger_m a_n, \quad (2)$$

corresponds to an exciton, e.g. the system without any disorder. The operator $a^\dagger_m$ (a_m) creates (annihilates) an exciton at site m, J_{mn} (for $m \neq n$) is the so-called transfer integral between sites m and n. The second term in (1),

$$H_{\rm ph} = \sum_q \hbar\omega_q b^\dagger_q b_q, \quad (3)$$

represents phonon bath in harmonic approximation (the phonon creation and annihilation operators are denoted by $b^\dagger_q$ and b_q, respectively). Last term in (1),

$$H_{\rm ex-ph} = \frac{1}{\sqrt{N}} \sum_m \sum_q G^m_q \hbar\omega_q a^\dagger_m a_m (b^\dagger_q + b_q), \quad (4)$$

describes exciton-phonon interaction which is assumed to be site-diagonal and linear in the bath coordinates (the term G^m_q denotes the exciton-phonon coupling constant).

Inside one ring the pure exciton Hamiltonian can be diagonalized using the wave vector representation with corresponding delocalized "Bloch" states α and energies E_α. Considering homogeneous case with only the nearest neighbour transfer matrix elements

$$J_{mn} = J_0(\delta_{m,n+1} + \delta_{m,n-1}) \quad (5)$$

and using Fourier transformed excitonic operators (Bloch representation)

$$a_\alpha = \sum_n a_n {\rm e}^{{\rm i}\alpha n},\ \alpha = \frac{2\pi}{N} l,\ l = 0, \pm 1, \ldots, \pm\frac{N}{2}, \quad (6)$$

the simplest exciton Hamiltonian in α - representation reads

$$H^0_{\rm ex} = \sum_\alpha E_\alpha a^\dagger_\alpha a_\alpha, \quad E_\alpha = -2J_0 \cos\alpha, \quad (7)$$

see Fig. 1 - right column. In case of the full Hamiltonian model (dipole-dipole approximation), energetic band structure slightly differs (Fig. 1 - left column). Differences of energies in lower part of the band are larger and in upper part of the band are smoller in comparison with the nearest neighbour approximation model.

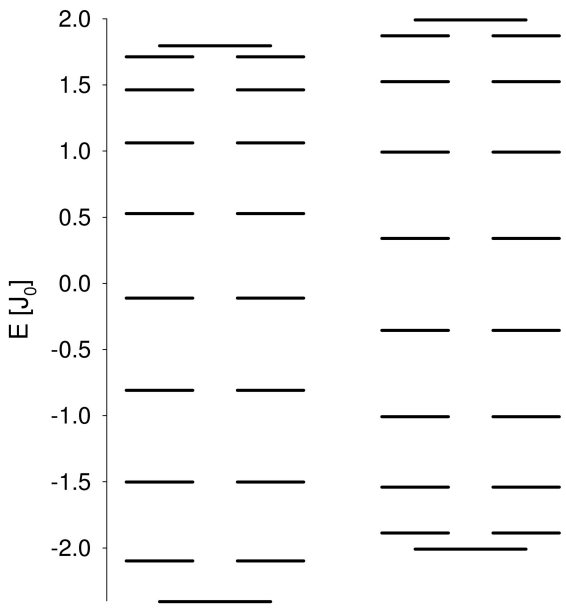


Fig. 1. Energetic band structure of the ring from LH2 (left column - full Hamiltonian model, right column - the nearest neighbour approximation model).

Influence of uncorrelated static disorder is modeled by the fluctuations of radial positions δr_n of bacteriochlorophylls on the ring with Gaussian distribution and standard deviation Δ_r. The Hamiltonian $H_{\rm s}$ of the uncorrelated static disorder adds to the Hamiltonian $H^0_{\rm ex}$ of the ideal ring.

The cumulant-expansion method of Mukamel et al. [30], [31] is used for the calculation of spectral responses of the system with exciton-phonon coupling. Absorption $OD(\omega)$ and steady-state fluorescence $FL(\omega)$ spectrum can be expressed as

$$OD(\omega) = \omega \sum_\alpha d^2_\alpha \times \times \ {\rm Re} \int_0^\infty dt {\rm e}^{{\rm i}(\omega-\omega_\alpha)t - g_{\alpha\alpha\alpha\alpha}(t) - R_{\alpha\alpha\alpha\alpha}t}, \quad (8)$$

$$FL(\omega) = \omega \sum_\alpha P_\alpha d^2_\alpha \times \times {\rm Re} \int_0^\infty dt {\rm e}^{{\rm i}(\omega-\omega_\alpha)t + {\rm i}\lambda_{\alpha\alpha\alpha\alpha}t - g^*_{\alpha\alpha\alpha\alpha}(t) - R_{\alpha\alpha\alpha\alpha}t}. \quad (9)$$

Here $\vec{d}_\alpha = \sum_n c^\alpha_n \vec{d}_n$ is the transition dipole moment of eigenstate α, c^α_n are the expansion coefficients of the eigenstate α in site representation and P_α is the steady state population of the eigenstate α. The inverse lifetime of exciton state $R_{\alpha\alpha\alpha\alpha}$ [32] is given by the elements of Redfield tensor $R_{\alpha\beta\gamma\delta}$ [33]. It is a sum of the relaxation rates between exciton states, $R_{\alpha\alpha\alpha\alpha} = -\sum_{\beta\neq\alpha} R_{\beta\beta\alpha\alpha}$. The g-function and λ-values in (9) are given by

$$g_{\alpha\beta\gamma\delta} = -\int_{-\infty}^{\infty} \frac{d\omega}{2\pi\omega^2} C_{\alpha\beta\gamma\delta}(\omega) \times \times \left[\coth\frac{\omega}{2k_{\rm B}T}(\cos\omega t - 1) - {\rm i}(\sin\omega t - \omega t) \right], \quad (10)$$

ISBN: 978-1-61804-251-4

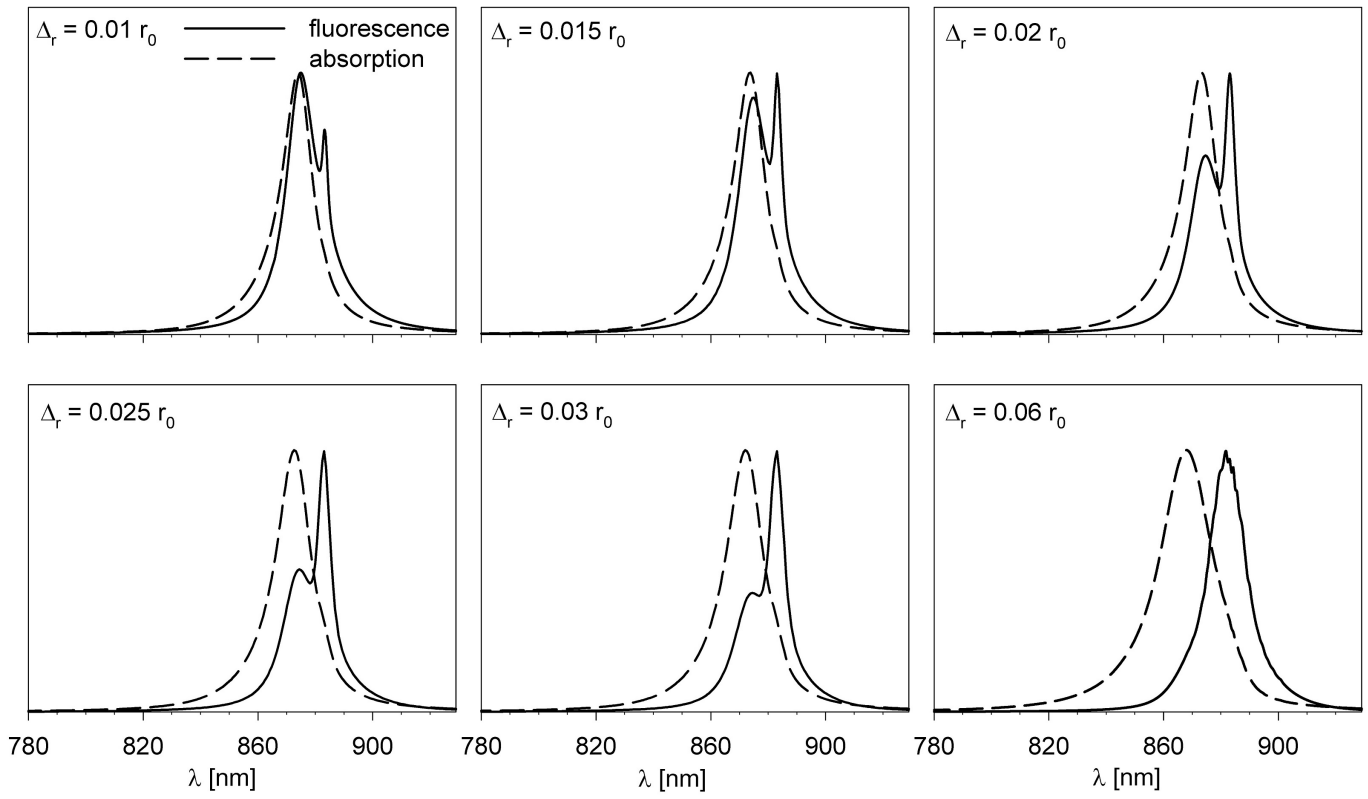


Fig. 2. Calculated $FL(\omega)$ and $OD(\omega)$ spectra averaged over 2000 realizations of static disorder in radial positions of molecules on the ring δr for full Hamiltonian model (low temperature $kT = 0.1\ J_0$, six strengths $\Delta_r = 0.010,\ 0.015,\ 0.020\ , 0.025,\ 0.030,\ 0.060\ r_0$).

$$\lambda_{\alpha\beta\gamma\delta} = -\lim_{t\to\infty}\frac{d}{dt}\mathrm{Im}\{g_{\alpha\beta\gamma\delta}(t)\} =$$

$$= \int_{-\infty}^{\infty}\frac{d\omega}{2\pi\omega}C_{\alpha\beta\gamma\delta}(\omega). \tag{11}$$

The matrix of spectral densities $C_{\alpha\beta\gamma\delta}(\omega)$ in the eigenstate (exciton) representation reflects one-exciton states coupling to the manifold of nuclear modes. In what follows only a diagonal exciton phonon interaction in site representation is used (see (1)), i.e., only fluctuations of the pigment site energies are assumed and the restriction to the completely uncorrelated dynamical disorder is applied. In such case each site (i.e. each chromophore) has its own bath completely uncoupled from the baths of the other sites. Furthermore it is assumed that these baths have identical properties [14], [34], [35]

$$C_{mnm'n'}(\omega) = \delta_{mn}\delta_{mm'}\delta_{nn'}C(\omega). \tag{12}$$

After transformation to the exciton representation we have

$$C_{\alpha\beta\gamma\delta}(\omega) = \sum_n c_n^\alpha c_n^\beta c_n^\gamma c_n^\delta C(\omega). \tag{13}$$

Various models of spectral density of the bath are used in literature [32], [36], [37]. In our present investigation we have used the model of Kühn and May [36]

$$C(\omega) = \Theta(\omega) j_0 \frac{\omega^2}{2\omega_c^3}\mathrm{e}^{-\omega/\omega_c} \tag{14}$$

which has its maximum at $2\omega_c$.

III. Computational point of view

To obtain absorption and steady state fluorescence spectra it is necessary to calculate single ring $OD(\omega)$ and $FL(\omega)$ spectra for large number of different static disorder realizations created by random number generator. Finally these results have to be averaged over all realizations of static disorder.

For our previous calculations of absorption and fluorescence spectra (for uncorrelated static disorder in local excitation energies δE_n and transfer integrals δJ_{mn}) software package *Mathematica* [38] was used. Standard numerical integration method used in *Mathematica* proved to be unsuitable in case of full Hamiltonian model and static disorder in radial positions of molecules δr_n. It was not possible to achieve satisfactory convergence by above mentioned integration method from *Mathematica*. This is the reason a procedure in Fortran was created for present calculations. Integrated function is oscillating and damped. That is why the function was integrated as a sum of contributions from individual cycles of oscillation. These contributions were added until required accuracy was achieved.

IV. Results

Above mentioned uncorrelated static disorder in radial positions of molecules on the ring has been taken into account in our simulations simultaneously with dynamic disorder in Markovian approximation. Dimensionless energies normalized to the transfer integral $J_{12} = J_0$ in B850 ring from LH2

ISBN: 978-1-61804-251-4

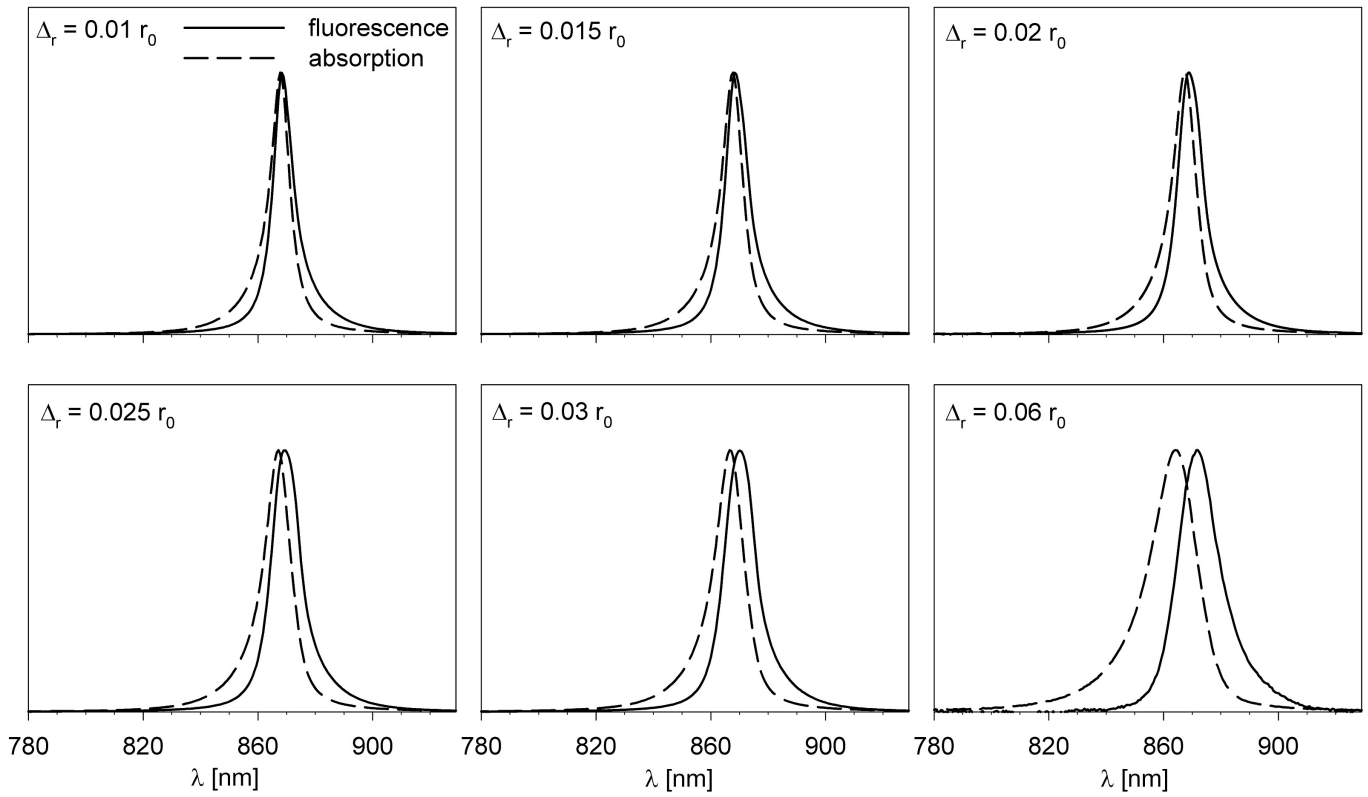


Fig. 3. Calculated $FL(\omega)$ and $OD(\omega)$ spectra averaged over 2000 realizations of static disorder in radial positions of molecules on the ring δr for the nearest neighbour approximation model (low temperature $kT = 0.1\ J_0$, six strengths $\Delta_r = 0.010,\ 0.015,\ 0.020\ , 0.025,\ 0.030,\ 0.060\ r_0$).

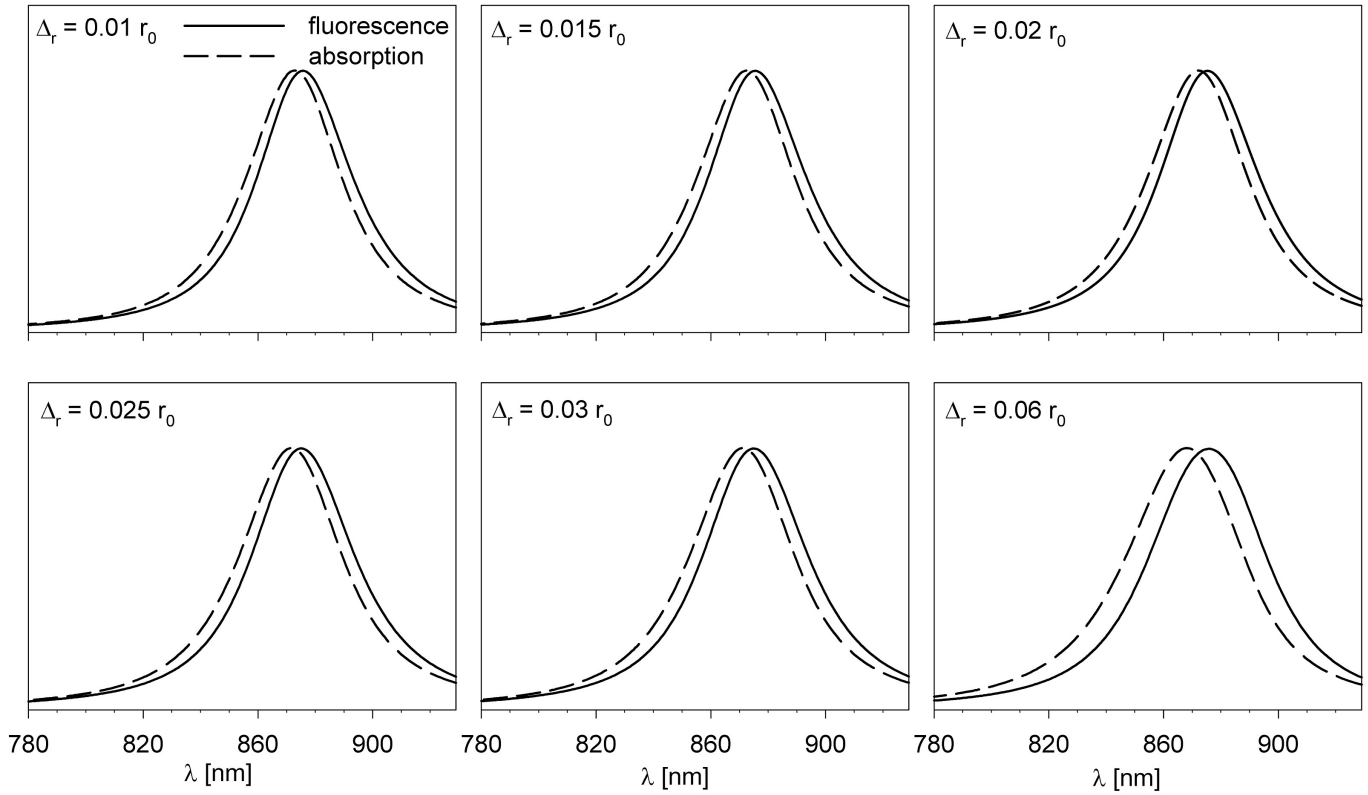


Fig. 4. Calculated $FL(\omega)$ and $OD(\omega)$ spectra averaged over 2000 realizations of static disorder in radial positions of molecules on the ring δr for full Hamiltonian model (room temperature $kT = 0.5\ J_0$, six strengths $\Delta_r = 0.010,\ 0.015,\ 0.020\ , 0.025,\ 0.030,\ 0.060\ r_0$).

ISBN: 978-1-61804-251-4

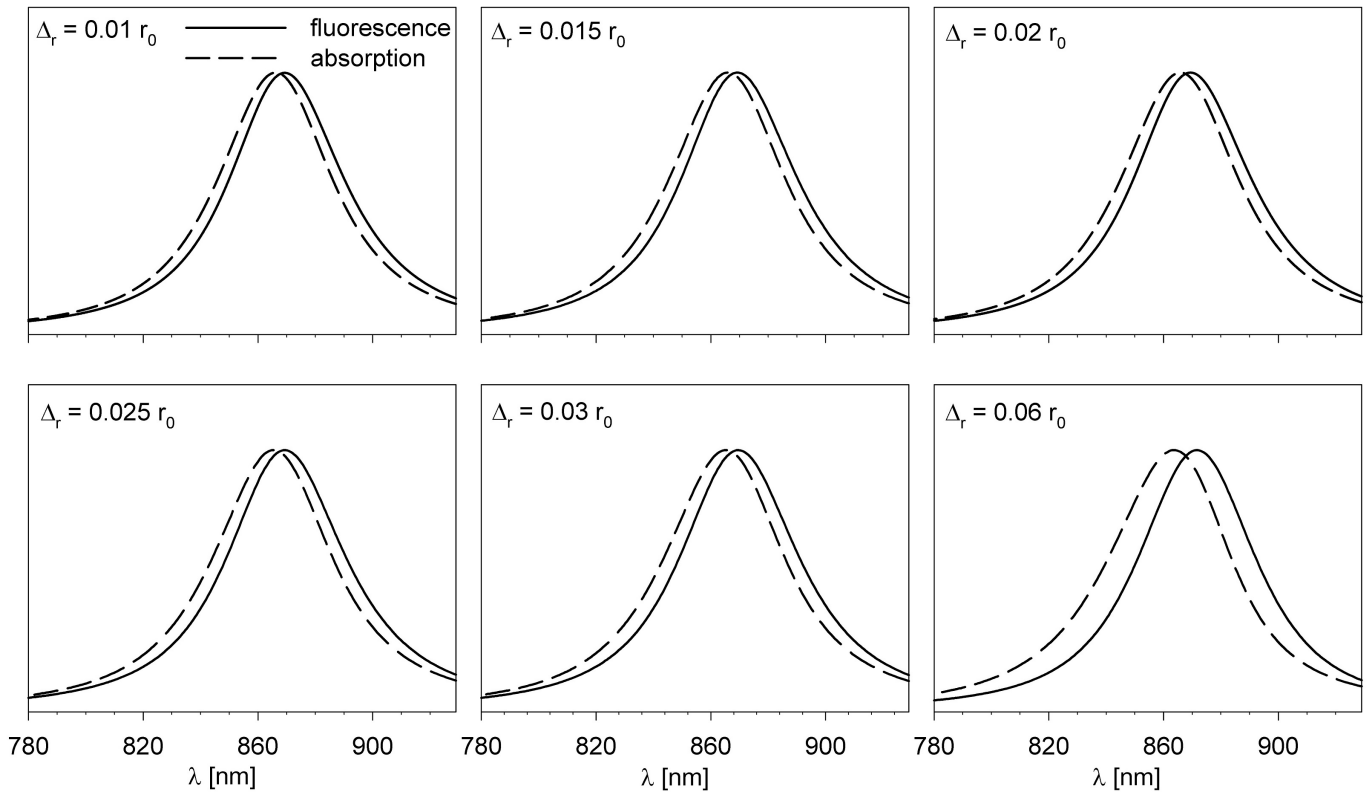


Fig. 5. Calculated $FL(\omega)$ and $OD(\omega)$ spectra averaged over 2000 realizations of static disorder in radial positions of molecules on the ring δr for the nearest neighbour approximation model (room temperature $kT = 0.1\ J_0$, six strengths $\Delta_r = 0.010,\ 0.015,\ 0.020\ , 0.025,\ 0.030,\ 0.060\ r_0$).

complex. have been used. Estimation of J_0 varies in literature between $250\ \mathrm{cm}^{-1}$ and $400\ \mathrm{cm}^{-1}$.

All our simulations of LH2 spectra have been done with the same values of J_0 and unperturbed transition energy from the ground state ΔE_0, that we found for LH2 ring in case of the nearest neighbour approximation model ($J_0 = 400\ \mathrm{cm}^{-1}$, $\Delta E_0 = 12300\ \mathrm{cm}^{-1}$) [22], [23].

Contrary to Novoderezhkin et al. [32], different model of spectral density (the model of Kühn and May [12]) has been used. In agreement with our previous results [16], [17] we have used $j_0 = 0.4\ J_0$ and $\omega_c = 0.212\ J_0$ (see (14)). The strength of uncorrelated static disorder has been taken in agreement with [19]: $\Delta_r \in (0.00, 0.12\, r_0)$.

Resulting absorption $OD(\omega)$ and steady state fluorescence spectra $FL(\omega)$ for LH2 ring averaged over 2000 realizations of static disorder for six strengths $\Delta_r = 0.01,\ 0.015,\ 0.02,\ 0.025,\ 0.03,\ 0.06\, r_0$ at low temperature ($kT = 0.1\ J_0$) can be seen in Figure 2 (full Hamiltonian model) and in Figure 3 (the nearest neighbour approximation model). The same but for room temperature ($kT = 0.5\ J_0$) can be seen in Figure 4 (full Hamiltonian model) and Figure 5 (the nearest neighbour approximation model).

V. Conclusions

Software package *Mathematica* was found by us very useful for the simulations of the molecular ring spectra in case of static disorder in local excitation energies and transfer integrals. But standard numerical integration method from *Mathematica* does not provide satisfactory results in case of the static disorder in radial positions of molecules on the ring and full Hamiltonian model. This problem has been solved by use of our procedure for integration of oscillating and damped function in Fortran.

We compare our new simulated FL and OD spectra for B850 ring from LH2 complex (full Hamiltonian model, static disorder in radial positions of molecules on the ring – Figure 2, Figure 4) from two aspects. At first the comparison with the spectra in case of the nearest neighbour approximation model (the same type of static disorder – Figure 3, Figure 5), then the comparison with our previous results (full Hamiltonian model, static disorder in local excitation enegies [27] and in transfer integrals [39]) is done. Following conclusions can be made.

Differences in spectral lines can be seen especially for low temperature $kT = 0.1\ J_0$ (Figure 2, Figure 3). Absorption spectral lines in case of full Hamiltonian model are wider (in particular on the right hand side of spectral profile) and their maxima are slightly shifted to higher wavelengths in comparison with the nearest neighbour approximation model. For growing strength Δ_r of static disorder, the absorption spectral peak positions move to lower wavelength (both models). On the other hand, any substantial shift is not visible for the fluorescence spectral line in case of the nearest neighbour approximation model. The most essential difference is fluorescence spectral line splitting ($\Delta_r \in (0.01\ r_0, 0.03\ r_0)$) for full Hamiltonian model. It is caused by different energetic band structure (Figure 1) and different distribution of the quantity

ISBN: 978-1-61804-251-4

$P_\alpha d_\alpha^2$ (see (9)). In case of the nearest neighbour approximation model any fluorescence spectral line splitting is not visible.

For room temperature $kT = 0.5\ J_0$, full Hamiltonian model does not give substantially different spectral lines in comparison with the nearest neighbour approximation model. The differences are hidden behind spectral line widening due to dynamic disorder.

As concerns the comparison of the case with static disorder in radial positions of molecules on the ring with other types of static disorder, following conclusions can be done. Fluorescence spectral line splitting is also visible in case of static disorder in local excitation energies [27] and in transfer integrals [39] (also for full Hamiltonian model and low temperature $kT = 0.1\ J_0$). Comparable splitting can be seen for the strengths $\Delta = 0.1\ J_0$ (static disorder in local excitation energies [27]), $\Delta_J = 0.05\ J_0$ (static disorder in transfer integrals [39]) and $\Delta_r = 0.015\ r_0$ (static disorder in radial positions of molecules on the ring).

REFERENCES

[1] R. van Grondelle , V. I. Novoderezhkin, "Energy transfer in photosynthesis: experimental insights and quantitative models," *Phys. Chem. Chem. Phys.*, vol. 8, no. 7, pp. 793–807, 2006.

[2] G. McDermott, S. M. Prince, A. A. Freer, A. M. Hawthornthwaite-Lawless, M. Z. Papiz, R. J. Cogdell, N. W. Isaacs, "Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria," *Nature*, vol. 374, pp. 517–521, 1995.

[3] M. Z. Papiz, S. M. Prince, T. Howard, R. J. Cogdell, N. W. Isaacs, "The structure and thermal motion of the B 800–850 LH2 complex from Rps. acidophila at 2.0 Å resolution and 100 K: new structural features and functionally relevant motions," *J. Mol. Biol.*, vol. 326, no. 5, pp. 1523–1538, 2003.

[4] W. P. F. de Ruijter, et al., "Observation of the Energy-Level Structure of the Low-Light Adapted B800 LH4 Complex by Single-Molecule Spectroscopy," *Biophys. J.*, vol. 87, no. 5, pp. 3413–3420, 2004.

[5] R. Kumble, R. Hochstrasser, "Disorder-induced exciton scattering in the light-harvesting systems of purple bacteria: Influence on the anisotropy of emission and band → band transitions," *J. Chem. Phys.*, vol. 109, no. 2, pp. 855–865, 1998.

[6] V. Nagarajan et al., "Ultrafast exciton relaxation in the B850 antenna complex of Rhodobacter sphaeroides," *Proc. Natl. Acad. Sci. USA*, vol. 93, no. 24, pp. 13774–13779, 1996.

[7] V. Nagarajan et al., "Femtosecond pump-probe spectroscopy of the B850 antenna complex of Rhodobacter sphaeroides at room temperature," *J. Phys. Chem. B*, vol. 103, no. 12, pp. 2297–2309, 1999.

[8] V. Nagarajan, W. W. Parson, "Femtosecond fluorescence depletion anisotropy: Application to the B850 antenna complex of Rhodobacter sphaeroides," *J. Phys. Chem. B*, vol. 104, no. 17, pp. 4010–4013, 2000.

[9] V. Čápek, I. Barvík, P. Heřman, "Towards proper parametrization in the exciton transfer and relaxation problem: dimer," *Chem. Phys.*, vol. 270, no. 1, pp. 141–156, 2001.

[10] P. Heřman, I. Barvík, "Towards proper parametrization in the exciton transfer and relaxation problem. II. Trimer," *Chem. Phys.*, vol. 274, no. 2-3, pp. 199–217, 2001.

[11] P. Heřman, I. Barvík, M. Urbanec, "Energy relaxation and transfer in excitonic trimer," *J. Lumin.*, vol. 108, no. 1-4, pp. 85–89, 2004.

[12] P. Heřman et al., "Exciton scattering in light-harvesting systems of purple bacteria," *J. Lumin.*, vol. 94-95, pp. 447–450, 2001.

[13] P. Heřman, I. Barvík, "Non-Markovian effects in the anisotropy of emission in the ring antenna subunits of purple bacteria photosynthetic systems," *Czech. J. Phys.*, vol. 53, no. 7, pp. 579–605, 2003.

[14] P. Heřman et al., "Influence of static and dynamic disorder on the anisotropy of emission in the ring antenna subunits of purple bacteria photosynthetic systems," *Chem. Phys.*, vol. 275, no. 1-3, pp. 1–13, 2002.

[15] P. Heřman, I. Barvík, "Temperature dependence of the anisotropy of fluorescence in ring molecular systems," *J. Lumin.*, vol. 122–123, pp. 558–561, 2007.

[16] P. Heřman, D. Zapletal, I. Barvík, "Lost of coherence due to disorder in molecular rings," *Phys. Stat. Sol. C*, vol. 6, no. 1, 89–92, 2009.

[17] P. Heřman, I. Barvík, "Coherence effects in ring molecular systems," *Phys. Stat. Sol. C*, vol. 3, no. 10, pp. 3408–3413, 2006.

[18] P. Heřman, D. Zapletal, I. Barvík, "The anisotropy of fluorescence in ring units III: Tangential versus radial dipole arrangement," *J. Lumin.*, vol. 128, no. 5-6, pp. 768–770, 2008.

[19] P. Heřman, I. Barvík, D. Zapletal, "Energetic disorder and exciton states of individual molecular rings," *J. Lumin.*, vol. 119-120, pp. 496–503, 2006.

[20] P. Heřman, I. Barvík, D. Zapletal, "Computer simulation of the anisotropy of fluorescence in ring molecular systems: Tangential vs. radial dipole arrangement," *Lecture Notes in Computer Science*, vol. 5101, pp. 661–670, 2008.

[21] P. Heřman, D. Zapletal, J. Šlégr, "Comparison of emission spectra of single LH2 complex for different types of disorder," *Physics Procedia*, vol. 13, pp. 14–17, 2011.

[22] P. Heřman, D. Zapletal, M. Horák, "Computer simulation of steady state emission and absorption spectra for molecular ring," in *Proc. 5th International Conference on Advanced Engineering Computing and Applications in Sciences (ADVCOMP2011)*, Lisbon: IARIA, 2011, pp. 1–6.

[23] D. Zapletal, P. Heřman, "Simulation of molecular ring emission spectra: localization of exciton states and dynamics, *Int. J. Math. Comp. Sim.*", vol. 6, no. 1, pp. 144–152, 2012.

[24] M. Horák, P. Heřman, D. Zapletal, "Simulation of molecular ring emission spectra - LH4 complex: localization of exciton states and dynamics," *Int. J. Math. Comp. Sim.*, vol. 7, no. 1, pp. 85–93, 2013.

[25] M. Horák, P. Heřman, D. Zapletal, "Modeling of emission spectra for molecular rings - LH2, LH4 complexes," *Phys. Proc.*, vol. 44, pp. 10–18, 2013.

[26] P. Heřman, D. Zapletal, "Intermolecular coupling fluctuation effect on absorption and emission spectra for LH4 ring," *International Journal of Mathematics and Computers in Simulation*, vol. 7, no. 3, 249–257, 2013.

[27] P. Heřman, D. Zapletal, M. Horák, "Emission spectra of LH2 complex: Full Hamiltonian model," *European Physical Journal B*, vol. 86, no. 5, art.no.215, 2013.

[28] P. Heřman, D. Zapletal, "Emission Spectra of LH4 Complex: Full Hamiltonian Model", *International Journal of Mathematics and Computers in Simulation*, vol. 7, no. 6, pp. 448-455, 2013.

[29] P. Heřman, D. Zapletal, "Simulation of Emission Spectra for LH4 Ring: Intermolecular Coupling Fluctuation Effect," *International Journal of Mathematics and Computers in Simulation*, vol. 8, pp. 74–81, 2014.

[30] W. M. Zhang et al., "Exciton-migration and three-pulse femtosecond optical spectroscopies of photosynthetic antenna complexes," *J. Chem. Phys.*, vol. 108, no. 18, pp. 7763–7774, 1998.

[31] S. Mukamel, *Principles of nonlinear optical spectroscopy*. New York: Oxford University Press, 1995.

[32] V. I. Novoderezhkin, D. Rutkauskas, R. van Grondelle, "Dynamics of the emission spectrum of a single LH2 complex: Interplay of slow and fast nuclear motions," *Biophys. J.*, vol. 90, no. 8, pp. 2890–2902, 2006.

[33] A. G. Redfield, "The Theory of Relaxation Processes," *Adv. Magn. Reson.*, vol. 1, pp. 1–32, 1965.

[34] D. Rutkauskas et al., "Fluorescence spectroscopy of conformational changes of single LH2 complexes," *Biophys. J.*, vol. 88, no. 1, pp. 422–435, 2005.

[35] D. Rutkauskas et al., "Fluorescence spectral fluctuations of single LH2 complexes from *Rhodopseudomonas acidophila* strain 10050," *Biochemistry*, vol. 43, no. 15, pp. 4431–4438, 2004.

[36] V. May, O. Kühn, *Charge and Energy Transfer in Molecular Systems*. Berlin: Wiley-WCH, 2000.

[37] O. Zerlauskiene et al., "Static and Dynamic Protein Impact on Electronic Properties of Light-Harvesting Complex LH2," *J. Phys. Chem. B*, vol. 112, no. 49, pp. 15883–15892, 2008.

[38] S. Wolfram, *The Mathematica Book*. 5th ed., Wolfram Media, 2003.

[39] D. Zapletal, P. Heřman, "Photosynthetic Complex LH2 - Absorption and Steady State Fluorescence Spectra," in *Proc. of 6th International Conference on Sustainable Energy and Environmental Protection (SEEP2013)*, Maribor: University of Maribor, 2013, pp. 284–290.

Gradient methods with the exponential relaxation

Igor G. Chernorutskiy

Abstract— For a class of matrix gradient methods a new concept of the relaxation function is suggested. This concept allows to evaluate the effectiveness of each gradient optimization procedure, and to synthesize new methods for special classes of ill conditioned (stiff) non-convex optimization problems. According to the suggested approach, it is possible to build relevant gradient method for any given relaxation function.

The theorem about the relaxation conditions of each matrix gradient method is proven. Based on the concept of the relaxation functions it is given the geometric interpretation of relaxation properties of gradient methods. According to this interpretation it is possible to build a relaxation area, and to evaluate the speed of the objective function values decreasing.

The analysis of classical matrix gradient schemes such as simple gradient method, Newton's methods, Marquardt-Levenberg method is given. It is shown that the relaxation function and its geometric interpretation gives almost full information about the properties and capabilities of relevant gradient optimization methods.

A new class of matrix gradient methods with the exponential relaxation function (ERF) is suggested. It is shown that ERF-method summarizes the classical gradient methods and have the relaxation functions, entirely located in the relaxation area, which significantly increases the computational efficiency.

The ERF-methods convergence for a wide class of non-convex objective functions is established.

Keywords—gradient method, Newton method, Marquardt-Levenberg method, relaxation function, stiff problems, methods with the exponential relaxation.

I. Introduction

The concept of the relaxation function (RF) for a class of matrix gradient methods, including such classical procedures as a simple gradient descent method, Newton method, Marquardt-Levenberg method is introduced in the article. It is shown that RF is a "passport" of each gradient method and it completely determines its local properties like the stability regions of methods of numerical integration of systems of ordinary differential equations. RF is not only a tool of the analysis of a given matrix gradient scheme but also the tool of new gradient methods with necessary properties synthesis. A new class of gradient methods with the exponential relaxation function is built in the paper.

I.G. Chernorutskiy is the Chair of Information and Control Systems Department, St. Petersburg State Polytechnical University, St. Petersburg, Russia (phone: +78122971600; e-mail: igcher1946@mail.ru).

II. Relaxation functions

A. Problem

Suppose we need to solve the problem

$$J(x) \to \min_x,\ x \in R^n,\ J \in C^2(R^n). \qquad (1)$$

Consider a class of matrix gradient methods

$$x^{k+1} = x^k - H_k(A_k, h_k) J'(x^k),\ h_k \in R^1, \qquad (2)$$

where $A_k = J''(x_k)$, H_k, – matrixes function. It is supposed, that in some ζ_k-region $\left\{x \in R^n / \left\|x - x^k\right\| \le \zeta_k\right\}$ of a point x^k the functional $J(x)$ accurately approximated by a quadratic functional

$$f(x) = 1/2\langle A_k x, x\rangle - \langle b_k, x\rangle + c_k, \qquad (3)$$

where A_k – symmetric, may be not positive definite matrix. No significant loss of generality we can assume $b_k = 0, c_k = 0$. Indeed, taking $\det A_k \neq 0, x = x^* + z$, where $x^* = A_k^{-1} b_k$, we will get an introduction

$$f_1(z) = f(x^* + z) = 1/2\langle A_k z, z\rangle + \overline{c}_k. \qquad (4)$$

Here constant $\overline{c}_k = c_k - 1/2\langle A_k x^*, x^*\rangle$ can not be considered as it is not influence to the optimization process.

Formula (2) has the property of invariance relative the shift of coordinates origin.

Being recorded for $f(x)$, it is converted in the same formula for $f_1(z)$. Really for $f(x)$ we have

$$x^{k+1} = x^k - H_k(A_k x^k - b_k). \qquad (5)$$

If $z^k = x^k - x^*$, we have $z^{k+1} = z^k - H_k A_k z^k$. And this is method (2) for functional f_1.

The goal is to build such a matrix functions H_k to satisfy the inequality $f\left(x^{k+1}\right) < f\left(x^k\right)$.

B. Definition

A scalar function

$$R_h(\lambda) = 1 - H(\lambda, h)\lambda; \quad \lambda, h \in R^1$$

is called the *relaxation function* of method (2), and its values $R_h(\lambda_i)$ on the matrix A_k spectrum - *multipliers of relaxation* for the point x_k. (In some cases, index h in the relaxation function expression will be dropped).

Here $H(\lambda, h)$ denotes the scalar function according to the matrix function $H(A_k, h_k)$ in the formula (2). Note that if A is a symmetric matrix and

$$A = Tdiag(\lambda_1, \lambda_2, ..., \lambda_n)T^T,$$

where T is orthogonal matrix with eigenvectors of the matrix A as columns , then

$$F(A) = Tdiag\left[F(\lambda_1), F(\lambda_2), ..., F(\lambda_n)\right]T^T.$$

The matrix function $F(A)$ is defined, if a scalar function $F(\lambda)$ is defined in points $\lambda_1, \lambda_2, ..., \lambda_n$.

III. RELAXATION CONDITIONS

A. Theorem 1

If the matrix A_k is not degenerate we have

$$f\left(x^{k+1}\right) \le f\left(x^k\right), \quad \forall x^k \in R^n, \qquad (6)$$

if and only if

$$\lambda_i < 0 \Rightarrow \left|R(\lambda_i)\right| \ge 1; \quad \lambda_i > 0 \Rightarrow \left|R(\lambda_i)\right| \le 1 \qquad (7)$$

for all matrix A_k eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$.

<u>*Proof*</u>. Let $\{u_i\}$ – orthonormal basis of matrix A_k eigenvectors. Then

$$x^k = \sum_{i=1}^{n} \xi_{i,k} u_i; \; x^{k+1} = x^k - H_k(A_k, h_k) f'\left(x^k\right) =$$

$$= \left(E - H_k A_k\right) x^k = \sum_{i=1}^{n} \xi_{i,k} \left(1 - H_k(\lambda_i, h_k)\lambda_i\right) u_i =$$

$$= \sum_{i=1}^{n} \xi_{i,k} R(\lambda_i) u_i.$$

From the comparison of expressions

$$f\left(x^k\right) = 1/2 \sum_{i=1}^{n} \xi_{i,k}^2 \lambda_i;$$

$$f\left(x^{k+1}\right) = 1/2 \sum_{i=1}^{n} \xi_{i,k+1}^2 \lambda_i = 1/2 \sum_{i=1}^{n} \xi_{i,k}^2 \lambda_i R^2(\lambda_i) \qquad (8)$$

we have that according to (7) every term of the sum in the representation $f\left(x^k\right)$ does not increase. Direct statement is proved. Further, suppose that there is an index $i = i_0$, for which

$$\lambda_{i_0} < 0, \left|R\left(\lambda_{i_0}\right)\right| < 1.$$

Let $x_k = u_{i_0}$.

Then

$$f\left(x^k\right) = 0,5\lambda_{i_0} < f\left(x^{k+1}\right) = 0,5\lambda_{i_0} R^2,$$

which contradicts the condition of relaxation (6). Similarly the second inequality (7) can be considered. The theorem is proved.

B. Comments

1. For strict implementation of the inequality (6) it is necessary and sufficient, in addition to fulfilling the conditions (7) require that we have such an index $i = i_0$ for which $\xi_{i,k} \ne 0$, and the corresponding inequality (7) is strict.

2) Formula (8) helps us to assess the speed of the functional f decreasing, depending on the level of the inequality (7) implementation.

Indeed, denote by the λ_i^+, λ_i^- the positive and negative eigenvalues of the matrix A_k. The same indices give to an appropriate own vectors. The summation on the relevant *i* denote as $\sum^+, \sum^-$. Then

$$2\left|f\left(x^k\right) - f\left(x^{k+1}\right)\right| = \Sigma^+ \xi_{i,k}^2 \lambda_i^+ \left(1 - R^2\left(\lambda_i^+\right)\right) +$$

$$+ \Sigma^- \xi_{i,k}^2 \left|\lambda_i^-\right| \left(R^2\left(\lambda_i^-\right) - 1\right).$$

Therefore the greatest suppression we will have for the components for which the multiplier relaxation values significantly differs from 1.

Next we will mainly consider functions $R_h(\lambda)$, for which we have

$$R_h(\lambda) \to 1, h \to 0. \qquad (9)$$

Then, from

$$\left\|x^{k+1}-x^k\right\|=\sum_{i=1}^{n}\xi_{i,k}^2\left[R_{h_k}\left(\lambda_i\right)-1\right]^2 \quad (10)$$

we will have that for $\forall\zeta_k\in R^1$, we can choose such h_k that $\left\|x^{k+1}-x^k\right\|\le\zeta_k$. Thus, by h_k settings we can adjust the vector advance norm in the search space to prevent output from the field of local quadratic model (3) justice.

Sometimes, for normal (10) regulation the damping parameter h can be entered into the scheme of optimization as a multiplier on the right part of (2):

$$x^{k+1}\left(h\right)=x^k-hH_kJ'\left(x^k\right),h\in\left[0,1\right]. \quad (11)$$

Then,

$$\left\|x^{k+1}\left(h\right)-x^k\right\|=h\left\|x^{k+1}\left(1\right)-x^k\right\|$$

and for the second equality in (8) we will have

$$f\left(x^{k+1}\right)=1/2\sum_{i=1}^{n}\xi_{i,k}^2\lambda_i\overline{R}^2\left(\lambda_i\right),$$
$$\overline{R}\left(\lambda_i\right)=\left(1-h\right)+hR\left(\lambda_i\right).$$

Thus, new relaxation multipliers $\overline{R}\left(\lambda_i\right)$ accept intermediate values between 1 and $R\left(\lambda_i\right)$, as required to ensure the relaxation properties according to (7).

So, the concept of the relaxation function allows us to evaluate local properties of various gradient search schemas. The convenience of this approach is also the ability to use visual geometric representations.

For any method (2), we can construct a relaxation function that characterizes the area of its relaxation in the set of eigenvalues. The relaxation area is presented in Fig. 1. It is shaded prohibited area, where the conditions of relaxation (7) are not met. The case of Hessian A_k degenerate considered separately for every relaxation function. In these cases we need for example to consider the uncertainty like $(\infty\times 0)$ in the expression $H\left(\lambda,h\right)\lambda$.

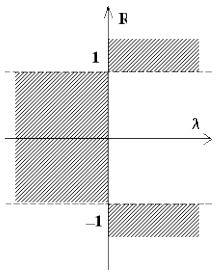


Fig. 1. Gradient methods relaxation area

A very important property of relaxation functions is the ability to use the appropriate views for the synthesis of new procedures of class (2) with some desirable properties for specific classes of optimization problems.

It is well known that an important class of finite-dimensional optimization problems is the class of *ill-conditioned* (*stiff*) problems. We have for local stiffness characteristic (index) in the point x [1]:

$$\eta\left(x\right)=\lambda_1\left(x\right)/\left|\min_i\lambda_i\left(x\right)\right|.$$

Here $\lambda_i\left(x\right)$- are eigenvalues of matrix $J''\left(x\right)$ that satisfy the following inequalities

$$\lambda_1\ge\ldots\ge\lambda_{n-r}\gg\left|\lambda_{n-r+1}\right|\ge\ldots\ge\left|\lambda_n\right|..$$

If we have such inequalities in some region of the point x then in this area we will have the well-known computational difficulties of the functional $J\left(x\right)$ minimizing. It is easy to see that the case of stiff matrix (matrix with high stiffness index) $J''\left(x\right)$ is a special case of matrix with large spectral number of conditionality. So we have here the particular case of ill-conditioned problems. If high stiffness index for goal functional $J\left(x\right)$ is stored in a certain area of search space, the functional level surface acquires the characteristic "valley" or “ravine” structure. It is well-known situation for specialists in computer optimization. Usually slow convergence of optimization methods occurs in the area called "bottom of the ravine" which approximated by subspace pulled down over eigenvectors corresponds to "small" eigenvalues $\lambda_i\left(x\right)$.

Using concepts RF we can quite clearly (geometrically) evaluate the computational capabilities of every gradient scheme in the case of the stiff functionals minimizing.

IV. Classical gradient methods

Consider some well known methods (2) and their relaxation functions.

A. Simple gradient descent method

Simple gradient descent method (SGD) looks as follows:

$$x^{k+1}=x^k-hJ'\left(x^k\right),h=const. \quad (12)$$

Its relaxation function

$$R\left(\lambda\right)=1-h\lambda \quad (13)$$

is linear and presented in Fig. 2.

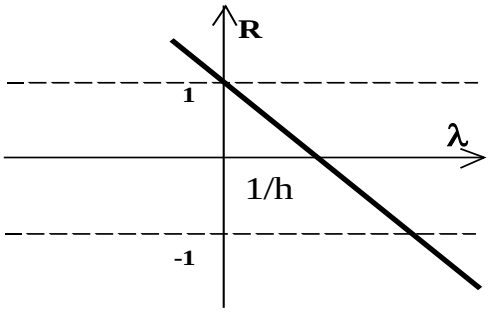


Fig. 2. Simple gradient descent relaxation function

Let matrix A_k eigenvalues are located in a closed interval

$$[m, M], 0 < m \ll M, \eta(x) = M / m \gg 1.$$

In this case, the condition (9) is true, and the inequality (7) is reduced to the requirement $|R(\lambda_i)| \le 1, i = 1, ..., n$ or

$$|1 - h\lambda_i| \le 1, i = 1, ..., n. \qquad (14)$$

From inequality (14) we have:

$$h \le 2 / M, R(m) = 1 - hm \cong 1.$$

So we have that the terms in (8) for small eigenvalues of the region $\lambda = 0$ are hardly will decrease, and progress will be very slow. This determines the low efficiency of the method (12). This fact is well known from linear algebra. There are different strategies of h choice, but if we have large η all of these methods, including the method of quickest descent

$$x^{k+1} = \arg\min_{h \ge 0} J\left[x^k - hJ'(x^k)\right],$$

are ineffective, even when the minimization of convex functionals.

B. Newton method

Newton method is based on the goal functional quadratic approximation using:

$$J(x) \cong f(x) = J(x^k) + \langle x - x^k, J'(x^k) \rangle + \\ + 1/2 \langle J''(x^k)(x - x^k), x - x^k \rangle$$

To find the next point x^{k+1} it is used the necessary condition of extremum: $f'(x) = 0$.

So we come to the well known formula

$$x^{k+1} = x^k - h_k \left[J''(x^k)\right]^{-1} J'(x^k). \qquad (15)$$

Added parameter h_k, as it was mentioned earlier, allows to change the norm of the promotion vector $\|x^{k+1} - x^k\|$. As a result we have the well known damped Newton method.

The damped Newton method relaxation function looks as follows:

$$R(\lambda) = 1 - H(\lambda, h_k)\lambda = 1 - h_k..$$

We supposed $J''(x)$ is not degenerate.

The main defect of Newton method is well known [1, 2]. The method does not work in non-convex situations. It is easy to see that the RF is adjudged to be in an area where we have not met inequalities (7) when λ is greater than zero, or when λ is less than zero. We have the classic version of Newton method if $h_k = 1$. Usually $h_k \in [0,1]$, and for $\lambda < 0$ RF is situated in the prohibited area.

Quasi-Newton algorithms and methods of conjugates lines have the similar lack. These methods efficiency approaches to the SGD method efficiency, if the functional $J(x)$ is not convex [1].

Traditional objection of Newton methods connected with the necessity of the second derivative calculate is less significant for real optimization problems.

C. Levenberg-Marquardt method (LM-method).

If we know that the matrix $J''(x^k)$ eigenvalues are located in the interval $[-m, M], M > m > 0$, then we can build method with nonlinear RF (Fig. 3):

$$R(\lambda) = h / (h + \lambda), \qquad (16)$$

meeting the requirements of (7) if : $\forall \lambda \in [-m, M], h > m.$

For stiff problems we have $M \gg m$.

The appropriate method was proposed by Levenberg [3] for

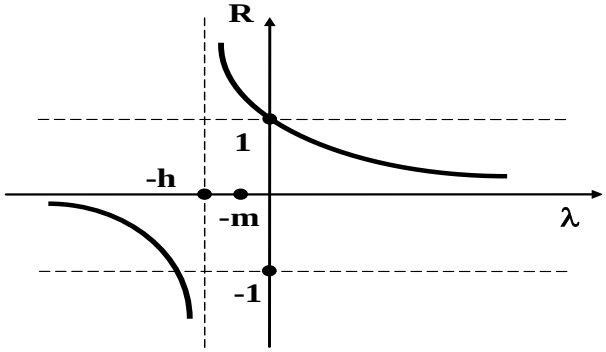


Fig. 3. Levenberg-Marquardt method relaxation function

other reasons and has the function

$$H(\lambda,h)=\left[1-R(\lambda)\right]/\lambda=(h+\lambda)^{-1}.$$

Method (2) scheme with the specified function H looks as follows:

$$x^{k+1}=x^k-\left[hE+J''(x^k)\right]^{-1}J'(x^k). \quad (17)$$

Here E – unit matrix. A scalar h to each step of the iterative process is chosen so that the matrix $hE+J''(x^k)$ was positively identified and that $\left\|x^{k+1}-x^k\right\|\le c_k$, where c_k can change from iteration to iteration.

From the last expression we can see that we have a regularized form of Newton method with the regularization parameter h. This form of Newton method applied by Levenberg for least squares method problems solving. Later this method was used by Marquardt to solve common problems of nonlinear optimization [4].

Method (17) implementation reduces to the linear algebraic system solution (on each step):

$$\left[hE+J''(x^k)\right]\Delta x^k=-J'(x^k),\ \Delta x^k=x^{k+1}-x^k. \quad (18)$$

In Fig. 3 it is easy to see the main defect of LM-method. Indeed, we need to have $h>m$. But the m value, as a rule, is unknown and cannot be computed with reasonable accuracy. The difficulties increase with dimension n increasing. The best that can normally be done in practice, is to take

$$h\ge\max\left\{\varepsilon_M n\left\|J''(x^k)\right\|,\left|\min\lambda_i\left(J''(x^k)\right)\right|\right\}. \quad (19)$$

Here ε_M – computer epsilon.

If the condition $h>m$ is false the system (18) can be degenerate. In addition, to the left of the point $\lambda=-h$ RF quickly enters the prohibited area and this method diverges or converges to a saddle point. Attempts to use algorithmic method for more accurate of h localization cause multiple solutions of ill-conditioned linear system (18) with different test h values. Also it is easy to see that too big h value involve slow convergence. Multiplier relaxation values will be near 1.

These difficulties increase when the approximations of derivatives by end differences are using. Indeed, for small values of vector $J'(x^k)$ components for the points x^k, located on the bottom of the ravine, we come to the necessity to obtain these components of the gradient vector as small differences of large quantities (of the order $J(x^k)$). As the result, the vector Δx^k components will have large relative errors of the order

$$\eta(x^k)\left|\varepsilon_M J(x^k)\right|/\left\|J'(x^k)\right\|.$$

Stiffness characteristic (index) η in this case plays the role of the error amplifier. For Newton method we have similar remark. At the same time for the SGD the $J''(x^k)$ calculation accuracy can be sufficient to find the correct direction of $J(x^k)$ decreasing.

Despite the noted defects, the method (18) is often quite effective, and its presence in the optimization methods library must be considered quite desirable.

V. Gradient Methods with Exponential Relaxation Function

Below with RF conception using we will suggest non-traditional gradient methods with *exponential relaxation* function. It is shown also that the methods generalize many of the already known gradient procedures.

In accordance with the main requirements for the relaxation functions it is natural to consider the exponential dependence of the form

$$R_h(\lambda)=R(\lambda)=\exp(-\lambda h),\ h>0, \quad (20)$$

for which the relaxation conditions are fulfilled for every values of the parameter h. In addition, we have the following relation:

$$R_h(\lambda)\to 1,\ h\to 0,$$

that allows to control the vector promotion norm $\left\|x^{k+1}-x^k\right\|$ for every matrix $J''(x^k)$ eigenvalues locations.

It is easy to see that the function (20) summarizes (in fact - gives) many famous methods relaxation functions and is in a sense the canonical (or optimal). Indeed, decomposing function (20) in a Taylor series and limited to the first two members we will have:

$$\exp(-\lambda h)=1/\exp(\lambda h)\cong 1/(1+\lambda h)=$$
$$=h'/(h'+1),\ h'=1/h,$$

that corresponds to the relaxation function of LM- method. And similarly, we have

$$\exp(-\lambda h)\cong 1-\lambda h.$$

That is SGD-method. For sufficiently large values of the parameter h we will have $\exp(-\lambda h)\cong 0$, that allows to speak about the degeneration of the method to the classical Newton method.

According to the exponential relaxation function matrix multiplier H_k from (2) can be built. Indeed, we have:

$$\lambda H(\lambda,h)=1-R(\lambda)=1-\exp(-\lambda h).$$

And for $\lambda \neq 0$

$$H(\lambda,h)=\lambda^{-1}\left[1-\exp(-\lambda h)\right]=\int_0^h \exp(-\lambda\tau)d\tau. \quad (21)$$

Assume $H(0,h)=h$ according to the function continuity requirement.

As the result the scheme of the method with the exponential relaxation (ER-method) will have the form

$$x^{k+1}=x^k-H(A_k,h_k)J'(x^k), \quad (22)$$

$$H(A,h)=\int_0^h \exp(-A\tau)d\tau, \quad (23)$$

$$h_k \in Arg\min_{h\geq 0} J\left[x^k-H(A_k,h)J'(x^k)\right]. \quad (24)$$

Additional ways of h_k choice may be used.

Principle scheme of ER-method was obtained from the analysis of the local quadratic model of the goal functional. It is interesting to clarify the possibilities of the method in a global sense, without taking the assumptions of $J(x)$ quadratic structure.

It is possible to prove that the algorithm (22), (23) converges almost under the same restrictions on the minimized functional, that the method of steepest gradient descent, having in certain conditions a significantly higher rate of convergence.

The following theorem establishes the fact of ER-method convergence for wide class of non convex functionals on the assumption of minimum point achievement (condition 2) and the points of local minima absence (condition 3).

Theorem 2

Let

1. $J(x)\in C^2(R^n)$;

2. $X_*=\left\{x^*/J(x^*)=\min J(x)\right\}\neq\varnothing$;

3. for every $\varepsilon>0$ we have $\delta>0$, that $\|J'(x)\|\geq\delta$, if $x\notin S(X_*)$, where

$$S(X_*)=\left\{x/d(x,X_*)\leq\varepsilon\right\},\ d(x,X_*)=\min_{x^*\in X_*}\|x-x^*\|;$$

4. for every $x,\ y\in R^n$

$$\|J'(x+y)-J'(x)\|\leq l\|y\|,\quad l>0;$$

5. matrix' $J''(x)$ eigenvalues are contained in the interval $[-M,M]$, where $M>0$ does not depend of x.

Then, for every initial point x^0 for the sequence $\{x^k\}$, built according to (3), (4), we will have:

$$\lim_{k\to\infty} d(x^k,X_*)=0, \quad (25)$$

$$\lim_{k\to\infty} J(x^k)=J(x^*). \quad (26)$$

The proof can be found in [1].

A. Note 1

The theorem is obviously right, if h_k will be chosen not from condition (5), but from the condition

$$J\left[x^k-H(A_k,h_k)J'(x^k)\right]=$$
$$=\min_{h\in[0,\bar{h}]} J\left[x^k-H(A_k,h)J'(x^k)\right],$$

where $\bar{h}>0$ is any number.

B. Note 2

Statements (25), (26) are right when replacing the condition (24) to the following:

$$J_{k+1}=J\left[x^k-H(A_k,h_k)J'(x^k)\right]\leq(1-\gamma_k)J'(x^k)+$$
$$+\gamma_k\min_{h\geq 0} J\left[x^k-H(A_k,h)J'(x^k)\right],\quad 0<\gamma<\gamma_k\leq 1.$$

In case of strong convexity of the functional $J(x)$, it is possible to obtain an estimate of the convergence rate.

C. Theorem 3

Let:

1. $J(x)\in C^2(R^n)$;

2. for any $x, y\in R^n$ we have

$$\lambda\|y\|^2\leq\langle J''(x)y,y\rangle\leq\Lambda\|y\|^2,$$
$$\|J''(x+y)-J''(y)\|\leq L\|x\|,\quad \Lambda>\lambda>0,\ L\geq 0.$$

Then, for every initial point x^0 for the method (3) we have (6), (7), and the rate of convergence

$$\|x^{k+1}-x^k\|\leq(\Lambda/\lambda)^{1/2}L\|x^k-x^*\|/(2\lambda).$$

The proof can be found in [1].

Thus, we have the quadratic speed of convergence, typical for Newton methods.

The problem of ER-methods algorithmic implementation is developed. It is a separate problem and we do not consider it here.

VI. CONCLUSIONS

1. The functions relaxation conception allows to produce a complete analysis of any matrix gradient scheme for finite-dimensional optimization. Analytical link between RF and matrix multiplier in the gradient method scheme allows for a given RF build the appropriate method of optimization with the desired relaxation properties. Essentially, we have a "generator" of new gradient methods.

2. On the basis of the relaxation functions conception it is considered a new class of matrix gradient methods with the exponent relaxation, generalizing the classical gradient methods, such as Newton methods and Levenberg-Marquardt method. Unlike classical prototypes built methods remain convergence for non convex nonlinear programming problems under conditions of high stiffness index of goal functionals.

REFERENCES

[1] Chernorutskiy, I.G.: Methods of optimization. Computer technologies. BXV, S. Petersburg, 384 p., 2011 (in Russian)

[2] Nesterov, Yu, Polyak B.: Cubic regularization of Newton method and its global performance. Math. Program., Ser. A 108. 177-205 (2006)

[3] Levenberg, K.: A method for the solution of sertain problems in least squares. Quart. Appl. Math. 2, 164-168 (1944)

[4] Marquardt, D.: An algorithm for least squares estimation of nonlinear parameters. SIAM J. Appl. Math. 11, 431-441 (1963)

I.G. Chernorutskiy currently is a Professor of Saint-Petersburg State Polytechnical University (SPbSPU). He earned all degrees at SPbSPU: Professor, 1990; Doctor of Technical Science, 1987; Associate Professor, 1982; Ph.D., 1978; M.S., 1970. Professor I.G.Chernorutskiy is the Chair of Information & Control Systems Division of Computer Science and Engineering School (CSES). His Research Interests include Applied Software Engineering, Optimization Tools, Real-time Systems Modeling and Simulation, Parameter Estimation and Adaptive Optimization, Decision Support Systems, Artificial Intelligence and Expert Systems.

ISBN: 978-1-61804-251-4

Dynamic response of a doubly curved shallow shell rectangular in plan impacted by a sphere

Yury A. Rossikhin, Marina V. Shitikova, and Muhammed Salih Khalid J. M.

***Abstract*—**Large amplitude (geometrically non-linear) vibrations of doubly curved shallow shells with rectangular base under the low-velocity impact by an elastic sphere are investigated. It is assumed that the shell is simply supported and partial differential equations are obtained in terms of shell's transverse displacement and Airy's stress function. The local bearing of the shell and impactor's materials is neglected with respect to the shell deflection in the contact region. The equations of motion are reduced to a set of infinite nonlinear ordinary differential equations of the second order in time and with cubic and quadratic nonlinearities in terms of the generalized displacements. Assuming that only two natural modes of vibrations dominate during the process of impact and applying the method of multiple time scales, the set of equations is obtained, which allows one to find the time dependence of the contact force and to determine the contact duration and the maximal contact force.

***Keywords*—**Doubly curved shallow shell rectangular in base, impact interaction, method of multiple time scales

I. Introduction

Doubly curved panels are widely used in aeronautics, aerospace and civil engineering and are subjected to dynamic loads that can cause vibration amplitude of the order of the shell thickness, giving rise to significant non-linear phenomena [1]–[4].

A review of the literature devoted to dynamic behaviour of curved panels and shells could be found in Amabili and Paidoussis [5], as well as in [3], wherein it has been emphasized that free vibrations of doubly curved shallow shells were studied in the majority of papers either utilizing a slightly modified version of the Donnell's theory taking into account the double curvature [1, 6] or the nonlinear first-order theory of shells [7, 8].

Large-amplitude vibrations of doubly curved shallow shells with rectangular base, simply supported at the four edges and subjected to harmonic excitation were investigated in [3], while chaotic vibrations were analyzed in [4]. It has been revealed that such an important nonlinear phenomenon as the occurrence of internal resonances in the problems considered in [3] and [4] is of fundamental importance in the study of curved shells.

In spite of the fact that the impact theory is substantially developed, there is a limited number of papers devoted to the problem of impact over geometrically nonlinear shells.

The nonlinear impact response of laminated composite cylindrical and doubly curved shells was analyzed using a modified Hertzian contact law in [9] via a finite element model, which was developed based on Sander's shell theory involving shear deformation effects and nonlinearity due to large deflection. A nine-node isoparametric quadrilateral element was used to model the curved shell. The nonlinear time dependent equations were solved using an iterative scheme and Newmark's method. Numerical results for the contact force and center deflection histories were presented for various impactor conditions, shell geometry and boundary conditions.

Later large deflection dynamic responses of laminated composite cylindrical shells under impact have been analyzed in [10] by the geometrically nonlinear finite element method based on a generalized Sander's shell theory with the first order shear deformation and the von Karman large deflection assumption.

Nonlinear dynamic response for shallow spherical moderate thick shells with damage under low velocity impact has been studied in [11] by using the orthogonal collocation point method and the Newmark method to discrete the unknown variable function in space and in time domain, respectively, and the whole problem is solved by the iterative method. Further this approach was generalized for investigating dynamic response of elasto-plastic laminated composite shallow spherical shell under low velocity impact [12] and nonlinear dynamic response for functionally graded shallow spherical shell under low velocity impact in thermal environment [13].

The nonlinear transient response of laminated composite shell panels subjected to low velocity impact in hygrothermal environments was investigated in [14] using finite element method considering doubly curved thick shells involving large deformations with Green-Lagrange strains. The analysis was carried out using quadratic eight-node isoparametric element.

This work was supported in part by the Ministry of Education and Science of the Russian Federation under Grant No. 2014/19.

Yu. A. Rossikhin is a Head of the Research Center on Dynamics of Solids and Structures, Voronezh State University of Architecture and Civil Engineering, Voronezh 394006, RUSSIA (phone: +7-4732-714220; fax: +7-4732-773992; e-mail: YAR@ vgasu.vrn.ru).

M. V. Shitikova is with the Research Center on Dynamics of Solids and Structures, Voronezh State University of Architecture and Civil Engineering, Voronezh 394006, RUSSIA (e-mail: MVS@vgasu.vrn.ru).

Muhammed Salih Khalid J.M. is a PhD student at the Research Center on Dynamics of Solids and Structures, Voronezh State University of Architecture and Civil Engineering, RUSSIA on leave from the Ministry of Higher Education of Iraq, IRAQ (e-mail: Khalid_bus@yahoo.com).

A modified Hertzian contact law was incorporated into the finite element program to evaluate the impact force. The nonlinear equation was solved using the Newmark average acceleration method in conjunction with an incremental modified Newton-Raphson scheme. A parametric study was carried out to investigate the effects of the curvature and side to thickness ratios of simply supported composite cylindrical and spherical shell panels.

The impact behaviour and the impact-induced damage in laminated composite cylindrical shell subjected to transverse impact by a foreign object were studied in [15] using three-dimensional non-linear transient dynamic finite element formulation. Non-linear system of equations resulting from non-linear strain displacement relation and non-linear contact loading was solved using the Newton-Raphson incremental-iterative method. Some example problems of graphite/epoxy cylindrical shell panels were considered with variation of impactor and laminate parameters and influence of geometrical non-linear effect on the impact response and the resulting damage was investigated.

In the present paper, a new approach is proposed for the analysis of the impact interactions of nonlinear doubly curved shallow shells with rectangular base under the low-velocity impact by an elastic sphere. It is assumed that the shell is simply supported and partial differential equations are obtained in terms of shell's transverse displacement and Airy's stress function. The local bearing of the shell and impactor's materials is neglected with respect to the shell deflection in the contact region. The equations of motion are reduced to a set of infinite nonlinear ordinary differential equations of the second order in time and with cubic and quadratic nonlinearities in terms of the generalized displacements.

Assuming that only two natural modes of vibrations dominate during the process of impact and applying the method of multiple time scales [16], the set of dynamic equations is obtained, which allows one to find the time dependence of the contact force and to determine the contact duration and the maximal contact force.

II. Problem Formulation and Governing Equations

Assume that an elastic or rigid sphere of mass M moves along the z-axis towards a thin walled doubly curved shell with thickness h, curvilinear lengths a and b, principle curvatures k_x and k_y and rectangular base, as shown in Fig. 1. Impact occurs at the moment t=0 with the velocity εV_0 (ε is a small value) at the point N with Cartesian coordinates x_0, y_0.

According to Donnell's nonlinear shallow shell theory, the equations of motion could be obtained in terms of lateral deflection w and Airy's stress function φ [17]

$$\frac{D}{h}\left(\frac{\partial^4 w}{\partial x^4}+2\frac{\partial^4 w}{\partial x^2\partial y^2}+\frac{\partial^4 w}{\partial y^4}\right)=\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 \phi}{\partial y^2}+\frac{\partial^2 w}{\partial y^2}\frac{\partial^2 \phi}{\partial x^2}$$
$$-2\frac{\partial^2 w}{\partial x\partial y}\frac{\partial^2 \phi}{\partial x\partial y}+k_y\frac{\partial^2 \phi}{\partial x^2}+k_x\frac{\partial^2 \phi}{\partial y^2}+\frac{F}{h}-\rho\ddot{w}, \tag{1}$$

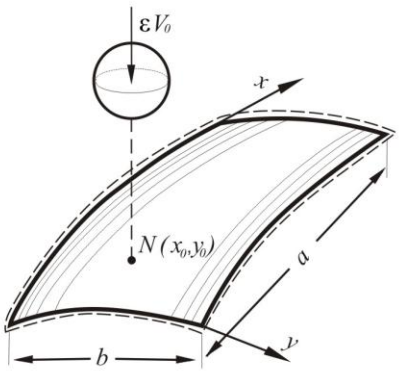


Fig. 1 Geometry of the doubly curved shallow shell

$$\frac{1}{E}\left(\frac{\partial^4 \phi}{\partial x^4}+2\frac{\partial^4 \phi}{\partial x^2\partial y^2}+\frac{\partial^4 \phi}{\partial y^4}\right)=-\frac{\partial^2 w}{\partial x^2}\frac{\partial^2 w}{\partial y^2}+\left(\frac{\partial^2 w}{\partial x\partial y}\right)^2$$
$$-k_y\frac{\partial^2 w}{\partial x^2}-k_x\frac{\partial^2 w}{\partial y^2}, \tag{2}$$

where $D=\dfrac{Eh^3}{12(1-\nu^2)}$ is the cylindrical rigidity, ρ is the density, E and ν are the elastic modulus and Poisson's ratio, respectively, t is time, $F=P(t)\delta(x-x_0)\delta(y-y_0)$ is the contact force, $P(t)$ is yet unknown function, δ is the Dirac delta function, x and y are Cartesian coordinates, overdots denote time-derivatives, $\phi(x,y)$ is the stress function which is the potential of the in-plane force resultants

$$N_x=h\frac{\partial^2 \phi}{\partial y^2},\quad N_y=h\frac{\partial^2 \phi}{\partial x^2},\quad N_{xy}=-h\frac{\partial^2 \phi}{\partial x\partial y}. \tag{3}$$

The equation of motion of the sphere is written as

$$M\ddot{z}=-P(t) \tag{4}$$

subjected to the initial conditions

$$z(0)=0,\quad \dot{z}(0)=\varepsilon V_0, \tag{5}$$

where $z(t)$ is the displacement of the sphere, in so doing

$$z(t)=w(x_0,y_0,t). \tag{6}$$

Considering a simply supported shell with movable edges, the following conditions should be imposed at each edge:

$$w=0,\quad \int_0^b N_{xy}dy=0,\quad N_x=0,\quad M_x=0,\quad \text{at } x=0,a, \tag{7}$$

$$w=0,\quad \int_0^a N_{xy}dx=0,\quad N_y=0,\quad M_y=0,\quad \text{at } y=0,b, \tag{8}$$

where M_x and M_y are the moment resultants.

The suitable trial function that satisfies the geometric boundary conditions is

$$w(x,y,t)=\sum_{p=1}^{\tilde{p}}\sum_{q=1}^{\tilde{q}}\xi_{pq}(t)\sin\left(\frac{p\pi x}{a}\right)\sin\left(\frac{q\pi y}{b}\right), \qquad (9)$$

where p and q are the number of half-waves in x and y directions, respectively, and $\xi_{pq}(t)$ are the generalized coordinates. Moreover, $\tilde{p}$ and $\tilde{q}$ are integers indicating the number of terms in the expansion.

Substituting (9) in (6) and using (4), we obtain

$$P(t)=-M\sum_{p=1}^{\tilde{p}}\sum_{q=1}^{\tilde{q}}\ddot{\xi}_{pq}(t)\sin\left(\frac{p\pi x_0}{a}\right)\sin\left(\frac{q\pi y_0}{b}\right). \qquad (10)$$

In order to find the solution of the set of equations (1) and (2), it is necessary first to obtain the solution of (2). For this purpose, let us substitute (9) in the right-hand side of (2) and seek the solution of the equation obtained in the form

$$\phi(x,y,t)=\sum_{m=1}^{\tilde{m}}\sum_{n=1}^{\tilde{n}}A_{mn}(t)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right), \qquad (11)$$

where $A_{mn}(t)$ are yet unknown functions.

Substituting (9) and (11) in (2) and using the orthogonality conditions of sines within the segments $0\le x\le a$ and $0\le y\le b$, we have

$$\begin{aligned}A_{mn}(t)&=\frac{E}{\pi^2}K_{mn}\xi_{mn}(t)\\&+\frac{4E}{a^3b^3}\left(\frac{m^2}{a^2}+\frac{n^2}{b^2}\right)^{-2}\sum_k\sum_l\sum_p\sum_q B_{pqklmn}\xi_{pq}(t)\xi_{kl}(t),\end{aligned} \qquad (12)$$

where

$$\begin{aligned}B_{pqklmn}&=pqklB^{(2)}_{pqklmn}-p^2l^2B^{(1)}_{pqklmn},\\B^{(1)}_{pqklmn}&=\int_0^a\int_0^b\sin\left(\frac{p\pi x}{a}\right)\sin\left(\frac{q\pi y}{b}\right)\sin\left(\frac{k\pi x}{a}\right)\\&\times\sin\left(\frac{l\pi y}{b}\right)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right)dxdy,\\B^{(2)}_{pqklmn}&=\int_0^a\int_0^b\cos\left(\frac{p\pi x}{a}\right)\cos\left(\frac{q\pi y}{b}\right)\cos\left(\frac{k\pi x}{a}\right)\\&\times\cos\left(\frac{l\pi y}{b}\right)\sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right)dxdy,\\K_{mn}&=\left(k_y\frac{m^2}{a^2}+k_x\frac{n^2}{b^2}\right)^2\left(\frac{m^2}{a^2}+\frac{n^2}{b^2}\right)^{-2}.\end{aligned} \qquad (13)$$

Substituting then (9)-(12) in (1) and using the orthogonality condition of sines within the segments $0\le x\le a$ and $0\le y\le b$, we obtain an infinite set of coupled nonlinear ordinary differential equations of the second order in time for defining the generalized coordinates

$$\begin{aligned}&\ddot{\xi}_{mn}(t)+\Omega^2_{mn}\xi_{mn}(t)\\&+\frac{8\pi^2E}{a^3b^3\rho}\sum_p\sum_q\sum_k\sum_l B_{pqklmn}\left(K_{kl}-\frac{1}{2}K_{mn}\right)\xi_{pq}(t)\xi_{kl}(t)\\&+\frac{32\pi^4E}{a^6b^6\rho}\sum_r\sum_s\sum_i\sum_j\sum_k\sum_l\sum_p\sum_q B_{rsijmn}B_{pqklij}\xi_{rs}(t)\xi_{pq}(t)\xi_{kl}(t)\\&+\frac{4M}{ab\rho h}\sin\left(\frac{m\pi x_0}{a}\right)\sin\left(\frac{n\pi y_0}{b}\right)\\&\times\sum_p\sum_q\ddot{\xi}_{pq}(t)\sin\left(\frac{p\pi x_0}{a}\right)\sin\left(\frac{q\pi y_0}{b}\right)=0,\end{aligned} \qquad (14)$$

where Ω^2_{mn} are natural frequencies of the target defined as

$$\Omega^2_{mn}=\frac{E}{\rho}\left[\frac{\pi^4h^2}{12(1-\nu^2)}\left(\frac{m^2}{a^2}+\frac{n^2}{b^2}\right)^2+K_{mn}\right]. \qquad (15)$$

The last term in each equation from (14) describes the influence of the coupled impact interaction of the target with the impactor of the mass M applied at the point with the coordinates x_0, y_0.

In order to study this additional nonlinear phenomenon induced by the coupled impact interaction, we suppose that only two natural modes of vibrations are excited during the process of impact, namely, $\Omega_{\alpha\beta}$ and $\Omega_{\gamma\delta}$. Then the set of equations (14) is reduced to the following two equations written in the dimensionless form:

$$\begin{aligned}&p_{11}\ddot{\zeta}_1+p_{12}\ddot{\zeta}_2+\zeta_1\Omega_1^2+p_{13}\zeta_1^2+p_{14}\zeta_2^2+p_{15}\zeta_1\zeta_2\\&\qquad+p_{16}\zeta_1^3+p_{17}\zeta_1\zeta_2^2=0,\end{aligned} \qquad (16)$$

$$\begin{aligned}&p_{21}\ddot{\zeta}_1+p_{22}\ddot{\zeta}_2+\zeta_2\Omega_2^2+p_{23}\zeta_2^2+p_{24}\zeta_1^2+p_{25}\zeta_1\zeta_2\\&\qquad+p_{26}\zeta_2^3+p_{27}\zeta_1^2\zeta_2=0,\end{aligned} \qquad (17)$$

where $\zeta_1=\frac{\xi_{\alpha\beta}}{a}$, $\zeta_2=\frac{\xi_{\lambda\delta}}{a}$, $\Omega_1=\Omega^*_{\alpha\beta}$ and $\Omega_2=\Omega^*_{\gamma\delta}$ are dimensionless natural frequencies

$$\Omega^{*2}_{mn}=\frac{\pi^4h^2}{12(1-\nu^2)a^2}\left(m^2+n^2\frac{a^2}{b^2}\right)^2+a^4K_{mn},$$ dimensionless coefficients p_{ij} (i=1, 2; j=1, 2,…,7) could be easily obtained from (14) using (13), wherein $x^*=x/a$, $y^*=y/b$, and $t^*=\frac{t}{a}\sqrt{\frac{E}{\rho}}$.

III. Method of Solution

In order to solve a set of two nonlinear equations (16) and (17), we apply the method of multiple time scales [16] for constructing the solution of Eqs. (13)

$$\zeta_1(t)=\varepsilon X^1_{\alpha\beta}(T_0,T_1)+\varepsilon^2X^2_{\alpha\beta}(T_0,T_1), \qquad (18)$$

$$\zeta_2(t)=\varepsilon X^1_{\gamma\delta}(T_0,T_1)+\varepsilon^2X^2_{\gamma\delta}(T_0,T_1), \qquad (19)$$

where $T_n=\varepsilon^n t$ are new independent variables, among them: $T_0=t$ is a fast scale characterizing motions with the natural frequencies, and $T_1=\varepsilon t$ is a slow scale characterizing the

modulation of the amplitudes and phases of the modes with nonlinearity.

Considering that

$$\frac{d^2}{dt^2}\zeta_i = \varepsilon(D_0^2 X_{ij}^1) + \varepsilon^2(D_0^2 X_{ij}^2 + 2D_0 D_1 X_{ij}^1),$$

where $ij = \alpha\beta$ or $\gamma\delta$, and $D_i^n = \partial^n / \partial T_i^n$ ($n = 1,2$, $i = 0,1$), and substituting the proposed solution (18) and (19) in (16) and (17), after equating the coefficients at like powers of ε to zero, we are led to a set of recurrence equations to various orders:

to order ε

$$p_{11}D_0^2 X_1^1 + p_{12}D_0^2 X_2^1 + \Omega_1^2 X_1^1 = 0, \tag{20}$$

$$p_{21}D_0^2 X_1^1 + p_{22}D_0^2 X_2^1 + \Omega_2^2 X_2^1 = 0; \tag{21}$$

to order ε^2

$$\begin{aligned} p_{11}D_0^2 X_1^2 + p_{12}D_0^2 X_2^2 + \Omega_1^2 X_1^2 = &-2p_{11}D_0 D_1 X_1^1 \\ &- 2p_{12}D_0 D_1 X_2^1 - p_{13}(X_1^1)^2 - p_{14}(X_2^1)^2 - p_{15}X_1^1 X_2^1, \end{aligned} \tag{22}$$

$$\begin{aligned} p_{21}D_0^2 X_1^2 + p_{22}D_0^2 X_2^2 + \Omega_2^2 X_2^2 = &-2p_{21}D_0 D_1 X_1^1 \\ &- 2p_{22}D_0 D_1 X_2^1 - p_{23}(X_1^1)^2 - p_{24}(X_2^1)^2 - p_{25}X_1^1 X_2^1, \end{aligned} \tag{23}$$

where for simplicity is it denoted $X_1^1 = X_{\alpha\beta}^1$, $X_2^1 = X_{\gamma\delta}^1$, $X_1^2 = X_{\alpha\beta}^2$, and $X_2^2 = X_{\gamma\delta}^2$.

A. Solution of Equations at Order of ε

Following Rossikhin and Shitikova [21], we seek the solution of (20) and (21) in the form:

$$X_1^1 = A_1(T_1)e^{i\omega_1 T_0} + A_2(T_1)e^{i\omega_2 T_0} + \text{cc}, \tag{24}$$

$$X_2^1 = \alpha_1 A_1(T_1)e^{i\omega_1 T_0} + \alpha_2 A_2(T_1)e^{i\omega_2 T_0} + \text{cc}, \tag{25}$$

where $A_1(T_1)$ and $A_2(T_1)$ are unknown complex functions, cc is the complex conjugate part to the preceding terms, and $\bar{A}_1(T_1)$ and $\bar{A}_2(T_1)$ are their complex conjugates,

$$\omega_{1,2}^2 = \frac{(p_{22}\Omega_1^2 + p_{11}\Omega_2^2) \pm \sqrt{(p_{22}\Omega_1^2 - p_{11}\Omega_2^2)^2 + 4\Omega_1^2\Omega_2^2 p_{12}p_{21}}}{2(p_{11}p_{22} - p_{12}p_{21})},$$

$$\begin{aligned} \alpha_1 &= -\frac{p_{11}\omega_1^2 - \Omega_1^2}{p_{12}\omega_1^2} = -\frac{p_{21}\omega_1^2}{p_{22}\omega_1^2 - \Omega_2^2} \\ \alpha_2 &= -\frac{p_{11}\omega_2^2 - \Omega_1^2}{p_{12}\omega_2^2} = -\frac{p_{21}\omega_2^2}{p_{22}\omega_2^2 - \Omega_2^2}, \end{aligned} \tag{26}$$

$$p_{11} = 1 + d_1, \qquad p_{22} = 1 + d_2, \qquad p_{12} = p_{21} = \frac{4M}{\rho hab}s_1 s_2,$$

$$d_1 = \frac{4M}{\rho hab}s_1^2, \qquad d_2 = \frac{4M}{\rho hab}s_2^2,$$

$$s_1 = \sin(\alpha\pi x_0^*)\sin(\beta\pi y_0^*), \qquad s_2 = \sin(\gamma\pi x_0^*)\sin(\delta\pi y_0^*).$$

Reference to relationships (26) shows that ω_1 and ω_2 are the frequencies of the coupled process of impact interaction of the impactor and the target. As the impactor mass $M \to 0$, the frequencies ω_1 and ω_2 tend to the natural frequencies of the shell vibrations Ω_1 and Ω_2, respectively. Coefficients s_1 and s_2 depend on the numbers of the natural modes involved in the process of impact interaction, $\alpha\beta$ and $\gamma\delta$, and on the coordinates of the contact force application x_0^*, y_0^*, resulting in the fact that their particular combinations could vanish coefficients s_1 and s_2 and, thus, coefficients $p_{12} = p_{21}$.

B. Solution of Equations at Order of ε^2

Substituting (24) and (25) in (22) and (23), we obtain

$$\begin{aligned} p_{11}D_0^2 X_1^2 + p_{12}D_0^2 X_2^2 + \Omega_1^2 X_1^2 = &-2i\omega_1(p_{11} + \alpha_1 p_{12})e^{i\omega_1 T_0}D_1 A_1 \\ &- 2i\omega_2(p_{11} + \alpha_2 p_{12})e^{i\omega_2 T_0}D_1 A_2 \\ &- (p_{13} + \alpha_1^2 p_{14} + \alpha_1 p_{15})A_1\left[A_1 e^{2i\omega_1 T_0} + \bar{A}_1\right] \\ &- (p_{13} + \alpha_2^2 p_{14} + \alpha_2 p_{15})A_2\left[A_2 e^{2i\omega_2 T_0} + \bar{A}_2\right] \\ &- 2\left[p_{13} + \alpha_1\alpha_2 p_{14} + (\alpha_1 + \alpha_2)p_{15}\right]A_1 \\ &\times\left[A_2 e^{i(\omega_1+\omega_2)T_0} + \bar{A}_2 e^{i(\omega_1-\omega_2)T_0}\right] + \text{cc}, \end{aligned} \tag{27}$$

$$\begin{aligned} p_{21}D_0^2 X_1^2 + p_{22}D_0^2 X_2^2 + \Omega_2^2 X_2^2 = &-2i\omega_1(p_{21} + \alpha_1 p_{22})e^{i\omega_1 T_0}D_1 A_1 \\ &- 2i\omega_2(p_{21} + \alpha_2 p_{22})e^{i\omega_2 T_0}D_1 A_2 \\ &- (p_{23} + \alpha_1^2 p_{24} + \alpha_1 p_{25})A_1\left[A_1 e^{2i\omega_1 T_0} + \bar{A}_1\right] \\ &- (p_{23} + \alpha_2^2 p_{24} + \alpha_2 p_{25})A_2\left[A_2 e^{2i\omega_2 T_0} + \bar{A}_2\right] \\ &- 2\left[p_{23} + \alpha_1\alpha_2 p_{24} + (\alpha_1 + \alpha_2)p_{25}\right]A_1 \\ &\times\left[A_2 e^{i(\omega_1+\omega_2)T_0} + \bar{A}_2 e^{i(\omega_1-\omega_2)T_0}\right] + \text{cc}. \end{aligned} \tag{28}$$

For the obtained set of coupled equations (27) and (28) all terms proportional to $e^{i\omega_1 T_0}$ and $e^{i\omega_2 T_0}$ are circular terms, so they should be eliminated from the further solution.

Thus, we obtain the following conditions of solvability:

$$D_1 A_1 = 0, \qquad D_1 A_2 = 0, \tag{29}$$

whence it follows that the functions A_1 and A_2 are T_1-independent.

C. Determination of the Contact Force

Representing A_1 and A_2 in the polar form

$$A_i = a_i e^{i\varphi_i} \quad (i = 1,2), \tag{30}$$

relationships (24) and (25) take the form

$$X_1^1 = 2a_1(0)\cos[\omega_1 t + \varphi_1(0)] + 2a_2(0)\cos[\omega_2 t + \varphi_2(0)], \tag{31}$$

$$X_2^1 = 2\alpha_1 a_1(0)\cos[\omega_1 t + \varphi_1(0)] + 2\alpha_2 a_2(0)\cos[\omega_2 t + \varphi_2(0)], \tag{32}$$

wherein the initial amplitudes $a_i(0)$ and phases $\varphi_i(0)$ should be determined from the initial conditions.

Considering (31) and (32), the solution for the shell deflection (9) at the point of impact and the contact force (10) is the following:

$$w(x_0, y_0, t) = \varepsilon(X_1^1 s_1 + X_2^1 s_2), \tag{33}$$

$$P(t) = -\varepsilon M(\ddot{X}_1^1 s_1 + \ddot{X}_2^1 s_2). \tag{34}$$

ISBN: 978-1-61804-251-4

IV. Conclusion

The procedure proposed in the present paper allows one to investigate the dynamic response of a nonlinear doubly curved shallow shell impacted by a sphere, to find the time dependence of the contact force and to determine the contact duration and the maximal contact force.

References

[1] A. W. Leissa and A. S. Kadi, "Curvature effects on shallow shell vibrations," *J. Sound Vibr.*, vol. 16(2), pp. 173–187, 1971.

[2] A. S. Volmir, *Nonlinear Dynamics of Plates and Shells* (in Russian). Moscow: Nauka, 1972.

[3] M. Amabili, "Non-linear vibrations of doubly curved shallow shells," *Int. J. Non-Linear Mech.*, vol. 40, pp. 683–710, 2005.

[4] F. Alijani and M. Amabili, "Chaotic vibrations in functionally graded doubly curved shells with internal resonance," *Int. J. Struct. Stability Dyn.*, vol. 12(6), pp. 1250047 (23 pages), 2012.

[5] M. Amabili and M. P. Paidoussis, "Review of studies on geometrically nonlinear vibrations and dynamics of circular cylindrical shells and panels, with and without fluid-structure interaction," *Appl. Mech. Rev.*, vol. 56, pp. 349-381, 2003.

[6] C. Y. Chia, "Nonlinear analysis of doubly curved symmetrically laminated shallow shells with rectangular platform," *Ing-Archive*, vol. 58, pp. 252--264, 1988.

[7] Y. Kobayashi and A. W. Leissa, "Large amplitude free vibration of thick shallow shells supported by shear diaphragms," *Int. J. Non-Linear Mech.*, vol. 30, pp. 57--66, 1995.

[8] A. Abe, Y. Kobayashi and G. Yamada, "Non-linear vibration characteristics of clamped laminated shallow shells," *J. Sound Vibr*., vol. 234, pp. 405--426, 2000.

[9] K. Chandrashekhara and T. Schoeder, "Nonlinear impact analysis of laminated cylindrical and doubly-curved shells," *J. Composie Mat.*, vol. 29(16), pp. 2160--2179, 1995.

[10] C. Cho, G. Zhao and C. B. Kim, "Nonlinear finite element analysis of composite shell under impact," *KSME Int. J.*, vol. 14(6), pp. 666--674, 2000.

[11] Y. M. Fu and Y.Q. Mao, "Nonlinear dynamic response for shallow spherical moderate thick shells with damage under low velocity impact" (in Chinese), *Acta Materiae Compositae Sinica*, vol. 25(2), pp.166-172, 2008.

[12] Y. M. Fu, Y. Q. Mao and Y. P. Tian, "Damage analysis and dynamic response of elasto-plastic laminated composite shallow spherical shell under low velocity impact," *Int. J. Solids Struct*., vol. 47, pp. 126-137, 2010.

[13] Y. Q. Mao, Y. M. Fu, C. P. Chen and Y. L. Li, "Nonlinear dynamic response for functionally graded shallow spherical shell under low velocity impact in thermal environment,' *Appl. Math. Mod*., vol 35, pp. 2887-2900, 2011.

[14] N. V. Swamy Naidu and P. K. Sinha, "Nonlinear impact behaviour of laminated composite shells in hygrothermal environments," *Int. J. Crashworthiness*, vol. 10(4), pp. 389-402, 2005.

[15] S. Kumar, "Analysis of impact response and damage in laminated composite cylindrical shells undergoing large deformations," *Struct. Eng. Mech.*, vol. 35(3), pp. 349-364, 2010.

[16] A. H. Nayfeh, *Perturbation Methods*. NY: Wiley, 1973.

[17] Kh. M. Mushtari and K. Z. Galimov, *Nonlinear Theory of Thin Elastic Shells* (in Russian). Kazan': Tatknigoizdat, 1957 (English translation NASA-TT-F62, 1961).

[18] J. Lennertz, "Beitrag zur Frage nach der Wirkung eines Querstosses auf einen Stab" (in German), *Eng.-Arch*., vol. 8, pp. 37-46, 1937.

[19] W. Goldsmith, Impact. *The theory and physical behaviour of colliding solids*. London: Arnold, 1960.

[20] V. X. Kunukkasseril and R. Palaninathan, "Impact experiments on shallow spherical shells*," J. Sound Vibr*., vol. 40(1), pp. 101-117, 1975.

[21] Y. A. Rossikhin and M. V. Shitikova, "Free damped nonlinear vibrations of a viscoelastic plate under the two-to-one internal resonance," *Materials Science Forum*, vols. 440-441, pp. 29-36, 2003.

Fractional viscoelastic model of the tooth root displacements in "noncompensable" periodontal ligament

Sergei Bosiakov, Sergei Rogosin
Belarusian State University
Nezavisimosti avenue 4, Minsk, 220030 Belarus
Email: bosiakov@bsu.by

Abstract—Analytical study of the viscoelastic periodontal membrane based on the classical viscoelastic model by Rabotnov. The geometrical shape of the tooth root and periodontal membrane are described by means the equations of elliptic hyperboloid. The outer surface of the periodontal membrane is moved in the normal direction to the outer surface of the root. The latter is rigidly fixed in the bone of dental alveolus. Relations between displacements and strains of the periodontal tissue are formulated with accounting "incompressibility" periodontium. Viscoelastic properties of the periodontal ligament are described by using nondifference kernel (an analog of the Rabotnov function related to Mittag-Leffler function). The movement equations for the translational displacements (in particular, in vertical direction) and rotation angles of the tooth root are obtained.

I. Introduction

Periodontal ligament is a thin membrane that holds the root of the tooth in the alveolar bone. It reduces and distributes the occlusal load on the tooth by means of collagen fibers. In normal conditions there is no contact between the tooth root and the bone tissue. The load acting on the tooth, is transmitted to the alveolar bone by the periodontal ligament strain. Periodontium can be loaded by long-term (orthodontics) forces or by short-term (occlusal) load. It is occurred an orthodontic movement of teeth as a result of the biological response of bone alveolar process [1], [2].

Linearly elastic (bilinear elastic), viscoelastic, hyperelastic and biphasic (multiphasic) models are used to predict the behavior of the periodontal ligament under the various loading conditions. Overview of the specific application of different models is given in [3]. The main drawback of the periodontal ligament simulation on the base of medium with complex properties is the lack of accurate quantitative data on physics and mechanical parameters. For the viscoelastic models it is compensated by existence of known values of the relaxation times and elasticity moduli [4], [5], [6], and the experimental data to determine the viscoelastic properties [7], [8], [9], [10], [11], [12]. In [13], it is shown that all the tissues involved in the reconstruction of bone tissue, demonstrate viscoelastic properties which do not depend on the applied forces. Experimental determination and modeling of properties of periodontal is discussed in [14].

Several viscoelastic models of the periodontal ligament behavior, based on the laws of Maxwell, Voigt, Kelvin-Voigt [3] have been proposed. In particular we have to point out those results which are related to use fractional approaches (see, e.g., [15], [16]). Rabotnov [17] presented a general theory of hereditary solid mechanics using integral equations, and Koeller [18] reviewed the use of integral equations for viscoelasticity and interjects fractional calculus into Rabotnov's theory by the introduction of the spring-pot, which he used to generalize the classical models (see also [19]). Rossikhin and Shitikova [20] (see also [21]) summarized Rabotnov's theory. Rabotnov's fractional exponential function is related to the well known Mittag-Leffler function and they showed the equivalence of Rabotnov's theory to Torvik and Bagley's fractional polynomial constitutive equation. The history of fractional modeling in rheology is presented in [22]. The fractional viscoelastic model is very natural for the study of periodontal membrane.

The aim of this work is to formulate equations of motion of the periodontal ligament. We use an approach based on viscoelastic model similar to Rabotnov's model. It allows us to determine the translational displacement and rotation angles of the periodontium under the action of a concentrated load. In further work we suppose to use the obtained equation to more accurate description of the periodontal ligament behavior.

II. Equations of movements for viscoelastic periodontal ligament

The outer surface of the tooth root (supposed to be an absolutely rigid body) and the adjacent inner surface of the periodontal ligament are modeling by an elliptic hyperboloid

$$F(x,y,z) = y - \frac{h}{\sqrt{1-p^2}-p}\times \\ \times\left(\sqrt{(1-e^2)\left(\frac{x}{b}\right)^2+\left(\frac{z}{b}\right)^2+p^2}-p\right)=0, \tag{1}$$

where h is the height of alveolar crest; $e=\sqrt{1-(b/a)^2}$ is the eccentricity of the ellipse in cross-section of tooth in alveolar crest; a and b are the axes of this ellipse; p is the parameter of rounding of the tooth root. The internal surface of the periodontal ligament adjacent to dental alveoli bone is shifted on the value δ along the normal to the surface of tooth root.

ISBN: 978-1-61804-251-4

Under the action of a concentrated force on a tooth, the points of the periodontal ligament contiguous to surface of the tooth root (1) begin the movement, which are equal to those of the root. The external surface of the periodontal ligament is fixed. There is no significant difference between the model considering the fixing of the outer surface of the periodontal ligament in the alveolar bone or its rigid fixing. Therefore, for calculating the initial movement of the teeth in the periodontal ligament, both the teeth and the alveolar bone could be considered as solids [23]. In what follows we suppose that the periodontal ligament is an incompressible material with Poisson's ratio equal to 0.49. This means that the periodontal tissue begins to flow around the surface of the root of the tooth when the root is displaced to the wall of the dental alveolus [24]. Hence the strains and relative shears associated with the normal, generatrix and guide to the external surface of the tooth root could be represented in the coordinate system as follows [24], [25]:

$$\begin{gathered}\varepsilon_{nn}=-\frac{u_n}{\delta},\varepsilon_{tt}=\varepsilon_{\theta\theta}=0,\\ \gamma_{n\theta}=-\frac{u_\theta}{\delta},\gamma_{nt}=-\frac{u_t}{\delta},\gamma_{t\theta}=0,\end{gathered}\tag{2}$$

where u_n, u_t and u_θ are displacements of the periodontium points in the direction of the $\vec{n}$ (the normal vector to the root surface), $\vec{t}$ (the generatrix vector to the root surface), $\vec{\theta}$ (the tangential vector to the root surface), and δ is the width of the periodontal ligament in the normal direction.

The strains and relative shears can be expressed in the coordinate system (x,y,z) by the components of the strain tensor in the coordinate system (n,t,θ) [25]:

$$\begin{pmatrix}\varepsilon_{11}&\varepsilon_{12}&\varepsilon_{13}\\ \varepsilon_{12}&\varepsilon_{22}&\varepsilon_{23}\\ \varepsilon_{13}&\varepsilon_{23}&\varepsilon_{33}\end{pmatrix}=\tag{3}$$

$$=T_2\cdot T_1\cdot\begin{pmatrix}\varepsilon_{nn}&\varepsilon_{tn}&\varepsilon_{\theta n}\\ \varepsilon_{tn}&0&0\\ \varepsilon_{\theta n}&0&0\end{pmatrix}\cdot T_1^T\cdot T_2^T,$$

$$\varepsilon_{tn}=\frac{1}{2}\gamma_{tn},\varepsilon_{\theta n}=\frac{1}{2}\gamma_{\theta n},1\equiv x,2\equiv y,3\equiv z.$$

The components of the vectors (u_n,u_t,u_θ) and (u_x,u_y,u_z) are related as follows:

$$\begin{pmatrix}u_n\\ u_t\\ u_\theta\end{pmatrix}=T_1^T\cdot T_2^T\cdot\begin{pmatrix}u_x\\ u_y\\ u_z\end{pmatrix},\tag{4}$$

$$T_1=\begin{pmatrix}\sin(\alpha)&\cos(\alpha)&0\\ -\cos(\alpha)&\sin(\alpha)&0\\ 0&0&1\end{pmatrix},$$

$$T_2=\begin{pmatrix}H&0&-G\\ 0&1&0\\ G&0&H\end{pmatrix},$$

$$H=\frac{x(1-e^2)}{\sqrt{x^2(1-e^2)^2+z^2}},G=\frac{z}{\sqrt{x^2+z^2}},$$

where T_1 is the rotation matrix relative to the guide $\vec{\theta}$; T_2 is the rotation matrix relative to the z-axis on angle φ; T_1^T, T_2^T are the transpose matrixes T_1 and T_2, respectively. The angle α between generatrix to the root surface and xz-plane is given by the relation

$$\tan(\alpha)=\frac{h\sqrt{(1-e^2)^2x^2+z^2}}{b(\sqrt{1+p^2}-p)\sqrt{(bp)^2+(1-e^2)x^2+z^2}}.$$

Any displacements of the tooth root can be described by a combination of the translational displacements u_{0x}, u_{0y}, u_{0z} and the angles of rotation θ_x, θ_y, θ_z relative of the coordinate axes. Since the thickness of periodontal is small, the rotation angles are small too. Therefore, we use the following linearized formula [25]:

$$\begin{gathered}u_x=u_{0x}+z\theta_y-y\theta_z,u_y=u_{0y}-z\theta_x+x\theta_z,\\ u_z=u_{0z}+y\theta_x-x\theta_y.\end{gathered}\tag{5}$$

The relationships (2)–(5) allow us to express the strains and relative shifts via translational displacements and rotation angles in the coordinate system (x,y,z).

The components of the stress tensor taking into account the viscoelastic properties of the periodontal ligament are represented in the following form:

$$\begin{gathered}\sigma_{ij}=\frac{E_\infty}{(2\nu-1)(1+\nu)}\Bigg\{(2\nu-1)\varepsilon_{ij}-\\ -\nu_\varepsilon\int_0^t\mathcal{E}_\gamma\left(-\frac{\tau}{\tau_\varepsilon}\right)\varepsilon_{ij}(t-\tau)d\tau+\nu\Bigg(\sum_{k=1}^3\varepsilon_{kk}-\\ -\nu_\varepsilon\int_0^t\mathcal{E}_\gamma\left(-\frac{\tau}{\tau_\varepsilon}\right)\sum_{k=1}^3\varepsilon_{kk}(t-\tau)d\tau\Bigg)\Bigg\},\end{gathered}\tag{6}$$

where τ_s is the relaxation time, $\nu_\varepsilon=\frac{E_\infty-E_0}{E_\infty}$, E_0 and E_∞ are the relaxed (prolonged modulus of elasticity, or the rubbery modulus) and nonrelaxed (instantaneous modulus of elasticity, or the glassy modulus) magnitudes of the elastic modulus, respectively [20], and $\mathcal{E}_\gamma\left(-\frac{\tau}{\tau_\varepsilon}\right)$ is Rabotnov's "fractional exponential function", which describes the relaxation of volume and shear stresses. It was introduced by Rabotnov in the form [26], [17]

$$\mathcal{E}_\gamma\left(-\frac{t}{\tau_\varepsilon}\right)=\frac{t^{\gamma-1}}{\tau_\varepsilon^\gamma}\sum_{n=0}^\infty(-1)^n\frac{(t/\tau_\varepsilon)^{\gamma n}}{\Gamma(\gamma(n+1))},$$

where $0<\gamma<1$ is a fractional parameter. Note that Rabotnov's function is a special case of the classical Mittag-Leffler function highly used in fractional models (see [19]).

To find the translational displacements and the rotation angles, the conditions of the dynamic equilibrium of the tooth root are involved:

$$\begin{gathered}\int\int_F(\vec{n}\cdot\sigma)dF+M\frac{d^2\vec{u_0}}{dt^2}-\vec{P}=0,\\ \int\int_F\vec{r}\times(\vec{n}\cdot\sigma)dF+J\frac{d^2\vec{\theta}}{dt^2}-\vec{m}=0,\end{gathered}\tag{7}$$

where $\vec{m}=(m_x,m_y,m_z)$ is the principal moment of external forces, $\vec{f}=(f_x,f_y,f_z)$ is the principal vector of external forces, $\vec{r}$ is the radius-vector, $\vec{n}=(n_x,n_y,n_z)$ is the unit

ISBN: 978-1-61804-251-4

normal vector to the surface (1), σ is the stress tensor, M is the mass of the tooth root (1), J is the axial moment of inertia of the tooth root, $\vec{u}_0 = (u_{0x}, u_{0y}, u_{0z})$ is the vector of translational displacements of the tooth root along the coordinate axes, and $\vec{\theta} = (\theta_x, \theta_y, \theta_z)$ is the vector of rotation angles of the tooth root with respect to the coordinate axes.

Taking into account relationships (6) one can reduce equations of motion (7) after the transformations to the following form

$$\begin{aligned}
c_x u_{0x} - \nu_\varepsilon c_x \int_0^t \mathcal{E}_\gamma u_{0x}(t-\tau)d\tau + c_{\theta xy}\theta_z - \\
-\nu_\varepsilon c_{\theta xy} \int_0^t \mathcal{E}_\gamma \theta_z(t-\tau)d\tau + M\frac{d^2u_{0x}}{dt^2} = f_x, \\
c_y u_{0y} - \nu_\varepsilon c_y \int_0^t \mathcal{E}_\gamma u_{0y}(t-\tau)d\tau + M\frac{d^2u_{0y}}{dt^2} = f_y, \\
c_z u_{0z} - \nu_\varepsilon c_z \int_0^t \mathcal{E}_\gamma u_{0z}(t-\tau)d\tau + c_{\theta yz}\theta_x - \\
-\nu_\varepsilon c_{\theta yz} \int_0^t \mathcal{E}_\gamma \theta_x(t-\tau)d\tau + M\frac{d^2u_{0z}}{dt^2} = f_z, \\
c_{\theta z} u_{0z} - \nu_\varepsilon c_{\theta z} \int_0^t \mathcal{E}_\gamma u_{0z}(t-\tau)d\tau + \mu_x\theta_x - \\
-\nu_\varepsilon\mu_x \int_0^t \mathcal{E}_\gamma \theta_x(t-\tau)d\tau + J_x\frac{d^2\theta_x}{dt^2} = y_f f_z - z_f f_y, \\
\mu_y\theta_y - \nu_\varepsilon\mu_y \int_0^t \mathcal{E}_\gamma \theta_y(t-\tau)d\tau + J_y\frac{d^2\theta_y}{dt^2} = \\
= z_f f_x - x_f f_z, \\
c_{\theta x} u_{0x} - \nu_\varepsilon c_{\theta x} \int_0^t \mathcal{E}_\gamma u_{0x}(t-\tau)d\tau + \mu_z\theta_z - \\
-\nu_\varepsilon\mu_z \int_0^t \mathcal{E}_\gamma \theta_z(t-\tau)d\tau + J_z\frac{d^2\theta_z}{dt^2} = x_f f_y - y_f f_x, \\
\mathcal{E}_\gamma \equiv \mathcal{E}_\gamma\left(-\frac{\tau}{\tau_\varepsilon}\right),
\end{aligned} \tag{8}$$

where c_x, c_y, and c_z are the stiffness coefficients of the periodontal ligament at the tooth root translation along the co-ordinate axes; $c_{\theta xy}$ and $c_{\theta yz}$ are the static moments of stiffness; $c_{\theta x}$ and $c_{\theta z}$ are the stiffness coefficients of the periodontal ligament at the tooth root rotations relative to the x-axis and z-axis, respectively under the force acting along this coordinate axis; μ_x, μ_y, and μ_z are the stiffness coefficients of the periodontal ligament at the tooth root rotations relative to the axes x, y and z, respectively; and x_f, y_f and z_f are the coordinates of the point where the load is applied. The coefficients of system (8) are defined as follows

$$\begin{aligned}
c_x = E_\infty \int\int_F (ABb(2\nu-1)\cos(\alpha) + \\
+h(2Hx(1-e^2)(\nu-1) + Gz(2\nu-1))\sin(\alpha))\frac{dF}{C}, \\
c_y = E_\infty \int\int_F (ABb(1-2\nu)\cos(\alpha) - \\
-h(Hx(1-e^2)(1-2\nu) + 2Gz(1-\nu))\sin(\alpha))\frac{dF}{C}, \\
c_z = E_\infty \int\int_F (ABb(1-2\nu)\cos(\alpha) - \\
-h(Hx(1-e^2)(1-2\nu) + \\
+2Gz(1-\nu))\sin(\alpha))\frac{dF}{C}, \\
c_{\theta x} = -E_\infty \int\int_F ((1-2\nu)((1-e^2)hx^2 + ABby) \times \\
\times\cos(\alpha) + (hy(2Hx(1-e^2)(1-\nu) + \\
+Gz(1-2\nu)) + 2ABHb\nu x)\sin(\alpha))\frac{dF}{C}, \\
c_{\theta z} = E_\infty \int\int_F ((1-2\nu)(hz^2 + ABby)\cos(\alpha) + \\
+(hy(Hx(1-e^2)(1-2\nu) + 2Gz(1-\nu)) + \\
+2ABGb\nu z)\sin(\alpha))\frac{dF}{C}, \\
c_{\theta xy} = E_\infty \int\int_F ((2B(1-e^2)h\nu x^2 - \\
-Ab(1-2\nu)yB^2)\cos(\alpha) - \\
-(Bhy(2Hx(1-e^2)(1-\nu) + Gz(1-2\nu)) + \\
+AB^2bhx(1-2\nu))\sin(\alpha))\frac{dF}{C}, \\
c_{\theta yz} = E_\infty \int\int_F ((ABby(1-2\nu) + 2hz^2\nu)\cos(\alpha) + \\
+(ABGbz(1-2\nu) + hy(Hx(1-e^2)(1-2\nu) + \\
+2Gz(1-\nu)))\sin(\alpha))\frac{dF}{C}, \\
\mu_x = E_\infty \int\int_F ((hyz^2 + ABb((1-2\nu)y^2 + \\
+2z^2(1+\nu)))\cos(\alpha) + \\
+(ABGbyz + h(Hx(1-e^2)(1-2\nu)(y^2+z^2) + \\
+Gz(2y^2(1-\nu) + z^2(1-2\nu))))\cos(\alpha))\frac{dF}{C}, \\
\mu_y = E_\infty \int\int_F (ABb(x^2+z^2)(1-2\nu)\cos(\alpha) - \\
-h(Gz(x^2(1+e^2-2\nu) + z^2(1-2\nu)) + \\
+BHx(x^2(1-e^2)(1-2\nu) + \\
+z^2(1-2e^2(1-\nu)-2\nu)))\sin(\alpha))\frac{dF}{C}, \\
\mu_z = E_\infty \int\int_F (hx^2y(1-e^2) + \\
+ABb(2x^2(1-\nu) + y^2(1-2\nu))\cos(\alpha) + \\
+(Hx(ABby + h(1-e^2)(x^2(1-2\nu) + \\
+2y^2(1-\nu))) + Ghz(x^2+y^2))\sin(\alpha))\frac{dF}{C}, \\
A = \sqrt{p^2+1} - p, B = \sqrt{(bp)^2 + (1-e^2)x^2 + z^2}, \\
C = 2ABb\delta(1+\nu)(2\nu-1).
\end{aligned}$$

III. The tooth root translation in vertical direction

To find the material constants and the relaxation time, the experimental data on the stress-strain state of the periodontal ligament could be used and, in particular, the time-dependence of the periodontal points displacements. Typically, such data

are obtained for the translational movement of the tooth root in the vertical and horizontal directions.

During the motion of the tooth root along the y-axis, the corresponding extrusion (or intrusion), the translational displacement along the x- and z-axes, as well as the angles of rotation are equal to zero, i. e., $u_{0x} = u_{0z} = 0$, and $\theta_x = \theta_y = \theta_z = 0$. Load is acting only in the y-axis direction. In this case, from (8) we obtain

$$c_y\left(u_{0y} - \nu_\varepsilon \int_0^t \mathcal{E}_\gamma\left(-\frac{\tau}{\tau_\varepsilon}\right) u_{0y}(t-\tau)d\tau\right) + \\ + M\frac{d^2u_{0y}}{dt^2} = f_y. \tag{9}$$

Note that (9) is similar to the equations of motion of a viscoelastic oscillator presented in [27], [28].

IV. Conclusion

The equations of motion of the tooth root with a fractional exponential function for generalization of viscoelastic models are proposed. The advantage of this model is the use of the fractional parameter to describe the various pathological processes and age-related changes in the periodontium. Fractional parameter allows us to take into account the different behavior of the periodontal tissue during the action of the short-term and the long-term loads. The experimental data [5], [7], [8] and [9] on the intrusion or extrusion of the tooth for the material constants assessment can be used together with the solution of equation (9).

Acknowledgment

The research is supported by the FP7 IRSES Marie Curie grant TAMER No 610547R. The authors are thankful to professor Francesco Mainardi for supporting of the above described approach.

References

[1] R. S. Masella and M. Meister, "Current concepts in the biology of orthodontic tooth movement," *Am. J. Orthod. Dentofacial. Orthop.*, vol. 129, no. 4, pp. 458–468, 2006.

[2] G. E. Wise and G. J. King, "Mechanisms of tooth eruption and orthodontic tooth movement," *J. Dent. Res.*, vol. 87, no. 5, pp. 414–434, 2008.

[3] T. S. Fill, R. W. Toogood, P. W. Major, and J. P. Carey, "Analytically determined mechanical properties of, and models for the periodontal ligament: Critical review of literature," *J. Biomech.*, vol. 45, pp. 9–16, 2012.

[4] K. Komatsu, "Mechanical strength and viscoelastic response of the periodontal ligament in relation to structure," *J. Dent. Biomech.*, vol. 18, 2010, doi:10.4061/2010/502318 (Article ID 502318).

[5] L. Qian, M. Todo, Y. Morita, Y. Matsushita, and K. Koyano, "Deformation analysis of the periodontium considering the viscoelasticity of the periodontal ligament," *Dent. Materials*, vol. 25, pp. 1285–1292, 2009.

[6] S. A. Wood, D. S. Strait, E. R. Dumont, C. F. Ross, and I. R. Grosse, "The effects of modeling simplifications on craniofacial finite element models: The alveoli (tooth sockets) and periodontal ligaments," *J. Biomech.*, vol. 44, pp. 1831–1838, 2011.

[7] M. Ferrari, R. Sorrentino, F. Zarone, D. Apicella, R. Aversa, and A. Apicella, "Non-linear viscoelastic finite element analysis of the effect of the length of glass fiber posts on the biomechanical behaviour of directly restored incisors and surrounding alveolar bone," *Dent. Mat. J.*, vol. 27, no. 4, pp. 485–498, 2008.

[8] A. N. Natali, P. G. Pavan, and C. Scapta, "Numerical analysis of tooth mobility: formulation of a non-linear constitutive law for the periodontal ligament," *Dent. Mater.*, vol. 20, pp. 623–629, 2004.

[9] S. R. Toms and A. W. Eberhardt, "A nonlinear finite element analysis of the periodontal ligament under orthodontic tooth loading," *Am. J. Orthod. Dentofacial Orthop.*, vol. 123, pp. 657–665, 2003.

[10] M. Bergomi, J. Cugnoni, M. Galli, J. Botsis, U. C. Belser, and H. W. A. Wiskott, "Hydro-mechanical coupling in the periodontal ligament: A porohyperelastic finite element model," *J. Biomech.*, vol. 44, pp. 34–38, 2011.

[11] G. R. S. Naveh, N. L.-T. Chattah, P. Zaslansky, R. Shahar, and S. Weiner, "Tooth-pdl-bone complex: Response to compressive loads encountered during mastication - a review," *Arch. Oral Biology*, vol. 57, pp. 1575–1584, 2012.

[12] N. Yoshida, Y. Koga, C.-L. Peng, E. Tanaka, and K. Kobayashi, "In vivo measurement of the elastic modulus of the human periodontal ligament," *Med. Eng. Phys.*, vol. 23, pp. 567–572, 2001.

[13] M. Cronau, D. Ihlow, D. Kubein-Meesenburg, J. Fanghanel, H. Dathe, and H. Nagerl, "Biomechanical features of the periodontium: An experimental pilot study in vivo," *Am. J. Orthod. Dentofacial Orthop.*, vol. 129, pp. 599.e13–599.e21, 2006.

[14] T. S. Fill, J. P. Carey, R. W. Toogood, and P. W. Major, "Experimentally determined mechanical properties of, and models for, the periodontal ligament: critical review of current literature," *J. Dent. Biomech.*, vol. 10, 2011, doi: 4061/2011/312980 (Article ID 312980).

[15] V. V. Uchaikin, *Fractional derivatives for physicists and engineers*. Berlin and Beijing: Springer and Higher Education Press, 2013, vol. I–II.

[16] R. C. Koeller, "A theory relating creep and relaxation for linear materials with memory," *J. Appl. Mech.*, vol. 77, pp. 031 008–1–031 008–9, 2010.

[17] Y. N. Rabotnov, *Elements of Hereditary Solid Mechanics*. Moscow: Mir Publishers, 1980.

[18] R. C. Koeller, "Application of fractional calculus to the theory of viscoelasticity," *ASME J. Appl. Mech.*, vol. 51, pp. 299–307, 1984.

[19] F. Mainardi, *Fractional Calculus and Waves in Linear Viscoelasticity*. London and Singapore: Imperial College Press and World Scientific, 2010.

[20] Y. A. Rossikhin and M. V. Shitikova, "Centennial jubilee of academician Rabotnov and contemporary handling of his fractional operator," *Fract. Calc. Appl. Analysis*, vol. 17, no. 3, pp. 674–683, 2014.

[21] ——, "Two approaches for studying the impact response of viscoelastic engineering systems: An overview," *Comp. Math. Appl.*, vol. 66, pp. 755–773, 2013.

[22] S. Rogosin and F. Mainardi, "George William Scott Blair - the pioneer of factional calculus in rheology," *Com. Appl. Ind. Math.*, 2014, arXiv:1404.3295v1 [math.HO] 12 Apr 2014.

[23] A. Hohmann, C. Kober, P. Young, C. Dorow, M. Geiger, A. Boryor, F. M. Sander, C. Sander, and F. G. Sander, "Influence of different modeling strategies for the periodontal ligament on finite element simulation results," *Am. J. Orthod. Dentofacial. Orthop.*, vol. 139, pp. 775–783, 2011.

[24] C. G. Provatidis, "An analytical model for stress analysis of a tooth in translation," *Int. J. Eng. Sci.*, vol. 39, pp. 1361–1381, 2001.

[25] A. V. Schepdael, L. Geris, and J. V. der Sloten, "Analytical determination of stress patterns in the periodontal ligament during orthodontic tooth movement," *Med. Eng. Phys.*, vol. 35, pp. 403–410, 2013.

[26] Y. N. Rabotnov, "Equilibrium of an elastic medium with after-effect," *Prikl. Matem. i Mekh. (PMM)*, vol. 12, no. 1, pp. 81–91, 1948, reprinted in English in Fract. Calc. Appl. Anal. Vol.17, no 3, pp. 684–696.

[27] Y. Rossikhin, M. V. Shitikova, and T. A. Scheglova, "Analysis of free vibrations of a viscoelastic oscillator via the models involving several fractional parameters and relaxation/retardation times," *Comp. Math. Appl.*, vol. 59, pp. 1727–1744, 2010.

[28] Y. Rossikhin and M. V. Shitikova, "Free damped vibrations of a viscoelastic oscillator based on Rabotnov's model," *Mech. Time-Depend. Mater.*, vol. 12, pp. 129–149, 2008.

Fractional model of electron diffusion in dye-sensitized nanocrystalline solar cells

Sibatov R. T., Svetukhin V. V., Uchaikin V. V., Morozova E. V.
Ulyanovsk State University
Email: ren_sib@bk.ru

Abstract—Dye-sensitized solar cells traditionally include a highly porous metal oxide nanocrystalline semiconductors, such as TiO_2 film on a transparent conductive electrode. Influence of topological disorder and energy distribution of localized states on transient currents is described in the framework of trap-limited diffusion on a comb structure simulating a percolation cluster of a porous structure. The integral transport equation with fractional order derivatives for the trap-limited diffusion in case of an arbitrary density of localized states is derived. Peculiarities of time-of-flight data for porous semiconductors are shortly discussed.

I. Introduction

The important component of modern dye-sensitized solar cells (DSSC) is a porous layer of metal oxide (often TiO_2) nanoparticles serving as an anode. Sunlight is absorbed by a molecular dye covering these nanocrystallites. Electrolyte is placed between an anode and a cathode. Sunlight passing through the transparent electrode excites electrons in the dye. The electrons flow in the metal oxide toward the electrode where they are collected for powering a load. The efficient electronic transport is responsible for overall performance of these solar cells and disordered structure of porous anode affects essentially on transport characteristics [1], [2].

The electron transport mechanism in nanoporous TiO_2 films is often described by the phonon-assisted hopping or multiple-trapping (MT) model [1], [3], and it is well known that the morphology of a material plays an essential role [3], [4], [5]. In present paper, we study joint action of percolation structure of porous electrode and trap distribution in the band gap to describe electron transport in dye-sensitized solar cells. We propose a fractional model of trap-limited diffusion on a comb structure simulating a percolation cluster.

II. Integral equation of multiple trapping

In recent years, there has been significant progress in understanding electron transport in nanoporous titania films [3], [6]. The models of variable-range hopping and multiple trapping (MT) are often used for electrons in an electrically homogeneous medium with traps.

In the multiple trapping model (see details in [7], [8], [9]), carriers are divided into mobile or free (f) and trapped (t) ones. First of them being in delocalized states are responsible for the charge transport: they are scattered on phonons and inhomogeneities, and drift in the external electric field $\mathbf{E}$ with an average velocity $\mathbf{v} = \mu\mathbf{E}$, where μ is mobility. We assume that the process is going on in a regular medium, i.e., particles in identical macroscopic volumes of a material are involved in a roughly similar mean number of capture events. Trap density is much higher than the number density of carriers involved in the transport process.

Denote the concentration of free holes by $p_f(x,t)$ and the concentration of trapped holes with activation energy in $d\varepsilon$ by $p_t(x,t;\varepsilon)d\varepsilon$, then the total concentration reads

$$p(x,t) = p_f(x,t) + \int_0^\infty p_t(x,t;\varepsilon)\ d\varepsilon. \tag{1}$$

The continuity equation in condition of the time-of-flight experiment with surface photoinjection of carriers

$$\frac{\partial p(x,t)}{\partial t} + \frac{\partial}{\partial x}\left\{\mu E\ p_f(x,t) - D\frac{\partial}{\partial x}p_f(x,t)\right\} = N\delta(x)\delta(t). \tag{2}$$

In the case of weak occupation, p_t and p_t are linked via interrelation

$$\frac{\partial p_t(x,t;\varepsilon)}{\partial t} = \omega_\varepsilon\rho(\varepsilon)p_f(x,t) - \frac{N_c}{N_t}\ \omega_\varepsilon e^{-\varepsilon/kT}p_t(x,t;\varepsilon).$$

It is the linear first-order differential equation for function $p_t(x,t;\varepsilon)$, solving which, we find the link between concentrations of localized and mobile carriers

$$p_t(x,t;\varepsilon) = \int_{-\infty}^t p_f(x,\tau)\ \omega_\varepsilon\rho(\varepsilon)\exp\left\{-\omega_\varepsilon\frac{N_c}{N_t}e^{-\varepsilon/kT}(t-\tau)\right\}d\tau. \tag{3}$$

Substituting it into (1), and then into the continuity equation (2), we obtain the following trap-limited diffusion equation

$$\frac{\partial p_f(x,t)}{\partial t} + \frac{\partial}{\partial t}\int_{-\infty}^t p_f(x,\tau)\ Q(t-\tau)d\tau + \frac{\partial}{\partial x}\left\{\mu E\ p_f(x,t) - D\frac{\partial}{\partial x}p_f(x,t)\right\} = N\delta(x)\delta(t). \tag{4}$$

Here N is the number of photoinjected carriers. Kernel Q is defined as the integral over trap energies

$$Q(t) = \int_0^\infty \omega_\varepsilon\exp\left\{-\omega_\varepsilon t\frac{N_c}{N_t}e^{-\varepsilon/kT}\right\}\rho(\varepsilon)d\varepsilon. \tag{5}$$

The Laplace transformation of equation (4) leads to

$$s[1+\tilde{Q}(s)\ ]\tilde{p}_f(x,s) + \frac{\partial}{\partial x}\left\{\mu E\tilde{p}_f(x,s) - D\frac{\partial}{\partial x}\tilde{p}_f(x,s)\right\}$$

ISBN: 978-1-61804-251-4

$$= N\delta(x). \quad (6)$$

From Eqs. (3) and (5),

$$\tilde{p}_t(x,s) = \tilde{Q}(s)\tilde{p}_f(x,s), \quad \tilde{p}_f(x,s) = \tilde{Q}^{-1}(s)\tilde{p}_t(x,s),$$

$$\tilde{p}_f(x,s) = \frac{\tilde{p}(x,s)}{1+\tilde{Q}(s)}.$$

Using the latter relation and equality (6), we obtain the equation for the transform of total concentration

$$s[1+\tilde{Q}(s)]\ \tilde{p}(x,s) + \frac{\partial}{\partial x}\left\{\mu E\tilde{p}(x,s) - D\frac{\partial}{\partial x}\tilde{p}(x,s)\right\}$$
$$= N\delta(x)[1+\tilde{Q}(s)], \quad (7)$$

and after inverting, we arrive at the integral equation of the multiple trapping

$$\frac{\partial p(x,t)}{\partial t} + \frac{\partial}{\partial t}\int_{-\infty}^{t} p(x,\tau)\ Q(t-\tau)d\tau$$
$$+\frac{\partial}{\partial x}\left\{\ \mu E\ p(x,t) - D\frac{\partial}{\partial x}p(x,t)\right\}$$
$$= N\delta(x)[\delta(t)+Q(t)]. \quad (8)$$

In case of a weak dependence of capture rate on energy $\omega_\varepsilon \approx \omega_0$ for exponential density of states $\rho(\varepsilon) = \varepsilon_0^{-1}\exp(-\varepsilon/\varepsilon_0)$, we obtain,

$$Q(t) = \frac{\omega_0}{\varepsilon_0}\int_0^\infty \exp\left\{-\omega_0 t\frac{N_c}{N_t}\ e^{-\varepsilon/kT}\right\}\exp(-\varepsilon/\varepsilon_0)d\varepsilon.$$

The substitution of variable $\xi = \omega_0 t\ (N_c/N_t)\ e^{-\varepsilon/kT}$ leads to the following kernel

$$Q(t) = \frac{\omega_0\alpha}{(\omega_0 t N_c/N_t)^\alpha}\int_0^{\omega_0 t\cdot N_c/N_t} e^{-\xi}\xi^{\alpha-1}d\xi$$
$$\sim \frac{\omega_0\alpha\Gamma(\alpha)}{(\omega_0 N_c/N_t)^\alpha}\ t^{-\alpha}, \quad t\to\infty, \quad \alpha = \frac{kT}{\varepsilon_0}.$$

Further, reduce the kernel to the form

$$Q(t) \sim \frac{\alpha\pi\omega_0}{(\omega_0 N_c/N_t)^\alpha\sin\pi\alpha}\ \frac{t^{-\alpha}}{\Gamma(1-\alpha)}$$
$$= \omega_0\frac{(c_\alpha t)^{-\alpha}}{\Gamma(1-\alpha)}, \quad c_\alpha = \frac{\omega_0 N_c}{N_t}\left(\frac{\sin\pi\alpha}{\pi\alpha}\right)^{1/\alpha}. \quad (9)$$

The Laplace transform of this kernel for $s \ll \omega_0 N_c/N_t$,

$$\tilde{Q}(s) \sim \frac{\omega_0}{c_\alpha^\alpha}s^{\alpha-1}. \quad (10)$$

After substitution of relation (9) into Eq. (4), we obtain the equation with fractional derivative of order, which is defined by the ratio of the Boltzmann temperature to the exponential band tail width ($\alpha = kT/\varepsilon_0$),

$$\frac{\partial p_f(x,t)}{\partial t} + \omega_0 c_\alpha^{-\alpha}\ {}_{-\infty}\mathsf{D}_t^\alpha\ p_f(x,\tau)$$
$$+\frac{\partial}{\partial x}\left\{\mu E p_f(x,t) - D\frac{\partial}{\partial x}p_f(x,t)\right\} = N\delta(x)\delta(t) \quad (11)$$

(for detail referred to fractional derivatives see [12]).

For the total concentration of carriers, as follows from (9) and (8), the following fractional Fokker-Planck equation takes place

$$\frac{\partial p(x,t)}{\partial t} + \omega_0 c_\alpha^{-\alpha}\ {}_{-\infty}\mathsf{D}_t^\alpha\ p(x,\tau)$$
$$+\frac{\partial}{\partial x}\left\{\mu E p(x,t) - D\frac{\partial}{\partial x}p(x,t)\right\} =$$
$$= N\delta(x)\left\{\delta(t) + \omega_0\frac{(c_\alpha t)^{-\alpha}}{\Gamma(1-\alpha)}\right\}. \quad (12)$$

Thus, the order of fractional derivative is equal to the dispersion parameter. In case $\alpha < 1$, the first item in Eq. (11) can be neglected, and diffusion is anomalous [10]. When $\alpha > 1$, for large times $t\to\infty$ the term with fractional derivative is negligible component, and we arrive at the standard Fokker-Planck equation and normal regime.

In frame of the CTRW-conception [11] the equation (8) can be generalized to the following one for the case of more than one independent delocalization mechanisms [12],

$$\frac{\partial}{\partial t}\int_0^t p(x,\tau)\ \sum_j w_j\Psi_j(t-\tau)\ d\tau + C\frac{\partial}{\partial x}p(x,t)$$
$$-D\frac{\partial^2}{\partial x^2}p(x,t) = \delta(x)\sum_j w_j\Psi_j(t). \quad (13)$$

Here, $\Psi(t) = \sum_j w_j\Psi_j(t)$, where j is a mechanism number, w_j is the corresponding probability of its realization.

Equation (13) can be applied to the multiple trapping. If the density of localized state energies of spatially heterogeneous system can be presented as the superposition $\rho(\varepsilon) = \sum_j w_j\rho_j(\varepsilon)$, then $\Psi_j(t)$ is the complementary distribution function of waiting times in states with density ρ_j.

III. Trap-limited diffusion on a comb

The simplest model of an infinite percolation cluster of dead ends is the so-called *comb model* [13], [14]. At each node of a conductive bond, the walker can go into a dead branch of a cluster, represented by a tooth of the comb, where it undergoes diffusion (in the simplest case, the normal Gaussian diffusion). The carrier return from the dead loop in the main channel occurs through a random time. The fractional equation of diffusion was first associated with a comb model by Nigmatullin [15]. A more detailed analysis of this model can be found in Arkhincheev's and coauthors works [16], [17].

Consider trap-limited diffusion on a comb structure. The conductive bond and dead branches are directed along x- and y-axis, respectively. Diffusion equation has the form

$$\frac{\partial}{\partial t}\int_0^t\sum_j w_j\Psi_j(t-\tau)\ G(x,y,\tau)d\tau - D_x\delta(y)\frac{\partial^2 p(x,y,t)}{\partial x^2}$$
$$-D_y\frac{\partial^2 p(x,y,t)}{\partial y^2} = N\delta(x)\delta(y)[\delta(t)+Q(t)],$$

where D_x and D_y stand for diffusivities along x- and y-axes correspondingly. The Laplace transformation on time and Fourier one on coordinate x leads to the expression

$$s\left[\sum_j w_j\tilde{\Psi}_j(s)\right]\tilde{p}(k,y,s)+D_x\delta(y)k^2\tilde{p}(k,y,s)$$
$$-D_y\frac{\partial^2}{\partial y^2}\tilde{p}(k,y,s)=N\delta(y)\left[\sum_j w_j\tilde{\Psi}_j(s)\right].$$

Its solution has the form

$$\tilde{p}(k,y,s)=$$
$$N\frac{\left[\sum_j w_j\tilde{\Psi}_j(s)\right]\exp\left\{-D_y^{-1/2}\sqrt{s\sum_j w_j\tilde{\Psi}_j(s)}|y|\right\}}{2\left\{D_y s\sum_j w_j\tilde{\Psi}_j(s)\right\}^{1/2}+D_xk^2}. \quad (14)$$

The equation for the concentration of particles on a bond:

$$\left\{2\left[D_y s\left(\sum_j w_j\tilde{\Psi}_j(s)\right)\right]^{1/2}+D_xk^2\right\}\tilde{p}(k,y=0,s)$$
$$=N\sum_j w_j\tilde{\Psi}_j(s).$$

Integrating equation (14) over y, we obtain the following expression

$$s^\gamma\left[\sum_j w_j\tilde{\Psi}_j(s)\right]^\gamma\tilde{p}(k,s)+Kk^2\tilde{p}(k,s)$$
$$=Ns^{\gamma-1}\left[\sum_j w_j\tilde{\Psi}_j(s)\right]^\gamma, \quad (15)$$

where $\gamma=1/2$, $K=D_x/2D_y^\gamma$

In case of exponential density of localized states, taking Eq. (10) into account, one obtain

$$s^{\alpha\gamma}\tilde{p}(k,s)+Kk^2\tilde{p}(k,s)=Ns^{\alpha\gamma-1}. \quad (16)$$

The inverse transformation leads to the fractional equation

$${}_0\mathsf{D}_t^{\alpha\gamma}\hat{p}(k,t)+Kk^2\hat{p}(k,t)=N\frac{t^{-\alpha\gamma}}{\Gamma(1-\alpha\gamma)}, \quad (17)$$

The comb-model is developed and modified, for example, by truncation of teeth length, using random lengths or random positions of teeth on the x-axis. So, the authors [18] obtained the following fractional equation for diffusion in the comb structure with random teeth lengths distributed according to the power law $P\{L>l\}=(l/l_0)^{-\beta}$:

$$I_\beta\frac{l_0}{d}\left(\frac{D_y}{l_0^2}\right)^{\frac{1-\beta}{2}}{}_0\mathsf{D}_t^{\frac{1+\beta}{2}}p(x,t)=D_x\frac{\partial^2p(x,t)}{\partial x^2}-\mu E\frac{\partial p(x,t)}{\partial x}.$$

Here, I_β is a constant defined by the integral (see [18]):

$$I_\beta=\beta\int_0^\infty\xi^{-\beta-1}\tanh\xi\ d\xi,$$

d distance between teeth, μ mobility of particles along bond.

Considering trap-limited diffusion on this structure, we obtain the equation

$$s^\gamma\left[\sum_j w_j\tilde{\Psi}_j(s)\right]^\gamma\tilde{p}(k,s)+D'k^2\tilde{p}(k,s)-ik\mu'E\ \tilde{p}(k,s)$$
$$=Ns^{\gamma-1}\left[\sum_j w_j\tilde{\Psi}_j(s)\right]^\gamma, \quad (18)$$

where

$$D'=D_x\left[I_\beta\frac{l_0}{d}\left(\frac{D_y}{l_0^2}\right)^{\frac{1-\beta}{2}}\right]^{-1},$$
$$\mu'=\mu\left[I_\beta\frac{l_0}{d}\left(\frac{D_y}{l_0^2}\right)^{\frac{1-\beta}{2}}\right]^{-1},\quad \gamma=\frac{1+\beta}{2}.$$

The inverse Fourier-Laplace transformation leads to the generalized diffusion equation with a specific integro-differential operator on time.

Using this equation, one can study competition between contributions of morphology and delocalization mechanisms at different time scales.

IV. Transient current in case of trap-limited transport on a comb

In the classical "time-of-flight" experiment, electrons and holes are usually generated in a sample by a pulse of laser radiation from the side of the semitransparent electrode. The voltage applied to the electrodes is such that the corresponding electric field inside the sample is significantly stronger than the field of nonequilibrium charge carriers. The electrons (or holes, depending on the voltage sign) enter the semitransparent electrode, while holes (or electrons) drift to the opposite electrode.

Calculate transient current in case of trap-limited transport on a comb. The Laplace transformation of equation (13) after neglecting by diffusion term leads to the following equation

$$s\left[\sum_j w_j\tilde{\Psi}_j(s)\right]^\gamma\tilde{p}(x,s)+v_d\frac{\partial}{\partial x}\tilde{p}(x,s)$$
$$=N\delta(x)\left[\sum_j w_j\tilde{\Psi}_j(s)\right]^\gamma, \quad (19)$$

which has solution $\tilde{p}(x,s)=$

$$\frac{N}{v_d}\left[\sum_j w_j\tilde{\Psi}_j(s)\right]^\gamma\exp\left(-\frac{xs}{v_d}\left[\sum_j w_j\tilde{\Psi}_j(s)\right]^\gamma\right). \quad (20)$$

The expression for the transient current

$$I(t)=\frac{e}{L}\frac{d}{dt}\int_0^L(x-L)\ p(x,t)dx \quad (21)$$

after Laplace transformation and substitution of Eq. (20) takes the form:

$$\tilde{I}(s) = \frac{eNv_d}{Ls\left[\sum_j w_j \tilde{\Psi}_j(s)\right]^{\gamma}}$$

$$\times \left[1 - \exp\left(-\frac{Ls}{v_d}\left[\sum_j w_j \tilde{\Psi}_j(s)\right]^{\gamma}\right)\right]. \qquad (22)$$

Using the latter expression and equation (18), we analyze the characteristics of dispersive transport and the temperature dependence of dispersion parameter in the dye-sensitized nanostructured solar cells. On the base these equations, it is possible to develop an algorithm for reconstructing the density of localized states from transient current curves, taking into account the percolation structure of a porous material.

V. Conclusion

We have proposed here a model, which takes into account the influence of percolation morphology, energy distribution of localized states and independent action of several transport mechanisms. The pecularity of the model is that it is based on the fractional calculus operations. Involving a new additional parameters makes the model to be more plastic in order to explain a number of experimental facts in porous nanostructured materials used, for example, in dye-sensitized nanocrystalline solar cells [1], [3]. So the classical MT-model explains main features of dispersive transport, but fails in explaining the temperature dependence of the diffusion coefficient and dispersion parameter in nanoporous semiconductors [3], [6]. Static comb model of a percolation structure leads to a temperature-independent dispersion parameters and the diffusion coefficient. The combination with the MT-model or phonon-assisted hopping allows us to vary the temperature dependence of parameters in wide ranges. It was not obvious how to produce this combination of mechanisms. Attempts have been made in a number of papers (see, eg, [19], [4], [5]). Note, that when considering multiple trapping on comb structures is equivalent to "serial connection" of mechanisms, and the model can be formalized in terms of the subordinated stochastic processes. Action of several independently operating mechanisms is described in the model of random walks with a mixture of distributions of waiting times. In the first case, the effective dispersion parameter is defined as the product of dispersion parameters α and γ, in the second one, we have a model described by differential equation with fractional derivative of distributed order.

Acknowledgment

We are also grateful to the Russian Foundation for Basic Research (project no. 13-01-00585) and the Ministry of Education and Science of the Russian Federation for financial support.

References

[1] Abdi, N., Abdi, Y., Oskoee, E. N., & Sajedi M. (2014). Electron diffusion in trap-contained 3D porous nanostructure: simulation and experimental investigation, *Journal of Nanoparticle Research*, 16(3), 1-8.

[2] Bai, Y., Zhang, J., Zhou, D., Wang, Y., Zhang, M., & Wang, P. (2011). Engineering organic sensitizers for iodine-free dye-sensitized solar cells: red-shifted current response concomitant with attenuated charge recombination, *Journal of the American Chemical Society*, 133(30), 11442-11445.

[3] Ansari-Rad, M., Abdi, Y., & Arzi, E. (2012). Monte Carlo random walk simulation of electron transport in dye-sensitized nanocrystalline solar cells: influence of morphology and trap distribution. *The Journal of Physical Chemistry C*, 116(5), 3212-3218.

[4] Benkstein, K. D., Kopidakis, N., Van de Lagemaat, J., & Frank, A. J. (2003). Influence of the percolation network geometry on electron transport in dye-sensitized titanium dioxide solar cells. *The Journal of Physical Chemistry B*, 107(31), 7759-7767.

[5] Cass, M. J., Walker, A. B., Martinez, D., & Peter, L. M. (2005). Grain morphology and trapping effects on electron transport in dye-sensitized nanocrystalline solar cells. *The Journal of Physical Chemistry B*, 109(11), 5100-5107.

[6] Kopidakis, N., Benkstein, K. D., van de Lagemaat, J., Frank, A. J., Yuan, Q., & Schiff, E. A. (2006). Temperature dependence of the electron diffusion coefficient in electrolyte-filled TiO 2 nanoparticle films: evidence against multiple trapping in exponential conduction-band tails. *Physical Review B*, 73(4), 045326.

[7] Arkhipov, V. I., Rudenko, A. I., Andriesh, A. M. et al. (1983) *Non-stationary injection currents in disordered solids*. Kishinev.

[8] Zvyagin, I. P. (1984) *Kineticheskie Yavleniya v Neuporyadochennykh Poluprovodnikakh* (*Kinetic Phenomena in Disordered Semiconductors*). Moscow: Izd. MGU.

[9] Nikitenko, V. R. (2011). *Non-stationary Processes of Transport and Recombination of Charge Carriers in Thin Layers of Organic Materials*. MEPhI (2011).

[10] Sibatov, R. T. & Uchaikin, V. V. (2009) Fractional differential approach to dispersive transport in semiconductors, *Physics Uspekhi*, 52, 1019-1043.

[11] Scher, H. & Montroll, E. W. (1975). Anomalous transit-time dispersion in amorphous solids, *Physical Review B*, 12, 2455-2477.

[12] Uchaikin, V. V. & Sibatov, R. T. (2013). *Fractional Kinetics in Solids: Anomalous Charge Transport in Semiconductors, Dielectrics and Nanosystems.* World Scientific.

[13] Weiss, G. H. & Havlin, S. (1986). Some properties of a random walk on a comb structure. *Physica A: Statistical Mechanics and its Applications*, 134(2), 474-482.

[14] Ben-Avraham, D. & Havlin, S. (2000). *Diffusion and Reactions in Fractals and Disordered Systems.* Cambridge: Cambridge University Press.

[15] Nigmatullin, R. R. (1986) The realization of the generalized transfer equation in a medium with fractal geometry, *Physica Status Solidi* (b), 133, 425.

[16] Arkhincheev, V. E. (1991). Anomalous diffusion and drift in the comb model of percolation clusters, *Journal of Experimental and Theoretical Physics*, 100, 292-300.

[17] Baskin, E., & Iomin, A. (2004). Superdiffusion on a comb structure, *Physical Review Letters*, 93(12), 120603.

[18] Lubashevskii, I. A., & Zemlyanov, A. A. (1998). Continuum description of anomalous diffusion on a comb structure, *Journal of Experimental and Theoretical Physics*, 87(4), 700-713.

[19] Bässler, H. (1993). Charge transport in disordered organic photoconductors, *Physica Status Solidi* (b) 175:15-56, 1993.

(M, 2)-Methods of Accuracy of a Maximal Order for Stiff Systems

Eugeny A. Novikov

Department of Computational Mathematics
Institute of Computational Modeling SB RAS
Krasnoyarsk, Russia
e-mail: novikov@icm.krasn.ru

Abstract— We study (m,2)-methods for solving stiff systems where at each step the right-hand side of a system of ordinary differential equations is calculated two times. It is shown that the maximal order of accuracy of an L-stable (m,2)-method is four and a method of the maximal order is constructed.

Keywords— Differential equations, numerical analysis, accuracy, numerical stability, Taylor series, algorithm design and analysis.

I. Introduction

When solving the Cauchy problem for a stiff system of ordinary differential equations, Rosenbrock type methods [1] are widely used due to simple implementation and reasonably good accuracy and stability. Rosenbrock type methods where the same Jacobi matrix is used in calculating each stage are in most common use It is known (see, for example, [2]) that in this case the maximal order of accuracy of the m-stage Rosenbrock method is (m+1), in addition, a scheme of maximal order can be only A-stable. If a maximal order is not required, an L-stable numerical formulae of order m can be constructed. In practical calculations, as a rule, a maximal order is abandoned in favour of L-stability. Notice that a scheme of order higher than two with freezing the Jacobi matrix can not be constructed on the basis of methods of Rosenbrock type [3]. This limits the application of these methods to calculations with moderate accuracy or to problems of a small dimension.

In [4–5] a class of (m,k)-methods where determining a stage does not involve calculation of a right-hand side of a system of differential equations is proposed. Implementation of (m,k)-methods is as simple as that of Rosenbrock methods, however, (m,k)-schemes have advantages for accuracy and stability. In the framework of (m,k)-methods the problem of freezing the Jacoby matrix and its numerical approximation is solved more easily.

In this paper (m,2)-methods for solving stiff systems where at each step the right-hand side of a system of ordinary differential equations is calculated two times are studied. It is shown that the maximal order of accuracy of the L-stable (m,2)-method is four. A method of maximal order is demonstrated.

II. Methods of Rosenbrock Type

We consider the Cauchy problem for a system of ordinary differential equations

$$y' = f(y),\ y(t_0) = y_0,\ t_0 \le t \le t_k, \tag{1}$$

where y and f are N-dimensional real vector functions, t is an independent variable which varies on a given finite interval. It is known that introducing an additional variable one can reduce a nonautonomous system to an autonomous form. Hence, the formulation (1) can be considered without loss of generality. Methods of Rosenbrock type for the problem (1) have the form

$$\begin{aligned} y_{n+1} &= y_n + \sum_{i=1}^{m} p_i k_i, \\ D_n k_i &= hf\left(y_n + \sum_{j=1}^{i-1} \beta_{ij} k_j\right), \\ D_n &= E - ahf'_n, \end{aligned} \tag{2}$$

where h is an integration step, E is the identity matrix, $f'_n = \partial f(y_n)/\partial y$ is the Jacobi matrix of a vector function $f(y)$, a, p_i, $1 \le i \le m$, and β_{ij}, $1 \le i \le m$, $1 \le j \le i-1$, are numerical coefficients. Nowadays methods of Rosenbrock type are treated in a wider sense [2]. The numerical formulae (2) can be obtained from a class of semiexplicit methods of Runge-Kutta type provided that only one iteration step of Newton's method is performed when calculating each stage. In Rosenbrock methods only linear systems of algebraic equations are solved when calculating a stage, whereas in implicit or semiexplicit Runge-Kutta method an iterative process of the Newton type is required that leads to additional problems in its implementation.

III. The Class of (M,K)-methods

A class of (m,k)-methods [4] is introduced as follows. Let integer positive numbers m and k, $k \le m$ be given. Denote the set of integer numbers i, $1 \le i \le m$ by M_m and the subsets of M_m of the form

$$M_k = \{m_i \in M_m \mid 1 = m_1 < \cdots < m_k \le m\},$$
$$J_i = \{m_{j-1} \in M_m \mid j > 1,\ m_j \in M_k,\ m_j \le i\}, \quad (3)$$
$$2 \le i \le m,$$

by M_k and J_i. Then (m,k)-methods can be represented in the form

$$y_{n+1} = y_n + \sum_{i=1}^{m} p_i k_i,$$
$$D_n = E - ahf_n',$$
$$D_n k_i = hf\left(y_n + \sum_{j=1}^{i-1} \beta_{ij} k_j\right) + \sum_{j \in J_i} \alpha_{ij} k_j,\ i \in M_k, \quad (4)$$
$$D_n k_i = k_{i-1} + \sum_{j \in J_i} \alpha_{ij} k_j,\ i \in M_m \setminus M_k$$

The set J_i, $2 \le i \le m$ serves to eliminate redundant coefficients α_{ij} by means of which we can not make effect on accuracy and stability of (4) and which can be expressed linearly in terms of other coefficients. In traditional one-step methods, to determine computational work per integration step a single constant m being the number of stages is sufficient because in these methods each stage is accompanied by obligatory calculation of the right-hand side of (1). In the methods (4) there are two types of stages. For some stages it is necessary to calculate the right-hand side but for other ones this is not required. As a result, to determine computational work per step in (4) two constants m and k are needed. The expense of a step is as follows: the Jacobi matrix is calculated once and the decomposition of a matrix D_n is performed once, function f is calculated k times, backward Gauss is performed m times. For $k=m$ and $\alpha_{ij}=0$ the numerical schemes (4) coincide with the methods (2) of Rosenbrock type. In other cases these methods differ and the methods (4) have advantages over the methods (2).

IV. The Maximal Order of Accuracy of (M,2)-methods

Consider *(m,2)*-methods of the form

$$y_{n+1} = y_n + \sum_{i=1}^{m} p_i k_i,$$
$$D_n = E - ahf_n',$$
$$D_n k_1 = hf(y_n),\ D_n k_i = k_{i-1},\ 2 \le i \le s_1 - 1, \quad (5)$$
$$D_n k_{s_1} = hf\left(y_n + \sum_{j=1}^{s_1-1} \beta_{s_1,j} k_j\right) + \alpha_{s_1,s_1-1} k_{s_1-1},$$
$$D_n k_i = k_{i-1} + \alpha_{i,s_1-1} k_{s_1-1},\ s_1 + 1 \le i \le m,$$

where s_1 and m, $s_1 \le m$ are arbitrary integer constants. It is easy to see that (5) describe all kinds of the *(m,2)*-methods (4).

Theorem. For any number m of stages and for any set M_2 it is impossible to construct an *(m,2)*-method of accuracy of order higher than four.

Without loss of generality we give a proof for a scalar problem (1) whose exact solution $y(t_{n+1})$ can be written in the form

$$y(t_{n+1}) = y(t_n) + hf + \frac{1}{2}h^2 f''f +$$
$$+\frac{1}{6}h^3\left[f'^2 f + f''f^2\right] +$$
$$+\frac{1}{24}h^4\left[f'^3 f + 4ff''f'f + f'''f^3\right] + \quad (6)$$
$$+\frac{1}{120}h^5\left[f'^4 f + 4ff'''f'^3 + 5f'^2 f''f^2 + f''^2 f^3 + f^{IV} f^4\right] +$$
$$+O\left(h^6\right),$$

where elementary differentials are calculated on the exact solution $y(t_n)$. Consider *(m,2)*-methods (5). Taking into account

$$D_n^{-1} = E + ahf_n' + a^2h^2 f_n'^2 + a^3h^3 f_n'^3 +$$
$$+ a^4h^4 f_n'^4 + O\left(h^5\right) \quad (7)$$

we observe that the second calculation of a function *f(y)* is performed at the point

$$y_{n,c} = y_n + \sum_{j=1}^{s_1-1} \beta_{s_1,j} k_j = y_n + \sum_{i=1}^{5} c_i h^i f_n'^{(i-1)} f_n + O\left(h^6\right),$$

where c_i, $1 \le i \le 4$ are determined in terms of the coefficients of the scheme (5) and elementary differentials are calculated on an approximate solution y_n. Taking into consideration (6) and (7), to prove the theorem it is sufficient to show that the Taylor expansion of a function $f(y_{n,c})$ in terms of powers of h does not involve the term $h^5 f_n'' f_n^3$. The Taylor expansion of $f(y_{n,c})$ in the neighbourhood of the point y_n up to terms of order h^5 inclusive has the form

$$f\left(y_{n,c}\right) = hf_n + c_1h^2 f_n' f_n + h^3\left[c_2 f_n'^2 f_n + \frac{1}{2} f_n'' f_n^2\right] +$$
$$+h^4\left[c_3 f_n'^3 f_n + c_1c_2 f_n' f_n'' f_n^2 + \frac{1}{6}c_1^3 f_n''' f_n^3\right] +$$
$$+h^5\left[c_4 f_n'^4 f_n + c_1c_3 f_n'^3 f_n'' f_n^2 + \frac{1}{2}c_1^2c_2 f_n' f_n''' f_n^3 + \frac{1}{24}c_1^4 f_n^{IV} f_n^4\right] +$$
$$+O\left(h^6\right),$$

which completes the proof.

V. An A-stable (M,2)-Method of Order Four

Let $k=2$. Take $M_k=\{1,2\}$, then $J_i=\{1\}$, $2 \le i \le m$. As a result we obtain numerical schemes of the form

$$y_{n+1} = y_n + \sum_{i=1}^{m} p_i k_i,$$
$$D_n k_1 = hf\left(y_n\right), \qquad (8)$$
$$D_n k_2 = hf\left(y_n + \beta_{21} k_1\right) + \alpha_{21} k_1,$$
$$D_n k_i = k_{i-1} + \alpha_{i1} k_1,\ 3 \le i \le m.$$

Case 1. Let m=2. Then the coefficients

$$a = \frac{6+\sqrt{12}}{12},\ p_1 = \frac{76a-3}{54a},\ p_2 = \frac{16}{27},$$
$$\beta_{21} = \frac{3}{4},\ \alpha_{21} = \frac{3-54a}{32a}$$

provide the third order of an *A*-stable scheme (8). Usually the maximal order is abandoned in favour of *L*-stability. For m=2 the coefficients of a second-order *L*-stable scheme have the form

$$a = \beta_{21} = 1 - \frac{\sqrt{2}}{2},\ p_1 = a,$$
$$p_2 = 1 - a,\ \alpha_{21} = 0.$$

Case 2. Let m=3 Introduce the notations:

$$c_1 = 1 - 4a + 2a^2,\ c_2 = \beta_{21} + a\alpha_{21},$$
$$c_3 = 6ac_2\beta_{21}^2,\ c_4 = 2c_2\left(\beta_{21} - a\right),$$
$$c_5 = 3c_1\beta_{21}^2\left(\beta_{21} - a\right).$$

Then the coefficients

$$p_1 = \frac{ac_3 + c_4 - c_5}{c_3},$$
$$p_2 = \frac{2c_2 - 3\beta_{21}^2 c_1}{6\beta_{21}^2 c_2},\ p_3 = \frac{c_1}{2c_2},$$
$$\alpha_{31} = \frac{2c_2\left[\left(1-a\right)c_3 - c_4 + c_5 - 2ac_2\left(1+\alpha_{21}\right)\right]}{c_2 c_3}.$$

provide third-order accuracy of an *L*-stable scheme (8). Here α_{21} and β_{21} are constant coefficients and the value of a is determined from the equation

$$6a^3 - 18a^2 + 9a - 1 = 0.$$

It is easy to verify that for any value of m it is impossible to construct an *L*-stable scheme (8) of order four. To do this, it is sufficient to write conditions of four-order consistency and *L*-stability. The simplest study of the obtained nonlinear system of algebraic equations shows its incompatibility. However, provided *L*-stability is not required, a method (8) of accuracy of order four can be constructed.

Case 3. Let m=4. Introduce the notations:

$$c_1 = \frac{60-81c_7}{81},\ c_2 = \frac{18-324c_7}{81},$$
$$c_3 = \frac{405c_7 - 64}{81},\ c_4 = \frac{19-162c_7}{81},$$
$$c_5 = \frac{64-81c_7}{81},\ c_6 = \frac{162c_7 - 16}{81},$$

where $c_7 \neq 0$ is a constant coefficient. Then the coefficients of a fourth-order method have the form:

$$a = \frac{3}{8},\ \beta_{21} = \frac{3}{4},\ p_1 = c_1,\ p_2 = c_5,\ p_3 = c_6,$$
$$p_4 = c_7,\ \alpha_{21} = \frac{c_4}{c_7},\ \alpha_{31} = \frac{c_3c_7 - c_4c_6}{c_7^2},$$
$$\alpha_{41} = \frac{c_2c_7^2 - c_4c_5c_7 - c_3c_6c_7 + c_4c_6^2}{c_7^2}.$$

VI. An L-stable (M,2)-method of Order Four

Let k=2. Take M_k={1, 3}, then J_i={2}, 3≤i≤4. As a result, we get a scheme of the form

$$y_{n+1} = y_n + \sum_{i=1}^{4} p_i k_i,$$
$$D_n k_1 = hf\left(y_n\right),$$
$$D_n k_2 = k_1, \qquad (9)$$
$$D_n k_3 = hf\left(y_n + \beta_{31}k_1 + \beta_{32}k_2\right) + \alpha_{32}k_2,$$
$$D_n k_4 = k_3 + \alpha_{42}k_2.$$

Introduce the notations

$$c_1 = \frac{76a^2 - 29a + 3}{27a^2},\ c_2 = \frac{-146a^2 + 89a - 12}{27a^2},$$
$$c_3 = \frac{32a-4}{27a},\ c_4 = \frac{72a^2 - 59a + 10}{18a^2},$$
$$c_5 = \frac{4-16a}{27a^2},\ c_6 = \frac{-18a^2 + 19a - 4}{18a^2}.$$

Then the coefficients of an *L*-stable method of order four have the form

$$p_1 = c_1,\ p_2 = c_2,\ p_3 = c_3,\ p_4 = c_5,$$
$$\alpha_{32} = \frac{c_6}{c_5},\ \alpha_{42} = \frac{c_4c_5 - c_3c_6}{c_5^2},$$
$$\beta_{31} = \frac{48a-9}{32a},\ \beta_{32} = \frac{9-24a}{32a},$$

where a satisfies the *L*-stability equation

$$24a^4 - 96a^3 + 72a^2 - 16a + 1 = 0.$$

This equation has the roots

$$a_1 = 0.10643879214266,\ a_2 = 0.22042841025921,$$
$$a_3 = 0.57281606248213,\ a_4 = 3.10031673511599.$$

In calculations it is wise to take a=0.57281606248213, and the coefficients of (9) have the form

$$p_1 = +1.27836939012447,\quad p_2 = -1.00738680980438,$$
$$p_3 = +0.92655391093950,\ p_4 = -0.33396131834691,$$
$$\beta_{31} = +1.00900469029922,\quad \beta_{32} = -0.25900469029921,$$
$$\alpha_{32} = -0.49552206416578,\quad \alpha_{42} = -1.28777648233922.$$

VII. Conclusions

From the above considerations we can make the following conclusions.

First, stability of (m,k)-methods depends on a choice of the sets (3), i.e., on a way of implementation of numerical schemes. This follows from a comparison of the formulae (8) and (9).

Second, a (4,2)-method which is competitive in accuracy with the implicit method of Runge-Kutta type with two calculations of the right-hand side of (1) can be constructed. In linear analysis of stability this method is also as good as the implicit method of Runge-Kutta type. In addition, in (9) a way of implementation is included, i.e., one can estimate computational work per integration step before calculations. Computational work of the implicit methods of Runge-Kutta type depends highly on a way of implementation. The use of two-stage scheme does not imply that at each step the right-hand side of (1) is calculated twice. Therefore for some problems the (4,2)-method is preferred over implicit numerical formulae of Runge-Kutta type.

Third, when a function f(y) of the problem (1) is calculated twice, an L-stable (4,2)-method of fourth-order accuracy can be constructed whereas corresponding L-stable method (2) of Rosenbrock type can be of order two only. For sufficiently high accuracy of calculations and large dimension of the problem (1), decomposition of the Jacobi matrix is responsible in fact for total computational work whereas impact of backward Gauss is unessential. In this case (4,2)-method is more efficient.

Notice that in the framework of (m,k)-methods the problem of freezing the Jacoby matrix can be solved easily [6].

Acknowledgment

This work is partially supported by Russian Foundation of Fundamental Researches (grants 14-01-00047).

References

[1] H.H. Rosenbrock, “Some general implicit processes for the numerical solution of differential equations,” Computer, vol. 5, pp. 329–330, 1963.
[2] E. Hairer and G. Wanner, Solving Ordinary Differential Equations, Vol. 2: Stiff and Differential–Algebraic Problems: monograph, 2nd ed., Berlin, Germany: Springer–Verlag, 1996, Moscow, Russia: Mir, 1999.
[3] E.A. Novikov, “Freezing of the Jacobi matrix in the Rosenbrock type method of the second order accuracy, “ E.A. Novikov, V.A. Novikov, L.A. Umatova, in proc. BAIL–IV Conf., Bool Press, pp. 380–386, 1986.
[4] E.A. Novikov, “On a class of one-step iteration-free methods for solving stiff systems,” E.A. Novikov, Urgent problems of computational and applied mathematics, Novosibirsk, Russia, pp. 138–139, 1987 (in Russian).
[5] E.A. Novikov, “One-step iteration-free methods of solving stiff systems,” E.A. Novikov, Yu.A. Shitov and Yu.I. Shokin, Soviet Math. Dokl., vol. 38, no.1, pp. 212–216, 1989.
[6] E.A. Novikov, “Approximation of the Jacobi matrix in (m,3)-methods for stiff systems,” E.A. Novikov, A.L. Dvinsky, Siberian Journal of Numerical Mathematics, vol..11, no. 3, pp. 283–295, 2009 (in Russian).

A macroeconomic model of consumption and investment spending: An econometric application for the economy of Cyprus.

Panayiotis L. Diacos

Abstract—A simple macroeconomic model of consumption and investment spending is specified and estimated, using time series data from the economy of Cyprus. The parameter estimates are found accurate and with plausible values. The model's dynamic behaviour seems realistic and its forecasting ability satisfactory.

***Keywords*—**Dynamic forecasts, full information maximum likelihood method, macroeconomic models, simultaneous equation systems, single period forecasts..

I. Introduction

THe main aim of this work is the development of a simple econometric model that can be employed for the study of aggregate consumption and aggregate investment decisions for the economy of Cyprus.

The bibliography is full of examples of economic studies where both aggregate consumption and aggregate investment behaviour are examined together. A theoretical investigation of consumption - investment decisions is often found in small economic models that are concerned with trade-cycle theory ([13], [11], [12], [8]) or in models that relate to macroeconomic growth ([14], [2]). An empirical analysis of such aggregates can be found in econometric models that focus on specific sectors in the economy or relate to the whole macroeconomic system. In most cases, such frameworks of analysis are relatively large with respect to the number of behavioural equations (and endogenous variables) which they include ([9], [10], [6], [3], [7]).

The econometric model developed in this paper assumes that the main determinants of aggregate income are aggregate consumption and investment spending. These variables, expressed in their first differences, are combined in a small macroeconomic system whose parameters are estimated by means of the Full Information Maximum Likelihood (FIML) method, using annual time series data over the period 1960-1996. The model is also employed for the production of two different types of forecasts.

The proposed model is the first econometric work for the macro-economy in Cyprus in which consumption and investment are analysed together. It can be regarded as an extension of an earlier model proposed by [4] where aggregate consumption spending and national income are studied simultaneously in logarithmic deviation form. Comparisons between the estimates and between the forecasts from the two cases will be provided in the sections that follow.

The plan of this paper is the following: In section 2 the model is specified. In section 3 a brief description of the econometric method is provided. Also in that section the reduced form of the model is derived. Section 4 relates to the data and section 5 presents the estimates. Section 6 summarises the forecasts and section 7 gives the conclusion.

II. The Model

Let us suppose that aggregate consumption and investment plans in the economy are determined by the simultaneous equation system,

$$\begin{aligned} \Delta C_t &= a1 .\Delta Y_t + a2 .\Delta Y_{t-1} \\ \Delta I_t &= b1 .\Delta C_t + b2 .\Delta C_{t-1} + b3 .\Delta I_{t-1} \\ \Delta Y_t &= c1 .\Delta Y_{t-1} + c2 .\Delta Y_{t-2} \end{aligned} \qquad (1)$$

where C_t is planned aggregate real consumption expenditure over period t,
I_t is planned aggregate real investment expenditure over period t,
Y_t is aggregate real income over period t,
$a1,\ a2,\ b1,\ b2,\ b3,\ c1$ and $c2$ are constant parameters,
$\Delta( )$ denotes the difference operator and
$t = 1,\ 2,\ 3 \ldots\ T$, denotes the time period.

In the above system the variables ΔC_t, ΔI_t and ΔY_t will be assumed to be endogenous while all the lagged variables will be assumed to be predetermined. In this way the model is properly identified and can be estimated by an appropriate econometric technique.

ISBN: 978-1-61804-251-4

Let us now consider the equations in a little more detail. The first equation is a consumption function showing that changes in current consumption depend on changes in current income and on changes in income one-period earlier. The second equation is an investment function indicating that changes in investment this period are affected by current and one-period lagged changes in consumption and also by one-period lagged changes in investment[1]. The last equation shows that changes in the level of income in any period could be related to past changes one period and two periods earlier.

If we ignore that the variables are in first difference, we observe that the first two equations of the model are similar, but dynamically more flexible, to those used in the common trade-cycle models. [13] and [8], for example, use a simple Keynesian consumption function with a Robertsonian lag. In the present case the function is extended to include also the current income variable. The investment function is also similar to that used by Samuelson, but with the assumption that $b1$ and $b2$ are not necessarily equal. In addition, the investment function in this model allows for a possible direct influence of last period's investment on current investment. Finally, the income identity constraint (which is commonly employed in trade-cycle and long run growth models) is replaced by the income autoregressive in order to allow for a more realistic representation for the movements in the available data set.

III. Estimation Method and Reduced Form

The parameters of the linear system (1) can be estimated, using the Full Information Maximum Likelihood (FIML) method. This method has more desirable asymptotic properties, compared to other system techniques[2]. Rearranging the system in stochastic form, we get,

$$\mathbf{YA} = \mathbf{XB} + \mathbf{U} \tag{2}$$

where $\mathbf{Y}$ is a T × N matrix of observable current endogenous random variables,
$\mathbf{X}$ is a T × K matrix of predetermined variables,
$\mathbf{U}$ is a T × N matrix of disturbances,
$\mathbf{A}$ and $\mathbf{B}$ are N × N and K × N matrices of unknown parameters that we wish to estimate,
T is the number of observations available,
N is the number of endogenous variables and
K is the number of predetermined variables.

It is important to assume that the vectors corresponding to the columns of $\mathbf{U}$, follow a multivariate normal distribution and for each value of t the matrix has mean zero and an unknown variance $\mathbf{\Sigma}$. In addition, the elements in each column vector are assumed serially independent.

Under the above conditions, the logarithmic likelihood function, $\ln L$, can be expressed as

$$\ln L = -(NT/2).\ln(2\pi) + T.\ln\|\mathbf{A}\| - (T/2).\ln|\mathbf{\Sigma}| - (1/2)\mathrm{tr}[\Sigma^{-1}(\mathbf{YA}-\mathbf{XB})''(\mathbf{YA}-\mathbf{XB})] \tag{3}$$

where $||A||$ is the absolute value of the determinant of $\mathbf{A}$ and tr[ ] denotes the trace.
Setting $(d\ln L/d\Sigma)$ equal to zero, it is possible to derive the expression

$$\mathbf{\Sigma} = T^{-1}(\mathbf{YA}-\mathbf{XB})'(\mathbf{YA}-\mathbf{XB}) \tag{4}$$

Substitution in (3), gives the concentrated log-likelihood, $\ln L^*$, given as

$$\ln L^* = -(T/2).\{\ln|T^{-1}(\mathbf{Y}-\mathbf{XBA}^{-1})''.(\mathbf{Y}-\mathbf{XBA}^{-1})|\} + \text{constant} \tag{5}$$

Maximising the last function with respect to $\mathbf{A}$ and $\mathbf{B}$, gives the FIML estimates $\mathbf{A}^*$ and $\mathbf{B}^*$. Substituting these into (4), gives $\mathbf{\Sigma}^*$, the FIML estimate of the variance.

Also form (2) can be simplified as

$$\mathbf{y(t)'A = x(t)'B + u(t)'} \tag{6}$$

where **y(t)**, **x(t)**, **u(t)** are of order N×1, K×1 and N×1 respectively and have elements vectors that correspond to the columns of **Y**, **X** and **U**.

Since the higher lag order in the predetermined variables is 2, the system can be transformed to

$$\mathbf{A'y(t) - B_a\, y(t-1) - B_b\, y(t-2) = u(t)} \tag{7}$$

where $\mathbf{B_a}$ and $\mathbf{B_b}$ are both square matrices of order N×N. We can attain equality in the order of these two matrices, by adding zeros to the elements that correspond to missing lagged endogenous.

Moreover, equation (7) could be expressed as,

$$\mathbf{B_0\, y(t) + B_1\, y(t-1) + B_2\, y(t-2) = u(t)} \tag{8}$$

where $\mathbf{B_0 = A'}$, $\mathbf{B_1 = -B_a}$ and $\mathbf{B_2 = -B_b}$.

Therefore the system has a reduced form,

$$\mathbf{y(t) = \Pi_1\, y(t-1) + \Pi_2\, y(t-2) + v(t)} \tag{9}$$

where $\mathbf{\Pi_1 = (A^{-1})'B_a}$, $\mathbf{\Pi_2 = (A^{-1})'B_b}$ and $\mathbf{v(t) = (A^{-1})'u(t)}$.

Also, its characteristic equation can be expressed as

[1] Over the relevant sample period, the Central Bank of Cyprus adopted a fixed interest rate policy. Thus the effect of interest rate variations on investment can be ignored in this paper.
[2] See [1], ch. 7.

$$| \mathbf{B_0}\lambda^2 + \mathbf{B_1}\lambda + \mathbf{B_2} | = \mathbf{0}. \quad (10)$$

IV. The Data

The parameters of the model are estimated by means of annual time series data from the economy in Cyprus. The data, obtained from government statistical publications, relates to the period 1960-2000. Estimation is carried out over the sample period 1960-1996 while four observations for the period 1997-2000 are preserved for each variable for forecasting purposes[3].

Data for the aggregate consumption variable, C_t, is obtained from the available published time series at constant 1980 prices[4].

Data for investment, I_t, is obtained from the available gross capital formation series at constant 1980 prices.

In the case of aggregate income, Y_t, data is obtained from the available series of GDP, also at constant 1980 prices.

Both the consumption and income series are expressed in terms of per capita values, using the published series of annual population figures.

Finally, for estimation, the three data series employed in the model are transformed to their first difference.

V. The Estimates

The FIML estimates of the model for the period 1960-1996, are presented in table I, below. We notice that all the estimates are significant at a level less than 10 percent.

The value of $a1$ is highly significant and shows that a unit rise in ΔY_t causes a rise in ΔC_t by 0.5441 units. The estimate in this model is higher than the marginal propensity estimate found in [4]. In that case the short-run MPC was found to lie in the range 0.40 to 0.47.

A positive value is also obtained for $a2$, the sensitivity of changes in consumption with respect to changes in income a period earlier. More specifically, a *ceteris paribus* rise in ΔY_{t-1} by one unit, makes ΔC_t to rise by 0.1475 units.

The values of $b1$ and $b2$ are also significant and both positive, showing a direct relation between changes in investment and changes in consumption either in the current or the previous period. More specifically, a unit change in ΔC_t causes ΔI_t to change by around 151 units, while a unit change in ΔC_{t-1} causes ΔI_t to change by around 209 units. The value of $b3$ is negative, showing that a unit rise in ΔI_{t-1} is followed by a fall in ΔI_t by about 0.44 units.

Finally, we observe that ΔY_t can also be related to its lagged values. A unit changes in ΔY_{t-1} causes a change in ΔY_t by 0.2479 units while a unit change in ΔY_{t-2} causes a change in ΔY_t by 0.3618 units.

Using the characteristic equation of the system, we derive the relevant latent roots which are shown in the last row in table 1. The first root, λ_1, corresponds to $b3$ and therefore relates to the investment function. The other two roots, λ_2 and λ_3, are common to the consumption and income equations. All the roots have modulus less than 1 and therefore the system is stable[5].

The estimates presented in table I, may not be so easy to use for policy making purposes since these are just measures of sensitivity for the variables. Elasticity measures on the other hand can be found more useful as they denote percentage changes. Such elasticity measures are derived using the estimated sensitivities and the structural form relations. Their values and standard errors are presented in table II.

Considering, for example, the income elasticity of consumption, $g_{y_t}^{c_t}$, we notice that this is positive and less than one. This is what it might be expected since this denotes the ratio of the marginal to the average propensity to consume. It shows that a unit percentage rise in real per capita income would cause real per capita consumption to rise by 0.8118 percent, over the same period[6].

The consumption elasticity of investment, $g_{c_t}^{I_t}$, is also positive indicating the positive relation between the two variables. A unit percentage rise in real per capita consumption would account for a 0.5798 percent rise in real investment over the same period.

[3] In [4] the model is estimated over the period 1960-1995 and excludes the 1975 observation. Thus the sample size for this project is larger only by two data observations. Using almost the same sample size, could allow us to make more accurate comparisons on the estimates and forecasts of the two models.

[4] All the data series can be collected from the following statistical booklets, published by the
Statistical Service of the Ministry of Finance in Cyprus:
-Economic Report 1995 & 1996,
-Historical Data on the Economy of Cyprus 1960 -1991,
-National Accounts 1999,
-National Accounts 2000.

[5] The value of R-sqr provides a measure of fit of the model and is statistically analogous to the square correlation coefficient of single equation regressions. In addition, a vector autocorrelation test of the first order conducted, showed no evidence of significant serial correlation. This is verified by the small value of the LR statistics.

[6] Notice that the values in table 2 refer to percentage changes of variables in levels whereas those in table 1 refer to unit changes of variables in differences.

ISBN: 978-1-61804-251-4

Table I. Structural form estimates

parameter	estimate	st. error	t-value	significance
$a1$	0.5441	0.0840	6.4731	0.000
$a2$	0.1475	0.0654	2.2564	0.030
$b1$	150.9100	80.5355	1.8738	0.069
$b2$	208.9730	89.5764	2.3329	0.025
$b3$	-0.4443	0.1293	-3.4353	0.002
$c1$	0.2479	0.1165	2.1287	0.040
$c2$	0.3618	0.1293	2.7980	0.008
max lnL = 95.0569, T = 34, λ_1 = -0.4443, λ_2 = 0.7381, λ_3 = -0.4901, L R (5) = 1.2258, R-sqr = 0.4987				

Table II. Derived consumption, investment and income elasticities

elasticity	value	st. error	elasticity	value	st. error
$g_{y_t}^{c_t}$	0.8118	0.1253	$g_{c_{t-2}}^{I_t}$	-0.6158	0.2640
$g_{y_{t-1}}^{c_t}$	-0.5434	0.1404	$g_{I_{t-1}}^{I_t}$	0.4869	0.1133
$g_{y_{t-2}}^{c_t}$	-0.1890	0.0838	$g_{I_{t-2}}^{I_t}$	0.2523	0.0734
$g_{c_{t-1}}^{c_t}$	0.9525	0.0000	$g_{y_{t-1}}^{y_t}$	1.1460	0.1070
$g_{c_t}^{I_t}$	0.5798	0.3094	$g_{y_{t-2}}^{y_t}$	0.0978	0.7213*
$g_{c_{t-1}}^{I_t}$	0.2022	0.4596*	$g_{y_{t-3}}^{y_t}$	-0.2686	0.0960

* Not significant

We notice that all the elasticity measures presented are significant (at a level less than 10 percent), apart from those for $g_{c_{t-1}}^{I_t}$ and for $g_{y_{t-2}}^{y_t}$.

Next, let us consider the long-run parameters of the model. The analysis will be restricted only to changes that arise from an initial change in income. In this model, such a change will not stop at the first period, but it will continue to have an influence on the system in subsequent periods too. An analogous treatment with this particular case can be found in [5].

In the cases of initial changes in consumption or investment, the long-run parameters can be determined in a way very similar to that which relates to a change in a variable of the system which is strictly exogenous.

The estimates and the standard errors are presented in table III.

The first parameter, $e_{c^* y}$, shows that a unit change in real income, from an equilibrium level Y^*, causes the equilibrium value of real consumption, C^*, to change by 1.2279 units over the long-run. Expressed as an elasticity, a unit percentage change in Y^*, causes C^* to change by 1.83 percent over the long-run.

The second parameter, $e_{I^* y}$, shows that an initial change in real income by one unit, makes the equilibrium value of real investment, I^* to change by 305.9623 units. Alternatively, a unit percentage change in Y^*, causes I^* to change by 1.84 percent over the long-run.

Finally, the income parameter, shows that an initial change in real income by one unit, causes the equilibrium value of

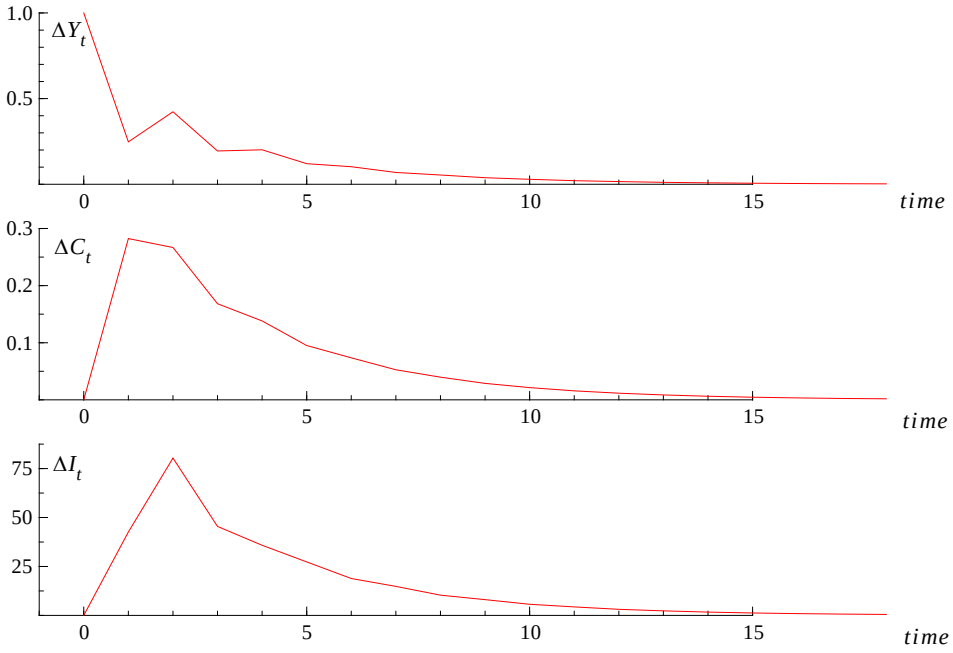
ISBN: 978-1-61804-251-4

Table III. Long-run parameters

parameter	sensitivity	st error	elasticity*	st error
$e_{c^* y}$	1.2279	0.5263	1.8322	0.7853
$e_{I^* y}$	305.9623	131.1510	1.8416	0.7894
$e_{y^* y}$	2.5622	1.0983	2.5622	1.0983

***** The elasticities are normalised on nearest value to sample means

Figure 1. Time plots of endogenous variables.



that variable to change by 2.5622 units over the long-run. Expressed as an elasticity, it shows that an initial unit percentage rise, makes that variable to change from its equilibrium by 2.56 percent. In [4], the logarithmic deviation model gives an elasticity value of about 2.88 percent.

It is also possible to demonstrate graphically how the time paths of the variables, change through time. For that purpose, at a starting period zero, ΔY_t can be assigned a value equal to 1 unit. That initial change will have an effect on the values of the endogenous variables (including ΔY_t) one period later, then a new effect two periods later and so on.

The plots, shown in figure 1, indicate that the change in ΔY_t continues through time but at smaller values, until it converges to zero. Changes in ΔC_t and also in ΔI_t, build up immediately after the initial distortion, reach a maximum and then start falling and eventually diminish.

The maximum for ΔC_t occurs with a delay of one period while that for ΔI_t occurs after a delay of two periods.

The results of course, are in agreement with the postulates of economic theory that require both consumption and investment spending to intensify after a significant upturn in economic activity.

ISBN: 978-1-61804-251-4

VI. FORECASTING

The model is also employed for the production of forecasts over the post-sample period 1997-2000. The forecasted values are then compared with the actual observations that have been preserved and the forecasting errors are derived. The results are presented in tables IV and V. So as to enable comparisons with the log-deviation model, both the forecasts and the actual values are expressed in logarithms.

Table IV. Single period forecasts for the period 1997-2000.

variable	year	actual value	forecast	error
$\ln c_t$				
	1997	0.6584	0.6486	-0.0098
	1998	0.7369	0.6664	-0.0705
	1999	0.7461	0.7555	0.0094
	2000	0.7941	0.7692	-0.0249
rmse = 0.0380				
$\ln I_t$				
	1997	6.1070	6.2192	0.1121
	1998	6.2162	6.1748	-0.0414
	1999	6.2096	6.2474	0.0378
	2000	6.2849	6.2352	-0.0496
rmse = 0.0674				
$\ln y_t$				
	1997	1.0097	1.0137	0.0040
	1998	1.0500	1.0165	-0.0336
	1999	1.0831	1.0644	-0.0187
	2000	1.1220	1.1048	-0.0172
rmse = 0.0211				

Table V. Dynamic forecasts for the period 1997-2000.

variable	year	actual value	forecast	error
$\ln c_t$				
	1997	0.6584	0.6486	-0.0098
	1998	0.7369	0.6583	-0.0786
	1999	0.7461	0.6663	-0.0798
	2000	0.7941	0.6717	-0.1179
rmse = 0.0815				
$\ln I_t$				
	1997	6.1070	6.2192	0.1121
	1998	6.2162	6.2277	0.0115
	1999	6.2096	6.2362	0.0266
	2000	6.2849	6.2421	-0.0427
rmse=0.0617				
$\ln y_t$				
	1997	1.0097	1.0137	0.0040
	1998	1.0500	1.0215	-0.0286
	1999	1.0831	1.0296	-0.0535
	2000	1.1220	1.0343	-0.0877
rmse=0.0533				

ISBN: 978-1-61804-251-4

Table VI. Root mean square errors of the forecasts

variable	single-period forecasts*	dynamic forecasts	trend forecasts
$\ln c_t$	0.0380	0.0815	0.1664
$\ln I_t$	0.0674	0.0617	0.1401
$\ln y_t$	0.0211	0.0533	0.1739

* Figures in parenthesis indicate the rmse values of the log-deviation model

Single period forecasts, listed in table IV, give the forecasted value for a variable in a specific period in the post-sample, by incorporating each time actual values of the variables from the earlier period. We observe that the errors are quite low for all the three variables. In absolute terms, the error for consumption ranges from around 0.01 to 0.07 while for investment it ranges from around 0.04 to 0.11. In the case of income the error is even lower ranging from around 0.004 to 0.034.

Dynamic forecasts are listed in table V. In this category, the forecasting procedure is initiated with some known past values and generates forecasts of the variables for a number of periods ahead. In absolute terms, the errors fall approximately in the ranges 0.01 to 0.12. for consumption, 0.01 to 0.11 for investment and 0.004 to 0.09 for income.

A better assessment of the model's predictive power can be made from the use of the root mean square errors (rmse) of the generated forecasts. These are listed in table VI together with the rmse that are obtained from naïve trend projections and relate to the same post-sample period. The values in the brackets below the consumption and income variables correspond to the rmse of the forecasts of the log-deviation model[7].

Considering first single period forecasts, we notice that the lower error of 0.0211 corresponds to the income forecasts, then comes the error of 0.0380 for consumption forecasts and last follows the error of 0.0674 for investment forecasts.

In the case of dynamic forecasts, income has again the lowest error value of 0.0533, then comes investment with 0.0617 and finally consumption with 0.0815. Therefore, single period income and consumption forecasts are superior to the corresponding dynamic forecasts, but single period

investment forecasts are slightly inferior to the relevant dynamic forecasts.

Both single period and dynamic forecasts are found superior to naïve trend forecasts for all the variables. In addition, the forecasts of the proposed model appear superior to those derived from the log-deviation model. In three out of the four comparisons the current model has lower errors. The exception is the error that relates to the single period forecast of consumption which is found marginally higher than that which corresponds to the log-deviation case.

VII. Conclusion

A simple model of the macro-economy of Cyprus was specified and estimated. The estimates were found highly significant and in agreement with the main principles of economic theory. Also some dynamic properties of the model were analysed and were found realistic.

An additional test of the validity of this work is the accuracy of the derived forecasts. In all the cases these were found to outperform the relevant naïve trend forecasts. Also in most of the comparisons these forecasts were found superior to those that have been derived by the log-deviation model for the consumption - income relations.

The results suggest that the model could be employed for economic analysis and also it could be considered as a good basis for further research.

[7] The post-sample period for the log-deviation model ranges from 1996 to 2000. Thus, it includes an additional forecast value, that of the year 1996.

ISBN: 978-1-61804-251-4

REFERENCES

[1] T. Amemiya, *Advanced Econometrics*. Cambridge, Massachusetts: Harvard University Press, 1985.

[2] A. R. Bergstrom, "Monetary Phenomena and Economic Growth: A Synthesis of Neoclassical and Keynesian Theories", in Bergstrom, A.R. (ed), *Continuous Time Econometric Modelling*. Oxford University Press, 1990.

[3] C. Christ, *Econometric Models and Methods*. New York: Wiley, 1968.

[4] P. Diacos, "A Simple Model of Consumption Decisions in Cyprus". *The XII International Conference on Economic Cybernetics: Sustainable Development Models for European Union Extension Process.* Bucharest, Romania, Nov 2-4, 2006.

[5] P. Diacos, & S. Hadjidakis,. "An Econometric Study of the Beef Meat Sector in Cyprus". *Ekonomia*, vol. 8, (no. 2), pp.210-244, 2005.

[6] J. S. Duesenberry, G. Fromm, L. R. Klein, & E. Kuh,. *The Brookings Quarterly Econometric Model of the United States*. Chicago: Rand McNally, 1965.

[7] G. Fromm & L. R. Klein, "A Comparison of Eleven Econometric Models of the United States". *American Economic Review*, vol. 63(2), pp. 385-93, 1973.

[8] J. R. Hicks. *A Contribution to the Theory of the Trade Cycle*. Oxford: Clarendon, 1950.

[9] L.R. Klein, L. R. *Economic Fluctuations in the United States, 1921-1941*. Cowles Commission Monograph 11. New York: Wiley, 1950.

[10] L.R. Klein, & A.S. Goldberger. *An Econometric Model of the United States,1929 - 1952*. Amsterdam: North-Holland, 1955.

[11] L.A. Metzler. "The Nature and Stability of Inventory Cycles". *The Review of Economics and Statistics*, vol. 23(3), pp. 113-129, 1941.

[12] H.P. Minsky. "A Linear Model of Cyclical Growth". *The Review of Economics and Statistics*, vol.41, pp. 133-145, 1959.

[13] P.A. Samuelson. "Interactions Between the Multiplier Analysis and the Principle of Acceleration". *The Review of Economics and Statistics*, vol. 21(2), pp. 75-78, 1939.

[14] J.K. Whitaker. "Neoclassical Economic Growth and the Consumption Function". *Oxford Economic Papers*, vol. 22(3), pp. 311-337, 1970.

Method of unbalanced power minimization in three-phase systems

Korovkin Nikolay[1], Quang Sy Vu[1], Yazenin Roman[1], Frolov Oleg[2], Silin Nikolay[3]
[1]St. Petersburg State Polytechnical University
[2]Joint Stock Company «Scientific and Technical Center of Unified Power System»
St. Petersburg, Russia
[3]Far Eastern Federal University
Vladivostok, Russia

***Abstract*—**The method of unbalanced power minimization in loaded nodes of three-phase AC systems with insignificant phase unbalance has been proposed. The method is based on innovative approach allowing receiving analytical expressions for unbalanced power in three-phase systems of arbitrary complexity. The method is focused on the application of modern information technologies for power grid management.

***Keywords*—** unbalance, optimization, steady-state mode, vibrations, electric energy quality, AC networks.

I. Introduction

ONE of the topical issues of power distribution AC networks of medium and low voltage is the phase unbalance of currents and voltages. A typical cause of phase unbalance is non-uniform distribution of loads per phase. It is rather frequently that it is impossible to achieve a uniform distribution of loads even at the stage of working design of the object power supply due to the multitude of household electric devices used under intermittent conditions.

Consequences of electrical imbalance may be negative both for single-phase and three-phase electrical receivers. The most critical consequences may occur when the neutral wire of network feeder is broken. In this case, the voltage on single-phase loads is distributed proportionally with their resistances. That is to say, that at the less loaded phase the voltage increases (by 1, 5 times in practice). In most cases, it may be the cause of ignition. One more important consequence of phase unbalance is the double frequency appeared in the network with phase unbalance. The unbalanced power (P_v) generates pulsations of electromagnetic torque on synchronous generator shafts in isolated power systems. The pulsations generate an adverse effect on the stability of power system operation, on electric power quality indices in loaded nodes and reduce essentially the service time of synchronous machines.

Unbalanced power (P_v) is composed of variable components of active power consumed per phase ($P_{v,A}$, $P_{v,B}$, $P_{v,C}$,).

$$p(t)=p_A(t)+p_B(t)+p_C(t)=P_c(t)+P_v(t)=$$
$$=\underbrace{U_A I_A\cos(\varphi_{uA}-\varphi_{iA})+U_B I_B\cos(\varphi_{uB}-\varphi_{iB})+U_C I_C\cos(\varphi_{uC}-\varphi_{iC})}_{P_c(t)}-$$
$$-\underbrace{U_A I_A\cos\left(2\omega t-(\varphi_{uA}-\varphi_{iA})\right)}_{P_{v,A}(t)}-\underbrace{U_B I_B\cos\left(2\omega t-\frac{2\pi}{3}-(\varphi_{uB}-\varphi_{iB})\right)}_{P_{v,B}(t)}-$$
$$-\underbrace{U_C I_C\cos\left(2\omega t-\frac{4\pi}{3}-(\varphi_{uC}-\varphi_{iC})\right)}_{P_{v,C}(t)};$$

$$P_v(t)=P_{v,A}(t)+P_{v,B}(t)+P_{v,C}(t)=P_v\cos(2\omega t+\varphi) \quad (1)$$

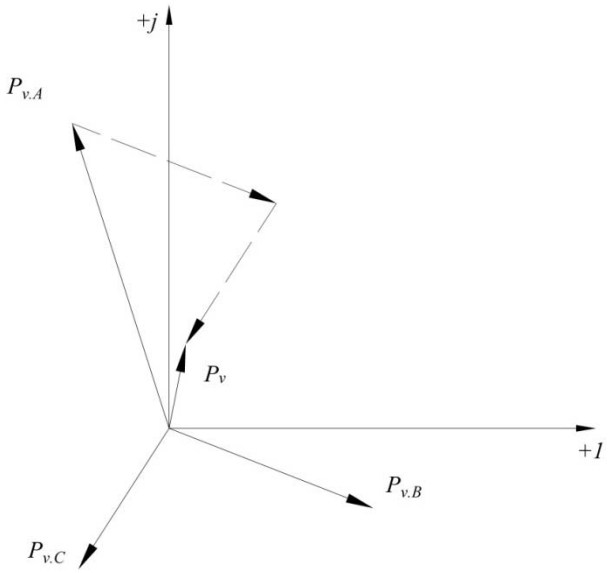


Fig. 1. Clock diagram of variable components of unbalanced power

A general approach to eliminate phase unbalance of a load requires the use of a step-down transformer with balancer set (BS) at the stage of object connecting to electrical grids or a three-phase balance-unbalance transformer. Disadvantages of this approach are as follows:

- high cost of step-down transformers with BS as compared to usual transformers;
- reduction of electric power supply reliability due to the complication of power supply circuit;

ISBN: 978-1-61804-251-4

- Increase of operating expenses.

As far as it concerns the small energy, the development of new types of devices performing the symmetrization of three-phase circuit in automatic mode is of interest now. The paper proposes a new method of minimization of unbalanced power. The method does not use balance-unbalance transformers or step-down transformers with balancer sets.

II. Minimization of Unbalanced Power

The use of devices with controllable reactivity (modification of FACTS devices) makes the basis of a proposer method. The above devices shall be installed in series into load phases [1]. In particular, modern FACTS devices and reactive shunts have a high response and a sampling interval of reactance change applicable for studied objectives.

The implementation of the method supposes the installation of one or several devices with controllable reactivity x (controllable reactive shunts) into the phase of one or several load arms in series with power receiver. The problem of optimal control synthesis may be reduced to the search of the relationship $P_{V(}x)$ between the value of unbalanced power P_v and the reactance of variable shunt x and to the search of the minimum of function $P_v(x)$. The value of shunt resistance corresponding to the minimum of the function shall be the solution of studied problem. Such problems are considered as the inverse class problems of electric circuit theory [2].

In previous paper [3], the authors received and showed the fractional and polynomial relationship between the parameters of steady-state operating conditions of power system and those of its equivalent circuit. Particularly, the dependence of the voltage in any node of studied circuit on the shunt resistance x is as follows:

$$\dot{U}(x) = \frac{a + bx}{1 + \alpha x}, \qquad (2)$$

where a, b, α – are complex constants. Here, the constants a and b are different for voltages of different nodes (arms) while the constant α is the same for all voltages. To find out these constants it shall be necessary to compose and to solve the system of three equations related to unknown constants. To do this, it is necessary to change three times the value x and to calculate the steady-state conditions of modified circuit.

The dependence of voltages in power system nodes on resistances x and y of two shunts is as follows:

$$\dot{U}(x, y) = \frac{a + bx + cy + dxy}{1 + \alpha x + \beta y + \gamma xy} \qquad (3)$$

where a, b, c, d, α, β, γ – are complex constants. Here, the constants a, b, c and d are different for voltages of different nodes (arms), while the constants α, β and γ are invariable for all voltages. These constants may be defined by solving a system of seven equations. However, the procedure of constant searching may be reduced to the composition and solution of several systems of third order equations.

It is evident that the relationship of any of currents on shunt resistance is similar. Upon receiving analytical expressions (2) or (3) for all currents and voltages, from (1) may be received an analytical expression for unbalanced power. Due to awkwardness of these expressions and to the simplicity of obtaining their solution, we shall not cite them here.

We shall consider further on the electric network shown on fig. 2. Loads 1-4 are three-phase, unbalanced. Let the shunt with changing resistance x be installed in series in the phase A of load 1. Unbalanced powers of loads 1-4 shall be marked as $P_{v1} - P_{v4}$, then the function is $P_{V(}x) = P_{v1}(x) + P_{v2}(x) + P_{v3}(x) + P_{v4}(x)$. Analytical expressions for relationships $P_v(x)$ and P_{v1} are obtained on the basis of (2). Dependency diagrams are given on fig. 3 and fig. 4.

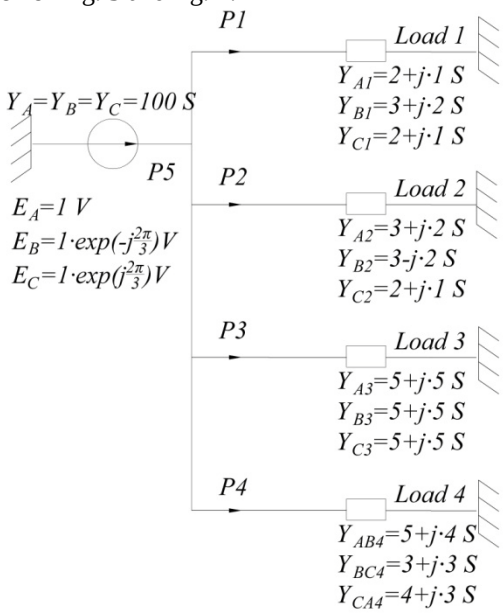


Fig. 2. Diagram of studied circuit.

The search of the minimum of function $P_{v5}(x)$ is quite simple. It may be seen from the diagram that for unbalanced conditions of studied electric circuit there is such a value of reactive shunt resistance x which allows reducing of the value of unbalanced power by 12% in feeding center. We must note that we failed to reduce P_{v5} to zero.

As can be seen from the diagram from fig. 4 – the accuracy of criterion function allows to use it for solving optimization problems.

Now let us consider the problem with two reactive shunts, one of which is installed in phase A of the load 1 and another 1 – between phases A and B of studied circuit. The dependence of the value of unbalanced power $P_v(x,y) = P_{v1}(x,y) + P_{v2}(x,y) + P_{v3}(x,y) + P_{v4}(x,y)$ on values x and y of these shunts is shown on fig. 5 and fig. 6.

ISBN: 978-1-61804-251-4

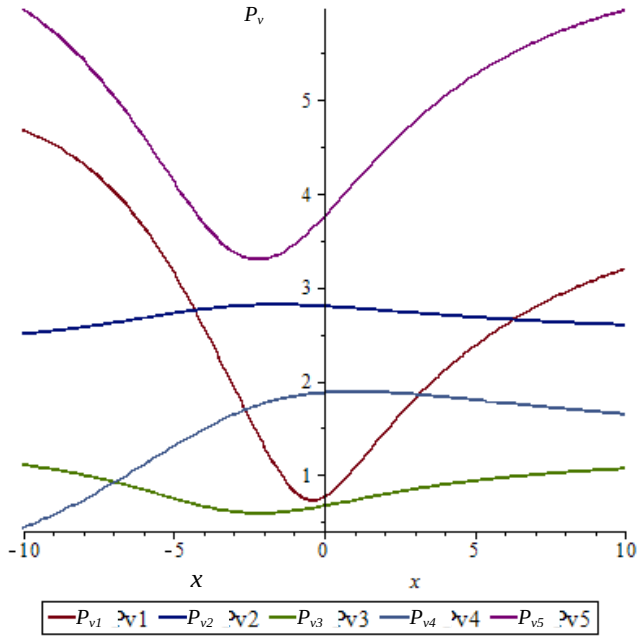


Fig. 3. Dependences of variable components of power on controllable shunt reactance x

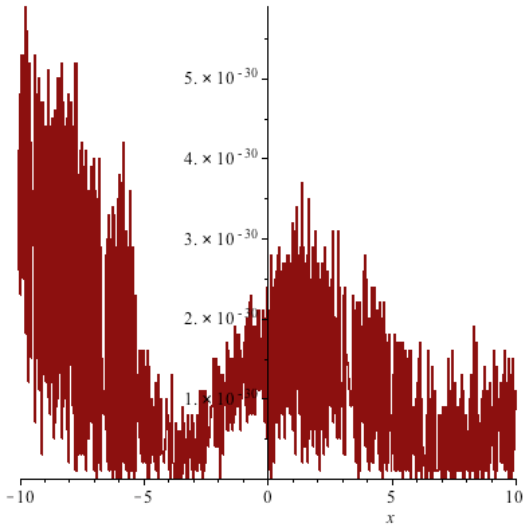


Fig. 4. Dependences of error in determining the value of the power P_{v5} on controllable shunt reactance x

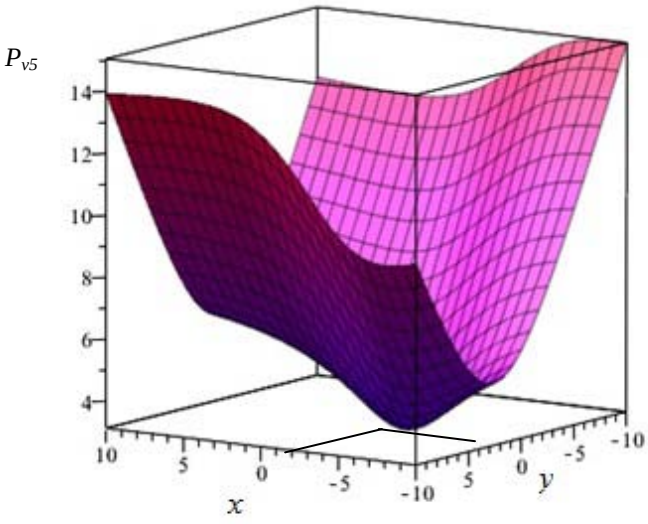


Fig. 5. Dependence of unbalanced power P_{v5} on reactances of two variable shunts (v.s.)

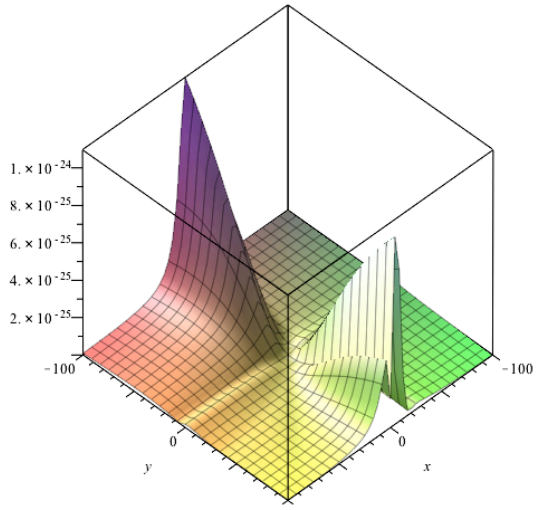


Fig. 6. Dependence of error in determining the value of the power P_{v5} on reactances of two variable shunts

It follows from the diagram that the use of two adjustable devices allows a complete compensation of unbalanced power. This approach may be summarized for a larger number of controllable shunts. It is evident that the increase in number of these devices makes possible to solve in more efficient way the electromagnetic torque pulsating with double frequency on the shaft of synchronous machine.

To solve the problem of optimizing of the value of unbalanced power we minimize criterion function which was received earlier.

For example, we shall consider electric network shown on fig. 2. In this scheme we set the v.s. in the different phases 1-4 loads. The optimal parameters v.s. and optimization results are shown in table I and table II.

ISBN: 978-1-61804-251-4

TABLE I. THE OPTIMUM VALUES OF CONTROL ACTIONS

Location v.s.		2 v.s.	3 v.s.	4 v.s.	5 v.s.	6 v.s.
		Optimum value, Ohm	*Optimum value, Ohm*	*Optimum value, Ohm*	*Optimum value, Ohm*	*Optimum value, Ohm*
Load 1	A	-	-	-	-	0
	B	-	-0.98835	-1	-1	-1
	C	-	-	-	-	-
Load 2	A	-	-	-1	-	-
	B	3.26937	3.51165	3	4	4
	C	-	-	-	1	1
Load 2	A	-	-	-	-	-
	B	-	-	-	-	-
	C	-	-	-	-	-
Load 3	AB	-1.09152	-1.04171	-1	-1	-
	BC	-	-	-	-	1
	AC	-	-	-	0	1

TABLE II. THE RESULTS OF OPTIMIZATION

	Without v.s.	2 v.s.	3 v.s.	4 v.s.	5 v.s.	6 v.s.
Pvar1	0.59557	0.59088	0.01761	3.86E-07	1.08E-06	0
Pvar2	2.35982	0.57690	0.54216	1.33E-05	2.49E-06	0
Pvar3	0.28735	0.06994	0.05722	1.07E-06	2.27E-07	0
Pvar4	1.37765	0.31336	0.20383	6.85E-06	2.62E-06	0
Pvar5	1.63397	0.39335	0.32172	6.01E-06	1.28E-06	0
K	3.25174	0.96941	0.66526	1.62E-05	3.99E-06	0

As can be seen from the table I and table II – two v.s. allow to decrease the value of unbalanced power in 4 times. Increasing of quantity of v.s. allows to decrease the value of unbalanced power.

III. CONCLUSIONS

The proposed method may be applied for the solution of other problems, among them are the following: balancing of the most sensitive with regard to electric energy quality part of power system, minimization of active power losses, stabilization of three-phase voltages, enhancement of asynchronous machine performance stability and reduction of errors occurring in power consumption measuring circuits.

In contrast with traditional balancer sets, the use of shunts with controllable reactivity possesses a series of essential advantages:

- low cost value as compared with balance-unbalance transformer;
- possibility to be connected to existing electric plant;
- low operating expenses
- possibility to continue power supply to consumers while the plant is out of service for repair.

However, it should be noted that the proposed method has a restricted scope of application versus to traditional balancer sets. This method may not be applicable to completely balance the three-phase load. The scope of application of this method covers three-phase consumers with insignificant phase unbalance.

The proposed approach fits well into the existing Smart grid conceptual design that declares the increased role of a consumer during the process of electric energy generation and supply. A common usage of similar devices for industrial and domestic loads shall help to transform large energy units into multi-agent systems with multitude of active consumers.

References

[1] Enrique Acha, Claudio R. Fuerte-Esquivel, Hugo Ambriz-Perrez, Cersar Angeles-Camacho, FACTS. Modelling and Simulation in Power Networks, England, John Wiley & Sons Ltd, 2004.

[2] Korovkin N.V., Chechurin V.L., Hayakawa M., Inverse problems in electric circuits and electromagnetics, USA, Springer, 2006.

[3] Korovkin N.V., Belyaev N.A., Chudny V.S., Frolov O.V. Power System State Optimization. Novel Approach, EMC Roma September 2012.

[4] Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[5] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

Nikolay V. Korovkin was born in Leningrad, Russia, on January 09, 1954. He received the M.S., Ph.D. and Doctor degrees in electrical engineering, all from St. Petersburg State Polytechnical University in 1977, 1984, and 1995 respectively. He worked as a Professor for the University of Electro-Communications (Tokyo, Japan), for the Swiss Federal Institute of Technology (EPFL) Switzerland, and for Otto-von-Guericke University (Magdeburg, Germany). His main scientific interests are EMC problems, stiff systems, impulse processes in linear and non-linear systems, "soft" methods of optimization. Now he is the head of the department of Theoretical Electrical Engineering in St. Petersburg State Polytechnical University. Prof. N. Korovkin is a Member of the Academy of Electrotechnical Science of the Russian Federation, (since 1996).

Quang Sy Vu got the master's degree EMC in power engineering at St. Petersburg State Polytechnical University in 2014.

Roman A. Yazenin is the postgraduate student in the department of Theoretical Electrical Engineering in St. Petersburg State Polytechnical University

Oleg V. Frolov is general manager in Joint Stock Company «Scientific and Technical Center of Unified Power System», Ph.D. in electrical engineering.

Nikolay V. Silin is the head of the department of Theoretical Electrical Engineering in Far Eastern Federal University, Doctor degree in electrical engineering

Integrated technology for industrial software verification and testing

Kotlyarov V., Drobintsev P., and Nikiforov I., St. Petersburg State Polytechnic University

***Abstract*—** This paper devoted to technology and software tools which allow automating of full cycle of software development from formalization of requirements provided in natural language and it's analysis with symbolic verification to tests generation and execution. The main achievement of technology is in checking of semantics consistency of requirements in generated code of target application. Area of technology applicability is very wide but now the target is telecommunication and distributed systems.

***Keywords*—** Formal model, software quality assurance, testing automation, verification.

I. Introduction

One of the main problems in development and testing automation of industrial applications' software is handling of complicated and large scale requirements specifications. Documents specifying requirements specifications are generally written in natural language and may contain hundreds and thousands of requirements. Thereby the task of requirements formalization to describe behavioral scenarios used for development of automatic tests or manual test procedures is characterized as a task of large complexity and laboriousness.

Applicability of formal methods in the industry is determined to a great extent by how adequate is the formalization language to accepted engineering practice which involves not only code developers and testers but also customers, project managers, marketing and other specialists. It is clear that no logic language is suitable for adequate formalization of requirements which would keep the semantics of the application under development and at the same time would satisfy all concerned people [1].

In modern project documentation the formulation of initial requirements is either constructive, when checking procedure or scenario of requirement coverage checking can be constructed from the text of this requirement in natural language, or unconstructive, when functionality described in the requirement does not contain any explanation of its checking method.

For example, behavioral requirements of telecommunication applications in case of described scenario of coverage are constructively specified and assume allow using of verification and testing for realization checking. Non-behavioral requirements are usually unconstructively specified and equire additional information during formalization which allows reconstructing the checking scenario, i.e. converting of non-constructive format of requirements specification into constructive one.

II. Requirements Coverage Checking

The procedure of requirement checking is exact sequence of causes and results of some activities (coded with actions, signals, states), the analysis of which can prove that current requirement is either covered or not. Such checking procedure can be used as a criterion of coverage of specific requirement, i.e. it can become a so-called criteria procedure. In the text below a sequence or "chain" of events will be used for criteria procedure.

Tracking the facts of criteria procedure coverage in system's behavioral scenario (hypothetical, implemented in the model or real system), it can be asserted that the corresponding requirement is satisfied in the system being analyzed.

Procedure of requirement checking (chain) is formulated by providing the following information for all chain elements (events):

- conditions (causes), required for activating of some activity;
- the activity itself, which shall be executed under current conditions;
- consequences – observable (measurable) results of activity execution.

Causes and results are described with signals, messages or transactions, commonly used in reactive system's instances communications [2], as well as with variables states in the form of values or limitations on admissible values. Tracking states' changes, produced by chains activities, lets observe the coverage of corresponding chains. While analysis it is acceptably to consider a direct transition from a state into a state with a null activity, and in case of non-determinism – alternative variants of states changes.

Problems with unconstructive formulations of requirements are resolved by development of requirement coverage checking procedures on user or intercomponent interfaces.

Thus, chains containing sequences of events can appear as criteria of requirements coverage; in addition, it is possible that criteria of some requirement coverage is specified not with one, but with several chains.

ISBN: 978-1-61804-251-4

III. Initial Documents Specifying Application Requirements

Usually initial requirements in technical documentation are formulated in natural language and can be presented in one of the following form:

1. in form of behavioral requirement, in this case scenario (procedure) of requirement checking can be restored based on requirement text;
2. in form of non behavior requirement, in this case only structure and sence of requirement can be restored without information about requirement checking.

Any behavioral requirement can be formalized with futher automatic analysis for this purposes VRS/TAT technology [3] can be used.

In VRS/TAT technology Use Case Maps (UCM) notation [4] (Fig 1.) is used for high-level description of the model, while tools for automation of checking and generation work with model in basic protocols language [5].

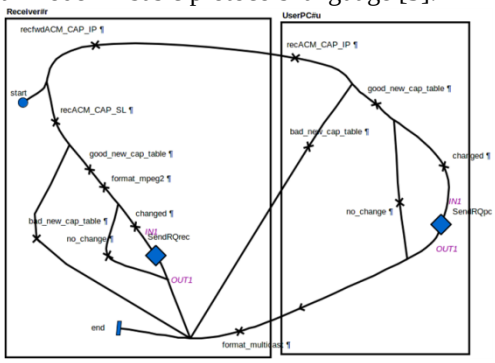


Figure 1. UCM model of two instances: Receiver and UserPC.

UCM model (Fig.1) contains two interacting instances model description. Each path on the graph from the event "start" to the event "end" represents one behavioral scenario. Each path contains specified number of events (Responsibilities). Events on the diagram are marked with × symbol, while Stub elements which encode inserted diagram – with ♦ symbol. As a result, each scenario contains specified sequence of events. Variety of possible scenarios are specified with variety of such sequences.

In these terms a chain is defined as subsequence of events which are enough to make a conclusion that the requirement is satisfied. A path on the UCM diagram, containing the sequence of events of some chain, is called trace, covering the corresponding requirement. Based on a trace tests can be generated which are needed for experimental evidence of requirement coverage.

IV. Traceability Matrix

Verification project requirements formalization starts with Traceability matrix (TRM) [1] creation (TRM for specific project is presented in table format in Fig.2). "Identifier" and "Requirements" columns contain requirement's identifier, used in the initial document with requirements, and text of the requirement, which shall be formalized. "Traceability" column contains chains of events sufficient for checking of corresponding requirement coverage and "Traces" column – traces or behavioral scenarios used for tests code generation.

Identifier	Requirements	Traceability	Traces
FREQ_GWR.1	Gateway shall transmit ACM_CAP messages repeatedly at an interval of T1 (TBD). This message will carry the ACM Capabilities table.	FREQ_GWR.3	
FREQ_GWR.2	Upon a change in the ACM Capabilities table, the Gateway shall send a new version of ACM_CAP message to the Satellite Terminal.	FREQ_GWR.3	
FREQ_GWR.3	Depending on configuration, ACM_CAP message shall be transmitted either in the MPEG2 private section or in a Multicast message sent across the satellite link.	recACM_CAP_SL recfwdACM_CAP_IP recACM_CAP_IP	FRED_GWR_3-1 FRED_GWR_3-2

Figure 2. Traceability matrix.

For example, in the third row of TRM there are 2 chains in "Traceability" column for covering FREQ_GWR.3 requirement. To satisfy the requirement it is enough to trace ACM_CAP signal sending in one of two possible scenarios:

3. FREQ_GWR.3-1:start,recACMCAP_SL, good_new_cap_table, format_mpeg2, no_chanes, end
4. FREQ_GWR.3-2: start, recfwdACM_CAP_IP, recACM_CAP_IP, good_new_cap_table, format_multicast, end

It should be noted, that during formulating of criteria chains a model of verified functionality is being created which introduces a lot of state variables, types, agents, instances, etc.

V. Development Integral Criteria Of Requirement Coverage

Mentioned above is distinctive feature of VRS/TAT technology – special criteria of each requirement's coverage checking. Below all criteria related to requirements are listed in ascending order of their strength:

1. Events criterion - coverage level in generated scenarios of subset of events used in criteria chains.
2. Chains criterion - coverage level in generated scenarios of subset of chains (consisting of events and states of variables) with at least one for each requirement.
3. Complex criterion - coverage level in generated scenarios of the whole set of chains specifying integral criteria (combined from criteria 1 and 2) of requirements coverage.

Criteria development shall be adaptive to specific project [6,12,13]. Criteria shall be applied flexibly and can be changed according to conditions of scenarios generation.

VI. Generating and Selecting Scenarios which Satisfy to Specify Coverage Criteria

Trace generation is performed by symbolic and concrete trace generators (STG and CTG) of VRS system, which implements effective algorithms of Model Checking [9,10]. The main problem of trace generation is "explosion" of variants combinations while generating traces from basic protocols [1], which formalize scenario events, conditions of their implementation and corresponding change of model state after their implementation. Solution here is filtration of generation variants based on numerous limitations specifically defined before trace generation cycle.

ISBN: 978-1-61804-251-4

There are general and specific limitations. For example, commonly used general limitations are maximum number of basic protocols used in a trace and maximum number of traces generated in a single cycle of generation. Also limitations on Goal and Visited states can be defined here. Specific limitations are defined by sequences of events in UCM model (so-called Guides) which guide the process of generation in user-preferable model behavior.

A guide is defining in terms of a state model which is presented in form of a transition system. A transition system TS is a tuple <Q, q0, T, P, f>, where Q is a set of states, $q_0 \in Q$ is initial state, T is a set of transition names, P is a set of agents, and f:Q$\rightarrow$P is a mapping which defines the current set of agents in the state Q.

To simplify, let's assume that events in the model are mapped with names of TS transitions and agents can be presented as one process or set of processes.

A path in TS from a state q_i to a state q_j is defined as a sequence of transitions $q_i \xrightarrow{t_i(a_i)} q_{i+1} \xrightarrow{t_{i+1}(a_{i+1})} q_{i+2} \ldots q_j$, where $q_k \in Q \wedge t_k \in T \wedge a_k \in f(q_k)$ for each $k \in i..j$.

A trace in TS is an ordered sequence $t_0(a_0)$, $t_1(a_1)$, … $t_n(a_n)$… such that there exists a path $q_0 \xrightarrow{t_0(a_0)} q_1 \xrightarrow{t_1(a_1)} \ldots \xrightarrow{t_n(a_n)} q_n \ldots$

A guide is a.n – transition a on the maximal distance n, which allows a set of traces {a, X_1 a, …, X_1…Xn a}, where X_1,…,Xn are any not empty symbols from {L\a},

~ a is a prohibition of transition any symbol from {L\a},

a; b (where a, b – guides) is a concatenation of guides, allows a set of traces {ab},

a $\vee$ b (where a, b guides) is a non deterministic choice of guides, allows a set of traces {a, b},

a || b is a parallel composition of guides a and b

join(a1,…,an) is a set {Sn} of all permutations of guides a1,…,an,

loop(a) is an iteration of guide *a*, i.e. { aa^* }.

There are two steps of test scenarios generation by Guides. On the first step guides are created which guarantee that the specified criteria of system behavior coverage are satisfied. On the second step guides in the UCM notation (Fig.3a) are translated into guides in the basic protocols language (Fig.3b) and control trace generation process.

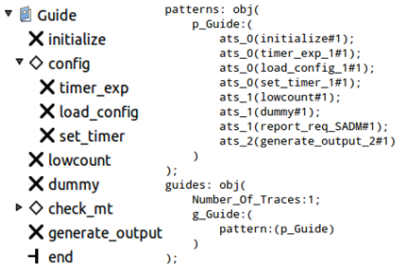


Figure 3. Guides: (a) in UCM notation,
(b) in basic protocols language (VRS Guide Language)

It is important, that only main control points in a behavior are specified in guides, while the trace generated from the guide contains a detailed sequence of behavioral elements. Such approach significantly reduces the impact of combinatorial explosion on the generation time during exploring a behavior UCM model of system of the system under development.

Usage of the method in real industrial projects reveals a set of problems, which are not trivial from the theoretical point of view. For instance, in case of a multi-search, when guides describe traces which belong to far branches of the behavioral tree, the storage of covered states grows very fast and the search process slows down accordingly. For such cases an efficient solution is partitioning the initial set of guides into subsets of different experiments.

Another problem is overstating the distance between transitions. In such cases the breadth-first search and greedy algorithms can help. To avoid permutations unnecessary from the test scenario point of view, the join operator is used. This operator allows permutation of arguments but prohibits interleaving.

VII. Automatic and Manual Process of Guides Creation

There are two possible approaches to guides creation from high-level system description in UCM language: manual and automatic.

Automatic approach allows to generate numerous guides [9.10] covering system behavior on branches criteria [12,13]. Each guide contains information about key points of behavioral scenario, starting from initial model state modeled by StartPoint element and ending in final state modeled by EndPoint element. Process of guides generation is performed by UCM to MSC generator [7].

The automatic approach to guides creation can be considered as a fast way to obtain a test set which satisfies branches criteria; however. this approach is not always suficient. Some functionality can be checked only using paths criteria. As this criteria often deals with uncertanties in selection of a test set suficient for the specified requirements coverage, it is usually partly applied only to cover specific requirements. Guides creation for covering a subset of requirements by paths criteria requires more information during generation which is added manually. Besides that, the customer and the test engineer may want to check specific scenarios of the system behavior for checking some specific requirements. Such scenarios are specified manually by and they are created with the UCM Events Analyzer (UCM EVA) tool [7].

In both cases, automatic or manual, problems may occur with guides' coverage by test scenarios.

In automatic mode this is due to the fact that VRS tool not always can successfully generate test scenario form guides because guides are created based on model's control flow and do not consider values of corresponding data flow. At the same time, traces generated based on control flow consider changes in variables values and accordingly model states.

Therefore the actual task is automated analysis of why some guides are not covered by traces and accordingly automation

of guides adjusting solved by guides or UCM model modification.

VIII. Technology Chain of Test Scenarios Generation

The process of tests generation with usage of VRS/TAT tools can be divided into following phases:

Manual creation of formal model for requirements in UCM language. Key actions are refined from initial requirements. Based on these actions a set of chains, which will be translated into behavioral scenarios are created. On the next step objects which interact in the scope of scenarios and structure are refined. Metadata for data flow are added into formal model. More detailed description of the phase is presented in [6].

The next phase is conversion of UCM model into BP notation [3] which is input language for VRS/TAT tools. Translation is performed with saving of model construction semantics and as result of this phase equivalent to initial UCM model is generated.

Guides which satisfy branch criterion are also generated based on initial UCM model. In process of generation all branches are traversed with VRS/TAT toolset.

In case if customer have to check some special behavior of the system then he creates guides manually in EVA toolset. It shall be noted that guides are created in terms of UCM element names and for future usage in VRS they shall be translated into sequence of basic protocols. Such translation is performed based on information which was collected on the first phase in process of UCM to BP translation.

On the next phase formal model in basic protocols notation and guides are used in trace generator of VRS tool for verification and automatic traversal of behavioral tree for symbolic scenarios generation.

After scenarios generation process finished it is necessary to analyze coverage of guides by traces with goal to define which guides are still not covered and identify the reasons. Such analysis can be performed in EVA module. Based on analysis results UCM model or guides shall be corrected.

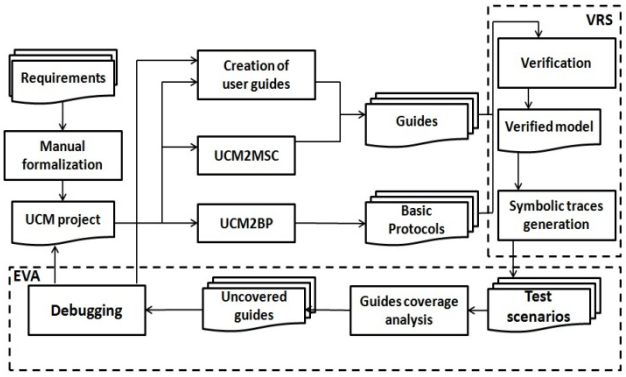


Figure 4. Process of test scenarios generation with usage of static analysis tool EVA.

IX. Method of Guides Adjustment Automation

Most often reasons of discrepancies are insufficient (not enough detailed) guides specification and mistakes in the sequence of UCM elements in the guide due to incorrect usage of variables, identified as a deadlock on the branch or ramification.

Consider the method of searching of places and reasons of discrepancies between a guide and a trace.

1. Guides and traces generated in VRS are presented as MSC diagrams containing sequences of basic protocols application, thus the first step of the algorithm is mapping basic protocols names on UCM elements.
2. Comparing the guide and the trace in terms of UCM elements it is possible to define the last trace element which satisfies the sequence of UCM elements specified in the guide. The next uncovered element of the guide will be referred to as the element of discrepancy.
3. For the element of discrepancy uncovered in the symbolic trace it is possible to explore corresponding data and precondition.
4. Then variables of precondition shall be singled out into separate list and those places on the UCM diagram where variables of this list are changed shall be analyzed. The analysis shall be performed from the bottom up, starting with the events closest to the element of discrepancy.
5. After revealing the reason of discrepancy, the guide shall be corrected or the UCM model shall be changed.

The steps above shall be repeated until all guides will be covered by traces.

Consider the process of searching for discrepancy on the example of telecommunication project (Fig. 5).

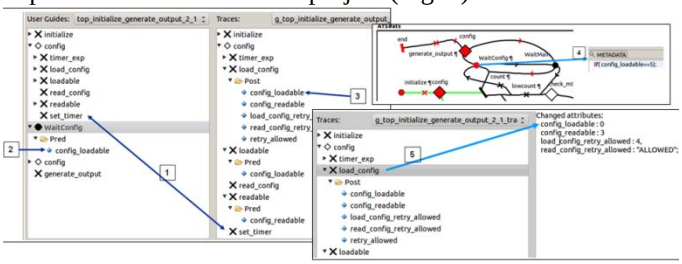


Figure 5. Revealing the reason of discrepancy between the guide and the trace due to variables values.

While searching for the reason of guide's non-coverage it is firstly required to find the last element of coincidence between the guide and the trace (1) – "set_timer" in this example. Then the guide element which can not be achieved in the trace (the discrepancy element) shall be found – "WaitConfig" in this example. Analyzing its metadata (2), detect the variable which affects the trace generation (3) and can be the reason of discrepancy – "config_loadable" in this example. After analysis of this variable's values in the UCM model (4) draw a conclusion that in order to apply this element the variable's value shall be 5. The analysis is held only for those points in the trace where "config_loadable" is used. In current case such point is "load_config" element (5), where 0 is assigned to "config_loadable" variable. Thus the conclusion can be made that in order to create a guide which will be successfully covered by a trace it is required to assign value 5 on "load_config" element or change the guide.

ISBN: 978-1-61804-251-4

Achieved is the reduction of laboriousness in searching for the reasons of errors due to decreasing of the number of points being analyzed which is actual for large industrial projects created in accordance with considered technology.

X. Concretization of Symbolic Traces

A hierarchical classification of tools used currently in software industry may be done depending on how they support the following features (Fig.6):

1. Automated run of a test suite based on a test plan and automated analysis of error causes and locations.
2. Automated generation of optimized test suites for a target platform from behavior scenarios (traces).
3. Using a behavior model for automated generation of scenarios which satisfy some given coverage criterion.
4. Automated generation of a behavior model of a software product from manually developed formal models of particular requirements (Hoare triples [14], Letichevsky basic protocols [5], etc.)

The last point needs some explanation. We distinguish between a formal model of the source requirements, which is created manually to render them in some formal language with some certain level of completeness and consistency, and a formal behavioral model, which describes the system behavior on a certain level of abstraction in terms of signal exchanges, delays, control flow synchronizations, etc.

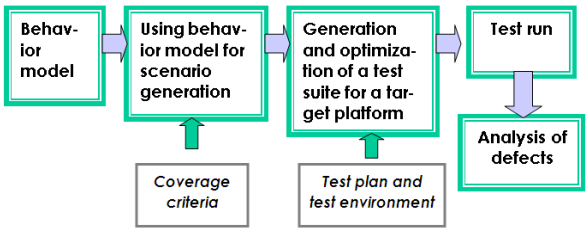


Figure 6. Features of the test automation tools.

Among a variety of existing testing technologies and tools, systems [15], [16] belong to the first most wide-spread class; systems [17], [18], [19] belong to the second class where manually developed scenarios are used; system [20] belongs to the third class, where powerful CASE-tools are used and traces are generated from manually developed behavior models; and VRS/TAT [3] belongs to the fourth class of systems with the highest level of automation with respect to this classification.

One of major problems of model-based behavior scenarios generation for industrial systems is the problem of explosion of the number of behavior variants to be tested. Indeed, the number of states which describe a set of concrete system behaviors and therefore, the number of generated traces, usually grows exponentially under a straightforward approach.

To reduce the behavior space, symbolic traces are used; they differ from concrete traces in only one aspect: variables of various types rather than concrete values are used as their parameters. Each such variable is accompanied with its tolerance range. Thus, each symbolic trace represents a bunch of concrete traces with equivalent behaviors. This means that one representative of each class of behavior equivalence may be used in the test suite.

Symbolic trace generator STG [21], which is a component of the verifier VRS, is a symbolic trace generator tool, which reduces the behavior space.

A set of symbolic traces which satisfies the coverage criterion, substantially reduces the test suite after such optimization; however, the delivered symbolic traces are not suitable for generation of platform-oriented tests. Concrete traces with concrete parameter values are needed for test generation, which poses the task of creation a toolset for automated concrete definition (Trace Concretization Tool) of symbolic traces.

It's noteworthy that the number of tests in current large and medium size industrial projects is thousands and more. Because of this large number, consistent concrete definition of symbolic traces using the respective tolerance ranges of their variables should be totally automated.

Besides that, real practice of testing does not allow for too formal concrete definition, e.g., by selecting parameter values from an acceptable tolerance range at random. This results in generation of a huge number of tests and in general inefficiency of the symbolic approach. It should be mentioned that tolerance ranges should be updated after each substitution of a concrete value, because ranges for next substitutions may depend on concrete value substitutions made before. When substituting, the test plan prepared upfront should be followed. Its major elements are parameter values known from practice; usually they are the boundary values, some value within the tolerance range, and some out-of-range value. Such plans are expected to be flexible, their major part being generated from standard templates, and the remaining part being user-edited plan-templates.

In the scope of the paper an approach to concretization problem solving is suggested. The structure of the concrete definition tool is represented in Fig.2. The tool consists of two modules Concretizator and Substitutor which communicate with each other in a dialog.

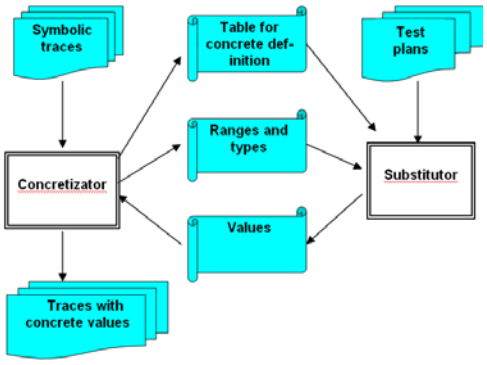


Figure 7. Tool for concrete definition.

Based on the set of symbolic traces, Concretizator creates a table for concrete definition (Fig.8) for each symbolic trace where it fills-in the columns of variables and signals names, types, and allowed tolerance ranges; and then while direct traversing each trace it calls Substitutor for a concrete value of

ISBN: 978-1-61804-251-4

each variable encountered in the trace.

Based on the commands and data of the plan for trace concrete definition, Substitutor calculates the value from the respective range to be substituted and returns it to Concretizator. The plan for automated concrete definition of a test suite is placed in the column “User option” and is formed in terms of the control commands R,M,L,O,C.

Var name	Signal name	Type	User option	Range	Value
Speed_Value	Current_Speed	I	L	-2147483648 <=Speed_Value &Speed_Value <=2	-2147483648
Speed_Value	Current_Speed	I	L	4<=Speed_Value&Speed_Value<=2147483647	4
Driving_Mode	MCS_Position	TRAIN_OPERATION_MODE	L	_AUTO_,_CS_,_RM_REV	_AUTO

Figure 8. Table of concrete definition.

Command R provides substitution of the right limit of the Range for the respective variable, command L provides substitution of the left limit, command M calculates and substitutes the mean value of the Range, command O provides substitution of a value out of the Range, and command C provides substitution of a concrete value specified by the user. Thus, means to control concrete definition are flexible enough to test any regimes of the software product functioning.

XI. Testing Automation

For automatic tests execution and results analysis TestCommander tool [7] is used. Testing automation process is presented in Fig.9.:

Full control over test execution and test result analysis are required to perform conformance testing of a system. Therefore, it is necessary to create a test suite, which is able to specify interaction with SUT in exactly the same way it is done in the requirements. Observing the behavior of the system during the execution of such test cases allows to establish whether the system implementation is correct or not.

For this purpose, Message Sequence Chart (MSC) language is used in TestCommander for test suite description. It automatically generates the code of test suite modules on a target programming language (C++, Java, tcl) from MSC charts. These modules interact with SUT and each other using predefined interfaces, reproducing test scenarios. These interactions can be unambiguously and in a human-oriented manned defined in terms of MSC and then represented in textual or graphical view.

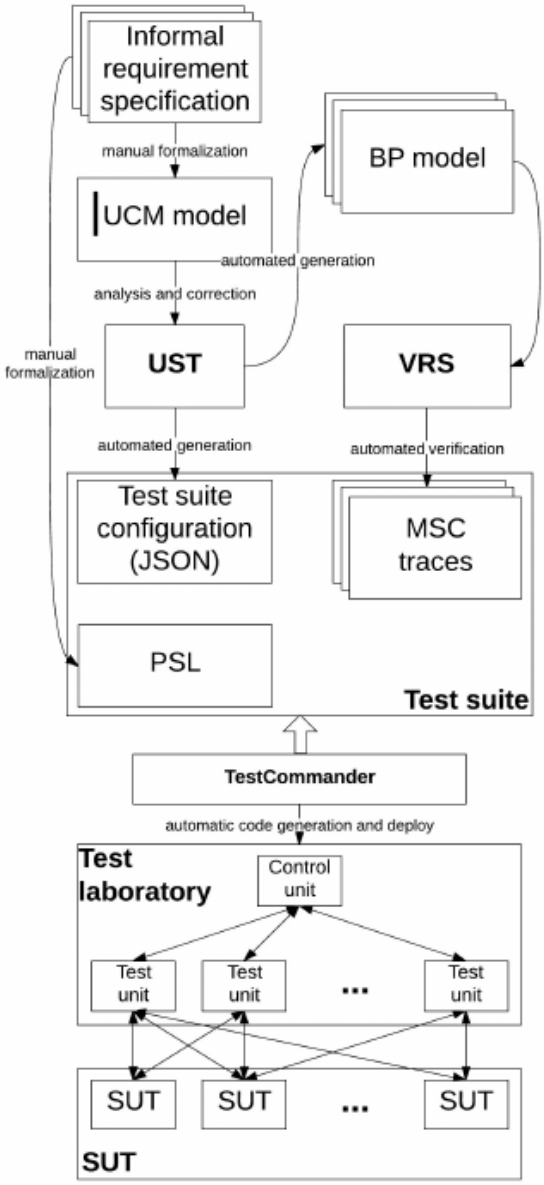


Figure 9. Distributed testing approach overview.

As a combination of requirements formalization technique described earlier and the technology of test scenario generation based on the verification of a model in the BP notation, this approach brings together the requirements management, verification and testing in one technology. Automatically generated tests define the complete description of the interaction of all the components of SUT and its environment. All of the above allows these specifications to be tested and verified.

TestCommander tool accepts MSC charts obtained with various methods (assuming that they correspond to MSC standard). However, the proposed approach has some significant advantages, such as:

- Automation of routine and time-demanding operations;
- Simple and human-oriented formal notations usage;

ISBN: 978-1-61804-251-4

- Requirements checking with verification methods

Generated test suite executable file set consists of one or more testing units and one control unit. Testing unit interacts with the SUT according to the test logic and exchanges the control signals with the control unit. The latter controls the test execution process and collects testing results. Protocols of interaction between testing system units and SUT are defined with Protocol Specification Language (PSL). This notation unambiguously specifies the format of the messages passed between the entities involved in the testing. PSL specification is created manually and is used for test suite code generation.

For test suite configuration, code generation and test suite deploy setup a configuration file is used. It is written in JSON and specifies the location of testing units and SUT components. Nonetheless, the main feature of this configuration is to specify, which of the SUT components are substituted with the test units. It allows testing of the behavior only of a part of the system. The configuration file is generated automatically from UCM model, but it can be adjusted manually.

After the test suite is configured, its code in target language is automatically generated and the test suite is deployed in the test laboratory. Testing starts with execution of control module of the test suite, which than controls the test unit threads and SUT. Test results are the MSC diagrams of test activities.

The suggested method is a combination of requirement management, verification and testing. It allows performing the checking of the correctness of the system implementation according to its specification within one technology. Testing approach is based on the automatically generated test suite, the correctness of which is proved during the system formal specification verification. It reduces the cost of regression testing needed in case of changing or refinement of specifications. All stages of this method are fully or partly automated. The developed software components used in these stages are independent; and all data formats are standardized. All of the above ensures that the whole method is scalable, highly flexible and adaptive and open for modernization.

XII. Results

Results of technology usage in three projects are presented in the table 1: SMTP – module of mail protocol; CDMA – module of base station which implements technology Control Division Multiple Access; «Satellite Terminal» - module of the telecommunication system.

Table 1. Results of experiments

Projects	Number of requirements	Found defects (total / significant)	Time saving (times)
SMTP	30	7/1	3,3
CDMA	205	129/17	2,9
Satellite Terminal	392	203/28	3

Results show effectiveness of defects detection with usage of tests generation. Also time saving is near 3 times in comparison with manual approach.

XIII. Conclusions

The result of this work is improved technology which integrates verification and testing of software projects and provides:

- Full automation of industrial software product development process with the control of requirements semantics realization.
- Generation of application's model and symbolic behavioral scenarios, which fully (100%) cover behavioral features of the application.
- Automated concretization of symbolic traces in accordance with test plan.
- Automated test-suite generation following concretized traces
- Automated execution of system and regression testing phases.
- High level of automation for the process of development and managing of software product quality.

Results of integrated design and testing technology appliance in the development of wireless telecommunication applications demonstrated 26% time-saving in software product creation.

References

[1] S.Baranov, V.Kotlyarov, A.Letichevsky. Industrial technology of mobile devices testing automation based on verified behavioral models of requirements project specifications// «Space, astronomy and programming» – SpbSU, Spb. – 2008. – pp. 134–145. (in Russian).

[2] Z.Manna, A.Pnueli.: The Temporal Logic of Reactive and Concurrent Systems. Springer-Verlag, 1992.

[3] S.Baranov, V.Kotlyarov, A.Letichevsky, P.Drobintsev. The technology of Automation Verification and Testing in Industrial Projects. / Proc. of St.Petersburg IEEE Chapter, International Conference, May 18-21, St.Petersburg, Russia, 2005 – pp. 81-86

[4] Recommendation ITU-T Z.151. User requirements notation (URN), 11/2008

[5] A. Letichevsky, J. Kapitonova, A. Letichevsky Jr., V. Volkov, S. Baranov, V. Kotlyarov, T. Weigert. Basic Protocols, Message Sequence Charts, and the Verification of Requirements Specifications. Proc of ISSRE04 Workshop on Integrated-reliability with Telecommunications and UML Languages (ISSRE04:WITUL), 02 Nov 2004: IRISA Rennes France.

[6] P.Drobintsev, V.Kotlyarov, I.Chernorutsky. Test automation based on user scenarios coverage. "Scientific and technical sheets", SpbSTU, vol.4(152)-2012, pp.123-126 (in Russian).

[7] I.Anureev, S.Baranov, D.Beloglazov, E.Bodin, P. Drobintsev, A.Kolchin,, V.Kotlyarov, A. Letichevsky, A. Letichevsky Jr., V.Nepomniashiy, I.Nikiforov, S.Potienko, L.Priyma, B.Tytin. Tools for support of integrated technology for analysis and verification of specifications telecom applications // SPIIRAN proceedings- 2013-№1-28P.

[8] I.Nikiforov, A.Petrov, V.Kotlyarov. Static method of test scenarios adjustment generated from guides // "Scientific and technical sheets", SpbSTU, vol.4(152)-2012, pp. 114-119 (in Russian)

[9] A.Kolchin, V.Kotlyarov, P. Drobintsev. A method of the test scenario generation in the insertion modelling environment // "Control systems and computers", Kiev: "Akademperiodika", vol.6-2012, pp.43-48 (in Russian)

[10] A.A. Letichevsky, J.V. Kapitonova , V.P. Kotlyarov, A.A. Letichevsky Jr., N.S.Nikitchenko, V.A. Volkov, and T.Weigert. Insertion modeling in distributed system design // Programming problems. – 2008. – pp. 13–38

[11] A. Letichevsky Jr., A. Kolchin. Test scenarios generation based on formal model // Programming problems. – 2010. – № 2–3. – pp. 209–215 (in Russian)

[12] V.P. Kotlyarov. Criteria of requirements coverage in test scenarios, generated from applications behavioral models // “Scientific and technical sheets”, SpbSTU. – 2011. – vol.6.1(138). – pp.202–207. (in Russian)

[13] Baranov S., Kotlyarov V., Weigert T. Varifiable Coverage Criteria For Automated Tesdting. SDL2011: Integrating System and Software Modeling // LNCS. –2012–Vol.7083 – P.79–89.

[14] Hoare C.A.R. Communicating sequential processes, Prentice Hall, 1985.

[15] Abbot framework for automated testing of Java GUI components and programs http://abbot.sourceforge.net/doc/overview.shtml

[16] Jameleon - An Automated Testing Tool – Overview
http://jameleon.sourceforge.net/index.html

[17] Silk Software Test Management, Test Automation and Performance Testing http://www.borland.com/us/products/silkline/index.aspx

[18] Open Source Software Engineering Tools http://maxq.tigris.org

[19] Software Testing Tools and other Products
http://www.parasoft.com/jsp/products.jsp

[20] IBM Rational software
http://www01.ibm.com/software/rational/?pgel=ibmhzn&cm_re=masthead-_-products-_-sw-rational

[21] Letichevsky A. Kapitonova Y, Volkov V. Letichevsky A(jr), Baranov S, Kotlyarov V. Specification of systems with usage of basic protocols notation // Cybernetics and system analysis. – 2005. – №4. – p. 256-268.

Vsevolod Kotlyarov - was born in Stavropol region of Russia on the 14 July 1944. Hold a master degree with specialty «Mathematical and computing instruments and devices» of Saint-Petersburg State Polytechnic University (SPbSPU) in 1968. Defended PhD thesis with specialty “Software engineering” in 1972. Main areas of interests - «Software engineering», «Technologies and tools of automated verification and testing».

Since 1972 he is working as associated professor in SPbSPU, since 1995 as senior researcher in St.Peterburg software development department of Motorola, since 2008 as full time professor of SPbSPU. He is scientific adviser of 20 PHD dissertations of post-graduate students. His scientific school of “Software Engineering” was included in the list of top schools of St.Petersburg.

Prof. Kotlyarov became a M of IEEE and ACM in 1993, M of SABA (Science Advisory Board Association) of Motorola Company in 2005. He is a member of the program committees of the following conferences: Microsoft Technology in Software theory and practice, SYRCOSE, Workshops of Ershov informatics conference (PSI).

ISBN: 978-1-61804-251-4

About detection substitutions in nonlinear algebraic equations with help of Tarjan's algorithm.

Isakov A.A., Senichenkov Yu.B., Distributed Computing and Networking department, Saint Petersburg state Polytechnical University, senyb@dcn.icc.spbstu.ru, SPBSTU, Polytehnicheskaj 29, St. Petersburg,195251, Russia

Abstract — MvStudium (www.mvstudium.com) is a tool for visual modelling and simulation of complex dynamical systems, including hybrid systems. While model building MvStudium forms and reduces large scale systems of differential-algebraic equations for each state of hybrid automation used for specification of hybrid systems. State equations written by user may contain substitutions. Numerical solution costs for current system are in proportion to a number of equations and their structure. Automated detecting substitutions among equations leads to decreasing size of solving equations. Taking into consideration a structure of built system allows calling most effective Solver. For example if system may be transformed to the system with block-triangular form, Solver should solve systems for diagonal blocks only.

The MvStudium's version of Newton's method for solving systems of nonlinear algebraic equations with detecting substitutions is considered. In this version Tarjan's algorithm of finding the strongly connected components is used for detecting substitutions and transforming of an initial system to block-triangular form.

***Keywords* —** complex dynamical systems, modeling languages, equation-based models, hybrid systems, Tarjan's algorithm, Newton's root-finding method.

I. INTRODUCTION

In this paper «Analyzer» means «module that can detect type and structure system of equations written by User», and «Solver» - «module that can choose most effective numerical method using information about type and structure about solved system getted from Analyzer».

Newton's method is the basic method for solving nonlinear algebraic equations

$$f_i(x_1, x_2, ..., x_n, time) = 0, \mathrm{i} = 1, \mathrm{n} \qquad (1)$$

Isakov A.A. and Senichenkov Yu.B. are with the National Research University «St. Petersburg State Polytechnical University», SPBSTU, Polytehnicheskaj 29, St. Petersburg,195251, Russia, senyb@dcn.icc.spbstu.ru.

used in MvStudium.

The MvStudium's modification of the method takes into account structure of system and can freeze Jacobi matrix if it is efficiently. For detecting matrix structure (band, block-triangular, sparse and so on) Analyzer uses structure matrix $\mathbf{S}$ of system with $s_{ij} \in \{0,1\}$. Non-zero element s_{ij} in matrix $\mathbf{S}$ says that j-th unknown is presented in i-th equation.

Let us consider an example.

Example 1. User wants to solve the system of nonlinear algebraic equation written in the form (1) respect x_i. Constants C_i are known.

$$\begin{cases} 10 \cdot x_3 + x_4 - 10 = 0 \\ x_3 + \sin(x_4) + \cos(x_4) - 10 = 0 \\ x_1 - C_1 = 0 \\ x_2 - \sin(x_1) - C_2 = 0 \\ x_3 + \sin(x_1) - x_1 - x_2 = 0 \\ x_{25} - \sin^2(x_4) - \cos^2(x_4) = 0 \end{cases}$$

It is obvious that among equations there are substitutions.

$$\begin{cases} x_1 = C_1 \\ x_2 = \sin(x_1) - C_2 \\ x_3 = \sin(x_1) - x_1 - x_2 \\ x_{25} = 1 \end{cases}.$$

There is no such conception as «substitutions» in Model Vision Modeling language (MVL), but MvStudium's Analyzer can detect formal substitutions of type

$$x_i = f_i(x_1, x_2, ..., x_n, time)$$

by itself.

Formal substitutions may be considered as equations of course, but it increases size of solved system and may cause difficulties for Newton's method (slow convergence, bad initial conditions).

Detection of formal substitutions among equations and reordering them if necessary is a goal of Analyzer. Find substitutions should form a sequence suitable for calculations.

Using different sequences of substitutions leads to different final systems.

ISBN: 978-1-61804-251-4

Example 2. Initial User's sequence

$$\begin{cases} x_1 = -16 \cdot x_4 + 1 \\ x_2 = \sqrt{|x_3|} \\ x_3 = 25 \cdot x_2 - \dfrac{1}{x_3 + 16} \\ x_4 = x_1 + 25 \end{cases}$$

may be transformed into three final systems with substitutions - **A**), **B**) or **C**).

Final system with substitutions A. Two substitutions and two equations. (Symbol «:= » is used for substitutions).

$$\begin{cases} x_1 := -16 \cdot x_4 + 1 \\ x_2 := \sqrt{|x_3|} \\ x_3 = 25 \cdot x_2 - \dfrac{1}{x_3 + 16} \\ x_4 = x_1 + 25 \end{cases},$$

Final system with substitutions B. One substitution and three equations.

Subroutine for calculation residuals may be written in the form

$$F(x_1, x_3, x_4) = \begin{pmatrix} x_3 - 25 \cdot x_2 + \dfrac{1}{x_3 + 16} \\ x_1 + 16 \cdot x_4 - 1 \\ x_4 - x_1 - 25 \end{pmatrix}, x_2 := \sqrt{|x_3|}$$

Final system with substitutions C. Block-diagonal system. First block: one substitution and one nonlinear equation. Second block: two linear equations.

$$\begin{cases} x_2 := \sqrt{|x_3|} \\ x_3 - 25 \cdot x_2 + \dfrac{1}{x_3 + 16} = 0 \end{cases}$$

$$\begin{cases} x_1 + 16 \cdot x_4 = 1 \\ x_4 - x_1 = -25 \end{cases}$$

Using equations as substitutions may be dangerously even for linear systems.

Let us consider a system of linear equations

$$\mathbf{A} \cdot \mathbf{x} = \mathbf{b}$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{A}_{12} \\ \mathbf{A}_{21}\mathbf{A}_{22} \end{bmatrix}; \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}; \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix},$$

$$\mathbf{A}_{12} = \mathbf{A}_{21} = \mathbf{E}$$

$\mathbf{E}$ - identity matrix, $\mathbf{A}$ - square matrix with square block matrices of the same dimension, and let $\mathbf{A}$ is not singular.

It is possible to build two new systems with substitutions

$$\begin{cases} \mathbf{x}_2 := \mathbf{b}_1 - \mathbf{A}_{11} \cdot \mathbf{x}_1 \\ \mathbf{x}_1 + \mathbf{A}_{22} \cdot \mathbf{x}_2 = \mathbf{b}_2 \end{cases}$$

or

$$\begin{cases} \mathbf{x}_1 := \mathbf{b}_2 - \mathbf{A}_{22} \cdot \mathbf{x}_2 \\ \mathbf{A}_{11} \cdot \mathbf{x}_1 + \mathbf{x}_2 = \mathbf{b}_1 \end{cases},$$

but their numerical properties depend now on conditional numbers of matrices

$$\mathbf{E} - \mathbf{A}_{22}\mathbf{A}_{11}$$
$$\mathbf{E} - \mathbf{A}_{11}\mathbf{A}_{22},$$

and even there is no insurance arrangements that they are not singular.

II. Construction formally calculable sequence of substitutions

Definition. *Sequence of formal substitutions*

$$x_i = f_i(x_1, x_2, ..., x_n, time), i = k, m \quad (2)$$
$$1 \le k \le m \le n$$

is called formally calculable if each its member contains in right-hand function only already calculated unknowns.

It means that square $n \times n$ structure matrix $\mathbf{S}$ for *right-hand* functions $f_i(x_1, x_2, ..., x_n, time), i = 1, n$ with n arguments x_i is square *lower triangular* matrix with zero diagonal.

Example 3. Formally calculable sequence of substitutions and its structure matrix.

$$\begin{cases} x_1 = C_1 \\ x_2 = \sin(x_1) + C_2 \\ x_3 = -\sin(x_1) + x_1 + x_2 \\ x_4 = 1 \end{cases} \qquad \mathbf{S} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

However a sequence of formal substitutions may contains equations:

$$\begin{cases} x_1 = C_1 \\ x_2 = \sin(x_1) + C_2 + x_3 \\ x_3 = -\sin(x_1) + x_1 + x_2 \\ x_4 = 1 \end{cases},$$

In this case Analyzer has to divide formal substitutions on formally calculable substitutions and equations.

Initial information about sequence of formal substitutions may be written in the form (3)

$$\mathbf{x} = \mathbf{S} \cdot \mathbf{x} + \mathbf{C}; \mathbf{x}, \mathbf{C} \in \Re^n; S \in \Re^{n \times n} \quad . \quad (3)$$

Let us concatenate with sequence of formal substitutions (3) a system of linear algebraic equations

$$\mathbf{S} \cdot \mathbf{y} = 0 \quad (4)$$

ISBN: 978-1-61804-251-4

with structure matrix $\mathbf{S}$. Simulating elimination of variables in (4) it is possible to detect is (3) *formally calculable* or not.

Example 3. Consider matrix

$$\mathbf{S} = \begin{bmatrix} 0\ 0\ 0\ 0\ 0\ 0 \\ 1\ 0\ 0\ 0\ 0\ 0 \\ 1\ 1\ 0\ 1\ 0\ 0 \\ 1\ 1\ 1\ 0\ 0\ 0 \\ 1\ 1\ 1\ 1\ 0\ 0 \\ 1\ 1\ 1\ 1\ 1\ 0 \end{bmatrix}.$$

First zero line says that $y_1 = C_1, (f_1(t))$. Let us eliminate unknown y_1, removing first line and first colon from the matrix

$$\mathbf{S}_1 = \begin{bmatrix} 0\ 0\ 0\ 0\ 0 \\ 1\ 0\ 1\ 0\ 0 \\ 1\ 1\ 0\ 0\ 0 \\ 1\ 1\ 1\ 0\ 0 \\ 1\ 1\ 1\ 1\ 0 \end{bmatrix}.$$

Then it will be possible eliminate y_2 and substitute y_3 into 4,5,6-th equations

$$\mathbf{S}_3 = \begin{bmatrix} 0\ 1\ 0\ 0 \\ 1\ 1\ 0\ 0 \\ 1\ 1\ 0\ 0 \\ 1\ 1\ 1\ 0 \end{bmatrix}.$$

Final matrix $\mathbf{S}_3$ is a block-triangular matrix. Its first diagonal block

$$\mathbf{S}_{11} = \begin{bmatrix} 0\ 1 \\ 1\ 1 \end{bmatrix}$$

corresponds to equations, and second

$$\mathbf{S}_{22} = \begin{bmatrix} 0\ 0 \\ 1\ 0 \end{bmatrix}$$

to substations.

This method of detection of substitutions leads to fill-in effect as Gaussian Elimination.

III. Using Tarjan's Algorithm for Detection Substitutions

Oriented graphs usually are used for description of a structure of sparse systems of equations [4].

Fig. 1 illustrates using oriented graphs for description structure of a) - systems of equations, b) - systems of equations with substitutions, c) – system of substitutions.

$$a)\ \begin{cases} x_1 = 10 \cdot x_1 + 16 \cdot x_2 - 1 \\ x_2 = 10 \cdot x_2 - x_1 - 25 \end{cases};\quad b)\ \begin{cases} x_1 = -25 \\ x_2 = x_3 - 15 + x_1 \\ x_3 = 10 \cdot x_2 - 15 \\ x_4 = x_1 + x_3 \end{cases};$$

$$c)\ \begin{cases} x_1 = -1 \\ x_2 = x_1 - 25 \\ x_3 = x_2 - 15 \end{cases}$$

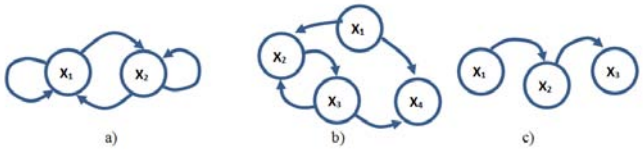


Fig. 1 Graphs for systems and substitutions

Tarjan's algorithm [1] for block triangularization of a matrix was used in MA28 [2]. We suggest using Tarjan's algorithm not only for reducing a matrix to block-triangular form but for detection of substitutions among equations.

Tarjan's algorithm will detect one strongly connected component for system a), three – for system b) and c). There are strongly connected component of dimension 1×1 in case of b), c). Those are substitutions.

Example 4. Matrix $\mathbf{S}$

$$\mathbf{S} = \begin{bmatrix} 1\ 1\ 0\ 0 \\ 1\ 0\ 0\ 0 \\ 1\ 1\ 0\ 0 \\ 1\ 1\ 1\ 0 \end{bmatrix} = \begin{bmatrix} S_{11} & 0 & 0 \\ S_{21} & S_{22} & 0 \\ S_{31} & S_{32} & S_{33} \end{bmatrix},$$

has block-triangular form. Each diagonal block S_{ii} corresponds to strongly connected component of oriented graph. Block S_{11} has dimensions 2×2, therefore there is one equation as minimum. If to solve the system corresponding to the first block, then sufficiently calculating only substitutions corresponding to blocks 2 and 3 (S_{22}, S_{33}).

ISBN: 978-1-61804-251-4

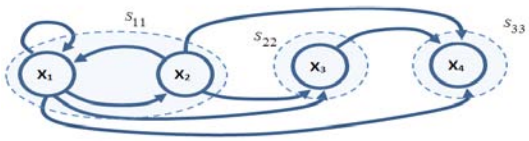


Fig. 2 Graph for the system with substitutions

Algorithm for detecting formally calculable sequence of substitutions.

Stage 1. Detect formal substitutions among equations

$$x_i = f_i(x_1, x_2, ..., x_n, time)$$

Stage 2. Delete from derived sequence explicit «members-equations»

$$x_i = f_i(x_1, .. \ \ x_i, \ \ ....x_n, time)$$

Stage 3. Detect strongly connected component with the help of Tarjan's algorithm.

Stage 4. Divide strongly connected component on substitutions (dimension 1×1) and equations (dimension greater then 1×1)

Stage 5. Reorder sequence and get *formally calculable* sequence.

Example 5. The system of formal substitutions

$$\begin{cases} x1 = -7 * x7 + x8 - 2 \\ x2 = x4 * 3 - 2 * x6 + 1 \\ x3 = \dfrac{x5}{2} + 8 \\ x4 = x6/x1 + 12 \\ x5 = (x7 + x8 + x3)^2 - 23 \\ x6 = x3 + x2 - 3 \\ x7 = 2 * x3 + 2 \\ x8 = (x7 - 2)^3 + 2 \end{cases}$$

contains among substitutions «algebraic loops» (for example, x2→x4→x6→x2), which should be detected [3] and removed («cutting algebraic loops»). Let us apply described algorithm to given system of formal substitutions.

Step 1.

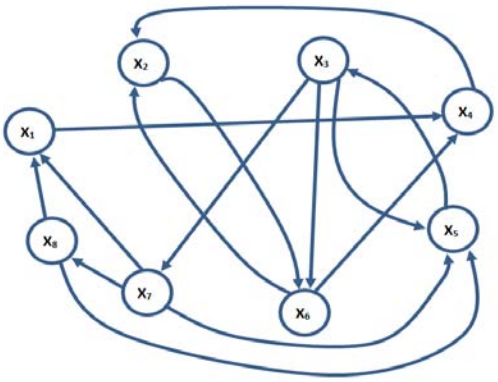


Fig. 3. Initial graph.

Step 2.

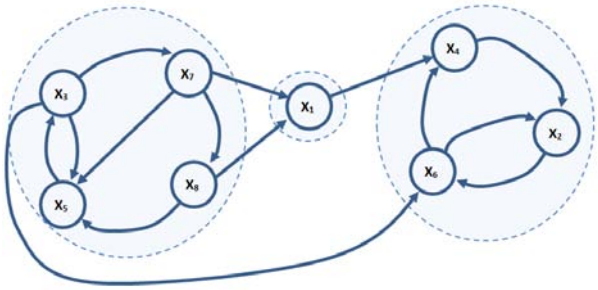


Fig. 4a. Strongly connected components detected by Tarjan's algorithm.

Edge's orientations (Fig 4.b) says about existance block-triangular form.

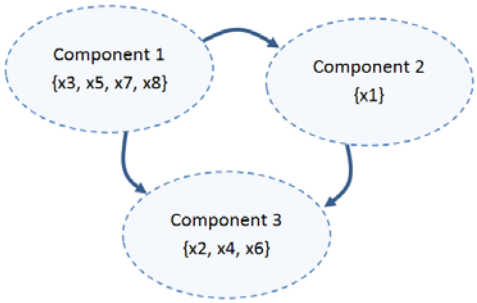


Fig. 4b. Links between strongly connected components.

Component 2 corresponds to substitution, components 1,3 – to equations.

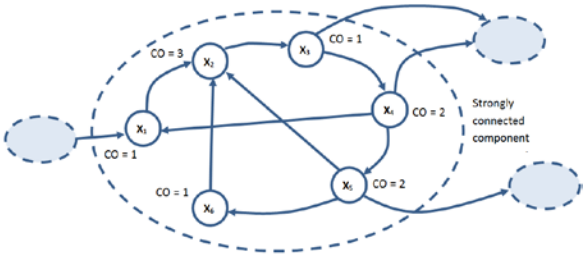
«Algebraic loops»

1. $x_3 \rightarrow x_5 \rightarrow x_7 \rightarrow x_8 \rightarrow x_3$
2. $x_2 \rightarrow x_4 \rightarrow x_6 \rightarrow x_2$

have to be founded into components 1,3.

It is obvious that «algebraic loop» may be cutting in different ways. We will choose equation applying principle of maximum cycling order for graph's nodes. Cycling order (CO) for node) is OC = max(number of outgoing arc, number of incoming arc) .

Outgoing and incoming arc have leading to nodes belong to strongly connected component (Fig. 5).



Fiq. 5. Strongly connected component with calculated cycling orders.

ISBN: 978-1-61804-251-4

Step 3. Calculating cycling orders. Fig. 6.a, 6.b.

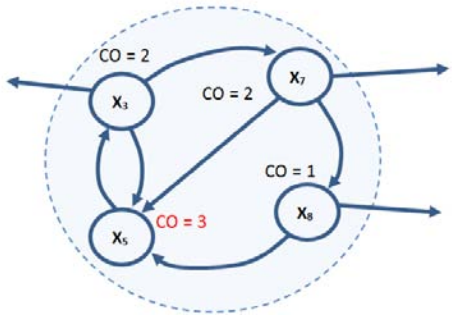


Fig. 6.a. Cycling orders for component 1.

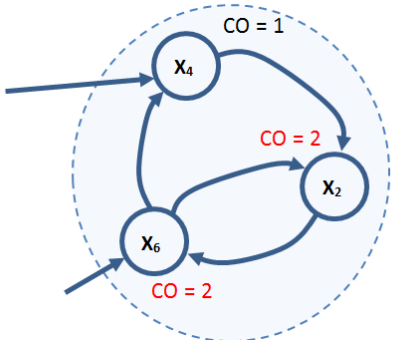


Fig. 6.b. Cycling orders for component 3.

Component 1 has only one node with maximum cycling order (node x_5, CO=3). Corresponding «substitution» becomes «equation».

Component 3 has two nodes with maximum cycling orders. Node x_2 was the first in Tarjan's algorithm list of nodes belong to Component 3. So corresponding «substitution» becomes «equation».

Step 4. Remove nodes marked as equations

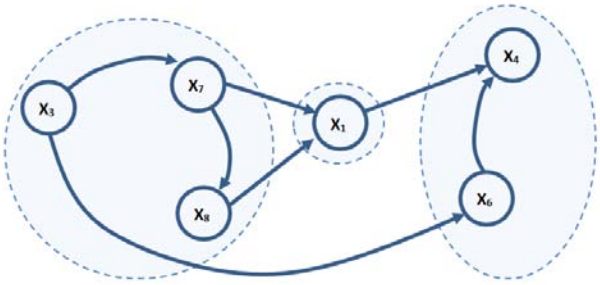


Fig. 7. Graph after removing «node-equation»

After removing nodes Component 1 and component 3 become not strongly connected components.

Step 5. Restart Tarjan's algorithm only for Components 1,3. Final graphs is shown on Fig. 8. Strongly connected components have dimensions 1×1, therefore all «algebraic loops» are detected and corresponding «substitutions» are announced as «equations».

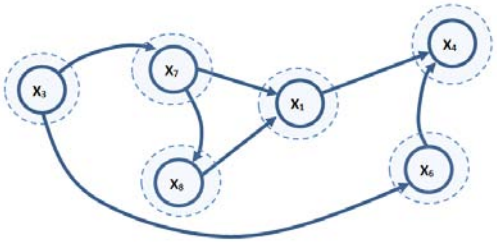


Fig. 8. Final graph.

Step 6. Building system of equations and substitutions

- Formal substitutions announced as equations:

$$\begin{cases} x2 = x4*3 - 2*x6 + 1 \\ x5 = (x7 + x8 + x3)^2 - 23 \end{cases}$$

- *formally calculable* sequance

$$\begin{cases} x3 = \dfrac{x5}{2} + 8 \\ x7 = 2*x3 + 2 \\ x8 = (x7 - 2)^3 + 2 \\ x1 = -7*x7 + x8 - 2 \\ x6 = x3 + x2 - 3 \\ x4 = x6/x1 + 12 \end{cases}$$

IV. Conclusion

Suggested algorithm was tested on TRANSAS' problems (www.transas.com) and demonstrated high performance. Extended version of this paper will contain results of computations experiments with TRANSAS models.

References

[1] Tarjan R. E. Depth-first search and linear graph algorithms. SIAM Journal on Computing. – 1972. – № 2. – p. 146–160

[2] Duff I. S. MA28 – a set of FORTRAN subroutines for sparse unsymmetric linear equations, 1977.

[3] The MathWorks Inc. Simulink. Simulation and Model-Based Design. – Eighth Printing – 2005.

[4] Pissanetzky S. Sparse matrix technology. – 1984.

Isakov Andrey A. -- MPhil (Computer Science, 2012, SPBSTU), postgraduate student of Distributed Computing and Networking Computing department, Saint Petersburg state Polythecnical University.

Senichenkov Yuri B. – DPhil (Numerical software, 2005, SPBSTU), Professor of Distributed Computing and Networking Computing Department, Saint Petersburg state Polythecnical University.

Dynamic model of the inverted pendulum on a mobile base with two active wheels and desing of an control law

J. E. Moisés Gutiérrez , *Research fellow, BUAP,* J. Gabriel Escamilla, *Student researcher, BUAP,*
J. Eladio Flores, *Research fellow, BUAP,* M. Montserrat Morín, *Research fellow, BUAP,*
and Josefina Castañeda, *Research fellow, BUAP*

***Abstract*—This paper presents the analysis of an inverted pendulum with mobile base, for obtain the dynamic model that describe the behavior ours system using the Newton-Euler method. In the analysis is considered an inertial frame reference fixed to floor, from this frame reference are measured all centers mass, for example; the center the axis of wheels, the wheels (left and right) and the mass of inverted pendulum. The motion of the wheels is considered independently, therefore, we have two frames of reference in each the wheels (left and right, respectively), further, there is a reference frame at the upper part which represents the mass of the pendulum.**
For represent the dynamic model the inverted pendulum in the space of states is necessary perform the linearization around the points equilibrium, this because it the dynamic model the inverted pendulum is an nonlinear system.
After obtaining the linear model of pendulum and the representation in the space state,we obtain a control law for stabilize around vertical position of equilibrium the inverted pendulum, as in the analysis are considered the motion of the wheels independent manner, so that has two signals of control.

***Index Terms*—dynamic model, space state, control law, Newton-Euler method, inverted pendulum, linealization.**

I. Introduction

The inverted pendulums are a family of devices that constitute a bank of test very comprehensive and interesting for nonlinear control engineering. The most studied of this family is the so-called control of pendulum inverted on a mobile base or vehicle. Consists of a pendulum or rod which rotates freely by one of their ends by a joint located on a carriage, that moves horizontally under the action of a controlling force which acts on the pendulum's position [1].
The inverted pendulum (Fig. 1) is an unstable system, because it can fall at any time unless is applied a suitable control force [2].

II. Deduction of the dynamic model

In this section is presented the Newton-Euler methodology for obtaining the mathematical model. Are made the following considerations:

- The wheels move in the plane **XY**.
- The wheels not have lateral slippage.
- The wheels fulfill the condition of bearing.
- The gravitational force is given by **mg**.
- The viscous force air is not considered.
- Is a *multi-bodies* system consisting of two wheels and the pendulum.

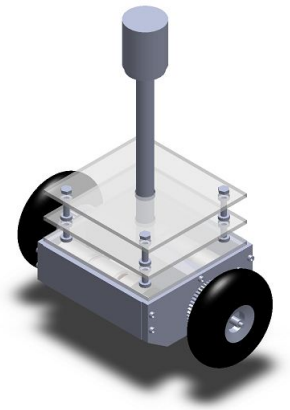


Fig. 1. Inverted pendulum with mobile base.

A. Approach of the Newton-Euler methodology

The equations describing the Newton-Euler methodology are given by

$$F_r = ma_{cm} \tag{1}$$

$$H_O = I_O\alpha \tag{2}$$

where F_r is the resultant force applied, m is the mass of the body, a_{cm} is the acceleration of the center of mass of the body, H_O is the sum of pares, I_O is inertia and α is the angular acceleration [3], [4], which are measured from an inertial reference frame.

B. Systems of reference

Due to that equations (1) and (2) are valid in an inertial system reference, we define this system as

$$\{O, \hat{\eta}_1, \hat{\eta}_2, \hat{\eta}_3\} \tag{3}$$

As is a *multi-bodies* system, we have four rigid bodies, which represent the axis center of wheels, the left wheel, right wheel and the mass of pendulum, a fixed reference system defined the center of mass of each body (which is a strategy that facilitates the description of the motion of each body).

ISBN: 978-1-61804-251-4

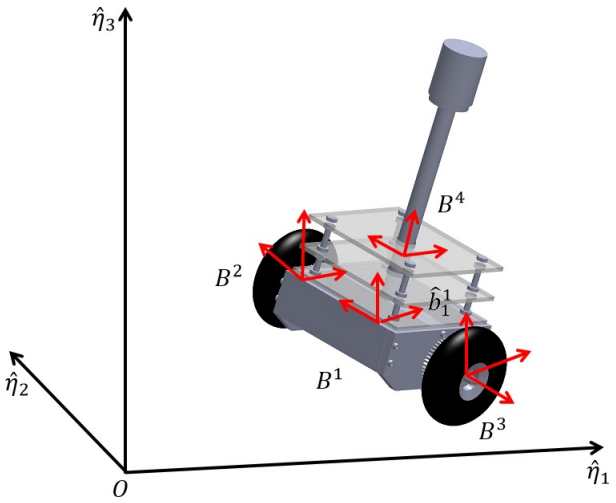


Fig. 2. Multi-bodies system of inverted pendulum.

$$\begin{array}{ll} \{B^1, \hat{b}_1^1, \hat{b}_2^1, \hat{b}_3^1\} & \text{Fixed to axis of the wheels.} \\ \{B^2, \hat{b}_1^2, \hat{b}_2^2, \hat{b}_3^2\} & \text{Fixed to wheel left.} \\ \{B^3, \hat{b}_1^3, \hat{b}_2^3, \hat{b}_3^3\} & \text{Fixed to wheel right.} \\ \{B^4, \hat{b}_1^4, \hat{b}_2^4, \hat{b}_3^4\} & \text{Fixed to pendulum.} \end{array} \tag{4}$$

The vectors $\hat{\eta}_i$ and $\hat{b}_j^i$ are unit vectors, while O and B^i are the origins of these reference systems. To determine the position of a body is required six coordinates; three for locating its center of mass and three to determine its orientation.

$$(q_1^i, q_2^i, q_3^i, q_4^i, q_5^i, q_6^i,) = (x_i, y_i, z_i, \phi_i, \theta_i, \psi_i) \tag{5}$$

C. Degrees of freedom

The degrees of freedom of any *multi-bodies* system is determined from the following expressions

$$DOF = 6n_b - n_c \tag{6}$$

where n_b is the number of rigid bodies that make up the system and n_c is the number of constraints to which the system is subjected.
For our $n_b = 3$ and $n_c = 5$

$$DOF = 6(3) - 15 = 3 \tag{7}$$

This by the following restrictions:

- At points B^2 and B^3 we have: three equations of constraint of translational in B^2 y three equations for rotation in B^2 .
- We have three constraint equations of translational and three rotational in B^3.
- We have an equation of bearing in the left wheel.
- We have an equation of bearing in the right wheel.
- We have an equation of no lateral sliding in both wheels.

D. Position

To find the 15 equations, we must first define the position of reference system relative to the inertial reference system.

$$\begin{aligned} q_1 &= (X_1^0, Y_1^0, Z_1^0, \phi_1, \theta_1, \psi_1)^T = R_1^0, \Omega_1^0 \\ q_2 &= (X_2^0, Y_2^0, Z_2^0, \phi_2, \theta_2, \psi_2)^T = R_2^0, \Omega_2^0 \\ q_3 &= (X_3^0, Y_3^0, Z_3^0, \phi_3, \theta_3, \psi_3)^T = R_3^0, \Omega_3^0 \\ q_4 &= (X_4^0, Y_4^0, Z_4^0, \phi_4, \theta_4, \psi_4)^T = R_4^0, \Omega_4^0 \end{aligned}$$

In each of these

$$R_i^0 = (x_i^0, y_i^0, z_i^0)^T \tag{8}$$

where R_i^O is the vector from O to B^i.

while

$$\Omega_i = (\phi_i, \theta_i, \psi_i)^T \tag{9}$$

It gives us the orientation of the axes of B^i with respect to the axes of O. The angles are called *angles Tait-Brayn* or *navigation angles*.

E. Rotation matrix

The rotation matrix represent the orientation of a moving system B^i with respect to an inertial system O and given by

$$J_i^0 = \begin{bmatrix} \cos\psi_i \cos\theta_i & -\sin\psi_i \cos\phi_i + \cos\psi_i \sin\theta_i \sin\phi_i & \sin\psi_i \sin\phi_i + \cos\psi_i \cos\phi_i \sin\theta_i \\ \sin\psi_i \cos\theta_i & \cos\psi_i \cos\phi_i + \sin\phi_i \sin\theta_i \sin\phi_i & -\cos\psi_i \sin\phi_i + \sin\theta_i \sin\psi_i \cos\phi_i \\ -\sin\theta_i & \cos\theta_i \sin\phi_i & \cos\theta_i \cos\phi_i \end{bmatrix} \tag{10}$$

Applying (8) and (10), we have

$$R_1 = \begin{cases} R_1^0 = (x_1^0, y_1^0, 0)^T \\ R_1^1 = (x_1^1, y_1^1, 0)^T \end{cases} \tag{11}$$

$$R_2 = \begin{cases} R_2^0 = (x_2^0, y_2^0, 0)^T \\ R_2^1 = (x_2^1, y_2^1, 0)^T \end{cases} \tag{12}$$

$$R_3 = \begin{cases} R_3^0 = (x_3^0, y_3^0, 0)^T \\ R_3^1 = (x_3^1, y_3^1, 0)^T \end{cases} \tag{13}$$

$$R_4 = \begin{cases} R_4^0 = (x_4^0, y_4^0, z_4^0)^T \\ R_4^1 = (x_4^1, y_4^1, z_4^1)^T \end{cases} \tag{14}$$

$$r_{12} = \begin{cases} r_{12}^0 = (-L\sin\psi_1, L\cos\psi_1, 0)^T \\ r_{12}^1 = (0, L, 0)^T \end{cases} \tag{15}$$

$$r_{13} = \begin{cases} r_{13}^0 = (L\sin\psi_1, -L\cos\psi_1, 0)^T \\ r_{13}^1 = (0, -L, 0)^T \end{cases} \tag{16}$$

$$r_{14} = \begin{cases} r_{14}^0 = (-C\sin\theta_4\cos\psi_1, -C\sin\theta_4\sin\psi_1, C\cos\theta_4)^T \\ r_{14}^1 = (-C\sin\theta_4, 0, C\cos\theta_4)^T \end{cases} \tag{17}$$

In Figure 3, is showing the reference systems the inverted pendulum.

ISBN: 978-1-61804-251-4

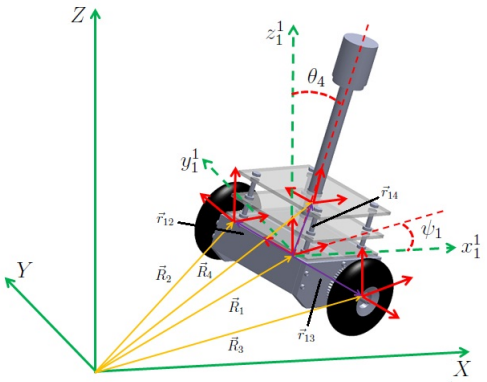


Fig. 3. Reference systems the inverted pendulum.

F. Velocity and Acceleration

The velocity of a point P an rigid body is calculated through an intermediate point in the body O^*, as shown in Figure 4, and by equation

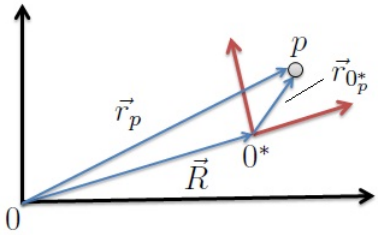


Fig. 4. Velocity compute for a rigid body.

$$r_P = R + r_{O^*P} \tag{18}$$

Temporal variations of both reference systems are related by

$$\frac{dB}{dt} = \frac{d^*B}{dt} + \omega \times \vec{B} \tag{19}$$

where ω is the angular velocity with which turn the O^* system compared to O. This relationship applies for any vector relation B_i.

Using the equation (18) into (19), we have

$$\frac{dr_P}{dt} = \frac{dR}{dt} + \frac{dr_{0^*P}}{dt} = \frac{dR}{dt} + \frac{d^*r_{0^*P}}{dt} + \omega \times r_{0^*P} \tag{20}$$

Applied the relationship (19) into expression (20), is obtained

$$\frac{d^2r_P}{dt^2} = \frac{d^2R}{dt^2} + \frac{d^2r_{0^*P}}{dt^2} = \frac{d^2R}{dt^2} + \frac{d^{*2}r_{0^*P}}{dt^2} + 2\omega \times \frac{d^*r_{0^*P}}{dt} + \frac{d\omega}{dt} \times r_{0^*P} + \omega \times \omega \times r_{0^*P} \tag{21}$$

given that

$$\frac{d\omega}{dt} = \frac{d^*\omega}{dt}$$

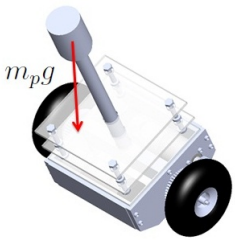


Fig. 5. Force of gravity on the mass of the pendulum.

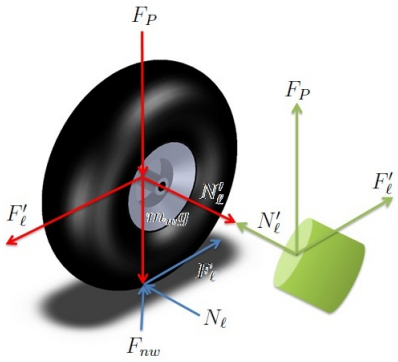


Fig. 6. Force diagram on the left wheel.

Note, the expressions (20) and (21) there is a double equality, in the first equation is more appropriate to use in inertial frame O, while the second equality is more appropriate to use in the representation fixed to body B^1.

G. Dynamic of inverted pendulum

To apply Newton's second law given by equation (1), need the free body diagram, as shown in Figure 5,6 and 7.

Where

F_ℓ	–	Force that exerted the floor on wheel left, that makes the wheel advance forward.
F'_ℓ	–	Force that exerted the pendulum upon wheel left, due to the advancement of the wheel.
F_r	–	Force that exerted the floor on wheel right, that makes the wheel advance forward.
F'_r	–	Force that exerted the pendulum upon wheel right, due to the advancement of the wheel.
F_{nw}	–	Normal force that exerted by the floor on the left and right wheel.
F_P	–	Force that exerted the pendulum upon the wheel left and right.
N_ℓ	–	Force that exerted by the floor on the left wheel, due to lateral movement.
N'_ℓ	–	Reaction force that exerted by the left wheel, before the lateral push of pendulum.
N_r	–	Force that exerted by the floor on the right wheel, due to lateral movement.
N'_r	–	Reaction force that exerted by the right wheel, before the lateral push of pendulum.
m_pg	–	Force that exerted by the earth on the pendulum.

ISBN: 978-1-61804-251-4

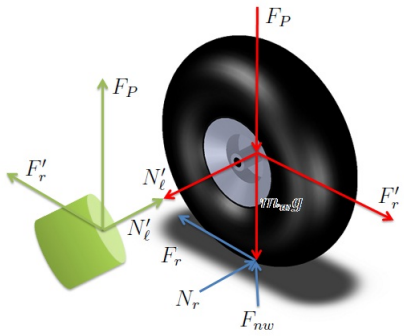


Fig. 7. Force diagram on the right wheel.

To apply the Euler equation, is employed the following representation

$$\tau_{0^*} = I_{cm}\frac{d\omega}{dt} + mr_{cm} \times \frac{d\vartheta_{cm}}{dt} \tag{22}$$

where O^* is the respect system to the which the applied torque is measured, I_{cm} is the tensor of inertia respect to the center of masses, ϑ_{cm} is the velocity of the center of masses measure from the inertial system, but in fixed representation to the body.

H. Dynamic model of inverted pendulum

After applying what mentioned previously, we have the equations describing the motion of the inverted pendulum

$$\frac{\tau_\ell + \tau_r}{r} = (m_p{+}3m_w)\ddot{x}_1^1{+}m_p\ell\sin\theta_4(\dot{\psi}_1^2{+}\dot{\theta}_4^2){-}m_p\ell\ddot{\theta}_4\cos\theta_4 \tag{23}$$

$$\frac{L(\tau_r - \tau_\ell)}{r} = \left\{3m_wL^2 + \frac{m_wr^2}{2} + I_{P3} + m_p\ell^2\sin^2\theta_4\right\}\ddot{\psi}_1 + 2m_p\ell^2\dot{\psi}_1\dot{\theta}_4\sin\theta_4\cos\theta_4 \tag{24}$$

$$-(\tau_\ell + \tau_r) = m_pg\ell\sin\theta_4 + m_p\ell\cos\theta_4\ddot{x}_1^1 + (I_{P2} - m_p\ell^2)\ddot{\theta}_4 + m_p\ell^2\dot{\psi}_1^2\cos\theta_4\sin\theta_4 \tag{25}$$

III. Representation in the State Space

Any dynamic system can be represented by

$$\dot{x} = f(x,u) \tag{26}$$

A set of ordinary linear differential equations is can represented by a set of first order equations. This representation is called *representation in the state space*. A general of expressing the dynamics of a linear system is [3]

$$\dot{x} = Ax + Bu \tag{27}$$

where x is the system state vector, u is the input vector to the system, A is the system matrix and B is the output matrix.

From the equation (23), (24) y (25) we cleared $\ddot{x}_1^1$, $\ddot{\psi}_1$ y $\ddot{\theta}_4$, respectively

$$\ddot{x}_1^1 = \frac{\frac{\tau_\ell - \tau_r}{r} - m_p\ell\sin(\theta_4)(\dot{\psi}_1^2 + \dot{\theta}_4^2) + m_p\ell\cos(\theta_4)\ddot{\theta}_4}{(m_p + 3m_w)} \tag{28}$$

$$\ddot{\psi}_1 = \frac{\frac{L(\tau_r - \tau_\ell)}{r} - 2m_p\ell^2\sin(\theta_4)\cos(\theta_4)\dot{\psi}_1\dot{\theta}_4}{3m_wL^2 + \frac{m_wr^2}{2} + I_{P3} + m_p\ell^2\sin^2(\theta_4)} \tag{29}$$

$$\ddot{\theta}_4 = \frac{-(\tau_\ell + \tau_r) - m_pg\ell\sin(\theta_4) - m_p\ell\cos(\theta_4)\ddot{x}_1^1}{(I_{P2} - m_p\ell^2)} - \frac{m_p\ell^2\cos(\theta_4)\sin(\theta_4)\dot{\psi}_1^2}{(I_{P2} - m_p\ell^2)} \tag{30}$$

we substituted in the equation (30) in (28) and we cleared $\ddot{x}_1^1$

$$\begin{aligned}\ddot{x}_1^1 = {} & \frac{(I_{P2}-m_p\ell^2)(\tau_\ell+\tau_r)}{r(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2r\cos^2(\theta_4)} - \\ & - \frac{(I_{P2}-m_p\ell^2)m_p\ell\sin(\theta_4)(\dot{\psi}_1^2+\dot{\theta}_4^2)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(\theta_4)} - \\ & - \frac{(\tau_\ell+\tau_r)m_p\ell\cos(\theta_4)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(\theta_4)} - \\ & - \frac{((m_p\ell)^2g\cos(\theta_4)\sin(\theta_4)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(\theta_4)} - \\ & - \frac{m_p^2\ell^2\cos^2(\theta_4)\sin(\theta_4)\dot{\psi}_1^2}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(\theta_4)}\end{aligned} \tag{31}$$

Now, we substituted (28) in (30) and we cleared $\ddot{\theta}_4$

$$\begin{aligned}\ddot{\theta}_4 = {} & -\frac{(m_p+3m_w)(\tau_\ell+\tau_r)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(\theta_4)} - \\ & - \frac{(m_p+3m_w)m_pg\ell\sin(\theta_4)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(\theta_4)} - \\ & - \frac{m_p\ell\cos(\theta_4)(\tau_e ll+\tau_r)}{r(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2r\cos^2(\theta_4)} + \\ & + \frac{(m_p\ell)^2\cos(\theta_4)\sin(\theta_4)(\dot{\psi}_1^2+\dot{\theta}_4^2)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(\theta_4)} - \\ & - \frac{(m_p+3m_w)m_p\ell^2\cos(\theta_4)\sin(\theta_4)\dot{\psi}_1^2}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(\theta_4)}\end{aligned} \tag{32}$$

We rewrite the equation (29), as

$$\begin{aligned}\ddot{\psi}_1 = {} & \frac{2L(\tau_r-\tau_\ell)}{6m_wL^2r+m_wr^3+2I_{P3}r+2m_p\ell^2r\sin^2(\theta_4)} - \\ & - \frac{4m_p\ell^2r\sin(\theta_4)\cos(\theta_4)\dot{\psi}_1\dot{\theta}_4}{6m_wL^2r+m_wr^3+2I_{P3}r+2m_p\ell^2r\sin^2(\theta_4)}\end{aligned} \tag{33}$$

Therefore, dynamic model of the inverted pendulum with mobile base is given by equations (31), (32) and (33). Now, we choose the state variables as

$$\begin{aligned}x_1 &= x_1^1 \\ x_2 &= \dot{x}_1^1 \\ x_3 &= \theta_4 \\ x_4 &= \dot{\theta}_4 \\ x_5 &= \psi_1 \\ x_6 &= \dot{\psi}_1\end{aligned} \tag{34}$$

and

$$\begin{aligned}u_1 &= \tau_\ell \\ u_2 &= \tau_r\end{aligned} \tag{35}$$

ISBN: 978-1-61804-251-4

Now, we derived the state variables and we substitute (34) and (35)

$$
\begin{aligned}
\dot{x}_1 &= f_1(x,u) = x_2 \\
\dot{x}_2 &= f_2(x,u) \\
&= \frac{(I_{P2}-m_p\ell^2)(u_1+u_2)}{r(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2 r\cos^2(x_3)} - \\
&- \frac{(I_{P2}-m_p\ell^2)m_p\ell\sin(x_3)(x_6^2+x_4^2)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{(u_1+u_2)m_p\ell\cos(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{((m_p\ell)^2 g\cos(x_3)\sin(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{m_p^2\ell^2\cos^2(x_3)\sin(x_3)x_6^2}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} \\
\dot{x}_3 &= f_3(x,u) = x_4 \\
\dot{x}_4 &= f_4(x,u) \\
&= -\frac{(m_p+3m_w)(u_1+u_2)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{(m_p+3m_w)m_p g\ell\sin(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{m_p\ell\cos(x_3)(u_1+u_2)}{r(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2 r\cos^2(x_3)} + \\
&+ \frac{(m_p\ell)^2\cos(x_3)\sin(x_3)(x_6^2+x_4^2)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{(m_p+3m_w)m_p\ell^2\cos(x_3)\sin(x_3)x_6^2}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} \\
\dot{x}_5 &= f_5(x,u) = x_6 \\
\dot{x}_6 &= f_6(x,u) \\
&= \frac{2L(u_2-u_1)}{6m_wL^2r+m_wr^3+2I_{P3}r+2m_p\ell^2 r\sin^2(x_3)} - \\
&- \frac{4m_p\ell^2 r\sin(x_3)\cos(x_3)x_6x_4}{6m_wL^2r+m_wr^3+2I_{P3}r+2m_p\ell^2 r\sin^2(x_3)}
\end{aligned}
\tag{36}
$$

The linearization of the equations is performed (36), therefore, the elements no null of the state matrix $A=(a_{ij})$, are

$$a_{12} = 1$$

$$
\begin{aligned}
a_{23} = &\frac{m_p\ell\sin(x_3)(u_1+u_2)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{m_p^2\ell^3 x_6^2\cos^3(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{(m_p\ell)^2 g\cos^2(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} + \\
&+ \frac{(m_p\ell)^2 g\sin(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{2(m_p\ell)^4 g\cos^2(x_3)\sin^2(x_3)}{[(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)]^2} + \\
&+ \frac{2m_p^2\ell^3 x_6^2\cos(x_3)\sin^2(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{2m_p^4\ell^5 x_6^2\cos^3(x_3)\sin^2(x_3)}{[(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)]^2} - \\
&- \frac{(I_{P2}-m_p\ell^2)m_p\ell\cos(x_3)(x_6^2+x_4^2)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{2m_p^3\ell^3\cos^2(x_3)\sin(x_3)(u_1+u_2)}{[(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)]^2} - \\
&- \frac{(I_{P2}-m_p\ell^2)2m_p^3\ell^3\cos(x_3)\sin^2(x_3)(x_6^2+x_4^2)}{[(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)]^2} + \\
&+ \frac{(I_{P2}-m_p\ell^2)2m_p\ell^2 r\cos(x_3)\sin(x_3)(u_1+u_2)}{r(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2 r\cos^2(x_3)}
\end{aligned}
$$

$$a_{24} = -\frac{(I_{P2}-m_p\ell^2)2m_p\ell x_4\sin(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)}$$

$$
\begin{aligned}
a_{26} = &-\frac{(I_{P2}-m_p\ell^2)2m_p\ell x_6\sin(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{2m_p^2\ell^3 x_6\cos^2(x_3)\sin(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)}
\end{aligned}
$$

$$a_{31} = 1$$

$$
\begin{aligned}
a_{43} = &\frac{(m_p\ell)^2\cos^2(x_3)(x_6^2+x_4^2)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{(m_p\ell)^2\sin^2(x_3)(x_6^2+x_4^2)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} + \\
&+ \frac{m_p\ell\sin(x_3)(u_1+u_2)}{r(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2 r\cos^2(x_3)} - \\
&- \frac{m_p\ell g\cos(x_3)(m_p+3m_w)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{(m_p+3m_w)m_p\ell^2 x_6^2\cos^2(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} + \\
&+ \frac{(m_p+3m_w)m_p\ell^2 x_6^2\sin^2(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} + \\
&+ \frac{2(m_p\ell)^4\cos^2(x_3)\sin^2(x_3)(x_6^2+x_4^2)}{[(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)]^2} - \\
&- \frac{(m_p+3m_w)2m_p^3\ell^4 x_6^2\cos^2(x_3)\sin^2(x_3)}{[(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)]^2} - \\
&- \frac{(m_p+3m_w)2(m_p\ell)^3\cos(x_3)\sin^2(x_3)}{[(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)]^2} - \\
&- \frac{(m_p+3m_w)2(m_p\ell)^2\cos(x_3)\sin(x_3)(u_1+u_2)}{[(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)]^2} - \\
&- \frac{2(m_p\ell)^3 r\cos^2(x_3)\sin(x_3)(u_1+u_2)}{[r(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2 r\cos^2(x_3)]^2}
\end{aligned}
$$

$$a_{44} = \frac{2(m_p\ell)^2 x_4\cos(x_3)\sin(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)}$$

$$
\begin{aligned}
a_{46} = &\frac{2(m_p\ell)^2 x_6\cos(x_3)\sin(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{(m_p+3m_w)2m_p\ell^2\cos(x_3)\sin(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)}
\end{aligned}
$$

$$a_{56} = 1$$

$$
\begin{aligned}
a_{63} = &\frac{4m_p\ell^2 r x_4 x_6\sin^2(x_3)}{6m_wL^2r+2m_p\ell^2 r\sin^2(x_3)+m_wr^3+2I_{P3}r} - \\
&- \frac{4m_p\ell^2 r x_4 x_6\cos^2(x_3)}{6m_wL^2r+2m_p\ell^2 r\sin^2(x_3)+m_wr^3+2I_{P3}r} + \\
&+ \frac{8m_p\ell^2 Lr\cos(x_3)\sin(x_3)(u_1-u_2)}{[6m_wL^2r+2m_p\ell^2 r\sin^2(x_3)+m_wr^3+2I_{P3}r]^2} + \\
&+ \frac{16m_p^2\ell^4 r^2 x_4 x_6\cos^2(x_3)\sin^2(x_3)}{[6m_wL^2r+2m_p\ell^2 r\sin^2(x_3)+m_wr^3+2I_{P3}r]^2}
\end{aligned}
$$

$$a_{64} = -\frac{4m_p\ell^2 r x_6\cos(x_3)\sin(x_3)}{6m_wL^2r+2m_p\ell^2 r\sin^2(x_3)+m_wr^3+2I_{P3}r}$$

$$a_{63} = \frac{4m_p\ell^2 r x_4\cos(x_3)\sin(x_3)}{6m_wL^2r+2m_p\ell^2 r\sin^2(x_3)+m_wr^3+2I_{P3}r}$$

Therefore, the elements of the output matrix $B=(b_{ij})$, are

$$
\begin{aligned}
b_{21} = &-\frac{m_p\ell^2+I_{P2}}{r(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2 r\cos^2(x_3)} - \\
&- \frac{m_p\ell\cos(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)}
\end{aligned}
$$

$$
\begin{aligned}
b_{22} = &-\frac{m_p\ell^2+I_{P2}}{r(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2 r\cos^2(x_3)} - \\
&- \frac{m_p\ell\cos(x_3)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)}
\end{aligned}
$$

$$
\begin{aligned}
b_{41} = &-\frac{(m_p+3m_w)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{m_p\ell\cos(x_3)}{r(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2 r\cos^2(x_3)}
\end{aligned}
$$

$$
\begin{aligned}
b_{42} = &-\frac{(m_p+3m_w)}{(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2\cos^2(x_3)} - \\
&- \frac{m_p\ell\cos(x_3)}{r(I_{P2}-m_p\ell^2)(m_p+3m_w)+(m_p\ell)^2 r\cos^2(x_3)}
\end{aligned}
$$

$$b_{61} = -\frac{2L}{6m_wL^2r+2m_p\ell^2 r\sin^2(x_3)+m_wr^3+2I_{P3}r}$$

$$b_{62} = \frac{2L}{6m_wL^2r+2m_p\ell^2 r\sin^2(x_3)+m_wr^3+2I_{P3}r}$$

ISBN: 978-1-61804-251-4

By the both matrix A and matrix B, are expressed as

$$A = \begin{bmatrix} 0 & a_{12} & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{23} & a_{24} & 0 & a_{26} \\ 0 & 0 & 0 & a_{34} & 0 & 0 \\ 0 & 0 & a_{43} & a_{44} & 0 & a_{46} \\ 0 & 0 & 0 & 0 & 0 & a_{56} \\ 0 & 0 & a_{63} & a_{64} & 0 & a_{66} \end{bmatrix} \tag{37}$$

$$B = \begin{bmatrix} 0 & 0 \\ b_{21} & b_{22} \\ 0 & 0 \\ b_{41} & b_{42} \\ 0 & 0 \\ b_{61} & b_{62} \end{bmatrix} \tag{38}$$

In the table 1, show the parameters of inverted pendulum, corresponding to the robot the Figure 1

Variable	Value	Descriptión
m_p	4.596	Robot mass $[Kg]$
m_w	0.204	Wheel mass $[Kg]$
L	0.130	Distance from B^1 to the center of the wheels $[m]$
ℓ	0.060	Distance from B^1 to the center mass of pendulum $[m]$
I_{P2}	0.020	Moment of inertia of the wheels $[Kgm^2]$
I_{P3}	0.080	Moment of inertia of the pendulum $[Kgm^2]$
r	0.114	Wheel radius $[m]$
g	9.8	gravitational constant $[m/s^2]$

Tabla 1: Parameters of inverted pendulum with mobile base

We evaluate the values of Table 1 and the equilibrium points in the matrices of equations (37) and (38), we have the linearized system

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -7.9251 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -149.6728 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{39}$$

$$B = \begin{bmatrix} 0 & 0 \\ -2.6103 & -2.6103 \\ 0 & 0 \\ -81.1083 & -81.1083 \\ 0 & 0 \\ -12.4400 & 12.4400 \end{bmatrix} \tag{40}$$

Our equation of state (27) is given of the form

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -7.9251 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -149.6728 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ -2.6103 & -2.6103 \\ 0 & 0 \\ -81.1083 & -81.1083 \\ 0 & 0 \\ -12.4400 & 12.4400 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \tag{41}$$

A. Control system desing

Consider control system, as

$$\dot{x} = Ax + Bu \tag{42}$$

we select the control signal

$$u = -Kx \tag{43}$$

This means that the control signal is determined through an instantaneous state. Such a scheme is called *state feedback*. The matrix K is called the gain matrix of state feedback.

B. Location of the poles

The necessary and sufficient condition for location of the poles, is that the system be completely state controllable. For the selection of the closed-loop poles, we chose a settling time $t_s = 2\ sec$ and an damping factor $\zeta = 0.8$, is calculated the undamped natural frequency ω_n with the criterion of 2%, therefore, settling time is given by

$$t_s = 4T = \frac{4}{\omega} = \frac{4}{\zeta\omega_n} \tag{44}$$

we cleared ω_n and we substituted values of t_s and ζ

$$\omega_n = \frac{4}{t_s\zeta} = \frac{4}{(2)(0.8)} = 2.5 \tag{45}$$

Now, the equation for a second order system is given by

$$\frac{C(S)}{R(S)} = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n + \omega_n^2} \tag{46}$$

we substituted values of ω_n and ζ

$$\frac{C(S)}{R(S)} = \frac{6.25}{s^2 + 4s + 6.25} \tag{47}$$

We take only the denominator of equation (47) and find the roots of polinomial

$$\mu_{1,2} = -2.0 \pm 1.5i \tag{48}$$

As the $\mu_{1,2}$ roots are the dominant poles of system, we elect the other poles $\mu_{3,4,5,6}$ in the complex plane as far apart as possible from the origin, as

$$\mu_{3,4} = -5 \qquad \mu_{5,6} = -10 \tag{49}$$

One time that the desired poles for our systems are chosen, we determined the feedback gain matrix K.

$$K = \begin{bmatrix} 0.6198 & 0.5826 & 0.1861 & -0.1359 \\ 0.6198 & 0,5826 & 0.1861 & -0.1359 \\ -2.0097 & -0,6029 \\ 2.0097 & 0.6029 \end{bmatrix} \tag{50}$$

ISBN: 978-1-61804-251-4

IV. Results

Once determined the feedback gain status matrix K, system performance is determined by simulation of control law. To simulate the dynamics of the system and get the answer to an initial condition given. The equation of state is given by (42) and the control equation given by (43), is substituted the control equation into the equation of state

$$\dot{x} = (A - BK)x \tag{51}$$

Substituting the numerical values of A, B y K in (51), as

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3.2358 & 3.0416 & -6.9536 & -0.7094 \\ 0 & 0 & 0 & 1 \\ 100.5422 & 94.5096 & -119.4858 & -22.0416 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & x_1 \\ 0 & 0 & x_2 \\ 0 & 0 & x_3 \\ 0 & 0 & x_4 \\ 0 & 1 & x_5 \\ -50 & -15 & x_6 \end{bmatrix} \tag{52}$$

The equation of state of the system is obtained by means of equation (52) and the initial conditions that proposed are

$$\begin{bmatrix} x_1(0) \\ x_2(0) \\ x_3(0) \\ x_4(0) \\ x_5(0) \\ x_6(0) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0.1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{53}$$

In Figure 8 show the displacement inverted pendulum x_1^1 to a perturbation in the pitch angle θ_4, in this graphic is seen as car tends to back to the starting position for to stabilize the inverted pendulum.

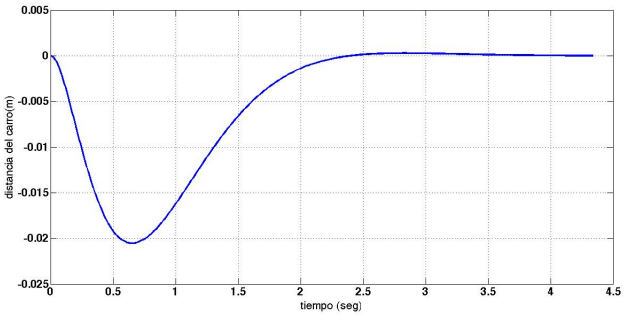


Fig. 8. Displacement inverted pendulum with mobile base.

In Figure 9 the disturbance is shown at the pitch angle θ_4 and as approximately in two seconds is the stabilization the inverted pendulum.

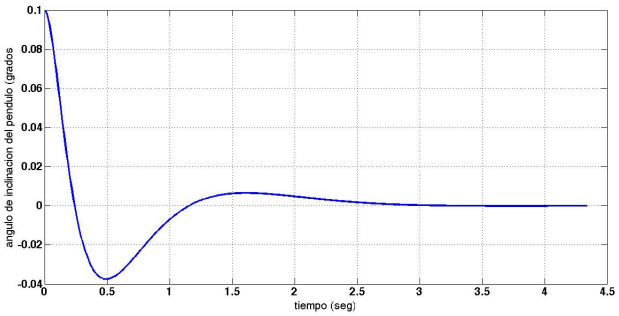


Fig. 9. Pitch angle of the inverted pendulum.

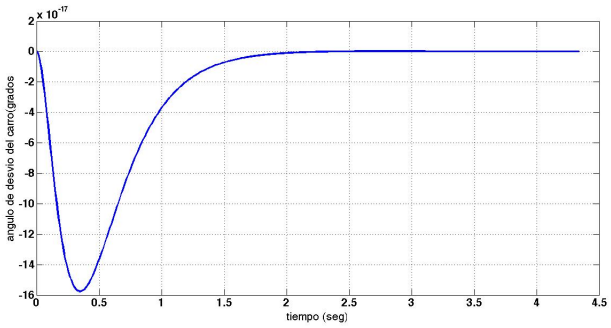


Fig. 10. Yaw angle of inverted pendulum.

In Figure 10 show the changes occurring in the yaw angle ψ_1 caused by the movement of inverted pendulum, although minimum, of the order of $10e^{-17}$.

In Figure 11 is show the velocities corresponding to pendulum, pich and yaw

In Figure 12 shows the graph for the initial condition of $\theta_4 = 0.3$, where you can see how the control system responds in 2 seconds.

The angle ψ_1 allow to parameterize trajectories in later works, as show in Figure 13.

V. Conclusion

Is presented deduction of the dynamic model for an inverted pendulum using the methodology of Newton-Euler,

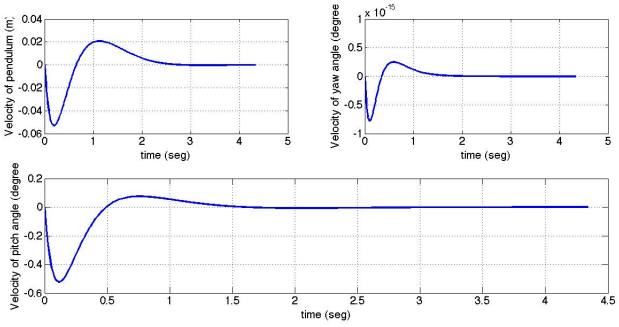


Fig. 11. Velocities of pendulum, yaw and pitch.

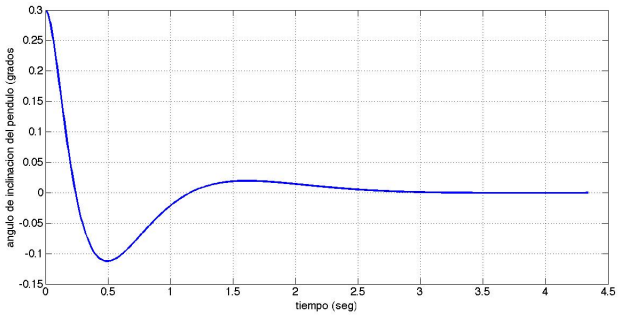


Fig. 12. Pitch angle of the inverted pendulum for $\theta_4 = 0.3$.

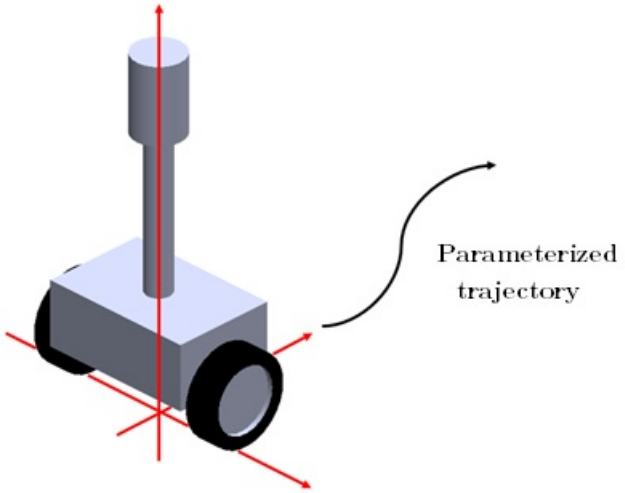


Fig. 13. Trajectory parameterized.

this has nonlinear terms and using control techniques in state-space is only applied to linear systems, accordingly, is realized the linearization of system. This model is unique as it is not yet documented in any publication, so importance for a detailed study of its stability of system and the controllability of system is necessary, furthermore find a control law to stabilize the inverted pendulum with mobile base.

In section of result is show a series of graphs of the pendulum position, the pitch angle of the pendulum and the yaw angle, and their respective velocities, resulting that in the system stabilizes at a time not exceeding two seconds, that is precisely the time that it was determined for the two complex-conjugate poles, which are precisely the dominant poles of the closed-loop system.

Note that this article is only the initial part of the work, as a future work to obtain a new control which allows us to describe parameterized paths in a surface,furthermore the use of dynamic programming and an optimal control algorithm.

ACKNOWLEDGMENT

This work is supported by *Consejo Nacional de Ciencia y Tecnología* (CONACYT), the *Benemérita Universidad Autónoma de Puebla* (BUAP). I appreciate the support by Dr. José E. Moisés Gutiérrez and Dr. José E. Flores,

REFERENCES

[1] J. Aracil and F. Gordillo, *El péndulo invertido, Un desafío para el control no lineal*, Revista iberoamericana de automática e informática industrial., 2005.
[2] Katsuhiko Ogata, *Ingenieria de control moderna*, 5 Ed, PEARSON, 2010.
[3] Terezio Soldovieri C. *Introdución a la mecánica de Lagrange y Hamilton*, La universidad de Zula, 2013.
[4] Ahmed A. Shabana, *Computational Dynamics, Departmet of Mechanical Engineering*, University of Illinois at Chicago, 2001.
[5] Antonio Flores T., *Linealización de funciones no lineales*, 2006.

José E. Moisés Gutiérrez Arias Puebla, Mexico, 1967. He received the titles of the Bachelor of Mathematics and the Master of Sciences Mathematics in 1997 from the Autonomous University of Puebla, the Doctorate in Mathematical Sciences also he received from the same institution in 2003. The purpose of his research is mathematical modeling and analysis of dynamic systems. He Researcher is professor at the Faculty of Electronic Sciences of the Autonomous University of Puebla.

Jesus G. Escamilla Reyes Born on October 21, 1986 in Puebla, Mexico. Graduate of Science in Electronics from the Autonomus University of Puebla (BUAP).
Student in Master of Electronic Engineering (MIE) of the Autonomus University of Puebla (BUAP).

José E. Flores Mena Research fellow in *Faculty of Science of the Electronics (FCE)* of the *Benemérita Universidad Autónoma de Puebla (BUAP)*

M. Montserrat Morín Research fellow in *Faculty of Science of the Electronics (FCE)* of the *Benemérita Universidad Autónoma de Puebla (BUAP)*

Josefina Castañeda Research fellow in *Faculty of Science of the Electronics (FCE)* of the *Benemérita Universidad Autónoma de Puebla (BUAP)*

Multiport Thevenen and Northon theorems analog for ARC-circuits with nonlinear and parametric R-elements

Anatoliy V. Bondarenko[1], Alla A. Lebedeva[2], and Nikolay V. Korovkin[3]
[1)] Saint-Petersburg State University of Architecture and Civil Engineering
, Russia
[2)]St. Petersburg State Polytechnical University, Russia
[3)]St. Petersburg State Polytechnical University, Russia

Keywords - circuit elements, Thevenin theorem, Northon theorem, circuit analyzes for multiports, active realization, linear and nonlinear parametric elements according with theorem.

I. INTRODUCTION.

Well known [1,2] that nonlinear and parametric L-and C-circuit elements can be represented by ARC-networks with nonlinear R-ports (two terminals). That is why suggested sufficiently general conception of nonlinear circuit realization look like circuit system, that is represented on Fig. 1

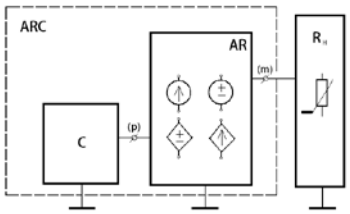


Fig. 1 common block system realization

II. Main part.

On Fig. 1 AR – is active multiport circuit with (m+p+1) output terminals (the last number addressed to common output note).

C – is "star" of C-elements (notes 1÷5); m-terminal belongs to star, consists of nonlinear R_n – one-ports. For realization unbalanced structures had been selected with common node that have some electrical engineering advantages for comparison of general type structures.

As active - elements with depend voltage and current sources can be used (look to block AR) or transforms to them – current and voltage inverters and immittans converters (multiports also included), mutators, nullors and any other systems (2÷4), realized by transistors, chips and hybrid realizations. Thevenen and Norton theorems application are oriented to n-nodes of R-block circuit AR with dependent energy sources and independent voltage and current circuits as well. For example, node quantity that connected with equivalent voltage energy sources may be equal m', and current-m", but m'+m"=m; also possible limit variants: m'=0, m"=m and m'=m, m"=0. If the problem decision for independent sours was formulated early [4], then presents of versatile ruled sources additional explanations needed [5], [6]. And also we receives additional possibilities of multiport system equivalent realization.

This article devoted to one of decision variation announced problem, that connected with introduction of tested independent voltage and current sources in accordance with model demands – equivalents voltage and current sources and/or current sources equivalents.

$$[A(s)][U(s)]=[B(s)]\,[I(s)]+[K(s)], \tag{1}$$

where [A(s)], [B(s)]-some system matrices transforms (s-Laplace operator) sircued structure with P-capacitance elements and [U(s)] and [I(s)] - vector-column voltage and current transforms but there are [K(s)] – some limitations of internal energy sources in AR-block. Equation (1) in other matrices view can be represented

$$[[A(s)],-[B(s)]\cdot[[U(s)]^t,[I(s)]^t]^t=[K(s)]. \tag{2}$$

Some other nonsingular (determinant is not equal zero) had to be introduced into (2) matrix Q, because variable changing is possible and they make combinations that usually used for network (circuit) descriptions through wave parameters. In that case we'll receive under $[Q]^{-1}[Q]=[1]$ (r-order unit matrix; t-transpose operation)

$$[[A(s)],-[B(s)]]\cdot[Q]^{-1}[Q]\,[[U(s)]^t,[I(s)]^t]^t=[K(s)]. \tag{3}$$

Let us introduce more general variable notice

$$[Q]\,[[U(s)]^t,[I(s)]^t]^t=[[V_1(S)^t],[V_2(s)]^t]^t,$$

N. V. Korovkin is a head of Electromagnetic Theory Department, St. Petersburg State Technical University, Russia (e-mail: nikolay.korovkin@gmail.com).
A. V. Bondarenko is Dr. of Sci. Tech., Professor academician of AES RF. (Saint-Petersburg State University of Architecture and Civil Engineering, Energy and Electrical Engineering Department) (e-mail: avb38@mail.ru)
A. A. Lebedeva. is Associate professor. (Saint-Petersburg State Pyrotechnical University, Theoretical Fundamentals of Electrical Engineering Department) (e-mail: alla280318@mail.ru)

(4)

And left part matrix multiplication (3) represented through

$$[[A(s)],-[B(s)]]\cdot[Q]^{-1}=[[N(s)],-[M(s)]], \quad (5)$$

then following (4), and excepting matrix[N(S)] nonsingular we'll receive

$$[[1],-[N(s)]^{-1}[M(s)]]\cdot\begin{bmatrix}[V1(s)]\\ [V2(s)]\end{bmatrix}=[N(s)^{-1}]\cdot[K(s)], \quad (6)$$

If we'll receive that independent sources have zero meanings then right part of (6) also transforms to zero matrix – matrix column. By other words that if we have independent source existence and also dependent source absent equation (6) transforms to expression:

$$[[1],-[N(s)]^{-1}[M(s)]]\cdot\begin{bmatrix}[V1_0(s)]\\ [V2_0(s)]\end{bmatrix}=[0], \quad (7)$$

In (7) measure of matrix column is m×1, but variable zeros indexes tell us about independent sources zeros absent. According (4) we receive

$$\begin{bmatrix}[V1_0(s)]\\ [V2_0(s)]\end{bmatrix}=[Q(s)]\cdot\begin{bmatrix}[U0(s)]\\ [I0(s)]\end{bmatrix}, \quad (8)$$

(6) and (7) uniting is

$$[[1]_1-[N(s)]^{-1}[M(s)]]\left(\begin{bmatrix}[V1(s)]\\ [V2(s)]\end{bmatrix}+\begin{bmatrix}[V1_0(s)]\\ [V2_0(s)]\end{bmatrix}\right)=$$

$$=\begin{pmatrix}[N(s)]^{-10}[K(s)]\\ [0]\end{pmatrix}, \quad (9)$$

That's why under zero influence taking in account (7÷9) we'll receive “dead” circuit. Let's introduce matrix column of energy sources testing with one-value currents and voltages $[e]_T=[1,1,...,1]^t_{m\times1}$,then in the right part realize testing (index “t” addresses to the testing operation)

$$\begin{bmatrix}[N(s)]^{-10}[K(s)]\\ [0]\end{bmatrix}=[e]=[[V_{1T}(s)]^t[V_{2T}(s)^t]^t=$$

$$=[1,1,...,1_{m'},1...1\ m^n\ ]^t \quad (10)$$

From that according to (7)

$$[[1],-[N(s)]^{-1}[M(s)]]\begin{pmatrix}[V1_T(s)]\\ [V2_T(s)]\end{pmatrix}=[0] \quad (11)$$

From (11) is clear that

$$[1]\cdot[V_{1T}(s)]-[N(s)]^{-1}[M(s)][V_{2T}(s)]=[0];$$

$$[1]\cdot[V_{1T}(s)]=[N(s)]^{-}1[M(s)][V_{2T}(s)] \quad (12)$$

III. Conclusion.

1) From (12) is clear that ,for example, variables $[V_{1T}(s)]$ are currents with $[V_{2T}(s)]$- sources then nonsingular conductivities matrix of multiport with m-ports is equal

$$[Y(s)]=[N(s)]^{-1}[M(s)]. \quad (13)$$

2) In other case $[V_{1T}(s)]$-voltage vector-column and $[V_{2r}(s)]$-current vector-column – we'll receive that [Z(s)]. In intermideal cases we'll have generalized hybrid matrix [H(s)] or wave [W(s)].
Using this descriptions is possible to construct calculate model system, received after disconnections nonlinear one-port. In this case on m-port realize dependent energy sources.

3) At last if we have a combinations of dependent and independent energy sources then according to circuit linearity after pospond disconnection nonlinear R-elements on m-port in general case there will be dependent and independent voltage and current sources according to fig.2, where dependent sources are represented by romp figure and independent by circle:

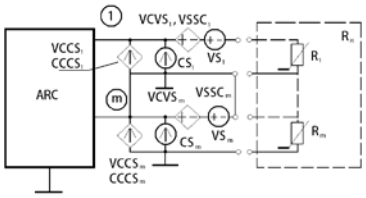


Fig.2 energy source on dedicated ports chain

Here there are some constructions:VS-voltage control voltage source, VSCC-voltage source current control, VCCS-voltage control current source, CCCS-current control current source; CS, VS- independent current and voltage sources respectively.

Lets illustrate by example.

Example. Lets combine two-port model after switch off two-nonlinear resistors with VC characteristics respectively ${}^{i}H_1=f_1(U_1)$ and ${}^{i}H_2=f_2(U_2)$ according to fig.3. Here we see VSCC ($\beta\cdot I_2$), VCVS ($\alpha\cdot U_1$).

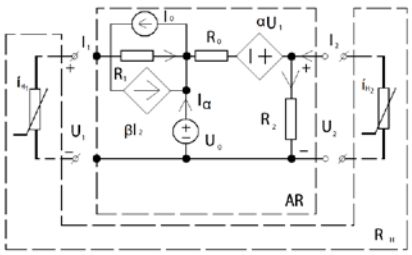


Fig.3 active-resistive (AR) and non-linear (R_H) sunchain

Lets take Norton theorem variant. For analyses' simplification we select pure resistive circuit represented on fig.3 through AR. Simple linear circuit analyses results (reactive element will not cardinally complicate decision with all variants are illustrated without s-operator.) to system equations:

$$\{I_1=\frac{U_1-O_0}{R_1}\ -I_0+\beta I_2$$

$$\{\ I_2=\left(\frac{1}{R_2}+\frac{1}{R_0}\right)U_2-\frac{\alpha}{R_0}\ U_1-\frac{U_0}{R_0}\ .$$

Though according (1) we receive:

$[B(s)]\cdot[I(s)]=[A(s)]-[K(s)]$ For this concept example

$$\begin{bmatrix}1-\beta\\0\ \ 1\end{bmatrix}\ \begin{bmatrix}I_1(s)\\I_2(s)\end{bmatrix}=\begin{bmatrix}\frac{1}{R_1}\ & 0\\ \frac{-\alpha}{R_0}\ & \left(\frac{1}{R_0}+\frac{1}{R_2}\right)\end{bmatrix}$$

$$\begin{bmatrix}U_1(s)\\U_2(s)\end{bmatrix}+\begin{bmatrix}\frac{-1}{R_1}\ & -1\\ \frac{-1}{R_0}\ & 0\end{bmatrix}\begin{bmatrix}U_0(s)\\I_0(s)\end{bmatrix}.$$

Here

$$[B(s)]=\begin{bmatrix}1 & -\beta\\0 & 1\end{bmatrix};[A(s)]=\begin{bmatrix}\frac{1}{R_1}\ & 0\\ \frac{-\alpha}{R_0}\ & \left(\frac{1}{R_0}+\frac{1}{R_2}\right)\end{bmatrix};[K(s)]=\begin{bmatrix}\frac{1}{R_1} & 1\\ \frac{1}{R_0} & 0\end{bmatrix}\begin{bmatrix}U_0(s)\\I_0(s)\end{bmatrix}.$$

Lets define currents on output circuit terminals under m=2

$$\begin{bmatrix}I_1(s)\\I_2(s)\end{bmatrix}=\begin{bmatrix}1 & -\beta\\0 & 1\end{bmatrix}^{-1}\cdot\begin{bmatrix}\frac{1}{R_1}\ & 0\\ \frac{-\alpha}{R_0}\ & \left(\frac{1}{R_0}+\frac{1}{R_2}\right)\end{bmatrix}$$

$$\begin{bmatrix}U_1(s)\\U_2(s)\end{bmatrix}+\begin{bmatrix}1 & -\beta\\0 & 1\end{bmatrix}^{-1}\begin{bmatrix}\frac{-1}{R_1}\ & -1\\ \frac{-1}{R_0}\ & 0\end{bmatrix}\begin{bmatrix}U_0(s)\\I_0(s)\end{bmatrix}.$$

Is not difficult to detect that according (6),

$$[V_1(s)]=\begin{bmatrix}I_1(s)\\I_2(s)\end{bmatrix}\ [V_2(s)]=\begin{bmatrix}U_1(s)\\U_2(s)\end{bmatrix}$$

$$-[N(s)]^{-1}\,[M(s)]=-\begin{bmatrix}\left(\frac{1}{R_1}-\frac{\alpha\beta}{R_0}\right)\ & \beta\left(\frac{1}{R_0}-\frac{1}{R_2}\right)\\ \frac{-\alpha}{R_0}\ & \frac{1}{R_0}+\frac{1}{R_2}\end{bmatrix};$$

$$[N(s)]^{-1}[k(s)]=\begin{bmatrix}\left(\frac{1}{R_1}+\frac{\beta}{R_0}\right)\ & 1\\ \frac{1}{R_0}\ & 0\end{bmatrix}\begin{bmatrix}U_0(s)\\I_0(s)\end{bmatrix}.$$

under $[Q]=\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$, with taking into account (12),(13) we get conductance matrix

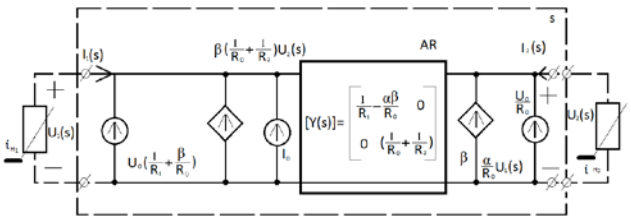
$$[Y(s)]=[N(s)]^{-1}[M(s)]=\begin{bmatrix}\left(\frac{1}{R_1}-\frac{\alpha\beta}{R_0}\right)\ & \beta\left(\frac{1}{R_0}-\frac{1}{R_2}\right)\\ \frac{-\alpha}{R_0}\ & \frac{1}{R_0}+\frac{1}{R_2}\end{bmatrix}.$$

That's finally we receive

$$\begin{bmatrix}I_1(s)\\I_2(s)\end{bmatrix}=[Y(s)]\begin{bmatrix}U_1(s)\\U_2(s)\end{bmatrix}+\begin{bmatrix}\left(\frac{-1}{R_1}-\frac{\beta}{R_0}\right)\ & -1\\ \frac{-1}{R_0}\ & 0\end{bmatrix}\begin{bmatrix}U_0(s)\\I_0(s)\end{bmatrix}$$

Realization [Y(s)] by concret ways is possible. One of then is shown on fig.4 AR-circuit detects by touches.

Fig.4 Norton theorem analog



Other variant may be describe by symmetric

conductance matrix

$$[Y(s)]I_1=\begin{bmatrix} \left(\frac{1}{R_1}-\frac{\alpha\beta}{R_0}+\frac{\alpha}{R_0}\right) & \frac{-\alpha}{R_0} \\ \frac{-\alpha}{R_0} & \left(\frac{1}{R_0}-\frac{1}{R_2}+\frac{\alpha}{R_0}\right) \end{bmatrix}.$$

In this case control source in right network part fig.4 escaped, but under passive reaction is necessary for realization conditions fullfilmed-matrix has to be dominant diagonal with negative out of diagonal elements that suggests unequalities:

$$\alpha(1-\beta)< \frac{R_0}{R_1}\ ;\ \alpha \geq 0;$$

$$\frac{R_0}{R_0} > \alpha\,\beta$$

When inside constant sources are absent ($U_{0,}I_0=0$), curcuid fig.4 become dead and for definitions inside immanences with control sources necessary to address to equations (11) and (12) under switch on to outside switch port unit voltage sources. Calculated currents make it possible to define nodecondactent parameters. In conclusion we want to say that suggested realization method with nonlinear resistors may be distributed also to time variable elements logistic subsidiarity and their combinations.

References:

[1] Charles A. Desolater and Ernest S. Kusch Basic Circuit Theory. MC Grow-Hill Book Company, NY. 1969. 876 p.

[2] Aram Budac Passive and Active Network Analysis and Syntheses, Boston. 1974. 733 p.

[3] A. V. Bondarenko Muliport realization with symmetric amplitude-frequency characteristic. Civi Engineering, SPb. 2011. 117-121p

[4] A. V. Bondarenko Electrotechnica.(Electrical engineering) Publication by SpbGACU, Spb. 2009, 406 p.

[5] N. V. Korovkin, A. A. Lebedeva, T. G. Minevich, K. I. Netreba , S. L. Shishigin Synthesis of RLC models grounding devices on the experimental and calculated transient response. Scientific and technical statements of the St. Petersburg State Polytechnic University. 2009. 202-207 p.

[6] K. S. Demirchan, L. R. Neiman, N. V. Korovkin, V. L. Chechurin. Theoretical Fundamentals of Electrical Engineering. Spb.: Peter. 2003.

Nikolay V. Korovkin, professor, is currently head of Electromagnetic Theory Department of St. Petersburg State Polytechnic University (SPBSPU). He received the M.S., Ph.D. and Doctor degrees in electrical engineering, all from SPbSPU in 1977, 1984, and 1995 respectively, academician of the Academy of Electrotechnical of Russian Federation, (1996) Invited Professor, Swiss Federal Institute of Technology (EPFL), Lausanne (1997), Professor, University of Electro-Communications, Department of Electronic Engineering, Tokyo, Japan (1999-2000), Professor EPFL (2000-2001), Otto-fon-Guericke University, Germany (2001-2004). Head of the Program Committee of the Int. Symp. on EMC and Electromagnetic Ecology in St. Petersburg, 2001-2011.

His main research interests are in the inverse problems in electro-magnetics, optimization of power networks, transients in transmission line systems, impulse processes in linear and non-linear systems, "soft" methods of optimization, systems described by stiff equations, the problems of the electromagnetic prediction of earthquakes and identification of the behavior of the biological objects under the influence of the electromagnetic fields

Anatoliy V. Bondarenko, professor, head of Electrical Engineering Department of Saint-Petersburg State University of Architecture and Civil Engineering, (1997-2012). He received M.S. Ph. D and Doctor degrees in electrical engineering in 1962, 1981 and 1983 respectively. Academician of the Academy of Electrotechnical of Russian Federation . Invited professor of Michigan university (1975-1976).

His main research interests are in the modern methods of analysis and synthesis active linear and non-linear , parametric and logic systems. Hibrid circuits realizations and problems of optimization.

Alla A. Lebedeva, Associate Professor of Theoretical Fundamentals of Electrical Engineering Department of St Petersburg State Polytechnic University (SPBSPU). She received M.S. Ph.D in electrical engineering in 1992, 2012 respectively.

His main research interests are in the modern methods of analysis and synthesis active linear and non-linear , parametric and logic systems. Hibrid circuits realizations and problems of optimization.

Time-dependent mesodiffusion through a boundary: the current inversion phenomenon

V. V. Uchaikin, R. T. Sibatov
Ulyanovsk State University
Email: vuchaikin@gmail.com

Abstract—We study the behavior of a diffusion packet near the boundary separating media characterized by inverse α-power type distribution (α-*medium*) and by exponential distribution (e-medium) of free path lengths. The inversion of current is observed near the boundary at the source side. The phenomenon can be interpreted as a stochastic reflection of the packet front from the dense medium.

I. Introduction

Traditionally, systems of interest to physicists have been divided into the macroscopic and the microscopic realms, where the latter implies atomic and molecular sizes or smaller. Recently, research in the intermediate – mesoscopic – regime has achieved significant scientific successes. The field is characterized by the need to use the microscopic laws of quantum mechanics, while, on the other hand, the samples can be made and operated by essentially ordinary macroscopic methods. This involves linear size scales from a few to thousand of atoms, and reliable fabrication and analysis methods exists down to the scale of about fifty atoms. The term "nano" characterizes the low end of this range.

A piece of solid can be considered as macroscopic body if all its properties are either scale independent, or can be written as intensities which are then scale independent. For example, a piece of Cu has a specific heat independent of its size, a metallic wire has a specific resistivity per unit length, etc. If a bulk solid is cut down into smaller pieces, there will be a critical size (in one, two or three dimension) below which some of the properties can no longer be described by intensities. This is the regime of mesoscopic physics. Trajectories of diffusing particles become differ from their brownian counterparts, intercollision free paths appear, inhomogeneities been invisible on macroscales become now an essential factor of the process, grew up sample-to-sample variations (mesoscopic fluctuations). Godoy and Garcia-Colin [1] named this process "mesoscopic diffusion" (we will use term *mesodiffusion*, for short).

Attention to this model grew up in connection with mass, charge and heat transport in mesoscopic systems or nanosystems described by so-called *extended irreversible thermodynamics* (EIT) (see Chapter 11 in [2]). EIT provides generalization of diffusion equations by incorporating memory and nonlocal effects with account of a finite speed of propagation of heat or charge pulses, and general features of ballistic behavior for long mean-free paths. Thus, they provide not only small corrections to the usual transport equations, but they turn out to be useful even under extreme non-equilibrium situations where the classical equations completely fail.

Some properties of the mesodiffusion process in a homogeneous media (because of exponential free path distribution we call it e-*medium*) were investigated in works [3], [4]. Later, we considered the case when the process is characterized not by exponential free path distribution as in e-medium but by inverse α-power type (α-*medium*) [5]. The present paper relates to inhomogeneous case, namely to the transport near the α-e boundary.

II. Normal and anomalous mesodiffusion

One-dimensional mesodiffusion of a particle in a uniformly homogeneous media conforms to the walk process along x-axis with a constant speed v changing its motion direction at the end of each random free path R such that

$$\mathsf{P}(R > x) = P(x) = \int_x^\infty p(x')dx'.$$

Let $X(t)$ be the random realization of the walker trajectory, $X(0) = 0$, and $p(x,t)dx = \mathsf{P}(X(t) \in dx)$. In case of a uniformly homogeneous media $P(x) = e^{-\sigma x}$ (call it e-*medium*) and pdf $p(x,t)$ obeys the telegraph equation in dimensionless time units having the form

$$\frac{\partial^2 p}{\partial t^2} + \frac{\partial p}{\partial t} = D\frac{\partial^2 p}{\partial x^2} + \delta(x)\delta(t), \tag{1}$$

where $D = v/2\sigma$ is the diffusion coefficient. Recall that it has resulted from coupling two first-order equations

$$\frac{\partial p}{\partial t} + \frac{\partial j}{\partial x} = 0,$$

$$j = -D\frac{\partial p}{\partial x} - \theta\frac{\partial j}{\partial t}.$$

The first of them is a continuity equation and the second one (with $\theta = D/v$) is the Cattaneo-Maxwell $j - p$ interrelation generalizing the classical Fick law.

The solution to Eq. (1) consists of two parts,

$$p(x,t) = p^{(0)}(x,t) + p^{(s)}(x,t).$$

first of which describes two scattering delta-pulses

$$p^{(0)}(x,t) = \frac{1}{2}\left[\delta(x - vt) + \delta(x + vt)\right] e^{-\sigma vt},$$

ISBN: 978-1-61804-251-4

while the second one gives the continuous component of solution filling the interval $(-vt, vt)$:

$$p^{(s)}(x,t) = \sigma \left[I_0 \left(\sqrt{(t^2 - x^2/v^2)/4} \right) + tI_1 \left(\sqrt{(t^2 - x^2/v^2)/4} \right) / \sqrt{t^2 - x^2/v^2} \right] e^{-\sigma vt}. \quad (2)$$

Observe that out the interval, $p(x,t) = 0$.

We considered another case as well: mesodiffusion in a quasihomogeneous (disordered) media possessing fractal signs [7], [8], [6], when the tail of the free path distribution has a power-law character with exponents $\alpha \in (0,1)$ (call it α-*medium*,

$$\mathsf{P}(R > r) \sim \frac{A}{\Gamma(1-\alpha)} r^{-\alpha}, \ r \to \infty.$$

This process is governed in long-time asymptotics by fractional differential equation

$$\left[\left(\frac{\partial}{\partial t} - v\frac{\partial}{\partial x} \right)^{\alpha} + \left(\frac{\partial}{\partial t} + v\frac{\partial}{\partial x} \right)^{\alpha} \right] \Phi(x,t) = \frac{t^{-\alpha}}{2\Gamma(1-\alpha)} \left[\delta(x - vt) + \delta(x + vt) \right].$$

Here,

$$\left(\frac{\partial}{\partial t} + v\frac{\partial}{\partial x} \right)^{\alpha} f(x,t) = = \frac{1}{\Gamma(1-\alpha)} \left(\frac{\partial}{\partial t} + v\frac{\partial}{\partial x} \right) \int_{-\infty}^{t} \frac{f(x - v(t-\tau), \tau)}{(t-\tau)^{\alpha}} d\tau, \quad 0 < \alpha < 1,$$

is the fractional material derivative.

Solution of the latter equation is expressed through the Lamperti distribution in terms of elementary functions (see [9], [5], [10], [11])

$$\Phi(x,t) = \frac{2\sin\pi\alpha}{\pi} \times \frac{\left(1 - x^2/v^2t^2\right)^{\alpha-1}}{(1 - x/vt)^{2\alpha} + (1 + x/vt)^{2\alpha} + 2\left(1 - x^2/v^2t^2\right)^{\alpha} \cos\pi\alpha}.$$

Typical distributions of particles from a point instantaneous source (i.e. *propagators*) in a homogeneous α-medium are presented in Figs. 1 and 2. They are confirmed by Monte Carlo simulation results.

Qualitative distinction of these U- and W-shaped distributions from unimodal diffusion solutions is explained by competition between two regimes: superdiffusion expanding according to the law $\propto t^{1/\alpha}$ in the absence of restrictions, and ballistic motion when the particle position is bounded by segment $[-vt, vt]$. When $\alpha > 1$, the first process dominates at large times: segment $[-vt, vt]$, expanding rapidly, ceases to influence on diffusion. If $\alpha < 1$, the role of kinematic restriction grows and the distribution begins to concentrate near the boundaries of segment $[-vt, vt]$. It should be noted that the reduced asymptotic propagators for $0 < \alpha < 1$

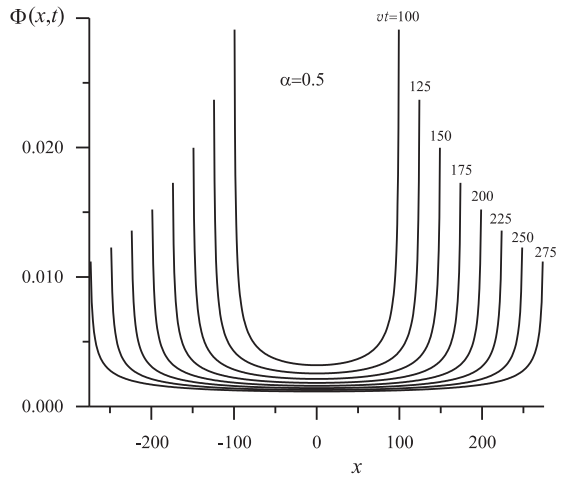


Fig. 1. Long-time evolution of propagator $\Phi(x,t)$ in a homogeneous α-medium ($\alpha = 0.5$).

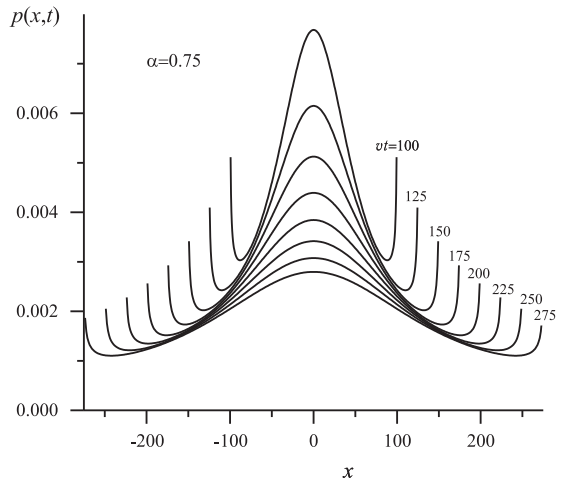


Fig. 2. Long-time evolution of propagator $\Phi(x,t)$ in a homogeneous α-medium ($\alpha = 0.75$).

are universal and do not depend on scale parameters of path length distributions. Scale coefficients determine the rapidity of convergence to the asymptotical Lamperti distribution.

III. Mesodiffusion currents in a homogeneous medium

The Cattaneo-Maxwell $j - p$ interrelation is valid only in case of exponential distribution $P(x)$. In order to generalize it to an arbitrary free path distribution, one should introduce the collision density rate functions $F_+(x,t)$ and $F_-(x,t)$ where + means the motion to the right meanwhile - relates to the particles moving to the left. Product $F_+(x,t)dx$ denotes the mean number of $(- \to +)$-events in dx per unit time and $F_-(x,t)$ denotes the same for $(+ \to -)$-events. These functions are linked via equations

$$F_+(x,t) = \int_0^{vt} F_-(x + \xi, t - \xi/v) p(\xi) d\xi + \frac{1}{2}\delta(x)\delta(t),$$

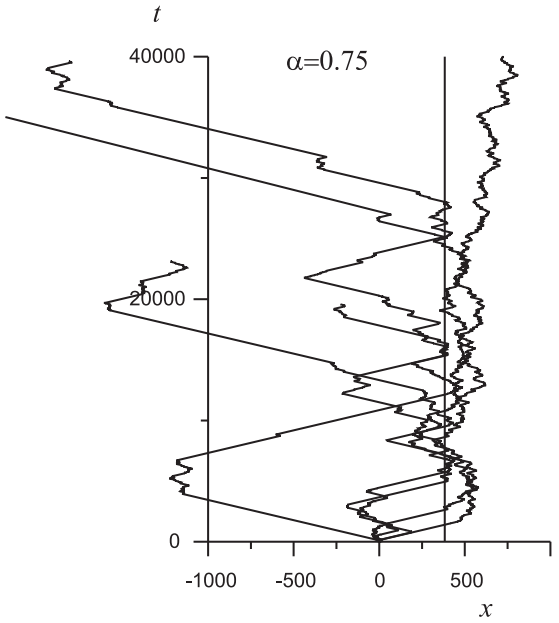


Fig. 3. Typical trajectories of random walks in a two-layer medium ($\alpha = 0.75$ for $x < 400$ and e-medium with $\sigma = 0.2$ for $x > 400$).

$$F_-(x,t) = \int_0^{vt} F_+(x-\xi, t-\xi/v)p(\xi)d\xi + \frac{1}{2}\delta(x)\delta(t),$$

where $p(x) = -P'(x)$. Recall, that in the classical theory (with exponential free path distribution) the functions are proportional to corresponding fluxes but in the general case the interrelation is more complicated. Nevertheless, they form a complete system of integral equations, and convert to a fractional equation system in case of the power-type free path distribution $P(x)$.

The forward and backward currents are expressed via relations

$$j_+(x,t) = \int_0^{vt} F_+(x-\xi, t-\xi/v)P(\xi)d\xi,$$

$$j_-(x,t) = \int_0^{vt} F_-(x+\xi, t-\xi/v)P(\xi)d\xi.$$

With the use $P(x) = e^{-\sigma x}$, the four equation yield for the total current $j = j_+ - j_-$ directly the Cattaneo-Maxwell formula. In case of a power type of $P(x)$, long time asymptotics leads to the fractional expressions

$$j_\pm(x,t) = Av^{1-\alpha}\left(\frac{\partial}{\partial t} \mp v\frac{\partial}{\partial x}\right)^{\alpha-1} F_\pm(x,t).$$

The simplest way to the final solution of the problem lays trough the numerical computing.

IV. Numerical calculations of the boundary transition process

The final aim of this work is investigation of behavior of the diffusion packet near the boundary separating $\alpha-$ and $e-$media and computing interrelation between direct (j_+) and inverse (j_-) components of the particle current. The normal diffusion theory states that the direct component always exceeds the inverse one. However, we suspected that more adequate mesodiffusion model may show departure from this rule. This idea motivated our work.

Evidently, this case is more difficult for analytic solving, but easy for direct simulation by Monte Carlo technique. Combining MC-cods for both media into one and performing calculations, we obtain the following results.

First-particles front reaches the boundary, dives into the second medium, diffuses there, and partially returns. Remarkably, that the inversion of current is observed at the source side near the boundary, when the front of incident packet is large enough. The phenomenon can be interpreted as a stochastic reflection of the packet front from the dense medium. Trajectories (Fig. 3) and bar charts presented in Fig. 4 confirm these reasoning. Concentration of particles moving in negative direction (back to the source) can significantly dominate over concentration of particles moving in positive direction in some region near the boundary. In this region, the anisotropy becomes negative.

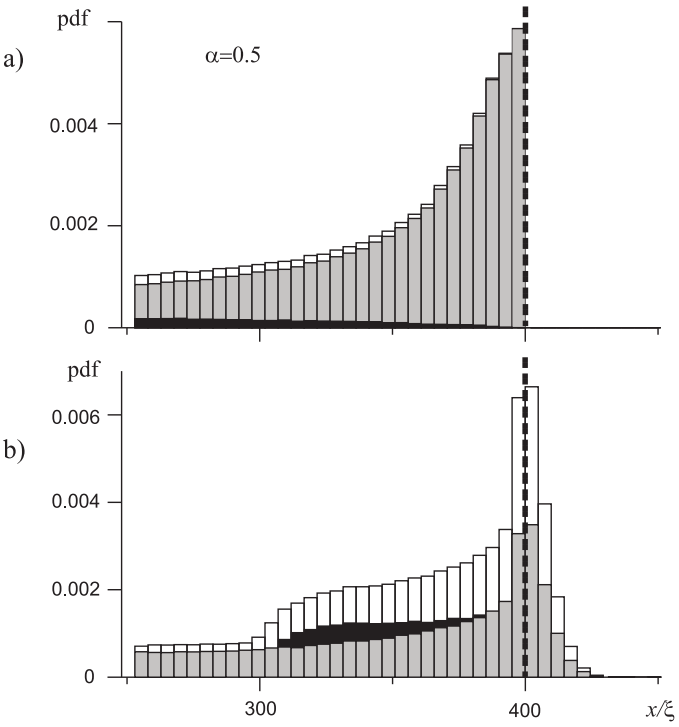


Fig. 4. Bar charts for all particles (white), particles moving in positive (grey) and negative (black) directions. a) The packet front has achieved but not traversed the boundary ($t = 400$, $v = 1$, $b = 400$). b) The packet front has traversed the boundary and partially reflected from the second medium.

The direct calculation of coefficient

$$\delta = \frac{j_+ - j_-}{j_+ + j_-} = \frac{\Phi_+ - \Phi_-}{\Phi_+ + \Phi_-}$$

confirms this idea (see Fig. 5). Here, Φ_+ and Φ_- are concentrations of particles moving in positive and negative directions,

ISBN: 978-1-61804-251-4

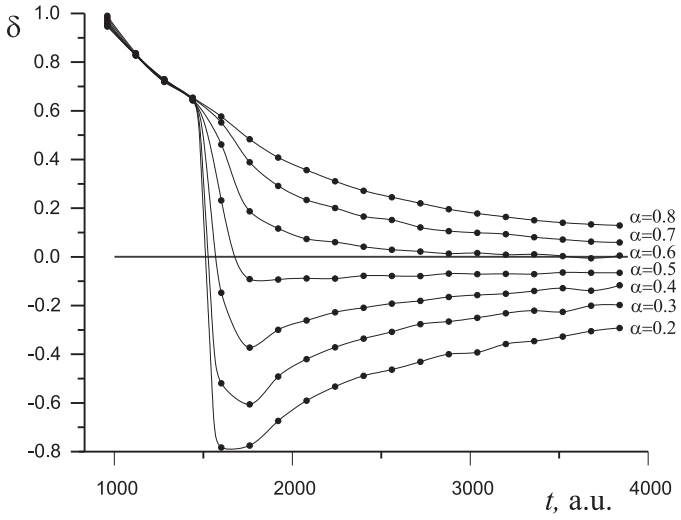


Fig. 5. The family of anisotropy time-dependence for different α values and short-lived source. The observation point is to the left of the boundary (that is, in α-medium). Distances: source-observer $d = 300$ and source-boundary $b = 400$, $v = 0.3$.

respectively.

The U- and W-shaped distributions for $\alpha < 1$ are convenient propagators to analyze this phenomenon, because they are universal and do not depend on scale parameters of α-medium if asymptotical regime is realized.

V. Conclusion

Thus, our calculations led to the following conclusion. The normal diffusion theory always shows prevalence of the direct current component (from a localized source) over the inverse one (to the source). The more adequate mesodiffusion theory confirms this conclusion only when the medium is homogeneous and/or the source doesn't change in the course of time. However, this improved diffusion theory has demonstrated the *real existing inversion current effect near the $\alpha - e$ boundary in case of a pulsed source*. This fact interpreted as a result of stochastic reflection from a more dense medium is discovered for the first time and possibly can shed light on some unsolved astrophysical [12], [13] and other problems.

Acknowledgment

We are much obliged to Prof. A. Wolfendale and Prof. A. Erlykin inspired us to do this work. We thank as well the Russian Foundation for Basic Research (project no. 13-01-00585) and the Ministry of Education and Science of the Russian Federation for financial support.

References

[1] Godoy S., Garcia-Colin L. S. (1998). Mesoscopic diffusion as a non-Markov process. *Physica A: Statistical Mechanics and its Applications*, 258(3), 414-428.

[2] Lebon G., Jou D. and Casas-Vázquez J. *Understanding Non-Equilibrium Thermodynamics*. Springer-Verlag Berlin, 2008.

[3] Uchaikin V. V., Saenko V. V. (2000). Telegraph equation in random walk problem. *Journal of Physical Studies*, 4(4).

[4] Uchaikin V. V., Saenko V. V. (2001). On the theory of classic mesodiffusion. *Technical Physics*. 46, 139146.

[5] Uchaikin V. V., Sibatov R. T. (2004). One-dimensional fractal walk at a finite free motion velocity. *Technical Physics Letters*, 30(4), 316-318.

[6] V.V.Uchaikin. (2013). *Fractional Derivatives for Physicists and Engineers*, Vol.I-II, Springer (Berlin) – High Education Press (Beijing).

[7] Shlesinger M. F., Klafter J. (1986). Lévy walks versus Lévy flights. In: *On Growth and Form* (pp. 279-283). Springer Netherlands.

[8] Sokolov I. M., Metzler R. (2003). Towards deterministic equations for Lvy walks: The fractional material derivative. *Physical Review E*, 67(1), 010101.

[9] Lamperti, J. (1958). An occupation time theorem for a class of stochastic processes. *Transactions of the American Mathematical Society*, 380-387.

[10] Rebenshtok A., Barkai E. (2008). Weakly non-ergodic statistical physics. *Journal of Statistical Physics*, 133(3), 565-586.

[11] Uchaikin V. V., Sibatov, R. T. (2009). Statistical model of fluorescence blinking. *Journal of Experimental and Theoretical Physics*, 109(4), 537-546.

[12] Erlykin A. D., Wolfendale A. W. (2006). The anisotropy of galactic cosmic rays as a product of stochastic supernova explosions. *Astroparticle Physics*, 25(3), 183-194.

[13] Erlykin A. D., Wolfendale, A. W. (2013). Cosmic rays in the inner galaxy and the diffusion properties of the interstellar medium. *Astroparticle Physics*, 42, 70-75.

Analysis of processes in DC arc plasma torches for spraying that use air as plasma forming gas

Vladimir Ya. Frolov, Dmitry V. Ivanov

Saint Petersburg State Polytechnical University

***Abstract*—**Developed in Saint Petersburg State Polytechnical University technological processes of air-plasma spraying of wear-resistant, regenerating, hardening and decorative coatings used in number of industrial areas are described. The article contains examples of applications of air plasma spraying of coatings as well as results of mathematical modeling of processes in air plasma torches for spraying.

***Keywords*—**air plasma, coatings, modeling of plasma processes, plasma spraying

I. Introduction

Application of DC arc plasma torches for spraying of coating is well known [1–3]. For that purpose they usually use a DC arc plasma torch with a variable arc length in which they use nitrogen, inert gases (argon, helium) and their mixtures with hydrogen as the plasma forming gas [2].

Studies being performed for many years in the Department of Electrical Power Engineering and Equipment of St. Petersburg State Polytechnical University were a basis for a development of DC arc plasma torches for spraying that have the following main features: a design is based on an interelectrode insert and an application of air as plasma forming gas. There were developed a large number of plasma spraying technologies for various purposes (heat-resistant coating, wear-resistant one, corrosion-resistant one, protective one, decorative one, etc.) using such plasma torches [3, 4].

The article presents examples of application of used plasma equipment as well as an analysis of processes in the DC arc plasma torch for spraying using a mathematical modeling.

II. Examples of Applications

Department of Electrical Power Engineering and Equipment of Saint Petersburg State Polytechnical University and Science and Educational Technological Centre "Electrotechnology" perform theoretical and experimental researches in the area of plasma technology over 50 years [3].

As a result of performed researches there were developed plasma torches and realized technologies of spraying of different protective coatings on new manufactured parts and technologies of recovery of out-of-repair items [4]. There were realized technologies of recovery of out-of-repair parts of industrial machines (spindles, seats for bearings), plain bearings of turbocharging compressor shafts for internal-combustion engines as well as technologies of air-plasma spraying of wear-resistant coatings on automotive crankshaft journal (freezer compressor) [4].

Fig. 1 shows the process of plasma spraying of aluminium coatings onto carbon fiber to improve the properties of supercapacitors [5].

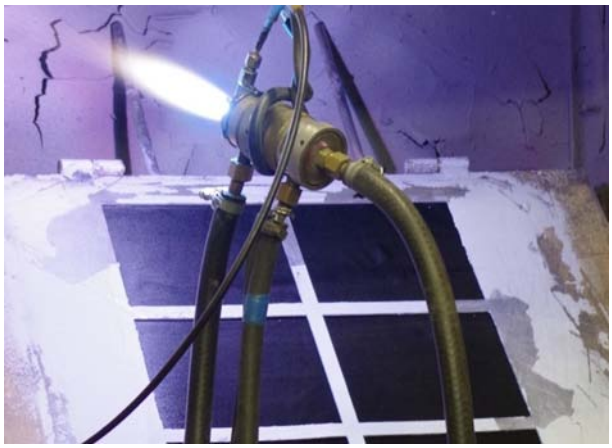


Fig.1. Air plasma spraying of aluminum coatings onto carbon fiber [5]

There were developed technologies of surface hardening of stop equipment of gas main lines. There were worked through conditions of spraying of wear-resistant and corrosion-resistant coatings on surfaces of plungers of autocrane elevators, face seals and rods of water pumps.

Air-plasma spraying provides to create coatings from metals that are usually exposed to considerable oxidation, for example, copper coating [4].

Fig. 2 presents parts of gas turbine GTN-25 with a heat-shielding zirconium dioxide coating created by air-plasma spraying.

V. Ya. Frolov is with Department of Electrical Power Engineering and Equipment of Saint Petersburg State Polytechnical University, 195251, St. Petersburg, Polytechnicheskaya, 29, Russia (corresponding author, e-mail: frolov.eed@gmail.com).

D. V. Ivanov is with Department of Electrical Power Engineering and Equipment of Saint Petersburg State Polytechnical University, 195251, St. Petersburg, Polytechnicheskaya, 29, Russia (e-mail: d.ivanov@list.ru).

ISBN: 978-1-61804-251-4

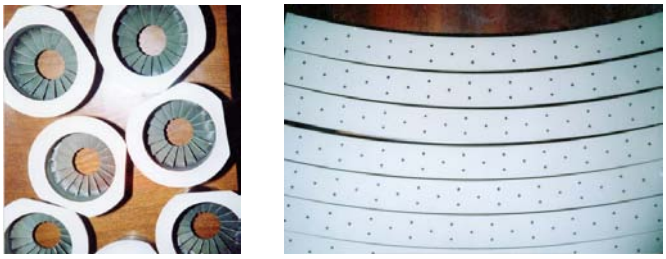


Fig. 2. Parts of gas turbine GTN-25 with a heat-shielding zirconium dioxide coating

Modern industry requires increasing of energetic branches, particularly in oil industry and power production. One of the problems, which were solved with air-plasma spraying, was a problem of wearing of threaded connection of pump-compressor pipes that used in oil industry. That problem was very important because during operations the threaded connection of pipe breaks down in the first place.

Coatings were created from powder materials of different composition with size of particles less than 50 μm. Optimal conditions of spraying were determined versus variation of following operational conditions: arc current, arc voltage, plasma forming air flow rate, spraying distance, torch velocity relative to pipe, angle between particles flow and pipe axis.

Example of pipes with threaded connection that was hardening by coatings were exposed to wear-resistant tests using a stand imitating pipe screwing. Coating wear-resistance as well as pipe breaking strength in an area of the threaded connection was estimated. Results of tests have shown that a wear resistance of thread with coating is increased in several times in comparison with an initial version without coating.

A new important trend is a coating for restoring of metal sculptures and monuments. In St. Petersburg, as in other historical centres of the world, there are a lot of monuments which have different damages because of aggressive influence of city's atmosphere.

There was elaborated an air-plasma technology of corrosion-resistant and decorative coatings for copper alloys. A preoxidised copper powder was chosen as a coating material. Previously, it was investigated a level of powder's oxidation and regimes of the spraying. One can choose the colour of the coating by varying the level of powder's preliminary oxidation (see. Fig. 3) [6].

Fig. 3. The colour palette of protective and decorative coatings based on copper powder sprayed by the air-plasma arc plasma torch [6]

Total technology of coatings includes several stages [7]:

- stream-vortex cleaning of surface;
- air-plasma spraying;
- impregnation an inhibitor of corrosion into the coating's porosity;
- surface treatment with natural wax.

Preliminary test was realized on bronze, brass and copper plates with size 150×150 mm. Thickness of the formed coatings was about 100 μm. The covering samples were processed by solution of inhibitor and fitted for test in climatic camera.

As a result based on accelerated test on corrosion stability it was determined an optimum regime of the air-plasma spraying. Test in climatic camera have shown that such covering can stand more than 70 years in our climatic condition without any trace of the corrosion.

The elaborated technology was used on practice in Saint Petersburg during last restorations of sculptures group "Tamer of horses" by P.Klodt (the Anichkov bridge), metal parts of "Aleksander's column" (the Palace square) and of sculptures group on the Senate and Synod Building (see Fig. 4).

Fig. 4. Air plasma spraying of protective and decorative coatings onto sculpture of the Senate and Synod Building (Saint Petersburg)

There is a possibility of applying of such covering not only for restoring purposes but as anti-corrosion covering in different branches of industry.

III. Mathematical Modeling

In conditions of widespread applications of air-plasma spraying it is very important to know the qualitative and quantitative relationship between technological efficiency of the spraying process, on the one hand, and the geometry of the plasma torch and its mode of operation (arc current, gas flow rate etc.) on the other hand. An effective way of obtaining this information is the mathematical modelling of plasma processes in the arc plasma torch for spraying.

Used mathematical model of plasma processes in the arc plasma torch for spraying is based on the following assumptions: plasma is in the state of local thermodynamic

equilibrium; plasma is laminar and optically thin. Those assumptions let us to consider plasma as a continuous medium.

Equations included in the model express fundamental conservation laws (energy, momentum, mass) and are given in [8]. The channel region of the plasma torch and a region of plasma jet were taken as the computational domain.

Plasma properties are also included in the mathematical model. Their dependencies on temperature (typically at atmospheric pressure) are given in the literature for main gases used in arc plasma torches for spraying, for example in [9].

A series of calculations of plasma processes in the air-plasma DC arc plasma torch PN-V1 was carried out with the following operational parameters: the arc current was varied between 150 and 200 A, the gas flow rate was varied between 0.6 and 1.2 g/s. Distributions of plasma temperature, velocity, pressure and electromagnetic functions were obtained as results of calculations. Examples of obtained distributions of plasma temperature and axial velocity are presented in Fig. 5 and 6.

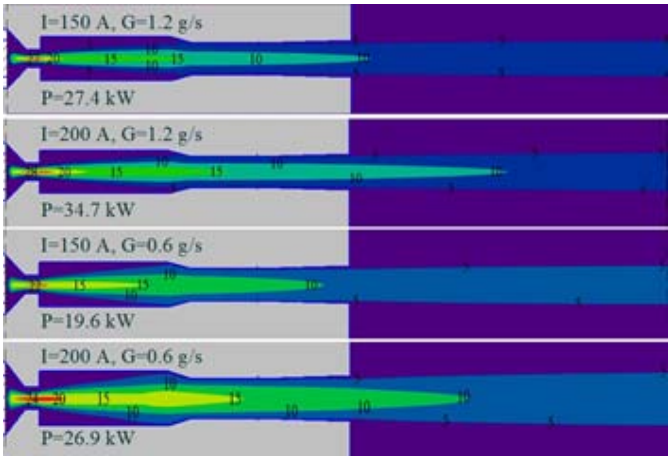


Fig. 5. Distributions of plasma temperature (in 103 K) at different modes of operations (I =150; 200 A, G = 0.6; 1.2 g/s)

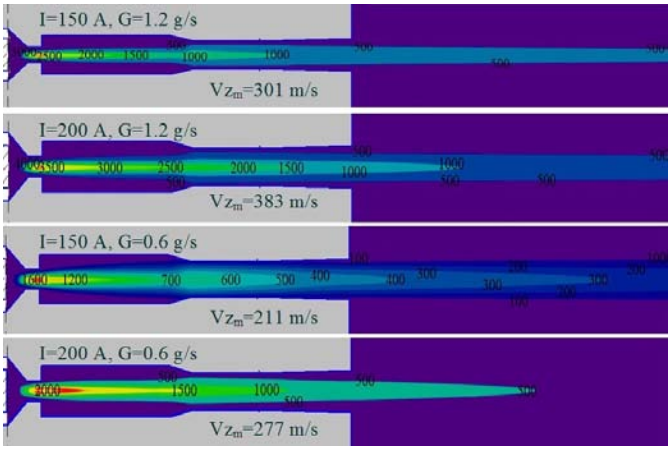


Fig. 6. Distributions of axial plasma velocity (in m/s) at different modes of operations (I =150; 200 A, G = 0.6; 1.2 g/s)

Comparing the results of calculations one can draw the following conclusions:

- temperature at the outlet of the plasma torch increases when the arc current increases (at a constant flow of gas) and when the gas flow rate increases (at a constant arc current);

- velocity at the outlet of the plasma torch (1256 m / sec) at the arc current of 200 A and the gas flow rate of 1.2 g/s is higher than for other operational conditions. Comparing the results at the same arc current one can see that the plasma velocity depends on the gas flow rate namely the plasma velocity is higher when the gas flow rate is higher.

IV. Conclusion

Developed equipment, techniques of experimental investigations, mathematical models of plasma processes and results of investigations are the basis for the implementation of the technology of air-plasma spraying of wear-resistant, corrosion-resistant, recovering and thermal barrier coatings as well as of protective and decorative coatings on the monuments of different materials.

References

[1] Pfender E., "Thermal plasma technology: where do we stand and where are we going?", *Plasma Chem. Plasma Process*, Vol. 19, 1999, p.1-31.

[2] Fauchais P. "Understanding plasma spraying", *J. Phys. D: Appl. Phys*., Vol.37, No.9, 2004, pp. R86-R108.

[3] Frolov V. Ya., Klubnikin V. S., Petrov G. K., Ushin B. A. *Technique and technology of coatings*, St. Petersburg Polytechnical Univ. Publ. House, 2008 (in Russian).

[4] Frolov V., Petrov G., Yushin B., Dubov M., Churkin I., Ivanov D. "Research and development of plasma technologies of spraying of coatings", *Proc. 18th Symposium on Physics of Switching Arc* (Nové Město na Moravě, Czech Republic, Sept. 7.-11., 2009) eds V.Aubrecht, M.Bartlova, 2009, pp 162–165.

[5] Isaeva E.M. "Investigation of the process of plasma spraying of aluminium coating onto the carbon fiber", Master Thesis, SPbSPU, St. Petersburg, 2014 (in Russian).

[6] Ushin B.A. "Development of air-plasma technology of spraying of protective and decorative coatings", PhD Thesis, SPbSPU, St. Petersburg, 2010.

[7] Petrov G.K., Frolov V.Ya., Ushin B.A. "Method of forming of protective and decorative coating on a metal surface", RF Patent 2486276, filed February 29, 2012, and issued June 27, 2013.

[8] Dresvin S.V., Ivanov D.V., Frolov V.Ya. "Method of calculation of thermal plasma processes", *Induction heating*, Vol. 4(22), 2012, pp. 22-25 (in Russian).

[9] Boulos M. I., Fauchais P., Pfender E. *Thermal Plasmas: Fundamentals and Applications*, Vol. 1 ,New York: Plenum Press, 1994.

Quantification of Selected Factors of Longevity

V. Pacáková, P. Jindrová

***Abstract*—**This article is devoted to the analysis of selected factors that have influence on life expectancy at birth and life expectancy at age 65, which represent longevity in the countries of European Union (EU). By applying selected multivariate statistical methods on 13 indicators from the Eurostat database and from the World Health Organization (WHO), we quantify the basic factors of life expectancy and of the changes in the period from 2000 to 2010, and further look for the causes of their different levels in various EU countries.

***Keywords*—**Common factors, clusters, life expectancy, longevity.

I. Introduction

Life expectancy has been increasing in almost all the countries of the world. Demographic trends in many countries show signs of ageing of population. This is due to declining birth rates and to longevity. This fact has strong social, cultural and economic consequences. The ageing process directly affects health care, long-term care and pension systems.

Although the longevity is often understood as economy and social problem, it is without a doubt the result and proof of a higher quality of life.

The indicators such as the Life expectancy at birth (*LE*) and Life expectancy at age 65 (*LE*65) by the inhabitants of the European Union (EU) in all 28 members' countries have a growing trend mainly as a result of improving living conditions and health care.

The columns in Table 1 for each country of the EU comprise successively life expectancy at birth in years 2000 and 2010, absolute (Δ) and relative growth rate (k) of life expectancy in 2010 compared to 2000, the average growth rate in time period 2000-2010 and rank of each country by ascending order of the growth rate k.

Spearman Rank Correlation between variables *LE* and *LE*65 is equal to 0,9647 and changes between 2010/2000 of both indicators are very similar.

In further analysis, by using selected multidimensional statistical methods, we will try to determine the main factors of differences in life expectancy in year 2010 and to compare the impact of life conditions on longevity in the EU countries.

Table 1 The changes of life expectancy at birth in the period 2000-2010 in EU countries

Country	LE-2000	LE-2010	Δ	k	$\bar{k}$	rank
AT-Austria	78,47	80,88	2,41	1,0307	1,0030	21
BE-Belgium	77,6	80,31	2,71	1,0349	1,0034	14
BG- Bulgaria	71,71	73,82	2,11	1,0294	1,0029	24
HR-Croatia	73	76,86	3,86	1,0529	1,0052	4
CY-Cyprus	77,89	82,19	4,3	1,0552	1,0054	2
CZ-Czech	75,21	77,81	2,6	1,0346	1,0034	16
DK-Denmark	77,22	79,3	2,08	1,0269	1,0027	26
EE-Estonia	70,95	76,03	5,08	1,0716	1,0069	1
FI-Finland	77,88	80,34	2,46	1,0316	1,0031	19
FR-France	79,35	81,98	2,63	1,0331	1,0033	18
DE-Germany	78,42	80,64	2,22	1,0283	1,0028	25
EL-Greece	78,23	80,69	2,46	1,0314	1,0031	20
HU-Hungary	71,93	74,78	2,85	1,0396	1,0039	9
IE-Ireland	76,61	80,8	4,19	1,0547	1,0053	3
IT-Italy	79,75	82,5	2,75	1,0345	1,0034	17
LV-Latvia	70,58	73,7	3,12	1,0442	1,0043	6
LT-Lithuania	72,21	73,57	1,36	1,0188	1,0019	28
LU-Luxembourg	79,07	81,49	2,42	1,0306	1,0030	22
MT-Malta	78,24	81,51	3,27	1,0418	1,0041	8
NL-Netherlands	78,29	81,15	2,86	1,0365	1,0036	11
PL-Poland	73,86	76,58	2,72	1,0368	1,0036	10
PT-Portugal	76,85	80,13	3,28	1,0427	1,0042	7
RO-Romania	71,25	73,83	2,58	1,0362	1,0036	12
SK-Slovakia	73,45	75,66	2,21	1,0301	1,0030	23
SI-Slovenia	76,27	79,96	3,69	1,0484	1,0047	5
ES-Spain	79,49	82,32	2,83	1,0356	1,0035	13
SE-Sweden	79,92	81,77	1,85	1,0231	1,0023	27
UK-United King.	78,06	80,78	2,72	1,0348	1,0034	15

II. Data and Methods

For statistical analysis and quantification of the main factors of longevity in EU countries, we have chosen these variables (Source: WHO, Eurostat):

LE Life expectancy at birth
*LE*65 Life expectancy at age 65
*X*1 (Gross domestic product (GDP), US$ per capita)
*X*2 (Unemployment rate (%))
*D*1 (Diseases of circulatory system, all ages, per 100 000)
*D*2 (Ischaemic heart disease, 0–64, per 100 000)
*D*3 (Malignant neoplasms, all ages, per 100 000)
*D*4 (Motor vehicle traffic accidents, all ages, per 100 000)

V. Pacáková is with Institute of Mathematics and Quantitative Methods, Faculty of Economics and Administration, University of Pardubice, Pardubice, Studentská 84, 532 10 Pardubice, Czech Republic (e-mail: Viera.Pacakova@upce.cz).

P. Jindrová is with Institute of Mathematics and Quantitative Methods, Faculty of Economics and Administration, University of Pardubice, Pardubice, Studentská 84, 532 10 Pardubice, Czech Republic (e-mail: Pavla.Jindrova@upce.cz).

ISBN: 978-1-61804-251-4

*D*5 (Suicide and self-inflicted injury, all ages, per 100 000)
*D*6 (Selected alcohol-related causes, per 100 000)
*D*7 (Selected smoking-related causes, per 100 000)
*H*1 (% population self-assessing health as good)
*H*2 (Total health expenditure as % of GDP, WHO estimates)
*S*1 (Pure alcohol consumption, litres per capita, age 15+)
HDI (Human Development Index)

Variables *X*1, *X*2 are the essential economic variables, variables *D*1 to *D*7 inform about the occurrence of serious diseases, *H*1 and *H*2 indicate the level of health care, *S*1 is an important social indicator and *HDI* cumulates information about the quality of life in each country.

According to the above mentioned aim for analysis of these variables we use the multiple regression analysis, factor analysis and cluster analysis. For calculation we use statistical package Statgrapics Centurion XVI.

The *Multiple Regression* procedure is designed to construct a statistical model describing the impact of selected thirteen variables on the dependent variables *LE* and *LE*65.

The goal of *Factor analysis* is to characterize the *p* variables in terms of a small number of common factors.

An important result of the above model is the relationship between the variances of the original variables and the variances of the derived factors. This variance is expressed as the sum of two quantities: the *communality* and the specific variance. The communality is the variance attributable to factors that all the origin variables have in common, while the specific variance is specific to a single factor.

An important concept in factor analysis is the rotation of factors. In practice, the objective of all methods of rotation is to simplify the rows and columns of the factor matrix to facilitate interpretation. The *Varimax criterion* centres on simplifying the columns of the factor matrix. With the Varimax rotation approach, the maximum possible simplification is reached if there are only 1's and 0's in a single column.

The correlation between the original variables and the factors show the factor loadings. They are the key to understanding the nature of a particular factor. Squared factor loadings indicate what percentage of the variance in an original variable is explained by a factor.

The *Factor Scores* in output of Factor analyse procedure display the values of the rotated factor scores for each of *n* cases, in our analysis in each of 28 countries of EU. Factor score show where each country falls with respect to the extracted factors.

The Cluster Analysis procedure is designed to group observations (countries) into clusters based upon similarities between them. A number of different algorithms is provided for generating clusters. We use the agglomerative algorithm, beginning with separate clusters for each observation or variable and then joining clusters together based upon their similarity. The results of the analysis are displayed in a *dendogram*.

The distance between two observations we calculate by *City Block distance*, defined as

$$d(x,y)=\sum_{i=1}^{n}|x_i-y_i|$$

and distance between two clusters by Ward's method. This method defines the distance between two clusters in terms of the increase in the sum of squared deviations around the cluster means that would occur if the two clusters were joint.

III. Results and Discussion

A. Multiple regression results

The output of this procedure shows the results of fitting a multiple linear regression model to describe the relationship between LE and 13 independent variables of life conditions in 28 countries of EU. The equation of the fitted model is

$$LE = 77{,}8 + 2{,}773\cdot E\,\hat{}\,(-7)\cdot X1 - 0{,}023\cdot X2 - 0{,}015\cdot D1 - 0{,}032\cdot D2$$
$$-0{,}022\cdot D3 + 0{,}032\cdot D4 - 0{,}017\cdot D5 - 0{,}018\cdot D6 + 0{,}006\cdot D7$$
$$-0{,}02\cdot H1 - 0{,}165\cdot H2 - 0{,}058\cdot S1 + 15{,}471\cdot HDI$$

The R-Squared statistic R = 98,74 % indicates that the model as fitted explains 98,74% of the variability in *LE*. Model confirms the positive impact of variables X1, and *HDI* on life expectancy at birth *LE*, positive impact of variables *D*4 and *D*7 and negative effect of all other variables. Regression coefficient of each independent variable indicates increase (decrease) in the value of *LE* if corresponding independent variable has increased (decreased) by one unit of measure.

The next linear regression model describes the relationship between *LE*65 and 13 selected independent variables. The value of R-squared R = 97,07 % indicate that also this model fit the data adequately.

$$LE65 = 23{,}45 - 0{,}00000255\cdot X1 + 0{,}013\cdot X2 - 0{,}009\cdot D1 + 0{,}020\cdot D2$$
$$-0{,}011\cdot D3 - 0{,}041\cdot D4 - 0{,}003\cdot D5 + 0{,}009\cdot D6 - 0{,}006\cdot D7$$
$$+0{,}006\cdot H1 + 0{,}046\cdot H2 - 0{,}045\cdot S1 - 0{,}093\cdot HDI$$

Interpretation of this model is analogous to the interpretation of the linear regression model of the variable *LE*.

B. Factor analysis results

The purpose of this analysis is to obtain a small number of factors, which account for most of the variability in the 13 variables of life conditions in EU countries.

Table 2 Factor Loading Matrix After Varimax Rotation

	Factor 1	*Factor* 2
*X*1	0,788345	-0,0889639
*X*2	-0,497086	0,244695
*D*1	-0,847056	0,375612
*D*2	-0,786431	0,481749
*D*3	-0,340812	0,712337
*D*4	-0,684894	0,0250224
*D*5	-0,187428	0,87721
*D*6	-0,515881	0,7778
*D*7	-0,775435	0,546034
*H*1	0,541965	-0,497831
*H*2	0,758679	-0,0991457
*S*1	0,126858	0,799122
HDI	0,913282	-0,0795356

In this case, two factors have been extracted, since two factors have eigenvalues greater than or equal to 1,0. Together

they account for 71, 787% of the variability in the original data. Since we have selected the principal components method, the initial communality estimates have been set to assume that all of the variability in the data is due to common factors.

Table Table of Factor Scores

Country	*Code*	*Factor 1*	*Factor 2*
Austria	AT	6,21032	-1,75297
Belgium	BE	4,90106	-1,50831
Bulgaria	BG	-8,71049	2,55031
Croatia	HR	-8,52015	5,42823
Cyprus	CY	3,97565	-6,90987
Czech Republic	CZ	-2,60874	3,2857
Denmark	DK	3,57181	-0,72485
Estonia	EE	-7,60153	5,68359
Finland	FI	3,65816	-1,322
France	FR	7,57466	-2,46918
Germany	DE	6,52678	-2,41902
Greece	EL	1,4221	-5,3375
Hungary	HU	-10,1226	8,91103
Ireland	IE	5,74519	-1,83929
Italy	IT	5,59178	-6,90799
Latvia	LV	-14,0058	8,38655
Lithuania	LT	-14,533	12,496
Luxembourg	LU	7,72952	-3,40073
Malta	MT	3,67656	-5,23497
Netherlands	NL	9,08067	-4,57608
Poland	PL	-5,83168	3,33399
Portugal	PT	1,40506	-1,88902
Romania	RO	-11,8394	4,59184
Slovakia	SK	-7,41602	4,75944
Slovenia	SI	1,22055	1,20971
Spain	ES	5,57447	-4,8472
Sweden	SE	7,76893	-5,87064
United Kingdom	UK	5,55623	-3,6268

Substantive interpretation of these factors is based on the significant higher loadings. Factor 1 has 4 significant loadings with positive signs with variables $X1$, $H1$, $H2$ and HDI and significant loadings with negative signs with 4 variables $D1$, $D2$, $D4$ and $D7$ of incidence of serious illnesses. Therefore, this factor can be interpreted as a factor of favourable conditions for life. Significant positive correlation with variables $D3$, $D5$, $D6$ and $S1$ with clearly negative impact on quality of life is the reason that we interpret Factor 2 as a factor of adverse living conditions.

This table shows the factor scores for each country of EU. A country with a high value of Factor 1, and also with a low value of Factor 2 has by interpretation of these factors favourable conditions for life; contrary, the conditions for life in a country with a low value of Factor 1 and a high value of Factor 2 are unfavourable.

Graphical display of EU countries in a two-dimensional coordinate system with axes $F1$ and $F2$ allows us to quickly assess the quality of life in each EU country and allows also compare living conditions in all EU countries. Such visual display we can see in Fig. 1.

It may be seen that quality of life is vastly different in the old and new Member States of the EU.

Life expectancy is undoubtedly the most reliable indicator of the quality of life. As can be seen in Fig. 2 and Fig. 3, LE increases with increasing values of Factor 1 and decreases with the decrease in values of the Factor 2.

These figures confirmed significantly worse quality of life and lower life expectancy in the former socialist countries in comparison with the old EU member states.

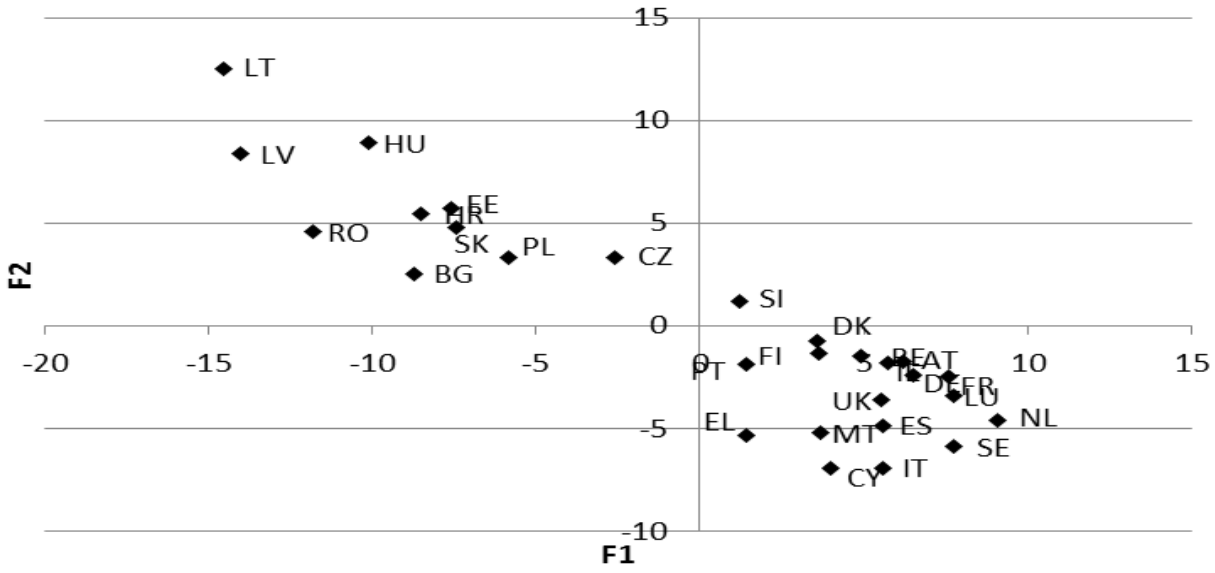


Fig. 1 Location EU countries in the coordinate system of the F1 and F2

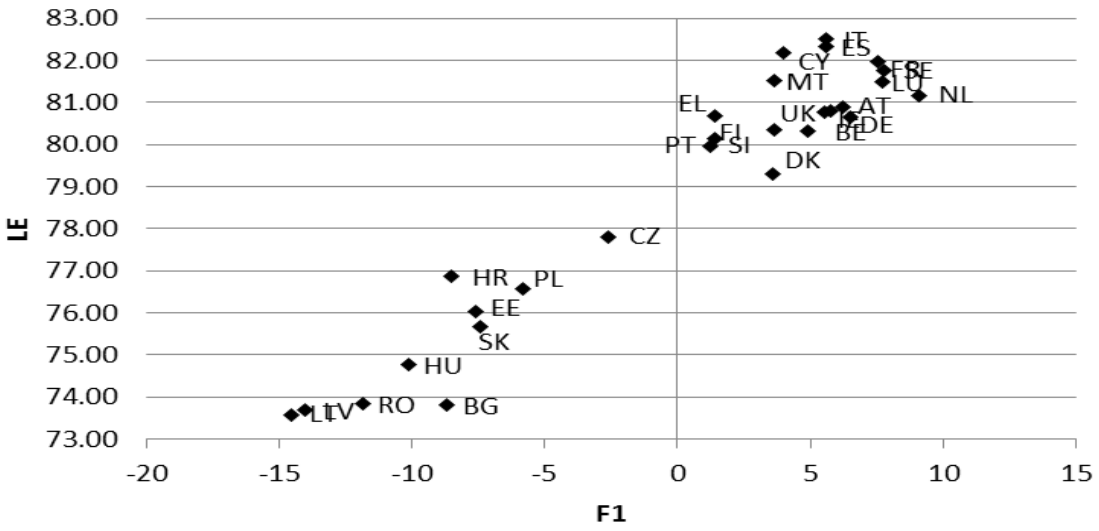


Fig. 2 Dependency of life expectancy at birth by a factor *F*1

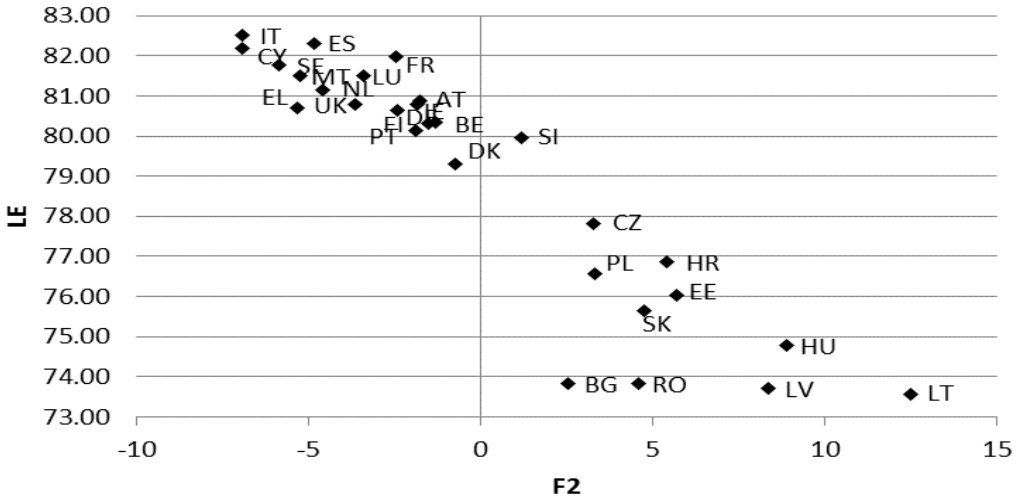


Fig. 3 Dependency of life expectancy at birth by a factor *F*2

C. Claster Analysis results

This procedure has created 2 clusters from the 28 observations (countries) supplied. The clusters are groups of observations with similar characteristics. To form the clusters, the procedure began with each observation in a separate group. It then combined the two observations which were closest together to form a new group. After re-computing the distance between the groups, the two groups then closest together were combined. This process was repeated until only 2 groups remained.

The results of cluster analysis are consistent with the results of factor analysis.

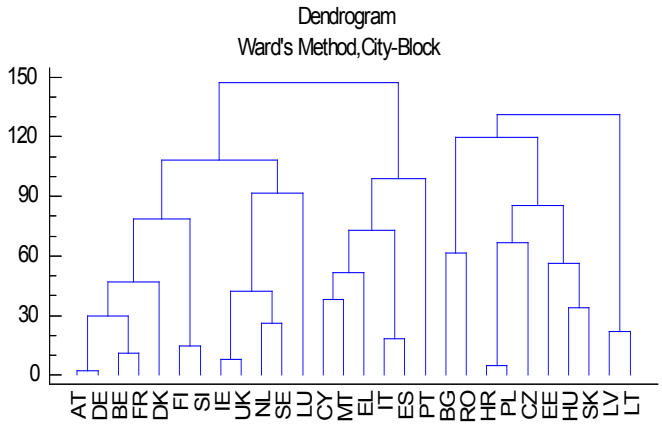


Fig.4 The result of Cluster Analysis

ISBN: 978-1-61804-251-4

IV. Conclusion

The results of the statistical analysis confirm the appropriateness of the methods used and the suitability of the chosen variables of living conditions in EU countries. The methods chosen enable to extract two common factors of quality of life instead of the original 13 variables. This allowed obtaining transparent and visual information about the impact of living conditions on the life expectancy in the EU countries and the possibility of graphical presentation of analysis results.

References

[1] B. Benjamin, J. H. Polard, *The Analysis of Mortality and Other Actuarial Statistics*. Reprint. Oxford: Butterworth-Heinemann, 1992.

[2] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, R. L. Tatham, *Multivariate Data Analysis*. Sixth Edition. New Jersey: Pearson Education, Inc., 2007.

[3] P. Hebák, J. Hustopecký, E. Jarošová, I. Pecáková, *Vícerozměrné statistické metody* (*Multivariate statistical methods*) (1). Praha: Informatorium, 2004.

[4] R. A. Johnson, D. W. Wichern, *Applied Multivariate Statistical Analysis*. Sixth Edition. New Jersey: Pearson Prentice Hall, 2007.

[5] J. Kubanová, *Statistické metody pro ekonomickou a technickou praxi* (*Statistical Methods for economic and technical practice*). Bratislava: Statis, 2008.

[6] V. Pacáková. et al., *Štatistiké metódy pre ekonómov* (*Statistical Methods for Economists*). Bratislava: IURA Edition, 2009.

[7] V. Pacáková, V., D. Sivašová, Multivariate Statistical Comparisons of the Economic and Social Situation of Chosen European Countries, In: Proceeding *Analysis an International Comparisons of Social Consequences of Transformation Processes in Post-Communist Countries*, Bratislava: STATIS, 2001, p. p. 141-151.

[8] I. Stankovičová, M. Vojtková, *Viacrozmerné štatistické metódy s aplikáciami* (*Multivariate statistical methods with applications*).. Bratislava: IURA Edition, 2007.

[9] Šoltés, E.: *Regresná a korelačná analýza s aplikáciami* (*Regression and correlation analysis with applications*). Bratislava: Iura Edition, 2008

[10] P. Jindrová, Quantification. of Risk in Critical Illness Insurance. In: Conference proceedings from 9th international scientific conference *Financial Management of Firms and Financial Institutions*, VŠB Ostrava, 2013. pp. 298-306.

[11] Eurostat [online]. http://epp.eurostat.ec.europa.eu/portal/page/portal/health/introduction http://epp.eurostat.ec.europa.eu/portal/page/portal/gdp_and_beyond/quality_of_life/context

[12] World Health Organization [online]. http://www.who.int/research/en/

Prof. RNDr. Viera Pacáková, Ph.D. graduated in Econometrics (1970) at Comenius University in Bratislava, 1978 - RNDr. in Probability and Mathematical Statistics at the same university, degree Ph.D at University of Economics in Bratislava in 1986, associate prof. in Quantitative Methods in Economics in 1998 and professor in Econometrics and Operation Research at University of Economics in Bratislava in 2006.

She was working at Department of Statistics Faculty of Economic Informatics, University of Economics in Bratislava since 1970 to January 2011. At the present she has been working at Faculty of Economics and Administration in Pardubice since 2005.
She has been concentrated on actuarial science and management of financial risks since 1994 in connection with the actuarial education in the Slovak and Czech Republic and she has been achieved considerable results in this area.

Mgr. P. Jindrová, **Ph.D.** graduated from Mathematical analysis at the Faculty of Science of Palacky University in Olomouc in 1993. She lectures mathematics, statistics and insurance and financial mathematics in branch of study of Insurance engineering. She finished her dissertation thesis which deals with problems of aggregation and disaggregation in economics and mathematics.

Specification and Analysis of Hybrid Systems with PDE in ISMA Simulation Environment

Yu.V. Shornikov, A.V. Bessonov, M.S. Myssak, D.N. Dostovalov

***Abstract*—** A class of hybrid systems (HS) with system of partial differential equations is considered. Architecture of instrumental environment for simulation of complex dynamic and hybrid systems is given. Algorithms of finite difference method for the transition from PDE to ODE system are given. Universal data structure for storing HS models has been designed and proved. Algorithms of variable step with accuracy and stability control of numerical scheme for solving high-dimensional Cauchy problems are proposed. The algorithms are based on explicit methods of Runge-Kutta type. Sequential and parallel implementation of numerical methods is presented. The example of specification and analysis of reaction-diffusion problem associated with Lotka-Volterra model is given.

Keywords— Autogenerated parsers, hybrid system, Runge-Kutta methods, sequential and parallel implementation, software architecture, system of PDE.

I. Introduction

MANY engineering problems are characterized by points of discontinuity in the first derivative of the phase variables. Such combined discrete-continuous problems called hybrid or switched systems [1], [2]. In this paper we propose a new system architecture of ISMA (that in Russian means Instrumental Facilities of Machine Analysis) based on the methodology of hybrid systems (HS). This article examines a new application of ISMA – systems of partial differential equations (PDE) with constraints. PDEs are used to describe processes in the chemical-technological systems, elasticity problems, etc. To achieve this goal a series of consecutive problems is solved: the textual specification of ISMA environment models expanded by constructions describing PDEs; a special data structure for storing the model in memory is developed; approximation algorithms are implemented for transition from a system of PDE to ODE system. In many cases, the problem is compounded by high dimensionality and stiffness of the considered system. To solve moderately stiff problems integration algorithms based on the explicit methods to control accuracy and stability of the numerical scheme can be applied [3], [4]. Furthermore when problem dimension reaches several thousands of equations and more its calculation by sequential methods becomes ineffective and requires the use of multiprocessor computer systems. In this situation parallel computation of local behaviors using cluster technologies can significantly improve the quality and efficiency of calculations. This paper discusses sequential and parallel implementation of algorithms of variable step based on two-stage and three-stage schemes of Runge-Kutta type of respectively second and third order of accuracy. These integration algorithms are well suited for solving hybrid problems including moderately stiff problems. As an example for specification and comparative analysis of the considered algorithms the reaction-diffusion problem associated with Lotka-Volterra model [5] is examined.

This work was supported by grant 14-01-00047-a from the Russian Foundation for Basic Research, RAS Presidium project № 15.4 "Mathematical modeling, analysis and optimization of hybrid systems".

Yu.V. Shornikov is with the Design Technological Institute of Digital Techniques Siberian Branch of Russian Academy of Science, Novosibirsk, Russia (e-mail: shornikov@inbox.ru).

A.V. Bessonov, M.S. Myssak, D.N. Dostovalov is with the Department of Automated Control Systems, Novosibirsk State Technical University, Novosibirsk, Russia (e-mails: abv.poste@gmail.com, maria_myssak@mail.ru, dostovalov.dmitr@mail.ru).

II. Class of Systems

There are many systems (mechanical, electrical, chemical, biological, etc.), the behavior of which can be conveniently described as a sequential change of continuous modes [1]. Each mode is given by a set of differential-algebraic equations with the following constraints:

$$\begin{aligned} &y' = f(x, y, t), x = \varphi(x, y, t), \\ &pr : g(x, y, t) < 0, \\ &t \in [t_0, t_k], x(t_0) = x_0, y(t_0) = y_0, \\ &x \in R^{N_x}, y \in R^{N_y}, t \in R, \\ &f : R^{N_x} \times R^{N_y} \times R \to R^{N_y}, \\ &\varphi : R^{N_x} \times R^{N_y} \times R \to R^{Nx}, \\ &g : R^{N_x} \times R^{N_y} \times R \to R^{S}. \end{aligned} \quad (1)$$

The vector-function $g(x,y,t)$ is referred to as event function or guard [2]. A predicate pr determines the conditions of existence in the corresponding mode or state. Inequality $g(x,y,t) < 0$ means that the phase trajectory in the current mode should not cross the border $g(x,y,t) = 0$. Events occurring in violation of this condition and leading to transition into another mode without crossing the border are referred to as one-sided.

This class of systems is expanding by the addition of

ISBN: 978-1-61804-251-4

boundary conditions for PDEs. Continuous behavior of HS is determined by the systems differential-algebraic equations and PDEs. In the proposed implementation the equations with the maximal order not higher than second are considered. Applied algorithms do not impose a restriction on number of variables – i.e. their number is theoretically unlimited. Nevertheless should take into account that the introduction of each new variable leads to a tremendous increase in the number of equations generated as a result of the finite differences method. Therefore the real limit on the number of variables is imposed by computing resources: computer software as well as computer itself. The considered equations are nonlinear type. The linear equation are also supported and regarded as a narrow equations type.

In this paper we consider the type of nonhomogeneous PDE. The coefficients used in partial differential equations considered in this paper can be either constant or variable. This paper discusses the parabolic type equations. Boundary conditions of considered problems must be rectangular area Ω – rectangle, parallelepiped, etc.

Thus, the proposed expansion of the instrumental environment designed for the analysis of PDE type equation (2): heterogeneous, non-linear, second order, with constant and variable coefficients, with an unlimited number of variables N and limited by N-dimensional rectangular grid.

$$
\begin{aligned}
&\frac{\partial z}{\partial t} = \psi\left(x, z, t, p, \frac{\partial z}{\partial p}, \frac{\partial^2 z}{\partial p^2}\right),\\
&x = \varphi(x, t),\\
&pr : g(x, t) < 0,\\
&x(t_0, p) = x_0, z(t_0, p) = z_0,\\
&t \in [t_0, t_k], p \in [p_0, p_m],\\
&\left.\frac{\partial z}{\partial n}\right|_p = \tilde{z}(p),\\
&x \in R^{N_x}, z \in R^{N_z}, t \in R, p \in R^{N_p},\\
&\varphi : R^{N_x} \times R \to R^{N_x},\\
&\psi : R^{N_x} \times R^{N_z} \times R \times R^{N_p} \times R^{2N_p} \to R^{N_z},\\
&g : R^{N_x} \times R \to R^{S},
\end{aligned}
\qquad (2)
$$

where n denotes the normal to the boundary and $\tilde{z}(p)$ is a given function to the boundary.

III. Architecture of ISMA

Simulation environment of complex dynamical and hybrid systems called ISMA is developed at the department of Automated control systems of Novosibirsk state technical university (Russia) [6].

Specification of hybrid systems is carried out using graphical and symbolic languages that are the system content of instrumental environment. Analytical content is provided by numerical methods and algorithms for computer analysis corresponding to the chosen class of systems and methods for solving these models. ISMA environment is developed subject to simplicity of description of dynamical and hybrid models in the language that is maximally close to the object language. Main features of ISMA are the following:

1) Composition of hybrid models is carried out in visual structural-textual form;
2) Structural form of model description corresponds to the classical description of systems by block diagrams and includes all necessary components such as integrators, accumulators, amplifiers, signal sources, nonlinear elements and others;
3) Language of symbolic specification is approached maximal to the language of mathematical formulas;
4) Special module for specification of problems of chemical kinetics in the language of chemical reactions which automatically translates them into a system of differential equations;
5) A variety of traditional and original numerical methods included methods that are intended for the analysis of ODE systems of medium and high stiffness;
6) Computer simulation in real time;
7) Graphic interpreter called GRIN provides a wide range of tools for analysis and visualization of simulation results such as scaling, tracing, optimization, displaying in the logarithmic scale and phase plane;
8) Extension of system functionality by adding new typical components and numerical methods.

Architecture of ISMA software package (Fig. 1) is designed to unify existing mathematical program software for analysis of problems in various object domains: chemical kinetics, automation, electricity, etc. Blocks that belong to this paper marked with a dark color.

IV. A New Description Language of HS Models with PDEs

One of the many approaches used is ISMA to describe HS models is textual representation. For this purpose a special language LISMA (Language of ISMA) is developed [7]. And context-free grammar of LL(2) class is designed for LISMA. However existing description tools are not suitable enough to model boundary value problems of PDE systems. Therefore new language structures are introduced to describe specific elements.

Before the development of grammar of language elements a comparative analysis of multiple peers was made. In these grammars the following characteristics were emphasized: flexibility and extensibility, usability, ease of perception, the corresponding mathematical description. Using these criteria multiple languages were evaluated with the ability to describe models with PDE equations. The main ones are the following: FlexPDE, Wolfram Mathematica, gPROMS, EMSO, ASCEND. In many languages to describe the partial derivative functional style is used, in which the derivative is a function of several arguments. Based on review and analysis of mentioned

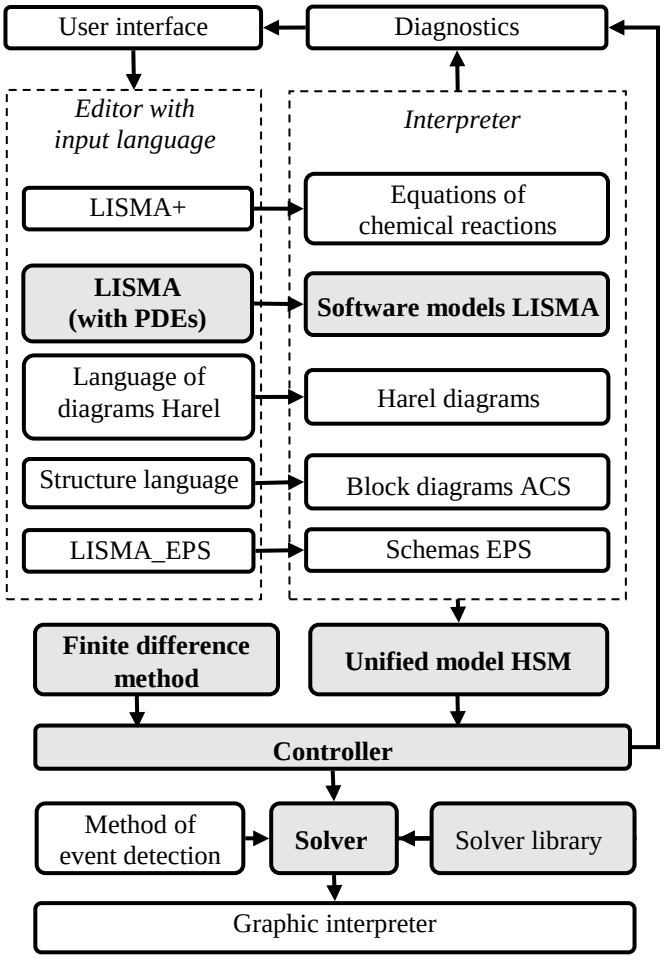


Fig. 1. Architecture of ISMA

simulation environments LISMA has been extended by description of boundary value problems with PDEs. This extended grammar was developed in the ideology of inheritance.

To describe a system of differential equations, boundary conditions and initial values a new LISMA language features are introduced. Explicit declaration of variables that should be subjected to discretization is introduced. For this purpose a special structure is used with grammar written in the Backus-Naur form as follows:

apx _var → **'var'** *var_ident* **' ('**
DecimalLiteral **',' ***DecimalLiteral* **')'**
apx_var_tail **';'**

apx_var_tail → **'apx'** *DecimalLiteral*
| **'step'** (*FloatingPointLiteral* | *DecimalLiteral*)

For example, the following expression corresponds to the given grammar:

var x[0, 20] **apx** 30;

var y[0, 30] **step** 0,5;

In this structure boundaries of the variable are defined in square brackets. The following constructions are the keyword **apx** (short for approximation) or the keyword **step**. Keyword **apx** used if we want to break the considered domain of definition for a fixed number of segments, thus realizing the priority execution speed. We use the keyword **step** if step size is important – an accuracy priority. Number of segments and step size are written following the keyword.

Several elements are introduced in the description of the equations. First, it is an explicit indication of variables that affect the equation on the left side of the equation. This record type is optional and can be omitted as before on the left side to specify variables in parentheses. Partial derivatives are described in a functional style. Letter **D** is used as function name – the most concise version, which is not lost in code, mostly lowercase. The arguments used name of a differentiable function, the variable on which the derivative is taken and the order of the derivative. If the derivative is first-order, the latter argument is omitted suggesting that it is taken equals to 1 by default. This approach to the description of the derivative satisfies all the previously entered criteria: it allows you to use all sorts of variables which should be differentiated; it does not contain descriptive information duplication and laconic, and thus it is practical and easy to perception; and finally it is easy to relate to mathematical description. As a result, the description of the PDE as follows:

partial_operand → **'D' '('** *Identifier* **','** *Identifier*
(**','** *DecimalLiteral*)? **')'**

Below is an example of this language construct:

c1' = Kh***D(c1,x,2)** + **D(Kv*D(c1,z), z)**;

c2' = **D(c1,x,2)***pow(x,2);

For numerical solution of PDEs by FDM is necessary to determine the boundary conditions – values of the derivative at the edges of the grid under consideration. It looks as follows:

edge → **'edge'** *edge_eq* **'on'** *Identifier edge_side* **';'**
edge_eq → *Identifier* **'='**
(*FloatingPointLiteral* | *DecimalLiteral*)
edge_side → **'left'** | **'right'** | **'both'**

Construction contains a partial derivative equation with a certain value on the right side. This allows you to specify a description for a particular variable value and the type of border: left, right or both. Below is an example of all three types of boundary conditions:

edge D(c1, x) = 0 **on left**;
edge D(c1, x) = 20 **on right**;
edge D(c1, y) =0 **on both**;

Lexical and syntactic analyzers for modified language are developed using the library Antlr4. This library is ideal for problems of fast and efficient automatic construction of parsers. Under the ISMA project it is decided to use antlr4,

ISBN: 978-1-61804-251-4

because it is quite applicable to the existing grammar and own work to create parsers require significant investment in the development and documentation. In addition, experience in using Antlr4 can be successfully applied in other projects.

V. Unified Description Section in ISMA Environment

After the language grammar is defined, a data structure for storing models in memory after parsing should be developed. This data structure must be universal for all types of model specification. With several ways to describe HS – graphics, text or block diagrams, we should be able to lead each of them to a single universal form, which subsequently may be transmitted to solver input. This approach simplifies the task of unification of ISMA software. For a more specific application of the simulation environment ISMA it is necessary to develop a graphical part responsible only for model specification. In this description module the modules of calculation and graphical interpretation remain unchanged.

When designing such a data structure many factors should be taken into account. This is necessary to ensure that upon presentation of new conditions the whole system of modeling was not inapplicable. Thus, the developed data structure should be easily expandable. Adding new elements to the model should not be a need to rewrite large parts of the system. Furthermore, the model should excludes redundancy. The appearance of redundancy in the description of such complex structures as the model of a hybrid system with a differential-algebraic equations and PDEs is a potential source of errors. In addition, the model must be easily divided into blocks and must be easy to use. In order to accommodate these properties subject-oriented approach is chosen in the design of the data structure for storing HS model. The implementation language is Java.

The data structure represents a set of closely related classes which can be divided into four types (Fig. 2):

1) equation description classes;
2) expression description classes;
3) discrete behavior (states and transition conditions) description classes;
4) entire model description classes.

Expressions are a sequence of elements called tokens. Tokens are both operands and operators. Two types of sequences to write expressions are supported: infix and postfix (inverse polish notation). Postfix notation is useful for computing values on the stack and is used for calculations in step semantic analysis to calculate values of the constants and initial conditions. If the expression contains only operators of algebraic is considered algebraic. If the expression contains logical operators it refers to a type of conditional expressions used in the description of the conditions of transitions between states. Algebraic expressions used in the description of the right sides of all considered types of equations: algebraic, differential, PDEs, and constants. All of them serve to describe continuous behavior of certain state of the hybrid system. Within each state there exists the so-called variable table that

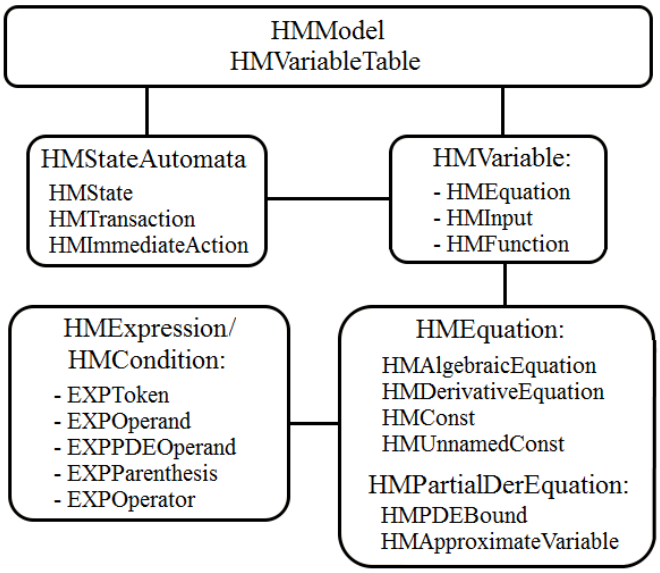


Fig. 2. Common HSM structure

stores equations and variables. For description of the states and conditions of the transitions between them responsible the appropriate type of classes. Collection of states, conditions of transitions between them, and a set of instantaneous action at the entrance to a condition is called hybrid automaton. Hybrid automaton has an initial state, which corresponds to the time t = 0. Complete set of all of the above classes called the HS model.

For solving PDE in the ISMA environment approximation of equations for difference grid by finite differences method is used. This method is used to solve systems with rectangular boundaries (ex. hydrodynamics and electrodynamics). In such problems FDM has a sufficiently high accuracy. At the same time it wins in speed other methods. Consider the algorithm transition from of PDE to a system of ordinary differential equations.

Step 1. Construct a list of all variables for discretization. At this stage special structure for describing variables of storing type of discretization and accuracy (number of elements of grid) is analyzed.

Step 2. Divide equations into two groups: permanent and approximated. Here you need to select the equation that must be converted to the difference analogue – is analyzed right-hand sides and identifies those who are in the right part of the required variables.

Step 3. Construct N-dimensional grid. For all variables for sampling the number of elements for which they are divided is defined. The product of these values is the dimension of the grid.

The number of equations to be obtained by the algorithm corresponds to the equation:

$$N = n_p \prod_i n_i, \tag{3}$$

where N is the number of approximating equations, n_p is the

number of PDEs, n_i, $i = 1,..,m$ is the number of grid points for each variable approximated.

Numerical approximation of derivatives is performed by the formulas:

$$\frac{\partial u_j}{\partial \zeta} = \frac{u_{j+1} - u_{j-1}}{2\Delta\zeta}, \frac{\partial^2 u_j}{\partial \zeta^2} = \frac{u_{j-1} - 2u_j + u_{j+1}}{(\Delta\zeta)^2}, 1 \le j \le N. \quad (4)$$

Step 4. Apply boundary conditions. For each variable boundary conditions under which the simulation and approximation takes place must be specified. At this stage the variables in the initial and final nodes of the grid (relative to the current variable) are replaced by the indicated boundary values. If not then the default is zero.

Step 5. Set initial conditions for all equations.

Step 6. Transition from the grid to the ODE system. Here you need to go over all grid points. Thus it is necessary to put a unique identifier for each new equation while maintaining a connection that has been set by difference equations. As a result for each equation in each grid a copy of it in the system of ODE will be created and the initial condition (t = 0) will be specified.

VI. Scheme of Model Interpretation and Solving

Four basic levels of working with model can be destinguished: the interpretation of the input specification of the model, controller, solver, graphic interpreter. The fig. 3 presents the first two levels in more detail.

Level of interpretation is responsible for converting the model described by the input specification of the universal representation HSM. After model is input to simulation environment in text specification before the numerical solution the model passes through a series of stages of analysis. The first two stages – the lexical and syntactic analysis. They are conducted by facilities of the parser generated library antlr4. If the model is correct, we have an abstract syntax tree (AST). Bypass AST and further retrieval of information allows you to fill a unified data structure a model description HS. This is done using a variety of services:

1) service of sequential approach syntax tree (design pattern "visitor");
2) conversion service from infix into postfix notation and vice versa;
3) value calculator of constants and initial conditions – a stack machine for the Polish-inverted notation;
4) service of model validation – semantic analysis.

In case the received data structure is correct in terms of semantics the interpretation is considered complete and can go to the level of controller.

Level of the controller is responsible for preparing the model HSM to the view which satisfies the conditions imposed by a solver. Also, the controller is responsible for loading of numerical methods, loading of library functions and other simulation settings. After the HSM model was obtained from the interpreter the finite difference method is performed according to the algorithm described earlier. As a result, the description of the PDE in the data structure is replaced by a system of ODE.

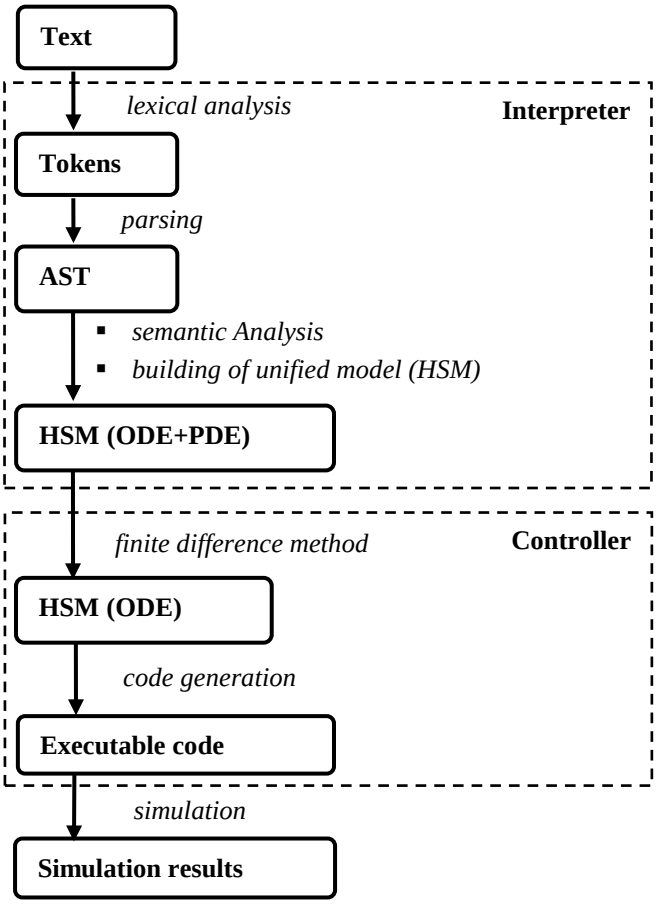


Fig. 3. Stages of interpreter and controller

Generated data structure describing the internal representation of the hybrid system model based on the input specification of the model (text or graphics) should be counted and get simulation results. For this data structure is go to the input of the controller. The controller is a bonding layer, which produces all the necessary transformation and sets the simulation parameters. In addition, the controller is responsible for connection and control of library functions.

At this stage, the model is finally formed and fed to the input of the controller solver. Depending on the selected model parameters numerical method and the method of event detection HS is a calculation model. It is worth mentioning state changes in the calculation. ISMA environment has a unique method of detecting events [1] that increases the accuracy of solutions to rigid problems. Subsequent step in the simulation is the calculation. It is produced using a unique library of numerical methods ISMA. The simulation shows the resulting set of points on a plane integrated graphical visualization tools.

VII. Solvers

This section is devoted to the integration algorithms of

variable step based on two-stage and three-stage explicit methods of Runge-Kutta type that implements schemes of second and third accuracy order respectively.

The algorithms are applied to numerical solving of Cauchy problem for ODE systems of the following form:

$$y' = f(y),\ y(t_0) = y_0,\ t_0 \le t \le t_k. \tag{5}$$

Consideration of autonomous problem does not reduce the generality because non-autonomous problem always can be cast to autonomous by introducing an additional variable.

Particular attention should be paid to the choice of the integration method. Fully implicit methods cannot be used because they require the calculation of $f(y)$ at a potentially dangerous area, where the model is not defined. Therefore here we will use explicit methods with solution: $y_{n+1} = y_n + h_{n+1}\varphi_n$, $n = 0, 1, 2, \ldots$ As a result we obtain the dependence of the predicted integration step h_{n+1}.

Considering that explicit methods are known by poor stability this paper examines integration methods with accuracy and stability control. Generally accuracy and stability control are used to limit the size of the integration step. As a result projected step h_{n+1} is calculated as follows.

The choice of the next integration step size is based on the proved theorem [4] and can be written as follows:

$$h_{n+1} = \max\left[h_n, \min\left(h^{ac}, h^{st}\right)\right] \tag{6}$$

where h^{ac} and h^{st} are step sizes obtained as a result of accuracy control and stability control respectively. This formula allows to stabilize the step behavior in the area of solution establishing where stability plays a decisive role. Because the presence of this area severely limits the use of explicit methods for solving stiff problems.

A. Two-stage Runge-Kutta method

Suppose that for numerical solving of problem (5) the following implicit two-stage method of Runge-Kutta type is used:

$$\begin{aligned} &y_{n+1} = y_n + 0.5(k_1 + k_2),\\ &k_1 = h_n f(y_n),\\ &k_2 = h_n f(y_n + k_1). \end{aligned} \tag{7}$$

where y and f are real N-dimensional vector-functions, t is an argument, h is an integration step, k_1 and k_2 are method stages and 0.5 is a numerical coefficient defining accuracy and stability properties of (7).

Inequality for accuracy control has the following form:

$$0.5\left\|k_2 - k_1\right\| \le \varepsilon. \tag{8}$$

And inequality for stability control looks as follows:

$$v_n = 2\max_{1\le i\le N}\left(\left|k_3^i - k_2^i\right| / \left|k_2^i - k_1^i\right|\right) \le 2, \tag{9}$$

where length of stability interval of the scheme is approximately equals to 2; k_1^i, k_2^i and k_3^i are the components of vectors k_1, k_2 and auxiliary vector k_3. Stage k_3 coincides with vector k_1 for next step and therefore does not lead to computational costs increasing.

B. Three-stage Runge-Kutta method

Consider implicit three-stage method of Runge-Kutta type for solving problem (5), which has the following form:

$$\begin{aligned} &y_{n+1} = y_n + \frac{1}{6}k_1 + \frac{2}{3}k_2 + \frac{1}{6}k_3,\\ &k_1 = hf(y_n),\ k_2 = hf(y_n + 0.5k_1),\\ &k_3 = hf(y_n - k_1 + 2k_2). \end{aligned} \tag{10}$$

Inequality for accuracy control has the following form:

$$\left\|k_1 - 2k_2 + k_3\right\| \le 6\varepsilon. \tag{11}$$

And inequality for stability control looks as follows:

$$v_{n,3} = 0.5\max_{1\le i\le N}\left(\left|k_1^i - 2k_2^i + k_3^i\right| \Big/ \left|k_2^i - k_1^i\right|\right) \le 2.5. \tag{12}$$

More detailed description of the designated methods can be found at [4].

VIII. Organization of Sequential Computations

Let the method (7) is used for numerical solving of problem (5) and let the approximate solution y_n is known at moment t_n with step h_n. Then to obtain the approximate solution y_{n+1} at moment t_{n+1} we have the following common algorithm:

Step 1. Calculate approximate solution y_{n+1} at moment t_n with step h_n according to performing method.

Step 2. Calculate approximate function value $f(y_{n+1})$.

Step 3. Obtain integration step accuracy characteristics.

Step 4. If solution is accurate then go to **Step 5**, else set integration step h_n equals to step h^{ac} corrected by accuracy according to performing method and go to **Step 1**.

Step 5. Obtain integration step stability characteristics.

Step 6. If solution is accurate then go to **Step 7**, else set integration step h_n equals to step h^{st} corrected by stability according to performing method and go to **Step 1**.

Step 7. Get size of next integration step using (6).

Step 8. Perform next integration step.

ISBN: 978-1-61804-251-4

IX. Organization of Parallel Computations

Developed parallel algorithms are based on the presented above sequential algorithms with the following differences.

For definiteness we assume that computer system consists of p processors, $N > p$ and let k is a number of equations per rank. Then all of N equations should be evenly allocated between computing nodes. To achieve this goal classical Round-Robin algorithm is chosen. Also parallel algorithm was designed to reuse sequential algorithm.

Taking into consideration assumptions about beginning of sequential method base parallel algorithms can be defined in the following way:

Step 1. Allocate equations evenly between ranks using Round-Robin algorithm.

Step 2. Calculate in each rank approximate solution y_{n+1}^j, $0 \le j \le k$ at moment t_n with step h_n according to performing method.

Step 3. Send obtained y_{n+1}^j from each rank to others.

Step 4. Calculate in each rank approximate function value $f^j\left(y_{n+1}\right)$, $0 \le j \le k$.

Step 5. Execute for each rank sequential algorithm from **Step 3** corresponding to accuracy control.

X. Reaction-Diffusion Problem Simulation

Let us consider specification features and solving of problem of the designated class on the example of reaction-diffusion problem in two-dimensional space, which is associated with competition model of Lotka-Voletrra [5].

Two kinds of variables $c^1(x,z,t)$ and $c^2(x,z,t)$ represent density of competing species in the habitat area $\Omega = \{(x,z): 0 \le x \le 1, 0 \le z \le 1\}$ and in time $0 \le t \le 3$:

$$\frac{\partial c^i}{\partial t} = d_i\left(\frac{\partial^2 c^i}{\partial x^2} + \frac{\partial^2 c^i}{\partial z^2}\right) + f^i\left(c^1, c^2\right), i = 1, 2 \qquad (13)$$

where $d_1 = 0.05$, $d_2 = 1.0$, $f^1\left(c^1, c^2\right) = c^1\left(b_1 - a_{12}c^2\right)$, $b_1 = 1$, $a_{12} = 0.1$, $f^2\left(c^1, c^2\right) = c^2\left(-b_2 + a_{21}c^1\right)$, $a_{21} = 100$, $b_2 = 1000$.

Boundary conditions are $\partial c^i / \partial x = 0$ at $x = 0$, $x = 1$ and $\partial c^i / \partial z = 0$ at $z = 0$, $z = 1$ Initial conditions are $c^1(x,z,0) = 10 - 5\cos(\pi x)\cos(10\pi z)$ and $c^2(x,z,0) = 17 + 5\cos(10\pi x)\cos(\pi z)$.

At $t \to \infty$ solution becomes spatially homogeneous and tend to periodically solve ODE system of Lotka-Volterra. This ODE system is alternately stiff and non-stiff depending on the solution position in the phase space.

Computer model in the ISMA is:

```
// Constants
const pi = 3.1415926;
const d1 = 0.05;
const d2 = b1 = 1;
const a12 = 0.1;
const a21 = 100;
const b2 = 1000;

// Variables to be sampling
var x[0, 1] apx 20;
var z[0, 1] apx 20;

// Equations
c1 '= d1 * (D(c1, x, 2) + D(c1, z, 2)) + f1;
c2 '= d2 * (D(c2, x, 2) + D(c2, z, 2)) + f2;

f1 = c1 * (b1 – a12*c2);
f2 = c2 * (-b2 + a21*c1);

// Boundary conditions
edge c1 = 0 on x both;
edge c2 = 0 on x both;
edge c1 = 0 on z both;
edge c2 = 0 on z both;

// Initial conditions
c1(0) = 10 – 5 * cos(pi * x) * cos (10 * pi * z);
c2(0) = 17 + 5 * cos (10 * pi * x)*cos(pi * z);
```

In this example blocks of model description are marked by comments: constants, variables to be sampling, algebraic equations, the system of PDE, boundary conditions, the initial value.

Turning to the grid of size $J \times K$ by x and z respectively we obtain $\Delta x = 1/(J-1)$ and $\Delta z = 1/(K-1)$ are grid steps by x and z coordinates, c_{jk}^i is approximation of $c^i\left(x_j, z_k, t\right)$, where $x_j = (j-1)\Delta x$, $z_k = (k-1)\Delta z$, $1 \le j \le J$, $1 \le k \le K$.

Thus we obtain differential equations system of $N = 2JK$ dimension:

$$\begin{aligned}\dot{c}_{jk}^i &= \frac{d_i}{\Delta x^2}\left(c_{j+1,k}^i - 2c_{jk}^i + c_{j-1,k}^i\right) + \\ &+ \frac{d_i}{\Delta z^2}\left(c_{j,k+1}^i - 2c_{jk}^i + c_{j,k-1}^i\right) + f_{jk}^i,\end{aligned} \qquad (14)$$

where $1 \le i \le 2$, $1 \le j \le J$, $1 \le k \le K$, $f_{jk}^i = f^i\left(c_{jk}^1, c_{jk}^2\right)$.

Boundary conditions on grid are the following: $c_{0,k}^i = c_{2,k}^i$, $c_{J+1,k}^i = c_{J-1,k}^i$ for $1 \le k \le K$ and $c_{j,0}^i = c_{j,2}^i$, $c_{j,K+1}^i = c_{j,K-1}^i$ for $1 \le j \le J$.

Fig. 4 show that solution diagrams are qualitatively similar.

ISBN: 978-1-61804-251-4

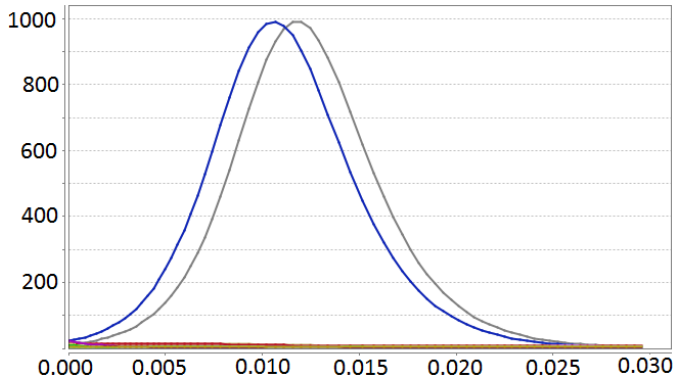


Fig. 4. Simulation results obtained by new version of ISMA solver

The comparative analysis results of implemented algorithms in sequential and parallel version are given on the Table 1.

Table 1. Computational Costs of the Algorithms

Dimension	Algorithm			
	RK2ST		RK3ST	
	Sequential time (ms)	Parallel time (ms)	Sequential time (ms)	Parallel time (ms)
20x20 (N = 800)	3645	750	6710	1092
40x40 (N=3200)	106545	7710	198540	10731
60x60 (N=7200)	1203345	91325	2400734	120289
80x80 (N=12800)	4709395	307575	8259426	4773248
100x100 (N=20000)	12109500	787905	24508700	107756

Problem (14) was calculated for equations from 800 to 20000. MPI is chosen as paralleling technology because this approach is focused on cluster system and in future will allow calculating system of much more higher dimension if needed.

Dependence of computational time from system dimension is shown in Fig 5. Such a significant increase of computational costs (especially by sequential algorithms) is related to costs of construction and inversion of Jacoby matrix of increasing dimension. Also the higher system dimension the advantage of parallel algorithm becomes even more clearly.

XI. Conclusion

In this paper a new class of hybrid systems, continuous dynamics of which is defined by a system of DAE and PDE with boundary conditions, within ISMA instrumental environment is introduced. The architecture of ISMA is considered. New elements of LISMA language for description of PDE and boundary conditions language are presented.

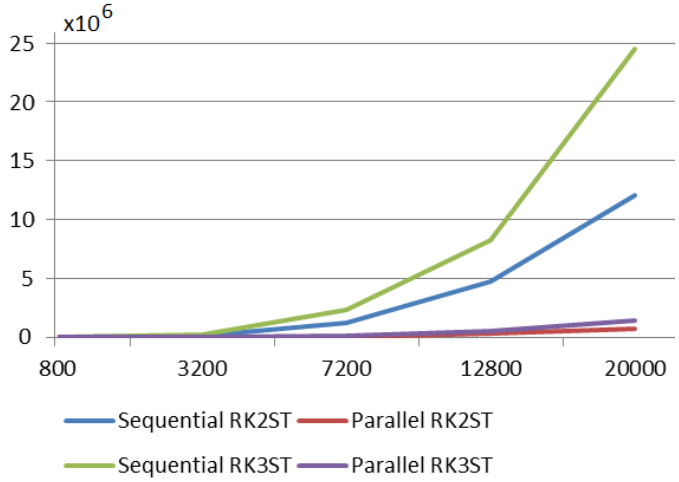


Fig. 5. Perfomances of implemented algortithms

Grammar of new language inherited from the old language and is also context-free. Data structure for storing unified representation model is designed. Features of the considered hybrid systems such as the increased stiffness and high dimension make difficult to apply the commonly used sequential methods of numerical analysis. Steps of parsing and transition to the solver are discussed in detail. An example of reaction-diffusion is given, text model is obtained and numerical calculations are carried out. The experiments show that parallel algorithms based on explicit schemes with accuracy and stability control are more suitable for analysis of system of the designated class.

References

[1] E.A. Novikov, Yu.V. Shornikov, “Computer simulation of stiff hybrid systems: monograph”, Novosibirsk, Russia: Publishing house of NSTU, 2012.

[2] J. Esposito, V. Kumar, G.J. Pappas “Accurate event detection for simulating hybrid systems”, Hybrid Systems: Computation and Control (HSCC), Springer-Verlag, vol. LNCS 2034, 1998.

[3] E.A. Novikov, Yu.V. Shornikov “Numerical simulation of hybrid systems by Runge-Kutta method of second accuracy order”, Computing technologies, vol. 13 #2, pp. 98-104, 2008.

[4] E.A. Novikov “Explicit methods for stiff systems”, Novosibirsk: Nauka, 1997.

[5] Peter N. Brown, Alan C. Hindmarsh, Matrix Free Methods in the Solution of Stiff systems of ODEs, Lawrence Livermore National Laboratory, 1983. – 38p.

[6] Yu.V. Shornikov “Instrumental facilities of machine analysis” Yu.V. Shornikov, V.S. Druzhinin, N.A. Makarov, K.V. Omelchenko, I.N.Tomilov. Certificate of official program registration # 2005610126, Moscow: Rospatent, 2005.

[7] Yu.V. Shornikov, I.N. Tomilov, “The Program of Language Processor from Language LISMA”. Certificate of official program registration # 2007611024, Moscow: Rospatent, 2007.

Piecewise-regular object recognition in real-time applications

Andrey V. Savchenko, and Vladimir R. Milov

***Abstract*—**Mathematical models and methods of complex (piecewise-regular) object recognition are reviewed. Examples in speech and image processing tasks are provided. The recognition methods are classified in dependence on the number of available models and the count of classes. We emphasize the issue of insufficient accuracy and computing efficiency of popular recognition methods (both statistical and machine learning) if the number of models per each class is not enough and the number of classes is large (thousands of alternatives). It is experimentally shown that the probabilistic neural network with homogeneity testing and the directed enumeration method allow to decrease the recognition speed in comparison with contemporary approximate nearest neighbor methods.

***Keywords*—**Image recognition, mathematical models and methods of statistical pattern recognition, piecewise-regular objects, speech recognition.

I. Introduction

Classification task is one of the most significant in the field of pattern recognition. In this task it is required to assign the query object *X* (facial photo, speech signal, image of natural scenes, text) to one of *C>1* classes [1]. Most part of contemporary research in this area is concentrated on the development of recognition algorithms by assuming that each class is specified by the given database (DB) $\{X_r\}$, $r=\overline{1,R}$ of $R \geq C$ cases. For each model object X_r class label $c(r) \in \{1,...,C\}$ is known.

It is possible to extract three possible directions of research in pattern recognition (Fig. 1). If analyzed objects are represented as feature vectors with the fixed dimension *M=const* then traditional (pointwise [2]) classification methods are applied. There are either simple methods, e.g., linear/quadratic discriminant analysis (LDA/QDA) or complex machine learning techniques, e.g., feed-forward multi-layer perceptron (MLP) [3] and support vector machine (SVM) [4], which approximate the given data by estimating parameters and, in general case, the net structure [1], [5].

Most complex is the task of group choice classification [2] in which the object X is represented by the group (collection) of independent identically distributed (i.i.d.) feature vectors. The decision of this task is usually achieved by aggregating the decisions of traditional classifiers mentioned above applied to each member of the group.

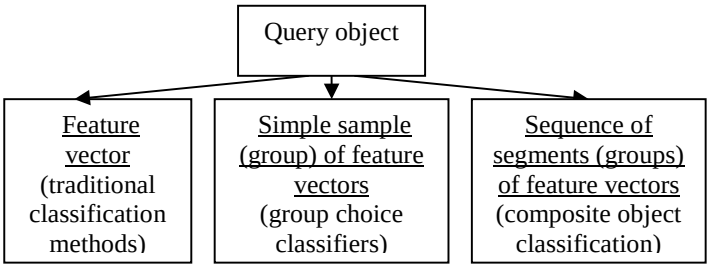


Fig. 1 classification of recognition systems in dependence on the object type

Recently the research have moved focus to the objects containing several independent homogeneous (regular, "stationary") parts. Each part (or segment) may be represented as a group of i.i.d. feature vectors [6]. Let's call such objects composite or piecewise-regular.

The recognition of piecewise-regular objects include image and speech processing tasks. For instance, the images of the whole object (in HOG (histogram of oriented gradients)-like methods [7], [8]) or the keypoint nethorhood (in SIFT (Scale-Invariant Feature Transform)-like methods [9]) are divided into a grid of blocks; each block is processed independently (compare with the known JPEG/MPEG compression algorithms).

Speech can be considered as a sequence of minimal units (phones, phonemes, triphones) which are independent of each other. The signal corresponded to particular word or phrase is assumed to be the realization of non-stationary random process. Various experimental studies show [10] that the application of traditional classification methods are characterized with low accuracy if feature vector is estimated for the *whole* analyzed signal. Thus, modern techniques include other approach which determines the classifier structure. Query object *X* and all models X_r are considered as sequences of, respectively, *K* and K_r homogeneous (regular) segments (phones). These segments are relatively independent (features of different segments inside one word may have nothing in common as they correspond to distinct phones).

The segmentation algorithms of practically important objects (images, speech) are well studied [1], [11] and are not

A. V. Savchenko is with the National Research University Higher School of Economics, Nizhny Novgorod, Russian Federation (phone: +79030434003; e-mail: avsavchenko@ hse.ru).

V. R. Milov is with N.Novgorod State Technical University, Nizhny Novgorod, Russian Federation (e-mail: vladimir.milov@gmail.com).

ISBN: 978-1-61804-251-4

discussed further in this article.

The rest of the paper is organized as follows: Section 2 presents the recognition procedure of composite objects more thoroughly. In Section 3, our experimental results in recognition of faces and Russian speech are presented. Finally, concluding comments are given in Section 4.

II. Composite Objects Recognition Methods

Methods of piecewise-regular object recognition are primarily determined by the characteristics of the available DB of models (Fig. 2). The decision is made in favor of the closest model in terms of the summary distance for all K segments. Each segment is recognized with pointwise or group choice classifiers. As the number of extracted segments is usually high, computing efficiency of algorithms of composite object recognition is rather low in comparison with other types of objects (Fig. 1).

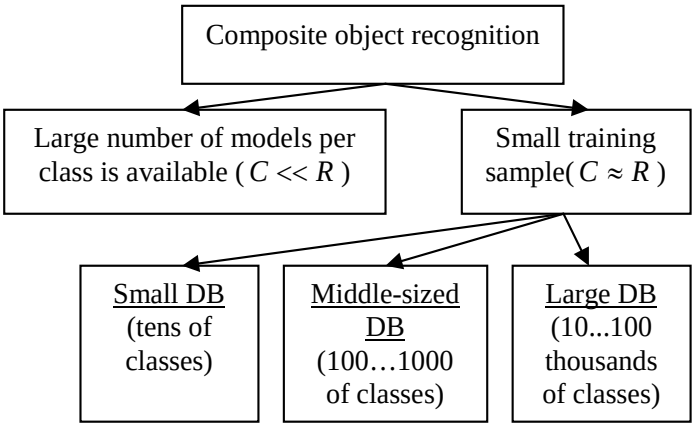


Fig. 2 classification of composite object recognition systems in dependence on the available model DB

A. Large Number of Models Per Class in the Database

The most well-studied are tasks with large number of available models for each class ($C << R$), e.g., optical character recognition, classification of traffic signs and phonemes. Modern recognition methods in such cases can be classified in the following way (Fig. 3).

Following traditional approach to pattern recognition [1], feature extraction is the crucial step to achieve high accuracy. Experimental studies clearly show that popular classifiers (LDA/QDA, MLP, SVM, etc.) are characterized by the best quality for uncorrelated features [1]. Hence, classical recognition procedures include normalization and decorrelation with, e.g., principal/independent component analysis (PCA/ICA) for primitive descriptions of analyzed objects (color matrix for images, signal of acoustic pressure amplitude of its fast Fourier transform (FFT) for speech) [3]. For instance, weighed histograms of gradient orientation are calculated in the image keypoints' neighborhoods [7]. For speech fragments the spectrum is estimated in several acoustic frequency bands (logarithmic mel-scale) and further

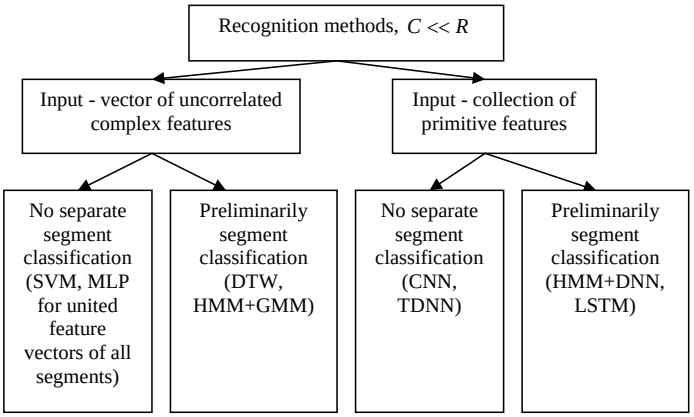


Fig. 3 classification of composite object recognition methods for $C << R$

decorrelated by the modified cosine transform. As a result, Mel-Frequency Cepstral Coefficients (MFCC) are obtained [10]. Thus, the most evident way to recognize composite object is to divide it into a *fixed* number of homogeneous segments, estimating features for all segments, uniting them into a single feature vector and classifying them with traditional MLP or SVM.

Unfortunately, homogeneous segments are extracted inaccurately, several important segments can be duplicated (e.g., vowels in speech signals) or missed (consonant phonemes). Hence, the described approach is ineffective for many tasks, such as automatic speech recognition (ASR). To overcome this drawback, preliminary segmentation of the query object and all models is performed. Next, the segments are dynamically aligned with the dynamic programming techniques (Dynamic Time Warping, DTW) [12]: each segment of the input object is compared with *several* model segments in some *neighborhood*. It is obvious that such alignment causes further increase of average recognition time. Unfortunately, such approach is known to be characterized with low accuracy if the number of classes C becomes high [10]. As a consequence, since 1980-th the most popular approach are based on the hidden Markov models (HMM), specially developed for classification of piecewise-stationary objects [13]. Let's discover their application to ASR task, for which HMM is a standard de-facto in modern recognition libraries (CMU Sphinx, HTK, Kaldi, etc.).

Originally, HMMs were built for each word from the vocabulary [13]. As it is required 100...1000 speech signals to train each model, their practical application was restricted only for tasks with small dictionary, e.g., recognition of digits. Hence, nowadays, other, phonetic, approach is widely used [10], according to which each word from the vocabulary is put in correspondence with the sequence of phones. For each phoneme of national language (or, more frequently, context phones such as triphones) separate HMM is built. For instance, HMM with 3 states is usually applied for triphone in which the context of the middle phone is specified with the previous and

next phones. Observed variables are the vector of MFCC features calculated in a frame of fixed duration 30-45 ms. It is assumed that this vector is generated by the Gaussian Mixture Model (GMM) with diagonal covariance matrix, i.e., the features are supposed to be uncorrelated. Training is done by using the large speech corpora with partially available phonetic transcription. Other part pf the corpora is used to clarify the HMM's parameters with the semi-supervised learning [14]. To improve the accuracy, the information of statistical correlation between segments (e.g., language model [10]) in syntactic, lexical and semantic levels in hierarchical systems (e.g., Hearsay-II [15]) is used.

The major restriction of these approaches is the requirement of features to be uncorrelated or independent (Fig. 3). It is not surprising that the recent research has been focused on the usage of primitive correlated features and more complex classifiers, e.g., deep neural network (DNN) which showed better accuracy in comparison with the state-of-the-art SVM for several model tasks [16]. Originally Restricted Boltzmann Machines are used as stacked auto encoders to extract features with final layers trained by the back propagation on the modern GPU's [16]. However, the best results are achieved with other neural network methods with deep architecture, such as Convolutional Neural Network (CNN) [17] and recurrent LSTM (Long Short-Term Memory) [18] which do not use preliminary unsupervised learning. Let's discuss these techniques for image and speech recognition tasks in more details.

CNN is an enhancement of the neocognitron ideas [19] and consists of sequentially mapped layers of convolution [20] and sub-sampling [17]. Recently, subsampling layer is replaced with the max pooling layer and several CNN's are united in a committee in the Multi-Column GPU Max-Pooling CNN, MC-GPU-MPCNN [22] which allowed to reach 0.54% error rate in traffic sign image recognition task (*C=43* classes, *R*=39209 models) which is 0.6% lower than the human's error rate. For digits character recognition task from the MNIST dataset (*C=10* classes, *R=60000*) standard DNN showed 99.65% accuracy [23], while the best is again the MC-GPU-MPCNN (99.77%) [24]. Application of the CNN to much complex ImageNET dataset (*C*=21841 classes, more than 14 million models) allowed to achieve 11% higher accuracy in comparison with the state-of-the-art methods [25].

Here we should highlight the need for testing if the difference in error rates are statistically meaningful with, e.g., sign rule or McNemar's test [26]. Moreover, the paper [27] discovers problems with estimation of the classifier accuracy so the simplest methods (LDA, naive Bayes, etc.) are sometimes better than the complex classifiers (MLP, SVM, DNN, etc.). Moreover, widely used datasets (MNIST, ImageNET, TIMIT, FERET etc.) are sometimes not representative and class labeling is sometimes quite arbitrary (e.g., in the tasks of emotion recognition from face and speech [28].

It is important to note the CNN can be applied not only in image recognition tasks. For instance, its special case, Time-Delay Neural Network (TDNN) [3] showed good accuracy of phoneme recognition with correlated spectrograms as the features [29], [30].

However, most widely-used in the ASR task are the methods with preliminary phoneme segmentation which allows to use phonetic approach widely developed in the HMM studies. First of all, the DNN with 3...8 layers and 1024...3072 neurons are applied in popular Kaldi ASR library instead of the HMM-GMM triphone model [31]. The DNN's input is usually the features of several (e.g., 9) sequential overlapped frames [32]. It was showed by Microsoft research team [33] that simple FFT-based features allow the DNN to achieve better accuracy than the conventional MFCC. DNN's usage is the very promising approach now as they decrease the word error rate at 10-15% in comparison with the state-of-the-art HMMs [31], [32].

Final alternative to HMM is the recurrent LSTM trained with the Connectionist Temporal Classification method [34]. In the task of phonetic transcription with TIMIT dataset network with 100 LSTM blocks in each hidden layer reaches the 17.7% error rate [35] and the phonetic labeling accuracy is 69.49% which is 4-8% higher in comparison with HMM [36].

Thus, we may conclude that classifiers with deep architecture and large number of parameters for several model tasks allows to achieve the accuracy comparable with the human recognition accuracy. However, the situation becomes quite more complicated if the training set contain small number of models per each class (in the worst case, 1 model per class, *C=R*) [37].

B. Small Training Sample

In case of the low number of models per class $C \approx R$ the classification procedures can be divided into deterministic and statistical (Fig. 4).

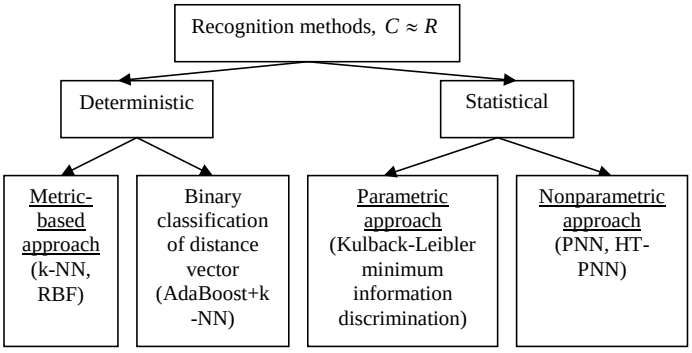


Fig. 4 classification of composite object recognition methods for $C \approx R$

Nowadays the most popular is the deterministic approach in which it is required to choose the proper distance between analyzed objects. To classify the query object, nearest neighbor (NN)-based rules, e.g., kNN [1] or RBF (Radial-Basis Function) [3], [38].

One variant of deterministic approach is the binary

classification of the distances between segments [39]. It is assumed that 2...5 models are available for each class. Despite the impossibility to train complex classifiers, e.g., CNN, for such training set, it is possible to assign the vector of distances between corresponding segments to one of two classes: are these distances calculated between objects of the same or distinct classes. The training is usually done with the AdaBoost [40]. Query object is segmented, for each model from DB the distance vector is estimated and recognized by trained AdaBoost classifier. The decision is made in favor of the class of the model with the highest confidence. Actually, it is the same NN rule, but the distance function is the AdaBoost's confidence. The usage of this approach to face recognition task from FRGC 2.0 dataset (16028 photos of 4444 persons) allowed to achieve 6-15% higher accuracy in comparison with conventional 1-NN rule [39].

The problem here is the variability of objects in each class. To overcome it, statistical approach [1] can be used. Each class is specified by the distribution of segments' feature vectors. Hence, pattern recognition task can be reduced to the statistical hypothesis testing of group's (sequence of primitive features, like in the DNN-based methods) distribution. In *parametric* approach the type of distribution is chosen by the researcher. Usually Gaussian approximation (in LDA/QDA) or polynomial distribution (histograms) for discrete features are used. It is possible to show that this approach is equivalent to the Kullback-Leibler minimum information discrimination principle [41], [42] if the segments are considered as a simple random sample of i.i.d. primitive features. However, in general case parametric approach is unreasonable [1]. It is preferable to us the nonparametric estimation of the class distribution with, e.g., Parzen kernel as it is done in the probabilistic neural networks (PNN) [43]. The PNN is characterized by extremely fast training procedure and convergence to the optimal Bayesian decision. Unfortunately, the property of optimality is missed if the class distributions are assumed to be unknown. In such case, it is possible to reduce the task to the testing of complex hypothesis and discover maximal likelihood decision which is known to be asymptotically (if the size of the group, i.e. the image resolution or the phoneme duration is large) equivalent to the minimax criterion [44]. Savchenko A.V. implemented this idea in the homogeneity testing PNN (HT-PNN) which is the enhancement for the PNN to a group-choice classification [45], [46]. HT-PNN showed higher accuracy than the PNN in the tasks of face recognition and authorship attribution [45], [47]. It is important to emphasize that the statistical approach for discrete features with either PNN or HT-PNN is equivalent to the 1-NN rule with special measures of similarity which are the generalization of the Kullback-Leibler divergence and chi-square distance, respectively [46].

Thus, practically all recognition methods for the case $C \approx R$ are implemented in k-NN rules which requires the brute force of the whole DB. Hence, they cannot be implemented in real-time applications even for middle-sized (thousands of classes) and, especially, very-large DBs. In the last case the recognition accuracy is known to be very low so that classifier is usually integrated into a decision-support system of object retrieval which returns several closest models to help the user to make better decision. To speed-up the classification, approximate NN methods [48], [49] are usually applied. Unfortunately, these methods are specially optimized for very-large DBs so they are not efficient for middle-size DB when the number of classes does not exceed several thousands [50]. To decrease the recognition speed for such training sets, ordering permutations (perm-sort) method has recently been proposed [49]. Another interesting approach, namely, the directed enumeration method (DEM) [51] is based on the asymptotic properties of the HT-PNN [8]. Experimental studies in face recognition [50] showed that the DEM outperforms the known approximate NN methods in face recognition with FERET and Essex datasets.

III. Experimental Results

In this section we present several experimental results of the HT-PNN, the DEM and the phonetic encoding method (PEM) proposed by Savchenko A.V. in [45], [8], [52] and [53]-[55], respectively.

A. Face recognition

In this section we present experimental results in the face recognition task. Popular FERET dataset was used: 2720 frontal facial images of *C=994* persons were selected. Faces are detected by OpenCV library. All facial images are divided into 100 segments (blocks) by 10x10 grid ($K_1 = K_2 = 10$). Next, HOG with *N=8* bins is evaluated for each segment. The following nearest-neighbor rule was used [51]

$$\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} \min_{\substack{|\Delta_1| \le \Delta, \\ |\Delta_2| \le \Delta}} \rho\left(H_r\left(k_1+\Delta_1, k_2+\Delta_2\right), H\left(k_1,k_2\right)\right) \to \min_{r \in \{1,\ldots,R\}},$$

where $H\left(k_1,k_2\right)$ is the HOG of the segment at the cell

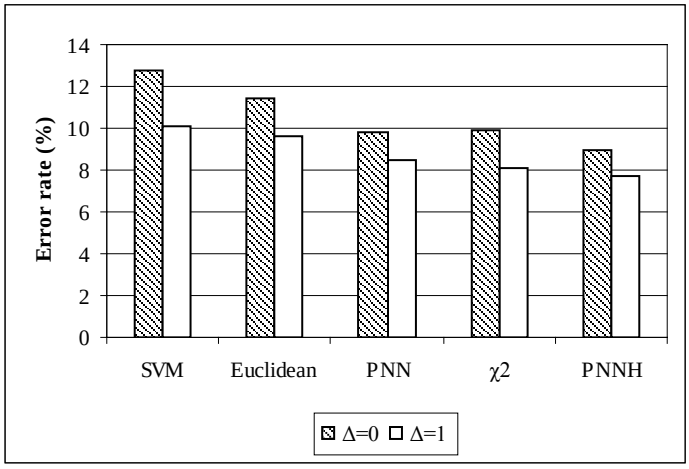


Fig. 5 Face recognition error rate, FERET dataset

$(k_1, k_2), k_1 \in \{1,...,K_1\}, k_2 \in \{1,...,K_2\}$, $H_r(k_1+\Delta_1, k_2+\Delta_2)$ is the HOG of *r*-th model segment at the cell $(k_1+\Delta_1, k_2+\Delta_2)$, $\rho(H_r(k_1+\Delta_1, k_2+\Delta_2), H(k_1, k_2))$ is an arbitrary distance between these HOGs, Δ - numeric size of the segment neighborhood. We tried both conventional value $\Delta = 0$ (no alignment) and $\Delta = 1$.

In the experiment the following popular distances were used: Euclidean, chi-square and PNN with Gaussian kernel [43] and smoothing parameter $\sigma = 3{,}5$ and compare them with our HT-PNN [45], [47]. Training set contained *R=1370* images, test set consisted of other 1350 photos of the same persons. Recognition error rate was estimated by 100-times repeated random sub-sampling cross-validation. Estimated error rates are shown in Fig. 5. McNemar's test with confidence level 0,05 showed that HOG's alignment ($\Delta = 1$) is characterized by statistically significant higher accuracy than the conventional approach ($\Delta = 0$). Improvements of error rate of our HT-PNN are significant for $\Delta = 0$. However, difference in the accuracy of the HT-PNN and chi-square distance for $\Delta = 1$ is not statistically significant and can be explained by random factors.

Based on Fig. 5 one can notice that error rate of traditional SVM and nearest-neighbor rule with Euclidean distance is too high. It seems that SVM needs more photos of one person to be trained efficiently. Moreover, we confirmed that the accuracy of the PNN is less than the accuracy of criterion based on homogeneity testing (chi-square, HT-PNN). Finally, in all cases HT-PNN showed the lowest error rate.

In the next experiment we measure the computing efficiency of the DEM and compare it with conventional brute-force and randomized kd-trees from FLANN library in both non-parallel and parallel environment (*T=8* threads). To estimate average recognition time modern laptop was used (Intel Core i7-2630QM, 4 kernels, 2GHz, RAM 6 Gb). The Windows ThreadPool API is used to implement parallel processing. The training set was randomly divided into *T* parts, each part is processed by separate task. The results are shown in Fig. 6 (conventional case $\Delta = 0$) and Fig. 7 (HOG's alignment, $\Delta = 1$).

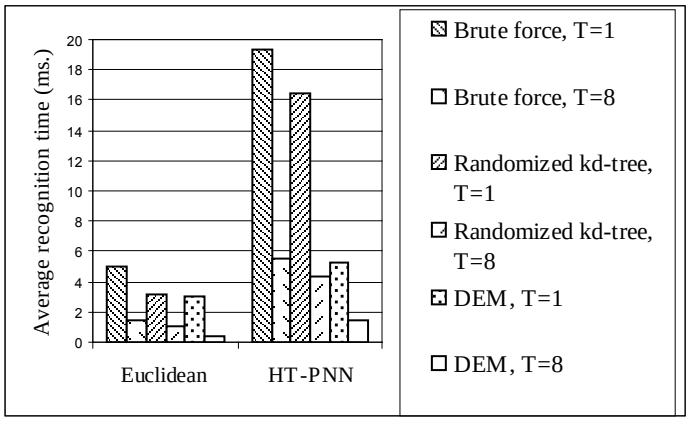


Fig. 6 Face recognition average recognition time, $\Delta = 0$

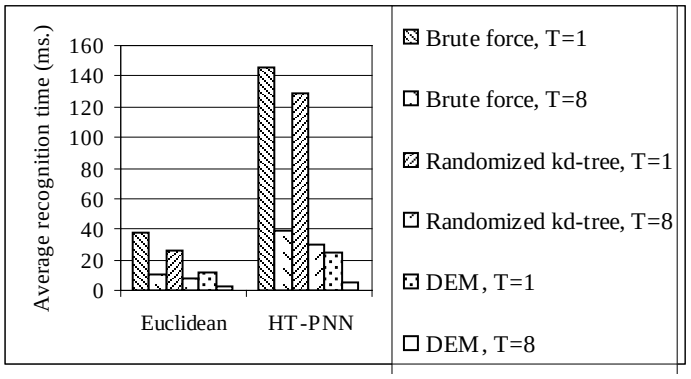


Fig. 7 Face recognition average recognition time, $\Delta = 1$

From these figures one can notice that the DEM allows to improve image recognition computing efficiency. At the same time, one of the best approximate nearest neighbor methods, namely, randomized kd-tree, in some cases shows practically the same computing efficiency as simple brute force. As a matter of fact, parallel DEM [51] is 6-10 times more efficient than nonparallel brute-force and 2-2,5 times better than the parallel one. For instance, recognition time of the exhaustive search implementation of the HT-PNN requires 38.7 ms which is quite high for real-time processing (as face detection is also quite hard procedure in terms of computing efficiency). However, the parallel DEM requires only 5.8 ms, while the accuracy is only 0.5% lower.

B. Voice command recognition

In this subsection we investigate the voice command recognition software [56] developed by A.V. Savchenko and V.V. Savchenko. It implements the PEM [53] which requires the user to pronounce the commands in isolated syllable mode. It has been shown that such requirement allows to build fast robust voice control system with automatic extraction of the pronounced command from continuous speech [56] and very fast speaker adaptation. In [57] it has been proposed to further improve the recognition accuracy by aggregating with fixed weight $0 \le \alpha < 1$ the output of the PEM with the posterior probability of extracted syllable estimated by any ASR library. Popular CMU Pocketsphinx library was used in the experiment.

The noise-canceling microphone A4Tech HS-12o was used to record speech in the following format: PCM wav, mono, sampling rate 8000 Hz, 16 bits per sample. The ASR quality was tested with the following vocabulary (in Russian): the list of 1830 Russian cities with the corresponding regions, e.g., "Kstovo (Nizhegorodskaya)" (hereinafter "Cities"). All vocabularies are available in text files, each line contains separate word/phrase. Ten speakers (five men and five women of different age) pronounced each word from these vocabularies twice. To train the system each speaker pronounce 10 vowels in isolated mode. The following parameters were chosen: frame length 30 ms, frame overlap 10 ms, autoregression model order *p*=12. To estimate the closeness of the speech signals, the implementation of the HT-

PNN, namely, the COSH distance between power spectral densities was used [10], [56].

In our experiments we compare the recognition accuracy of the proposed approach with conventional Pocketsphinx and popular Google Voice Search service [58]. We added an artificially generated white noise to each test signal (with signal-to-noise ratio 30 dB, 20 dB, 10 dB). The error rates are shown in Fig. 8.

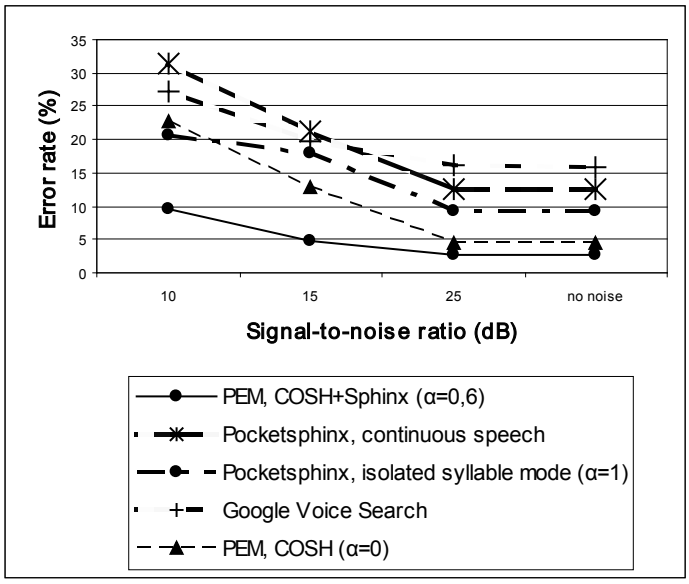


Fig. 8 ASR error rates

Based on this figure we could draw the following conclusions. First, isolated syllable mode allow to not only extract voice commands but to increase the ASR accuracy [53]. Second, addition of noise leads to high error rate of the original PEM without classifier fusion ($\alpha = 0$) as the vowels from the speaker's phonetic database obtained during training became not similar to the noised vowels, so adaptation is not efficient. However, the fusion of general ASR with the user vowel recognition ($\alpha = 0.6$) is the best choice in our experiment with isolated syllables [57]. McNemar's test showed that the gain in the accuracy is statistically significant in all cases except the absence of noise. In the last case, fusion is not quite better than the original PEM ($\alpha = 0$).

IV. Conclusion

In this paper we analyzed the methods of piecewise-regular object classification. The dependence of the classifier choice on the number of classes and models in DB is highlighted (Fig. 2). Our brief survey showed that the current trends in the development of composite object recognition methods are connected with the refusal of complex algorithms of uncorrelated feature extraction and complication of the classifiers (Fig. 3). We emphasized one of the most exciting challenges in this field, namely, small number of models per each class (Fig. 4). Most of recognition procedures in this case implement the nearest neighbor rule with various distances. If the number of classes C is large, brute force solution is not computing efficient. Hence, approximate nearest neighbor algorithms can be applied. Unfortunately, most of such algorithms allows to make the recognition faster only for very-large DBs and do not work better than an exhaustive search for middle-sized DBs (Fig. 6, 7). However, our experimental study show that there is the DEM which improves the recognition speed in several times.

References

[1] S. Theodoridis, K. Koutroumbas, *Pattern Recognition*, 4th ed. Burlington, MA; London: Academic Press, 2008.

[2] R. A. Abusev, "On Group Choice Procedures for Problems of Classification and Reliability in the Case of Lognormal Variance", *Journal of Mathematical Sciences*, vol. 189, no. 6, pp. 911–918, 2013.

[3] S. O. Haykin, *Neural Networks and Learning Machines*. 3th ed.. Harlow: Prentice Hall, 2008.

[4] C. Cortes, V. N. Vapnik, "Support-Vector Networks", *Machine Learning*, vol. 20, no.3, pp. 273-297, 1995.

[5] L. Rutkowski, *Computational Intelligence: Methods and Techniques*. Springer, 2010.

[6] P. Prandoni, M.Vetterli, "Approximation and compression of piecewise smooth functions", *Philosophical Transactions of the Royal Society*, vol. 357, no. 1760, pp. 2573-2591, 1999.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", *in 2005 Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) Conf.*, pp. 886–893.

[8] A. V. Savchenko, "Directed enumeration method in image recognition", *Pattern Recognition*, vol. 45, no 8, pp. 2952–2961, 2012.

[9] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", *International Journal of Computer Vision*, vol. 60, no. 2. pp. 91–110, 2004.

[10] J. Benesty, M. M. Sondhi and Y. Huang (eds.) *Springer Handbook of Speech Processing*. Berlin u.a.: Springer, 2008.

[11] Y. Qiao, N. Shimomura, N. Minematsu, "Unsupervised optimal phoneme segmentation: Objectives, algorithm and comparisons", *in 2008 Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, pp. 3989–3992.

[12] C. S. Myers and L. R. Rabiner, "A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word Recognition", *Bell System Technical Journal*., vol. 60, no. 7, pp. 1389–1409, 1981.

[13] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, N.J: Prentice Hall, 1993.

[14] O. Chapelle, B. Schölkopf, A. Zien *Semi-Supervised Learning*. Cambridge, Mass: The MIT Press, 2010.

[15] L. D. Erman, F. Hayes-Roth, V. R. Lesser and D. R. Reddy, "The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty", *ACM Computing Surveys*, vol. 12, no 2, pp. 213–253, 1980.

[16] G. E. Hinton, S. Osindero, Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets", *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[17] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[18] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", *Neural Computation*, vol. 9, no 8, pp. 1735–1780, 1997.

[19] K. Fukushima, "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position", *Biological Cybernetics*, vol. 36, pp. 193–202, 1980.

[20] L. G. Shapiro and G. C. Stockman, *Computer Vision*. Upper Saddle River, NJ: Prentice Hall, 2001.

[21] A. V. Savchenko, "Adaptive video image recognition system using a committee machine", *Optical Memory and Neural Networks*, vol. 21, no. 4, pp. 219–226, 2012.

[22] D. Cireşan, U. Meier, J. Masci and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification", *Neural Networks*, vol. 32. pp. 333–338, 2012.

[23] D. Cireşan, U. Meier, L.M. Gambardella and J. Schmidhuber, "Deep, Big, Simple Neural Nets for Handwritten Digit Recognition", *Neural Computation*, vol. 22, no. 12, pp. 3207–3220, 2010.

[24] J. Schmidhuber, "Multi-column Deep Neural Networks for Image Classification", *in 2012 Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3642–3649.

[25] A. Krizhevsky, I. Sutskever and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", in *Advances in Neural Information Processing Systems* 25 / ed. Pereira F. et al. Curran Associates, Inc., pp. 1097–1105, 2012.

[26] L. Gillick and S.J. Cox, "Some statistical issues in the comparison of speech recognition algorithms", *in 1989 Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP-89)*, pp. 532–535.

[27] D. J. Hand, "Classifier Technology and the Illusion of Progress", *Statistical Science*, vol. 21, no. 1, pp. 1–14, 2006.

[28] B. Schuller, A. Batliner, S. Steidl and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge", *Speech Communication*, vol. 53, no. 9-10, pp. 1062–1087, 2011.

[29] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang, "Phoneme recognition using time-delay neural networks", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[30] L. Bottou, F. Fogelman Soulié, P. Blanchet and J.S. Liénard, "Speaker-independent isolated digit recognition: Multilayer perceptrons vs. Dynamic time warping", *Neural Networks*, vol. 3, no. 4, pp. 453–465, 1990.

[31] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups", *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[32] A. Ghoshal, P. Swietojanski and S. Renals, "Multilingual training of deep neural networks, *in 2013 Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pp. 7319–7323.

[33] J.-T. Huang, J. Li, D. Yu, L. Deng and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers", *in 2013 Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pp. 7304–7308.

[34] A. Graves, S. Fernández and F. Gomez "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks", *in 2006 Proc. International Conference on Machine Learning (ICML 2006)*, pp. 369–376.

[35] C. Chow, "On optimum recognition error and reject tradeoff", *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, 1970.

[36] A. Graves, A. Mohamed and G. E. Hinton, "Speech recognition with deep recurrent neural networks", *in 2013 Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, pp. 6645–6649.

[37] X. Tan, S. Chen, Z.-H. Zhou and F. Zhang, "Face recognition from a single image per person: A survey", *Pattern Recognition*, vol. 39, no. 9, pp. 1725–1745, 2006.

[38] V. R. Milov, "Synthesis of a nonparametric classifier on the basis of RBF artificial neural networks", *Radiophysics and Quantum Electronics*, vol. 46, no. 2, pp. 128-133, 2003.

[39] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S.Z. Li, "Learning Multi-scale Block Local Binary Patterns for Face Recognition", in *Advances in Biometrics* / ed. Lee S.-W., Li S.Z. Springer Berlin Heidelberg, pp. 828–837, 2007.

[40] G. Zhang, X. Huang, S. Z. Li, Y. Wang and X. Wu, "Boosting Local Binary Pattern (LBP)-Based Face Recognition", *in Advances in Biometric Person Authentication* / ed. Li S.Z. et al. Springer Berlin Heidelberg, pp. 179–186, 2005.

[41] S. Kullback, *Information Theory and Statistics*. Mineola, N.Y: Dover Publications, 1997.

[42] V. V. Savchenko, "Automatic speech processing according to the minimum-information-mismatch criterion based on the whitening filter method", *Journal of Communications Technology and Electronics*, vol. 50, no. 3, pp. 286-291, 2005.

[43] D. F.Specht, "Probabilistic neural networks", *Neural networks*, vol. 3, no. 1, pp. 109–118, 1990.

[44] A. A. Borovkov, *Mathematical Statistics*. Gordon and Breach Science Publishers, 1998.

[45] A. V. Savchenko, "Probabilistic neural network with homogeneity testing in recognition of discrete patterns set", *Neural Networks*, vol. 46, pp. 227–241, 2013.

[46] A. V. Savchenko, "Statistical Recognition of a Set of Patterns Using Novel Probability Neural Network", *in 2012 Proc. of International Conference on Artificial Neural Networks and Pattern Recognition (ANNPR-2012)*, LNCS/LNAI, vol. 7477, pp. 93-103, 2012.

[47] A. V. Savchenko, "Nonlinear Transformation of the Distance Function in the Nearest Neighbor Image Recognition", *in 2014 Proc. International Conference on Computational Modeling of Objects Presented in Images (CompIMAGE 2014)*, Y. Zhang and J.M.R.S. Tavares (Eds.): CompIMAGE 2014, LNCS, vol. 8641, pp. 261-266, 2014.

[48] C. Silpa-Anan and R. Hartley, "Optimised KD-trees for fast image descriptor matching", *in 2008 Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pp. 1–8.

[49] E.C. Gonzalez, K. Figueroa and G. Navarro, "Effective Proximity Retrieval by Ordering Permutations, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 9, pp. 1647–1658, 2008.

[50] A. V. Savchenko, "Face Recognition in Real-Time Applications: A Comparison of Directed Enumeration Method and K-d Trees", *in Proc. 2012 International Conference on Perspectives in Business Informatics Research (BIR 2012)* / ed. Aseeva N., Babkin E., Kozyrev O. Springer Berlin Heidelberg, LNBIP, vol. 128, pp. 187–199, 2012.

[51] A. V. Savchenko, "Real-Time Image Recognition with the Parallel Directed Enumeration Method", *in 2013 Proc. of International Conference on Vision Systems (ICVS 2013)*, M. Chen, B. Leibe, and B. Neumann (Eds.), LNCS, vol. 7963, pp. 123-132, 2013.

[52] A. V. Savchenko, "Image Recognition with a Large Database Using Method of Directed Enumeration Alternatives Modification", *in 2011 Proc. of International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC-2011)*, LNCS/LNAI, vol. 6743, pp. 338-341, 2011.

[53] A. V. Savchenko, "Phonetic encoding method in the isolated words recognition problem", *Journal of Communications Technology and Electronics*, vol. 59, no. 4, pp. 310-315, 2014.

[54] L.V. Savchenko and A.V. Savchenko, "Fuzzy phonetic decoding method in a phoneme recognition problem", *in 2013 Proc. International Conference on Nonlinear Speech Processing (NOLISP 2013)*, eds. Dragman, T., Dutoit, T., LNCS, vol. 7911, pp. 176-183. Springer, Heidelberg, 2013

[55] A. V. Savchenko, L. V. Savchenko "Classification of a Sequence of Objects with the Fuzzy Decoding Method", *in 2014 Proc. of International Conference on Rough Sets and Current Trends in Computing (RSCTC 2014)*. C. Cornelis et al. (eds.): LNCS/LNAI, vol. 8536, pp.309-318, 2014

[56] A.V. Savchenko, "Phonetic words decoding software in the problem of Russian speech recognition", *Automation and Remote Control*, vol. 74, no, 7, pp. 1225-1232, 2013.

[57] A. V. Savchenko, "Semi-automated Speaker Adaptation: How to Control the Quality of Adaptation?", *in 2014 Proc. of International Conference on Image and Signal Processing (ICISP 2014)*, A. Elmoataz et al. (Eds.), LNCS, vol. 8509, pp. 638-646, 2014.

[58] M. Schuster, "Speech recognition for mobile devices at Google", *in 2010 Proc. of the Pacific Rim International Conference on Artificial Intelligence (PRICAI 2010)*, LNCS/LNAI, vol. 6230, pp. 8-10, 2010.

Methods of assessing and predicting the energy efficiency of electrical complexes of urban distributive power grids

V. Frolov, A. Korotkov

***Abstract*—**Energy efficiency of electrical complexes of urban distributive power grids is determined with an error caused by changes in the characteristics of the main equipment in process of exploitation and the lack of information of such level grids due for the specific features. Development methodology based on experimental research which would improve the accuracy of the estimation and forecasting of energy efficiency in the development and planning measures of energy saving is expedient.

***Keywords*—**Energy efficiency, load schedules, losses of idle transformers, urban distributive power grids.

I. Introduction

NOWADAYS a special attention is paid to energy conservation, energy efficiency and saving energy resources.

Operating efficiency of electrical complexes of urban distributive power grids is largely dependent on condition of the equipment and information about its operating modes. The efficacy variable is losses of electricity. Equipment of electrical complexes of urban distributive power grids, as no other power grids, is extremely massive. Even for the electrical complexes of urban distributive power grids of regional centers of the European part of Russia, the number of units of this equipment is the hundreds, thousands and even tens and hundreds of thousands of units.

There are problems with forecasting the state of the equipment of electric grids and modes of its operation at a low level of monitoring. As a result there are problems with forecasting the structure of energy losses as a measure of the efficiency of the electric grids.

Development of methods for assessing and predicting the energy efficiency of electrical complexes of urban distributive power grids is an relevant task, which is primarily associated with the refinement of the basic components of the values of electricity losses and addresses to the problems of assessing and predicting of the structure of electricity losses as the main indicator of energy efficiency. Rational and efficient exploitation of electrical complexes of urban distributive power grids is impossible without a clear understanding of the structure and quantitative components of regulatory and excessive losses [1]. These indicators are expedient to define by calculation. Special investigations should be carried out.

The research is carried out with financial support from the state represented by the Ministry of Education of Russia. Now the unique identifier of applied research RFMEFI57614X0055.

II. Problem Formulation

Existing methods for assessing the energy efficiency of electrical complexes of urban distributive power grids do not account for changes in the characteristics of the equipment during operation and the general low level of information security electrical power grids 6-10 kV and 0.4 kV. It leads to nonobjective results.

The main components of the total losses of electricity of urban complexes of urban distributive power grids are conditional permanent losses in power transformers, load losses in the lines of 0.4 kV and the losses due to accuracy and disadvantages of electricity metering system. Values of the constituent determined by calculation and requires an increase in the accuracy of calculations.

Currently, a significant part of tenure of employment of the power transformers of distributive grids with highest voltage of 6-10 kV exceeds 25–30 years, and the actual values of the power-load losses are substantially different from the certified value. Existing methods of assessing the energy efficiency does not take account of this fact, it leads to an error estimates.

There are research results in electric systems and municipal electric grids of the Central Federal District of Russia, which suggest that the fact of exceeding the actual values of power load losses of idle power transformers over real values of losses occurs and can reach significant values [2, 4].

Variation of power of load losses of idle power transformers is rated with the excess of passport value $\Delta P_{IR.PASS}$ over the real value of load losses of idle power transformers $\Delta P_{IR.REAL}$ in conducting research. The excess amount is determined by the formula:

$$\Delta P_{IR}^{*}(T_E) = \frac{\Delta P_{IR.REAL} - \Delta P_{IR.PASS}}{\Delta P_{IR.PASS}}, \qquad (1)$$

After the culling the sample with results of experimental researches was composed to analyze them. The sample

ISBN: 978-1-61804-251-4

included the results of researches of 682 power transformers with a rated power of 20 kWA to 630 kVA.

Results of experimental researches are presented in Fig. 1.

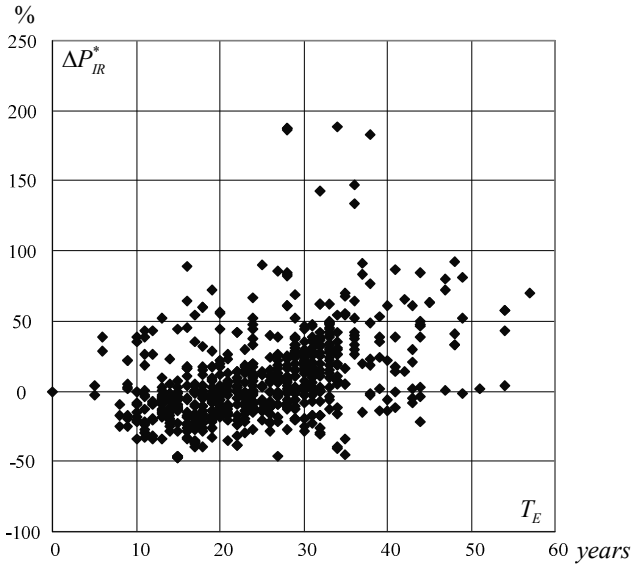


Fig.1. Experimental data about power of load losses of idle power transformers

Obtained values of power of load losses of idle power transformers are characterized by a large spread, while values of power of load losses of idle power transformers increase with increasing of tenure of their employment.

A result of research proved that the most significant factor that affects on the change of power of load losses of idle power transformers is tenure of their employment, therefore the influence of other factors that determine values of power of load losses of idle power transformers was not taken in statistical processing.

According to experimental data regression mathematical models of changing of power of load losses of idle power transformers in service were developed. Goal was to obtain a simple method that allows to determinate by calculation the average value $\Delta P^*_{IR}(T_E),\%$ for power transformers with an arbitrary tenure of employment.

Comparative analysis of the obtained models led to the conclusion that as a relevant and statistically significant mathematical model that most accurately describes the relationship $\Delta P^*_{IR},\% = f(T_E)$ must be selected approximating model in the form of a power equation:

$$\Delta P^*_{IR}(T_E) = -25,3201 + T_E^{1,0935}, \qquad (2)$$

because:

– all the coefficients of the equation are statistically significant;

– determination coefficient R^2 reaches the highest value;

– obtained model is adequate in relation to the experimental data.

Mathematical model (2) was the basis for the development of an original method for calculating power of load losses of idle power transformers with different tenure of employment.

Practical and simple method of constructing models $\Delta P^*_{IR},\% = f(T_E)$ was proposed in the form of two linear dependencies $\Delta P^*_{IR},\% = f(T_E)$ for different intervals of tenure of employment:

– from 0 to 20 years – a linear dependence $\Delta P^*_{IR},\% = 0$ $(\Delta P_{IR} = \Delta P_{IR.PASS})$, this seems absolutely logical for new transformers and transformers with low tenure of employment;

– from 20 years and more – linear dependence $\Delta P^*_{IR},\% = a_1 \times \Delta P_{IR.PASS}$, where value of coefficient a_1 obtained from experimental data.

For power transformers with tenure of employment more than 20 years dependence $\Delta P^*_{IR},\% = f(T_E)$ can be described by the equation $\Delta P^*_{IR},\% = 1.75 \times (T_E - 20)$ according to which calculated value of power of load losses of idle power transformers increased by 1.75% for each year of their employment.

Obtained method of determining of power of load losses of idle power transformers and its comparison with the power mathematical model that best describes the relationship $\Delta P^*_{IR},\% = f(T_E)$, graphically presented in Fig. 2.

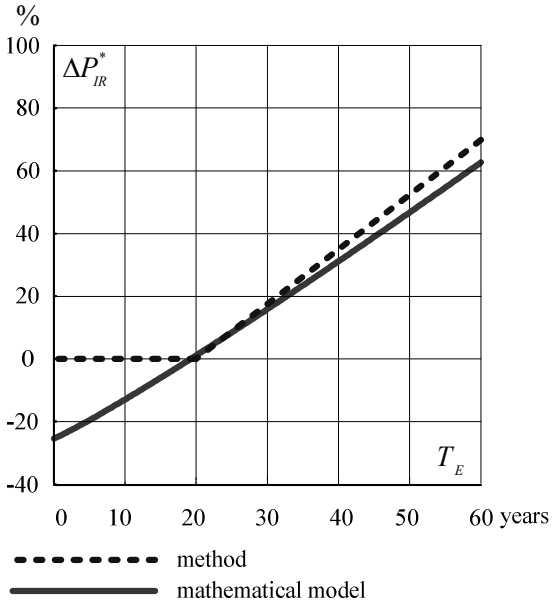


Fig.2 A graphical representation of the method and its comparison with the developed mathematical model

Random measurements of power of load losses of idle power transformers of grids of 6-10 kV, that were put into operation between 1962 and 1993, were carried out for assess of proposed method. A comparison of the measurement values

$\Delta P_{IR.REAL}$ with calculated values $\Delta P_{IR.CALC}$ was carried out.

The results of control tests have confirmed the legitimacy of the proposed method of calculating of power of load losses of idle power transformers. Here criterion of depending of a single parameter T_E works for the average of a group of indicators. The greater the size of the group, the more accurate determination of the desired result. With a significant difference between the calculated and experimental values of the power of load losses of idle power transformers in groups, the difference $\Delta P_{IR.REAL}$ and $\Delta P_{IR.CALC}$ on all transformers in sample was 0.39% compared to $\Delta P_{IR.CALC}$ The obtained result confirms the accuracy of the calculations of power of load losses of idle power transformers by the proposed method for a group of power transformers with the chosen T_E as a criterion [2].

Very close values were obtained and the results of [4], fragments of which are shown in Fig. 3 for transformers with highest voltage of 35 kV.

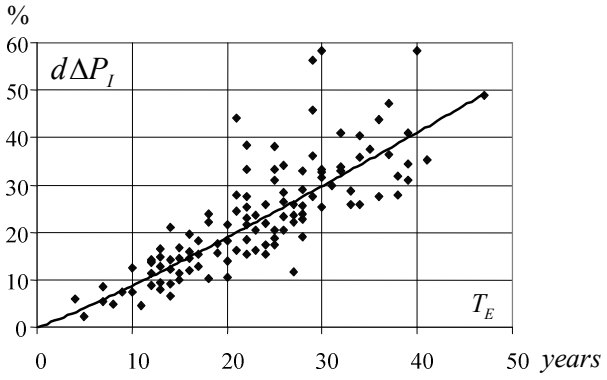


Fig.3. Approximation of deviations of power losses in magnetic conductors of transformers of 35 kV with power function

It is possible to lead the formula:

$$d\Delta P_{I.i} = 0.66 \cdot T_{E.i}^{1,12}, \quad (3)$$

The designations $d\Delta P_{I.i}$ and $T_{E.i}$ adopted by the authors of [4] identical to designations $\Delta P_{IR}^*, \%$ and T_E adopted by the authors of abstract accordingly.

Comparison shows that for the 50 years of service the calculated values $\Delta P_{IR}^*, \%$ for both methods are of the close to 50%. Obtained fact indicates that not only the values $\Delta P_{IR}^*, \% = f(T_E)$ exceed passport values of power of load losses of idle power transformers, but that for transformers of different types and capacities power variation $\Delta P_{IR}^*(T_E), \%$ obeys certain general laws, which can be installed as a result of special researches.

The general level of condition monitoring of electrical complexes of urban distributive power grids is low, therefore significant problems of calculating of load electricity losses associated with the lack of initial information about the parameters of the equipment and operating conditions of 0.4 kV. This is due to a lack of recording devices at 0.4 kV feeders supplying urban substations and, as a consequence, the lack of data about the load schedules and indicators used in the calculations [3].

Analysis of scientific literature, techniques and methods of calculation, and the practice of the work on the calculation of losses in complexes of urban distributive power grids showed insufficient research problems, associated with obtaining accurate data about size and structure of electricity losses.

Rules of exploitation of urban distributive power grids do not provide registration of load schedules of 0.4 kV grid, so it is recommended to use the value of form coefficient of graph k_f obtained by calculation according to the formula:

$$k_f^2 = \frac{1 + 2k_z}{3k_z}, \quad (4)$$

where $k_z = 0.5$, by $k_f = 1.15$ $(k_f^2 = 1.33)$, and $k_f = 1.33$ $(k_f^2 = 1.67)$ by $k_z = 0.3$.

Schedules of consumption of active power $P(t)$ and reactive power $Q(t)$ are presented as graphs $P^*(t)$ and $Q^*(t)$ in relative units with respect to the average daily values $P_{AV.NIGHT}$ and $Q_{AV.NIGHT}$ in the expressions:

$$P^*(t) = \frac{P(t)}{P_{AV.NIGHT}}, \; Q^*(t) = \frac{Q(t)}{Q_{AV.NIGHT}}. \quad (5)$$

Example of received daily schedules of active load $P^*(t)$ of residential consumers is shown in Fig.4

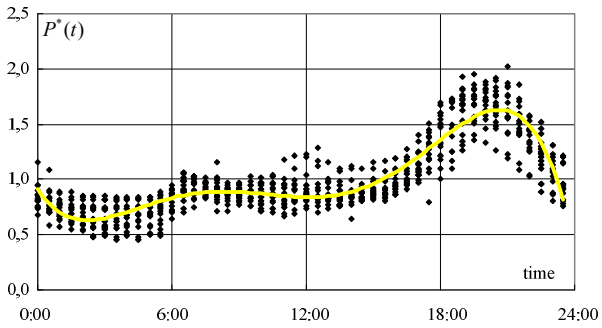


Fig. 4. Daily schedules of active load of residential consumers

Dependence of the change shape of the schedules from: season of the year, the locality and the day of the week is set with a result of research of daily schedules of electrical loads $P^*(t)$ The absence of such dependencies allows to combine

schedules in groups for further research.

Preliminary research showed that in contrast to the schedules of active power $P^*(t)$, reactive power schedules $Q^*(t)$ have not pronounced peaks, although in the form of graphics, they repeat the active power (Fig. 5).

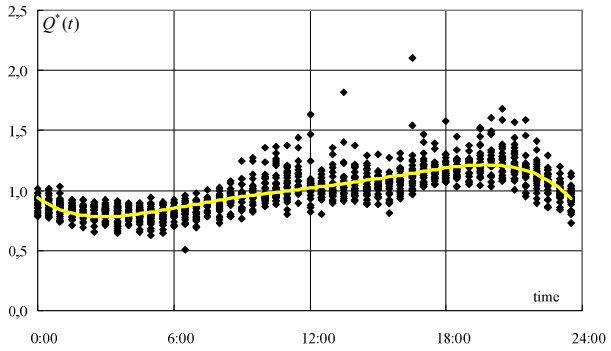


Fig. 5. Daily schedules of reactive load of residential consumers

It was established, that values of reactive power factor $tg\varphi(t) = Q^*(t) / P^*(t)$ have considerable variation depending on time of day and can be described by only an interval of values which is characterized by an average value. Such a notion is reflected in the current regulations.

A problem to develop methods for assessing and predicting the energy efficiency of electrical systems of urban distributive power grids is posed. Methods provide clarification of the calculated values of power losses as an index of energy efficiency in conditions of changing the characteristics of the equipment at a low level of information security.

III. Problem Solution

Analysis of the structure of electricity losses in urban distributive power grids with the highest voltage of 6-10 kV of the North-West Federal District of the Russian Federation shows that the energy losses in power transformers are the most significant and primary depend of tenure of their employment, but it ignored in modern methods of calculation. The accuracy of calculations of losses is particularly important due to the fact that currently the share of power transformers with a tenure of employment of 25–30 years more than 50% in urban distributive power grids, while changes in power load losses of power transformers reaches 50% or more from the certified value during exploitation.

Plans and graphics of measurements of power load losses of idle run of significant part of power transformers of Northwestern Federal District, commissioned in the period from 1941 to 2004, were developed.

Regression mathematical model of change in capacity-load losses of idle run of power transformers during exploitation will be developed based on these results. The goal is to obtain a simple method to determine the average value of load losses of idle run of group of power transformers with a random tenure of employment by calculations.

A problem to determine the actual values of form coefficient for the calculation of load losses in electrical power complexes of urban distributive power grids of 0.4 kV is posed.

Actual material about the load schedules of residential consumers in towns of the North-West Federal District was received by experimental studies of daily schedules of electrical loads feeders providing electricity power characteristic groups of residential customers.

Analysis of schedules of active and reactive load residential customers, united similar living conditions, will provide an idea of the features of electricity in different social living conditions and classify electricity consumers based on the type of building groups. According to obtained data regression mathematical model that describe changes in the active power during the day for the groups will be developed.

Obtained mathematical model will be the basis for the development of methods for determining the basic indicators of daily schedules of active load: electricity supply to the consumer group, the time of using the maximum load, the absolute minimum and maximum of active power of consumption for any period of time. Indicators are determined by mathematical calculations and analysis of expressions describing graphs and their combinations.

Methods and algorithms for determining indicators load schedules of different groups of consumers of urban distributive power grids and the values of the coefficients will improve the accuracy of calculations of losses of electricity for each case in a low level of monitoring of 0.4 kV grids, reasonably solve issues of forecasting and assessing the energy efficiency of these grids and its modernization.

IV. Conclusion

The analysis of existing evaluation methods of energy efficiency of urban distributive power grids shows the need for research to improve the accuracy of the determination and prediction of energy efficiency of grids in modern conditions.

Receiving the results of the change of power load losses of power transformers during exploitation, and also information about the characteristics of load schedules of consumers of urban power grids will improve the accuracy of calculations of losses in distributive power grids high voltage of 6-10 kV. They can be used to assess and predict the energy efficiency of electrical systems of urban distributive grids for the development of energy conservation and energy efficiency at a low level of equipment condition monitoring, its features and modes of operation.

References

[1] A. V. Korotkov, V. Y. Frolov, “About the improving of accuracy of definition of load losses of electrical energy and structure of actual losses”, in *St. Petersburg State Polytechnical University Journal*, no.1, 2012, pp. 41–44.

[2] Y. B. Kazakov, A. B. Kozlov, V. V. Korotkov “Accounting changes of losses of idle run of transformers in tenure of employment of distributive electrical grids”, in *Elektrotekhnika*, vol.5, 2006, pp. 11–16.

[3] A. V. Korotkov, V. Y. Frolov, "Graphs of active and reactive loads of residential consumers", in "*Vestnik IGEU" journal*, vol.5, 2011, pp. 29–31.
[4] A. A. Balabin, V. F. Zaugolnikov, A. A. Savinkov, "Some aspects of the economic operation of power transformers" in *Promyshlennaya energetika*, vol.4, 2006, pp. 10–14.

Modeling silicon spintronics

Viktor Sverdlov, Joydeep Ghosh, Dmitri Osintsev, and Siegfried Selberherr
Institute for Microelectronics, Technische Universität Wien
Gusshausstrasse 27–29, A-1040 Vienna, Austria
Email: {sverdlov|ghosh|osintsev|selberherr}@iue.tuwien.ac.at

***Abstract*— Silicon, the main material of microelectronics, is perfectly suited for spin-driven applications. All-electrical spin injection in silicon has been demonstrated, however, the magnitude of the corresponding signal is larger than theoretically predicted. We analyze the influence of electrostatic charge screening on the efficiency of spin injection at the ferromagnet-semiconductor interface. We show that the spin-injection efficiency cannot exceed the value obtained at the charge neutrality condition. Finally, we demonstrate that a large enhancement of the electron spin lifetime in silicon thin films can be obtained by applying shear strain, which is routinely used to boost the electron mobility in MOSFETs.**

***Index Terms*— Spin injection modeling, spin lifetime modeling, valley splitting modeling**

I. Introduction

Miniaturization of CMOS devices has made possible a tremendous increase in performance, speed, and density of modern integrated circuits. However, difficulties to reduce the supply voltage V_{DD} result in an approximately constant power dissipation per a single MOSFET. This leads to a rapid increase of generated heat with increasing transistor density, which results in a saturation of MOSFET miniaturization and puts limitations on the performance of integrated circuits. Therefore, research for finding alternative technologies and computational principles becomes urgently needed.

The MOSFET operation is fundamentally based on the charge degree of freedom of an electron. Another intrinsic electron property, the electron spin, attracts at present much attention as a possible candidate for complimenting or even replacing the charge degree of freedom in future electron devices.

Until recently, silicon was remaining aside from the main stream of spin-related applications: even a demonstration of basic elements necessary for spin related applications, such as injection of spin-polarized currents in silicon, spin transport, spin manipulation, and detection, was missing. The first demonstration of coherent spin transport through an undoped 350μm thick silicon wafer [1] has triggered a systematic study of spin transport properties in silicon [2]. The use of silicon for spin driven devices would greatly facilitate their integration with MOSFETs on the same chip.

II. Spin injection

Spin injection in silicon and other semiconductors by purely electrical means from a ferromagnetic metal electrode was not very successful until recently. The fundamental reason has been identified as an impedance mismatch problem [3]. A solution to the impedance mismatch problem is the introduction of a potential barrier between the ferromagnetic metal and the semiconductor [4]. A successful experimental demonstration of a signal which should correspond to spin injection in doped silicon at room temperature was first performed in 2009 [5] using an $Ni_{80}Fe_{20}/Al_2O_3$ tunnel contact. Electrical spin injection through silicon dioxide at temperatures as high as 500K has been reported in [6].

Regardless of a success in demonstrating spin injection at room temperature, there are unsolved challenges which may compromise the results obtained. According to theory, in a three-terminal scheme [2] the value of the voltage signal ΔV due to spin accumulation divided by the current density j flowing through the injecting contact is proportional to

$$\Delta V/j = P^2 \rho_S \sqrt{D_{DIFF}\tau_S}. \quad (1)$$

Because of the injection and detection, the tunnel spin polarization P enters squared, and the silicon resistivity ρ_S multiplied with the spin diffusion length $l = \sqrt{D_{DIFF}\tau_S}$, where τ_S is the spin lifetime, determines the additional area resistance of the contact due to spin accumulation under it. However, there is a several orders of magnitude discrepancy between the signal measured and the theoretical value (1). It turns out that the signal is stronger in three-terminal measurements, while it is weaker in the non-local scheme [2]. The reasons for the discrepancies are heavily debated [7], [8] and it is apparent that more research is needed to resolve this controversy.

A. Spin injection in silicon through a space-charge layer

In a quite recent publication a ten-fold spin injection efficiency increase was predicted [9], which is attributed to electrostatic screening effects. In a conventional approach the presence of a space charge layer at the interface is ignored [10]. When the space charge layer is absent (charge neutrality), analytical expressions for the spin injection efficiency through a ferromagnetic-non-magnetic semiconductor interface can be obtained. The density of states in both materials is considered similar to avoid the impedance mismatch problem. When the charge current J_n flows through the junction, the spin accumulation in the semiconductor appears, which is characterized by the spin current J_s injection efficiency $\alpha = J_s/J_n$ and the spin density s injection efficiency $\beta = s/n$, where the carrier density n is equal to the doping level N_D under the charge neutrality conditions. The analytical expressions at the

This work is supported by the European Research Council through the grant #247056 MOSILSPIN.

ISBN: 978-1-61804-251-4

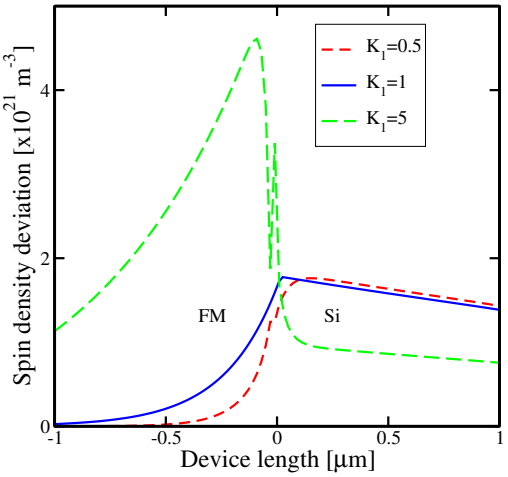


Fig. 1. Spin density distribution, when the charge current density is fixed to $23.4\,\mathrm{MA/m^2}$. $P = 0.2$. K_1 is the doping ratio in the ferromagnet to the non-magnetic material.

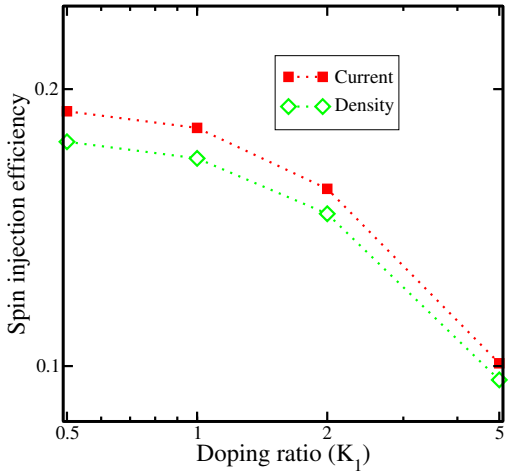


Fig. 2. Spin density and spin current injection efficiencies (α_D, and β_D), taken at the screening length λ_D away from the interface in the non-magnetic material, for $P = 0.2$.

Si interface for α and β are cumbersome. The simplified expressions, valid for small values of P, are written as:

$$\alpha = P\,\frac{l_d}{l_d + (1-P^2)\,l_u} \tag{2}$$

$$\beta = (1 - \frac{l_u}{l_d})\,\alpha \tag{3}$$

where $P = s/n$ is the equilibrium bulk spin polarization in the ferromagnet. The spin diffusion lengths $l_{u(d)}$ against (along) the electric field E are [10]

$$l_{u(d)} = \left(-(+)\frac{|eE|}{2k_BT} + \sqrt{\frac{|eE|}{2k_BT} + \frac{1}{l}} \right)^{-1}, \tag{4}$$

For the chosen current direction from the non-magnetic to ferromagnetic semiconductor, l_d is the spin diffusion length in the ferromagnet, while l_u is the length in the non-magnetic semiconductor. For simplicity the intrinsic spin diffusion length l is taken the same on both sides of the junction.

To violate the charge neutrality and introduce the space charge layer at the ferromagnetic-non-magnetic semiconductor interface, we modify the carrier concentration in the ferromagnet by assuming the ferromagnet to be doped to a concentration proportionally with a factor K_1 to the doping value N_D in the semiconductor. Thus, when K_1=1, the charge neutrality condition is recovered, while a charge accumulation and a charge depletion are introduced when K_1>1 and K_1<1, respectively, at the non-magnetic side of the junction. We investigate the carrier distribution and the spin current variation along the junction considering a fixed charge current density J_n=$23.4\,\mathrm{MA/m^2}$. The spin density s and the spin current J_s behave differently at the interface and in the bulk. When K_1>1 (K_1<1), s gradually piles up (drops down) in the bulk of the ferromagnet and drops down (piles up) in the bulk of the non-magnetic semiconductor (Fig.1), compared to the charge neutrality condition. This phenomenon happens due to the difference in the material conductivity proportional to the doping concentration, and the bulk electric field, which eventually modifies the effective spin diffusion length. On the contrary, when K_1>1 (K_1<1), s develops a dip (peak) at the ferromagnetic interface followed by a sharp peak (dip) at the non-magnetic semiconductor interface (Fig.1). These features are correlated with the charge depletion (accumulation) at the ferromagnetic/non-magnetic interface, which results in the formation of a potential profile with a barrier for electrons. These interface effects give rise to an alteration in the spin current at the interface, however persisting only up to the charge screening length (λ_D. The spin injection efficiencies at a distance λ_D away from the interface in the non-magnetic semiconductor displayed in Fig.2 shows an increment in both α_D, and β_D, compared to the charge neutrality case K_1 =1, if the spins are injected into a non-mgnetic material from the ferromagnet with doping level lower than in the non-magnetic material. However, its value is always limited by the bulk spin polarization of the ferromagnetic contact. Under similar conditions, the spin injection efficiency in the non-magnetic semiconductor bulk decreases, if the spins are injected from a highly doped ferromagnetic source.

III. Modeling Spin Relaxation

For a spin-based device the possibility to transfer the excess spin injected from the source to the drain electrode is essential. The excess spin is not a conserved quantity, in contrast to charge. While diffusing, it gradually relaxes to its equilibrium value which is zero in a non-magnetic semiconductor. In a ground breaking experiment it was demonstrated that spin can propagate through a 350μm silicon wafer at liquid nitrogen temperatures. A lower estimation for the spin lifetime at room temperature obtained within the three-terminal injection scheme was of the order 0.1-1ns [2]. This corresponds to an intrinsic spin diffusion length l=0.2-0.5μm. The spin lifetime is determined by the spin-flip processes. Several important spin relaxation mechanisms are identified [11], [12]. In silicon the spin relaxation due to the hyperfine interaction of spins with the magnetic moments of the ^{29}Si nuclei is important at low temperature. Because of the inversion symmetry in the

ISBN: 978-1-61804-251-4

silicon lattice the Dyakonov-Perel spin relaxation mechanism is absent in bulk systems [11], [12]. At elevated temperatures the spin relaxation due to the Elliot-Yafet mechanism [11], [12] becomes important.

The Elliot-Yafet mechanism is mediated by the intrinsic interaction between the orbital motion of an electron and its spin. Due to the spin dependence, the microscopic spin-orbit interaction does not conserve the electron spin, thus it generates spin flips, which is the Yafet process. When the microscopic spin-orbit interaction is taken into account, the Bloch function with a fixed spin projection is not an eigenfunction of the total Hamiltonian. Because the eigenfunction always contains a contribution with an opposite spin projection, even spin-independent scattering with phonons generates a small probability of spin flips, which is the Elliot process.

In order to analyze the spin relaxation in silicon, both, the Elliot and the Yafet processes must be taken on equal footing. In this way a good agreement between the experimentally observed and calculated spin life time as a function of temperature has been achieved confirming that in bulk silicon the Elliot-Yafet mechanism is the dominant spin relaxation mechanism at ambient temperatures [13]. The spin lifetime in undoped silicon at room temperature is about 10ns, which corresponds to a spin diffusion length of 2μm. In case of heavily doped silicon the spin lifetime is determined by the Elliot-Yafet mechanism due to ionized impurity scattering and is expected to be around 1ns at $N_D = 10^{19}\text{cm}^{-3}$, in agreement with experiments.

The main contribution to spin relaxation was identified to be due to optical phonon scattering between the valleys residing at different crystallographic axis, or f-phonons scattering [14], [15]. This scattering is enhanced at high electric field due to the accelerated f-phonon emission process to counteract a further deviation of the electron system from thermal equilibrium [16], which results in an unusual experimentally observed behavior, when the reduction of the carrier transition time between the injector and the collector is accompanied by a reduction in spin polarization.

The relatively large spin relaxation experimentally observed in electrically-gated lateral-channel silicon structures [17], [18] indicates that the extrinsic interface induced spin relaxation mechanism becomes important. This may pose an obstacle in realizing spin driven CMOS compatible devices, and a deeper understanding of fundamental spin relaxation mechanisms in silicon inversion layers, thin films, and fins is needed.

The theory of spin relaxation must account for the most relevant scattering mechanisms which are due to electron-phonon interaction and surface roughness scattering. In order to evaluate the corresponding scattering matrix elements, the wave functions must be provided.

To find the wave functions, we employ the Hamiltonian describing the valley pairs along the [001]-axis [19]. The Hamiltonian includes confinement, a spin-orbit effective interaction term with the effective constant Δ_{so}, and shear strain ε_{xy} entering with the deformation potential D_{xy}. It is possible to accurately describe the valley bulk dispersion in the presence of strain including shear strain dependent effective masses [20].

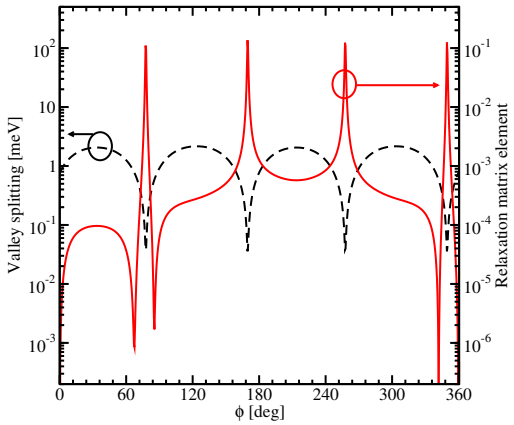


Fig. 3. Dependence of the normalized spin relaxation matrix elements and valley splitting on the angle between the incident and scattered waves for a quantum well of 4nm thickness, k_x=0.5nm^{-1}, k_y=0.1nm^{-1}, ε_{xy}=0.01%.

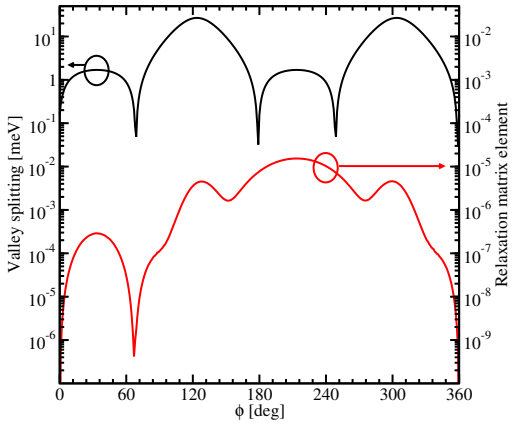


Fig. 4. Dependence of the normalized spin relaxation matrix elements and valley splitting on the angle between the incident and scattered waves for a quantum well of 4nm thickness, k_x=0.5nm^{-1}, k_y=0.1nm^{-1}, ε_{xy}=0.92%.

The Hamiltonian accounts for the unprimed subband (valley) splitting. In confined silicon systems it is usually assumed that the unprimed subbands, because they are originating from the two equivalent [001] valleys, are double degenerate. However, this is true only in the parabolic band approximation when the two valleys are independent. Due to the presence of the off-diagonal terms in the Hamiltonian, the [001] valleys are coupled, which results in an unprimed subband degeneracy lifting. In the case when the confinement potential is approximated with an infinite square well, the difference between the unprimed subband energies is as [19]

$$\Delta E = \frac{2y^2\sqrt{\Delta_{\text{so}}^2\left(k_{\text{x}}^2+k_{\text{y}}^2\right)+\left(D_{xy}\varepsilon_{\text{xy}}-\frac{\hbar^2 k_{\text{x}}k_{\text{y}}}{M}\right)^2}}{k_0 t\sqrt{\left(1-y^2-\eta^2\right)\left(1-y^2\right)}} \times\left|\sin\left(\sqrt{\frac{1-y^2-\eta^2}{1-y^2}}k_0 t\right)\right|, \tag{5}$$

with $y=\frac{\pi}{k_0 t}$, $\eta=\frac{m_1 B}{k_0^2\hbar^2}$, t is the film thickness, and $k_0 = 0.15(2\pi/a)$ is the position of the valley minimum relative to the X-point.

The minimum of the $\sqrt{...}$ term in (5) reveals a very strong increase of the intersubband spin relaxation shown in Fig.3.

ISBN: 978-1-61804-251-4

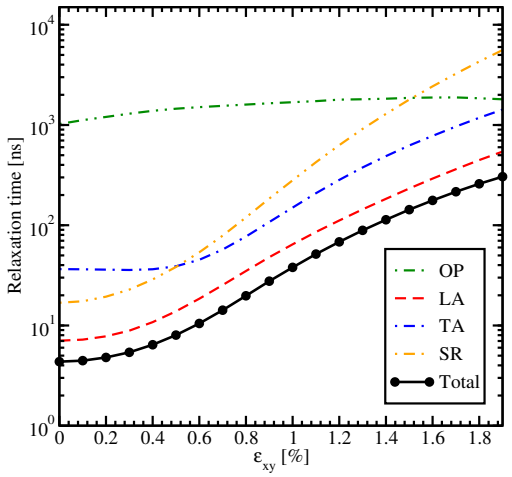


Fig. 5. Dependence of spin lifetime on shear strain for T=300K and a film of 4nm thickness. Optical (OP), longitudinal (LA) and transversal (TA) acoustic phonon, and surface roughness spin relaxation contributions are also shown.

Under these conditions the subband splitting is purely determined by the effective spin-orbit interaction term and is linear in $\Delta_{SO}\sqrt{k_x^2+k_y^2}$, where k_x, k_y are the components of the in-plane electron wave vector. This linear dependence of the splitting is similar to the Zeeman splitting in a magnetic field. Thus, the spin-orbit interaction term $\Delta_{SO}\mathbf{k}$ with $\mathbf{k}=(k_x,-k_y)$ can be interpreted as an effective magnetic field, while the pairs of states $(X_1,\uparrow),(X_{2'},\downarrow)$ and $(X_{2'},\uparrow),(X_1,\downarrow)$ it couples have similarities with the Zeeman spin-up, spin-down states split because of the effective field. Spin along the z-direction starts precessing in the in-plane effective field $\Delta_{SO}\mathbf{k}$, which results in a large mixing between the opposite spin states from the different valleys. This mixing results in large spin relaxation matrix elements defining hot spin relaxation spots seen in Fig.3. The origin of the spin relaxation hot spots in thin films lies in the unprimed subband degeneracy in a confined electron system. Because the hot spots are determined by the minimum of the $\sqrt{...}$ prefactor, they are located in the middle of the two-dimensional Brillouin zone in an unstrained film, thus contributing strongly to the spin relaxation.

When shear strain is applied, the spin relaxation hot spots are pushed towards higher energies and do not contribute significantly to spin relaxation. The minimum splitting between the subbands seen in Fig.4 does not result in any peculiarities of the spin relaxation matrix elements (Fig.4). We have checked that the valley splitting at the minima shown in Fig.4 is exactly zero. Thus the degeneracy between the subbands at these points is precisely recovered due to the oscillating $\sin\left(\sqrt{\frac{1-y^2-\eta^2}{1-y^2}}k_0t\right)$ term. However, this degeneracy is insignificant, because it does not result in any peculiar behavior of the spin relaxation scattering matrix elements.

Moving the hot spots above the Fermi energy outside the occupied states region results in a sharp reduction of spin relaxation and in an increase of the spin lifetime with shear strain. Fig.5 demonstrates an order of magnitude enhancement of the spin lifetime at the stress values comparable achieved in advanced MOSFETs for boosting the electron mobility. Therefore, shear strain now routinely used to enhance the performance of modern MOSFETs can also be used to influence the spin propagation in the channel by enhancing the spin lifetime and the spin diffusion length significantly.

IV. Summary and Conclusion

Recent ground-breaking experimental and theoretical findings regarding spin injection and transport in silicon make spin an attractive option to supplement or to replace the charge degree of freedom for computations. The large discrepancy between the spin injection signal observed and predicted cannot be attributed to space charge effects. Mechanical stress routinely used to enhance the electron mobility can also be used to boost the spin lifetime.

Acknowledgment

This work is supported by the European Research Council through the grant #247056 MOSILSPIN.

References

[1] B. Huang, D. J. Monsma, I. Appelbaum, Coherent Spin Transport through a 350 Micron Thick Silicon Wafer, Phys. Rev. Lett. 99 (2007) 177209.
[2] R. Jansen, Silicon Spintronics, Nature Materials 11 (2012) 400–408.
[3] G. Schmidt, D. Ferrand, L. W. Molenkamp, A. T. Filip, B. J. van Wees, Fundamental Obstacle for Electrical Spin Injection from a Ferromagnetic Metal into a Diffusive Semiconductor, Phys. Rev. B 62 (2000) R4790–R4793.
[4] E. I. Rashba, Theory of Electrical Spin Injection: Tunnel Contacts as a Solution of the Conductivity Mismatch Problem, Phys. Rev. B 62 (2000) R16267–R16270.
[5] S. P. Dash, S. Sharma, R. S. Patel, M. P. de Jong, R. Jansen, Electrical Creation of Spin Polarization in Silicon at Room Temperature, Nature 462 (2009) 491–494.
[6] C. Li, O. van 't Erve, B. Jonker, Electrical Injection and Detection of Spin Accumulation in Silicon at 500K with Magnetic Metal/Silicon Dioxide Contacts, Nature Communications 2 (2011) 245.
[7] R. Jansen, A.M. Deac, H. Saito, S. Yuasa, Injection and Detection of Spin in a Semiconductor by Tunneling via Interface States, Phys. Rev. B 85 (2012) 134420.
[8] Y. Song, H. Dery, A. Lemaitre, Magnetic-Field-Modulated Resonant Tunneling in Ferromagnetic-Insulator-Nonmagnetic Junctions, Phys. Rev. Lett. 113 (2014) 047205.
[9] M.R. Sears, and W.M. Saslow, Spin Accumulation at Ferromagnet/Nonmagnetic Material Interfaces, Phys. Rev. B 85 (2012) 014404.
[10] Z.G. Yu, and M.E. Flatte, Spin Diffusion and Injection in Semiconductor Structures: Electric Field effects, Phys. Rev. B 66 (2002) 235302.
[11] J. Fabian, A. Matos-Abiaguea, Ch. Ertlera, P. Stano, I. Zutic, Spintronics: Fundamentals and Applications, Rev. Mod. Phys. 76 (2004) 323–410.
[12] I. Zutic, J. Fabian, S. Das Sarma, Semiconductor Spintronics, Acta Phys. Slovaca 57 (2007) 567–907.
[13] J. L. Cheng, M. W. Wu, J. Fabian, Theory of the Spin Relaxation of Conduction Electrons in Silicon, Phys. Rev. Lett. 104 (2010) 016601.
[14] P. Li, H. Dery, Spin-Orbit Symmetries of Conduction Electrons in Silicon, Phys. Rev. Lett. 107 (2011) 107203.
[15] Y. Song, H. Dery, Analysis of Phonon-Induced Spin Relaxation Processes in Silicon, Phys. Rev. B 86 (2012) 085201.
[16] J. Li, L. Qing, H. Dery, I. Appelbaum, Field-Induced Negative Differential Spin Lifetime in Silicon, Phys. Rev. Lett. 108 (2012) 157201.
[17] J. Li, I. Appelbaum, Modeling Spin Transport in Electrostatically-Gated Lateral-Channel Silicon Devices: Role of Interfacial Spin Relaxation, Phys. Rev. B 84 (2011) 165318.
[18] J. Li, I. Appelbaum, Lateral Spin Transport through Bulk Silicon, Appl. Phys. Lett. 100 (16) (2012) 4704802.
[19] D. Osintsev, O. Baumgartner, Z. Stanojevic, V. Sverdlov, S. Selberherr, Subband Splitting and Surface Roughness Induced Spin Relaxation in (001) Silicon SOI MOSFETs, Solid-State Electronics 90 (2013) 34 – 38.
[20] V. Sverdlov, Strain-Induced Effects in Advanced MOSFETs, Springer, Wien - New York, 2011.

A simulation based decision-making support approach for foundry plants investment projects estimation of efficiency

Mikhail V. Zenkovich and Yury G. Drevs

Abstract— Methods and software enabling the estimation of efficiency and the comparisons of alternative designs of foundry plants on the basis of moulding lines are discussed. Problem of estimation of efficiency of investment projects of foundry plants is formulated in the terms of decision theory. Presented approach is based on the reduction of multicriterion problem of estimation of investment project to one-criterion problem. Application of simulation for estimation of technological and structural decisions, which was made during the plant design, is the central feature of presented approach. Model of moulding line refers to discrete-event class. Object-oriented approach was applied for designing of the model and programming language C++ for its implementation. Application of detailed simulation model of moulding line allows carrying out an accurate estimation of technological and structural characteristics of involved projects.

Keywords— decision-making support, simulation, investment projects estimation of efficiency, moulding line.

I. Introduction

In this paper methods and software enabling the estimation of efficiency and the comparisons of alternative designs of foundry plants on the basis of moulding lines are discussed. Estimation of efficiency is caring out with respect to specific of produced castings, current market situation and individual preferences of decision-makers. Estimation of efficiency could be conducted as for one individual project, as for group consisted of several alternative projects. In case of several alternative projects the most preferable project is chosen. As results of the estimation the following decisions could be made: if values of all characteristics of the best project are satisfying for decision-makers than follows decision of this project implementation, otherwise "bottlenecks" of the project are analyzed, some corrections implemented and the procedure of estimation for this project is repeated.

In general investment project P could be presented by the following model [1]:

$P = \{IC_j, CF_k, p, r\}$,

where: IC_j – investments in the year j, $j = 1, 2,\ldots, q$, $q \le p$; CF_k – cash flow in the year k, $k = 1, 2, \ldots, p$; p – project's length (period of time for the implementation of the project); r – discount rate.

For the efficiency estimation of such projects usually the following criterions are used [1], [2]: *Net Present Value* (*NPV*), *Profitability Index* (*PI*), *Internal Rate of Return* (*IRR*), *Payback Period* (*PP*), *Discounted Payback Period* (*DPP*), *Accounting Rate of Return* (*ARR*) and *Modified Internal Rate of Return* (*MIRR*).

II. Decision-making problem definition

Decision-making problem can be formulated conceptually in a following way: there is a set of decision variants (alternatives), every alternative realization leads to some event (outcome) each outcome is characterized by a set of vectorial estimations. It is needed after studying all decision-maker's preference to design a model of alternative choice better in some specific sense.

Decision-making problem can be described formally by the following tuple [3]:

$\langle A, \Omega, E, F, P_s, D, T\rangle$,

where A – a set of admissible alternatives Ω – a set of outcomes of admissible alternatives, E – a set of vectorial estimations of outcomes, F – mapping of a set Ω to a set E, $F: \Omega \to E$; P_s – structure of decision-maker's preferences.

It is necessary to find some decision rules or algorithm D to provide needed action T on a set of alternatives A: to select a set of non-dominating alternatives, to find the most preferable alternative, to produce linear ordering of admissible alternatives and etc.

Needed action T: on a set of alternatives A characterizes the type of decision-making problem (choice, ordering and etc). Environment and a system of preferences are granted with elements Ω, E, F, P_s, D. Single result (deterministic or random) which is characterized with vector estimation corresponds to each alternative. The system of preferences is described by some total combination of sets (criterions, alternatives, results, for example) with preferences relations and is some empirical system with relations. Structural representation of decision-maker's preferences as a system with relations will be named decision-maker's preferences structure. This structure defines the procedure of estimations comparison $e(\omega)$ and the decision rule or algorithm – the principle of elements choice from set A on the basis of comparison results in conformity with required action T.

In the considered problem elements of the tuple above are [5]–[7]:

ISBN: 978-1-61804-251-4

1. The set of admissible alternatives outcomes Ω.

Outcome $\omega \in \Omega$, corresponding to alternative $a \in A$ is characterized with the vector of following type [4]–[7]:

$$\omega = \begin{pmatrix} \omega_1, \ldots, \omega_{ec}, \omega_{ec+1}, \ldots, \omega_{con+ec}, \omega_{con+ec+1}, \ldots, \\ \omega_{tec+con+ec}, \omega_{tec+con+ec+1}, \ldots, \omega_{pro+tec+con+ec} \end{pmatrix},$$

where $\omega_1, \ldots, \omega_{ec}$ – components which are describing economic parameters of project (costs for castings, raw materials, energy and etc); *ec* – is a number of components describing economic parameters of project; $\omega_{ec+1}, \ldots, \omega_{con+ec}$ – components which are describing structural parameters of project (a number of continuous-handling systems for cooling, devices for transporting of moulds and etc); *con* – is a number of components describing structural parameters of project; $\omega_{con+ec+1}, \ldots, \omega_{tec+con+ec}$ – components which are describing technological parameters of project (number of moulding sand components, recommended values of technological characteristics for all issued casting types and etc); $\omega_{tec+con+ec+1}, \ldots, \omega_{pro+tec+con+ec}$ – are components describing parameters which characterize line throughput (a number of definite type good castings produced on moulding line during a year; capacity factors for equipment in production sites of a line and etc); *pro* – is a number of components describing line throughput.

2. Mapping F: $\Omega \to E$ is following vector function [5]–[7]:

$$F(\omega) = \begin{pmatrix} NPV(\omega), PI(\omega), IRR(\omega), PP(\omega), \\ DPP(\omega), ARR(\omega), MIRR(\omega) \end{pmatrix},$$

where *NPV*(ω) – is a function of the criterion Net Present Value, *PI*(ω) – is a function of the criterion Profitability Index, *IRR*(ω) – is a function of the criterion Internal Rate of Return, *PP*(ω) – is a function of the criterion Payback Period, *DPP*(ω) – is a function of the criterion Discounted Payback Period, *ARR*(ω) – is a function of the criterion Accounting Rate of Return and *MIRR*(ω) – is a function of the criterion Modified Internal Rate of Return.

3. A set of vectorial estimations of outcomes *E*. Set elements are vectors $e(\omega) \in E$, which components values correspond to criterions values (*NPV, PI, IRR, PP, DPP, ARR* and *MIRR*), calculated for the corresponding outcomes [5]–[7].

4. Needed action *T* over a set of admissible alternatives *A*. It is necessary to find the most preferable alternative $a^* \in A$ [5]–[7].

5. Decision rule *D* [5]–[7]. It is necessary to find such alternative $a^* \in A$, for which corresponding outcome $\omega^* \in \Omega$, ensures the maximum meaning of efficiency function:

$$U(\omega) = \sum_{i=1}^{7} \rho_i \cdot UN_i(F_i(\omega)),$$

where *U*(ω) – is project (alternative) efficiency $a \in A$ corresponding to the outcome $\omega \in \Omega$;

$\rho_1, \ldots, \rho_7$ – are weight coefficients reflecting the relative impotence of corresponding criterions values. They are assigned processing from individual decision-maker's preferences reflecting his preferences structure P_s,

$$\rho = \{\rho_i\} = \left\{ \rho_i : \rho_i \geq 0, i = 1,\ldots,7, \sum_{i=1}^{7} \rho_i = 1 \right\}.$$

Criterion function *NPV*(ω), *PI*(ω), *IRR*(ω), *ARR*(ω) and *MIRR*(ω) are maximized and *PP*(ω) and *DPP*(ω) are minimized. To maximize the value of selected efficiency function *U*(ω) it is necessary to have all criterion functions maximized. That is why it is necessary to change the purpose direction (replacement «min» to «max») for criterions *PP*(ω) and *DPP*(ω). For this we use the following transformations: $F_4(\omega) = -PP(\omega)$ and $F_5(\omega) = -DPP(\omega)$.

Now it is necessary to conduct the procedure of criterions normalization and ranking because we propose using multicriteria choice of economically rational investment project of foundry plant, but criterions chosen for its evaluation have different dimensions. The given procedure means taking criterions to none-dimensional view with the help of certain transformation. That transformation has to satisfy the following qualities: 1) to have the mutual beginning of counting out and single change values order for the whole set of admissible alternatives; 2) to be monotonous (that is to say this transformation has to keep preference relation for whole set of admissible alternatives).

$UN_i(F_i(\omega)), i = 1,\ldots,7, \omega \in \Omega$ – are monotonous functions transporting every criterion function $F_i(\omega), i = 1,\ldots,7, \omega \in \Omega$ to normalized (non-dimensional) view, $F_1(\omega) = NPV(\omega)$; $F_2(\omega) = PI(\omega)$; $F_3(\omega) = IRR(\omega)$; $F_4(\omega) = -PP(\omega)$; $F_5(\omega) = -DPP(\omega)$; $F_6(\omega) = ARR(\omega)$; $F_7(\omega) = MIRR(\omega)$.

For criterion normalization let us use the procedure of full normalization:

$$UN_i(F_i(\omega)) = \frac{F_i(\omega) - F_i^{\min}}{F_i^{\max} - F_i^{\min}}, \; i = 1,\ldots,7, \; \omega \in \Omega,$$

where $F_i^{\min}$ and $F_i^{\max}$ – the least and the greatest (correspondingly) criterion function value $F_i(\omega)$ at the set of admissible alternatives results Ω. This normalization reflects initial criterion values to a segment [0, 1]. The best value of normalized criterion equals 1, the worst one equals 0.

6. The relation of preference.

Let us consider that the alternative a_1 is more preferable than alternative a_2 ($a_1 \succ a_2$) if for corresponding outcomes ω_1 and $\omega_2 \in \Omega$ the following inequality is true: $U(\omega_1) > U(\omega_2)$. In case $U(\omega_1) = U(\omega_2)$ we consider alternative a_1 and a_2 are equal or equivalent ($a_1 \sim a_2$).

III. Criterions Computation and Variant Generation

There are two most widespread approaches to the computation of mentioned above criterions (*NPV, PI, IRR, PP, DPP, ARR* and *MIRR*) of investment projects evaluation [1], [2], [8]–[10]: deterministic and stochastic (related upon statistical tests method). When using the deterministic approach the values of all cash flow parameters sets on the bases of experts' estimations. When we use stochastic

approach we can divide these parameters into two groups: 1) meanings of those arranged by decision-maker personally and 2) random values for which decision-maker sets only intervals of change, random distribution types and parameters reflecting (according to decision-maker's opinion) certain regularity of given parameter value change. Thus in general view correlation for *NPV* criterion computation will be as follows (it is possible to produce correlation for computation other criterions in the same way):

$NPV = f(\chi_1, ..., \chi_i, ..., \chi_l, \xi_1, ..., \xi_j, ..., \xi_s)$,

where χ_i – are stochastic parameters (components of cash flow; they are random values); l – is a number of stochastic parameters; ξ_j – are deterministic parameters (components of cash flow which after analysis were defined as independent values or weakly depending on environment and so will be considered as deterministic values); s – is a number of deterministic parameters. Then with the help of special software statistical modeling is provided and on this basis the valuations of criterions sought values are obtained.

Essential shortage of above approaches is great estimation result dependence on decision-maker's opinion: all deterministic parameters values, intervals, types and random distributions characteristics for stochastic parameters are fixed by decision-maker on a subjunctive basis. One of the ways out of this situation is using simulation model for throughput parameters values estimations of moulding line project under consideration [4]–[7]. When using this approach decision-maker sets the values of economic parameters on the basis of experts' estimations. Values of structural parameters are set in accordance with technological regulations, cards and expert's evaluations. Values of structural and technological parameters influence throughput parameters values. Simulation model of moulding line allows estimating throughput parameters values of considering moulding line project changing technological and structural parameters.

Quantity of good castings producing in a year is the main parameter among all moulding line throughput parameters. With market requirements and price this parameter influences very much on income value of production realization in a year. In its turn realization production income for the year, summary production costs for the year and profit tax pay in a year are the main parameters which are taking in account when computation yearly cash flow (CF_k) is taking place. Annual yearly cash flows depending on investment project under realization are taken in account when criterions *NPV, PI, IRR, PP, DPP, ARR* and *MIRR* are calculated.

We shell name casting as a good one if all values of its technological characteristics are in the certain limits [4]–[6]. Let's name the technological characteristics of casting: 1) time from semimould production till mould assembly; 2) time from mould assembly till its casting; 3) metal temperature when mould was cast; 4) duration of casting cooling in a mould; 5) duration of casting cooling after its shaking-out; 6) content of bentonite and 7) content of a special technological addition in moulding sand which this mould was produced from. Structural features of specific moulding line, equipment stoppage in the moulding line, staff qualification and some other factors influence values of those parameters. We shall consider a casting bad even if only one of its characteristics will be out of permissible meanings.

IV. Main principles of simulation model design

The theory of aggregative system [11] has been chosen for formal description of moulding line. In this approach state of each unit is described by a vector which components are time functions. Time dependence can be continuous (casting temperature, for example) and discrete (positions in continuous-handling system, for example).

Let us consider moulding line as an aggregative system consisting of four aggregative subsystems. They are corresponding to production sites of moulding line (casting and cooling, shaking-out and cooling after this, moulding sand preparation and moulding). In its turn each aggregative subsystem consists of limited number of aggregates describing equipment included in the production sites of given subsystem. Each aggregate in any aggregative system can be classified from one of the following groups [4]: 1) transporting device: device for transporting semimoulds, moulds and castings; 2) continuous-handling system: system for semimoulds, casting, cooling and castings cooling after shaking-out; 3) device for making object: moulding machine and device for assembly of moulds; 4) mixer/bunker for moulding sand: mixer and bunker for moulding sand; 5) casting machine with flooding scoop; 6) device for object disassembling: shaking-out device and device for flask disassembling; 7) belt feed conveyor. Algorithm for presentation of aggregates belonging to each groups is being made on the base of general aggregate model which describes common features for all aggregates in this group features.

Simulation model of moulding line is built on the basis of four autonomous models of production sites [12]. Models of all moulding line production sites consist of two modules: structural module and algorithms of its elements interactions. Common modules of moulding lines models are modeling monitor and user's interface. Discrete-events method was used for model design. The mechanism of time advancement with a constant step was used as a principle of time changing. Object-oriented approach was used for design and language C++ was used for model implementation.

All model elements were described as classes (in C++ notation). The library of these classes was designed and this permits easily to add new elements into the model. All library elements are the heirs of basic class or the heirs of basic class heirs. Heirs of basic class are classes describing groups of devices (specified above) such as transporting device, continuous-handling system and etc. Heirs of classes describing groups of elements are classes describing devices of moulding line (such as devices for mould transporting, continuous-handling system for cooling and etc). Description of model elements interaction algorithm was based on conditionally-events principle. Such approach to the

ISBN: 978-1-61804-251-4

implementation of control mechanism permits to easily modify the function system algorithm and to model any non-permanent situation. New elements could be integrated into structural part of the model without changing of already existed function algorithm [4], [12].

V. Example of Projects Estimation

Let's take a good look at the implementation of discussed methodology on the following example. Two alternative projects of foundry plants on the basis of moulding lines are estimated. Let's mark them – Project 1 and Project 2. So in this case A – set of admissible alternatives consists of two elements a_1 and a_2. Range of produced castings for both two projects is the same. The main differences of these two projects are specifications of casting and cooling site of moulding line design, the amount of initial investments and costs of castings manufacture. The amount of initial investments and costs of castings manufacture for Project 1 are higher (in compare to Project 2). Estimation of these projects conducted by described above methodology gave the following results:

	Project 1	*Project 2*
NPV	4,22 mil. USD	3,018 mil. USD
PI	2,34	1,97
IRR	25,1%	23,5%
PP	2 years	2,2 years
DPP	2,5 years	2,8 years
ARR	55,3 %	52,8 %
MIRR	23,1 %	21,7 %
U(ω)	0,83	-0,16

It is evident from the presented data that variant a_1 is more preferable than variant a_2 ($a_1 \succ a_2$), because for corresponding to them outcomes ω_1 and $\omega_2 \in \Omega$ the following inequality is true: $U(\omega_1) > U(\omega_2)$.

Values of all criterions and value of efficiency function in case of Project 1 implementation are more preferable than values of the same criterions and efficiency function in case of Project 2 implementation. Values deterioration of all criterions and efficiency function in case of Project 2 implementation were analyzed. It was reveled that this significant deterioration was conditioned by specifications of casting and cooling site of moulding line design. On the projected moulding line it is supposed to produce castings for which permissible meaning of technological characteristic "duration of casting cooling in a mould" is above 3 hours. On the Figures 1 are presented histograms of durations of casting cooling in a mould distributions for Project 1 and Project 2 respectively.

In case of the implementation of Project 1 duration of casting cooling in a mould for all moulds would be above 3 hours. In case of the implementation of Project 2 duration of casting cooling in a mould for 14.44% of moulds would be less than 3 hours. Because technological characteristic "duration of casting cooling in a mould" for this castings is out of permissible meanings we consider this castings as wasted. Decreasing of produced good casting amount leads to values deteriorations of all criterions and efficiency function in case of Project 2 implementation.

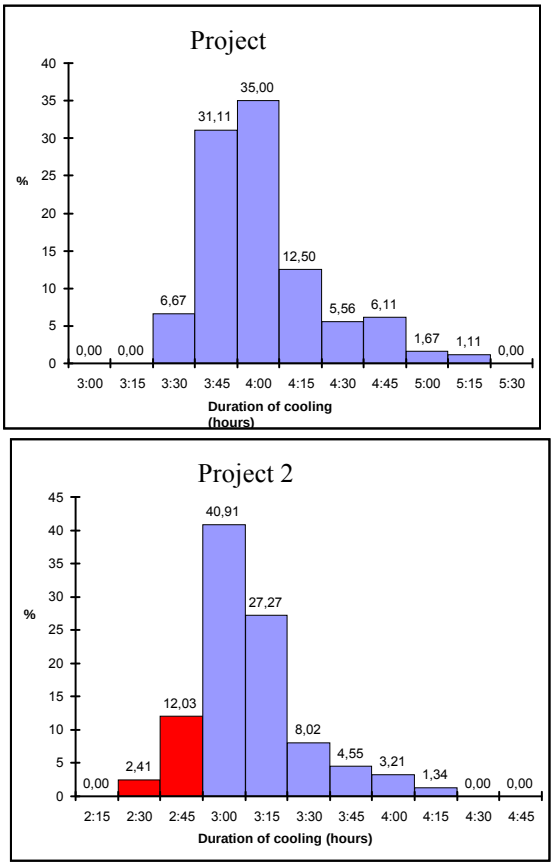


Fig. 1. Durations of casting cooling in a mould distribution.

From this example you can see that adopted on the earlier stages of project implementation construction concept could lead to production of considerable amount of wasted castings, which in turn leads to values deteriorations of all criterions and efficiency function. In spite of the fact that on the first account it was supposed that this construction concept could allow shortening expanses significantly, without any negative effect. Discussed above traditional methods of investment projects estimation (deterministic and stochastic) aren't permit to take into account structural and technological parameters of project. Because production efficiency significantly depends from values of these parameters it is better to use different approaches for estimation of such kind of projects. Presented in this paper approach is based on the application of simulation model of moulding line for the estimation of structural and technological parameters of the considered project. Application of this approach improves decision-making efficiency, especially on the earlier stages of project implementation.

ISBN: 978-1-61804-251-4

VI. Conclusion

Presented evaluation method for investment projects of foundry plants on the base of moulding lines has proved effective on designing and engineering stages of project's implementation. Problem of estimation of efficiency of investment projects of foundry plants on the basis of moulding lines is formulated in the terms of decision theory. Presented approach is based on the reduction of multicriterion problem of estimation of investment project to one-criterion problem. This paper describes: the structure of set of outcomes of admissible alternatives, set of vectorial estimations of outcomes, mapping of set of outcomes of acceptable alternatives to set of vectorial estimations of outcomes and structure of decision maker's preferences. Decision rule which allows to carry out required operation over the set of admissible alternatives is formulated. Application of simulation for estimation of technological and structural decisions, which was made during the plant design, is the central feature of presented approach. Model of moulding line refers to discrete-event class. Object-oriented approach was applied for designing of the model and programming language C++ for its implementation. Application of detailed simulation model of moulding line allows carrying out an accurate estimation of technological and structural characteristics of involved projects. Presented methodology of estimation of investment projects of foundry plants on the basis of moulding lines is tried-and-true method which applies on the phase of designing and engineering of foundry plant.

We have successful results of using considered method and never the less we have some plans for its improvement. Now the thorough revision of moulding line simulation model is made in accordance with agent modeling principles. Agent technologies are connected with the concept of intellectual agent as some intellectual robot (active element) purposely interacting with other such elements and environment under taking conditions. There are a lot of successful examples of implementation of agent-based simulation models of different production systems [13], [14]. It is very impotent for us because moulding line is also a production system.

References

[1] V.V. Kovalev. Introduction to financial management. Moscow: Finance and statistics, 2004.

[2] R.A. Brealey, S.C. Myers. Principles of Corporate Finance. 7th edition. The McGraw−Hill Companies, 2003.

[3] A.N. Borisov, A.V. Alekseev, G.V. Merkureva, N.N. Sliadz, V.I. Glushkov. Processing of fuzzy data in the decision-making support systems. Moscow: Radio and communications, 1989.

[4] M.V. Zenkovich, Y.G. Drevs. Decision-making support of moulding lines design // Automatization in industry. 2010. №11.

[5] M.V. Zenkovich, Y.G. Drevs. Supported decision making in estimation of investment projects of foundry plants // Applied informatics. 2012. №5(41).

[6] M.V. Zenkovich, Y.G. Drevs. Problem of decision-making support in estimation of investment projects of foundry plants on the basis of moulding lines // Software and systems. 2012. №4(100).

[7] M.V. Zenkovich. Estimation of investment projects of foundry plants with application of simulation models // System analysis and information technologies: 15-th International conference SAIT 2013, Kyiv, Ukraine, May 27–31, 2013. Proceedings. – ESC "IASA" NTUU "KPI", 2013.

[8] Jay April, Marco Better, Fred Glover, James Kelly. New advances and applications for marrying simulation and optimization // Proceedings of the 2004 Winter Simulation Conference, 2004, Washington, D.C., USA.

[9] Gerald W. Evans, Suraj M. Alexander. Using multi-criteria modeling and simulation to achieve lean goals // Proceedings of the 2007 Winter Simulation Conference, 2007, Washington, D.C., USA.

[10] Bernard J. Kornfeld, Sami Kara. Project portfolio selection in continuous improvement // International Journal of Operations & Production Management. 2011. Volume 31 issue 10.

[11] N.P. Buslenko. Simulation and modeling of complex systems. Moscow: Nauka. 1978.

[12] Y.G. Drevs., M.V. Zenkovich, A.S. Lubchenko. Simulation of moulding lines // Automatization in industry. 2008. №7.

[13] Barbosa J., Leitao, P. Simulation of multi-agent manufacturing systems using Agent-Based Modelling platforms // 2011 9th IEEE International Conference on Industrial Informatics (INDIN) pp. 477 – 482.

[14] V. Marik, D. McFarlane. Industrial Adoption of Agent-Based Technologies // IEEE Intelligent Systems, 20 (1), 2005, pp. 27-35.

Mikhail V. Zenkovich is a senior researcher at the National Research Nuclear University "MEPhI". Moscow. Russia.

Yury G. Drevs, Doctor of Science (Tech.) is a full professor at the National Research Nuclear University "MEPhI". Moscow. Russia.

Decision-making support tools in data bases to improve the efficiency of inventory management for small businesses

SVETLANA V. SHIROKOVA, OKSANA Y. ILIASHENKO
Department "Information Systems in Economics and Management"
Saint-Petersburg State Polytechnical University
Polytekhnicheskaya str. 29, 195251
RUSSIAN FEDERANION
swchirokov@mail.ru, ioy120878@gmail.com
http://www.isem-fem.spb.ru

Abstract: The article highlights the questions of the use of mathematical methods and control of current balances by databases for solution the problems of inventory management. Also considered the game-theoretic approach to multiobjective optimization for resource allocation. Proposed means of information systems are invariant under the mathematical methods that are used in the planning of inventory management.

Key-Words: inventory management tools, support decision-making, the game-theoretic approach to multiobjective optimization, OLAP tools, economic order quantity model, SQL-instructions.

1. Introduction

Efficient management at all levels will greatly improve efficiency of any company. The area of great importance is an inventory management. Effective inventory management increases "frozen" in stocks asset's turnover. Due to effective inventory management product (necessary resource) will always be available and easy to access. That will prevent loosing possible clients/customers. Information Systems Supply Chain Management (SCM), as well as warehouse management modules (Inventory Management System) contain serious mathematical apparatus designated for inventory management optimization. Small businesses, however, prefer to employ less expensive but less efficient software such as Microsoft Office, compare to specialized information systems. In this article we discuss about the tools of decision support systems based on mathematical methods, game theory and databases.

Inventory management tools (IMT) are important part of the company's management. IMT could be divided in two parts. One of them consists of mathematical models and methods of planning of the optimal size of the stocks, calculation of optimal reserve supply chain, and methods of optimal service in reserve. Another includes an information system, which provides a continuous registration data of the goods accounting, stocks calculation and visualize results for decision-making support.

Decision-making support is an important part of IMT for enterprises of any size. Online analytical processing (OLAP) are key technologies for enterprise decision support systems. They provide sophisticated technologies for data collection, integration, retrieval, and analysis; query optimization, and advanced user interfaces. There are many OLAP tools, which are widely used in this field.

The history of the development of OLAP technologies rather significant [18]. Most OLAP tools were developed for decision support at large enterprises, and require significant funding. Small businesses, most probably, will be interested in the tools based on standard software such as spreadsheet or desktop database manage system.

This article provides tools for the decision support systems based on mathematical methods, game theory and databases.

The article highlights the issues of implementation of decision-making support tools involving methods of database manage systems.

2. Problem Formulation

Effective inventory management should include a comprehensive solution. This solution must include proper software. When it comes to small business, software should be affordable and ease of use. As mentioned above, small businesses prefer to use

Microsoft Office, rather than expensive specialized information systems.

Due to the limited financial capacity, small business often faces necessity for the allocation of resources or reserves, for example, by projects within the company. In general, any mathematical programming concludes selection from a given feasible set of variables that achieves maximum or minimum of the objective function [11]. Here we are talking about single-criterion optimization. However, in practice, often there are cases when there is a diversity of purposes, the extent to which expressed a number of criteria, ie vector. So, when deciding on the allocation of the resource (reserve) between projects or departments, company needs to consider multiple criteria facing the problem of multicriteria (or vector) optimization [2].

The concept of OLAP (On-Line Analytical Processing) was formulated by E. F. Codd in 1993 [5]. The main idea of this system is based on Multidimensional conceptual view, and consists of the construction of multidimensional cubes (multidimensional tables) available by user's request. These multidimensional cubes are based on the source and aggregated data (are you sure about source data, in biology we say "row" data), all of them could be stored in relational as well as multidimensional databases. Software on the base of the OLAP system allows flexible and fast information search, variety of analytical operations including multiple data reports, and data comparison analysis over time. All work with OLAP-system is in terms of the subject area.

Currently a market offers a huge variety of OLAP-systems. There are several classifications of this type of products: for example, classification by type of data storage, by the degree of accessibility. Let's review the first of the above classifications. There are three ways of data storage in OLAP-systems [18]:

- MOLAP (Multidimensional OLAP). This type of storage provides high performance OLAP operations. However multi-dimensional cube most often be redundant and depend on the number of measurements.
- ROLAP (Relational OLAP). The source data are stored in a relational database on the file server. Aggregate data may be placed in special tables in the same database. Converting data from a relational database in multidimensional cubes happens by request of OLAP tools. Based on that the cube building speed will strongly depend on the type of data source, decreasing response time of the system and, sometimes, making the process unacceptably slow.
- HOLAP (Hybrid OLAP). In this type of OLAP original data remains in a relational database and aggregates are placed in a multidimensional. Construction of OLAP cube happens in response of OLAP tools request on the base of relational and multidimensional data. This approach allows you to avoid the explosive growth of data, and achieve optimal runtime of client's request.

There are main advantages of OLAP [19]:

- Subject orientation. That means that information in cubes is collected and stored based on various aspects of business: purchasing, sales, etc. That's distinguishes the OLAP database from the operational database, where data is organized according to different processes such as registration or issue of documents, registration of orders, etc.
- Multi user mode. Client-server architecture of OLAP products provides simultaneous access for a large number of users. Despite the large number of users the analysis is equally fast in all aspects of information regardless of the size and complexity of the database structure.
- Direct access to data allows a user to see all the information at once, not filtered out by the report.
- Concentration of the necessary data in one place. For example, all sale information including calculations, contractor, manager, date, type of transaction, etc. is stored in the same cube, and does not require referral to extra sources (manuals and etc).
- Convenient means of access, view and analyze business information (friendly interface). The user receives an intuitive model of data, organized in the form of multidimensional cubes. This allows comparative analysis of indicators, analysis of different scenarios on the principle of "what-if", based on forecasting and statistics of the company.
- Convenient data navigation with a mouse allows to move through the hierarchy within the dimension (the transition from analysis by year to the analysis by quarter, sales analysis by managers to the analysis of sales by region and so on); as well as to move between dimensions (transition from

ISBN: 978-1-61804-251-4

analysis by the dimension of "Time" to the analysis of "Goods", etc.).

However, OLAP applications are quite expensive even for large enterprises, and also require staff training. Therefore use of such tools for the small enterprises is problematic. Small enterprises widely use universal software like MS Excel. MS Excel tools use OLAP functions and provides the implement for support decision making. There are main advantages of OLAP tools in universal software like MS Excel [6]:

- easy of using;
- no requires special staff training;
- low cost;
- versatility

As a rule, the technology of using OLAP functions of MS Excel provides data export from one DBMS, for example, from SQL server or MS Access. Data table is exported from the database to Excel, following an employment of the OLAP functions in Excel [8]. But the disadvantages can include some complexity data export technologies (two-level technology): preparation of the data in the database and export to Excel. Alternative MS Excel can be a simple solution using general-purpose DBMS.

The database is a system for storing information in a structured form. Database management system (DBMS) is the software that controls access to the database and interact with the user, other applications, and the database itself to capture and analyze data. Database programs enable manipulation of data that is stored to be presented in various ways. Mathematical operations can also be performed on the data stored in these devices. Database management systems allow extremely complex operations to be done on massive amounts of stored data.

A general-purpose DBMS is a software system designed to allow the definition, creation, querying, update, and administration of databases. Well-known DBMSs include MySQL, Microsoft SQL Server, Oracle, LibreOffice Base, Microsoft Access etc.

A typical DBMS has the following features [10]:

- Provides a way to structure data as records, tables, or objects.
- Accepts data input from operators and stores that data for later retrieval.
- Provides query languages for searching, sorting, reporting, and other decision-making support activities that help users correlate and make sense of collected data
- Provides multiuser access to data, along with security features that prevent some users from viewing and/or changing certain types of information.
- Provides data integrity features that prevent more than one user from accessing and changing the same information simultaneously.
- Provides a data dictionary (metadata) that describes the structure of the database, related files, and record information

A general-purpose DBMS use SQL queries as queries for complex data views and decision-making support activities. So, in general-purpose databases, OLAP tools implemented with SQL queries.

The article describes one of the ways of the structural organization of tools for database. That includes requests for specific data grouping, specific SQL-calculation, and graphical interpretation of the results for facilitating decision-making by managers of enterprises. As an example, we use one of the calculations of inventory management from economic order quantity model (EOQ).

3. Problem Solution

3.1. Wilson`s model. Economic order quantity model (EOQ)

There are two types of inventory management models - deterministic and stochastic inventory management models. Among stochastic models produce stable and non-stable models.

Wilson`s model (EOQ - Economic order quantity model) is one of stable models with determined demand [1, 13].

Model with high intensity of completion reserves and big penalties for deficiency.

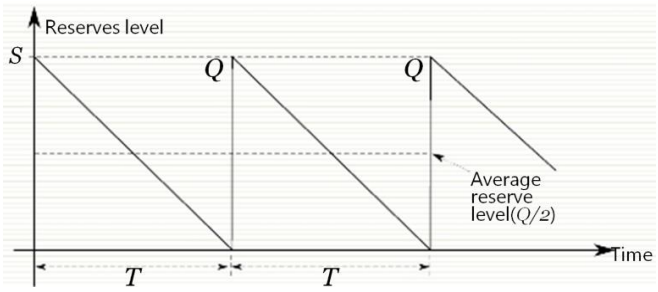


Fig. 1. Model of optimal order quantity

Main points:

- Demand is known;
- Instant receipting of product;
- Discounts aren`t considered;

- Deficit isn`t admitted;
- Resources may be analyzed separately

Total cost (LT), rub = supply costs + storage costs

Total annual cost, rub/year:

D – average consumption of resource, unit/year;

$$L_c = \frac{L_T}{T} = \frac{K+\frac{hTQ}{2}}{T} = \frac{DK}{Q} + \frac{hQ}{2} \to min \qquad (1)$$

K – supply cost, rur/order;
Q – size of supply (order), unit;
T – supply period, year;
h –marginal cost on storage of stocks during the year, rur./(unit ∗ year).
Optimal supply size:

$$Q^* = \sqrt{\frac{2DK}{h}}\ ; \qquad (2)$$

Optimal order period

$$T^* = \frac{Q^*}{D} = \sqrt{\frac{2K}{Dh}}\ ; \qquad (3)$$

Supply quantity:

$$n^* = \sqrt{\frac{2K\mu}{h}}\ ; \qquad (4)$$

K – fixed outgoings, independent of order rub/order;
μ – intensity of resources consumption, unit/t;
h – marginal costs of maintenance, rub/ (unit ∗ t);

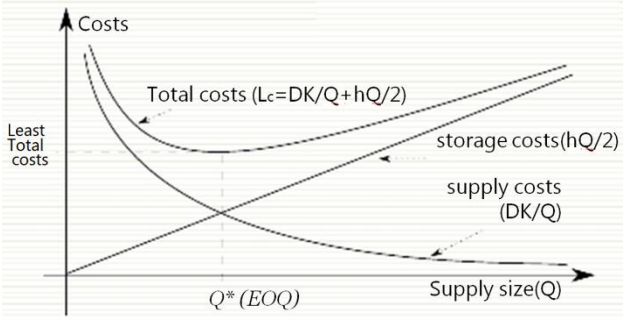


Fig.2. Economic order quantity model

There are many type of Wilson`s models: model of optimal order quantity EOQ, model inventory management with the assumption deficit, the inventory model without an admission of a deficit, etc [11].

3.2. Problem multicriteria (or vector) optimization.

Multiobjective optimization problem is formulated as follows:

Let there be a controlled event (operation), the outcome of which depends on the chosen side A strategy $X \in \mathbf{X} \subset \mathbf{R}^m$. Effectiveness of side A vector valued criterion $e = (e_1, e_2, \dots, e_n)$, a local criterion e_i $(1 \leq i \leq n)$ is associated with a certain relation strategy X

$$e_i = e_i(X), \quad i = 1, \dots, n. \qquad (5)$$

Assume that the goal is to maximize all local criteria by selecting a strategy $X \in \mathbf{X}$ (the problem of minimizing some local criteria is easily converted into their maximization problem by changing the sign of these criteria). However, in vector optimization problems has non-antagonistic contradiction between local criteria: improving the quality of decisions on some criteria may degrade the quality of decisions on others. If this should not be given to local criteria corresponding weights, ie harmonize local criteria, to solve the problem, you can use an arbitration scheme [17].

Going to arbitration scheme reduces vector problem of decision-making to the equivalent (in the sense of the principle of optimality for both Nash arbitration schemes) scalar problem, which has a unique solution. It is convenient to move from the space $\mathbf{R}^m$ strategy sets X to the space $\mathbf{R}^n$ criteria vectors $e = (e_1, e_2, \dots, e_n)$ on the coordinate axes which are deposited values of local criteria.

Let U denote the image of the set of all admissible strategies X under the mapping defined by formula (5).

Now, to reduce the problem of multicriteria optimization to arbitration scheme $G =< I, U, u^* >$, where $I = \{1, 2, \dots, n\}$ - a set of local criteria, it is necessary to determine the point u *. This point can be determined from the content of the problem of multicriteria optimization or common sense, as well as enlisting the skills and intuition. Also as a point of status quo often take $u^* \epsilon U, u_i^*$ components are equal to the guaranteed value of the local criterion e_i $(1 < i < n)$ for all values of all other criteria. Then the solution of any multiobjective problem can be obtained using computational optimization scheme. Of course, in this case the solution is optimal [12].

3.3. Inventory management software

The mathematical methods allow to plan the size of supply, the minimum allowable residues and the supply period. Considering the distribution in real time, some specifications can change. For example, the supply period can change depending on demand. It's necessary to consider such factors as, for example, changing the dynamics of demand for

making purchasing decisions. The information systems requirements should decide tasks of accumulation of data on merchandise and contain the tools for decision support [4].

Companies use information systems to reduce their carrying costs. Inventory management software is an important component of computer system to monitor levels of enterprise sales, deliveries and stocks [3].

Inventory management software allow you to calculate the parameters: the size of supply, supply period, reorder point [14].

Also, computerized systems allow to track the real state of the company to ensure the stocks.

As a tool for the data used spreadsheets. The database are a tool for the accumulation of data on current inventories. It's necessary to have tools for intelligent decision support. All these qualities correspond to the requirements for Big Data [16].

The main disadvantages of the inventory management software based on Big Data are its cost and complexity. The high cost of inventory management software allow to use these tools only by large companies. Small businesses can't to afford to use expensive software. In addition, the management of companies, which use Big Data technology, should provide training of stuff to use of Big Data in the future.

As said before, we describe problem's solution of account balances and the use of tools for decision-making support using simple tools. The classic example is the task based on model of optimal order quantity EOQ.

3.4. Practical solution of inventory management for small business

The sale of the certain store is 500 units of production in the year. The level of demand is distributed evenly throughout the year. For the delivery of the order the store owner has to pay 1000 rubles. Order delivery time from the supplier is 12 working days (6-day work week). According to experts, the cost of storage is 40 rubles per year for a package. It is necessary to determine: how many bars must store owner order for one delivery; frequency of orders; reorder point and the annual cost of managing inventory. It is known that the shop is open 300 days a year.

This mathematical tools allows to calculate the re-ordering time and optimum batch size.

Stable models with determined demand:

$$Q^* = \sqrt{\frac{2K\mu\cdot\left(1+\frac{h}{d}\right)}{h}}; \qquad (6)$$

$$T^* = \sqrt{\frac{2K}{h\mu}}\cdot\sqrt{1+\frac{h}{d}}; \qquad (7)$$

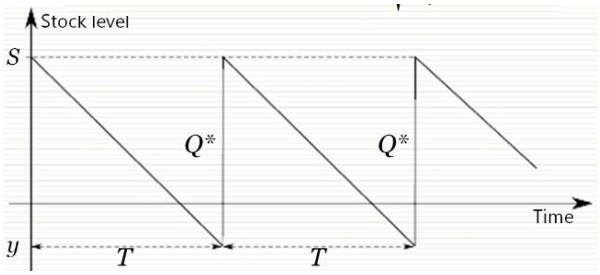


Fig.3. Static model with deterministic demand.

The optimum batch size

$$Q^* = \sqrt{\frac{2K\mu}{h}} = \sqrt{\frac{2\cdot 10\cdot 500}{0{,}4}} = \\ = 158{,}11 \approx 158\ (unit) \qquad (8)$$

Annual costs are

$$L_c = K\frac{\mu}{Q^*} + h\frac{Q^*}{2} = \\ = 1000\frac{500}{158} + 0{,}4\frac{158}{2} = 6325(rub/year) \qquad (9)$$

Delivery of each new order should be made through

$$T^* = \frac{Q^*}{\mu} = \frac{158}{500} = 0{,}316\ (year) \qquad (10)$$

Since in this case the year is 300 working days, the order should be served at the level of reserve:

$$S_0 = \mu T_\partial = \frac{500}{300}\cdot 12 = 20\ (unit) \qquad (11)$$

i.e., these 20 units will be sold for 12 days until the first order will be delivered.

3.5. Solution of practical task with the use of multiobjective optimization (game-theoretic approach)

Consider a concrete example for a small enterprise. Suppose that you want to allocate 100 unit of some resource between the two projects in the company's vector criterion

$$u = (u_1,\ u_2) = \left(1-\frac{x}{100}, \sqrt{\frac{x}{100}}\right), \qquad (12)$$

where x∈ [0, 100] - amount of resource, allocated to the second draft.

Using an arbitration scheme <I, U, u*>, where I = {1, 2}, and u = (0,0).
Seen that if x = 100, than $u_1 = 0$ and $u_2 = 1/10$,
and if x = 0,than $u_1 = 1$ and $u_2 = 0$. After we may express through local criterion u_1 u_2, i.e. we find

$$u_1 = 1 - 100u_2^2 \qquad (13)$$

and define the set $U = \{u | u_1, u_2 \in [0,1]\}$.
Then we find the arbitration solution $\tilde{u} = (\widetilde{u_1}, \widetilde{u_2})$, this solution maximizes:

$$g(u', U', u^*) = u_1 u_2 = (1 - 100u_2^2)u_2 \qquad (14)$$

Differentiating this equation and equating to zero the derivative, we obtain $\widetilde{u_2} = (10\sqrt{3})^{-1}$, and then find $\widetilde{u_1} = 2/3$. Consequently,

$$\widetilde{u} = \left(\frac{2}{3}, \frac{1}{10\sqrt{3}}\right) = (1 - \frac{x}{100}, \frac{\sqrt{x}}{100}), \qquad (15)$$

where $x = 100/3$.

Thus, according to the principle of optimality for general Nash [15] arbitration schemes first project in the company should allocate 200/3 unit, and the second 100/3 unit of resource.

There are many software to solution such task.

This task can be solved by OLAP tools in DBMS. Such tools allow you to use cheaper technologies of decision-making support for small business.

3.6. Decision-making support system tools based on databases

Information systems of enterprises solve the problem of continuous registration data of the goods accounting . In addition, it requires the development of tools to carry out of estimating size of stocks in real-time and to support decision-making [9].

It is known solution, based on SQL-instruction using the WHERE predicate [7].

The problem of representing data on current stocks propose to solve consistently using two SQL-instructions:

1) The query GrouppingOperation carries out the group all data according to the chronology and calculates the current amount of the issue and receipt of goods:

```
SELECT Month([Date]) AS [Month],
ItemsRegistration.Date,
Sum(ItemsRegistration.IncomeQuantity) AS
[In],
Sum(ItemsRegistration.ExpensesQuantity) AS
[Ex]
FROM ItemsRegistration
GROUP BY Month([Date]),
ItemsRegistration.Date
ORDER BY ItemsRegistration.Date;
```

2) The query RunningTotal calculates current stocks as progressive total:

```
SELECT GroupingOperations.Date,
Sum(GroupingOperations_1.[In]) AS Income,
Sum(GroupingOperations_1.[Ex]) AS Expenses,
[Income]-[Expenses] AS Stocks
FROM GroupingOperations AS
GroupingOperations_1 INNER JOIN
GroupingOperations ON
GroupingOperations_1.Date<=
GroupingOperations.Date
GROUP BY GroupingOperations.Date
ORDER BY GroupingOperations.Date;
```

We use this formula for the calculation of current stocks:

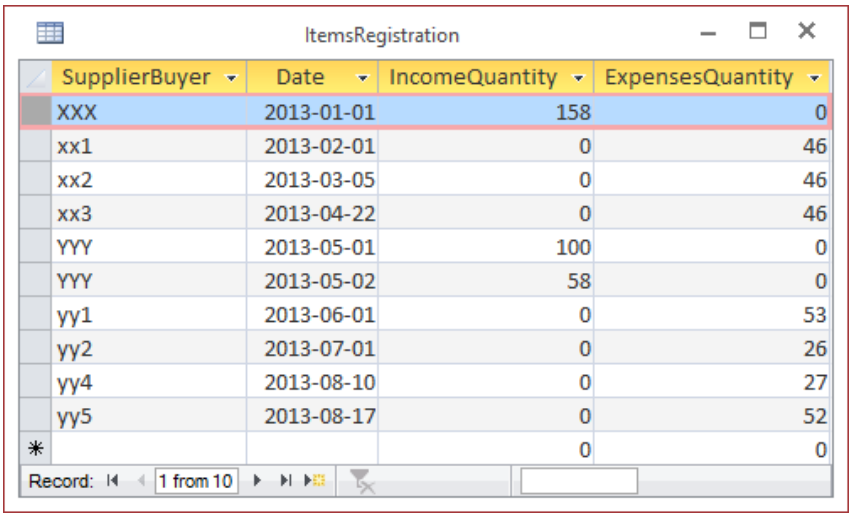
$$S = \sum_{i=1}^{n} s_i = \sum_{i=1}^{n} (I_i - E_i), \qquad (16)$$

s_i – current stocks;
I_i – current income;
E_i – current expenses.

In Fig. 4 it presents the dataset reflecting the chronological flow accounting (income and expenditure) of goods in the first period residues according to the calculated data above.

Implementation of the proposed queries for calculation of current balances shown in Fig.5.

ISBN: 978-1-61804-251-4

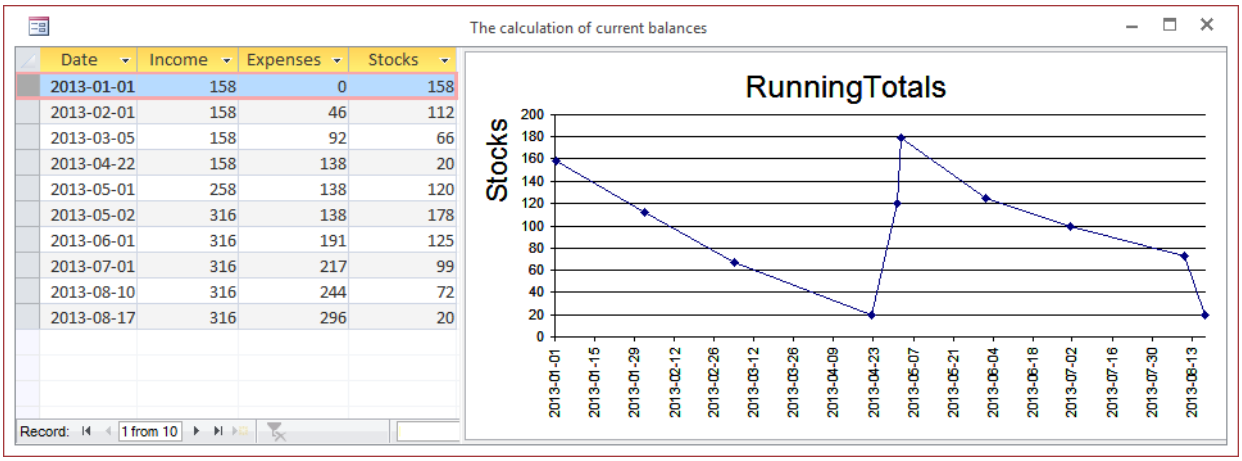


ItemsRegistration

SupplierBuyer	Date	IncomeQuantity	ExpensesQuantity
XXX	2013-01-01	158	0
xx1	2013-02-01	0	46
xx2	2013-03-05	0	46
xx3	2013-04-22	0	46
YYY	2013-05-01	100	0
YYY	2013-05-02	58	0
yy1	2013-06-01	0	53
yy2	2013-07-01	0	26
yy4	2013-08-10	0	27
yy5	2013-08-17	0	52
		0	0

Record: 1 from 10

Fig.4. A set of data tables accounting in the first period (six months)



The calculation of current balances

Date	Income	Expenses	Stocks
2013-01-01	158	0	158
2013-02-01	158	46	112
2013-03-05	158	92	66
2013-04-22	158	138	20
2013-05-01	258	138	120
2013-05-02	316	138	178
2013-06-01	316	191	125
2013-07-01	316	217	99
2013-08-10	316	244	72
2013-08-17	316	296	20

Record: 1 from 10

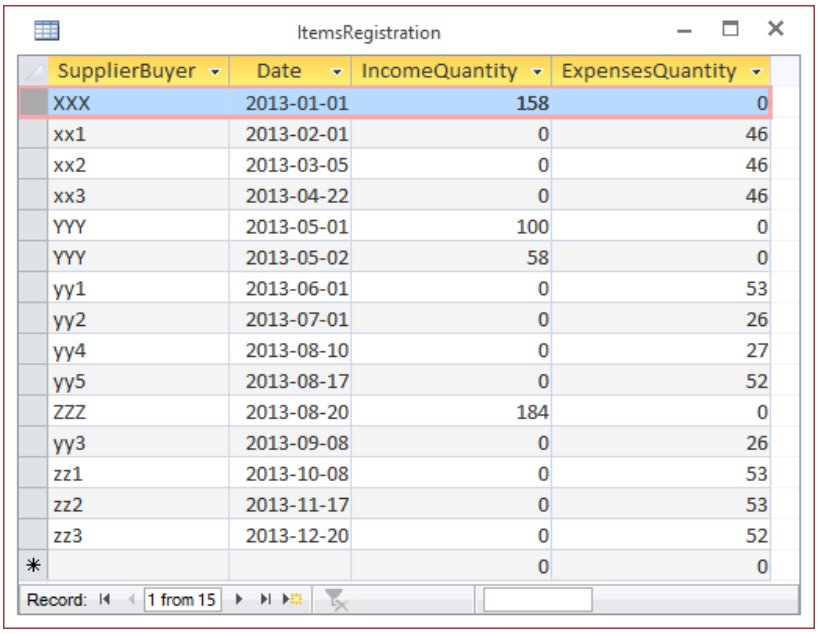
Fig.5. Current balances and the graph for the first period.

The graph shows, the current stocks correspond to the calculated: the supply - 158 units, uniform sales - for the period of 4 months, the stocks is 20 units on the eve of the second supply. Then the model reflects the supply 158 units (in 2 periods: 100+58). Last points of the graph show the situation of increasing of goods sales to a marginal stocks size of 20 units. This fact supports the decision about the new order. Taking into account the dynamics of demand, a possible solution is to rise the size of the order (158+26).

The model of implementation of this order and further change in stocks up to the end of the year you can see in Fig. 6 and Fig.7.

The graph of inventory changes and dynamics of reserves is a good way to make decisions on adjusting the time and size of supply.

ISBN: 978-1-61804-251-4

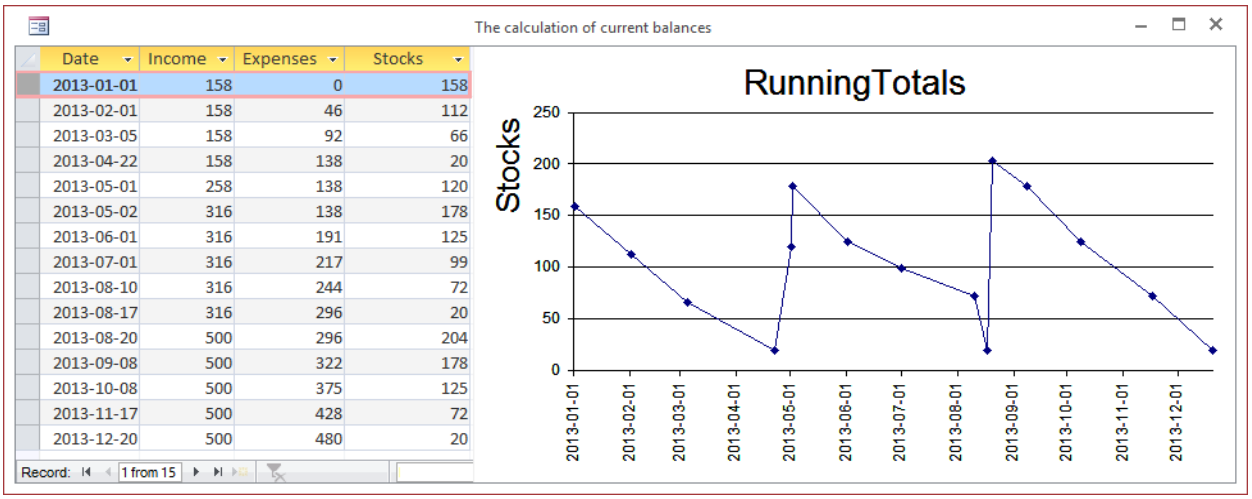


ItemsRegistration

SupplierBuyer	Date	IncomeQuantity	ExpensesQuantity
XXX	2013-01-01	158	0
xx1	2013-02-01	0	46
xx2	2013-03-05	0	46
xx3	2013-04-22	0	46
YYY	2013-05-01	100	0
YYY	2013-05-02	58	0
yy1	2013-06-01	0	53
yy2	2013-07-01	0	26
yy4	2013-08-10	0	27
yy5	2013-08-17	0	52
ZZZ	2013-08-20	184	0
yy3	2013-09-08	0	26
zz1	2013-10-08	0	53
zz2	2013-11-17	0	53
zz3	2013-12-20	0	52
*		0	0

Record: 1 from 15

Fig.6. A set of data tables accounting for the year.



The calculation of current balances

Date	Income	Expenses	Stocks
2013-01-01	158	0	158
2013-02-01	158	46	112
2013-03-05	158	92	66
2013-04-22	158	138	20
2013-05-01	258	138	120
2013-05-02	316	138	178
2013-06-01	316	191	125
2013-07-01	316	217	99
2013-08-10	316	244	72
2013-08-17	316	296	20
2013-08-20	500	296	204
2013-09-08	500	322	178
2013-10-08	500	375	125
2013-11-17	500	428	72
2013-12-20	500	480	20

Record: 1 from 15

Fig.7. Current balances and the graph for the year.

As this example demonstrates, running totals are simply to create.

In addition, such approach improves the query processing. The speed of processing is raised, because the query RunningTotal contains operator INNER JOIN to join object GroupingOperations and its replica GroupingOperations_1.

4. Conclusion

The paper proposes a solution to the problem of accounting balances and the use of the decision support with simple tools. Also invited to the game-theoretic approach to multiobjective optimization when deciding how to allocate resources (reserve) between projects or departments within the company. The suggested queries, providing registration of balances in real time, displaying the results in the form of graphics, which facilitate a decision on procurement. The considered example demonstrates full compliance of basic parameters between the characteristics of inventory management processes, described using mathematical models, and practical implementation of a process by means of databases. This proves the correctness of using this mathematical model in the inventory management tasks, within the limits of the subject area. The offered means of information systems invariant with respect to the mathematical methods, which is using in the planning of stocks management.

ISBN: 978-1-61804-251-4

References:

[1] Axsäter, S., *Inventory Control*. Springer, 2006.

[2] Berezhnaya E.V., *Mathematical methods of modeling of economic systems: Proc. allowance,* Moscow: Finance and Statistics, 2001.

[3] Boronenko S.D., Iliashenko O.Y., *Application design based on the underlying data model.* // The XV Tsarskoye Selo readings: 20-21 Oct. 2011 - I. II. - SPb.: Pushkin Leningrad State University, 2011, pp.151-155.

[4] Cios K. J., *Data Mining: A Knowledge Discovery Approach*, Springer, 2007.

[5] Codd E. F., Codd S. B., Salley C. T. *Providing OLAP (On-line Analytical Processing) to User-analysts: An IT Mandate,* Codd & Associates, 1993

[6] Ferrari, A. & Russo M., *Microsoft Excel 2013 Building Data Models with PowerPivot (Business Skills)*, O'Reilly Media, Inc., 2013

[7] Fuller, A., *Running totals in SQL server queries*, http://www.techrepublic.com/article/running-totals-in-sql-server-queries

[8] Jelen, B. & Michael A., *Excel 2013 Pivot Table Data Crunching*, MrExcel Library, 2013

[9] Kimball, Ralph; Margy Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd ed.), Wiley, 2013.

[10] Kroenke D.M., Auer D.J., *Database processing. Fundamentals, design, and implementation*, Pearson, 2011

[11] Kuzin B.I., Yuriev V.N., Chakhdinarov G.M., *Methods and models of management of the company*, St. Petersburg, Special literature, 2001.

[12] Labkaster L.G., Babeshko L.O. *Game methods to manage the economy and business: Textbooks.* - M.: Business.

[13] Muckstadt, J. A., & Sapra, A., *Principles of Inventory Management.* Springer, 2010.

[14] Nagabhushana, S., *Data Warehousing: OLAP and Data Mining*, New Age International, 2006.

[15] Petrosyan L.A., Zenkevich N.A., Semina E.A. *Game Theory: A Tutorial for University Comrade*, M.: Higher School, Book House "University", 1998.

[16] Reeves L. L., A Manager's Guide To Data Warehousing, Wiley, 2009.

[17] Shirokova S.V., *Game Theory. Implementing game approach in the management of the company*, St. Petersburg. Univ Polytechnic University, 2008.

[18] Thomsen, Erik. OLAP Solutions: Building Multidimensional Information Systems, Wiley, 2002.

[19] Wrembel, Robert, Koncilia, Christian, Data Warehouses and OLAP: Concepts, Architectures, and Solutions, Idea Group Inc., 2007.

Adjustment semantics of real time constructions in UCM language for implementation in translator of UCM to Basic Protocols

Kotlyarov V., Drobintsev P., and Nikiforov I., St. Petersburg State Polytechnic University

***Abstract*—** The paper describes an approach to adjustment of semantics for UCM real time constructions in implementation of translator into Basic Protocols notation. The following constructions and their adjustment are described: multithreading, delays and interruptions. The main problem of such constructions is that initial version of UCM standard allows to create semantically incorrect models. Proposed extensions and restrictions of UCM semantics allowed solving of these problems for different types of projects.

***Keywords*—** UCM semantics, timers, delays, behavioral tree, synchronization, threads generation.

I. Introduction

SOFTWARE system development starts with creation of requirements. Documents specifying requirements specifications are generally written in natural language and may contain hundreds and thousands of requirements. Initial specifications often contain errors related to discrepant, incomplete and nondeterministic system behavior. Locating and fixing errors in requirements are more effective at early stages of the development [1].

It is almost impossible to manually analyze industrial systems specifications on errors presence without supporting toolset. Existing systems of verification and testing do not work with informal specifications. Thus the actual task is formalization of initial textual requirements using input languages of the tools for verification and testing.

One of the perspective integrated technologies of testing automation and verification based on formal models is VRS/TAT technology [2], where Use Case Maps (UCM) [3] notation is used for high level description of behavioral models while tools for checking and generation automation work with the model in basic protocols language [4].

UCM specifications language is standardized, but however contains a number of inaccuracies which do not allow displaying the modeled systems semantics unambiguously and correctly.

Proposed in this work are restrictions on development of multithread models of the systems as well as adjustments of UCM language constructions semantics modeling time delays and interruptions.

II. Use Case Map

Use cases describe sequences of actions performed by a system in response to external impact from users or other software systems (components). Use cases reflect system functionality from system architecture description point of view. They introduce important components in software systems development process [5], namely:

- fill in the gap between textual requirements description and detailed system design;
- - allow developing system architecture on high level of abstraction as well as specifying system behavior when architecture is already defined;
- - help a developer to predict complicated systems behavior;
- - provide convenient notation for depicting parallel structures, timers, interruption points on the diagram and aspects using.

System design in UCM language is presented as a set of diagrams interacting between each other. Each diagram in turn is focused on description of components (agents, system processes), objects, observers and subsystems interaction. Each component and subsystem contains elements of responsibility (Responsibilities) corresponding to some events in the system as well as strictly defined sequence of their occurrence.

Using elements of UCM notation not only linear behavior can be specified but also parallel scenarios (AndFork) with their further synchronization (AndJoin) can be described. FailurePoint element participates in description of interruptions generation and processing mechanism. Timer element is used to specify system timer behavior both for cases with simple time delay and for cases with complicated logical behavior.

Also structuring element (Stub) is worth noticing which allows creating hierarchical system representation and conducting the development by components from the highest level of abstraction to detailed description of low level diagrams.

Thus the aggregation of components and diagrams provides visible representation of system behavior and system's components interaction to the user.

UCM diagram is developed using UCM Navigator [6]

ISBN: 978-1-61804-251-4

graphical editor. In figure 1 a fragment of UCM diagram for real telecommunication project is shown where high level behavior of the agent modeling automatic telephone station is described.

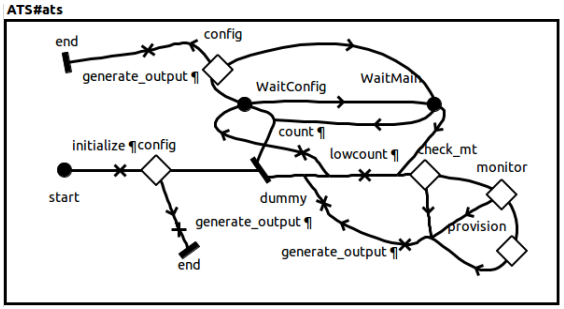


Figure.1. UCM diagram with automatic telephone station module behavior.

III. Restrictions of Multithread Systems Development

Despite the advantages of UCM language and its elements semantics usage, the existing standard Z.151 [3] contains a number of inaccuracies hampering multithread systems modeling.

Brackets balance in parallel threads specification

Consider a case when syntactically correct elements of threads generation and synchronization can cause a violation of parallel threads structure and consequently an incorrect system behavior.

In figure 2 after *AndFork_A* and *AndFork_E* elements generation of threads B, E and F, G respectively is performed while on *AndJoin_C* and *AndJoin_D* elements synchronization of threads B, F and C,G respectively is performed.

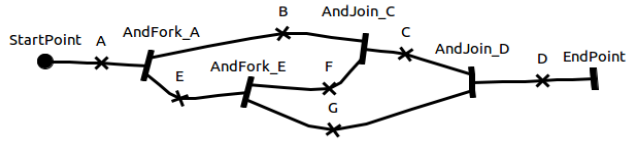


Figure 2. Violation of parallel threads structure.

It is easy to notice that synchronization of threads generated by different elements significantly complicates the mechanism of errors detection and fixing in the system as well as complicates the tracking of parent/child connection in threads hierarchy. Such connections are useful when child thread keeps executing after parent thread has finished.

System behavioral graph with correct structure of threads generation and synchronization is depicted in figure 3.

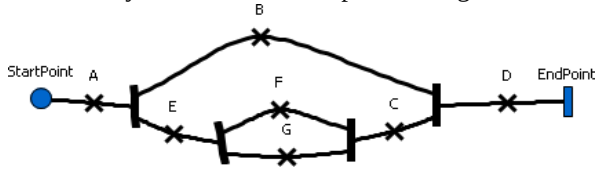


Figure 3. The graph with correct threads structure.

Threads structure analysis can be compared with the analysis of mathematical expressions brackets format. If expression brackets format is violated, it is considered to be syntactically incorrect. This is also valid for parallel threads modeling: if threads structure is violated, the whole system is considered to be syntactically incorrect.

Analyzing threads for errors detection and fixing allows creating syntactically correct system models.

Unlimited generation of threads

Consider the case shown in Fig.4. Threads B and E are generated after D element. Thread B is finished on EndPoint element. Thread E returns through the cycle, which has no condition of iterations limits, and D element and generates new threads B' and E'. The behavioral scenario is repeated for thread E'.

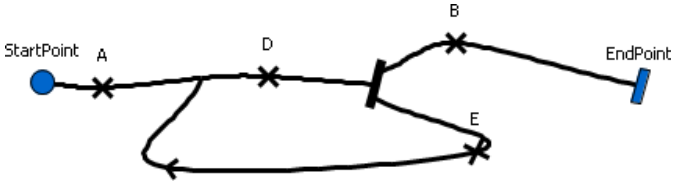


Figure 4. Unlimited generation of threads

Unlimited cycles lead to generation of unlimited number of unfinished threads which leads to shortage of memory and other resources. Thus it is important to introduce restrictions on usage of such constructions in developed models.

Data racing while accessing shared resources by parallel processes

Consider the case when shared resources are used on parallel branches without synchronization. Figure 5 depicts two parallel threads using "**var**" shared resource without synchronization. Such formalization leads to racing while data access [7].

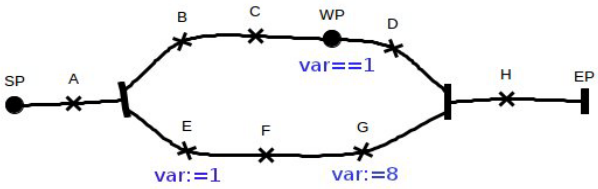


Figure 5. Shared resources without synchronization.

There are two executable scenarios in the model worth noticing.

1) If "**E**"-"**F**"-"**G**"-"**WP**" scenario is executed, "**D**" element will never be applied. This scenario leads to a deadlock.
2) If "**E**"-"**F**"-"**WP**"-"**G**" scenario is executed, "**D**" element can be applied and this scenario will reach EP end point.

Deadlock can be avoided by introducing synchronization and thus excluding parallel access to shared system resource (figure 6).

ISBN: 978-1-61804-251-4

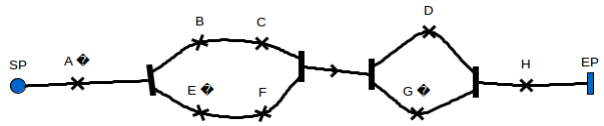


Figure 6. System model with synchronization.

Proposed restrictions on multithread systems development in UCM

- Using parallel constructions with violated threads structure is not allowed.
- Using unlimited recursive threads generation is not allowed.
- Using shared resources on parallel execution paths without synchronization is not allowed.

IV. Features of Time Delays Modeling

Requirements of time delays creation are often met in industrial systems. Note that used in the considered case is event modeling with relative time – the time between events. Events are the change in system attributes values.

Features of timer usage

According to the standard [3] two outgoing paths are connected with Timer element (figure7): regular path (RP) and timeout path (TOP). For selecting each path there are conditions CRP and CTOP respectively. Also there is a trigger path (or trigger counter) which affects timer behavior and allows to cancel the delay.

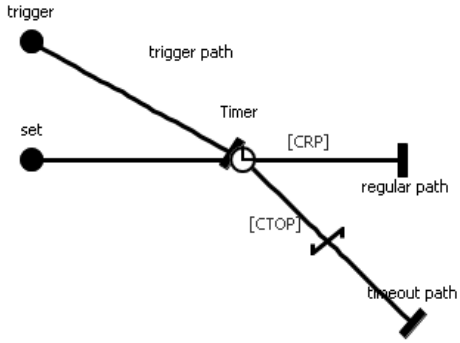


Figure 7. UCM diagram with Timer element

In Z.151 standard semantics of the elements modeling time delays contains cases description of model possible behaviors depending on occurred events, but does not describe which types of events are associated with timers and does not specify types of some events specific for telecommunication applications specification.

Extend timer semantics description

Extend UCM timer semantics description with the following events:

- Timer set: TIMER_SET <timer name>. The event occurs when Timer element is reached.
- Timer expiration: TIMER_EXPIRE <timer name>. The event occurs after CTOP condition has been executed.
- Timer reset: TIMER_RESET <timer name>. The event occurs after RP or TOP path has started execution or at trigger event occurrence.

Using semantics of Timer element and associated events three types of time delays can be stated:

1) simple delay, whose modeling feature is strictly specified conditions of outgoing paths (false) and absence of trigger event;
2) interruption delay, whose modeling feature is presence of trigger event;
3) interrupted execution delay, whose modeling feature is presence of FailurePoint interruption on timeout path.

Proposed extension of UCM timer semantics by timer set, expiration and reset allowed solving the delay description problem for telecommunication projects.

V. Features of Interruptions Modeling

There are two types of interruptions in requirements: local interruptions, affecting behavior of specific function or object, and interruptions, affecting behavior of other system threads. Each interruption shall has a corresponding handler.

Interruptions are modeled by the group of elements [3]: FailurePoint, AbortStartPoint and FailureStartPoint. Figure 8 depicts a simple UCM diagram modeling an interruption with handler.

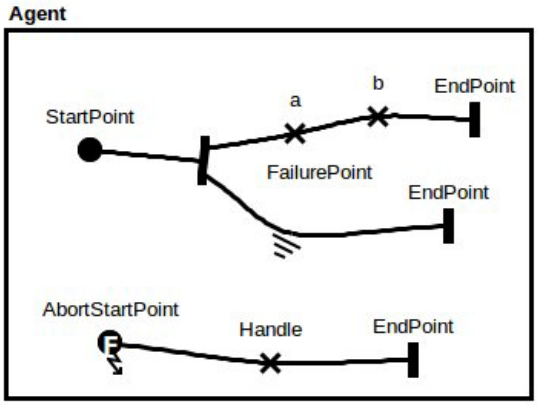


Figure 8. UCM diagram modeling an interruption with handler

When FailurePoint (grounding symbol) is reached, interruption occurrence condition is calculated. Call it *FailureCondition*. If calculation result is true, then the execution flow with FailurePoint will be interrupted and a flag of interruption occurrence (call it *FailureFlag*) will be enabled.

As soon as *FailureFlag* equals **true**, conditions calculation on all interruptions handlers is performed: FailureStartPoint and AbortStartPoint. Handlers types behavior is described in the standard [3].

The main difference between two types of handlers is the impact on parallel threads, affected by the interruption. In case

ISBN: 978-1-61804-251-4

of FailureStartPoint only interruption of the thread which reached FailurePoint is performed. In case of AbortStartPoint interruption of all threads which belong to FailurePoint activity area is performed. Call this set AbortScope.

Usage of any of the considered elements introduces the enormous number of system behaviors, which need to be checked during verification. In general case checking of all possible execution variants is impossible due to states explosion problem.

For all elements from the AbortScope set it is required to check interruption occurrence which means to perform interleaving of all cases where interruption can occur.

Proposed are three approaches which can be used either separately or supplement each other:

1. Checking behaviors on the bounds of linear parts of paths and in the points of common resources sharing. For this purpose the analysis of the paths and elements set from FailurePoint activity area is performed as well as key points where behavior shall be checked are specified. Combination of all possible behaviors is performed for these points only.
2. User check. User marks his check points on the diagram with a marker.
3. Default check, i.e. verification will be performed for all elements of the set. This case can be only used after manual introduction of restrictions on the set of verified elements [8], otherwise states explosion is inevitable.

Worth noticing, that the first two approaches are enough to balance the time of verification and required coverage level. In general case usage of proposed approaches allowed to effectively solve the problem of interruption description in telecommunication projects.

VI. Conversion of time delays into basic protocols

For VRS/TAT toolset for verification and testing a tool for translation of models in UCM language into models in basic protocols language was developed [9,10,11]. UCM→BP translator implements the conception of time delays and interruption conversion as well as checking of formulated restrictions on multithread systems development.

Consider the features of some constructions translation important for specification of real-time applications.

For Timer element there is *timer_var* attribute, which is responsible for timer state and assigned two possible values: **true** — if the timer is set, **false** – if the timer is reset. By default the value is **false**.

In basic protocol for Timer element responsible for timer set *timer_var:=true* expression is generated in postcondition while TIMER_SET expression is generated in the process field of basic protocol.

For each outgoing path (RP and TOP) from Timer element a single basic protocol is generated.

Precondition of the basic protocol for RP path (expression for selection of regular path) is generated in accordance with logic formula derived based on [5]:

(timer_var=true)&(CRP)∨

(timer_var=true)&(trigger)&(CTOP) ***(1)***

where trigger is a logic expression for trigger event.

TIMER_RESET action is generated in the process field of basic protocol.

In postcondition of this basic protocol the expression modeling timer reset is generated: *timer_var:=false.*

Consider a basic protocol for timeout path (TOP). In general case the following expression is generated in precondition:

(timer_var=true)&(~CRP)&[(CTOP∨

(~trigger)&(~CTOP)] ***(2)***

TIMER_EXPIRE and TIMER_RESET operations are generated in the process field of the basic protocol for TOP path.

Thus, conversion of time delays implies generation of three basic protocols with different logical expressions in precondition.

For each considered case of timer modeling optimization of logical expressions is possible as the values of used conjuncts and disjuncts is known beforehand.

VII. Conversion of interruptions in UCM language into Basic Protocols

Two basic protocols are generated in basic protocols notation for elements modeling interruptions (FailurePoint). The first protocol is for regular execution path with negation of interruption occurrence *~(FailureCondition)* in precondition.

The second protocol contains checking of interruption occurrence *FailureCondition* in precondition and a flag in postcondition signaling that the interruption has occurred – *FailureFlag:=true*.

For all handlers of interruptions: FailureStartPoint and AbortStartPoint a new execution flow will be created, the first protocol for each of them will contain checking of interruption occurrence in the system in precondition as well as expression for this handler enabling, call it *HandlingCondition*.

(FailureFlag_1 = true)∨ (FailureFlag_2 = true∨...∨ (FailureFlag_n = true) & (HandlingCondition) ***(3)***

For all protocols generated for elements of AbortScope set *~(FailureFlag=true)* expression is added to precondition which means that flow execution will continue until exception will occur.

Using approaches to translation of time delays and interruptions a conversion principle of time delay with FailurePoint element can be described. For this case generated are a basic protocol for timer set, two basic protocols for timer exit and two protocols for FailurePoint element.

Herewith the condition of interruption event occurrence is added to protocols which lead to FailurePoint.

VIII. Usage example

Using of proposed UCM semantics adjustments in the project for SMTP mail protocol presented on figure 9 allowed translating of the set of UCM behavioral diagrams into 56 basic protocols, performing of model verification, generating

241 test scenarios and testing the mail protocol which reduced the efforts on 26% in comparison with traditional approach of manual testing.

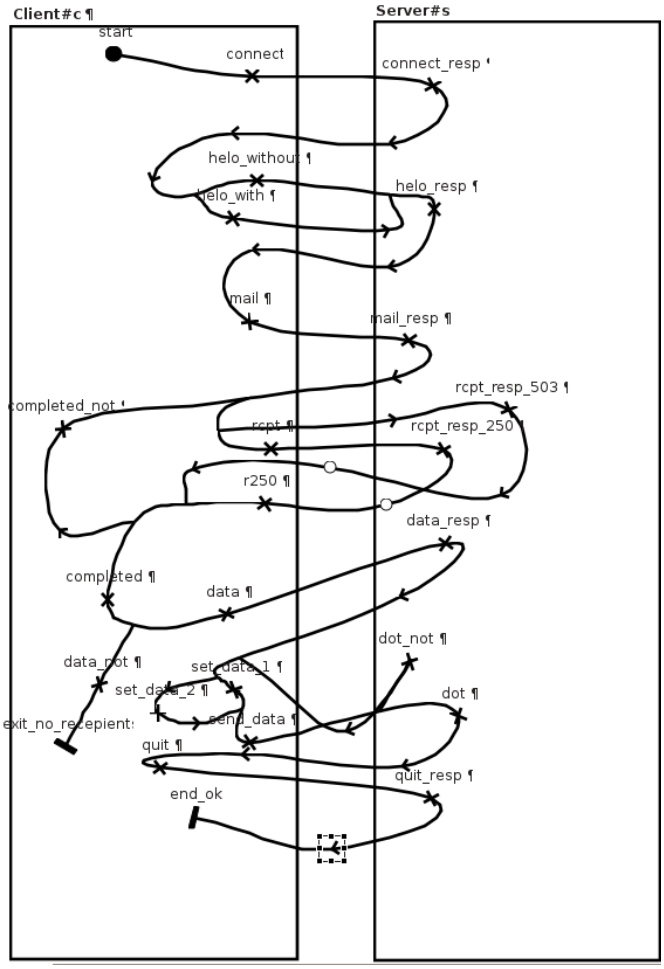


Figure 9. UCM diagram for SMTP project

IX. Conclusion

Considered in this work methods of semantics adjustments of UCM standard elements, modeling time delays and interruptions as well as restrictions on multithread systems development allow modeling of complex telecommunication systems reducing possibility to create semantically incorrect model behaviors.

Methods are implemented in UCM→BP translator which allows using of VRS/TAT technological chain more convenient and effective for projects of middle and high complexity.

Proposed translator together with supporting toolset of VRS/TAT technology was applied in modules development of telecommunications applications and has shown a significant reduction of efforts on quality industrial software project development.

References

[1] Booch Gr., Maksimchuk R., Engel M., Young B., Conallen J., Houston K. Object-Oriented Analysis and Design with Applications. Addison-Wesley Professional; 3rd edition, 2007. 720 p.

[2] Veselov A.O, Kotlyarov V.P. Test automation in telecommunication area // “Scientific and technical sheets of SpbSTU”. №4(103). Spb.: SpbSTU publishing, -2010. - pp. 180-185 (in Russian)

[3] Recommendation ITU-T Z.151. User requirements notation (URN), 11/2008

[4] Letichevsky A.A., Kapitonova Yu.V., Letichevsky A.A. (Jr) and others. Systems specification using basic protocols // Cybernetics and system analysis, 2005, №4. pp.3-21 (in Russian)

[5] Buhr R. J. A., Casselman R. S., “Use Case Maps for Object-Oriented Systems.” Prentice Hall, 1995.

[6] UCM Navigator - http://jucmnav.softwareengineering.ca/ucm/bin/view/ProjetSEG/WebHome

[7] Gergel V.P. High-performance calculations for multiprocessor multicore systems. Nizhny Novgorod: University of Nizhny Novgorod, 2010. – 544 p. (in Russian)

[8] P.Drobintsev, V.Kotlyarov, I.Chernorutsky. Test automation based on user scenarios coverage. “Scientific and technical sheets”, St.Petersburg university, vol.4(152)-2012, pp.123-126 (in Russian)

[9] Nikiforov I.V., Petrov A.V., Yusupov Yu.V. Generating formal model of the system based on requirements specified in USE CASE MAP notation // “Scientific and technical sheets of SpbSTU”. №4(103). Spb.: SpbSTU publishing, -2010. - pp. 191-195 (in Russian)

[10] I.Anureev, S.Baranov, D.Beloglazov, E.Bodin, P.Drobintsev, A.Kolchin, V. Kotlyarov, A. Letichevsky, A. Letichevsky Jr., V.Nepomniaschy, I.Nikiforov, S. Potienko, L.Pryima, B.Tyutin. Tools for supporting integrated technology of analysis and verification of specifications for telecommunication applications // SPIIRAN works- 2013-№1-28P (in Russian)

[11] 8. I.Nikiforov, A.Petrov, V.Kotlyarov. Static method of test scenarios adjustment generated from guides // “Scientific and technical sheets”, SpbSTU, vol.4(152)-2012, pp. 114-119 (in Russian)

Vsevolod Kotlyarov - was born in Stavropol region of Russia on the 14 July 1944. Hold a master degree with specialty «Mathematical and computing instruments and devices» of Saint-Petersburg State Polytechnic University (SPbSPU) in 1968. Defended PhD thesis with specialty “Software engineering” in 1972. Main areas of interests - «Software engineering», «Technologies and tools of automated verification and testing».

Since 1972 he is working as associated professor in SPbSPU, since 1995 as senior researcher in St.Peterburg software development department of Motorola, since 2008 as full time professor of SPbSPU. He is scientific adviser of 20 PHD dissertations of post-graduate students. His scientific school of “Software Engineering” was included in the list of top schools of St.Petersburg.

Prof. Kotlyarov became a M of IEEE and ACM in 1993, M of SABA (Science Advisory Board Association) of Motorola Company in 2005. He is a member of the program committees of the following conferences: Microsoft Technology in Software theory and practice, SYRCOSE, Workshops of Ershov informatics conference (PSI).

Bayesian Probability Models for Critical Illness Insurance

P. Jindrová, V. Pacáková

***Abstract*—**Critical Illness Insurance (CII) is a type of long term insurance that provides a lump sum on the diagnosis of one of a specified list of critical illnesses within the policy conditions. Knowledge of the probability of an insured event is the basis for the valuation of the products in life and non-life insurance companies. The aim of this article is to use Bayesian binomial/beta model for estimation of event probability for critical illness insurance. In Bayesian approach to estimation we should always start with a priori distribution for unknown parameter. In this paper we have used the algorithm for such a priori estimation of the binomial probability, which allows Bayesian estimates with less square error compared with classical estimates. In article the algorithm has been applied on the data submitted by the Decree No. 20/2008 to the National Bank of Slovakia from Slovak insurance companies giving exposure to the critical illness risk.

***Keywords*—**Bayesian estimation, binomial/beta model, event probability, posterior distribution, prior distribution.

I. Introduction

Critical illness insurance (CII) first came to the scene in South Africa early in the 1980s under the name of Dread Disease Insurance. However, before this, in the USA, Japan and Israel some life insurance policies were extended to cover cancer. CII has been very popular in the UK. Although CII policies have been issued since the 1980s in the UK, the number of policies increased dramatically in the early 1990s. Currently critical illness insurance is common product of many insurance companies around the world, although these insurance products vary in number and set up of diseases, they cover.

CII covers pay an insurance benefit if the insured person suffers a serious condition, depending on the definitions stipulated in the policy wording, such as cancer, heart attack, stroke, coronary artery (bypass) surgery or kidney failure. The number of diseases covered varies considerably depending on the market and provider concerned.

This kind of insurance products covers against the financial consequences of the serious condition. People affected are given financial support to enable them to better manage their changed circumstances of life.

Insurance products differ in their specifications and in premiums. In their creation there is necessary knowledge about the probabilities of claims that are covered by critical illness policy. These probabilities need to know for different homogeneous groups of clients.

This article aims to explain and apply methods of classical and Bayesian statistical inference to estimate the probability of critical illness, specifically for the men and women and for various age groups. It also included a comparison of results obtained by above mentioned two approaches and shows the advantages of Bayesian estimates. Article investigates the Bayesian estimators of the parameters of binomial distribution using quadratic loss function. The possibility to express the Bayesian estimators in the form of credibility formulas allows easy application of these models in insurance practice.

II. The Binomial/Beta Model of Event Probability

The classical approach to point estimation treats parameters as something fixed but unknown. The essential difference in the Bayesian approach to inference is that parameters are treated as random variables and therefore they have probability distributions.

Suppose $x = (x_1, x_2, ..., x_n)$ is a random sample from a population specified by density function $f(x/\theta)$ and it is required to estimate parameter θ. By Tse, Y. K. [13] or Waters, H. R. [14] prior information about θ that we have before collection of any data is the prior distribution $f(\theta)$ which is probability density function or probability mass function. The information about θ provided by the sample data $x = (x_1, x_2, ..., x_n)$ is contained in the likelihood

$$f(\mathbf{x}/\theta) = \prod_{i=1}^{n} f(x_i/\theta) \tag{1}$$

Bayes theorem combines this information with the information contained in $f(\theta)$ in the form

$$f_{\Theta}(\theta/\mathbf{x}) = \frac{f(\mathbf{x}/\theta)\, f(\theta)}{\int_{\theta} f(\mathbf{x}/\theta)\, f(\theta)\, \mathrm{d}\theta} \tag{2}$$

which determines the posterior distribution. A useful way of expressing the posterior density is to use proportionality. We can write

P. Jindrová is with Institute of Mathematics and Quantitative Methods, Faculty of Economics and Administration, University of Pardubice, Pardubice, Studentská 84, 532 10 Pardubice, Czech Republic (e-mail: Pavla.Jindrova@upce.cz).

V. Pacáková is with Institute of Mathematics and Quantitative Methods, Faculty of Economics and Administration, University of Pardubice, Pardubice, Studentská 84, 532 10 Pardubice, Czech Republic (e-mail: Viera.Pacakova@upce.cz).

ISBN: 978-1-61804-251-4

$$f(\theta/\mathbf{x}) \propto f(\mathbf{x}/\theta)\, f(\theta) \tag{3}$$

or simply *posterior* $\propto$ *likelihood * prior*.

The posterior distribution contains all available information about θ and therefore should be used for making decisions, estimates or inferences. The following procedure of Bayesian estimation of the binomial parameter is explained for example in Boland [1], Pacáková, [9]–[11], Jindrová [5].

For estimation of a binomial probability θ from a single observation X with the prior distribution of θ being beta with parameters α and β, we will investigate the form of the posterior distribution of θ. Prior beta density function by assumption and omitting the constant is

$$f(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad 0<\theta<1. \tag{4}$$

Note that the uniform distribution on (0,1) is a special case of the beta distribution with $\alpha = 1$ and $\beta = 1$. This corresponds to the non-informative case. Omitting the constant likelihood is given by

$$f(x/\theta) \propto \theta^{x}\,(1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n \tag{5}$$

where n is number of independent trials (in our case number of policies) and x is number of events.

By (3) we get the posterior density of θ in the form

$$f(\theta/x) \propto \theta^{x}\,(1-\theta)^{n-x}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1} = \theta^{\alpha+x-1}\,(1-\theta)^{\beta+n-x-1} \tag{6}$$

Apart of the appropriate constant it is the posterior beta density function of θ with new parameters

$$\alpha' = \alpha + x \tag{7}$$

$$\beta' = \beta + n - x \tag{8}$$

By minimizing the quadratic loss the Bayesian estimator of θ can be expressed as the mean of this posterior distribution as follows:

$$\theta_B = \frac{\alpha + x}{(\alpha + x) + (\beta + n - x)} = \frac{\alpha + x}{\alpha + \beta + n} \tag{9}$$

We can rewrite the Bayesian estimator of θ in the form of credibility formula by Gogola [3], Šoltés [12] or Gray and Pitts [4]:

$$\theta_B = Z \cdot \frac{x}{n} + (1 - Z)\cdot \mu \tag{10}$$

where factor credibility Z can be expressed as

$$Z = \frac{n}{\alpha + \beta + n} \tag{11}$$

and μ is the mean of the prior beta distribution expressed as

$$\mu = \frac{\alpha}{\alpha + \beta}. \tag{12}$$

We note that as n (number of insurance policies) increases, the weight Z attaching to the data-based estimator increases and the weight attaching to the prior mean correspondingly decreases.

III. Source of Data

To estimate the probability θ that the insured person is diagnosed a critical illness we have found the data about the number of insurance agreements on a CII and the number of claims from these insurance contracts from Slovak insurance companies giving exposure to the critical illness risk. To estimate this probability we have found the data about the number of claims x and risk exposure n in the years 1999-2010 from dataset of NBS [8].

Data covering the period 1999-2010 were submitted to the National Bank of Slovakia based its Decree No. 20/2008 on submitting of actuarial data and statistical data of insurance company and branch of a foreign insurance company, on the basis of which it started to gather statistical data about insured people from insurance undertakings in 2009. The data were gathered in classification according to gender, age and thirteen insurance risks. Among them there are also the critical illness risks.

IV. Results and Discussion

Let θ is unknown probability of diseases on critical illness. To estimate this probability we have found the data about the number of claims x and risk exposure n in the years 1999-2010 from dataset of NBS [8].

Before collecting these data there are no information about this risk. In case of no prior knowledge if θ is a binomial probability, then a prior distribution which is beta distribution with parameters $\alpha = 1$, $\beta = 1$ often would seem appropriate. This assumption leads to the a priori estimate $\theta = 0{,}5$ in the first year 1999, which highly overstates real value of the probability of critical illness diagnosis in Slovak republic.

To eliminate this drawback, instead of interval <0;1> for a priori estimate of probability θ need to propose more realistic interval in which we assume a uniform prior distribution. Such interval and the algorithm for its use in Bayesian estimation of the binomial probability of random event θ was published in Kotlebová and Láska [6]. The proposed procedure is as follows:

We set the interval $(\theta_{\min}, \theta_{\max})$, in which we suppose to get a better estimate.

We denote by the symbol s the mean of beta prior distribution, which is the center of this interval:

$$s = \frac{x_{\min} + x_{\max}}{2}.$$

We mark as θ_0 the more distant boundary from the value of 0,5 of the interval $(\theta_{\min}, \theta_{\max})$.

Calculate the allowable error as $h_B = |\theta_0 - s|$.

We calculate q according to the formula

$$q = \frac{2n\theta_0(1-\theta_0)}{nh_B^2 - \theta_0(1-\theta_0)}. \tag{13}$$

We estimate the parameters α, β of the a priori beta distribution as follows:

$$\alpha = qs, \quad \beta = q - qs. \tag{14}$$

Bayesian estimator of probability θ we obtain by the formula (9).

According to the Material of Munich Re Group: Critical Illness Insurance [7] we selected the interval (0,0001; 0,0003) for a priori estimate of parameter θ for category of men to 30 years old. Following the procedure described above, we have obtained a priori estimate of the probability θ of diseases on critical illness in the year 1999. In calculation by (13) we used the number n = 5390866 of the population in Slovak republic in the year 1998.

The procedure of prior distribution parameters estimation for the category of men less than 30 years is as follows:

1. $\theta_{\min} = 0{,}00001;\ \theta_{\max} = 0{,}0003$
2. $s = \dfrac{0{,}00001 + 0{,}0003}{2} = 0{,}000155$
3. $\theta_0 = 0{,}00001$
4. $h_B = |0{,}00001 - 0{,}000155| = 0{,}000145$
5. $q = \dfrac{2 \cdot 5390866 \cdot 0{,}00001 \cdot (1 - 0{,}00001)}{5390866 \cdot 0{,}000145^2 - 0{,}00001 \cdot (1 - 0{,}0001)} = 951{,}3229$
6. $\alpha = 951{,}3229 \cdot 0{,}000155 = 0{,}147455$
   $\beta = 951{,}3229 - 951{,}3229 \cdot 0{,}000155 = 951{,}1755$

Table 1 Updated Bayesian estimation of critical illness probability for men less than 30 years

Year	n	x	x/n	α	β	θ_B
1999	103,5	0	0,000000	0,147	951,18	0,000155
2000	1054,2	0	0,000000	0,147	1054,74	0,000140
2001	4014,7	0	0,000000	0,147	2109,00	0,000070
2002	7671,3	6	0,000782	0,147	6123,78	0,000024
2003	11868,8	10	0,000843	6,147	13789,12	0,000446
2004	16393,0	9	0,000549	16,147	25647,98	0,000629
2005	21749,1	7	0,000322	25,147	42032,05	0,000598
2006	28121,6	9	0,000320	32,147	63774,23	0,000504
2007	34005,4	13	0,000382	41,147	91886,87	0,000448
2008	40944,8	10	0,000244	54,147	125879,36	0,000430
2009	48404,2	18	0,000372	64,147	166814,21	0,000384
2010	48766,7	27	0,000554	82,147	215200,42	0,000382
2011				109,14	263940,13	0,000413

Source: own calculations based on data with prior Beta(0,147; 951,18)

Results of posterior parameters α and β estimates according to formulas (7), (8) and Bayesian estimations of probabilities θ_B for men less than 30 years according to (9) based on data x, n from Slovak insurance companies for the years 1999 to 2011 there are in the Table 1.

Maximum likelihood estimates x/n and Bayesian estimates θ_B from Table 1 in successive years 1999-2011 is shown in Fig. 1. We can see that the Bayesian estimates are not so strong affected by randomness as a maximum likelihood estimates, because Bayesian estimates contain also a priori information from the previous years. Therefore the Bayesian estimates are more suitable for actuarial calculations in comparison with maximum likelihood estimates.

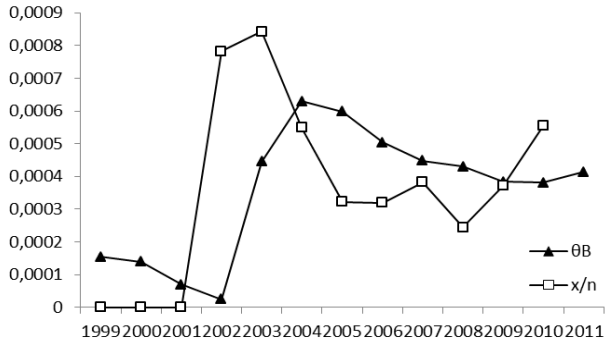


Fig. 1 Maximum likelihood and Bayesian estimations of critical illness probabilities for men less then or 30

Bayesian estimations obtained by the same procedure for men aged 31 to 60 years (θ_{B1}), for women less than or 30 years (θ_{B2}) and for women aged 31-60 (θ_{B3}) are presented in Table 2.

Table 2 Comparison of Bayesian estimations of critical illness probabilities for different groups of insured persons

Year	θ_{B1} -men 31-60	θ_{B2}-women less then or 30	θ_{B3} -women 31-60
1999	0,1255	0,1505	0,1255
2000	8,5E-05	4,6182E-06	0,00397096
2001	0,001373	0,00012933	0,00867027
2002	0,001906	0,00029244	0,00904263
2003	0,002198	0,00025284	0,00893824
2004	0,001889	0,00016749	0,00815181
2005	0,001936	0,00016684	0,00749277
2006	0,001826	0,00014043	0,00672558
2007	0,001776	0,0001335	0,00603558
2008	0,001683	0,00015397	0,00555998
2009	0,001562	0,00017739	0,00482478
2010	0,001778	0,0002267	0,00410486
2011	0,001814	0,00028119	0,00358628

Source: own calculations

Bayesian estimations of critical illness probabilities from Table 2 together with adequate maximum likelihood estimates show Fig. 2-4.

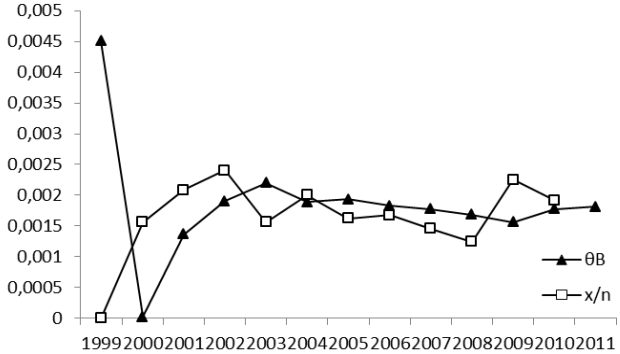


Fig. 2 Maximum likelihood and Bayesian estimations of critical illness probabilities for men in age 31-60

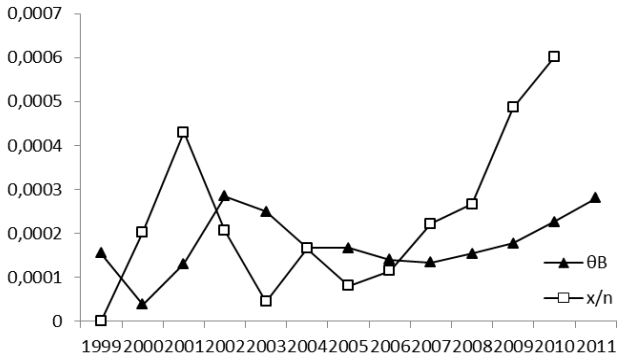


Fig. 3 Maximum likelihood and Bayesian estimations of critical illness probabilities for women less then or 30

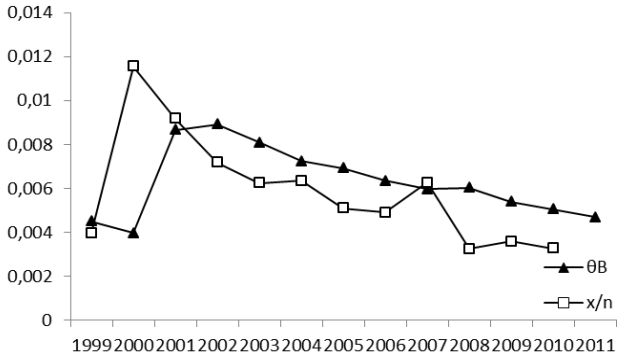


Fig. 4 Maximum likelihood and Bayesian estimations of critical illness probabilities for women in age 31-60

V. Conclusion

Bayesian estimation theory provides methods for permanently updated estimates of the event probability for each coming year in insurance company. Bayesian approach combine prior information that are known before collected of any data and information provided by the sample data, which are in our case number of concluded insurance contracts and number of claims in previous n years. Probabilities of the claims which are the subject of insurance contracts are necessary to know for insurance company especially when calculating premiums for next year.

The insurance company can correctly determine premiums only if use correctly estimates probabilities of claims. This article is both theoretical and practical demonstration of permanently updated Bayesian estimates of event probability which in this case is critical illness. This procedure has of course general use and provides better estimates of probabilities as maximum likelihood method.

The maximum likelihood estimate is assigned to period which has already expired, while Bayesian estimate of event probability is for next period. This is undoubtedly advantage for premium calculation. The possibility to express Bayesian estimate of binomial probability in the form of credibility formulas by expression (10) allow easy application of this theory in insurance practice.

The weakest point of Bayesian estimation is the choice of parameters of prior distribution and the associated a priori estimate of the parameter. Article also presents an algorithm to improve the a priori estimates.

References

[1] P. J. Boland, *Statistical and Probabilistic Methods in Actuarial Science*, London: Chapman&Hall/CRC, 2007.

[2] H. Bühlmann, Experience rating and credibility, *ASTIN Bull*. 4, 119, 1967.

[3] J. Gogola, Spôsob permanentnej úpravy výšky poistného v neživotnom poistení (Method for Permanent Adjustments of Premium in Non-Life Insurance). E+M *Economics and Management*, No. 4/2013, 2013, pp. 134-142.

[4] R. J. Gray, S. M. Pitts, *Risk Modelling in General Insurance*. Cambridge: Cambridge University Press, 2012, ch. 4.4.

[5] P. Jindrová, Quantification. of Risk in Critical Illness Insurance. In: Conference proceedings from 9th international scientific conference *Financial Management of Firms and Financial Institutions*, VŠB Ostrava, 2013. pp. 298-306.

[6] E. Kotlebová, I. Láska, Využitie bayesovského prístupu pri odhade podielu a možnosti jeho aplikácie v ekonomickej praxi (Use of Bayesian approach in estimating the proportion and its possible applications in the economic practice). *Slovenská štatistika a demografia* (*Slovak Statistics and Demography*), Volume 24, No 2, 2014, pp.3-17.

[7] Material of Munich Re Group: Critical Illness Insurance, 2001, pp. 56-58: http://www.munichre.com.

[8] National Bank of Slovakia (2012). Data Disclosure according to Directiv 2004/113/EC: Undrlying data_1999_2010.xls. Retrieved from: http://www.nbs.sk/en/financial-market-supervision/insurance-supervision/data-disclosure-according-to-directive-2004-113-ec.

[9] V. Pacáková, The Bayesian Inference in Actuarial Sciences, *Central European Journal for Operations Research and Economics*, Volume 5, Number 3-4, 1997, pp. 255-268.

[10] V. Pacáková, Bayesian Estimations in Insurance Theory and Practice, *Proceeding of the 14th WSEAS International Conference on Mathematical and Computational Methods in Science and Engineering* (MACMESE'12), Sliema, Malta, September 7-9, 2012, pp. 127-131.

[11] V. Pacáková, Credibility models for permanently updated estimates in insurance. *International Journal of Mathematical Models and Methods in Applied Sciences*, Issue 3, Volume 7, 2013, pp. 333-340.

[12] E. Šoltés, *Modely kredibility na výpočet poistného* (Models for calculating credibility premiums), Bratislava: (Publisher) Vydavateľstvo EKONÓM, 2009.

[13] Y. K. Tse, *Nonlife Actuarial Models*, Cambridge: University Press, 2009.

[14] H. R. Waters, *An Introduction to Credibility Theory*, London and Edinburgh: Institute of Actuaries and Faculty of Actuaries, 1994.

Mgr. Pavla Jindrová, Ph.D. graduated from Mathematical analysis at the Faculty of Science of Palacky University in Olomouc in 1993. She lectures mathematics, insurance and financial mathematics in branch of study of Insurance engineering. She finished her dissertation thesis which deals with problems of aggregation and disaggregation in economics and mathematics.

Prof. RNDr. Viera Pacáková, Ph.D. graduated in Econometrics (1970) at Comenius University in Bratislava, 1978 - RNDr. in Probability and Mathematical Statistics at the same university, degree Ph.D at University of Economics in Bratislava in 1986, associate prof. in Quantitative Methods in Economics in 1998 and professor in Econometrics and Operation Research at University of Economics in Bratislava in 2006.

She was working at Department of Statistics, Faculty of Economic Informatics, University of Economics in Bratislava since 1970 to January 2011. At the present she has been working at Faculty of Economics and Administration in Pardubice since 2005.

She has been concentrated on actuarial science and management of financial risks since 1994 in connection with the actuarial education in the Slovak and Czech Republic and she has been achieved considerable results in this area.

On a method of texture analysis

Natalia B. Ampilova, Igor P. Soloviev

***Abstract* —** A method of texture analysis based on the using Kullback-Leibler divergence is discussed. A digital image is described by a discrete probability distribution. We consider a group of direct multifractal transforms relating to the distribution and for two given images calculate a vector of Kullback-Leibler divergences between the initial distributions and their direct multifractal transforms. The method is demonstrated on the example of the Serpinsky carpet. It proves to be appropriate for textures from different classes but having similar structures and may be applied for classification of biomedical preparation images.

***Keywords* —** Direct multifractal transform, discrete probability distribution, Kullback-Leibler divergences, texture analysis.

I. Introduction

THE problem of analysis and classification of digital images having structural features (textures) is very important and actual. In the literature by textures are meant images having both regular and irregular structure. Hence to analyze such images one have to use a number of methods — statistical, fractal, multifractal, morphological, spectral ones. It is very often several methods should be applied to classify textural images. The detail survey and bibliography on this subject are given in [10].

Any method of textural analysis results in obtaining a numerical characteristic (or a feature set) that may be used as a classifying sign to refer the images under investigations to some classes.

Now it is widely accepted that textural images may be considered as phase portraits of complex dynamical systems. Hence, when analyzing these images in some points of time we can obtain a system characteristic. For dynamical systems various types of invariant sets are subjects of much interest. Considering an image as a portrait of a process one can find its stationary state. A method of a classification of images relating to a substance propagation process was proposed in [1]. The image is considered as a lattice formed by pixels of given intensities. Then an oriented graph corresponding to the image is constructed in the following way. Every vertex is connected with its neighbours, and all outcoming edges have the value of intensity divided on the number of neighbours. The constructed flow is normed. Hence we obtain Markov chain on the graph: for every vertex its weight equals sum of weights of outcoming edges. Denote the initial distribution on graph edges by p_{ij}. Our purpose is to find such a distribution u_{ij} that the flow on the graph be stationary: for every vertex the sum of weights of incoming edges equals the sum of weights of outcoming ones. It is well known that such a problem has a solution if there is a cycle on a graph. Moreover, this solution minimizes weighted entropy $g(u) = -\sum_{i,j} u_{ij} \ln {}^{p_{ij}}/_{u_{ij}}$. It is weighted entropy that is used as a classifying sign when images relating to different doses of a substance are analyzed. In fact, weighted entropy may be considered as a time that is required for a distribution process to achieve a stationary state. (It should be noted that weighted entropy is the Kullback-Leibler divergence.)

The structure of any digital image is defined by pixel intensities. For a given digital image it is reasonable to relate a measure distribution that describes densities of different parts of the image. With this object in view the image is partitioned on a set of cells by a given size, and for every cell its measure is calculated. In the simple case the measure of the cell is defined as the sum of pixel intensities, but sometimes various filters may be applied [11]. The obtained distribution is normed.

So, we match a discrete probability distribution to a given digital image. Such a distribution (a measure) is a basis of multifractal analysis. One can obtain the so called multifractal spectrum (MFS) [5] which is very important characteristic for images with complex structures. It may be obtained by using Regny dimensions and the following Legandre transform [11], or by the direct calculation [4, 2].

To obtain an image characteristic means to extract information described by the discrete probability distribution. In information theory the notion *information* was originally concerned with Shannon entropy. Later A. Regny introduced a set of entropies depending on a real parameter α, which goes into the Shannon entropy when α tends to unity. As it mentioned in [7], ”one of the fundamental observation of information theory is that the most general functional form for the mean transmitted information (i.e., information entropy) is that of Regny”. A. Bashkirov [3] also supposed to consider namely Regny entropy as statistical characteristic of complex systems. Moreover, he proved that for systems with power-series distribution Regny entropies for $\alpha < 1$ are mostly representative. So, for a digital image and concerning to it discrete probability distribution one can use a set of Regny

This work was supported in part by the Russian Foundation of Basic Research (RFBR) under Grant 13-01-00782.

N. B. Ampilova, Saint-Petersburg State University, St.Petersburg, 192504, Russia (corresponding author; e-mail: n.ampilova@spbu.ru).

I. P. Soloviev, Saint-Petersburg State University, St.Petersburg, 192504, (e-mail: i.soloviev@spbu.ru).

ISBN: 978-1-61804-251-4

entropies as the distribution characteristics.

But it should be marked that the images with different structures may have identical entropy characteristics, because the entropy does not depend on the order of component in a probability distribution vector. For example two Serpinsky carpets constructed by different methods but having the same number of black and white cells have the same entropy. (It is interesting to note that they have the same similarity dimensions).

In such cases it is reasonably to use Regny divergences, in particular the Kullback-Leibler divergence. But as experiments show, for many classes of biomedical preparations images the constructed distributions are close and KL divergence is small. Hence it would be difficult to use its value as a classification sign.

In practice it proves to be fruitful to consider not only the image, but also its additional modifications. For example, in [9] the authors applied *fractal signature* method to analyze textures. It is based on the construction of a sequence of special *blankets* [5] for the image and calculation a numerical characteristic, fractal signature, which is closely connected with the Minkovsky dimension. Every such a blanket corresponds to an image obtained from a given one by resolution changing. The vector of fractal signatures is the image characteristic, and the distance between two images may be defined as the Euclid distance between their vectors. In [2] the method was successfully applied to classify the biomedical preparations images of 4 classes.

We suppose to consider both a given distribution and its direct multifractal transform [11], which is a renormalization of the given measure. Such transforms form a group. For two given distributions we construct their direct multifractal transforms and calculate Kullback-Leibler divergences between both initial distributions and their renormalizations. The obtained divergence vectors may be considered for comparing given images. The method is demonstrated on the Serpinsky carpet.

II. Main Part

A. Definitions

Let a discrete probability distribution $p = \{p_i\}$, $p_i \geq 0, \sum p_i = 1,\ i = 1,...,n$ be given. Following [3] we define the Regny entropy of order α as

$$H_\alpha(p) = \frac{1}{1-\alpha} \ln \sum_1^n p_i^\alpha . \qquad (1)$$

It is known [7] that the entropy is nonincreasing function of α. When α tends to 1 the Regny entropy turns into Shannon entropy. The most generally used Regny entropies are

$H_0(p) = \ln n$ (Hartley entropy),

$H_1(p) = -\sum_1^n p_i \ln p_i$ (Shannon entropy),

$H_2(p) = -\ln \sum_1^n p_i^2$,

$H_\infty(p) = -\ln \max_i p_i$ (min-entropy). (2)

It is easy to note that entropy characteristics do not depend on the position component p_i in the distribution vector p. In other words, using them one cannot distinguish between two images with different structures but having the same number values p_k .

Let p and q be discrete probability distributions. Define Regny (or α -) divergences as

$$D_\alpha(p,q) = \frac{1}{\alpha - 1} \ln \sum_1^n p_i^\alpha q_i^{1-\alpha} . \qquad (3)$$

The Kullback-Leibler divergence (for α =1) is defined as

$$D_1(p,q) = \sum_1^n p_i \ln \frac{p_i}{q_i} . \qquad (4)$$

B. Serpinsky carpet

Consider the following example. Let us construct the Serpinsky carpet (the first step) by two ways. We take unit square, divide every side onto 7 parts and obtain 49 small squares. Then delete 9 small squares as shown on Fig.1.

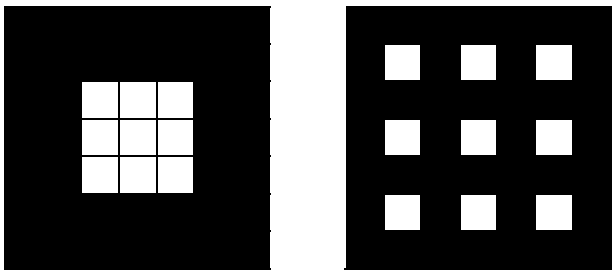


Fig. 1 Two types of the Serpinsky carpet (the first step of the construction)

We see that the obtained images have different structures. According to (1) entropy characteristics do not allow distinguishing between the images. In addition, as it was marked in [6], these images have the same similarity dimension, namely $D_s = \frac{\ln 40}{\ln 7}$.

On the second step on the Serpinsky carpet construction we have to repeat the described procedure for every of black squares. The results are shown on Fig.2 (Figure after [8].)

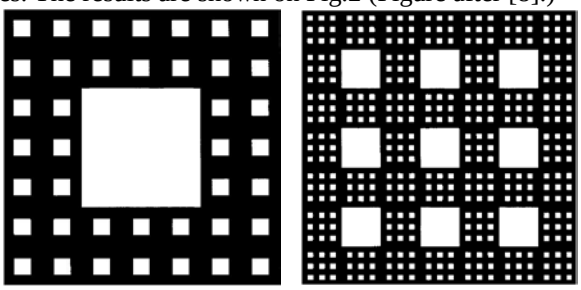


Fig. 2 Two types of the Serpinsky carpet (the second step of the construction)

In this case the distribution vector is the union of vectors constructed on the first step. Hence the entropy characteristics will be the same.

In what follows we need some denotations. Let us suppose that for Serpinsky carpet the normalized intensity (measure) of a small black square is b, and the measure of white square is w, so that $40b+9w=1$. We suppose that w is a small number and $b=wm$, where m is a real number.

To calculate Kullback-Leibler divergence for the example we have to find the sum of divergence between rows of the matrix constructed in accordance with images: every square is coded by b or w. Divergences for 1 and 7 rows are equal zero, because the rows are identical. Denoting the divergence in *k*th rows by $D_1^k(p,q)$ we have

$$D_1^2(p,q)=3b\ln m, D_1^3(p,q)=-3w\ln m,$$
$$D_1^4(p,q)=2(b-w)\ln m, D_1^5(p,q)=-3w\ln m,$$
$$D_1^6(p,q)=3b\ln m.$$

So,

$$D_1(p,q)=8(b-w)\ln m \qquad (5)$$

and two types of the Serpinsky carpet are different images.

C. Direct multifractal transform

Consider a distribution $p=\{p_i\}$ that defines some measure on the image and consider the following *direct multifractal transform* $f_k(p)=\dfrac{p_i^k}{\sum_i p_i^k}$ [11]. It is easy to understand that such transforms form a group, because

$$f_{k_1}\circ f_{k_2}=f_{k_1k_2},$$
$$f_k\circ Id=f_k,$$
$$f_k\circ f_{1/k}=Id,$$
$$Id=f_1.$$

Every transform results in a renormalization of the initial measure and we can consider a new image relating to the transform. Hence we analyze not only initial image, but a set of its modifications as well.

Denote the measure obtained by means of f_k as $p(k)$. In [11] the author considers the set of Kullback-Leibler divergences between the initial measure p and $p(k)$, namely

$$D_1(p,p(k))=(1-k)\sum_i p_i\ln p_i+\ln\sum_i p_i^k.$$

Hence the obtained vector is a characteristic of the image by means of the initial measure transform. However, for two types of the Serpinsky carpet these vectors turn to be the same.

D. Divergences between direct multifractal transforms

Consider the following method to compare distributions p and q. We construct multifractal transforms $p(k)$ and $q(k)$, and calculate

$$D_1(p(k),q(k))=\sum_i\frac{p_i^k}{\sum_i p_i^k}\ln\frac{p_i^k}{\sum_i p_i^k}\frac{\sum_i q_i^k}{q_i^k}. \qquad (6)$$

We see that for Serpinsky carpets $\sum_i p_i^k=\sum_i q_i^k$, because the images have the same number of black and white squares. By (5) for initial distribution we have

$$D_1(p,q)=8b(1-\frac{1}{m})\ln m.$$

In this case (6) has the form

$$D_1(p(k),q(k))=\frac{k}{\sum_i p_i^k}\sum_i p_i^k\ln\frac{p_i}{q_i}.$$

By using (5) we have

$$\sum_i p_i^k\ln\frac{p_i}{q_i}=8b^k(1-\frac{1}{m^k})\ln m,$$

and finally

$$D_1(p(k),q(k))=\frac{8k}{u_k}(1-\frac{1}{m^k})\ln m,$$

where $u_k=40+\dfrac{9}{m^k}$.

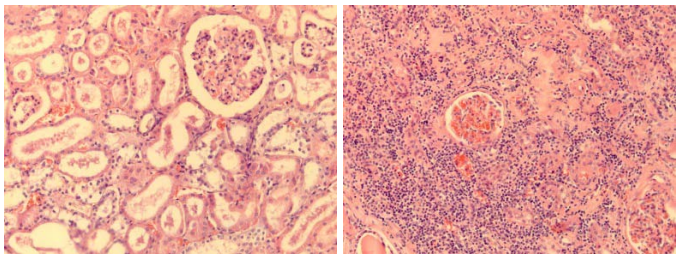
For k large enough

$$D_1(p(k),q(k))\approx\frac{k}{5}\ln m,$$

i.e. the divergence increases.

III. Numerical Experiments

Experiments were performed for two classes of biomedical preparations images: healthy kidney and kidney with chronic pyelonephritis. The size of the partition cell is 100x100 pixels. The order of the comparison: left picture is supposed to have the distribution p, and right picture – q. Divergences were computed for (p,q) and (q,p).



ISBN: 978-1-61804-251-4

a b

Fig. 3 Healthy kidney (a) and kidney with chronic pyelonephritis (b)

k	$D_1^k(p,q)$	$D_1^k(q,p)$
0	0,005971	0,003865
1	0,005180	0,005188
2	0,020604	0,020668
3	0,046085	0,046297
4	0,081406	0,081909
5	0,126324	0,127325
6	0,180562	0,182356
7	0,243810	0,246811
8	0,315723	0,320499
9	0,395925	0,403237
10	0,484009	0,494852

We note that the method is very sensitive to small difference in structures. On the Fig.4 the images belonging the class of healthy kidneys are given.

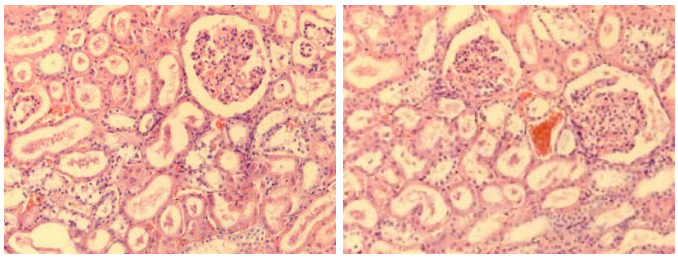


Fig. 4 Two images of healthy kidney

The divergence vectors for orders (p,q) and (q,p) show that divergence increases.

k	$D_1^k(p,q)$	$D_1^k(q,p)$
0	0,003484	0,006955
1	0,005483	0,005502
2	0,021725	0,021886
3	0,048428	0,048996
4	0,085299	0,086680
5	0,132032	0,134759
6	0,188299	0,193009
7	0,253744	0,261159
8	0,327973	0,338883
9	0,410558	0,425809
10	0,501036	0,521526

IV. Conclusion

The described method is a natural generalization of the Kullback-Leibler divergence application. By considering modifications of a given image we use additional information and obtain more grounded classification sign. Such a method may be successfully applied for analysis and classification of complex textures. At the same time the divergence increasing does not mean that the images are in different classes. Hence an additional methods should be applied.

Acknowledgment

Authors wish to express their thanks to postgraduate student of SPbGU V. Sergeev for numerical experiments.

References

[1] N.B. Ampilova, " Stationary processes on graphs and image analysis", *Computer instruments in education*, no.3, pp. 27-32, 2013 (in Russian).
[2] N. Ampilova, I. Soloviev, Y. Shupletzov," On some aspects of the fractal signature method", in *Proc. 8th Int. Conf. CEMA13*, Sofia, 2013, pp.80-84.
[3] A. Bashkirov A. G. "Regny entropy as a statistical entropy for complex systems", *Theoretical and Math. Physics*, 149(2), 2006, pp.299-317 (in Russian).
[4] A. Chhabra, C. Meneveau, R. Jensen and K.R. Sreenivasan, "Direct determination of the f(α) singularities spectrum and its application to fully developed turbulence", *Physical Review A*, vol. 40, no. 9, November 1,1989. pp. 5284-5294.
[5] K.J. Falconer, *Fractal Geometry. Mathematical Foundations and Applications.* John Wiley & Sons, 1990.
[6] D.L. Jaggard, A.D. Jaggard and P. Frangos, "Fractal Electrodynamics : Surfaces and Superlattices" , in *Frontiers in Electromagnetics*, D.H. Werner and Raj Mittra, Eds., IEEE Press, 2000, pp. 1-47.
[7] P. Jizba, T. Arimitsu, "The world according to Rényi: Thermodynamics of multifractal systems", *Annals of Physics* (Elsevier), 312,2004, pp. 17–59.
[8] B. B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, San Francisco, 1983.
[9] S. Peleg, J. Naor, R. Hartley, D. Avnir, "Multiple Resolution Texture Analysis and Classification", *IEEE transactions on pattern analysis and machine intelligence*, vol. PAMI-6, no.4, 1984. pp.518-523.
[10] M. Tuceryan, Anil K. Jain, Texture analysis, in *The Handbook of Pattern Recognition and Computer Vision (2nd Edition)*, by C. H. Chen, L. F. Pau,P. S. P. Wan, Eds., pp. 207-248, World Scientific Publishing Co., 1998.
[11] G. Vstovsky, *Elements of information physics.* Moscow.: MGIU, 2002.
[12] Y.Xu, H. Ji, C. Fermüller, " Viewpoint Invariant Texture Description Using Fractal Analysis", *International Journal of Computer Vision*, no. 83, 2009, pp. 85–100.

Knowledge representation in the category of unformalized decision-making problems

Lyudmila V. Borisova, Inna N. Nurutdinova, Valery P. Dimitrov

***Abstract* —** Some aspects of representing fuzzy expert knowledge in decision-making problems on technological adjustment of machines have been considered. The technique based on application of different criteria of conformity while representing fuzzy knowledge, including the ones with regard to different hierarchy of expert knowledge, has been suggested. The given technique makes it possible to determine a rational term-set of a linguistic variable while describing input and output attributes. A justified choice of terms of linguistic variables makes it possible to optimize base knowledge parameters based on fuzzy production laws.

***Keywords* —** Criteria of coherence, fuzzy expert knowledge, weighting coefficient.

I. Introduction

The most important component of intelligent decision support systems in the sphere of operating complex machines is a subsystem of knowledge acquisition and updating [1]. Expert data, as a rule, are difficult to formalize in terms of traditional mathematical approaches and this has caused application a theory of fuzzy sets and widespread use of knowledge bases founded on fuzzy production laws [2], [3]. A decision-making system operates with fuzzy knowledge and allows making conclusions on the basis of fuzzy logic rules and it makes the problem of adequate representation of fuzzy expert data actual. To form such data it is necessary to determine membership functions (MF) of linguistic variables (LV) of a domain model, as well as to determine an optimal number of LV terms. At the same time, the natural question is the question of criteria according to which the choice of an optimal set of values of the linguistic scale should be made when estimating this or that attribute.

The requirements for the models of expert estimation of attributes are formulated, as a rule, in terms of each specific problem, and the quality of the developed models depends on experience and skills of the researchers. The reason for this dependency is evidently not only the fact that the methods of formalization are limited by both the ways of obtaining data from experts and the type of data, but also the absence of common natural demands for MF used for formalizing the collection of fuzzy sets.

This work was supported by the Ministry of Education and Science of the Russian Federation under the project 1.12.12 – registration number 01201255338.

Lyudmila V. Borisova is with the Department of Economics and Management in Engineering, Don State Technical University, Rostov-on-Don, Russia (phone: +7(863)258-92-10, e-mail: borisovalv09@mail.ru).

Inna N. Nurutdinova is with the Department of Mathematics, Don State Technical University, Rostov-on-Don, Russia (phone: +7(863)258-91-45, e-mail: nurut.inna@yandex.ru).

Valery P. Dimitrov is with the Department of Quality Management, Don State Technical University, Rostov-on-Don, Russia (phone: +7(863)238-15-10, e-mail: kaf-qm@donstu.ru).

II. Problem Definition

The criterion for optimal choice of values (terms) of linguistic variables must meet the requirements of minimal uncertainty for experts when describing input and output factors and maximal conformity of expert data. From the practical point of view this problem is reduced to establishing an optimal set of the linguistic scale used for estimating factors of the model and the optimal number of LV terms. Determination of basic and extended term-sets is a key moment when forming a MF. In the general case a basic LV term-set has the form [3], [4]

$T_i = \{T_1^i, T_2^i, \ldots T_m^i\}$, ($i \in K = \{1, 2, \ldots, m\}$).

Here $\langle T_i, X; \tilde{C}_i \rangle$ is a fuzzy variable corresponding the term $T_i \in T$; C_i is the carrier of the fuzzy set $\tilde{C}_i$. LV terms are determined on the real axis R (according to physical meaning of the variable). On top the number of terms is limited for reasons of measuring accuracy of the factor under consideration. And the lower bound must be such one that it would be possible to recognize and describe interaction of the stated input factor with output factors. When solving the stated problem it is necessary to estimate conformity of fuzzy expert knowledge.

We will consider normal fuzzy sets A for which the height equals 1, i.e. upper bound of its membership function is equal to 1 ($\sup_{x \in E} \mu_A(x) = 1$). Fuzzy sets may be both unimodal, i.e. $\mu_A(x) = 1$ only on one x of E, and have domain of tolerance.

To choose an optimal model as a criterion of conformity the indices of general and pair conformity can be used [5], [6]. In the first stage additive and multiplicative indices of general conformity are usually used, and according to their values a conclusion about the conformity of the studied models is stated. In the next stage the analysis of matrix of pair conformity of the models X_i and X_j of experts is carried on.

General conformity of a set of models of expert estimation of an attribute is determined by additive *k* and multiplicative $\tilde{k}$ indices:

ISBN: 978-1-61804-251-4

$$k = \frac{1}{m}\sum_{l=1}^{m}\frac{\int_0^1 \min_{\forall i=1,2,...,k}\mu_{il}(x)dx}{\int_0^1 \max_{\forall i=1,2,...,k}\mu_{il}(x)dx} \qquad \tilde{k} = \sqrt[m]{\prod_{l=1}^{m}\frac{\int_0^1 \min_{\forall i=1,2,...,k}\mu_{il}(x)dx}{\int_0^1 \max_{\forall i=1,2,...,k}\mu_{il}(x)dx}} \tag{1}$$

where $l = 1,2,...,m$ – is the term number, $i = 1,2,...,k$ – is an expert number, $\mu_{il}(x)$ – is MF, which was defined by the *i*-st expert for the *l*-st term.

The indices of conformity k_{ij} between the models of two experts, *i* – th and *j* – th, in terms of *l* – th term are the elements of matrix K_m^{ij} of pair conformity of the models X_i and X_j of experts:

$$k_{ij} = \frac{\int_0^1 \min[\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)]dx}{\int_0^1 \max[\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)]dx} \tag{2}$$

The elements of models conformity matrix for all terms are determined by the formula:

$$\tilde{k}_{ij} = \frac{1}{m}\sum_{l=1}^{m}\tilde{k}_{ij}^{l} \tag{3}$$

where *m* is the number of terms.

The choice of a rational term-set of MF is defined as a result of analysis of additive and multiplicative indices and also pair conformity matrices for the models with different number of terms.

As an alternative approach to estimation of expert data conformity we can use a minimization method of weighted mean quadratic deviation F_m of the parameters estimated by experts, from average values of these parameters:

$$F_m = \sum_{l=1}^{m}\sum_{i=1}^{k}\omega_i\sum_{j=1}^{4}\left(a_j^{il} - a_j^{l}\right)^2 \to \min, \tag{4}$$

where a_1^{il} and a_2^{il} are tolerance limits of a fuzzy number $\mu_{il}(x)$, a_3^{il} and a_4^{il} are its left and right coefficients of fuzziness accordingly, a_j^{l} – their averaged values, ω_i – measuring coefficients of experts.

From the necessary condition of the function extremum F_m we obtain:

$$a_j^{l} = \sum_{i=1}^{k}\omega_i a_j^{il}. \tag{5}$$

With the given weight coefficients and constant number and composition of experts F_m depends only on the number of the model terms.

Thus formed expert data serves for obtaining a generalized MF which is then applied in the mechanism of fuzzy logical derivation. An adequacy of the obtained solutions is provided, to a great extent, by maximal conformity of the expert data. In the given approach there appears a question as to the choice of weight coefficients that is not a trivial one. As an initial approximation for solving application problems it is commonly accepted to use equal weight coefficients for all experts that is natural only with the same qualification of experts. It is this approach that is applied in the suggested and implemented software system for entering and updating expert knowledge [7]. However, expert assessments are based not only on their qualification, that is often different, but also on the use of indirect means of objective control of different accuracy. The necessity of implementing different weighting coefficients of experts is obvious. In the present paper we suggest to apply the numbers of Fishburn for calculating weighting coefficients [8]. Application of the rule of Fishburn will make it possible to take into account a significance level of estimation by experts, for this we introduce ranks r_i of experts and establish a relation $r_1 \geq r_2 \geq ... \geq r_k$. The collection of Fishburn weights for the system of strict preferences is determined by the formula:

$$\omega_i = \frac{2(N - i + 1)}{N(N+1)}, \tag{6}$$

where *N* is the number of experts, *i* – is the number of an expert by significance.

For the mixed system of preferences, when along with preferences the system incorporates indifference ratios, weighting coefficients of Fishburn have the form:

$$\omega_i = \frac{a_i}{b}, \tag{7}$$

where $a_{i-1} = \begin{Bmatrix} a_i, & \text{if } r_{i-1} \approx r_i \\ a_{i+1}, & \text{if } r_{i-1} > r_i \end{Bmatrix}$, $r_N = 1$, $i = N, \ldots, 2$; $b = \sum_{i=1}^{N} a_i$.

Different considerations can be used for experts ranking, for example, as in the present paper, the degree of conformity of their data with that of others.

III. Problem Solution

For the purposes of illustration of the suggested technique for estimating conformity of expert data we will consider a subject domain "Combine-harvesting of grain crops" [9]. By virtue of formalism of the applied body of mathematics this choice doesn't limit communities of consideration. A priori analysis of the subject domain has shown that for considering the question of choice of optimal set of the linguistic scale used for estimating factors of external environment, adjustable parameters of the machine and quality index of operation, the performance of the analysis of expert data conformity is expedient.

Let us consider the problem of representation of 3 LVs: "stand of grain humidity" in the form of 2-term, 3-term and 4-term models, "grain humidity" in the form of 2-term, 3-term and 4-term models, "grain dockage" in the form of 3-term, 4-term and 5-term models.

The estimations of MF for the first two LVs were given by four experts: for 2 terms ("dry", "humid"), 3 terms ("dry", "normal", "humid") and 4 terms ("dry", "normal", "humid", "very humid"). The estimations of MF for the LV of "grain dockage" were given by 5 experts: for 3 terms ("low", "average", "high"), 4 terms ("low", "average", "high", "very high"), and 5 terms ("very low", "low", "average", "high", "very high"). For description of the terms typical trapezoidal functions have been applied [10], and definitional domain for x is from 0 to 1 (normalized values), but definitional domain for coefficients of MF equation: a, b, c, d – from 0 to1. For calculations the values of MF coefficients have been chosen,

ISBN: 978-1-61804-251-4

which are presented in [11].

Additive k, multiplicative $\tilde{k}$ indices and matrices of pair conformity (formulas (1) – (3)) have been determined with help of the software subsystem of acquiring knowledge of the expert system [7], [12]. The results of calculations are given in table I.

Table I – Values of k and $\tilde{k}$ indices

LV	Model	k	$\tilde{k}$
"stand of grain humidity"	2-term	0,8	0,8
	3-term	0,781	0,778
	4-term	0,689	0,684
"grain humidity"	2-term	0,807	0,806
	3-term	0,699	0,677
	4-term	0,537	0,517
"grain dockage"	3-term	0,572	0,562
	4-term	0,479	0,466
	5-term	0,469	0,436

As a result of calculations we can make a conclusion that the most conformable model for the first two LVs is a 2-term model, and for the third LV – 3-term model.

The results of calculations of matrices of pair conformity for all the models make it possible to calculate weighted-average quadratic deviations F_m of the parameters, estimated by experts, from the averaged values of these parameters. With this we use both equal weight coefficients and those calculated according to the rule of Fishburn. Table II presents one of the matrices of pair coherence for the 3-term model of the LV "grain dockage".

Table II - Matrix of pair conformity for the 3-term model of the LV "grain dockage"

Matrix					Expert rank	Weight
1	0,741	0,9	0,9	0,741	1	1/3
0,741	1	0,667	0,667	0.964	3	1/9
0,9	0,667	1	0,946	0,667	2	2/9
0,9	0,667	0,946	1	0,667	2	2/9
0,741	0,964	0,667	0,667	1	3	1/9

The experts ranking was carried out on the basis of the criterion of the greatest pair conformity, for this we used sums of matrix lines elements. The parameters of the generalized MF and the values F_m from the condition (5) for all the models have been calculated. The results of calculations of the value F_m are presented in table III.

Table III – F_m values

LV	Model	F_m	
		Equal weight coef.	Fishburn weight coef.
"stand of grain humidity"	2-term	0,01625	0,01195
	3-term	0,009063	0,0074
	4-term	0,011875	0.01035
"grain humidity"	2-term	0,007813	0,00525
	3-term	0,015469	0,00975
	4-term	0,021875	0,01535
"grain dockage"	3-term	0,005664	0,004968
	4-term	0,006864	0,005436
	5-term	0,005744	0,003155

In table IV we have compared the results of definition of the optimal number of LV terms which were obtained with help of conformity index analysis and applying the minimization method of weighted mean quadratic deviation F_m of individual parameters, given by experts, from the averaged values of these parameters.

Table IV – Optimal models for LV

LV	Optimal number of terms		
	By conformity indices	By F_m values	
		Equal weight coef.	Fishburne weight coef.
"stand of grain humidity"	2	3	3
"grain humidity"	2	2	2
"grain dockage"	3	3	5

The data of tables III and IV allow us to make a conclusion that for developing a generalized MF with the purpose of maximal conformity of expert data it is more preferable to apply Fishburn weight coefficients with experts ranking according to the degree of conformity of their data with those of others. It is also obvious that under the conditions requiring greatest conformity of expert assessments base the application of Fishburn weight coefficients gives the best results. We should note that the suggested method is relevant for other variants of ranking expert data, for example, by level of experts qualification.

IV. CONCLUSION

We have considered one of the stages of forming knowledge base of the expert system (fuzzification stage), including analysis of fuzzy expert data on the basis of criteria of conformity. In order to achieve maximal proximity of the formalized data of the real situation it is suggested to introduce different weight coefficients for the experts, and Fishburn numbers have been applied for their ranking. Examples from the subject domain "Combine-harvesting of grain crops" have been presented, several LVs with 2-, 3-, 4-, and 5-term MF and with different number of experts have been considered. The characteristics of general and pair conformity of expert models, parameters of generalized MF in case of equal weight coefficients and Fishburn weights have been calculated. The method of choosing an optimal model used in next stages (composition and defuzzification) of developing mechanism for displaying solutions have been illustrated. As a result of application of the studied procedure the basic term-sets have been determined and membership functions for 12 factors of external environment and 8 adjustable parameters of a grain harvester have been defined.

REFERENCES

[1] D. Waterman, "Expert systems manual: transl. from English, Moscow, Mir, 1989, p. 388.

[2] L. A. Zadeh, "Fuzzy sets" (Fuzzy sets and systems), 1965, № 8, pp. 338–353.

[3] A. N. Averkin, I. Z. Batyrshin, A. F. Blishun, V. B. Silov, V. B. Tarasov. Under the editorship of D. A. Pospelov, "Fuzzy sets in the models of management and artificial intelligence", Moscow, Nauka, 1986, p. 312.

[4] A. N. Borisov, A. V. Alekseev, G. V. Merkuriev and others, “Processing of fuzzy data in the systems of decision-making”, Moscow, Radio and communication, 1989, p.304.
[5] V. Y. Pivkin, Y. P. Bakulin, D. I. Korenkov, “Fuzzy sets in the systems of management”. Under the editorship of prof. Y. I. Zolotukhin,
Available: http:/www.idisys.iae.nsk.su/fuzzy_book/content.html
[6] V. P. Dimitrov, L.V. Borisova, I. N. Nurutdinova, ”The methods for evaluating conformity of models of fuzzy expert knowledge”, Rostov-on-Don, DSTU, 2010, Vol. 10, #2(45), pp. 205–216.
[7] V. P. Dimitrov, L.V. Borisova, P. V. Aleksandrov, I. N. Nurutdinova, Certificate of state registration of software # 2011611212, “Evaluation of conformity of fuzzy expert knowledge”, # 2010617854, of 04.02.11.
[8] A. Nedosekin, “Fuzzy Financial Management”, Russia, Moscow, AFA Library, 2003, p. 183.
[9] L. V. Borisova, V. P. Dimitrov, K. L. Khubiyan, “Mechanical systems of the products of JSC “Rostselmash” for the families Don-680, CK-5M-1, Don-1500B: Design, technical maintenance, adjustment and malfunction diagnostics”, Rostov-on-Don, 2003, p. 116.
[10] L. Kofman, “Introduction in the theory of fuzzy sets”, Moscow, Radio and communication, 1982, p. 432.
[11] V. P. Dimitrov, L. V. Borisova, “Formalization of fuzzy expert knowledge during linguistic description of technical systems” Rostov-on-Don, DSTU, 2011, p. 209
[12] V.P. Dimitrov, L.V. Borisova, I.N. Nurutdinova, E.V. Bogatyreva, “Software system for inputting expert knowledge”, Rostov-on-Don, DSTU, Vol. 1111 #1 (52), 2011, pp. 83–90.

Methods to Choosing Subcontexts in Good Maximally Redundant Tests Inferring

Xenia Naidenova, Vladimir Parkhomenko

Abstract—Good maximally redundant classification tests (GMRTs) inferring is considered. A decomposition of classification contexts into two kinds of subcontexts is defined: into attributive (or value) and object ones. Two methods of forming and reducing subcontexts are given. Various possibilities of constructing algorithms for GMRTs inferring with the use of both kinds of subcontexts are considered depending on the nature of GMRTs features. The results obtained are supplied with examples.

Keywords—Good classification test, Galois lattice, decomposition, implications, functional dependencies, subcontexts.

I. Introduction

THE paper presents a development of the Good Test Analysis (GTA) as a Machine Learning method based on inferring Functional Dependencies and Implications as well as on Plausible Reasoning. It deals with the formation of the best descriptions of a given object class (positive objects) against the objects which do not belong to this class (negative objects) on the base of lattice theory. Assume that objects are described in terms of values of a given set U of attributes; see an example of object descriptions in Tab.I. The principle concept of GTA is the concept of classification. To give a target classification of objects, we use an additional attribute $\mathrm{KL} \notin U$. A target attribute partitions a given set of objects into disjoint classes the number of which is equal to the number of values of this attribute. In Tab.I, we have two classes: the objects in description of which the target value k appears and all the other objects.

Let M be the set of attribute values such that $M = \{\cup \mathrm{dom}(\mathrm{attr}), \mathrm{attr} \in U\}$, where $\mathrm{dom}(\mathrm{attr})$ is the set of all values of attr. Let G be the set of objects; $G = G_+ \cup G_-$, where G_+ and G_- are the sets of positive and negative objects respectively. Denote by $P(B), B \subseteq M$, the set of all the objects in description of which B appears. $P(B)$ is called the interpretation of B in the power set 2^G. If $P(B)$ contains only G_+ objects and the number of these objects more than 2, then we call B a description of some positive objects or a **test** for G_+ [1].

Note that the definition of (maximally redundant) tests for G_+ does not differ from the definition of positive hypotheses given in [2]. Let us recall the definition of a good test or good description for a subset of G_+ (via partitions of objects). A subset $B \subseteq M$ of attribute values is a **good test** for a subset of positive objects if it is a test and no such subset $C \subseteq M$ exists, so that $P(B) \subset P(C) \subseteq G_+$.

For the first time, the definitions of good tests (maximally redundant and irredundant ones) have been given in 1986 [3]. In [3] an algorithm of inferring all good functional dependencies (as good descriptions of object classifications) from the many-valued data has been advanced.

TABLE I
EXAMPLE OF CLASSIFICATION

No	Height	Color of Hair	Color of Eyes	KL
1	Low	Blond	Blue	$k(+)$
2	Low	Brown	Blue	not $k(+)$
3	Tall	Brown	Hazel	not $k(+)$
4	Tall	Blond	Hazel	not $k(+)$
5	Tall	Brown	Blue	not $k(+)$
6	Low	Blond	Hazel	not $k(+)$
7	Tall	Red	Blue	$k(+)$
8	Tall	Blond	Blue	$k(+)$

The equivalence of diagnostic tests and FDs has been proven in 1982 [4] with the use of a partition model of tests based on the partition lattice. An algorithm of inferring FDs (as good classification tests) is given in [5], where it is shown that the task of inferring good classification tests covers inferring both functional and implicative dependencies. In what follows, we should deal with searching for good descriptions of positive or negative classes of objects. For this goal, we use the initial set of object descriptions with a partition of this set into two disjoint classes as follows: for a given value $v \in M$, the positive class is the set of objects in descriptions of which value v appears and the negative class is the set of all other objects. We consider the set of object descriptions with the partition of it into two classes as an initial classification context for the task of GMRTs inferring. In [5], it is proven that this problem is reduced to searching for causal dependencies in the form $B \to v$, $B \subseteq M$ and $v \notin B$.

Sec.II is devoted to defining a concept of good diagnostic (classification) test. Sec.III gives the decomposition of good tests inferring based on two kinds of subcontexts of initial classification context. Sec.IV is devoted to an analysis of algorithms based on using subcontexts including the evaluation of the number of sub-problems to be solved, the depth of recursion, the structure of sub-problems and their ordering, and some others.

II. The Concept of Good Classification Test

Let $G = \overline{1, N}$ be the set of objects indices and $M = \{m_1, m_2, \ldots, m_j, \ldots m_m\}$ be the set of attributes values (objects and values respectively). Each object is described by a set of values from M. The object descriptions are represented by rows of a table the columns of which are associated with the attributes taking their values in M.

Assume $A \subseteq G$, $B \subseteq M$. Denote by B_i, $B_i \subseteq M$, $i = \overline{1, N}$ the description of object with index i. The Galois

ISBN: 978-1-61804-251-4

connection between the ordered sets $(2^G, \subseteq)$ and $(2^M, \subseteq)$ is defined by the following mappings called derivation operators: for $A \subseteq G$ and $B \subseteq M$, $A' = \mathrm{val}(A) = \{$intersection of all $B_i |\ B_i \subseteq M, i \in A\}$ and $B' = \mathrm{obj}(B) = \{i |\ i \in G, B \subseteq B_i\}$. Of course, we have $\mathrm{obj}(B) = \{$intersection of all $\mathrm{obj}(m) |\ \mathrm{obj}(m) \subseteq G, m \in B\}$.

It is worth noticing that, for defining these operators, we do not use any procedure to transform many-valued context to two-valued one [6]. We introduce two generalization operations that are closure operators [7]: $\mathrm{generalization\_of}(B) = B'' = \mathrm{val}(\mathrm{obj}(B))$ and $\mathrm{generalization\_of}(A) = A'' = \mathrm{obj}(\mathrm{val}(A))$. A set A is closed if $A = \mathrm{obj}(\mathrm{val}(A))$. A set B is closed if $B = \mathrm{val}(\mathrm{obj}(B))$. For $g \in G$ and $m \in M$, $\{g\}'$ is denoted by g' and called object intent, and $\{m\}'$ is denoted by m' and called value extent. Let us recall the main definitions of GTA [3].

A Diagnostic Test (DT) for the positive examples G_+ is a pair (A, B) such that $B \subseteq M$, $A = B' \neq \emptyset$, $A \subseteq G_+$, $B \not\subseteq g'\ \forall g \in G_-$. A diagnostic test (A, B) for G_+ is **irredundant** if any narrowing $B_* = B \setminus m$, $m \notin B$ implies that $(obj(B_*), B_*))$ is not a test for G_+. A diagnostic test (A, B) for G_+ is **maximally redundant** if $obj(B \cup m) \subset A$ for all $m \notin B$ and $m \in M$. A diagnostic test (A, B) for G_+ is **good** if and only if any extension $A_* = A \cup i$, $i \notin A$,$i \in G_+$ implies that $(A_*, val(A_*))$ is not a test for G_+.

In the paper, we deal with GMRTs. If a good test (A, B) for G_+ is maximally redundant, then any extension $B_* = B \cup m$, $m \notin B$, $m \in M$ implies that $(\mathrm{obj}(B_*), B_*)$ is not a good test for G_+. Any object description d of $g \in G$in a given classification context is a maximally redundant set of values because for any value $m \notin d$, $m \in M, \mathrm{obj}(d \cup m)$ is equal to $\emptyset$.

In Tab.I, $((1, 8)$, Blond Blue) is a GMRT for $k(+)$ but it is irredundant one, simultaneously; $((4, 6)$, Blond Hazel) is a DT for $k(-)$ but it is not a good one; and $((3, 4, 6)$, Hazel) is a good irredundant test for $k(-)$.

III. The Decomposition of Good Maximally Redundant Test Inferring into Subtasks

There are two possible kinds of subtasks of GMRTs Inferring for a set G_+ [8]:

1) given a set of values $B \subseteq M$, $\mathrm{obj}(B) \neq \emptyset$, B is not included in any description of negative object, find all GMRTs $(\mathrm{obj}(B_*), B_*)$ such that $B_* \subset B$;
2) given a non-empty set of values $X \subseteq M$ such that $(\mathrm{obj}(X), X)$ is not a test for positive objects, find all GMRTs $(\mathrm{obj}(Y), Y)$ such that $X \subset Y$.

For solving these subtasks we need only to form subcontexts of a given classification context. The first subtask is useful to find all GMRTs intents of which are contained in the description d of an object g. This subtask is considered in [9] for fast incremental concept formation, where the definition of subcontexts is given.

We introduce **projection of a positive object description** t on the set D_+, i.e. descriptions of all positive objects. The $\mathrm{proj}(t)$ is $Z = \{z |\ z = t \cap t_* \neq \emptyset, t_* \in D_+$ and $(\mathrm{obj}(z), z)$ is a test for $G_+\}$.

We also introduce a concept of **value projection** $\mathrm{proj}(m)$ of a given value m on a given set D_+. The projection is $\mathrm{proj}(m) = \{t |\ m$ appears in $t,\ t \in D_+\}$ or $\mathrm{proj}(m) = \{t$ for $g_* |\ g_* \in (\mathrm{obj}(m) \cap G_+)\}$.

Algorithm ASTRA, based on value projections, has been advanced in [10]. Algoritm DIAGaRa, based on object projections, has been proposed in [11]. A family of fast algorithms for constructing a Concept (Galois) Lattice are advanced in [12] to compute intersections $\{\mathrm{obj}(X) \cap \mathrm{obj}(A), X \subseteq M, A \in M\}$. In what follows, we are interested in using both kinds of subcontexts for inferring all GMRTs for a positive (or negative) class of objects. The following **theorem** gives the foundation of reducing subcontexts [10]. Let $X \subseteq M$, $(\mathrm{obj}(X), X)$ be a maximally redundant test for positive objects and $\mathrm{obj}(m) \subseteq \mathrm{obj}(X)$, $m \in M$. Then m can not belong to any GMRT for positive objects different from $(obj(X), X)$.

Consider some example of reducing subcontext (see, please, Tab.I). Let $\mathrm{splus}(m)$ be $\mathrm{obj}(m) \cap G_+$ or $\mathrm{obj}(m) \cap G_-$ and SPLUS be $\{\mathrm{splus}(m) |\ m \in M\}$. In Tab.I, we have for values "Hazel, Brown, Tall, Blue, Blond, and Low" respectively $\mathrm{SPLUS} = \mathrm{obj}(m) \cap G_- = \{\{3, 4, 6\}, \{2, 3, 5\}, \{3, 4, 5\}, \{2, 5\}, \{4, 6\}, \{2, 6\}\}$.

We have $\mathrm{val}(\mathrm{obj}(\text{Hazel})) = $ Hazel, hence $((3, 4, 6), \text{Hazel})$ is a test for G_-. Then value "Blond" can be deleted from consideration, because $\mathrm{splus}(\text{Blond}) \subset \mathrm{splus}(\text{Hazel})$. Delete values Blond and Hazel from consideration. After that the description of object 4 is included in the description of object 8 of G_+ and the description of object 6 is included in the description of object 1 of G_+. Delete objects 4 and 6. Then for values "Brown, Tall, Blue, and Low" respectively $\mathrm{SPLUS} = \{\{2, 3, 5\}, \{3, 5\}, \{2, 5\}, \{2\}\}$. Now we have $\mathrm{val}(\mathrm{obj}(\text{Brown})) = $ Brown and $((2, 3, 5), \text{Brown})$ is a test for G_-. All values are deleted and all GMRTs for G_- have been obtained.

The initial information for finding all the GMRTs contained in a positive object description is the projection of it on current set D_+. It is essential that the projection is a subset of object descriptions defined on a certain restricted subset t_* of values. Let s_* be the subset of indices of objects the descriptions of which produce the projection. In the projection, $\mathrm{splus}(m) = \mathrm{obj}(m) \cap s_*,\ m \in t_*$.

It is useful to introduce the characteristic $W(t)$ of any collection t of values named by the weight of t in the projection: $W(t) = ||\mathrm{obj}(t) \cap s_*|| = ||\mathrm{splus}(t)||$ is the number of positive objects of the projection containing t. Let W_{min} be the minimal permissible value of weight. We assume that $W_{min} = 1$ in the paper.

Let STGOOD be the partially ordered set of elements s satisfying the condition that $(s, \mathrm{val}(s))$ is a good test for D_+. The basic recursive procedure for solving any kind of subtask consists of the following steps:

1) Check whether $(s_*, \mathrm{val}(s_*)$ is a test and if so, then s_* is stored in STGOOD if s_* corresponds to a good test at the current step; in this case, the subtask is over. Otherwise the next step is performed.
2) For each value m in the projection, the weight $W(m)$ is determined and if the weight is less than W_{min}, then

ISBN: 978-1-61804-251-4

the value m is deleted from the projection. We can also delete the value m if $W(m)$ is equal to W_{min} and $(\mathrm{splus}(m), \mathrm{val}(\mathrm{splus}(m)))$ is not a test — in this case m cannot appear in any GMRT satisfying W_{min} (we use the function $\mathrm{to\_be\_test}(t)$: if $\mathrm{obj}(t) \cap \mathrm{splus}(t) = \mathrm{obj}(t)$ $(\mathrm{obj}(t) \subseteq s_*)$ then "true" else "false").

3) The value m can be deleted from the projection if $\mathrm{splus}(m) \subseteq s$ for some $s \in \mathrm{STGOOD}$.
4) For each value m in the projection, check whether $(\mathrm{splus}(m), \mathrm{val}(\mathrm{splus}(m)))$ is a test and if so, then value m is deleted from the projection and $\mathrm{splus}(m)$ is stored in STGOOD if it corresponds to a good test at the current step.
5) If at least one value has been deleted from the projection, then the reduction of the projection is necessary. The reduction consists in checking, for each element t of the projection, whether $(\mathrm{obj}(t), t)$ is not a test (as a result of previous eliminating values) and if so, this element is deleted from the projection. If, under reduction, at least one element has been deleted, then Step 2, Step 3, Step 4, and Step 5 are repeated.
6) Check whether the subtask is over or not. The subtask is over when either the projection is empty or the intersection of all elements of the projection corresponds to a test (see, please, Step 1). If the subtask is not over, then the choice of an object (value) in this projection is selected and the new subtask is formed. The new subsets s_* and t_* are constructed and the basic algorithm runs recursively.

Forming the Set STGOOD. Let 2^S be the set of all subsets of the set S. 2^S is the set lattice [13]. The ordering determined in the set lattice coincides with the set-theoretical inclusion. Subset s_1 is absorbed by subset s_2, i.e. $s_1 \leq s_2$, if and only if the inclusion relation is hold between them, that is $s_1 \subseteq s_2$. Under formation of STGOOD, a collection s of object indices is stored in STGOOD if and only if it is not absorbed by any element of this set. It is necessary also to delete from STGOOD all the elements that are absorbed by s if s is stored in STGOOD. Thus, when the algorithm is over, the set STGOOD contains all the collections of objects that correspond to GMRTs and only such collections. The algorithm is based on topological sorting of partially ordered sets. The set TGOOD of all the GMRTs is obtained as follows: $\mathrm{TGOOD} = \{\mathrm{tg} | \ \mathrm{tg} = (s, \mathrm{val}(s)),\ s \in \mathrm{STGOOD}\}$.

IV. Selecting and Ordering subcontexts and GMRTs Inferring

Algorithms of GMRTs inferring are constructed by the rules of selecting and ordering subcontexts of the main classification context. Before entering into the details, let us recall some extra definitions. Let t be a set of values such that $(\mathrm{obj}(t), t)$ is a test for G_+. We say that **the value** $m \in M, m \in t$ **is essential** in t if $(\mathrm{obj}(t \setminus m), (t \setminus m))$ is not a test for a given set of object. Generally, we are interested in finding the maximal subset $\mathrm{sbmax}(t) \subset t$ such that $(\mathrm{obj}(t), t)$ is a test but $(\mathrm{obj}(\mathrm{sbmax}(t)), \mathrm{sbmax}(t))$ is not a test for a given set of positive objects. Then $\mathrm{sbmin}(t) = t \setminus \mathrm{sbmax}(t)$ is a minimal set of essential values in t. Let $s \subseteq G_+$, assume also that $(s, \mathrm{val}(s))$ is not a test.

The object t_j, $j \in s$ **is said to be an essential** in s if $(s \setminus j, \mathrm{val}(s \setminus j))$ proves to be a test for a given set of positive objects. Generally, we are also interested in finding the maximal subset $\mathrm{sbmax}(s) \subset s$ such that $(s, \mathrm{val}(s))$ is not a test but $(\mathrm{sbmax}(s), \mathrm{val}(\mathrm{sbmax}(s)))$ is a test for a given set of positive objects. Then $\mathrm{sbmin}(s) = s \setminus \mathrm{sbmax}(s)$ is a minimal set of essential objects in s.

An Approach for Searching for Initial Content of STGOOD. In the beginning of inferring GMRTs, the set STGOOD is empty. Next we describe the procedure to obtain an initial content of it. This procedure extracts a quasi-maximal subset $s_* \subseteq G_+$ which is the extent of a test for G_+ (maybe not good).

We begin with the first index i_1 of s_*, then we take the next index i_2 of s_* and evaluate the function $\mathrm{to\_be\_test}(\{i_1, i_2\}, \mathrm{val}(\{i_1, i_2\}))$. If the value of the function is true, then we take the next index i_3 of s_* and evaluate the function $\mathrm{to\_be\_test}(\{i_1, i_2, i_3\}, \mathrm{val}(\{i_1, i_2, i_3\}))$. If the value of the function is false, then the index i_2 of s_* is skipped and the function $\mathrm{to\_be\_test}(\{i_1, i_3\}, \mathrm{val}(\{i_1, i_3\})))$ is evaluated. We continue this process until we achieve the last index of s_*.

The complexity of this procedure is evaluated as the production of $||s_*||$ by the complexity of the function $\mathrm{to\_be\_test}()$. To obtain the initial content of STGOOD, we use the set $\mathrm{SPLUS} = \{\mathrm{splus}(m) | m \in M\}$ and apply the procedure described above to each element of SPLUS. To illustrate this procedure, we use the sets D_+ and D_- represented in Tab.II and III (our illustrative example). In these tables, $M = \{m_1, \ldots, m_{26}\}$. The set SPLUS_0 for positive class of examples is in Tab.IV. The initial content of STGOOD_0 is {(2,10), (3, 10), (3, 8), (4, 12), (1, 4, 7), (1, 5,12), (2, 7, 8), (3, 7, 12), (1, 2, 12, 14), (2, 3, 4, 7), (4, 6, 8, 11)}.

TABLE II
THE SET D_+ OF POSITIVE OBJECT DESCRIPTIONS

G	D_+
1	$m_1\ m_2\ m_5\ m_6\ m_{21}\ m_{23}\ m_{24}\ m_{26}$
2	$m_4\ m_7\ m_8\ m_9\ m_{12}\ m_{14}\ m_{15}\ m_{22}\ m_{23}\ m_{24}\ m_{26}$
3	$m_3\ m_4\ m_7\ m_{12}\ m_{13}\ m_{14}\ m_{15}\ m_{18}\ m_{19}\ m_{24}\ m_{26}$
4	$m_1\ m_4\ m_5\ m_6\ m_7\ m_{12}\ m_{14}\ m_{15}\ m_{16}\ m_{20}\ m_{21}\ m_{24}\ m_{26}$
5	$m_2\ m_6\ m_{23}\ m_{24}$
6	$m_7\ m_{20}\ m_{21}\ m_{26}$
7	$m_3\ m_4\ m_5\ m_6\ m_{12}\ m_{14}\ m_{15}\ m_{20}\ m_{22}\ m_{24}\ m_{26}$
8	$m_3\ m_6\ m_7\ m_8\ m_9\ m_{13}\ m_{14}\ m_{15}\ m_{19}\ m_{20}\ m_{21}\ m_{22}$
9	$m_{16}\ m_{18}\ m_{19}\ m_{20}\ m_{21}\ m_{22}\ m_{26}$
10	$m_2\ m_3\ m_4\ m_5\ m_6\ m_8\ m_9\ m_{13}\ m_{18}\ m_{20}\ m_{21}\ m_{26}$
11	$m_1\ m_2\ m_3\ m_7\ m_{19}\ m_{20}\ m_{21}\ m_{22}\ m_{26}$
12	$m_2\ m_3\ m_{16}\ m_{20}\ m_{21}\ m_{23}\ m_{24}\ m_{26}$
13	$m_1\ m_4\ m_{18}\ m_{19}\ m_{23}\ m_{26}$
14	$m_{23}\ m_{24}\ m_{26}$

In these tables we denote subsets of values $\{m_8, m_9\}$, $\{m_{14}, m_{15}\}$ by m_* and m_+, respectively. Applying operation $\mathrm{generalization\_of}(s) = s'' = \mathrm{obj}(\mathrm{val}(s))$ to $\forall s \in \mathrm{STGOOD}$, we obtain STGOOD_1 = {(2,10), (3, 10), (3, 8), (4, 7, 12), (1, 4, 7), (1, 5,12), (2, 7, 8), (3, 7, 12), (1, 2, 12, 14), (2, 3,

ISBN: 978-1-61804-251-4

TABLE III
THE SET D_- OF NEGATIVE OBJECT DESCRIPTIONS

G	D_-
15	$m_3m_8m_{16}m_{23}m_{24}$
16	$m_7m_8m_9m_{16}m_{18}$
17	$m_1m_{21}m_{22}m_{24}m_{26}$
18	$m_1m_7m_8m_9m_{13}m_{16}$
19	$m_2m_6m_7m_9m_{21}m_{23}$
20	$m_{19}m_{20}m_{21}m_{22}m_{24}$
21	$m_1m_{20}m_{21}m_{22}m_{23}m_{24}$
22	$m_1m_3m_6m_7m_9m_{16}$
23	$m_2m_6m_8m_9m_{14}m_{15}m_{16}$
24	$m_1m_4m_5m_6m_7m_8m_{16}$
25	$m_7m_{13}m_{19}m_{20}m_{22}m_{26}$
26	$m_1m_2m_3m_5m_6m_7m_{16}$
27	$m_1m_2m_3m_5m_6m_{13}m_{18}$
28	$m_1m_3m_7m_{13}m_{19}m_{21}$
29	$m_1m_4m_5m_6m_7m_8m_{13}m_{16}$
30	$m_1m_2m_3m_6m_{12}m_{14}m_{15}m_{16}$
31	$m_1m_2m_5m_6m_{14}m_{15}m_{16}m_{26}$
32	$m_1m_2m_3m_7m_9m_{13}m_{18}$
33	$m_1m_5m_6m_8m_9m_{19}m_{20}m_{22}$
34	$m_2m_8m_9m_{18}m_{20}m_{21}m_{22}m_{23}m_{26}$
35	$m_1m_2m_4m_5m_6m_7m_9m_{13}m_{16}$
36	$m_1m_2m_6m_7m_8m_{13}m_{16}m_{18}$
37	$m_1m_2m_3m_4m_5m_6m_7m_{12}m_{14}m_{15}m_{16}$
38	$m_1m_2m_3m_4m_5m_6m_9m_{12}m_{13}m_{16}$
39	$m_1m_2m_3m_4m_5m_6m_{14}m_{15}m_{19}m_{20}m_{23}m_{26}$
40	$m_2m_3m_4m_5m_6m_7m_{12}m_{13}m_{14}m_{15}m_{16}$
41	$m_2m_3m_4m_5m_6m_7m_9m_{12}m_{13}m_{14}m_{15}m_{19}$
42	$m_1m_2m_3m_4m_5m_6m_{12}m_{16}m_{18}m_{19}m_{20}m_{21}m_{26}$
43	$m_4m_5m_6m_7m_8m_9m_{12}m_{13}m_{14}m_{15}m_{16}$
44	$m_3m_4m_5m_6m_8m_9m_{12}m_{13}m_{14}m_{15}m_{18}m_{19}$
45	$m_1m_2m_3m_4m_5m_6m_7m_8m_9m_{12}m_{13}m_{14}m_{15}$
46	$m_1m_3m_4m_5m_6m_7m_{12}m_{13}m_{14}m_{15}m_{16}m_{23}m_{24}$
47	$m_1m_2m_3m_4m_5m_6m_8m_9m_{12}m_{14}m_{16}m_{18}m_{22}$
48	$m_2m_8m_9m_{12}m_{14}m_{15}m_{16}$

TABLE IV
THE SET SPLUS_0

$\mathrm{splus}(m), m \in M$
$\mathrm{splus}(m_*) \to \{2, 8, 10\}$
$\mathrm{splus}(m_{13}) \to \{3, 8, 10\}$
$\mathrm{splus}(m_{16}) \to \{4, 9, 12\}$
$\mathrm{splus}(m_1) \to \{1, 4, 11, 13\}$
$\mathrm{splus}(m_5) \to \{1, 4, 7, 10\}$
$\mathrm{splus}(m_{12}) \to \{2, 3, 4, 7\}$
$\mathrm{splus}(m_{18}) \to \{3, 9, 10, 13\}$
$\mathrm{splus}(m_2) \to \{1, 5, 10, 11, 12\}$
$\mathrm{splus}(m_+) \to \{2, 3, 4, 7, 8\}$
$\mathrm{splus}(m_{19}) \to \{3, 8, 9, 11, 13\}$
$\mathrm{splus}(m_*) \to \{2, 8, 10\}$
$\mathrm{splus}(m_{13}) \to \{3, 8, 10\}$
$\mathrm{splus}(m_{16}) \to \{4, 9, 12\}$
$\mathrm{splus}(m_1) \to \{1, 4, 11, 13\}$
$\mathrm{splus}(m_5) \to \{1, 4, 7, 10\}$
$\mathrm{splus}(m_{12}) \to \{2, 3, 4, 7\}$
$\mathrm{splus}(m_{18}) \to \{3, 9, 10, 13\}$
$\mathrm{splus}(m_2) \to \{1, 5, 10, 11, 12\}$
$\mathrm{splus}(m_+) \to \{2, 3, 4, 7, 8\}$
$\mathrm{splus}(m_{19}) \to \{3, 8, 9, 11, 13\}$
$\mathrm{splus}(m_{22}) \to \{2, 7, 8, 9, 11\}$
$\mathrm{splus}(m_{23}) \to \{1, 2, 5, 12, 13, 14\}$
$\mathrm{splus}(m_3) \to \{3, 7, 8, 10, 11, 12\}$
$\mathrm{splus}(m_4) \to \{2, 3, 4, 7, 10, 13\}$
$\mathrm{splus}(m_6) \to \{1, 4, 5, 7, 8, 10\}$
$\mathrm{splus}(m_7) \to \{2, 3, 4, 6, 8, 11\}$
$\mathrm{splus}(m_{24}) \to \{1, 2, 3, 4, 5, 7, 12, 14\}$
$\mathrm{splus}(m_{20}) \to \{4, 6, 7, 8, 9, 10, 11, 12\}$
$\mathrm{splus}(m_{21}) \to \{1, 4, 6, 8, 9, 10, 11, 12\}$
$\mathrm{splus}(m_{26}) \to \{1, 2, 3, 4, 6, 7, 9, 10, 11, 12, 13, 14\}$

4, 7), (4, 6, 8, 11)}.

By Theorem above, we can delete value m_{12} from consideration, see $\mathrm{splus}(m_{12})$ in Tab.IV. The initial content of STGOOD allows to decrease the number of using the procedure $\mathrm{to\_be\_test}()$ and the number of putting extents of tests into STGOOD.

The number of subtasks to be solved. This number is determined by the number of essential values in the set M. The quasi-minimal subset of essential values in M can be found by a procedure analogous to the procedure applicable to search for the initial content of STGOOD. We begin with the first value m_1 of M, then we take the next value m_2 of M and evaluate the function $\mathrm{to\_be\_test}(\mathrm{obj}(\{m_1, m_2\}), \{m_1, m_2\})$. If the value of the function is false, then we take the next value m_3 of M and evaluate the function $\mathrm{to\_be\_test}(\mathrm{obj}(\{m_1, m_2, m_3\}), \{m_1, m_2, m_3\})$. If the value of the function is true, then value m_2 of M is skipped and the function $\mathrm{to\_be\_test}(\mathrm{obj}(\{m1, m3\}), \{m1, m3\})$ is evaluated. We continue this process until we achieve the last value of M. The complexity of this procedure is evaluated as the production of $||M||$ by the complexity of the function $\mathrm{to\_be\_test}()$. In Tab.II,III we have the following list of essential values LEV: $\{m_{16}, m_{18}, m_{19}, m_{20}, m_{21}, m_{22}, m_{23}, m_{24}, m_{26}\}$.

Proposition 1: Each essential value is included at least in one positive object description.

Proof: Assume that for an object description $t_i, i \in G_+$, we have $t_i \cap LEV = \emptyset$. Then $t_i \subseteq M \setminus \mathrm{LEV}$. But $M \setminus \mathrm{LEV}$ is included at least in one of negative object descriptions and, consequently, t_i also possesses of this property. But it contradicts to the fact that t_i is a description of positive object. ■

Proposition 2: Assume that $X \subseteq M$. If $X \cap \mathrm{LEV} = \emptyset$, then $\mathrm{to\_be\_test}(X) = \mathrm{false}$. This proposition is the consequence of Proposition 1.

Note that the description of $t_{14} = \{m_{23}, m_{24}, m_{26}\}$ is closed because of $\mathrm{obj}(\{m_{23}, m_{24}, m_{26}\} = \{1, 2, 12, 14\}$ and $\mathrm{val}(\{1, 2, 12, 14\} = \{m_{23}, m_{24}, m_{26}\}$. We also know that $s = \{1, 2, 12, 14\}$ is closed too (we obtained this result during generalization of elements of STGOOD. So $(\mathrm{obj}(\{m_{23}, m_{24}, m_{26}\}))$, $\{m_{23}, m_{24}, m_{26}\})$ is a maximally redundant test for positive objects and we can, consequently, delete t_{14} from consideration. As a result of deleting m_{12} and t_{14}, we have the modified set SPLUS (Tab.V).

The main question is how we should approach the problem of selecting and ordering subtasks (subcontexts). Consider Tab.VI with auxiliary information. It is clear that if we shall have all the intents of GMRTs entering into descriptions of objects 1, 2, 3, 5, 7, 9, 10, 12, then the main task will be over because the remaining object descriptions (objects 4, 6, 8, 11) give, in their intersection, the intent of already known test (see, please, the initial content of STGOOD). Thus we have to consider only the subcontexts of essential values associated with object descriptions 1, 2, 3, 5, 7, 9, 10, 12, 13. The number

ISBN: 978-1-61804-251-4

TABLE V
THE SET SPLUS_1

$\mathrm{splus}(m), m \in M$
$\mathrm{splus}(m_*) \to \{2, 8, 10\}$
$\mathrm{splus}(m_{13}) \to \{3, 8, 10\}$
$\mathrm{splus}(m_{16}) \to \{4, 9, 12\}$
$\mathrm{splus}(m_1) \to \{1, 4, 11, 13\}$
$\mathrm{splus}(m_5) \to \{1, 4, 7, 10\}$
$\mathrm{splus}(m_{18}) \to \{3, 9, 10, 13\}$
$\mathrm{splus}(m_2) \to \{1, 5, 10, 11, 12\}$
$\mathrm{splus}(m_+) \to \{2, 3, 4, 7, 8\}$
$\mathrm{splus}(m_{19}) \to \{3, 8, 9, 11, 13\}$
$\mathrm{splus}(m_{22}) \to \{2, 7, 8, 9, 11\}$
$\mathrm{splus}(m_{23}) \to \{1, 2, 5, 12, 13\}$
$\mathrm{splus}(m_3) \to \{3, 7, 8, 10, 11, 12\}$
$\mathrm{splus}(m_4) \to \{2, 3, 4, 7, 10, 13\}$
$\mathrm{splus}(m_6) \to \{1, 4, 5, 7, 8, 10\}$
$\mathrm{splus}(m_7) \to \{2, 3, 4, 6, 8, 11\}$
$\mathrm{splus}(m_{24}) \to \{1, 2, 3, 4, 5, 7, 12\}$
$\mathrm{splus}(m_{20}) \to \{4, 6, 7, 8, 9, 10, 11, 12\}$
$\mathrm{splus}(m_{21}) \to \{1, 4, 6, 8, 9, 10, 11, 12\}$
$\mathrm{splus}(m_{26}) \to \{1, 2, 3, 4, 6, 7, 9, 10, 11, 12, 13\}$

of such subcontexts is 39. But this estimation is not realistic.

TABLE VI
AUXILIARY INFORMATION

No	m_{16}	m_{18}	m_{19}	m_{20}	m_{21}	m_{22}	m_{23}	m_{24}	m_{26}	$\sum m_{ij}$
1					×		×	×	×	4
2						×	×	×	×	4
3		×	×					×	×	4
5							×	×		2
7				×		×		×	×	4
9	×	×	×	×	×	×			×	7
10		×		×	×				×	4
12	×			×	×		×	×	×	4
13		×	×				×		×	4
4	×			×	×			×	×	
6				×	×				×	
8			×	×	×	×			×	
11			×	×	×	×			×	
$\sum d_i$	2	4	3	4	4	3	5	6	8	39

We begin with ordering index of objects by the number of their entering in tests in STGOOD_1, see Tab.VII.

TABLE VII
ORDERING INDEX OF OBJECTS IN STGOOD_1

Index of object	**9**	**13**	**5**	**10**	**1**
The number of entering in STGOOD_1	0	0	1	2	3
Index of object	**2**	**3**	**12**	**7**	
The number of entering in STGOOD_1	4	4	4	5	

Then we continue with object descriptions t_9 and t_{13}. Now we should select the subcontexts (subtasks), based on $\mathrm{proj}(t \times m)$, where t is object description containing the smallest number of essential values and m is an essential value in t, entering in the smallest number of object descriptions. After solving each subtask, we have to correct the sets SPLUS, STGOOD, and auxiliary information. So, the first sub-task is $t_9 \times m_{16}$. Solving this sub-task, we have not any new test, but we can delete m_{16} from t_9 and then we solve the sub-task $t_9 \times m_{19}$. As a result, we introduce $s = \{9, 11\}$ in STGOOD and delete t_9 from consideration because of m_{16}, m_{19} are the only essential values in this object description.

Then we solve sub-tasks $t_{13} \times m_{19}$ and $t_{13} \times m_{18}$. The result is introducing $s = \{13\}$ in STGOOD and deleting t_{13} because m_{18} is the only essential value in this object description. After deleting t_9, t_{13}, we can modify SPLUS and delete from it $\mathrm{splus}(m_{16}) = \{4, 12\}$ and $\mathrm{splus}(m_{18}) = \{3, 10\}$. This means that we delete from consideration values m_{16}, m_{18}. Tabs VIII and IX contain the modified set SPLUS and auxiliary information.

TABLE VIII
THE SET SPLUS_2

$\mathrm{splus}(m), m \in M$
$\mathrm{splus}(m_*) \to \{2, 8, 10\}$
$\mathrm{splus}(m_{13}) \to \{3, 8, 10\}$
$\mathrm{splus}(m_1) \to \{1, 4, 11\}$
$\mathrm{splus}(m_5) \to \{1, 4, 7, 10\}$
$\mathrm{splus}(m_2) \to \{1, 5, 10, 11, 12\}$
$\mathrm{splus}(m_+) \to \{2, 3, 4, 7, 8\}$
$\mathrm{splus}(m_{19}) \to \{3, 8, 11\}$
$\mathrm{splus}(m_{22}) \to \{2, 7, 8, 11\}$
$\mathrm{splus}(m_{23}) \to \{1, 2, 5, 12\}$
$\mathrm{splus}(m_3) \to \{3, 7, 8, 10, 11, 12\}$
$\mathrm{splus}(m_4) \to \{2, 3, 4, 7, 10\}$
$\mathrm{splus}(m_6) \to \{1, 4, 5, 7, 8, 10\}$
$\mathrm{splus}(m_7) \to \{2, 3, 4, 6, 8, 11\}$
$\mathrm{splus}(m_{24}) \to \{1, 2, 3, 4, 5, 7, 12\}$
$\mathrm{splus}(m_{20}) \to \{4, 6, 7, 8, 10, 11, 12\}$
$\mathrm{splus}(m_{21}) \to \{1, 4, 6, 8, 10, 11, 12\}$
$\mathrm{splus}(m_{26}) \to \{1, 2, 3, 4, 6, 7, 10, 11, 12\}$

TABLE IX
AUXILIARY INFORMATION (2)

No	m_{19}	m_{20}	m_{21}	m_{22}	m_{23}	m_{24}	m_{26}	$\sum m_{ij}$
1			×		×	×	×	4
2				×	×	×	×	4
3	×					×	×	4
5					×	×		2
7		×		×		×	×	4
10		×	×				×	4
12		×	×		×	×	×	4
4		×	×			×	×	
6		×	×				×	
8	×	×	×	×			×	
11	×	×	×	×			×	
$\sum d_i$	1	3	3	2	4	6	6	25

Then we solve sub-tasks $t_5 \times m_{23}$ and $t_5 \times m_{24}$. The result is deleting t_5 without obtaining any new test. However we can

modify SPLUS and delete from it $\text{splus}(m_{23}) = \{1, 2, 12\}$. This means that we delete from consideration value m_{23}. Now we solve sub-tasks $t_{10} \times m_{20}$, $t_{10} \times m_{21}$, and $t_{10} \times m_{26}$. The result is deleting t_{10} with introducing $s = \{8, 10\}$ into STGOOD. We delete values m_*, m_{13}, m_4, m_5. The modified set SPLUS and Auxiliary information are in Tabs X, XI.

TABLE X
THE SET SPLUS_3

$\text{splus}(m), m \in M$
$\text{splus}(m_1) \to \{1, 4, 11\}$
$\text{splus}(m_2) \to \{1, 11, 12\}$
$\text{splus}(m_+) \to \{2, 3, 4, 7, 8\}$
$\text{splus}(m_{19}) \to \{3, 8, 11\}$
$\text{splus}(m_{22}) \to \{2, 7, 8, 11\}$
$\text{splus}(m_3) \to \{3, 7, 8, 11, 12\}$
$\text{splus}(m_6) \to \{1, 4, 7, 8\}$
$\text{splus}(m_7) \to \{2, 3, 4, 6, 8, 11\}$
$\text{splus}(m_{24}) \to \{1, 2, 3, 4, 7, 12\}$
$\text{splus}(m_{20}) \to \{4, 6, 7, 8, 11, 12\}$
$\text{splus}(m_{21}) \to \{1, 4, 6, 8, 11, 12\}$
$\text{splus}(m_{26}) \to \{1, 2, 3, 4, 6, 7, 11, 12\}$

TABLE XI
AUXILIARY INFORMATION (3)

No	m_{19}	m_{20}	m_{21}	m_{22}	m_{24}	m_{26}	$\sum m_{ij}$
1			×		×	×	3
2				×	×	×	3
3	×				×	×	3
7		×		×	×	×	4
12		×	×		×	×	4
4		×	×		×	×	
6		×	×			×	
8	×	×	×	×		×	
11	×	×	×	×		×	
$\sum d_i$	1	2	2	2	5	5	17

Then we solve the subtasks $t_1 \times m_{21}$ and $t_1 \times m_{24}$ and after that we can delete t_1 without obtaining a new test, but with modifying the set SPLUS and the Auxiliary information (Tabs XII, XIII).

TABLE XII
THE SET SPLUS_4

$\text{splus}(m), m \in M$
$\text{splus}(m_+) \to \{2, 3, 4, 7, 8\}$
$\text{splus}(m_{19}) \to \{3, 8, 11\}$
$\text{splus}(m_{22}) \to \{2, 7, 8, 11\}$
$\text{splus}(m_3) \to \{3, 7, 8, 11, 12\}$
$\text{splus}(m_6) \to \{4, 7, 8\}$
$\text{splus}(m_7) \to \{2, 3, 4, 6, 8, 11\}$
$\text{splus}(m_{24}) \to \{2, 3, 4, 7, 12\}$
$\text{splus}(m_{20}) \to \{4, 6, 7, 8, 11, 12\}$
$\text{splus}(m_{21}) \to \{4, 6, 8, 11, 12\}$
$\text{splus}(m_{26}) \to \{2, 3, 4, 6, 7, 11, 12\}$

TABLE XIII
AUXILIARY INFORMATION (4)

No	m_{19}	m_{20}	m_{21}	m_{22}	m_{24}	m_{26}	$\sum m_{ij}$
2				×	×	×	3
3	×				×	×	3
7		×		×	×	×	4
12		×	×		×	×	4
4		×	×		×	×	
6		×	×			×	
8	×	×	×	×		×	
11	×	×	×	×		×	
$\sum d_i$	1	2	1	2	4	4	14

Now we solve the subtasks $t_2 \times m_{22}$, $t_2 \times m_{24}$, and $t_2 \times m_{26}$. After that we can delete t_2 and m_{22} with introducing $s = \{7, 8, 11\}$ into STGOOD. Tabs Tabs XIV, XV contain the modified set SPLUS and Auxiliary information.

TABLE XIV
THE SET SPLUS_5

$\text{splus}(m), m \in M$
$\text{splus}(m_+) \to \{3, 4, 7, 8\}$
$\text{splus}(m_{19}) \to \{3, 8, 11\}$
$\text{splus}(m_3) \to \{3, 7, 8, 11, 12\}$
$\text{splus}(m_6) \to \{4, 7, 8\}$
$\text{splus}(m_7) \to \{3, 4, 6, 8, 11\}$
$\text{splus}(m_{24}) \to \{3, 4, 7, 12\}$
$\text{splus}(m_{20}) \to \{4, 6, 7, 8, 11, 12\}$
$\text{splus}(m_{21}) \to \{4, 6, 8, 11, 12\}$
$\text{splus}(m_{26}) \to \{3, 4, 6, 7, 11, 12\}$

TABLE XV
AUXILIARY INFORMATION (5)

No	m_{19}	m_{20}	m_{21}	m_{24}	m_{26}	$\sum m_{ij}$
3	×			×	×	3
7		×		×	×	3
12		×	×	×	×	4
4		×	×	×	×	
6		×	×		×	
8	×	×	×		×	
11	×	×	×		×	
$\sum d_i$	1	2	1	3	3	10

Now we solve the subtasks $t_3 \times m_{19}$ with introducing $s = \{3, 11\}$ into STGOOD. After that we can delete $\text{splus}(m_{19})$, this means that we delete m_{19}. Then we solve the subtask $t_3 \times m_{24}$. It leads to deleting $\text{splus}(m_{24})$ and m_{24}. It implies deleting $t_{12}$and t_7. Then we solve the subtask $t_3 \times m_{26}$ After deleting t_3, the main problem is over.

In the example (**method 1**), we have the following subtasks (Tab. XVI).

Tab.XVIII shows the sets STGOOD and TGOOD. All subtasks did not require a recursion. More simple method of

ISBN: 978-1-61804-251-4

TABLE XVII
THE SEQUENCE OF SUBTASKS (METHOD 2)

N	Context, associated with	Extents of tests obtained	Values deleted from context	Object descriptions deleted from context
1	m_{26}	$(2,10),(3,10),$ $(2,3,4,7),(1,4,7)$	$m_*, m_{13}, m_+,$ m_5, m_6	t_{10}
2	m_{26}, m_{24}	$(3,7,12),$ $(4,7,12)$	$m_3, m_{20}, m_{23}, m_1,$ $m_2, m_4, m_7, m_{16},$ m_{18}, m_{19}, m_{22}	
	Subtask is over; return to the previous context and delete m_{24}			
3	m_{26}, not m_{24}, m_{23}	(13)	$m_3, m_7, m_{16}, m_{18},$ m_{19}, m_{20}, m_{22}	
	Subtask is over; return to the previous context, delete m_{23}			
4	m_{26}, not m_{24}, not m_{23}		$m_2, m_3, m_4, m_{16},$ $m_{18}m_{19}, m_{21}$	
5	$m_{26}, m_{22},$ not $m_{24},$ not m_{23}	(9,11), (7,11)		t_2, t_7
	Subtask is over; return to the previous context and delete m_{22}			
6	$m_{26},$ not $m_{24},$ not $m_{23},$ not m_{22}	(3,11), (4,6,11)	$m_2, m_3, m_4, m_{16},$ m_{18}, m_{19}	t_7, t_9, t_2, t_3
	Subtask is over; we have obtained all GMRTs, intents of which contain m_{26}			
7	Context t_5	(1,5,12)		t_5
	Subtask is over; we have found all GMRTs intents of which are contained in t_5			.
8	Context $t_8 \times m_{22}$	(7,8,11), (2,7,8)	$m_3, m_{20}, m_+, m_6,$ $m_*, m_{13}, m_{19}, m_{21}$	
	Subtask is over; return to the previous context and delete m_{22}			
9	Context t_8 without m_{22}	(8,10)	m_*	t_2, t_7
10	Context $t_8 \times m_{21}$ without m_{22}	(4,6,8,11)	m_7, m_{13}, m_{19}	t_6, t_{10}, t_{11}
	Subtask is over; return to the previous context and delete m_{21}, m_{20}			
11	Context t_8 without m_{22}, m_{21}, m_{20}	(3, 8)		t_4, t_6, t_{10}, t_{11}
	Subtask is over; we have found all GMRTs intents of which are contained in t_8.			

TABLE XVI
THE SEQUENCE OF SUBTASKS (METHOD 1)

N	Subcontext	Extent of New Test	Deleted values	Deleted objects
1	$t_9 \times m_{16}$			
2	$t_9 \times m_{19}$	$(9,11)$		t_9
3	$t_{13} \times m_{18}$			
4	$t_{13} \times m_{19}$	(13)	m_{16}, m_{18}	t_{13}
5	$t_5 \times m_{23}$		m_{23}	
6	$t_5 \times m_{24}$			t_5
7	$t_{10} \times m_{20}$	$(8,10)$		
8	$t_{10} \times m_{21}$			
9	$t_{10} \times m_{26}$		$m_*, m_{13},$ m_4, m_5	t_{10}
10	$t_1 \times m_{21}$			
11	$t_1 \times m_{24}$		m_1, m_2	t_1
12	$t_2 \times m_{22}$	$(7,8,11)$	m_{22}	
13	$t_2 \times m_{22}$			
14	$t_2 \times m_{24}$			t_2
15	$t_3 \times m_{19}$	$(3,11)$	m_{19}	
16	$t_3 \times m_{24}$		m_{24}	t_{12}, t_7
17	$t_3 \times m_{26}$			t_3

ordering contexts is based on the basic recursive procedure for solving any kind of subtask described in the previous section. At each level of recursion, we can select value entering into the greatest number of object descriptions; the object descriptions not containing this value generate the contexts to find GMRTs intents of which are included in them. For our example, value m_{26} does not cover two object descriptions: t_5 and t_8. The initial context is associated with m_{26}. The sequence of subtasks in the basic recursive procedure is in Tab.XVII (**method 2**). We assume, in this example, that the GMRT intent of which is equal to t_14 has been already obtained.

TABLE XVIII
THE SETS STGOOD AND TGOOD

N	STGOOD	TGOOD
1	13	$m_1m_4m_{18}m_{19}m_{23}m_{26}$
2	2,10	$m_4m_*m_{26}$
3	3,10	$m_3m_4m_{13}m_{18}m_{26}$
4	8,10	$m_3m_6m_*m_{13}m_{20}m_{21}$
5	9,11	$m_{19}m_{20}m_{21}m_{22}m_{26}$
6	3,11	$m_3m_7m_{19}m_{26}$
7	3,8	$m_3m_7m_{13}m_+m_{19}$
8	1,4,7	$m_5m_6m_{24}m_{26}$
9	2,7,8	m_+m_{22}
10	1,5,12	$m_2m_{23}m_{24}$
11	4,7,12	$m_{20}m_{24}m_{26}$
12	3,7,12	$m_3m_{24}m_{26}$
13	7,8,11	$m_3m_{20}m_{22}$
14	2,3,4,7	$m_4m_{12}m_+m_{24}m_{26}$
15	4,6,8,11	$m_7m_{20}m_{21}$
16	1,2,12,14	$m_{23}m_{24}m_{26}$

We consider only two possible ways of GMRTs construction

ISBN: 978-1-61804-251-4

based on decomposing the main classification context into subcontexts and ordering them by the use of essential values and objects. It is possible to use the two sets $\mathrm{QT} = \{\{i,j\} \subseteq G_+| \ (\{i,j\}, \mathrm{val}(\{i,j\}) \text{ is a test for } G_+\}$ and $\mathrm{QAT} = \{\{i,j\} \subseteq G_+| \ (\{i,j\}, \mathrm{val}(\{i,j\}) \text{ is not a test for } G_+\}$ for forming subcontexts and their ordering in the form of a tree structure. In this case, the algorithm of GMRTs inferring can be constructed in the form of decision tree without recursion and with obtaining each test only once.

V. Conclusion

The decomposition of inferring good classification tests into subtasks of the first and second kinds is presented in the paper. It allows to transform the process of inferring good tests into an incremental reasoning process.

Two methods of forming and reducing subcontexts are given. Various possibilities of constructing algorithms for GMRTs inferring with the use of both subcontexts are considered depending on the nature of GMRTs features.

References

[1] I. Chegis and S. Yablonskii, "Logical methods of electric circuit control," *Trudy Mian SSSR*, vol. 51, pp. 270–360, 1958, (in Russian).

[2] B. Ganter and S. O. Kuznetsov, "Formalizing hypotheses with concepts," in *Proceedings of the Linguistic on Conceptual Structures: Logical Linguistic, and Computational Issues*. Springer-Verlag, 2000, pp. 342–356.

[3] X. A. Naidenova and J. G. Polegaeva, "An Algorithm of Finding the Best Diagnostic Tests," in *The 4-th All Union Conference "Application of Mathematical Logic Methods"*, G. Mintz and E. Lorents, Eds., 1986, pp. 87 – 92, (in Russian).

[4] K. Najdenova, "A relational model of the analysis of experimental data," *Engineering Cybernetics*, vol. 20, no. 4, pp. 99–115, 1982.

[5] X. Naidenova, "Machine Learning as a diagnostic task," in *Knowledge-Dialog-Solution, Materials of the Short-Term Scientific Seminar*, I. Arefiev, Ed. St Petersburg, Russia: State North-West Technical University Press, 1992, pp. 26 – 36.

[6] ——, "Good classification tests as formal concepts," in *Formal Concept Analysis*, ser. Lecture Notes in Computer Science, F. Domenach, D. Ignatov, and J. Poelmans, Eds. Leuven: Springer Berlin Heidelberg, 2012, vol. 7278, pp. 211–226.

[7] O. Ore, "Galois connections," *Trans. Amer. Math. Soc*, vol. 55, pp. 494–513, 1944.

[8] X. Naidenova and A. Ermakov, "The decomposition of good diagnostic test inferring algorithms," in *"Computer-Aided Design of Discrete Devices" (CAD DD2001), Proceedings of the 4-th Inter. Conf.*, J. Alty, L. Mikulich, and A. Zakrevskij, Eds., Minsk, 2001, vol. 3, pp. 61 – 68.

[9] S. Ferré and O. Ridoux, "The use of associative concepts in the incremental building of a logical context," in *ICCS*, ser. Lecture Notes in Computer Science, U. Priss, D. Corbett, and G. Angelova, Eds., vol. 2393. Springer, 2002, pp. 299–313.

[10] X. A. Naidenova, M. V. Plaksin, and V. L. Shagalov, "Inductive Inferring All Good Classification Tests," in *"Knowledge-Dialog-Solution", Proceedings of International Conference*, J. Valkman, Ed., vol. 1. Kiev, Ukraine: Kiev Institute of Applied Informatics, 1995, pp. 79 – 84.

[11] X. A. Naidenova, "DIAGARA: An Incremental Algorithm for Inferring Implicative Rules from Examples," *Inf. Theories and Application*, vol. 12 - 2, pp. 171 – 196, 2005.

[12] V. Choi, "Faster algorithms for constructing a concept (galois) lattice," *CoRR*, 2006. [Online]. Available: http://arxiv.org/abs/cs/0602069

[13] H. Rasiowa, *An Algebraic Approach to Non-classical Logics*, ser. Studies in logic and the foundations of mathematics. North-Holland Publishing Company, 1974.

Xenia Naidenova obtained Ph.D. in Computer Science from the St.Petersburg Electrotechnical University. Xenia is a senior researcher of the Group of Psycho Diagnostic Systems Automation at the Military Medical Academy, St.Petersburg, Russia. Email: ksennaid@gmail.com

Vladimir Parkhomenko is a software engineer in the St.Petersburg State Polytechnical University, St.Petersburg, Russia. Email: parhomenko.v@gmail.com

Colombian Manufacturing Sector: Industrial Structures 2000 – 2012

KARINA MANRIQUE LOPEZ
INDUSTRIAL ENGINEERING PROFESSOR
UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS
BOGOTÁ, COLOMBIA
kmanriquel@udistrital.edu.co

SERGIO ARDILA RODRIGUEZ
INDUSTRIAL ENGINEERING STUDENT
UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS
BOGOTÁ, COLOMBIA
sergioardilar@gmail.com

CARLOS JULIO CASTILLO RINCON
INDUSTRIAL ENGINEERING STUDENT
UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS
BOGOTÁ, COLOMBIA
carlos.castillo18@live.com

The purpose of the current article is to analyze the industrial structures of the manufacturing sector in Colombia between the years 2000-2012, to accomplish this it was made a structure in 14 sectors, from the ISIC (International Standard Industrial Classification) rev. 3. The analysis is based on the changes that have occurred in the number of the enterprises by sector, for this it will be used the industrial concentration indexes, the Herfindahl-Hirschman index and the CR4, the data for this study were obtained by the Sistema de Información y Reporte Empresarial (SIREM) with the enterprises that take part in the industry real sector under inspection, watch and control by the Superintendencia de Sociedades, that includes commercial companies, branches of foreign companies and sole propetorships.[1] The target population of the Encuesta Anual Manufacturera consists by the establishments working on the country and is defined by industrials; this one must have ten or more employees or a value higher than that stipulated annual production.[2]
This article is developed in five parts, in the first one is presented the background of the evolution of the Colombian manufacturing industry, the second one will have a theoretical revision of the industrial concentration indexes, in the third one will be presented the results for the Colombian industry by sector by the 2000-2012 period, the fourth part will show the conclusions and the last one will have the literature review.

Key words: Industrial concentration, manufacturing industry, Herfindahl-Hirschman Index, CR4.

I. Background

According to Garay, Colombian industry has its beginnings in the early 19th century, the structure that was consolidating corresponded with a import substitution model, the first laws (1904-1909) had protectionist measures that supported final products trough duties. Coffee was the main export and was constituted the main source of foreign exchange for the country as same as foreign investment, money that was being used for construction of roads and railways, a situation that was exploited to allow an industrial growth and expansion, this first half of the century was affected historically by two world events: the crisis of the 29 and the second world war that had an important impact on the country. After the war the industrial development policies were consolidated in a own import substitution model accompanied by serious measures of that limited the import, the objective was to settle the early industry foundations: food, drinks, tobacco, clothing, and at the same time giving basis to the substitutions process in intermediate industries. The development of these industries was supported by foreign investment, the protectionist policies and the government support; on the other hand, despite the promotion of industrial production there wasn't a big promotion in export, which became known as the anti-export slant. This situation limited the possibility of foreign exchange earnings to the only product that have outcome opportunities: coffee, this product falling prices in 1955 the growth that had been brought became slower. From this decade is sought

[1] SIREM; *Sistema de Información y Reporte Empresarial- Guía de usuario.* Taken on May 27th from Superintendencia de Sociedades web : www.supersociedades.gov.co/

[2] DANE (2013); *Ficha Metodológica Encuesta Anual Manufacturera-EAM.* Taken on May 27th from DANE's web del sitio web del DANE: www.dane.gov.co/index.php/industria/encuesta-anual-manufacturera-eam

the promotion of the export as embodied in a mixed model of export promotion that was established in 1967, trying to solve the over-reliance on foreign currency export coffee. At the same time, in order to decrease the constant exposition to the currency crises it is adopted a devaluation system crowling peg (dropwise) and it is increased the incentives to the export trough the CAT and creation PROEXPORT, agency to promote exports. These policies have a considered effect on the economy, achieving the initial purposes until the mid- 70s, where is seen a loss of leadership of the industrial sector and at the same time a few attempts at trade liberalization, the economy is influenced by external processes seen and macroeconomic adjustment. In the following period between 1985 and 1990 some manufacturing companies were financially consolidated and in some periods continued the duty protectionist measures in response to the Latin-American economical crisis. In 1990 a shift to the protectionist policy of the state is given and is given the economic opening with the aim to of strengthening the industry and make it more competitive, the first results (1992 and 1993) led to a substantial increase in imports compared to growth marginal exports, some of the producing capital goods and intermediate goods sectors were benefited by to acquire raw materials, machinery and equipment at a lower cost, this favored economic development, particularly during the period 1990-1995 the GDP had a average 4.5% annual growth, the industry step by 1.2% in 1990 and 6.3% in 1993 to 1995 there was a strong performance in the manufacturing sector. But from the moment conditions of the country presented a scene where there was a consumer demand suppressed, and an inflated economy supporting surpluses of transitory income, economic conditions in the country were affected by a real appreciation of the peso, a high interest rates, an increase in the levels of smuggling and political instability.

"Industrial activity was affected by the situation, to the point that in 1996 it recorded a negative growth rate of 3.1%, suggesting structural adjustment problems in the process of adopting the new model -closures, substantial increase unemployment rates, among others- and, more importantly, the industrial structure still had not developed real competitive advantages that allow solidly face foreign competition and penetrate more aggressively in international markets."

II. Industrial concentration indexes

The degree of concentration of a productive branch provides valuable information about its organizational structure and, together with the information that they can provide other variables; it is relevant to determine the degree of competition. In this way, when quantitatively assessing the level of competition in a given sector, the most commonly used method is to obtain some measure indicating the degree of industrial concentration that exist in this sector and its deviation from the situation of perfect competition. Industrial concentration this refers to the size distribution of firms operating in a given sector, which is defined primarily by their market share or size depending on different parameters. [3] This descriptive variable of the industry concentration has a great importance in determining the predominant industrial structure of a sector because it plays a key role in determining the behavior that will have the companies as well as measures from the public sector are implemented in relation to the sector.

In this way, the degree of concentration in a given branch of activity depends on two variables: the number of companies and unequal size. Thus, an activity is more concentrated the smaller the number of firms operating in it, and the larger the difference in their size. The first aspect does not cast doubt, as the number of firms is reduced by market or productive branch, its concentration increases. However, determining the degree of similarity or inequality of companies that make a market entails greater difficulties. On the one hand, it is difficult to determine which variable should be taken to establish the degree of corporate inequality, and on the other hand, we should assess what would be the best indicator to quantify this aspect properly. The main variables are commonly used value added, turnover or number of employees.

Coefficient of concentration

This coefficient is defined as the cumulative market share of the n largest firms in an industry, and therefore it would be worth taking the concentration curve at point n. Its expression is:

$$CR_n = \sum_{i=1}^{n} Z_i$$

$Zn = Market\ share\ of\ every\ firm\ in\ the\ sector$
$n = Number\ of\ firms\ in\ the\ sector$

H*erfindahl-Hirschman Index*

Are considered as characteristic under study two variables reference the personnel employed and the number of establishments, the HH index calculation is done using the following expression:

[3] Furió, E. y Alonso, M. (2008*). Concentración Económica. Algunas consideraciones sobre su naturaleza y medida*, Boletín Económico del ICE nº 2947. Madrid, España.

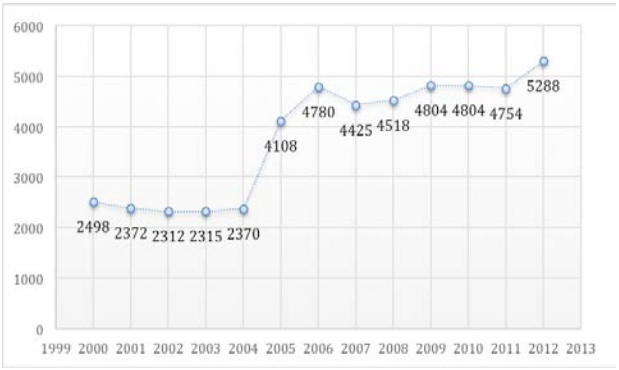
$$H - H_j = \sum_{n=1}^{N} Z_n^2$$

Zn=Market share of every firm in the sector
n=Number of firms in the sector

III. Colombian industrial manufacturing sector structures (2000-2012)

The period after economic openness fostering the growth of the Colombian economy steadily, consequently, GDP in 2013 was an increase of 4.3%, this growth according to President Juan Manuel Santos is due to PLAN BOOST PRODUCTIVITY AND EMPLOYMENT – PLAN DE IMPULSO A LA PRODUCTIVIDAD Y EL EMPLEO (PIPE), although the data reported by DANE will show that this is due in large part to higher growth that occurred in construction, 9,8%; social, community and personal services 5.3% and 5.2 % agriculture and other services [4]. This had an important effect of the number of constituted firms during the period 1999 to 2013 as you can appreciate in figure 1:

Figure 1. Total number of companies of Manufacturing Industry by year.



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

During the early part of the decade the number of companies remained stable, thanks to the impact that this was coming from generating economic openness and uncertainly of entrepreneurs generated by the possible signing of free trade agreements. After that it can be seen that during the 2004, the number of companies presented a significant increase as a result of low interest rates, an increasing exchange rate and with an average close to $ 2.800 in 2003, at the same time of a decline in the rate of consumer prices, this fostered establishment of companies in support to the export to the produced goods, however during 2004, economic conditions changed the exchange rate, which had a value close to $ 1.800 and the debt that had been generated in 2003 for the creation of companies, could not be amortized by the dependent movements in domestic demand, which was evidenced by "stagnation" of the movement of societies as well as a technological backwardness, as the investment made by the Colombian entrepreneurs was destined to raw material, not a technology upgrade that would have a competitiveness looking for the trade agreements.

For this study was taken account industrial clustering of Banco de la República[5], which is presented in table 1:

Table 1. Industrial groups equivalence based on ISIC rev. 3

Classification ISIC rev. 3	Industrial Group
15 - Manufacture of food products and beverages	
16 - Manufacture of tobacco products	Food, beverage and tobacco
17 - Manufacture of textiles	Yarn and fabrics
18 - Manufacture of wearing apparel; dressing and dyeing of fur	Clothing
19 - Tanning and dressing of leather; manufacture of luggage, handbags, saddlery, harness and footwear	Leather and its manufactures
20 - Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials	Wood and its manufactures
21 - Manufacture of paper and paper products	
22 - Publishing, printing and reproduction of recorded media	Graphics and editorial
23 - Manufacture of coke, refined petroleum products and nuclear fuel	
24 - Manufacture of chemicals and chemical products	Chemical Industry
25 - Manufacture of rubber and plastics products	Plastic and rubber products
26 - Manufacture of other non-metallic mineral products	Non- metallic mineral products
27 - Manufacture of basic metals	
28 - Manufacture of fabricated metal products, except machinery and equipment	Basic metals industry
29 - Manufacture of machinery and equipment n.e.c.	
30 - Manufacture of office, accounting and computing machinery	
31 - Manufacture of electrical machinery and apparatus n.e.c.	
32 - Manufacture of radio, television and communication equipment and apparatus	Machinery and equipment
33 - Manufacture of medical, precision and optical instruments, watches and clocks	Optical devices, cinema and others
34 - Manufacture of motor vehicles, trailers and semi-trailers	
35 - Manufacture of other transport equipment	Transport materials
36 - Manufacture of furniture; manufacturing n.e.c.	Other industries

The present analysis was base on criteria of industrial concentration as the Herfindahl-Hirschman index and concentration ratios (RC4). This way it can be set the

[4] DANE (2013). *"Producto interno bruto, primer trimestre de 2013 base 2005"*. Taken on May 22nd from DANE's web: www.dane.gov.co

[5] Banco de la República; *Balanza de pagos de Colombia a partir de 1994. Metodología contemplada en la quinta edición del manual del Fondo Monetario Internacional*; Taken on May 19th from Banco de la República web: www.banrep.gov.co

ISBN: 978-1-61804-251-4

behavioud of market shares, and will identify market structures present in the Colombian industry sectors.

Manufacture of food products and beverages

The food industry and beverages accounts for over 20% of total domestic industry consists of processing and preserving of meat and fish, fruit, vegetables, oils and fats, bakery and grain mil products, coffee products, sugar, cocoa, alcoholic beverages and non- alcoholic beverages (ISIC Rev. 3). The behavior of the food and beverage industry is highly related to the demand patterns of Colombian households, and is linked directly to the agricultural sector as it is the main supplier of raw materials.

Figure 2 allows us to appreciate the change in the number of companies in the recent periods.

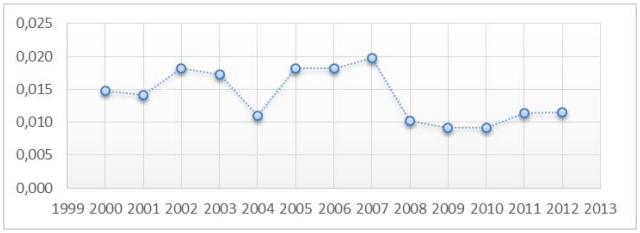
Figure 2. Manufacture of food products and beverages - Total number of enterprises

Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 2 shows an increase in business creation during the period, consistent with the momentum presented by industry, which arose between 2001 and 2006, increasing performance of 9,22% in its production. At the end of 2007, the trend continued and reached a growth of 5.82%, lower than the national aggregate[6]. It is the sector with the highest level of imports, an increase of 80% between 2012 and 2013, which highlights the 137% increase in imports of fruits, vegetables, oils and fats[7].

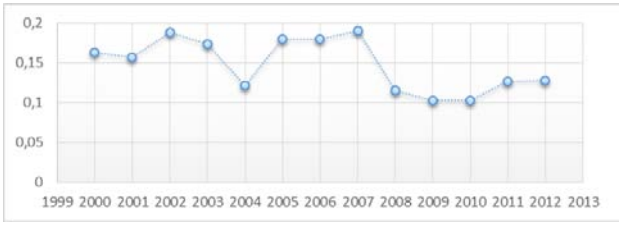
According to the Herfindahl-Hirschman index presented in figure 3, the industrial concentration in the food and beverage industry in the early part of the century ranges between 0.009 and 0.015, indicating a low level of concentration, this means that there is a market with great competence in their companies; so according to Dominguez and Brown (1997), can be demonstrated with the use of concentration ratios, as shown in figure 4, which indicates that the four major companies in the sector are in a smaller market share to 40% for this case; this establishes a competitive market structure.

Figure 3. Herfindahl- Hirschman Index Manufacture of food products and beverages



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 4. CR4 Manufacture of food products and beverages



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Yarn and fabrics Sector

The textile and apparel sector in Colombia represents 8% of industrial GDP and 3% of the national total[8], is an intensive workforce sector and it looks highly affected by smuggling and high rates of informality, this sector has characterized as one of the most traditional sector nationally due to the most important characteristics is that it is vertical integration, which has allowed the joint development of garments and compliance with international standards. This sector comprises the subsectors of yarns, fabrics and garments.

The textile sector has been affected by the increase in informal market, smuggling and entry of China to the world market, the figures according to the Central Bank show that exports have been decreasing in 2008 were 1961.8 million dollars in 2011 and 1116.6 in 2013 to 970.7 million dollars which represents a decrease of 43% respectively.

[6] DANE (2008); Muestra mensual manufacturera; Boletín especial; Bogotá D.C. Taken on May 22nd DANE web: www.dane.gov.co

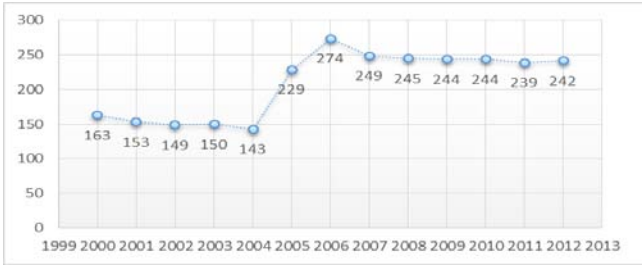
[7] ANDI (2012); *Balance 2012 y perspectivas 2013*; Taken on May 22nd from La República's website: www.larepublica.com.co

[8] MAPFRE (2010). *"Informe sector textil y confecciones colombiano"*.

ISBN: 978-1-61804-251-4

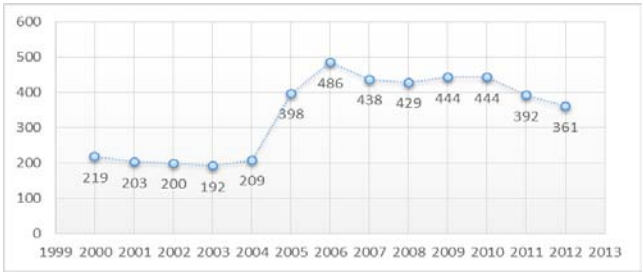
The number of companies associated with the component sectors, over the last years is presented in figure 5 and 6.

Figure 5. Total companies Yarn and fabrics Sector



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 6. Total companies- clothing sector

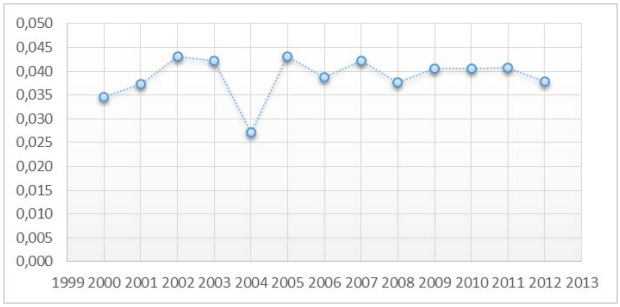


Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

In the charts above can be seen as the sector follows the general behavior of the industry in the period 2004-2006.

The values of the Herfindahl-Hirschman Index figures 7 show that industrial concentration in the yarn and weaving industry during the early part of the century ranges between 0.027 and 0.042, which shows a low level of concentration, indicating a market with great competence in their companies; according to Dominguez and Brown (2007), the concentration ratios, Figure 8, indicate that the four major companies in the sector occupy a share market less than 40% for this case, which identifies a competitive market structure throughout the decade, it can be seen that during the signing of the free trade agreements with the United States and Chile, the values of the graph CR4 decrease, which indicates a lower market share for major companies.

Figure 7. Herfindahl- Hirschman index Yarn and fabrics Sector



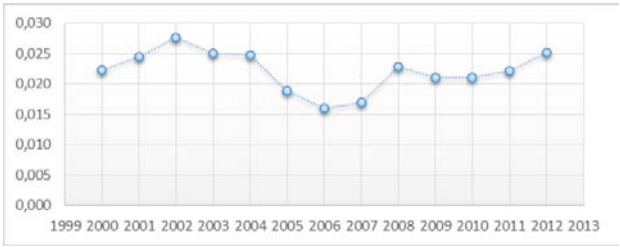
Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 8. CR4 Yarn and fabrics Sector

Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

For the clothing industry, the Herfindahl- Hircshman and CR4 Indexes in figures 9 and 10 respectively have a significant relation, as evidenced as early in the century, industrial concentration was low, and had competitive structure, as Dominguez and Brown (1997), but from 2008, the CR4 is among the range of 0.4 and 0.6, indicating the transition to a competitive oligopoly, where there is an abundance of small producers since barriers to entry are relatively low.

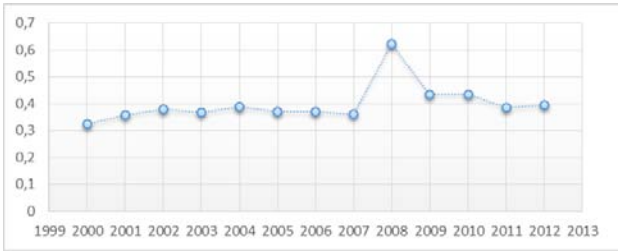
Figure 9. Herfindahl- Hirschman Index Yarn and fabrics Sector



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

ISBN: 978-1-61804-251-4

Figure 10. CR4 Manufacture Yarn and fabrics Sector



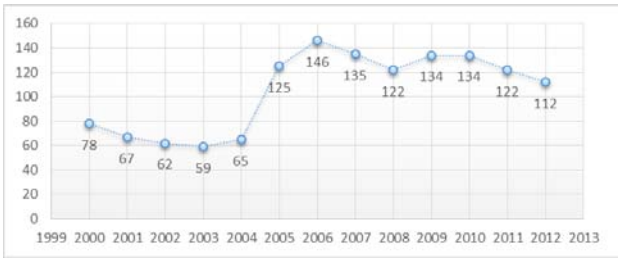
Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Leather and it manufactures

The leather industry and manufacturing is one of the sectors that has been lagging growth more severely over time, in 2012 only had a share of 0.26% of GDP[9], is an intensive labor sector with a low technological index, besides being a highly recognized by the negative impact of production on the environmental industry.

The number of companies in the sector is presented in figure 11, as it shows the sector follows the general behavior of the aggregate. Its behavior in recent years has had a downward trend as a result of conflicts with Venezuela who was one of the main destinations of exports, China's entry to the market and lower domestic demand, so it is a sector that cleary depends on their level of exports but does not have the production conditions to complete in international markets.

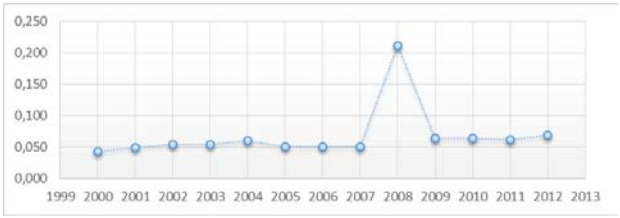
Figure 11. Total companies sector Leather and it manufactures



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

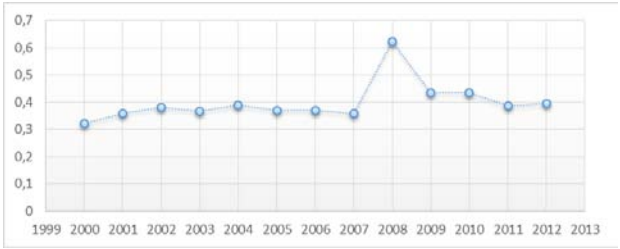
When analyzing the leather industry and it manufactures trough their market structures, is evident a competitive structure in the early years of the century, but in 2008, both the HH index (Herfindahl-Hirschman) in figure 12, as the figure 13, shows an increase in concentration, ant the last one indicates a change in structure to a concentrated oligopoly, as the CR4> 60% and the industry is not intensive in technology, from 2009, presents the near competitive oligopoly structure and is confirmed as the HH indicates a drecrease in the concentration of the industry trend.

Figure 12. Herfindahl- Hirschman Index Leather and it manufactures



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 13. CR4 Leather and it manufactures



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Wood and it manufactures

The Wood industry has products such as sheets and panels, doors, lumber and plywood. In terms of raw materials, although Colombia has 17 million hectares suitable for reforestation, only 1.5%[10] had been used. Colombia has a climate deal that favors the competitiveness of the sector, in addition to the tax benefits for companies within that highlights the income tax exemption for the use of new plantations.

The sector has had a low growth in 2011, especially explained by the strong and increasing competition from products from China. It has also influenced the impact it has had a decrease in exports to Venezuela, which, despite the best efforts of enterprises, have not yet been able to completely replace[11].

Amog it's a main problem the lack of infrastructure and transportation highlights, triggering this in on costs for producers, plus the informatily that is presented in the sector, as most of the companies are really small workshops with a semi-industrial character or artisanal quality with low rates.

[9] AKTIVA (2013). "*El cuero y sus manufacturas en Colombia*".

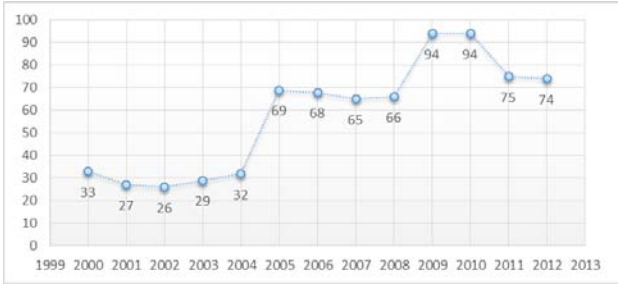
[10] PROEXPORT (2010). "*Sector forestal en Colombia*". Taken on May 22nd from PROEXPORT website: www.proexport.com.co

[11] DNP (2011); *Balance sector industrial 2011*; Bogotá D.C. Pág. 56.

ISBN: 978-1-61804-251-4

The sector is characterized by a variable behavior over time, and although in the last year took a downward trend, according to the national development Plan 2010-2014 recognizes the forestry sector by economic, social and environmental benefits generated this activity. This variability is evident in the movement of companies, which through the last 12 years has shown an increase in their industries, and the turning point of the greater number of companies not presented between the years 2004-2006, as in most of the industry, this were presented in the years 2009-2010.

Figure 14. Total companies- Wood and it manufactures

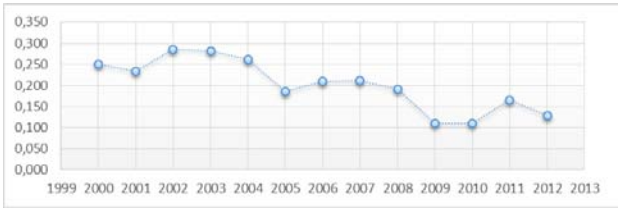


Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

An additional component that has affected the dynamism of the sector in 2011 is the reductuin in the supply of its main raw material, the Wood-winter due to the impact on the Access roads to transport it[12].

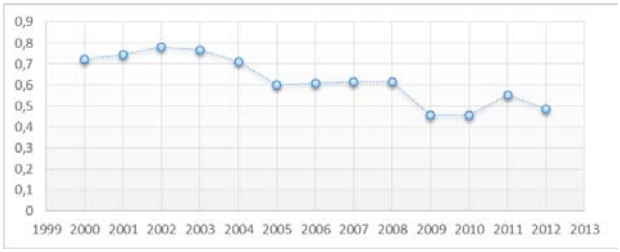
The identification tools market structures indicate that the wood industry and its manufactures present an concentrated oligopolyc structure until 2008, due to low barriers of entry and low technological capacity, subsequently shown by CR4, Graphic 16 shows a competitive oligopoly structure, presenting this trend until 2012, the above analysis is corroborated by the HH index, Figure, which shows a progressive decline in industrial concentration.

Figure 15. Herfindahl- Hirschman Index Wood and it manufactures



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 16. CR4 Wood and it manufactures



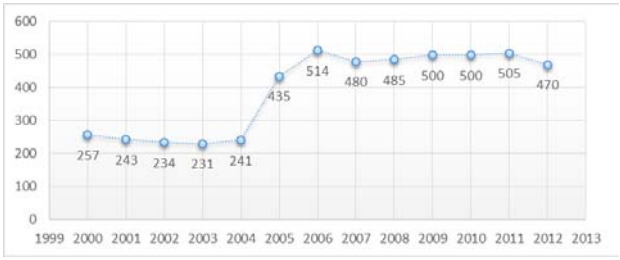
Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Graphics and editorials

The graphics and editorials industry is one of the most affected by the policies adopted by the national government in addition to the inability to compete internationally; odds according to the Banco de la República shows that exports have been falling in 2008 they were $ 862.38 million in 2011 to 727.2 in 2013 and 487 million representing a fall of 15.6% and 33% respectively. This sector is characterized by having a seasonal demand mainly in shool and election periods but as in other industrie, is heavily affected by informality and smuggling.

The ovement of progressive societies indicates a rise from 2004 to 2006, with a stabilization in 2007, as shown in figure 17.

Figure 17. Total companies Graphics and editorials

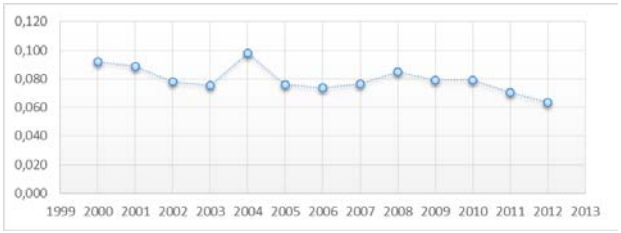


Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Within the sub-sector is the manufacturing of paper and paperboard which has a structure of competitive and differentiated oligopoly for the period 2000-2011 which is transformed at the end of one period of competition, this is evident in the HH index (Figure 18) and CR4 (Figure 19).

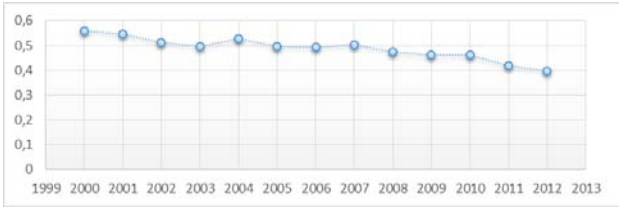
[12] Íbid. P. 56.

Figure 18. Herfindahl- Hirschman Index paper and paperboard products manufacturing



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

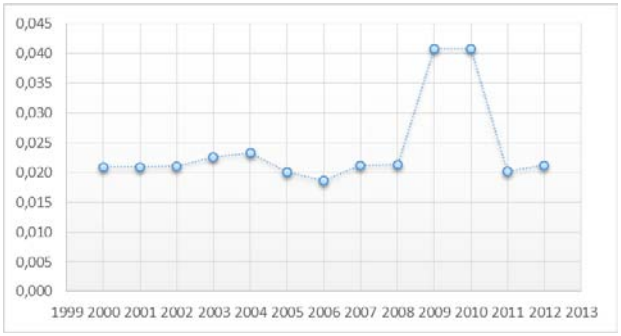
Figure 19. CR4 Manufacture paper and paperboard products manufacturing



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

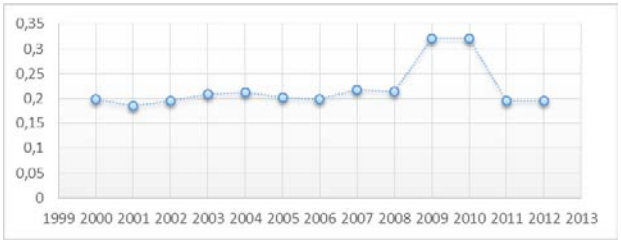
Also, is part of the sector the editorial and printing sub sector, where both HH and CR4 index, figures 20 and 21 respectively agree to show that is in competitive structure.

Figure 20. Herfindahl- Hirschman Index editorials



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 21. CR4 Manufacture sector editorials



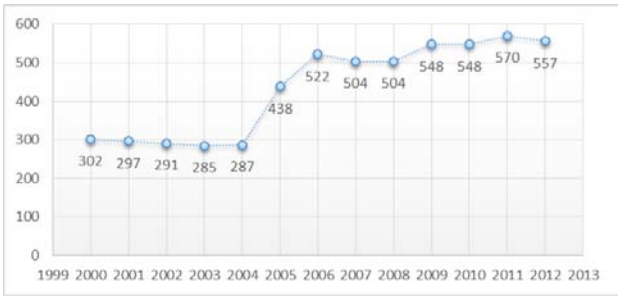
Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Chemical Industry

The chemical industry was one the sectors adversely affected GDP in 2013 with a change od -2.0%[13] due to the production capacity of the products in very low compared to market demand, so that necessary to resort to foreign suppliers, usually from China, Japan and United States of America.

The Chemical industry in general tends persistently to developing new products and improving and refining existing processes chemical production form. The movement of societies indicates a considerable increase in the number of firms between 2004 and 2006, and although it has submitted a further growth, has not been able to maintain the same rate of growth, leveling off in the last years of the decade, as show in figure 22.

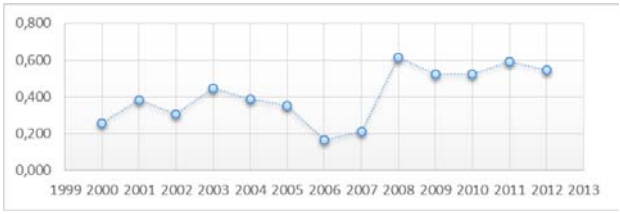
Figure 22. Total companies Chemical Industry



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Belong to the chemical industry sector and subsector coking and oil refining products has a concentrated market structure of concentrated and differenciated oligopoly, as is evidenced by the HH index values, and CR4 Figure 23 and 24 respectively, can be evidenced that from 2007, the concentration has been increased, this behavior continued until 2012.

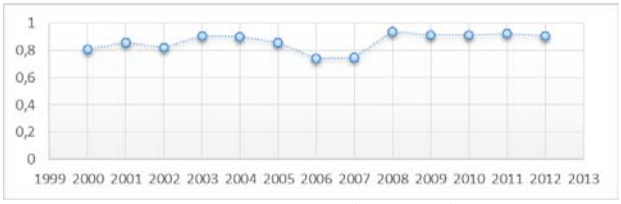
Figure 23. Herfindahl- Hirschman Index coking and oilg refining products



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

[13] Ministerio de Comercio Industria y Turismo (2013). *"Informe de Industria"*. Taken from Ministerio de Comercio, Industria y Turismo web: www.mincit.gov.co

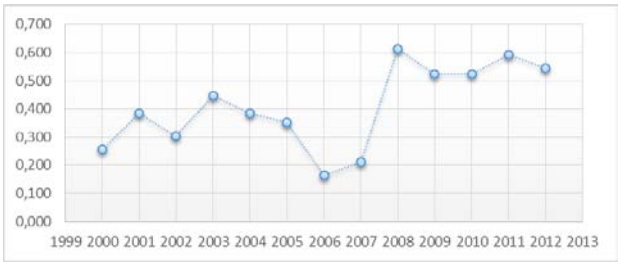
Figure 24. CR4 Manufacture coking and oilg refining products



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

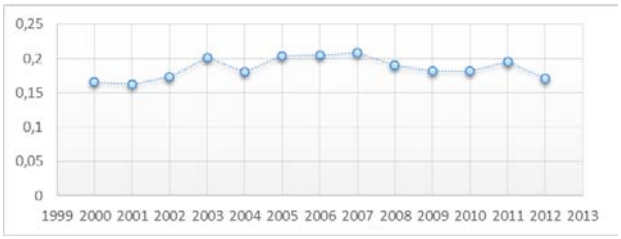
Also, the subsector in charge of chemical products manufacturing takes part of this industry, for which HH and CR4 index show a different behavior from the coking sector, as the market structure is competitive, demonstrated by CR4<0.4, as shown in figure 26, despite the concentration increase in 2007 evidenced by the HH index in figure 25.

Figure 25. Herfindahl- Hirschman Index chemical products manufacturing



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 26. CR4 chemical products manufacturing



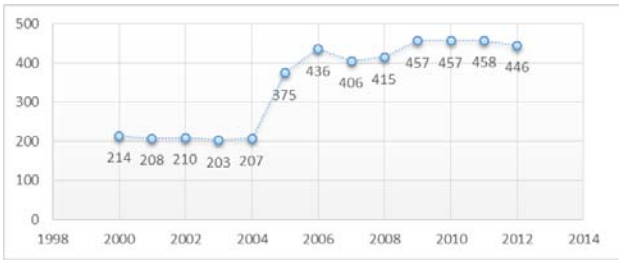
Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Plastic and rubber products

While in 2010 and 2011 the sector grew by 7.2% and 6.4% respectively[14], in 2013 was the most adversely affecting sector to the manufacturing industry with a variation of 6.5%[15] this as a result of domestic demand was affected by lower demand for containers, packaging by sectors such automotive, furniture and construction using inputs and outputs of this subsector.

The number of companies in the sector of plastic and rubber increased between 2004 and 2006 (Figure 27), being consistent with the general behavior of the industry , followed by a stabilization of the number of companies to 446 in 2012.

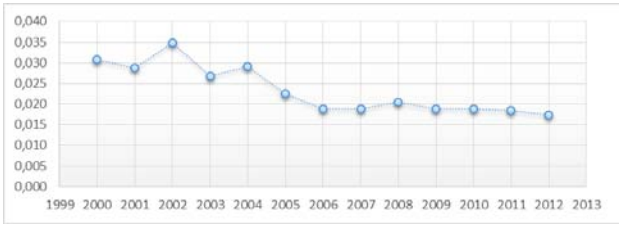
Figure 27. Total companies – plastic and rubber products



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

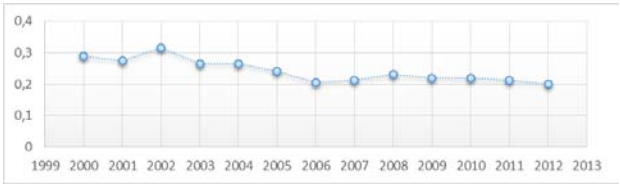
According to the HH index (Figure 28) and CR4 (Figure 29), the performance of the sector relates with a competitive structure, with a decrease in the industrial concentration from 2000 to 2012.

Figure 28. Índice Herfindahl- Hirschman plastic and rubber products



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 29. CR4 Manufacture plastic and rubber products



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Non- metalic minerals products

The contribution of mining to the Colombian economy has evidenced increases in GDP, mineral exports and foreign direct investment in mining, that dynamic has been due mainly to investor interest in the activities of the mininf cycle (exploration and exploitation), both from the expansion of existing mining projects, mainly in

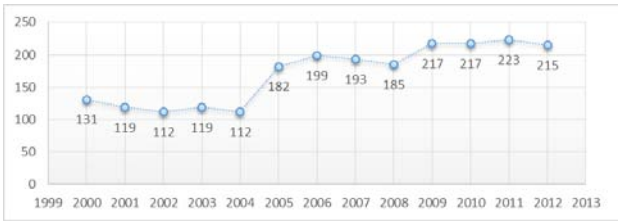
[14] Munera, D., Molina, L. & Montoya, C. "*Caracterización económica del sector envases y empaques en Colombia.*

[15] Ministerio de Comercio Industria y Turismo (2013). Óp. Cit.

coal mining in the north of Colombia (La Guajira and Cesar), and the initiatinon of new explotarion projects primarily for precious metals and base metals.

The National mining GDP about GDP accounted on average in 2011 indicated annual share of 2.27% [16], mainly driven by coal mining projects and nickel in the north of the country. At present this sector is one of the main engines of the Colombian economy, mainly driven by the extraction and export of coal, but one of its biggest problems is the illegal mining as currently estimated that 63% of mines still illegal[17], a situation that cries out for a new mining policy for the country. In figure 27 can be evidenced an increase in the number of firms between 2004 and 2011, when ip to 223 companies presented.

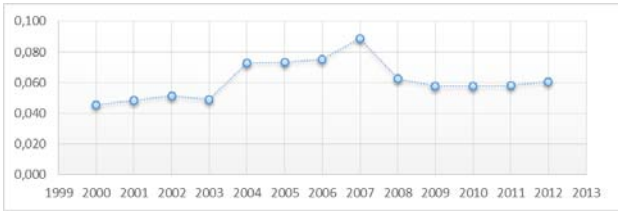
Figure 30. Total companies - Non- metalic minerals products



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

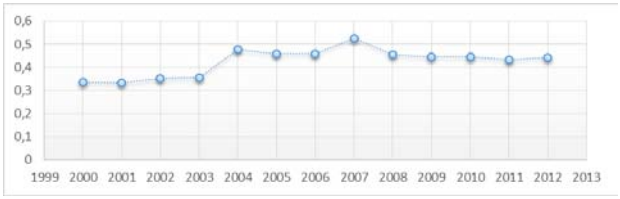
As the HH and CR4 index shows similar behavior in terms of industrial concentration, but since the CR4 shows market structures present, which, in the case of non-metalic minerals sector was competitive until 2003, then shows competitive oligopoly behavior until the end of the period in 2012.(Figure 31).

Figure 31. Herfindahl- Hirschman Index Non- metalic minerals products



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 32. CR4 Non- metalic minerals products



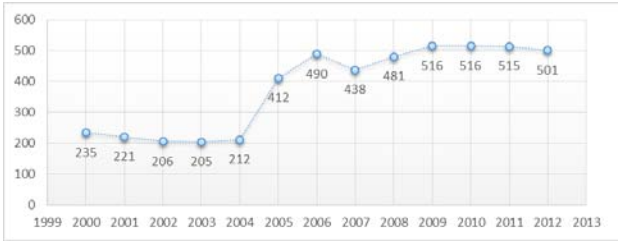
Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Basic metal industry

Industries belonging to base metal in 2010 reached 8% on Colombian manufacturing (0.9% of the industry's GDP) and concentrated approximately 14% of employment in the manufacturing industry [18]. It is a sector with strong investment, but net profits are reduced due to the expected recovery periods investment, but net profits are reduced due to the expected recovery periods investment, steel companies generally have a greater capacity to stand relatively large and higher than in other sectors with negative earning periods, but not sustainable in time given the high level of competition.

The number of companies in the sector is presented in figure 33, as can be seen, it has the same behavior of aggregate industry.

Figure 33. Total companies Basic metal industry



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

The results of the HH index, Figure 31, show a relatively high concentration of the sector of the year 2004 but declined significantly by 2005, when it increased the number of companies according to figure 30; according to CR4, figure 34, the market structure was an concentrated and differentiated oligopoly between 2000 and 2003 and particularly in 2005, 2008 and 2012 given the high entry barriers and requiring technological level, the structure is transformed the competitive and differentiated oligopoly.

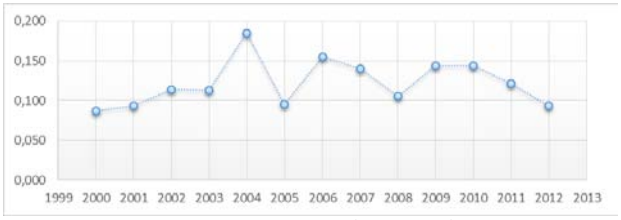
[16] Ministerio de minas y energía (2014). *"Minas.* Taken from the Ministerio de minas y energía website: www.minminas.gov.co

[17] Hurtado, J. (2014). *¿Qué pasa con la minería ilegal en Colombia?* Las dos Orillas. Bogotá, Colombia.

[18] IDOM consulting (2013). *"Plan de Negocio para el sector siderúrgico, metalmecánico y astillero en Colombia"*. RESUMEN EJECUTIVO. Bogotá, Colombia.

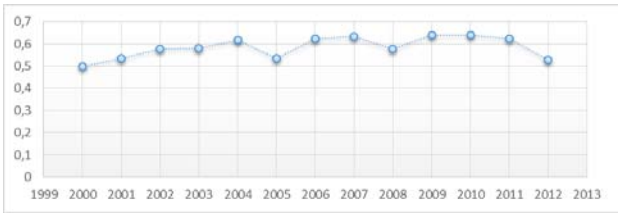
ISBN: 978-1-61804-251-4

Figure 34. Herfindahl- Hirschman Index Basic metal industry



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 35. CR4 Basic metal industry

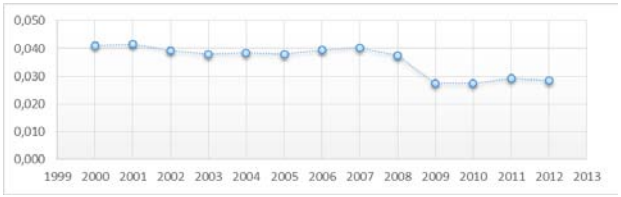


Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

"Metal products except machinery and equipment" also takes part of this sector.

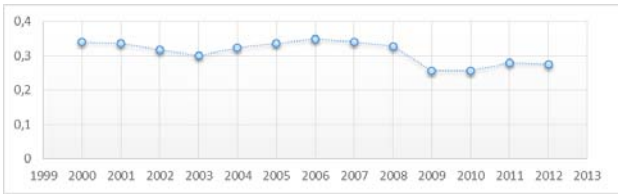
Behaviors evidenced by graphic HH and CR4 indexes (Figures 34 and 35 respectively) regarding industrial concentration, are consistent, have decreased over the period, with a low point in 2009, the CR4 shows a competitive market structure.

Figure 36. Herfindahl- Hirschman Index Metal products except machinery and equipment



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 37. CR4 Metal products except machinery and equipment

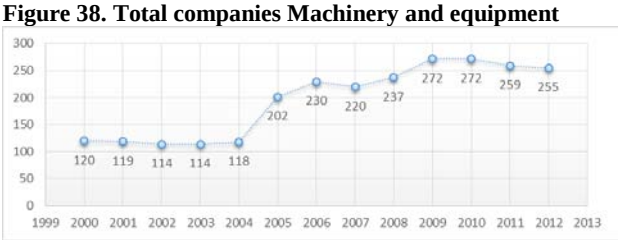


Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Machinery and equipment

The machinery and equipment sector has been an increase in the number of films, starting from 2004, but has stabilized at a value of 255 in recent years, and hasn't reached the top of 272 companies in 2009 again, as is shown in Figure 33.

Figure 38. Total companies Machinery and equipment

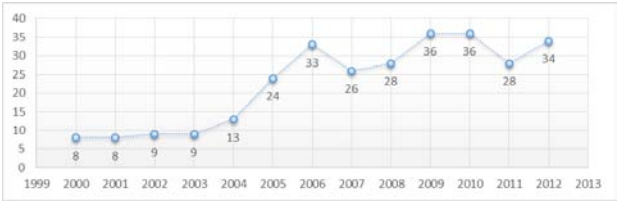


Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Optical devices, cinema and others

The optical devices, cinema and others sector currently has a high participation in Colombian manufacturing industry, as has 34 companies incorporated under the Information System and Business Report, however has increased the number of companies since 2004, reaching a turning point in 2009 to 36 companies.(Figure 39).

Figure 39. Total companies - Optical devices, cinema and others



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

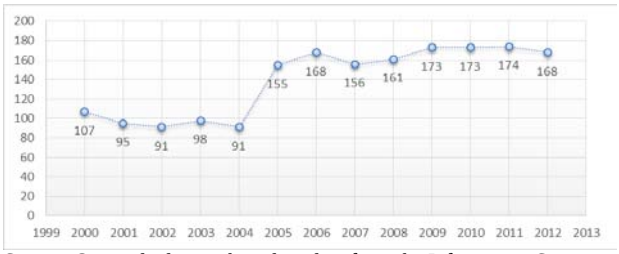
Transport materials

In 2012, Colombia was the fourth largest producer of vehicles in Latin America, using 2.6% (24.783 direct jobs) of employees in manufacturing. Additionally, the sector accounts for 4% of industrial GDP. The automotive sector in Colombia includes assembly activity (light vehicles, trucks, buses and motorcycles) and manufacture of parts and components used in the process as well as the aftermarket. Likewise, input suppliers in other industries as metallurgy, petrochemical (lastics, rubbers) and textiles are involved. Today the country has a fleet of about 4 million units of vehicles of which about 59.5% are imported[19]. The evolution of the

[19] PROEXPORT (2012); *Industria automotriz en Colombia*; Taken on May 25th from Proexport website: www.proexport.com.co

sector in terms of number of companies are presented in Figure 40.

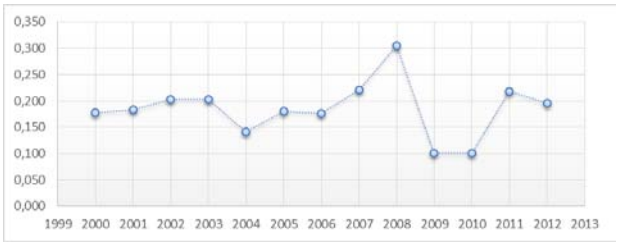
Figure 40. Total companies Transport materials- vehicles, auto parts, trailers and semi trailers



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

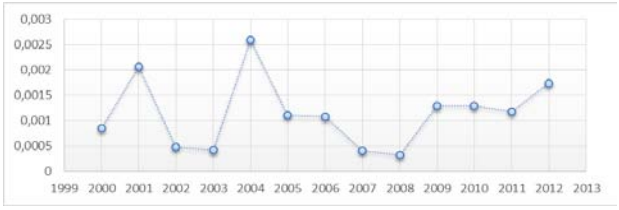
Industrial concentration increased significantly in 2008 as the HH index in figure 41, which could be explainde by the economic crisis that occurred and the output of several companies in the manufacture of motor vehicles, while the CR4 in figure 42, indicates that the prevailing market structure is competitive, since CR4<0.4.

Figure 41. Herfindahl- Hirschman Index Transport materials - vehicles, auto parts, trailers and semi trailers



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

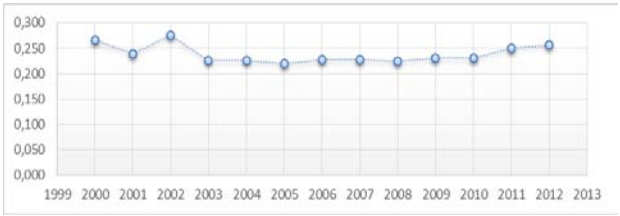
Figure 42. CR4 Transport materials- vehicles, auto parts, trailers and semi trailers



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

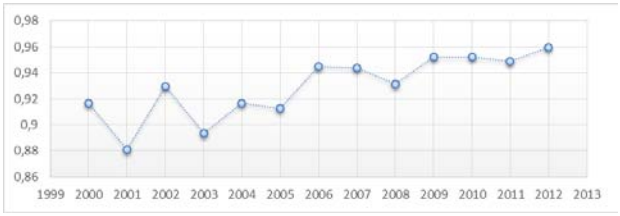
Also, the subsector “other types of transport equipment” takes part in this sector, where, according to the HH index in figure 43 the industrial concentration has remained at a value 0.25, and is consistent with what is whotn in the CR4 index in figure 44, which indicates that the prevailing market structure is concentrated and differenciated oligopoly.

Figure 43. Herfindahl- Hirschman Index other types of transport equipment



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 44. CR4 other types of transport equipment

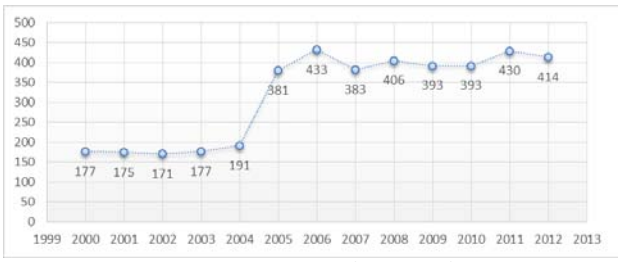


Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Other industries

The sector identified as “other industries” shows an increased in the number of companies, this is reflected in figure 45, and between 2004 and 2006 period in which the greatest number of incorporated companies appeared, peaking in 2006 with 433 companies, value has been adjusted and has not presented substantial movement organizations.

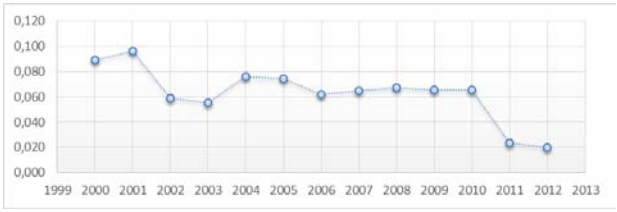
Figure 45. Total companies sector otras industrias



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

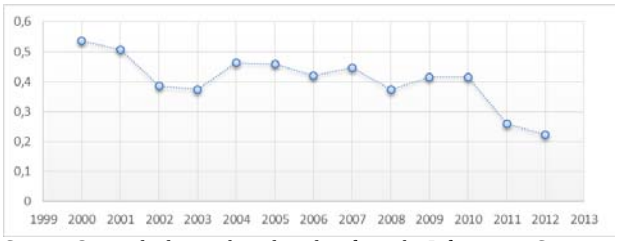
According to the HH index, figure 46, a decrease is evident in the industrial concentration in the sector “other industries” between 2000 and 2012, only a variation of this behavior between 2003 and 2004; according to the methodology for studying the structure of this market sector, competitive oligopoly prevails during the start of the century, in 2002 and 2003 that transformed the structure of competition, while by 2919 the structure showed the trend again competitive oligopoly, the end of the study period ended with a competitive structure, (Figure 47) although the four major industries have a lower market share in recent years.

Figure 46. Herfindahl- Hirschman Index sector otras industrias



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

Figure 47. CR4 Manufacture sector otras industrias



Source: Own calculations based on data from the Information System and Business Report (SIREM 2014)

IV. Conclusions

Given the results of the concentration ratios for the 4 major companies in each sector of the Colombian manufacturing industry shows that for 2012, 11 of the 22 industrial sectors have a competitive market structure, this indicates, according (Koutsoyiannis, 1985) that the number of firms is high, the product is homogeneous, is it has no obstacles in the entry barriers and the prices are settled by the market forces, not by companies.

The results indicate 3 of 22 sectors have a competitive and differentiated oligopoly market structure, while 2 of the 22 has competitive oligopoly, the difference between them is given by their entry barriers and technological capacity. The predominant characteristics of these groups is a smaller number of companies in each sector, the product can be homogeneous or differentiated and may present entry barriers or obstacles (Koutsoyiannis, 1985).

It is evident that 6 of the 22 sectors have concentrated and differentiated Oligopoly market structure, because the degree of concentration is very high in its four main businesses and fairy high entry barriers are presented, along with a requirement of high technological capacity.

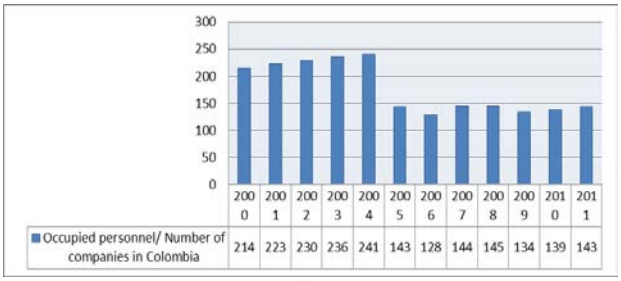
Table 2. CR4 Results and Colombian industry market structures

CR4	Market Structure	Industrial Sector
0,128	Competitive	15 - Manufacture of food products and beverages
1,000	Concentrated oligopoly	16 - Manufacture of tobacco products
0,287	Competitive	17 - Manufacture of textiles
0,244	Competitive	18 - Manufacture of wearing apparel; dressing and dyeing of fur
0,395	Competitive	19 - Tanning and dressing of leather; manufacture of luggage, handbags, saddlery, harness and footwear
0,484	Concentrated oligopoly	20 - Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials
0,396	Competitive	21 - Manufacture of paper and paper products
0,196	Competitive	22 - Publishing, printing and reproduction of recorded media
0,905	Concentrated and differencied oligopoly	23 - Manufacture of coke, refined petroleum products and nuclear fuel
0,171	Competitive	24 - Manufacture of chemicals and chemical products
0,201	Competitive	25 - Manufacture of rubber and plastics products
0,442	Concentrated oligopoly	26 - Manufacture of other non-metallic mineral products
0,529	Concentrated oligopoly	27 - Manufacture of basic metals
0,276	Competitive	28 - Manufacture of fabricated metal products, except machinery and equipm
0,531	Concentrated oligopoly	29 - Manufacture of machinery and equipment n.e.c.
0,916	Concentrated and differencied oligopoly	30 - Manufacture of office, accounting and computing machinery
0,506	Concentrated oligopoly	31 - Manufacture of electrical machinery and apparatus n.e.c.
1,000	Concentrated and differencied oligopoly	32 - Manufacture of radio, television and communication equipment and apparatus
0,902	Concentrated and differencied oligopoly	33 - Manufacture of medical, precision and optical instruments, watches and clocks
0,002	Competitive	34 - Manufacture of motor vehicles, trailers and semi-trailers
0,960	Concentrated and differencied oligopoly	35 - Manufacture of other transport equipment
0,224	Competitive	36 - Manufacture of furniture; manufacturing n.e.c.

Source: Own calculations based on data from the Information System and Business Report (SIREM-2014)

About the employed population it can be seen that although there was an increase in the number of enterprises, the relationship had a strong variation in the 2004-2006 period, implying a decrease in value. The period of the signing of the free trade agreements with the United States and Chile (Years 2006 and 2007), show that the personnel employed respect to the number of companies in Colombia has stabilized, and it is observed a growth in the employed persons and a decrease in the number of companies. This result could have negative consequences, as it expresses that as occupied personnel in Colombia population increases, the demand for labor (expressed in the number of films) is decreasing.

Figure 43. Occupied personnel/Number of companies in Colombia: Years 2000-2011.



Source: Encuesta Anual Manufacturera- EAM. Chart: Variables principales según departamentos, Bogotá D.C. y grupos industriales Total Nacional (2000-2011).

The raid on the industrial market from China has generated a steady decrease in some of the Colombian industrial sectors such as textiles and clothing, leather and leather goods, and publishing sector. Since these sectors are labor intensive and have a low coefficient technological, constant changes in monetary policies do not give a sense of stability for the employer towards creating a more competitive international environment.

International relations with Venezuela have a direct impact on the leather sector, textiles and clothing. This is why policies for these sectors' development have affected negatively the contribution of industry to the GDP, and they should be based on an international policy that builds confidence in entrepreneurs, guided by monetary policies to advancement in technology as this is one of the biggest problems of the Colombian industry.

Industrial Concentration indexes such as the Herfindahl-Hirschman and Concentration Ratios showed consistency in the Colombian industry behavior, showing similarities in the change in concentration depending on the studied period, according to the characterization found these tools evidenced the consequences of the environment in which the Colombian manufacturing industry were developed between 2000 and 2012.

It is expected that the Colombian industry despite currently having a slight recovery compared to previous years thanks to the mining boom, will not have a sustainable development in the coming years as there are no clear policies by the government supporting the industry development, Colombian road infrastructure generates one of the highest cost for employers, informality and illegality present in most industry and poor preparation fot the free trade agreements goes against a recovery in the industry, concluding in a lack of competitiveness and high entry barriers to a market that is not sustainable with the domestic demand but also isn't able to compete internationally.

REFERENCES

[1] AKTIVA (2013). "*El cuero y sus manufacturas en Colombia*". Taken on May 23rd from Aktiva's web Aktiva: www.aktiva.com.co

[2] ANDI (2012); *Balance 2012 y perspectivas 2013*; Taken on May 22nd from La Republica's web: www.larepublica.com.co

[3] Banco de la República; *Balanza de pagos de Colombia a partir de 1994. Metodología contemplada en la quinta edición del manual del Fondo Monetario Internacional*; Taken on may 19th from Banco de la república's web www.banrep.gov.co

[4] DANE (2013); *Ficha Metodológica Encuesta Anual Manufacturera-EAM.* Taken on May 27th from DANE's web del sitio web del DANE: www.dane.gov.co/index.php/industria/encuesta-anual-manufacturera-eam

[5] DANE (2008); *Muestra mensual manufacturera. Boletín especial*; Bogotá D.C. Taken on May 27th from DANE's web: www.dane.gov.co

[6] DANE (2013). "*Producto interno bruto, primer trimestre de 2013 base 2005"*. Taken on May 22nd from DANE's web www.dane.gov.co

[7] Departamento Nacional de planeación (2011); *Balance sector industrial 2011*; Taken on May 22nd from DNP's web. www.dnp.gov.co

[8] Hurtado, J. (2014). *¿Qué pasa con la minería ilegal en Colombia?* Las dos Orillas. Bogotá, Colombia.

[9] IDOM consulting (2013). "*Plan de Negocio para el sector siderúrgico, metalmecánico y astillero en Colombia"*. RESUMEN EJECUTIVO. Bogotá, Colombia.

[10] GARAY, LUIS JORGE (2004); *¨Colombia: Estructura industrial e internacionalización 1967 -1996¨*. Banco de la República's Virtual library.

[11] Ministerio de Comercio Industria y Turismo (2013). *"Informe de Industria"*. Taken on May 20th from Ministerio de Comercio, Industria y Turismo web: www.mincit.gov.co

[12] Ministerio de minas y energía (2014). "*Minas.* Taken on May 19th from del Ministerio de minas y energía web: www.minminas.gov.co

[13] Munera, D., Molina, L. & Montoya, C. "*Caracterización económica del sector envases y empaques en Colombia";*

[14] PROEXPORT (2010). "*Sector forestal en Colombia"*. Taken on May 22th from PROEXPORT's web: www.proexport.com.co

[15] PROEXPORT (2012); *Industria automotriz en Colombia*; Taken on May 25th from Proexport's web: www.proexport.com.co

[16] SIREM; *Sistema de Información y Reporte Empresarial- Guía de usuario*. Taken on May 22th Superintendencia de Sociedades web: www.supersociedades.gov.co/

The effect of the variation of Popov's parameter on the size of the region of absolute robust stability of a monotonous nonlinear impulsive control system

N.A. Tseligorov, G.M. Mafura

***Abstract*—** This paper focuses on the effect of the variation of Popov's parameter on the size of the region of absolute robust stability of monotonous nonlinear impulsive control systems. To evaluate the effect of Popov's parameter, the program complex "Stability" is used. This program complex plots the regions of absolute robust stability on the complex plane. An illustrative example is given to demonstrate the effect of varying Popov's parameter on the region of absolute robust stability.

***Keywords*—** absolute robust stability, nonlinear impulsive system, transfer function, perturbed polynomial, Popov's parameter, monotonous nonlinearity, root locus.

I. Introduction

TODAY, the process of designing any technically complex system is accompanied with the creation of its mathematical model. The mathematical model allows the engineer to get information about the stability of the object of control. Mathematical models, which are based on the application of the criteria of absolute stability to the system under design, focused mainly on control systems with standard nonlinear characteristics, with theoretical forms of nonlinear characteristics, which differed considerably from actual real world systems. To get results which are closer to the real world conditions of exploitation, it is necessary to take into account the parametric uncertainty of the control system in question. Consequently, the software used for such models changes and expands in order to solve the various problems and challenges faced. This justifies the use of computer algebra systems to work with symbolic data.

II. Statement of the Problem

The Academic Y.Z. Tsipkin proposed the criteria for absolute stability of nonlinear control systems (NICS) with monotonous characteristics in the form of the following inequalities[1]

$$\mathrm{Re}[1+q(1-e^{-j\varpi})]W(j\varpi,0)+k^{-1}>0 \qquad (1)$$

This inequality must be satisfied for all frequencies ϖ within the interval $[0,\pi]$ for real V.M. Popov's parameters $q \geq 0$. Characteristics $\Phi(\sigma)$ nonlinear elements (NE) fulfill the following condition (The absolute stability of the equilibrium point)

$$0 \leq \frac{\Phi(\sigma)}{(\sigma)} \leq k, \quad \Phi(0)=0 \qquad (2)$$

The above criterion (2) can be interpreted geometrically as plotting a modified amplitude-phase characteristics graph $\tilde{W}(j\varpi,0)$ with Popov's line plotted on the $\tilde{W}(j\varpi,0)$ plane [1]

$$U^*(\varpi,0)+q\tilde{V}(\varpi,0)+1/k>0,$$

$$\text{Where } U^*(\varpi,0)=\mathrm{Re}W^*(j\varpi,0),$$

$$V^*(\varpi,0)=-\mathrm{Re}[e^{-j\varpi}W^*(j\varpi,0)]+\mathrm{Re}W^*(j\varpi,0)$$

The plotted line crosses through the point -1/k, on the real axis and at an angle of arctg 1/q. A graphical illustration of the criterion of absolute stability is given below in Fig.1.

In well-known scientific publications the test for absolute

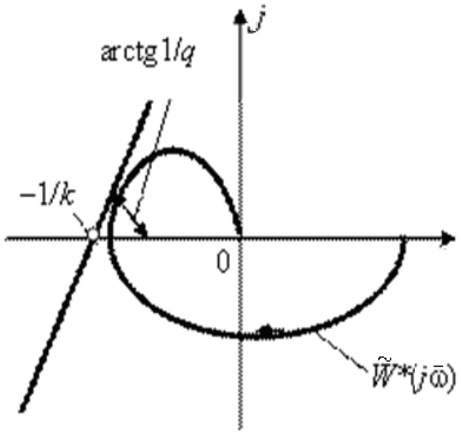


Fig. 1 Criterion of absolute stability for monotonous nonlinearities

N.A. Tseligorov is with Rostov on Don's affiliate of The Russian Customs Academy , 344000, Rostov on Don, Budenonvsky Av. D.20,Russia (phone: +7(632)218-07-12; e-mail: nzelig@rambler.ru).

G.M. Mafura is with LLC Rostovgiproshaht, 344000, Rostov on Don, St. Krasnoarmeickaya, D.157, Russia. (phone: +7928 764 48 15, e-mail: mafurag@hotmail.com).

stability of NICS is carried out with the condition that Popov's parameter q=0. In [2] the values of parameter q in certain intervals with respect to the characteristics of NE:

- $-\infty < q < \infty$, if the NE characteristic are definite and stationary;
- $0 \le q < \infty$, if the NE characteristic are stationary, but not definite and with negative hysteresis;
- $-\infty < q \le 0$, if the NE characteristics are stationary, but not definite with positive hysteresis;
- $q = 0$, if the NE characteristics are not stationary (indefinite or definite).

In [3] it is proposed to select values of q, using the recursive relation, found when Popov's criteria is geometrically interpreted:

$$q_i^+ = \frac{1}{tg(\frac{\pi}{2}\frac{l-i}{l+i})}, \quad q_{i+1}^- = -q_i, \quad q_0 = 0, \quad i = 1,3,...l-1 \qquad (3)$$

where $l+1$ – number of selected values of q;

q_i^+, q_{i+1}^- –corresponding number of positive and negative values of Popov's parameter;

l – an even number.

However to the best of our knowledge, there is no literature which graphically illustrates the region of absolute robust stability and the effect of varying Popov's parameter on the extent of the region of stability.

It is necessary to evaluate the change in the region of absolute robust stability and define the effect Popov's parameter on the region of absolute robust stability of a control system under evaluation.

III. SOLUTION OF PROBLEM

Transcendence of the frequency response characteristics $W(j\varpi)$ complicates the use of the criteria for testing the absolute robust stability of NICS. Using w-transform it is possible to simplify the test for absolute stability of NICS to testing if the resultant real polynomial is Hurwitz and for robust stability to testing the Kharitonov's polynomials [4] i.e. checking if the resultant group of real polynomials are Hurwitz.

Using w-transform, the test for absolute stability of NICS is reduced to testing if the resultant real polynomial is Hurwitz[3]. Criterion (1) takes the following form in w-plane

$$\mathrm{Re}[(1+\mathrm{q}\frac{2w}{1+w})\mathrm{W}(w)] + \mathrm{k}^{-1} > 0 \quad \text{or}$$

$$\mathrm{Re}[(1+\mathrm{q}\frac{2\mathrm{j}\nu}{1+\mathrm{j}\nu})\mathrm{W}(\mathrm{j}\nu)] + \mathrm{k}^{-1} > 0, \quad \forall \nu \in [0,\infty], \qquad (4)$$

where $w = j\nu$, $\nu = \mathrm{tg}\frac{\omega \mathrm{T}_0}{2}$ – relative pseudo frequency, T_0 – sampling interval.

If the transfer function's frequency response characteristics are presented as shown below

$$W(j\nu) = \frac{\alpha_1(\nu) + j\beta_1(\nu)}{\alpha_2(\nu) + j\beta_2(\nu)}, \qquad (5)$$

then after substituting (5) in (4) we get the following polynomial expression which corresponds to the criterion (4) [5].

$$k\{(\alpha_1(\nu)\alpha_2(\nu) + \beta_1(\nu)\beta_2(\nu))(1+\nu^2) + 2q[(\alpha_1(\nu)\alpha_2(\nu) + \beta_1(\nu)\beta_2(\nu))\nu^2 + (\alpha_1(\nu)\beta_2(\nu) - \alpha_2(\nu)\beta_1(\nu))\nu]\} + [\alpha_2^2(\nu) + \beta_2^2(\nu)](1+\nu^2) = 0 \qquad (6)$$

The resultant polynomial expression (6) is the general polynomial form of the criterion for transfer function of NICS of any degree. To get the symbolic expression of (6) one can use any of the known computer algebra systems (CAS) for example Mathematica, Mathcad, Maxima etc. Using any of the above CAS systems allows us to get coefficients in symbolic form for a given transfer function degree. Substituting the symbolic equation (6) for criteria of stability of NICS with the numeric values from the numerator and denominator of the transfer function under evaluation, the polynomial expression in numeric coefficients is gotten. The resultant polynomial expression in numeric form can now be used to test the absolute stability of NICS in question without intermediate calculations. To test the robust stability of NICS, it is necessary to have interval values of the coefficients in the numerator and denominator of the transfer function. This interval values are used in the strong Kharitonov's theory [6] for testing the robust stability of NICS in question.

Evaluating the effect of Popov's parameter q on the region of stability can be done by giving q a concrete value then plotting the region of absolute robust stability.

When concrete values of *i* are applied to formulae (3), then fan symmetrical and uniform values of Popov's parameter as shown in Table.1.

An analysis of the values in table.1 shows that the positive values, that can be used while evaluating NICS using criterion (4), lie within the interval from q=1 to q=12. To get a graphic representation of the region of absolute stability, the program complex "Stability" is used to plot the region of stability[7].

Table 1

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
q+	0,29		0,89		1,73		3,52		12,07		-13,35		-4,81	
Angle + (degrees)	16,40		41,56		60,01		74,12		85,26		-85,71		-78,2	
q-		-0,29		-0,89		-1,73		-3,52		-12,07		13,35		4,81
Angle - (degrees)		-16,4		-41,5		-60,0		-74,1		-85,26		85,71		78,26

IV. ILLUSTRATIVE EXAMPLE

Consider a NICS, with the following transfer function with perturbed coefficients

$$W(w)=\frac{(0{,}12..0{,}18)w^3+(0{,}41..0{,}59)w^2-(0{,}22..0{,}54)w+(0{,}09..0{,}15)}{(0{,}55..1{,}99)w^3+(2{,}21..3{,}22)w^2+(0{,}45..1{,}21)w+(0{,}11..0{,}16)} \quad (11)$$

The nonlinear elements characteristics lie within the interval [0; 1.5].

The values of V.M. Popov's parameter, while taking into account the values in table.1, fall between q=0.1 and q=14.

Criterion expression for the transfer function (11) in symbolic form is written as follows

$$\begin{aligned} P(x)|_{x=\nu^2} &= k[a_3b_3x^4+(a_3b_3+a_2b_2-a_1b_3-a_3b_1)x^3+ \\ &+(-a_3b_1-a_2b_0-a_1b_3+a_2b_2-a_0b_2+a_1b_1)x^2+ \\ &+(a_0b_0-a_0b_2-a_2b_0+a_1b_1)x+a_0b_0]+ \\ &+2qk[a_3b_3x^4+(a_2b_2-a_1b_3-a_3b_1+a_2b_3-a_3b_2)x^3+ \\ &+(-a_0b_2-a_2b_0+a_1b_1-a_0b_3-a_2b_1+a_3b_0+a_1b_2)x^2+ \\ &+(a_0b_0+a_0b_1-a_1b_0)x]+ \\ &+[b_3^2x^4+(b_2^2-2b_3b_1+b_3^2)x^3+(b_1^2+b_2^2-2b_2b_0-2b_1b_3)x^2+ \\ &+(b_0^2-2b_0b_2+b_1^2)x+b_0^2] \end{aligned} \quad (12)$$

The variation of numeric coefficient values of criterial equation (6) as the value of parameter k changes is illustrated on the graphs below.

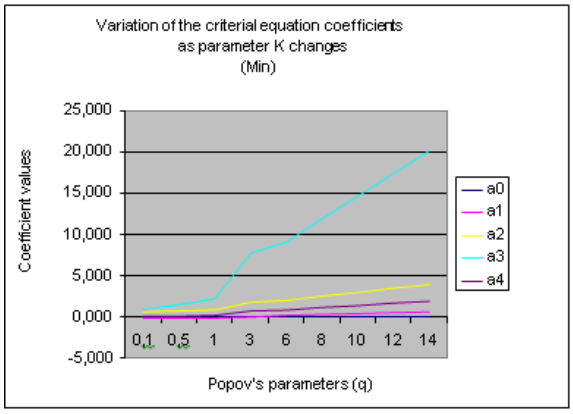


Fig. 2. The variation of polynomial coefficients as parameters of the criterial equation change (minimum values of the transfer function

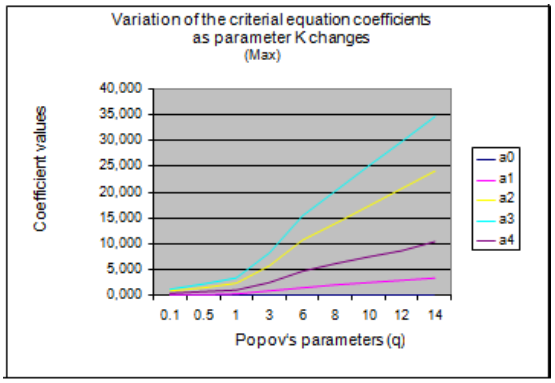


Fig. 3. The variation of polynomial coefficients as parameters of the criterial equation change (maximum values of the transfer function

The variation of coefficient values of the free polynomial in the criterial equation (6) is illustrated below.

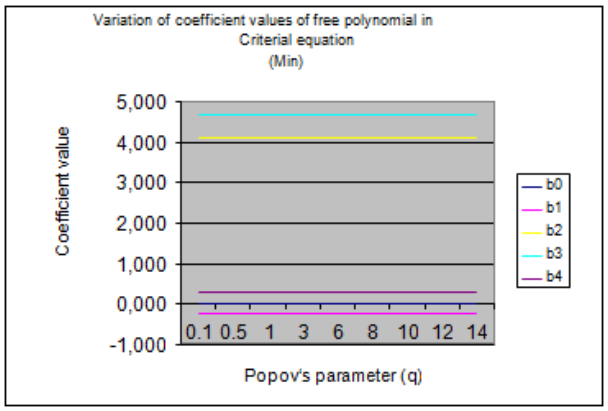


Fig. 4. Variation of coefficient values of the free polynomial in the criterial equation (minimum values of the transfer function)

The calculated coefficient values are typed into the form, as shown in Fig.7, of the program complex "Stability". The program complex then plots a modified root locus diagram, which takes into account the perturbed nature of the coefficients, for concrete values of q.

ISBN: 978-1-61804-251-4

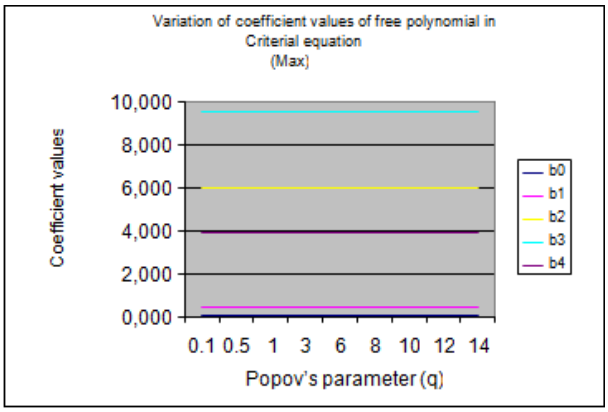


Fig. 5 Variation of coefficient values of the free polynomial in the criterial equation (minimum values of the transfer function).

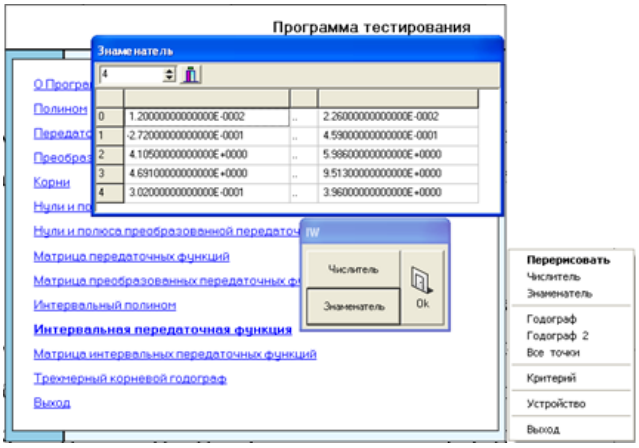
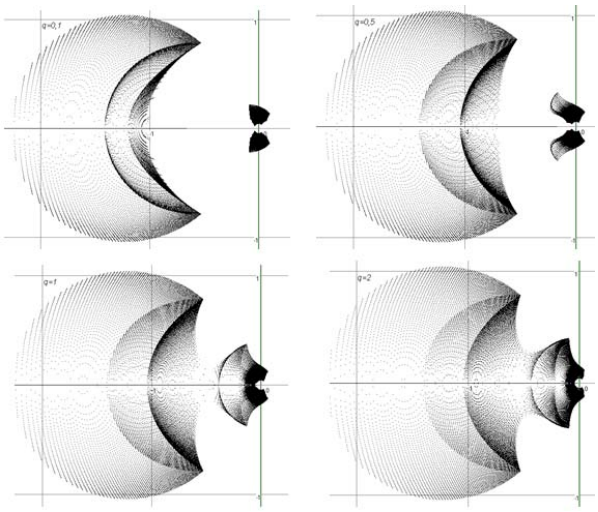


Fig. 6. User interface of program complex “Stability”

Screenshots of the regions of robust stabilities for several values of q. As shown in Fig. 8.



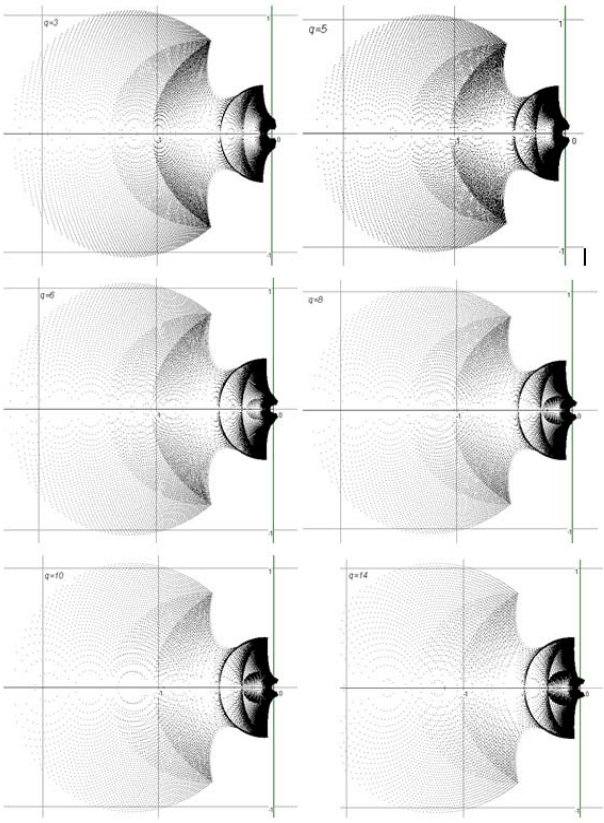


Fig.7. The shapes and sizes of regions of stabilities for various concrete values of q.

After comparison of the above regions of stabilities, one can see that the regions of stabilities tend to increase. This is especially evident when small values of q are applied i.e. q=0 to q=2. As the values of q further increase, the region of stability remains almost unchanged.

V. Conclusion

The proposed approach enables one to get the symbolic coefficient expression of the criteria, for a nonlinear impulsive control system, with monotonous nonlinearity, with transfer function of given degree, in analytical form and the resultant criterial expression. The resultant criterial expression, with concrete numeric values from the transfer function, can be used to plot a modified root locus plot, which takes into account the perturbed nature of coefficients. The resultant root locus diagram proves that the region of stability increases while the value of Popov’s parameter increases from 0.1 to 14.

References

[1] Cypkin, Ja.Z. Teorija nelinejnyh impul'snyh sistem/ Ja.Z. Cypkin, Ju.S. Popkov. – M.: Nauka, 1973. - 414s.

[2] Mutter, V.M. Analogo-cifrovye avtomaticheskie sistemy: Proektirovanie i raschet/ V.M. Mutter. – L. Mashinostroenie, 1981. –199s.

ISBN: 978-1-61804-251-4

[3] Diduk, G.A. Analiz i optimal'nyj sintez na JeVM sistem upravlenija/ G.A. Diduk, A.S. Konovalov, I.A. Orurk, L.A. Osipov. M.: Nauka. Glavnaja redakcija fiziko-matematicheskoj literatury, 1984. - 344s.

[4] 4. Tseligorov, N.A. Analiz absoljutnoj ustojchivosti nelinejnyh impul'snyh avtomaticheskih sistem analiticheskimi metodami/ V.I. Serkov, N.A. Tseligorov//– Avtomatika i telemehanika, 1975, #9, s. 60–65.

[5] Tseligorov, N.A. Podhod k issledovaniju robastnoj absoljutnoj ustojchivosti nelinejnyh impul'snyh sistem upravlenija s monotonnymi harakteristikami/ N.A. Tseligorov, G.M. Mafura// Izv. vuzov. Jelektromehanika. 2013. #5. s. 44-48.

[6] 6. Haritonov, V.L. Ob asimptoticheskoj ustojchivosti polozhenija ravnovesija semejstva sistem linejnyh differencial'nyh uravnenij/ V.L. Haritonov // Differencial'nye uravnenija.-1978. - #11. - s.2086-2088.

[7] Tseligorov, N.A. Primenenie modelirujushhego kompleksa «Ustojchivost'» dlja issledovanija nelinejnyh impul'snyh sistem upravlenija s neopredeljonnostjami/ N.A. Tseligorov, G.M. Mafura// «Komp'juternoe modelirovanie 2013»: trudy mezhdunarodnogo seminara. – SPb.: Izd–vo Politehnicheskogo un-ta.

Computer Simulation of Hybrid Systems by ISMA Instrumental Facilities

Yu.V. Shornikov, M.S. Myssak, D.N. Dostovalov

Abstract— A class of hybrid systems (HS) unresolved with respect to the derivative is considered. Architecture of instrumental environment is designed in accordance with CSSL standard. Library of original numerical solvers, embedded in simulation environment, is presented. Algorithm of numerical analysis of HS modes is given. Theorem about the choice of the integration step considering the HS event function dynamic has been formulated and proved. Algorithm of accurate HS event detection with implicit continuous behaviour models is designed. Testing of the proposed algorithms in the ISMA instrumental environment is performed. Example of specification and analysis of electric power systems models is given.

Keywords— computer aided analysis, software architecture, numerical simulation, differential equations, event detection, circuit simulation.

I. Introduction

Hybrid systems (HS) theory is a modern and versatile apparatus for mathematical description of the complex dynamic processes in systems with different physical nature. Such systems are characterized by the composition of the continuous and discrete behaviours. Earlier the ISMA instrumental environment [1, 2] examined models and methods of HS analysis, continuous modes of which are described by the Cauchy problem with constraints. In this paper the extension of class of systems by models unresolved with respect to the derivative is proposed. Numerical analysis of the new class of problems requires using a specific integration and HS event detection algorithms. The described algorithms are implemented in the ISMA and successfully tested.

II. CLASS OF SYSTEMS

There are many systems (mechanical, electrical, chemical, biological, etc.), the behavior of which can be conveniently described as a sequential change of continuous modes. These systems are referred to as hybrid or event-continuous. Each mode is given by a set of differential-algebraic equations with the following constraints:

$$\begin{aligned} &y' = f(x, y, t), x = \varphi(x, y, t), \\ &pr : g(x, y, t) < 0, \\ &t \in [t_0, t_k], x(t_0) = x_0, y(t_0) = y_0, \\ &x \in R^{N_x}, y \in R^{N_y}, t \in R, \\ &f : R^{N_x} \times R^{N_y} \times R \to R^{N_y}, \\ &\varphi : R^{N_x} \times R^{N_y} \times R \to R^{Nx}, \\ &g : R^{N_x} \times R^{N_y} \times R \to R^{S}. \end{aligned} \quad (1)$$

The vector-function $g(x,y,t)$ is referred to as event function or guard. A predicate pr determines the conditions of existence in the corresponding mode or state. Inequality $g(x,y,t) < 0$ means that the phase trajectory in the current mode should not cross the border $g(x,y,t) = 0$. Events occurring in violation of this condition and leading to transition into another mode without crossing the border are referred to as one-sided. Many practical problems are characterized by stiff modes, and the surface of boundary $g(x,y,t) = 0$ has sharp angles or solution has several roots at the boundary [2]. Numerical analysis of such models by traditional methods is difficult or impossible, as it gives incorrect results. Therefore it is necessary to use special methods to detect events accurately.

Computer analysis of these systems is typically performed in simulation tools, best of which are Charon (USA), AnyLogic (Russia), Scicos (France), MVS (Russia), Hybrid Toolbox and HyVisual (USA), DYMOLA (Sweden) and etc.

In the simulation of electrical circuits, processes of chemical kinetics, electromechanical processes and many other applications a necessity arises to numerically analyze HS, modes of which are given by stiff implicit systems of high-dimensional differential equations with strict one-sided constraint:

$$F(x, x', t) = 0, pr : g(x, t) < 0, t \in [t_0, t_k], x(t_0) = x_0, \quad (2)$$

where $x \in R^N$ is the vector of state variables, $t \in R$ is the argument, $F : R^N \times R^N \times R \to R^N$ is a continuous vector-function for given mode of HS, $g : R^N \times R \to R$ is the event-scalar function or the guard, x_0 is the value at the initial point t_0.

The problem (2) is usually stiff that leads to the application of implicit numerical formulas required Jacobi matrix inversion. Due to the ease of implementation and good

This work was supported by grant 14-01-00047-a from the Russian Foundation for Basic Research, RAS Presidium project № 15.4 "Mathematical modeling, analysis and optimization of hybrid systems".

Yu.V. Shornikov is with the Design Technological Institute of Digital Techniques Siberian Branch of Russian Academy of Science, Novosibirsk, Russia (e-mail: shornikov@inbox.ru).

M.S. Myssak, D.N. Dostovalov is with the Department of Automated Control Systems, Novosibirsk State Technical University, Novosibirsk, Russia (e-mails: maria.myssak@gmail.com, dostovalov.dmitr@gmail.com).

accuracy and stability properties Rosenbrock type methods [3, 4] are widely used in solving stiff problems.

III. ARCHITECTURE OF INSTRUMENTAL ENVIRONMENT

Development of simulation languages, simulators, simulation systems, etc. is essentially influenced by the CSSL (continuous system simulation language) Standard 1968 [4]. Although forty years old, the structures defined in CSSL Standard are used up to now. End of 90ties, CSSL extended to implicit systems, while a new modelling language, Modelica, was introduced. In principle, the modelling paradigm changed from signal flow − oriented modelling (explicit systems) to power − oriented modelling (implicit systems), from "causal" signal modelling to "acausal" physical modelling. The early CSSS standard determined basic necessary features for a simulator, the late developments to implicit systems fixed extended features for simulation systems − both referred as classical CSSL features. In 1968, the CSSL standard set first challenges for features of simulation systems, defining necessary basic features for simulators and a certain structure for simulators.

The CSSL standard also defines segments for discrete actions, first mainly used for modelling discrete control. So-called DISCRETE regions or sections manage the communication between discrete and continuous world and compute the discrete model parts. For incorporating discrete actions, the simulation engine must interrupt the ODE solver and handle the event. For generality, efficient implementations set up and handle event lists, representing the time instants of discrete actions and the calculations associated with the action, where in-between consecutive discrete actions the ODE solver is to be called. In order to incorporate DAEs and discrete elements, the simulator' s translator must now extract from the model description the dynamic differential equations (derivative), the dynamic algebraic equations (algebraic), and the events (event i) with static algebraic equations and event time, as given in Fig. 1 [5] (extended structure of a simulation language due to CSSL standard). In principle, initial equations, parameter equations and terminal equations (initial, terminal) are special cases of events at time t = 0 and terminal time. Some simulators make use of a modified structure, which puts all discrete actions into one event module, where CASE - constructs distinguish between the different events.

Simulation environment of complex dynamical and hybrid systems called ISMA (translated from Russian "Instrumental Facilities of Machine Analysis") is developed at the department of Automated control systems of Novosibirsk state technical university (Russia) [6].

Specification of hybrid systems is carried out using graphical and symbolic languages that are the system content of instrumental environment. Analytical content is provided by numerical methods and algorithms for computer analysis corresponding to the chosen class of systems and methods for solving these models. ISMA environment is developed subject to simplicity of description of dynamical and hybrid

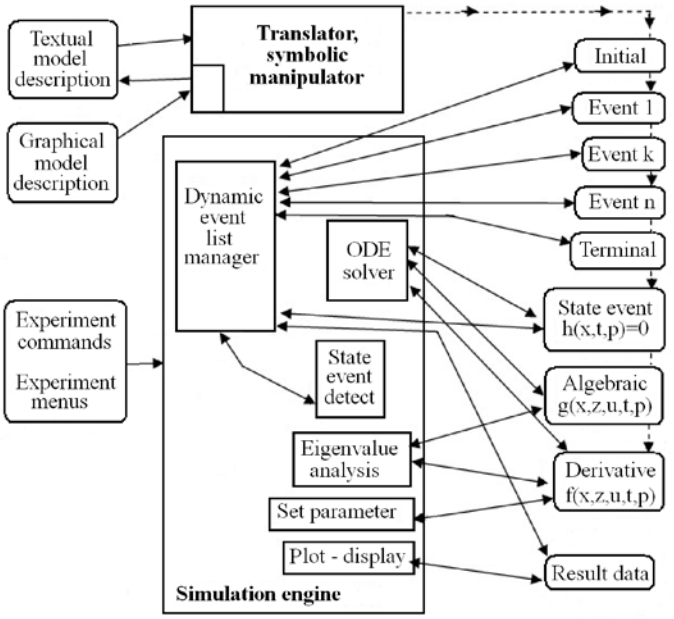


Fig. 1. Extended structure of a simulation system due to extensions of the CSSL standard with discrete elements and with DAE modelling

models in the language that is maximally close to the object language. Main features of ISMA are the following:

- Composition of hybrid models is carried out in visual structural-textual form;
- Structural form of model description corresponds to the classical description of systems by block diagrams and includes all necessary components such as integrators, accumulators, amplifiers, signal sources, nonlinear elements and others;
- Language of symbolic specification is approached maximal to the language of mathematical formulas;
- Special module for specification of problems of chemical kinetics in the language of chemical reactions which automatically translates them into a system of differential equations;
- A variety of traditional and original numerical methods included methods that are intended for the analysis of ODE systems of medium and high stiffness;
- Computer simulation in real time;
- Graphic interpreter called GRIN provides a wide range of tools for analysis and visualization of simulation results such as scaling, tracing, optimization, displaying in the logarithmic scale and phase plane;
- Extension of system functionality by adding new typical components and numerical methods.

Architecture of ISMA software package (Fig. 2) is designed [7] in accordance with CSSL to unify existing mathematical program software for analysis of problems in various object domains: chemical kinetics, automation, electricity, etc.

ISBN: 978-1-61804-251-4

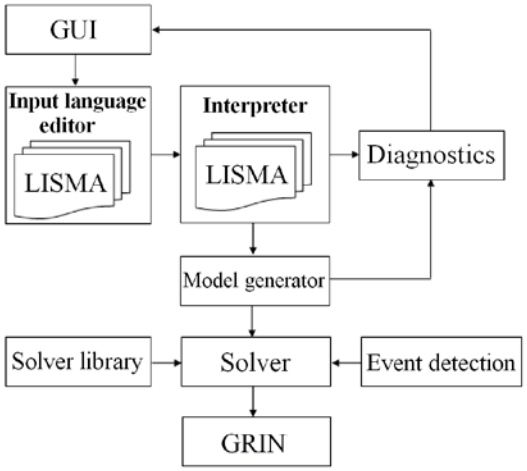


Fig. 2. ISMA architecture

IV. LIBRARY OF NUMERICAL METHODS

Peculiarities of numerical analysis are defined by the configuration and implementation of the solver in the scheme interpreter. Solver is configured to numerical analysis not only of smooth dynamical systems but also systems with ordinary discontinuity and stiff systems [2, 8]. For the analysis of the stiff modes new original m- phasic methods of p - order (Table I), developed by the authors, are included in the solver library.

TABLE I. LIBRARY OF NUMERICAL METHODS

Method (p , m)	**Description**
DISPF (5, 6)	Stability control, systems of medium and low stiffness
RADAU5 (3, 3)	Stiff systems
DISPF1_RADAU	Adaptive method DISPF in combination with RADAU5 with stiffness control, essentially stiff systems
DP78ST (8, 13)	Stability control, variable order and step, systems of medium stiffness and high precision
RKF78ST (7, 13)	Stability control, variable order and step, systems of medium stiffness and high precision
RK2ST (2, 2), RK3ST (2, 3)	Explicit methods with stability control for analysis of non-stiff systems
DISPS1	Algorithm of variable order with adaptive stability region
MK22 (2, 2), MK21 (2, 2)	Freezing of Jacobean matrix, stiff systems
MK11F	Algorithm of analysis of implicit problems

Libraries of standard blocks and numerical methods are implemented as independent application modules that are loaded at run time.

This approach allows to allocate in the application programming interface (API) a set of functions and classes required for the implementation of element libraries and numerical methods. Using the API any user with basic knowledge of object-oriented programming able to develop and built in the system new typical elements and numerical methods without recompiling the entire system.

V. EVENT DETECTION IN HYBRID SYSTEMS

The correct analysis of hybrid models is significantly depends on the accuracy of detection of the change of the local states of the HS. Therefore, the numerical analysis is necessary to control not only the accuracy and stability of the calculation, but also the dynamics of the event-function. The degree of approximation by the time the event occurred is defined by the behavior of event driven function.

Analyze the behavior of the event function $g(x,t)$. Let the method of the form $x_{n+1} = x_n + h_n\varphi_n$, where function φ_n is calculated in point t_n, is used for calculations.

Then the event-function $g(x,t)$ at point t_{n+1} has a form $g_{n+1} = g(x_n + h_n\varphi_n, t_n + h_n)$. Decomposing the g_{n+1} in a Taylor series and taking into account the linearity of g_{n+1}, we obtain the dependence of g_{n+1} of the projected step h_n:

$$g_{n+1} = g_n + h_n\left(\frac{\partial g_n}{\partial x}\cdot\varphi_n + \frac{\partial g_n}{\partial t}\right). \qquad (3)$$

Theorem. The choice of the step according to the formula

$$h_n = (\gamma - 1)g_n \Big/ \left(\frac{\partial g_n}{\partial x}\cdot\varphi_n + \frac{\partial g_n}{\partial t}\right), \gamma \in (0,1), \qquad (4)$$

provides the event-dynamics behavior as a stable linear system, the solution of which is asymptotically approaching to the surface $g(x,t) = 0$.

Proof. Substituting (4) in (3), we have $g_{n+1} = \gamma g_n$, $n = 0,1,2,...$ Converting recurrently this expression we get $g_{n+1} = \gamma^{n+1} g_0$. Given that $\gamma < 1$, then $g_n \to \infty$ takes place when $n \to \infty$. In addition, condition $\gamma > 0$ implies that function g_n does not change sign. Therefore, when $g_0 < 0$, $g_n < 0$ will be valid for all n. Then the guard condition will never cross the potentially dangerous area $g(x_n, t_n) = 0$, which completes the proof.

A. Control of event function in the integration algorithm

We complete the implicit problem's integration algorithm by the algorithm of the step control that takes into account the event function dynamics.

Let the solution x_n and $y_n = x_n'$ at the point t_n is calculated with the step h_n. In addition, the new accuracy step

h_{n+1}^{ac} is computed by the formula (4). Then the approximate solution at the point t_{n+1} is calculated as follows

Step 1. Calculate the functions

$$g_n = g(x_n, t_n), \frac{\partial g_n}{\partial x} = \frac{\partial g(x_n, t_n)}{\partial x}, \frac{\partial g_n}{\partial t} = \frac{\partial g(x_n, t_n)}{\partial t}.$$

Step 2. Calculate $g_n' = \frac{\partial g_n}{\partial x}\varphi_n + \frac{\partial g_n}{\partial t}$, where $\varphi_n = y_n$.

Step 3. If $g_n' < 0$, then $h_{n+1} = h_{n+1}^{ac}$ and go to the Step 6.

Step 4. Calculate the new "Event" step h_{n+1}^{ev} by the formula

$$h_{n+1}^{ev} = (\gamma - 1)\frac{g_n}{g_n'}.$$

Step 5. Calculate the new step h_{n+1} by the formula

Step 6. Go to the next integration step.

In the Step 3, unlike the previously presented algorithm [9], we determine the direction of event-function change. Near the boundary regime denominator (4) will be positive, and away from the boundary $g(x,t) = 0$ it becomes negative. Then, defining the direction of event-function change, we do not impose any further restrictions on the integration step if the event-function is removed from the state boundary.

B. Tests

To illustrate the capacity for work of the proposed algorithms consider a simple hybrid system – a bouncing ball. Modal behavior can be given by an implicit system of differential equations

$$y' - v = 0, v' + a = 0, \quad (5)$$

where y is the height from the surface of the ball rebound, v is the ball velocity, a is the free fall acceleration. An event occurs at the moment when the ball touches the rebound surface $y = 0$, therefore the event function takes the form $g = -y$, and the predicate $pr: -y < 0$. At the moment of rebound the ball changes the moving direction. Let the rebound is inelastic, then when the event occurs the velocity changes according to the rule $v = -\alpha \cdot v$, where α is the retention rate of speed.

Moments of the ball rebound from the surface and values of the variable h when the event occurs are shown in Figure 3. A significant error $\varepsilon_1 \approx 0.75$ in detection of event changes is made when calculating event function without dynamics control (Fig. 3a). This leads to violation of condition of one-sidedness of the events and as a result to erroneous global solution. Using of algorithm of asymptotic approximation to the regime border (Fig. 3b) provides about an order more accurate detection of the moment when regime of HS has changed $\varepsilon_2 \approx 0.06$.

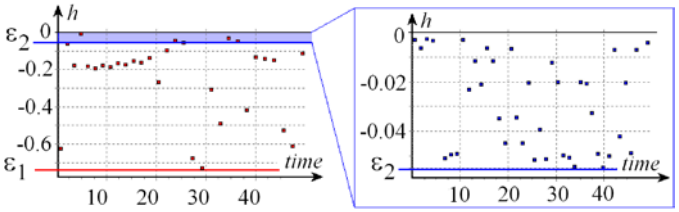


Fig. 3. Moments of event detection: a) without dynamics control; b) with asymptotic approximation to regime border

VI. SIMULATION OF FAULT IN AN ELECTRICAL NETWORK

As an illustration of a new class of systems and as a test case for the formulated algorithms analyze the model of three-phase fault in an electrical network. Schematic diagram of the electrical power system (EPS) built in graphics editor of the ISMA instrumental environment is shown in Fig. 4.

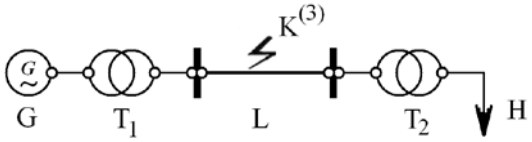


Fig. 4. Schematic diagram of the electrical network

Considered scheme consists of generator G, transformers T_1, T_2, line L and load H. In the equivalent circuit in Fig. 5 capacitive conductivity of the line and transformer non-load loses are not taken into account and the load is taken into account by approximately active and inductive reactance.

Transient is initiated by the contact closure K. In this case previously established mode of power system is changed to the new mode corresponding fault and another system configuration. Thus, the model is a two-mode hybrid system (HS) [2].

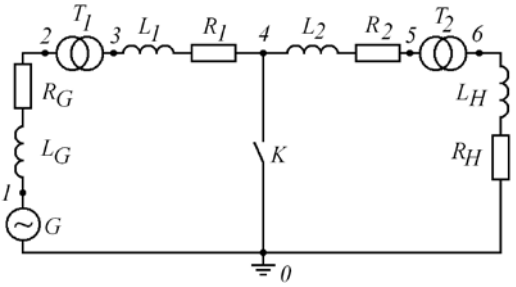


Fig. 5. Schematic diagram of the electrical network

The discrete behavior of the hybrid system is illustrated by the state chart shown in Fig. 6. State init corresponds to the functioning of EPS before the fault. Switching to state short corresponded to the fault condition occurs when a logical predicate pr is carried out.

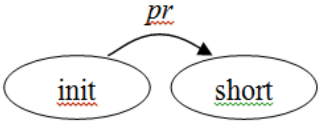


Fig. 6. Behavior map

In the graphics editor of schematic diagrams of EPS hybrid behavior is specified in the configuration editor window for the equivalent circuit of a transmission line L as shown in Fig. 7.

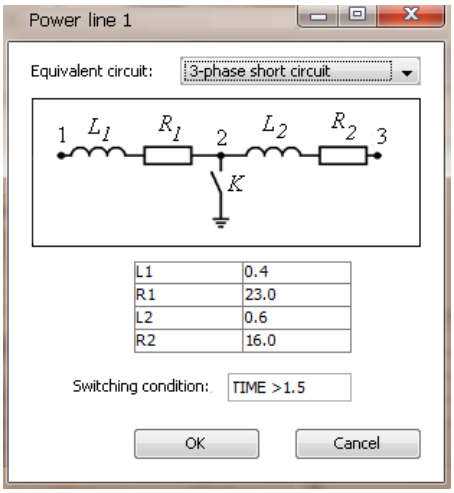


Fig. 7. Configuring the parameters of the equivalent circuit

The mathematical model is composed by Park-Gorev equations in rotating coordinate system (d,q) associated with the generator rotor G. Let the axis q is ahead of the axis d. Obtain a system of equations for the generator G

$$u_{1d}\cos(\theta-t)+u_{1q}\sin(\theta-t)+ri_{Gd}+$$
$$+L_d\frac{di_{Gd}}{dt}-L_{ad}\frac{di_f}{dt}-L_{ad}\frac{di_g}{dt}-\left(L_q i_{Gq}-L_{aq}i_h\right)\omega=0,$$
$$-u_{1d}\sin(\theta-t)+u_{1q}\cos(\theta-t)+ri_{Gq}+$$
$$+L_q\frac{di_{Gq}}{dt}-L_{aq}\frac{di_h}{dt}+\left[L_d i_{Gd}-L_{ad}\left(i_f+i_g\right)\right]\omega=0,$$
$$-u_f+r_f i_f+L_f\frac{di_f}{dt}+L_{ad}\frac{di_g}{dt}-L_{ad}\frac{di_{Gd}}{dt}=0,$$
$$r_g i_g+L_g\frac{di_g}{dt}+L_{ad}\frac{di_f}{dt}-L_{ad}\frac{di_{Gd}}{dt}=0,$$
$$r_h i_h+L_h\frac{di_h}{dt}-L_{aq}\frac{di_{Gq}}{dt}=0,$$
$$\frac{d\omega}{dt}=\frac{T_\partial+\left[\left(L_d-L_q\right)i_{Gq}+L_{aq}i_h\right]i_{Gd}-L_{ad}i_{Gq}\left(i_f+i_g\right)}{T_J},$$
$$\frac{d\theta}{dt}=\omega.$$

Here the index f refers to the excitation winding and indices g and h refers to the longitudinal and transverse damper contours respectively.

Equations for the area of the equivalent circuit 1-2:

$$u_{1d}-u_{2d}=R_G i_{12d}+L_G\left(\frac{di_{12d}}{dt}-i_{12q}\right),$$
$$u_{1q}-u_{2q}=R_G i_{12q}+L_G\left(\frac{di_{12q}}{dt}-i_{12d}\right).$$

For the areas 3-4, 4-5 and 6-0 equations will have a similar form.

Equations for the transformer T_1:

$$u_{3d}=K_{T1}\left(u_{1q}+\sqrt{3}u_{1d}\right),\ u_{3q}=K_{T1}\left(\sqrt{3}u_{1q}-u_{1d}\right),$$
$$i_{34d}=K_{T1}\left(i_{12q}+\sqrt{3}i_{12d}\right),\ i_{34q}=K_{T1}\left(\sqrt{3}i_{12q}-i_{12d}\right).$$

Here K_{T1} is a transformation ratio. Equations for the transformer T_2 are treated similarly.

Equations of the first Kirchhoff's law for point 1:

$$i_{Gd}\cos(\theta-t)-i_{Gq}\sin(\theta-t)=i_{12d},$$
$$i_{Gd}\sin(\theta-t)+i_{Gq}\cos(\theta-t)=i_{12q}.$$

When an event corresponded to the fault occurs in HS, the voltage in point 4 is equated to zero $u_{4d}=u_{4q}=0$. In this case in the equivalent circuit two independent contours are formed. The equations for sections of the contours remain the same.

Plots of some state variables obtained in ISMA are shown in Figure 8. Calculation results correspond to theoretical statements and coincide with results obtained in MATLAB.

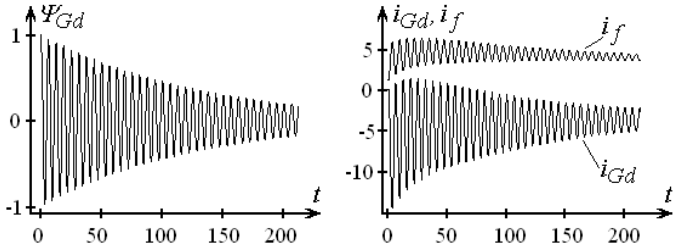


Fig. 8. Simulation results in ISMA

VII. CONCLUSIONS

In this paper the new class of hybrid systems within the ISMA instrumental environment, the modal behavior of which is defined by a system of ODE unresolved with respect to the derivative, is introduced. Architecture of instrumental environment is designed in accordance with CSSL standard. The new original method of switching point's localization is proposed. The algorithm easily complements the existing numerical solvers based on explicit and semi-explicit schemes including the proposed algorithm of implicit problem's analysis. Model of new HS system class is presented and studied in ISMA.

ISBN: 978-1-61804-251-4

REFERENCES

[1] Yu.V. Shornikov, "Numerical modeling of dynamic processes in electric power systems as a strategic management tool," Yu.V. Shornikov, I.N. Tomilov, D.N. Dostovalov, M.S. Denisov, Scientific bulletin of the NSTU, vol. 4, no. 45, pp.129-134., 2011.

[2] E.A. Novikov, Yu.V. Shornikov, Computer simulation of stiff hybrid systems: monograph, Novosibirsk, Russia: Publishing house of NSTU, 2012.

[3] H.H. Rosenbrock, "Some general implicit processes for the numerical solution of differential equations," Computer, vol. 5, pp. 329–330., 1963.

[4] Yu.V Shornikov, D.N. Dostovalov, M.S. Myssak "Simulation of hybrid systems with implicitly specified modal behavior in the ISMA environment ", Humanities and scinence university journal, no. 5, pp. 175-182, 2013.

[5] F. Breitenecker, N.Popper, "Classification and evaluation of features in advanced simulators," Proceedings MATHMOD 09 Vienna, Full papers CD Volume, 2009.

[6] Yu.V. Shornikov, "Instrumental tools of computerized analysis (ISMA)," Yu.V. Shornikov, V.S. Druzhinin, N.A. Makarov, K.V. Omelchenko and I.N. Tomilov, Official registration license for computers 2005610126, Moscow, Rospatent, 2005.

[7] Yu.V. Shornikov, M.S. Myssak, D.N. Dostovalov et al, "Using ISMA Simulation Environment for Numerical Solution of Hybrid Systems with PDE", Proc. Computer Modeling and Simulation, Sankt-Petersburg, Rusia, pp. 101-108, 2014.

[8] D.N. Dostovalov "Computer simulation and algortithms of numerical analysis of hybrid systems", Control system and information technologies, vol. 53, no 3.1 pp. 128-133, 2013.

[9] Yu.V. Shornikov, D.N. Dostovalov, M.S. Myssak et al, "Specification and analysis of discrete behavior of hybrid systems in the workbench ISMA", Open Journal of Applied Sciences, vol. 3, no. 2b, pp. 51-55, 2013.

ISBN: 978-1-61804-251-4

Mathematical Modelling as Analytical Instrument of Research of Innovative Processes

GALINA YU.SILKINA
Department "Information systems in economy and management"
St. Petersburg State Polytechnical University
195251, St. Petersburg Polytechnicalst., 29
RUSSIA
e-mailgalina.silkina@gmail.com

Abstract:- Paper is devoted to the analysis of fundamental properties of innovative processes and their reflection in mathematical models. The general regularities of innovative processes as a basis of conceptual modeling are considered. Models of formation of knowledge are presented. Analogies between the physical phenomena of transfer and diffusion of the innovations, the distributions of innovations represented in models are proved.

Key-Words:- Innovative process, innovation, knowledge, addition of knowledge, diffusion, equation of transfer, communication network

1 Introduction

The present time we quite reasonably call an era of innovations. To that there are strong reasons. Innovations often act as a key factor of success,make a basis of long-term competitiveness of the enterprises and firms.

Innovative activity allows to win competitions for sales markets by development and release of new types of production more attractive to potential consumers, to increase efficiency of a production activity due to modernization of technologies, provide market success with realization of new forms of cooperation and transition to more perfect business models.

But the economic world is arranged in such a way that any progress in technologies, a grocery variety, ways of the organization of production and management is carried out through rejection or adjustment before existing ways of action. These processes of creative destruction, i.e. continuous updating of economic activity on the basis of innovations, were described still by Joseph Schumpeter, who considered this theory as the most adequate interpretation of economy during an era of big business.

2 Key directions in research of innovative processes

Essential (and ambiguous) influence of an innovative component on economic development proves relevance of its academic studying. In scientific researches of innovative processes the priority belongs to conceptual models. Development of the concept precedes mathematical modeling and is an important investigation phase, where lines of model, its elements and communications are defined. Besides, conceptual discussion of innovative problems and methods of their permission is capable to create essential prerequisites for progress in many spheres of public life.

It is possible to allocate the following key directions in researches of innovative processes:

- innovations as those;
- sources of innovations;
- diffusion and replacement of innovations.

These questions are discussed in the present paper.

3 Analysis and modeling of innovative processes

The fundamental properties of innovative processes and their representation in mathematical models are studied in this section.

3.1 General regularities of innovative processes

There are bases to claim that a necessary condition of development of production and progress of society is the accumulated scientific knowledge. Development of scientific ideas and knowledge creates theoretical base for improvement of products, modernization of technologies, developments of organizational forms. Scientific progress is the prerequisite not only various improvements, but also more radical transformations in economic and, first of all, in the production sphere.

This situation is postulated as an axiom in existing international documents. In particular, the OsloManual specifies that knowledge in all forms plays a crucial role in economic progress [4].

In the modern world knowledge acquire the status of a specific product, and the science turns into a special resource which reduces expenses of traditional resources and at the same time opens new opportunities. The science role admits economic progress everywhere. Many economists consider today a contribution of scientific knowledge to the long-term economic growth by more powerful, than a contribution of such classical factors, as work and the capital.

The circumstance that continuous scientific progress is the heart of innovative activity, gives the grounds to characterize it as process. Dictionaries define process as consecutive change of the phenomena, states in development something, i.e. in the concept of process the idea of a continuity is put. To a continuity of innovative process especially points also the Oslo Manual [4].

At the same time continuous process of scientific creation and the subsequent embodiment of its results in concrete production, organizational and administrative decisions, proceeds in the form of realization of separate innovations. Complete innovative process breaks up to the individual innovations, connected in a whole as is substantial, on their structural and hierarchical subordination, and dynamically by means of existential communications. In total formulated arguments define the discrete and at the same time continuous nature of innovative process and differentiates analysis methods on dominating nature of process.

The local aspect (at the chosen priority of discrete nature of process) closely correlates with the point of view of J. Schumpeter, who considered that high-quality changes of economy have essentially discrete character and are based on realization of innovations. Here an innovation is allocated as the independent unit, separate object of supervision and studying, and its essence is investigated. The principles of complexity and system approach have to become a methodological basis of this approach. Each innovation needs to be considered as a certain unity, integrity, system phenomenon.

Recognition of a dominating role of continuous nature of process in a context of system approach (integrated aspect of the analysis) assumes studying of interaction of innovations.

Within integrated approach study coexistence of innovations, mechanisms of their parallel and consecutive connection, the competition, change. In other words, laws, under which innovations in interaction with each other form a continuous stream are investigated. Research of separate innovations as their properties substantially define properties of innovative process as a whole has to become primary. Besides, despite variety of their forms, initiation conditions, specifics of implementation and distribution in various spheres, innovations possess the whole set of properties, invariant concerning these forms, conditions and specifics.

Basis of each innovation, as well as innovative process as a whole, is the new knowledge. Potential of the scientific idea which is cornerstone of an innovation, substantially defines its success. However,very seldom only one scientific theory or the unique fact lies at the heart of an innovation. Usually it represents result of a combination of several types of knowledge. In other words, important feature of innovations is convergence of knowledge on which it is based. Until all necessary knowledge won't connect together, time counting, it is necessary in order that the innovation became reality, at all won't begin.

Probably, as the most impressive example of convergence of knowledge and the technologies constructed on their basis the phenomenon which noticed by researchers at the end of the XX century and has received the name of NBIC convergence (N – Nano, B – Bio, I – Info, C – Cogno) serves [3].

This scientific concept treats convergence not only as association and mutual influence of knowledge, but also interpenetration of technologies. Borders between separate technologies are erased, and the most significant results arise within interdisciplinary work on a joint of scientific areas. Concerning NBIC convergence speak even about partial connection of branches expected in the long term in uniform scientific and technical area of the knowledge which objects of studying there are practically all levels of the organization of a matter – from the molecular

ISBN: 978-1-61804-251-4

nature of substance before processes of information exchange.

This circumstance allows founders and adherents of the concept of NBIC convergence to characterize the present stage of scientific and technical progress as the new evolutionary defining factor, capable to provide high-quality change of technological capabilities, individual development of the person and society as a whole. However it is obvious that a practical embodiment of this concept –event of more or less long-term future.

One more characteristic of an innovation is its duration. For realization of an innovation it is necessary to gain, at first, new knowledge, secondly, to use them for creation of new equipment, technology, a method, and, at last, to find expedient methods of application created in economy or public life. The retrospective analysis shows that usually a lot of time passes before all necessary knowledge will be gained, will connect together, and on their basis there will be a new technology. It is followed sometimes in addition by longer period, while the new technology won't turn into goods, processes or the services recognized as society and demanded in the market.

Only the practical demand defines essence and success of an innovation. The task consists not in opening, inventing and realizing something new, but in introducing that will give real effect. The idea only then becomes an innovation, when the decision based on this idea seeks for public recognition through large-scale use in practical activities of people.

For innovative processes dependence on previous development and the reached results ("the history matters") that will quite be coordinated with provisions of the evolutionary economic theory is characteristic. Russian scientists B. Kuzyk and Yu. Yakovetch analyzed the saved up experience in comparison to biological evolutionary processes and studied the similar phenomena in natural-science areas [6]. They came to a conclusion, thatin dynamics of innovations action of regularities of heredity, variability and selection is observed. Each innovation leans on the reserve of innovative development saved up by the previous development, inherits a genotype of transformed system and alters it in relation to the changed external and internal conditions, clearing of outdated elements and enriching new. Selection innovations from a set of possible is thus carried out. Many regularities of innovative process are explained by these provisions, its practical demand and public recognition are defined.

The innovation, which has gained recognition starts extending in space, i.e. innovation diffusion takes place and is observed. The Oslo Manual defines diffusion as a way what innovations extend by market and non-market channels from a place of their first realization to various consumers – the countries, to regions, branches, the markets and the enterprises [4]. Process of realization of innovations serves as the catalyst of the subsequent inventions and innovations. It belongs, first of all, to the innovations, being cornerstone of a new technological paradigm and possessing high potential of market penetration.

When using innovation in various areas of human activity its modifications adapted for specifics of these areas are created, the experience, allowing to improve it in these or those aspects and details is saved up. In other words, the innovation doesn't remain identical to that initial "image" which it had at the time of emergence of idea and its first realization. Moreover, modifications improving the first sample promote the fastest diffusion of an innovation in the economic environment. It is manifestation of one more phenomenon – the divergence of innovation. In synergetics a synonym of a divergence is the concept of bifurcation.

Concepts of convergence and divergence were initially used within separate natural-science areas (mechanics, biology) for their specific requirements. Today they get the status of general scientific categories. The understanding of was so far created that these phenomena designate features of emergence, existence, developments of objects of any nature, are inseparably linked with the general concepts of stability/instability, balance/nonequilibrium, integrity/discontinuity, and are shown in functioning practically all systems, their structures, the relations, communications.

These concepts are antonymous in essence. Convergence is a manifestation of unity, discovery of integrity, potential or real. Divergence conducts to instability when the object, before whole, shares on components and branches (as on a tree), more precisely, on the new directions of further changes. In this situation the image of the branching tree, used in science and practice for the description of evolution and development, arises not so incidentally – it visually reflects possible conditions of object during various changes. This approach can be quite applied to representation of innovative processes and separate innovations taking into account one more feature of the last – their limitation.

ISBN: 978-1-61804-251-4

Each innovation has the limit. R. Foster, within 22 years (1982 – 2004) being one of heads of the consulting company McKinsey, especially emphasized that if the limit, how try is reached, advance forward is impossible. Existence of a limit can depreciate all earlier existing practices and cause a gap in development; misunderstanding of it often leads to wrong decisions.

Properties of continuous innovative processes considerably are defined by types and forms of a combination of separate innovations. Essentially these forms can be divided into two classes: parallel and consecutive connection.

Parallel connection of innovations is caused by that in one innovation, owing to the principle of convergence some scientific facts, opening and inventions are, as a rule, put. They partially in combination with additional knowledge become a basis for other innovations. As a result, at various innovations the general theoretical roots that conducts in formation of treelike structures are possible.

The hypothesis that innovations appear in economic system not evenly, and in the form of clusters was stated still by J. Schumpeter. The cluster is a set of the radical innovations concentrated on a certain interval of time and in certain area of economic space.

Rather convenient instrument of model representation of similar processes is the device of the theory of counts. Expediency of application of models of counts is caused by their presentation and possibility of display of a large number of dependences of qualitative type. Innovations in such model form some kind of family tree each branch on which can both flush, and to be deadlock. The branching high probability, the develops economy quicker. In a form of a crown it is possible to distinguish a growing tree from the perishing. To innovations there is the same: they or give the next portion of novelties, or dry.

The idea that technological opening are connected by some tree of communications, stopped being long ago the specific scientific concept. The studied scientific hypothesis consists that the crown has to be added with "root system". The root system is made by the knowledge which is cornerstone of innovations.

3.2 Properties of knowledge and their representation in mathematical models

Knowledge does not only represent independent value, but actively influences other factors of production, generates and reproduces multiplicative effect in relation to them. Scientific ideas and development create a necessary reserve, are the base of innovative potential – abilities to production of new knowledge, technical and technological decisions and, thereof, possibilities of further improvement. It defines need of studying of properties of knowledge, including, on the basis of model representations.

Knowledge is a special type of product. About it FritzMachlup wrote in his famous book "TheProduction and Distribution Knowledge in the United States" (1962), where three specific properties of this product were especially allocated:

- not disappearance at consumption;
- uniqueness;
- indivisibility (discretization).

Property of discretization defines a way of measurement of knowledge. It is necessary to record some nomenclature of knowledge (a discrete scale) to measure amount of knowledge in integers (pieces). For the first time this idea sounded in V. L. Makarov's works, where the classical model of intersectoral balance was adapted for the analysis of processes of formation and use of knowledge [7]. For creation of this model n of types of knowledge (the scientific directions) which current level of development was characterized by non-negative variables $x_i, i = 1, 2, \ldots, n$, and a science level of development as a whole – the n-dimensional vector $X = (x_1, x_2, \ldots, x_n)$, were allocated.

Further the list of properties of knowledge as product was expanded with inclusion of additional elements:

- invariancy concerning space;
- durability;
- sensitivity to time factor;
- expanded reproduction of knowledge;
- possibility of their repeated sale.

Invariancy concerning space means ability of knowledge rather freely, practically without expenses to extend on modern networks of communications. Property it is durable the fundamental knowledge which is based on the checked facts, generalizing them and scientifically reasonable possesses, first of all. At the same time the value of applied knowledge significantly depends on time. Expanded reproduction of knowledge is closely connected with their not disappearance at consumption. It essentially distinguishes knowledge from other products, first of all, the material. Moreover, the greatest value of knowledge consists in their abundance while material resources are appreciated on their rarity.

ISBN: 978-1-61804-251-4

Ideas give rise to the new ideas, the imparted (sold) knowledge remains with those who shares them (sells).

That for knowledge other algebra is characteristic, other than traditional arithmetics, was noticed and recorded already for a long time [2, 9]. The phenomenon of convergence of the knowledge, being cornerstone of an innovation, demands prime definition of operation of addition of knowledge.

We offer the next way of a graphic representation of operation of addition. From the allocated nomenclature we will connect a coordinate axis with each type of basic knowledge. We will represent the current level of development of knowledge a point on axes, and a knowledge increment – the directed piece (vector) going along this axis. Length of a vector shows increment size.

The result of addition of knowledge depends on a relative positioning of summable vectors concerning which two hypotheses are possible:

- vectorsare directed along one coordinate axis;
- vectorsare directed along different axes.

In the first case the property opposite to not disappearance at consumption – property of idempotence is shown. Formally mathematical law of idempotence is expressed by equality $a + a = a$. Substantially in the annex to knowledge means that in science including applied, already received result doesn't matter. Repeatedly formulated statement bears in itself no more information, than the statement formulated once. The invention of that is already invented, enriches with nothing mankind.

In fig. 1 two summable vectors and result of their addition are shown.

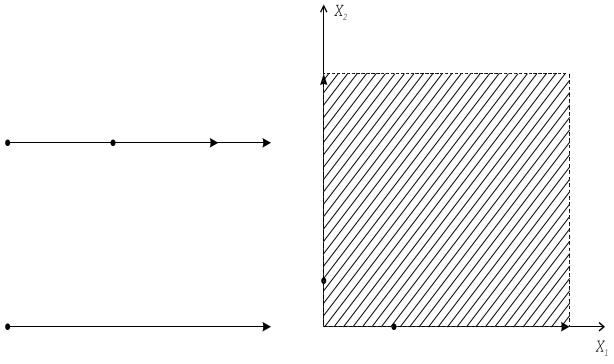


Fig.1: a) idempotent addition of basic knowledge; b) synergetic additionof basic knowledge

With the greatest community idempotent addition of knowledge can be formalized is presented by formula

$$\begin{aligned}\mu(a+b) &= \max\{\mu(a), \mu(b)\} \\ &= \mu(a)+\mu(b)-\mu(ab)\end{aligned}, \quad (1)$$

where $\mu(ab)$ – the general part of increments of knowledge*a*and *b*.

At addition of increments of knowledge in various key branches of their direct summation doesn't occur, but there is a synergetic effect. The essence of effect consists that as a potential point of the appendix of efforts on accumulation of knowledge there is available each point from formed by vectors of increments of area (fig.1b).

Thus the greatest potential has synergetic addition. V.L. Makarov entered the non-negative coefficients a_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, n$, showing intensity of use of each knowledge by "production" of other knowledge. For ensuring scientific progress the level of development of each demanded knowledge (and only the demand determines the value of any product, and knowledge including) has to be not below necessary value [7}.

The operation of summation traditional for classical intersectoralbalance, is replaced with operation of a capture of a maximum.

$$x_i \geq \max_{j=1,2,\ldots,n} \{a_{ij} x_j\}, \quad i = 1, 2, \ldots, n. \qquad (2)$$

At the same time need of the balanced development of science located. Until, when all necessary for development of some scientific direction of knowledge won't be received, progress in this direction is simply impossible.

Back, development of some applied, synthetic knowledge provides an increment of knowledge in basic fundamental sciences. It is equivalent to design of the directed piece representing an increment of knowledge, on basic vectors (fig. 2a).

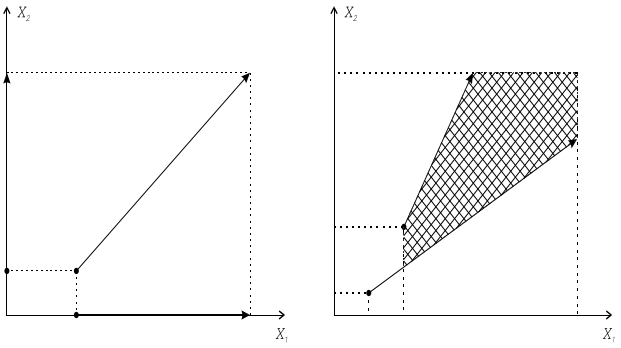


Fig 2: a)influence of applied knowledge on development of the basic knowledge; b) synergetic additionof applied knowledge

ISBN: 978-1-61804-251-4

At last, the result of parallel carrying out applied researches in two adjacent directions is presented in fig. 2b.

The shaded part of the plane shows new the field of knowledge, suitable for application

3.3 Models of diffusion of innovations

Within the concept of innovative development the central place is taken by a phenomenon of diffusion of innovations. The Oslo Manual especially emphasizes that without diffusion the innovation has no economic value [4]. If the innovation isn't new to the whole world (that, as a rule, is result of own unique development), its realization can be only result of transfer, i.e. diffusion.

The term "diffusion" came to the economic theory and economic practice from natural sciences – physics and chemistry. Historically interest to a problem of diffusion arose at the beginning of the XX century thanks to works J. Schumpeter. The first analytical models of this process were constructed only in the 1960th years.

According to J. Schumpeter's theory diffusion of innovations is process of cumulative increase in number of the managing subjects which have accepted an innovation. The subsequent researches of processes of distribution of innovations confirmed its scientific hypothesis [5, 8, 9]. They showed that dynamics of diffusion of an innovation is characterized from the point of view of a configuration in time by the following features:

- at the initial stage of production scales of distribution extend not too considerably. Rates of this expansion can be very high;
- at the main stage of the period the innovation wins the potential sphere of the effective application. Rates of expansion of scales of its production or use are at the level of the previous stage. They, as a rule, decrease eventually whereas pure gains considerably during the whole period are very high in scales of the potential sphere of distribution. The innovation extends on branch or group of consumers or forms new branch of a economy;
- at the final stage the innovation gets into economic spheres, limit from the point of view of comparative efficiency of its use. Growth rates are insignificant. In process of saturation of requirement or filling of the sphere of distribution gradually decrease to zero.

Dynamics of absolute majority of cumulative values submits to the laws, described by logistic curves. The elementary logistic curve is set by the differential equation

$$\frac{dz}{dt} = kz(b - z). \qquad (3)$$

Here the variable t is independent argument, $z = z(t)$ the current value of the analyzed parameter; $k > 0$ positive constant (proportionality coefficient); b the positive constant limiting from above value $z = z(t)$, which lower bound is equal to zero.

The differential equation (3) goes back to researches of E.M. Rogers and is usually treated as quantitative expression of action of the law of mutual transition of quantitative and high-quality changes in relation to cumulative processes [5]. In this scheme processes of diffusion of innovations at appropriate interpretation of parameters of this model lay down also.

The differential equation (3) is integrated in an explicit form and its decision is analytically set by a formula

$$z(t) = \frac{b}{1 + ce^{-bkt}}. \qquad (4)$$

Function graph is the *S*-shaped logistic curve. The geometry of a curve is defined by parameters of the equation and will be coordinated with the logic of process of distribution of innovations given above.

So, function $z = z(t)$ monotonously increases:

$$\frac{dz}{dt} = ckb^2 \frac{e^{-bkt}}{\left(1 + ce^{-bkt}\right)^2} > 0. \qquad (5)$$

Its schedule contains three obviously distinguishable sites – a site of initial lifting, a site of vigorous growth and a site of the fading growth. In initial timepoints when $z(t)$b there is much less, it practically coincides with exponenty and grows with an increasing speed. Eventually existence of limiting factors and limit existence is more and more shown.

Formally influence of limiting factors is shown that the site of vigorous growth also breaks into two parts – the increasing growth rate in some timepoint starts decreasing. This moment corresponds to an inflection point of a logistic curve $t_0 = \frac{\ln c}{bk}$ which is defined from a condition

$$\frac{d^2z}{dt^2} = 2ck^2b^3 \frac{e^{-bkt}\left(1 - ce^{-bkt}\right)}{\left(1 + e^{-bkt}\right)^3} = 0. \qquad (6)$$

The straight line $z = b$ is a horizontal asymptote of a logistic curve as approaching which both the

growth rate, and its absolute values decrease practically to zero.

That circumstance is essentially important that the coefficient of proportionality can change eventually: $k = k(t)$ [8, 9]. It is confirmed, in particular, by named researches where on the basis of the analysis of empirical data two liftings, two ascending waves on a classical logistic curve are revealed.

Authors connected the first of these liftings with internal properties of the innovation, its technical and technological features. The second – with the economic reasons, external in relation to an innovation and expressing readiness of society for introduction of the corresponding innovations and natural increase of demand for them.

The assumption of possible changes of coefficient leads to generalization of the basic equation of diffusion in shape

$$\frac{dz}{dt} = k(t)(z - b_1)(b_2 - z). \quad (7)$$

The solution of the generalized logistic equation is function

$$z(t) = b_1 + \frac{(b_2 - b_1)\theta(t)}{\theta(t) + c}, c > 0, \quad (8)$$

$$\theta(t) = e^{(b_2 - b_1)\int_0^t k(u)du}. \quad (9)$$

In the generalized logistic model time passes not linearly, and is "somewhat proportional" to function $k(t)$. The type of the decision essentially depends on a type of this function. Than less function $k(t)$ reminds a constant, especially the events described by this model not linearly develop.

The basic logistic equation and its modifications are the differential equations with the concentrated parameters. In them time acts as the only independent variable. It models only a configuration of process of diffusion in time and doesn't reflect spatial features of this process.

At the same time innovative process is characterized not only temporary, but also spatial parameters. The interrelation between natural-science processes of transfer and diffusion of innovations proves prospects of adaptation of the categories borrowed from fundamental natural sciences and models to the analysis of innovative processes.

The physical phenomena of transfer are based on the general regularities (existence of a gradient of some physical value and aspiration of system to an equilibrium state). Their existential features can be presented by the generalized equation of transfer.

In thermodynamics and molecular physics processes of transfer are generally represented by the equation in partial derivatives

$$\frac{\partial a}{\partial t} = \nabla \cdot [D(a,r)\nabla a(r,t)] + f(r,t). \quad (10)$$

The value $a(r,t)$, depending on spatial coordinates and time, characterizes a certain property of physical system, $D(a,r)$ – the generalized coefficient of transfer in point *r*(physical parameter of the environment in which transfer process), ∇ – the vector differential Hamilton's operator in *n*-dimensional space set by components $\left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n}\right)$, $f(r,t)$ – the function describing a source of substance (energy) proceeds; the point designated a scalar product of vectors.

At constant coefficient of transfer the *D* equation is reduced to the linear differential equation in private derivatives

$$\frac{\partial a}{\partial t} = D\Delta a(r,t) + f(r,t), \quad (11)$$

where $\Delta = \nabla^2 = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$ – Laplace's operator.

In case of one-dimensional diffusion process (when size and depends only on one spatial coordinate *x*) with transfer coefficient the *D* equation has an appearance

$$\frac{\partial}{\partial t} a(x,t) = \frac{\partial}{\partial x} D \frac{\partial}{\partial x} a(x,t) + f(x,t), \quad (12)$$

at constant *D* – a form

$$\frac{\partial}{\partial t} a(x,t) = D \frac{\partial^2}{\partial x^2} a(x,t) + f(x,t). \quad (13)$$

Just as the equation of logistic dynamics in theories of innovations, the equation of transfer arises for empirical reasons. It expresses proportionality of a stream of substance (energy) to the difference of concentration (temperatures) in areas.

Application of the equation of transfer to the analysis of diffusion of innovations assumes exact identification and information filling of parameters of model, first of all, distances between elements (subjects of managing). This distance has to be defined differently, than in geographical sense. R. Cowan, P.Jonard offered the following approach to intellectual identification of subjects of managing and measurement (cognitive) distance between them [1].

Through level $v_{i,k}^t$ of knowledge of category*k*, $k \in \{1,2,\ldots,K\}$, which the agent *i*in *t*timepoint possesses is designated. The distance between *i* and *j* in a considered timepoint is measured by size

$$\Delta^t(i,j) = \max\left\{w^t,\ 1/w^t\right\} - 1, \qquad (14)$$

$$w^t = \left|v_i^t\right| / \left|v_j^t\right|, \qquad (15)$$

where $\left|v_\alpha^t\right|$ – standard Euclidean norm of a vector $v_\alpha^t = \left(v_{\alpha,1}^t, v_{\alpha,2}^t, \ldots, v_{\alpha,K}^t\right)$:

$$\left|v_\alpha^t\right| = \sqrt{\sum_{k=1}^{K}\left(v_{\alpha,k}^t\right)^2}, \quad \alpha \in \{i,j\}. \qquad (16)$$

Other options of a task of a metrics are possible also.

Prospects of adaptation of the classical equation of transfer to the analysis of processes of distribution of innovations substantially are defined by that this equation is integrated in an explicit form.

In the elementary (one-dimensional) case the fundamental solution of the uniform equation with constant, not dependent from *x* and *t* coefficient of transfer of *D*, the initial condition expressed δ-function of Dirac $a(x,0) = \delta(x) = \begin{cases}\infty, & x = 0\\ 0, & x \neq 0\end{cases}$, and a boundary condition $a(\infty,t) = 0$ is set by a formula

$$a(x,t) = \frac{1}{2\sqrt{\pi D t}} e^{-\frac{x^2}{4Dt}}. \qquad (17)$$

At data initial and boundary conditions an average square of removal the particles (or the corresponding characteristic of distribution of temperature) from the initial point

$$Mx^2 = \int_{-\infty}^{+\infty} x^2 a(x,t)dx = 2Dt. \qquad (18)$$

Function (15), like a logistic curve, allows research by means of the classical analysis. So, at each fixed *t* it monotonously decreases:

$$\frac{\partial a(x,t)}{\partial x} = -\frac{1}{4D\sqrt{\pi D t}} x e^{-\frac{x^2}{4Dt}} < 0. \qquad (19)$$

Its schedule (section) has the inflection point $x_0 = 2Dt$, calculated from a condition $\frac{\partial^2 a(x,t)}{\partial x^2} = 0$

.

The last will be coordinated with logic of transfer process. Concentration of particles is especially great near the indignation center, but quickly decreases in process of removal from it. Nature of process changes (its speed falls) when passing the inflection point which situation is defined by a present situation of time and transfer coefficient.

Data initial and boundary conditions are quite adequate to processes of diffusion of innovations. The innovation plays a role of the dot impulse, entering indignation into the system. Influence of an impulse is really observed only within some territory. During considerable removal (that is entered distance) from a source its role is negligible. These and other above-stated reasons far don't settle substantial analogies between the physical phenomena of transfer and distribution of innovations.

Speed and configuration of physical processes of transfer depends on properties of the environment – its permeability for diffusion, which is represented in diffusion coefficient. The indicator of innovative permeability of the economic environment, defined by the parameter of its innovative conductivity and considering existence of the factors constraining distribution of innovations can become analog of this coefficient.

At the heart of the mechanism of diffusion process of data transmission, information and knowledge lies. Exchange of information about advantages and shortcomings of innovations leads to that uncertainty in relation to the last decreases. As a result of an innovation are used by more wide range of managing subjects.

The central value thus gets process of acceptance of innovations in the group of persons, connected with each other and with environment a communication network. The network covers all channels on which the information on an innovation, its properties, effect and experience of use is spread.

On nature of distribution of information allocate communication networks of two types:

- simple imitation people around;
- the network including specialized channels of communications [1].

The second version is most characteristic for production and economic innovations. Here the speed of diffusion of an innovation doesn't depend on contacts between persons, and is defined only by the factors, characterizing properties of channels of communications. Individuals or the enterprises receive data on an innovation mainly from external sources. Some intermediary in this case acts as information source. The individual or such sources of information, as printing editions, television, advertising of any kind can be the intermediary. Functions of the intermediary consist in the

ISBN: 978-1-61804-251-4

message of data on an innovation, its advantages and shortcomings, experience of use to the persons, potentially capable to accept it.

At the first case the main thing is the effect of imitation at which persons in considered group or set borrow an innovation each other, communicating. Diffusion of an innovation is other things being equal accelerated in comparison with the previous case as the speed of diffusion depends and on number of the subjects which have already accepted an innovation.

The special importance is gained by the questions connected with formation of optimum structure of a communication network, the principles and methods of the organization of the general information space. In this sense the optimality is understood as economic efficiency of functioning of a network.

Probably, the specified criterion of an optimality is answered most fully by a combination of two types of communication structures which is shown, in particular, in industrial clusters. Speed of diffusion of innovations in clusters such factors, as territorial proximity of a source of information, advantages of an innovation, its coherence with the general economic, organizational, social and cultural norms influence. In total these factors increase innovative conductivity of the environment.

The physical phenomena of transfer contain a stochastic component. From positions of the molecular and kinetic theory thermal chaotic movement of elementary particles of substance (molecules) is their main reason. Intensity of use of existing channels of communications can become analog of this component in the analysis of innovative networks. Not always even in the presence of communication information on an innovation is really transferred.

Similar interpretation more corresponds to the device not to probability theory, and the theory of opportunities. Opportunities, unlike probabilities, don't submit to standard normalizing equality – their sum shouldn't be equal to unit.

The numerous researches which traditions go back even to works of J. Schumpeter, it is proved that among the factors favoring or counteracting distribution and widespread introduction of innovations, value judgment of their potential effects and definiteness or reliability of a similar assessment have essential value.

The saved up experience of studying of innovativeprocesses demonstrates that in any set of subjects of acceptance of an innovation there is their whole range from innovators to conservatives depending on that, how fast they perceive an innovation, in what measure are inclined to risk at decision-making in the conditions of incomplete information.

Susceptibilities of subjects of managing to innovations allows an assessment through innovative potential.

4 Conclusion

The analysis of innovative processes on the basis of their model representation not only reflects the general tendencies of modern knowledge, but also is essentially necessary for development of reasonable innovative solutions.

Advantage and force of mathematical methods consists in possibility of the conclusions confirmed with calculations about a course and characteristics of these processes, designing of mechanisms of management by innovative activity.

In conclusion, we will notice the following: the basic conclusion about applicability of some formal models to research of a certain circle of questions yet doesn't give the grounds to claim that this or that relating to this circle of questions a specific objective actually can be solved on the basis of this scheme. For carrying out the analysis and finding of rational decisions it isn't enough to state correctness of the description of an available problem and adequacy of the chosen model.

The model needs to be set quite accurately and unambiguously, having identified the parameters characterizing it components, on quantitative and besides rather exact level. It is however connected with difficulties of quantitative measurement of characteristics of the corresponding technical, economic or technical and economic events which overcoming represents an independent task.

References:

[1] 1. Cowan R., Jonard P. Network Structure and the Diffusion of Knowledge // *Journal of Economic Dynamics and Control*,Vol. 8, No 28, 2004, pp. 1557-1575.

[2] 2. Kozyrev A.N. *Economics of Intellectual Capital.Discussion Paper*, Institute of Management, St. Petersburg State University, 2006.

[3] 3. *Managing Nano-Bio-Info-Cogno Innovations.Convergence Technologies in Science*. William Sims Bainbridge and Mihail C. Roco (Eds.), Springer, 2005.

[4] 4. Oslo Manual. 3rd ed., OECG and Eurostat, 2005

[5] 5. Rogers E.M. *Diffusion of Innovations* (4th ed.), The Free Press, 1983.
[6] 6. Kuzyk B. N., YakovetchYu.V. *Russia – 2050.Strategy of innovative break*.Publishing House «Economics», 2005.
[7] 7. Makarov V. L. Review of mathematical models of economy and innovations//*Economics and Mathematical Methods*, Vol. 45, No. 1, 2009, pp. 3-14.
[8] 8. NizhegorodcevR. M.*Information economy*.*Book 1*.*Information bases of economic growth*. *Moscow – Kostroma*, 2002.
[9] 9. Frolov I.E. Chaplygina I.G. Modernproblemsofcreationofmodelsinthescientificandtechnicalsphereofeconomy//*Economicscienceofmodern Russia*, No. 1, 2009, pp. 1-7

TROPICAL CRYPTOGRAPHY AND ANALYSIS OF IMPLEMENTATION OF NEW MATRIX ONE-WAY FUNCTION

Richard P. Megrelishvili

***Abstract* —** In this article we first announced about two versions of the new matrix one-way function (With respect to the issue of relevance, we repeat, that the main advantage of the matrix one-way function is high speed operation). The first variant is the result of the natural development of cryptography and is associated with the use in the cryptography of new tropical arithmetic operations. The results their applications may be named as "Tropical Cryptography." But at the same time, regardless of the general algebraic values "Tropical Cryptography", it is fact, that the construction of multiplicative groups, based on the our tropical operations, may be accepted as an integral part of the realization of the matrix one-way function. Therefore, its adoption and an implementation can be associated with its recognition.

The second option, at this stage, is the result of repeated analysis of matrix one-way function and is associated with the use of exponential one-way function within a certain time frame, as well as ElGamal used the exponential function for its Digital Signature (Assuming the exponential one-way function, which Diffie and Hellman took from Number Theory). However it is obvious that the use of the degree (exponential) one-way function, in a certain time interval is not associated with loss of speed for the matrix one-way function, therefore, and also - for the corresponding key exchange algorithm via an open channel communication or to perform other actions.

Keywords — Cryptography, matrix one-way function, key exchange algorithm, Tropical Operations

I. INTRODUCTION

The analysis showed that the matrix function is broken, if it is used without a joint application with Tropical cryptography or without the use of one-way function (ie, the function is not a carrier of properties one-way function if it is applied without any special versions of, see below).

Matrix function is as follows:

$$v A' = u. \quad (1)$$

Where $A' \in \check{A}$, a $\check{A}$ is a set of high power from an n-dimensional quadratic commutative matrices [1]. Along with this, $v, u \in V_n$. Where V_n vector space of dimension n (For simplicity $\check{A}$ and V_n is considered over the Galois field GF(2)). In expression (1) v and u are open (without

any special versions) and A' is secret, although A - initial matrix is open with which may be formed a plurality $\check{A}$ (e.g., a plurality $\check{A}$ can be produced with degrees of matrix A). Therefore, if the expression (1) is considered as a one-way function, then it can break down in the following ways:

If the matrix set $\check{A}$ contains recursion (that was identified by us), then the expression (1) can easily be broken with the help Companion matrices, that is, the set of n^2 unknown can be lead to a matrix with n unknowns, for any square matrix $A' \in \check{A}$ can be bring to n unknown, i.e., using the equation (1) can obtain a system of n equations in n unknowns, etc. These issues have been discussed in [2-5, 6].

II. ON THE POSSIBILITY OF BREAKING THE MATRIX ONE-WAY FUNCTION

We want to show that though (1) the matrix function is broken without additional versions, but this is exceptional function. It is special function because of its speed and therefore deserves special attention. We are convinced that the additional versions will not reduce the speed and efficiency of the entire system. It is interesting, how it is can be possible with additional means maintained the speed, the efficiency and the strength of the system? In addition, for this article we consider the ability to break of matrix one-way function, and then we will discuss the possibilities of using tropical cryptography and exponential one-way function. We'll look at how break the matrix one-way function with the use, of said, of basis matrixes (other questions, how to hack the function (1), were considered in [2-5,6]). We will consider breaking this function in the particular example.

Suppose, it is given the multiplicative group $\check{A}$ of the commutative matrices of dimension 3x3 (the group has a maximal order, $e = 2^3 - 1 = 7$):

$$A=\begin{pmatrix}1&0&1\\1&0&0\\0&1&1\end{pmatrix},\quad A^2=\begin{pmatrix}1&1&0\\1&0&1\\1&1&1\end{pmatrix},\quad A^3=\begin{pmatrix}0&0&1\\1&1&0\\0&1&0\end{pmatrix},\ldots,A^7=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}.$$

(2)

Suppose, the two subjects X (Alice) and Y (Bob) can form the secure key k with matrix one-way algorithm via public channel (This algorithm is based on a matrix one-way function (1)). Then Alice selects matrix $A_1 = A^2$ as the secret matrix in (2). Bob, for his part, chooses the matrix $A_2 = A^3$, we also assume that v = (110). Then our algorithm will be functioning as follows:

– Alice computes and sends to Bob the following vector:

$$u_1 = v A_1 = (011). \quad (3)$$

– Bob computes and sends to Alice the following vector:

$$u_2 = vA_2 = (111). \quad (4)$$

– Elice computes the exchanged key:

$$k_1 = u_2A_1 = (100). \quad (5)$$

– Bob computes the exchanged key:

If the matrix of set of Ă does not contain recursion (or hard to find), then the matrix one-way function can be broken with the use of the basic matrixes of A^0, A^1, A^2,..., A^{n-1} which is not hard to get, if we know the initial matrix A.

III. Two Embodiment of the One-Way Function Matrix

As stated above, this paper first announced two special versions of the matrix one-way function. First option, as a result of the natural development of cryptography, involves the use of new tropical arithmetic operations in cryptography. When applying was found that the new tropical operations apart from a general purpose can be thought integral part of our matrix one-way function. Therefore, if earlier, for the construction of matrices Ă had to use classical arithmetic operations, it is now necessary to apply our new tropical arithmetic. With new tropical operations, we must build a set of matrices Ă with the properties with the same as before: high dimension and order, i.e. we should construct a multiplicative group Ă that is formed by degrees of an initial matrix A of new form (of a new structure). Construction of a new matrix of Ă, as noted above, is already a meaningful (traditional) problem and we would not have shown any effect if there was not having contact with her. Consider the issues of the first option, that we have introduced, or questions about Tropical Cryptography.

The obtained tropical operations, for simplicity, considered over the Galois field GF (2). Additive operations, in this case, are the same as the classical operations:

$$0 + 0 = 0;\ 0 + 1 = 1;\ 1 + 0 = 1;\ 1 + 1 = 0. \quad (12)$$

But the multiplicative operations are fundamentally different from the classical operations [7]:

$$0 * 0 = 0;\ 0 * 1 = 1;\ 1 * 0 = 1;\ 1 * 1 = 1. \quad (13)$$

Interestingly, what feature and utility of our proposed tropical operations? Must be stated that the new operations cause so impressive effect in their application that raises another question? It is about ensuring the stability of the matrix function (1), i.e. on the solubility or insolubility of the system of equations (11), depending on what kind of arithmetic operations will be applied - the classic or offered by us? For example, in our opinion, the system of equations (11) does not have a unique solution. Matrix function (1), with tropical operations, is one-way function, it will not be broken in real time, and satisfies the conditions of stability (under appropriate conditions, implying the proper dimension and higher order for a set of matrices Ă). Indeed, when using the new operations (12) and (13), a system (14) has not a unique solution (to the counterweight (11)), since by multiplication coefficients of c_0, c_1, c_2 on the w_0, w_1, w_2 will not cause the formation of null values but on the contrary, causes the formation of new unknowns (While, in the classical operations and using the Gauss method, the system (11) is rapidly soluble):

$$k_2 = u_1A_2 = (100). \quad (6)$$

As we see $k = k_1 = k_2$ and the results are correct (The matrixes are commutative: $vA_1A_2 = vA_2A_1$).

As noted above, we plan to break the algorithm by means of the basis matrix comprising a multiplicative set $Ă = \{ c_0 A^{2^0}. c_1 A^{2^1} \dots c_{n-1} A^{2^{n-1}}\}$ (where $\{c_0, c_1, \dots, c_{n-1}\} \in$ GF(2)). For a set of (2) we form an appropriate basis:

$$A^0 = I,\ A^1, A^2, \quad (7)$$

Where $A^0 = I$ is the identity matrix. In the beginning we define the matrix $A_1 = A^2$ selected by Alice. The required matrix is denoted by $A_1(x)$, then we will have:

$$A_1(x) = c_0 A^0 + c_1A^1 + c_2 A^2. \quad (8)$$

Since Ellis opened calculates the value of $u_1 = v A_1(x)$, then we have:

$$u_1 = v A_1(x) = c_0 vA^0 + c_1vA^1 + c_2vA^2 = c_0w_0 + c1w_1 + c_2w_2. \quad (9)$$

Considering (2), (3) and (9) we can determine the values of u_1 and w_0, w_1, w_2:

$$\begin{aligned} vA^0 &= (110)\ A^0 = (110) = w_0, \\ vA^1 &= (110)A^1 = (001) = w_1, \\ vA^2 &= (110)\ A^2 = (011) = w_2, \\ u_1 &= (011). \end{aligned} \quad (10)$$

Using (9) and (10) we may form a system of equations for the coefficients c_0, c_1, c_2:

$$\begin{aligned} 1c_0 + 0c_1 + 0c_2 &= 0, \\ 1c_0 + 0c_1 + 1c_2 &= 1, \\ 0c_0 + 1c_1 + 1c_2 &= 1. \end{aligned} \quad (11)$$

Solving the system of equations (11), we define the values of the coefficients: $c_0 = 0$, $c_1 = 0$, $c_2 = 1$. Then, from (8) we obtain the value of the ratio of the desired matrix: $A_1(x) = A^2$, i.e. get the matrix A^2 of (2). The answer is correct. (Similar we can find the matrix A_2, chosen by Bob).

IV. References

[1] R.Megrelishvili, M.Chelidsze, K.Chelidze, "On the construction of secret and public key cryptosystems," Iv.Javakhishvili Tbilisi State University, I.Vekua Institute of Applied Mathematics, Informatics and Mechanics (AMIM), v. 11, No 2, 2006, pp. 29-36.

[2] R.Megrelishvili, A.Sikharulidze, "New matrix sets generation and the cryptosystems," Proceedings of the European Computing Conference and 3rd International Conference on Computational Intelligence, Tbilisi, Georgia, June, 26-28, 2009, pp. 253-255.

[3] R.Megrelishvili, M.Chelidze, G.Besiashvili, "Investigation of new matrix-key function for the public cryptosystems". Proceedings of The Third International Conference, Problems of Cybernetics and Information, v.1, September, 6-8, Baku, Azerbaijan, 2010, pp. 75-78.

[4] R.Megrelisvili, M.Chelidze, G.Besiashvili, "One-way matrix function - analogy of Diffie-Hellman protocol", Proceedings of the Seventh International Conference, IES-2010, 28 September-3 October, Vinnytsia, Ukraine, 2010, pp. 341-344.

[5] R.Megrelishili, M.Jinjikhadze, Matrix one-way function for the exchange of cryptographic keys and method for the generation of multiplicative matrix groups ", in Proceedings of The International Conference SAIT 2011, May 23-28, Kyiv, Ukraine, in 2011. p. 472.

[6] W.P.Wardlaw, Matrix Reprezentacion of Finite Fields, U.S. Navy, March 12, 1992, pp. 1-10, NRL/MR/5350.1-92-6953.

$$1^* c_0 + 0 * c_1 + 0 * c_2 = 0,$$
$$1^* c_0 + 0^* c_1 + 1^* c_2 = 1, \quad (14)$$
$$0^* c_0 + 1^* c_1 + 1^* c_2 = 1.$$

For example, the first line of system (14) has the six unknowns, therefore, when dimension has high order (and there are used our tropical operations), the system (14) does not has a solution in real time. Therefore, our matrix one-way function according to the first embodiment ensures durability, since it is not can to break in real time (Take into account the fact that tropical group (15) is a multiplicative group and not a field). As an example we present the multiplicative group (15). For the key exchange algorithm are used: A is an Initial Matrix of (15) and the corresponding $u = vA^3$, where $v = (110)$:

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad A^2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$

$$A^3 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \ \ldots, \ A^7 = A^0 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

(15)

The implementation of the algorithm according to (15) does not differ from the implementation of the algorithm (3) - (6), since the main issue here - the generation of the multiplicative group of maximal order, which meets the requirements of Tropical Cryptography (12) - (13).

Interestingly than can one explain that - the second embodiment has, too, a high efficiency and durability as the first, whereas radically different from the first? In a second embodiment, with respect to the matrix of our one-way function is used a different one-way function (i.e. there is a new problem), but as a method of processing, it shows identity with the decision of other cryptography tasks, which, in our opinion, deserves attention (see. below). For example, ElGamal uses an exponential one-way function to solve their problems, but the thing is - how? He uses a one-way function periodically, for a certain length of time [8]. The similarity with our second option is a period of time for which use the function [9]. In the algorithm of ElGamal degree (exponential) one-way function is used within a certain time period, to meet the challenges of authentication and verification. We use it also within a certain time period, to resolve the problem of the stability of our matrix one-way function. For this, by using exponential one-way function occurs a key exchange via the open channel. The result of this key exchange is a secret.

parameter k = v. In this same time period occurs the key exchange, or other operations carried out, with our algorithm. In this case, in (1) parameters v, A' are secret and only parameter u is open. This change defines the stability of one-way function (1) and also of algorithm (3) - (6), and it does not cause decrease the rate of operations.

[7] R.P.Megrelishvili, New Direction in Construction of Matrix One-Way Function and Tropical Ctyptography, Archil Eliashvili Institute of Control Systems of The Georgian Technical University, Proceedings, N 16, 2012, pp.244-248.

[8] T.ElGamal. "A Public-Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms," IEEE Transaction on Information Theory, v. IT-31, n. 4, 1985, pp. 469-472.

[9] W.Diffie and M.E.Hellman. New Direction in Cryptography, IEEE Transaction on Information Theory, IT-22, n. 6, Nov. 1976, pp. 644-654.

Analysis of non-stationary transport of electrical charge in polymer and composite materials

Borisova M.E.

St.-Peterburgs State Polytechnical University

Abstract -The measurements of ramp voltage and reverse voltage I-V characteristics were used to study of charge carrier mobility and polarization phenomena in polymer films and in blends of polymers.

From reverse and ramp voltage I-V characteristics we calculated the carrier mobility which was found to follow Arrhenius behavior. The experimental data were analysed on the basis of different physical models. The nature of nonlinear effects is discussed using of these models.

Keywords -polymer films, nonlinear effects, polarization, charge mobility, temperature dependence.

I. INTRODUCTION

The nonlinear polarization phenomena in thin polymer films (PETF, PVDF, ABC, PC) and in different blends (TPU/SAN) are discovered by measurements of the ramp voltage and reverse voltage I-V characteristics after polarization of samples under the action of voltage – U_p during time t_p at high temperature T_p. The appearance of maxima on I(U) and I(t) curves are related to presencein film of charge carriers for wich blocking contact between dielectric and electrodes occurs [1-6]. In this case the charge carriers move from one electrode to another in every cycle of measuring U(t) without discharging on electrodes.

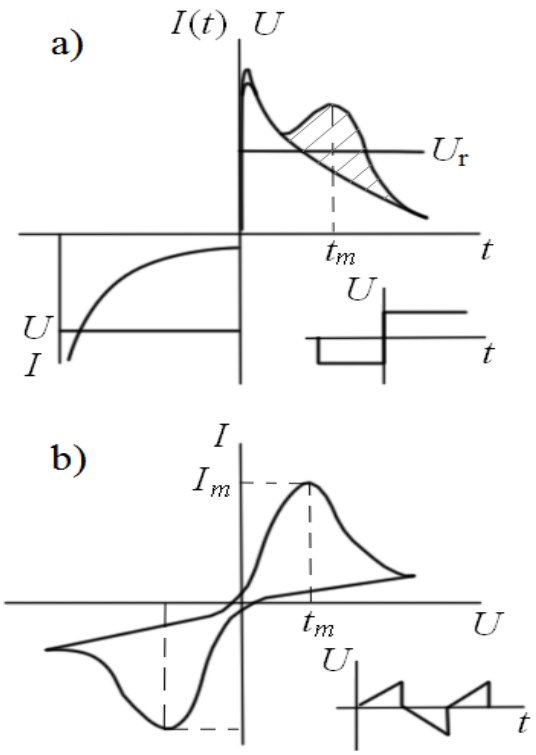


Fig.1 Reverse (a) and ramp (b) curves

II.EXPERIMENTAL

The ramp voltage I-V characteristics were measured by linearly rising of voltage $U = \beta t$ (Fig.1b). In this case upon changing voltage U(t) with the rate $dU/dt = \beta$ maximum of current is observed. The measuring of ramp voltage I-V may be reproduced at periodical change of voltage.

For measurement of reverse voltage I-V characteristics after polarization of the sample the voltage was abruptly changed from $– U_p$ to U_r(Fig.1a). Theresults of measuring of reverse voltage I-V characteristics for PETF films at high temperature 170-200^{0}C are shown on Fig.2.

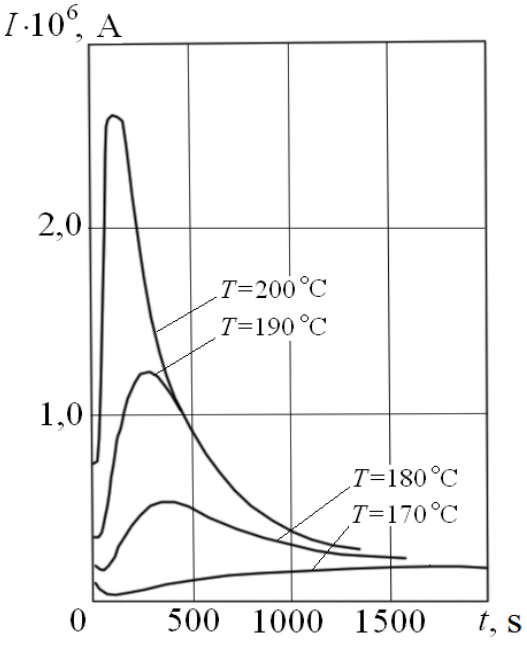


Fig.2 Reverse voltage I-V characteristics for PETF films(h=12 mkm)

It is demonstrated by Fig.2 that with rising of temperature the maximum of current I_m increases and the time of maximum t_m decreases. It is established that the time $t_m$grows with increasing thepolarization Fig.1 Reverse (a) and ramp (b) curves

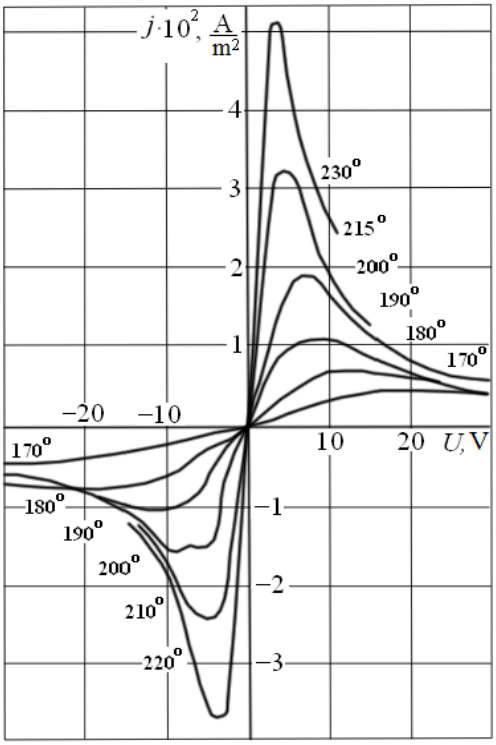
ISBN: 978-1-61804-251-4

voltage.The value of initial reverse current I(0) at moment of the voltage inversion decreases with increasing the polarization voltage U_p and with increasing the time polarization of the sample.

With raising of $-U_p$ and t_p the net charge Q increases. The value of Q is determined as anintegral under the curve I(t). However upon further increase in U_p and t_p the charge Q, t_m, $I_{rev}(t)$ and I(0) are stabilized.

The analogouschanges are observed for ramp voltage I-V curves uponvariation of $-U_p$, t_p, T and β. It is established that at constant temperature T the maximum on curves I(U) shifts tohigh value U_m or to lower t_m =

Fig.3 The ramp voltage I-V characteristics by different temperature (β = 18 mV/s).



U_m/β upon an increasein the rate of the voltage variation β = dU/dt. The value I_m increases with raising the voltage increase rate β [7]. The current at maximum $I_m$increases and position of the maximum shifts to region of lower values of U with increasing average electrostatic field E = -U/h.

As in the case of reverse current I_{rev} the value ramp current (I_m) and the position of maximum (U_m, $t_m = U_m/\beta$) depend on polarization regime, i.e. on the values $-U_p$, t_p.

With raising temperature T at constant value of β the maximum of the ramp voltage I-V characteristics shifts to lower value U_m, i.e. to lower $t_m = U_m/\beta$. The value of current in maximum increases. The typical results of these measurements are

shown on Fig.3.

Dielectrics measurements by means of thermally stimulated depolarization currents (TSDC) and reverse I-V characteristics were used to study the carrier mobility and polarization phenomena in blends of thermoplastic polyurethane (TPU) and styrene-acrylonitrile copolymer (SAN) with different SAN contents.

The reverse voltage I-V characteristics were measured for polymer blends at high temperatures of 120-160 ^{0}C. The choice of U_p and U_r was determined by the sample conductivity. The curves of reverse voltage characteristics have maxima which indicate the nonlinear effect of electro transfer. These characteristics for the blend 50/50 are presented on Fig.4.

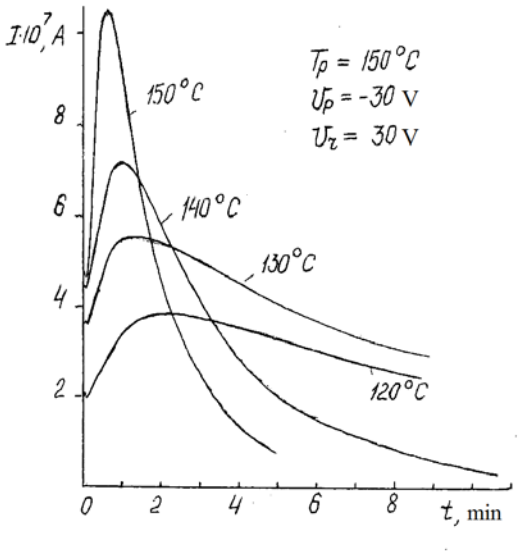


Fig.4 The reverse voltage I-V characteristicsmeasured on 50/50 TPU/SAN blend.

With increasing temperature the peaks shift to shorter times. The analysis of I-V characteristics was made by the assumption that at relatively high temperatures the contact between electrodes and sample is blocking for the carriers which provide the current through the volume of the sample.

By the method of thermally stimulated depolarization currents it was shown that in the high temperature range a heterocharge is accumulated in the sample of blends TPU/SAN. On the TSDC thermograms we observe one peak which is related to the heterocharge relaxation (Fig.5). We may assume that the heterocharge accumulated near the blocking electrode. Fig.5 shows that the maximum of TSDC curves shifts to lower temperature within increasing TPU content in the blend. This is in agreement with an increase in the film conductivity [8].

From the reverse voltage I-V characteristics we calculated the carrier mobility μ by the approximate formulae

$$\mu = \frac{h^2}{U_r t_m} \quad (1)$$

ISBN: 978-1-61804-251-4

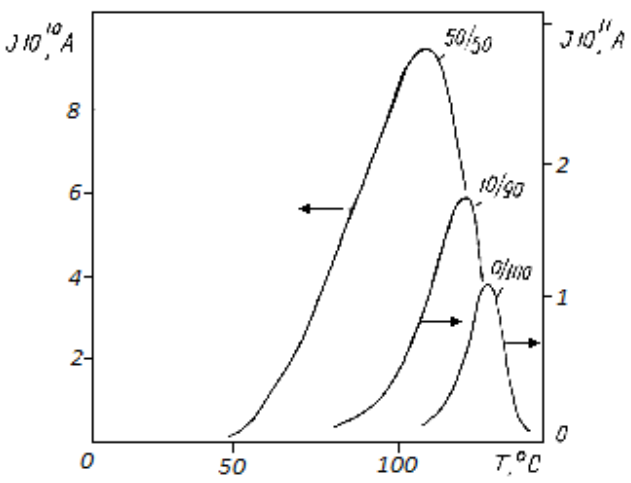


Fig.5 High temperature range TSDC thermograms measured on 50/50, 10/90 and 0/100 TPU/SAN blends.

$$\mu = j_m \frac{h^2}{U_m Q}, \quad (2)$$

where h is the film thickness, U_r– is the reverse voltage, t_m – is the peak time, Q- is the total charge in external circuit and j_m- is the peak current. The calculated from these equations values of μ were found to be close to each other. The μ value calculated from eq.1for50/50 TPU/SAN varies from $1.2\text{x}10^{-12}$ $m^2/V{\cdot}s$ at T= 120 ^{0}C to $4.16\text{x}\ 10^{-12}$ $m^2/V{\cdot}s$ at T= 150^0C.

The carrier mobility follows Arrhenius behavior withtemperature.From the Arrhenius plot we obtain the activationenergy of the carrier mobility equal to 0.65 eV.This value is very close to the conductivity activation energy, which was found to be E_a = 0.70 eV. Taking into account the small value of μ we may assume that ionic conductivity in TPU/SAN blend is predominant at high temperatures..

The values of μ calculated for PET films varied from 10^{-14} to 10^{-15} $m^2/V{\cdot}s$. At temperature 100 ^{0}C μ = $1.5\text{x}10^{-16}$ $m^2/V{\cdot}s$.

Such values of mobility are typical for ionic charge carriers or for electron carriers moving with strong retraping.

III. THEORETICALANALYSIS

The theoretical analysis of the experimental results was usually based on the assumption that contacts between electrodes and dielectric are blocking for the carriers which provide the current through the bulk of polymer. Analysis of ramp and reverse curves on the base of near-electrode polarization model did not take into account the carrier diffusion. It was assumed that sample of the dielectric polymer material comprises the charge carriers of one sign and fixed charge of another sign distributed uniformly in a bulk of polymer $\rho = -\rho_f = \rho_0$. We propose the model in which the carriers concentration in the initial state is equal n and the net charge of carriers per unite area is Q = n e h. This charge in the initial state of the dielectric is compensated by the fixed charge Q_f which is caused by the charge of shallow traps. These traps are immobile in the electrical field and Q_f = - Q. We assume that in the polarized dielectric sample the part of volume is free of carriers and this part possesses only the charge of immobile traps.

To determine of charge carrier mobility were analyzed the ramp and the reverse curves. The problem may be reduced to the solution of the equations of boundary moving for charge cloud. From comparison of experimental (Fig. 2, Fig.3 and Fig.4) and theoretical [9] curves we obtained the following equations for calculating the carriers mobility μ

$$\mu = j_m \frac{h^2}{U_r Q} \quad (\text{for } U = U_r);$$

$$\mu = j_m \frac{h^2}{U_m Q} \quad (\text{for } U = \beta t).$$

The value of Q may be evaluated from experimental data.

The correct analysis of the ramp and reverse curves on the basis of considered above model may be fulfilled by solution of the system differential equation by using of numerical method:

$$j = \mu\rho E - D\frac{\partial \rho}{\partial x} + \varepsilon_0 \varepsilon \frac{\partial E}{\partial t};$$

$$\frac{\partial E}{\partial x} = \frac{\rho + \rho_\phi}{\varepsilon_0 \varepsilon};$$

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x}\left(\mu\rho E - D\frac{\partial \rho}{\partial x}\right)$$

with boundary condition

$$\int_0^h E\,dx = -U;$$

$$j = \varepsilon_0 \varepsilon \frac{dE_0}{dt} = \varepsilon_0 \varepsilon \frac{dE_h}{dt},$$

where E_0 and E_h – electrical fields nearly electrodes x = 0 and x = h, $U = -U_r$ or $U = \beta{\cdot}t$.

For stationary state of near-electrode polarization the system of equation may be solved by analytical method. The results of calculation of potential and electrical field distributions in the stationary state are presented in Fig.6.

It is shown that at small value of U_p the charge is localized in a narrow near-electrode region of the dielectric forming the double electrical layer with charge on the electrode. The results of calculation show that with increasing voltage U_p the thickness of the near-electrode layer is sharply reduced.

IV. CONCLUSION

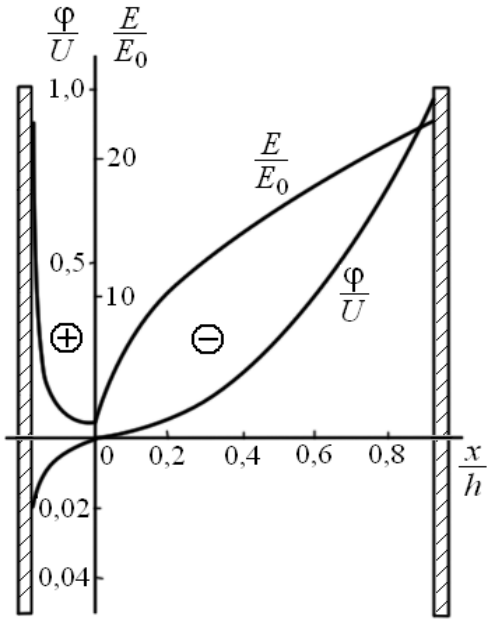
For some polymer films and polymer blends at high temperature on reverse and ramp I-V characteristics the

ISBN: 978-1-61804-251-4

current maxima were observed. These maxima are caused nonlinear processes of electrotransfer.
It is established that at high temperature where maxima $I_{rev}(t)$ and $I_{ramp}(U)$ have place the hetero-charge

Fig.6 The distribution of electricalfield E and potential φ for the dielectric model with one type of carriers and$\rho = - \rho_f$.

accumulated in polymer films.These facts indicate on



processes near-electrode polarization in film which have place at blocking contact between polymer film and electrodes.

The ramp and reverse characteristics were analyzed on the base model of dielectric which content some charge of one sign and fix charge of another sign. It was received the formulae for evaluation of carrier mobility. The comparatively low values of mobility point out ionic type of charge carrier. However one cannot strike off and electron nature of carriers.
The distributions of electrical field E and potential φ were calculated on the base proposed model of dielectric.

REFERENCES

[1] M. Onoda, H. Nakayama, K. Amakawa, «Transient current of plasticized polyvinylchloride,» Jap.J. Appl. Phys.,vol.19, no.2, pp.381-382,1980.
[2] K. Iida, H. Ishiguro, S. Nakamura, G. Sawa, M. Ieda,«Current peaks following voltage reversal in chlorinated polyethylene», J. Appl. Phys., vol.24,
no.6, pp.666-668, 1985.
[3] S. Nakamura, G. Sawa, M. Ieda, «Electrical conduction of nylon 6 at the high temperature», Jap. J. Appl. Phys., vol.20, no.1, pp. 47-53, 1981.
[4]K. Miyairi, M. Ikeda, «Current peaks observed in polyethylene terephthalate films with linearly increasing voltage,» J.Appl.Phys.,vol.19, no.6,pp.1067-
1071, 1980.
[5] K. Miyairi, M. Ikeda, «Investigation of current-voltage characteristics in polymers by new method,» ,3rd Int.conf. Dielec. mater.,meas. and appl.,
Birmingham, pp.314-317, 1979, London-New York.
[6] M.E.Borisova, O.V. Galjukov, S.N. Kojkov, «Nonlinear effects of electrotransfer in polymer films,» Electrotechnica, no.7, pp.69-71, 1991.
[7]M.E.Borisova, O.V. Galjukov, S.N. Kojkov, «The migratory polarization in polymer and current maxima of dynamic current-voltage characteristics,»
Izvestiya of high educational school. Physics, no.8, pp. 29-34, 1988.
[8] A.Kanapitsas, P. Pissis, A. Garcia Estrella. Eur. Polym. J. vol.35, p.923,1999.
[9] M.E.Borisova, O.V. Galjukov, «The investigation of the transfer parameters in polymer films by method of dynamic current-voltage characteristics,»
Izvestiya of high educational school. Physics, no.10, pp.101-106, 1987.

Mathematical simulation of thermal contact of the thermocouple for research of an error of measurements

Olga S. Yashutina, Yuliana K. Atroshenko, and Pavel A. Strizhak

***Abstract*—**The two-dimensional model of heat transfer is developed for research of regularities of warming up of a sensitive element of typical contact thermoelectric transformers (thermocouples). It is shown that in the presence of air gap between the thermocouple and a surface of object of measurements duration of heating up of the thermocouple significantly increases. Results of numerical research of duration of heating up of thermocouples are given in case of different values of value of air gap. Theoretical dependences of the relative error of temperature measurement by means of the thermocouple of K type of runtime of measurements are set.

***Keywords*—**error, heating time, mathematical modeling, nonstationary process of heat transfer, thermocouple.

I. Introduction

CONTACT methods of temperature measurement were widely adopted in control and management systems in all fields of activity of the person: in the industry, medicine, technique, etc. Among the means implementing such methods thermocouples and resistance temperature detectors are the most applied. Thus the error of temperature measurements has essential impact on quality of system operation of control as a whole [1, 2]. Despite all advantages of [3] of these gages reliability of received results in case of surface measurements in many respects depends on conditions of thermal contact with object of measurement [4–6].

The type selection of the thermocouple depends on conditions of execution of measurements, their purposes (requirements imposed to accuracy of measurements, duration of measurements, etc.). The type selection of the thermocouple also defined by the range of taken temperatures. Most often in the range of taken temperatures from -200 °C to 1100 °C *K*, *E* and *L* thermocouples types are used. For execution of high-precision measurements thermocouples of types *R* or *S* are most often used.

This work is performed with financial support from the Russian Foundation for Basic Research (project No. 14–08–00057).

Olga S. Yashutina is with the National Research Tomsk Polytechnic University, Tomsk, Russia (corresponding author to provide phone: 8-913-879-60-62; e-mail: julie55@tpu.ru).

Yuliana K. Atroshenko is with the National Research Tomsk Polytechnic University, Tomsk, Russia (e-mail: yusina_kr@rambler.ru).

Pavel A. Strizhak is with the National Research Tomsk Polytechnic University, Tomsk, Russia (e-mail: pavelspa@tpu.ru).

In this operation the thermocouple model is provided, also results of numerical research of influence of conditions of thermal contact on an error of thermocouples are given.

II. Physical Model of Heat Transfer

The area of the solution of the task represents the non-uniform system "a thermocouple seal – powder – a protective cover air" which geometrical representation is given in fig. 1.

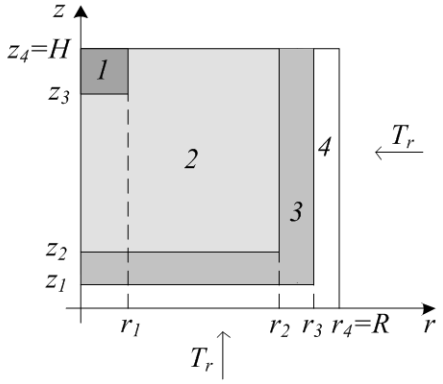


Fig. 1 Area of the solution of the task of heat transfer: *1* – thermocouple junction, *2* – the powder Al_2O_3, *3* – a protective cover, *4* – air gap

In case of the solution of the task of heat conduction the following assumption is accepted: heatphysical characteristics of elements of system in the field of the solution of the task of heattransfer don't depend on temperature.

The initial temperature of all system corresponds to reference conditions and is equal 20 °C. Heating up is made from the surface separated from the thermocouple by air gap of 4 (fig. 1). Thus the moment of the end of process of heating up is defined by achievement by a thermocouple junction of 1 temperature different from temperature of T_r on value not exceeding value of an admissible error. For researched thermocouples in [7] the following values of admissible errors are set: for the thermocouple of *S* type the admissible error makes ±1,5 °C in the range of temperatures 0…1100 °C; for the thermocouple of *K* type the admissible error makes ±1,5 °C in the range of temperatures -40 … 375 °C and ±0,004·*t* in the range of temperatures 375…1000 °C; for the thermocouple of *L* type the admissible error makes ±2,5 °C in the range of temperatures -40…300 °C and ±0,0075·*t* in the range of temperatures 300…800 °C.

ISBN: 978-1-61804-251-4

Research was conducted for the thermocouple with a diameter of 5 mm; height of a simulated section of the thermocouple is restricted 5 mm high from lower bound; value of air gap changed from 1 to 10 mm.

III. Mathematical Model

The two-dimensional model of heattransfer (fig. 1) is described by the following differential equations:

$$c_1\rho_1\frac{\partial t_1}{\partial t}=\lambda_1\left(\frac{\partial^2 t_1}{\partial r^2}+\frac{1}{r}\frac{\partial t_1}{\partial r}+\frac{\partial^2 t_1}{\partial z^2}\right) \quad (1)$$

$t>0$, $0<r<r_1$, $z_3<z<H$;

$$c_2\rho_2\frac{\partial t_2}{\partial t}=\lambda_2\left(\frac{\partial^2 t_2}{\partial r^2}+\frac{1}{r}\frac{\partial t_2}{\partial r}+\frac{\partial^2 t_2}{\partial z^2}\right) \quad (2)$$

$t>0$, $0<r<r_2$, $z_2<z<z_3$; $r_1<r<r_2$, $z_3<z<H$;

$$c_3\rho_3\frac{\partial t_3}{\partial t}=\lambda_3\left(\frac{\partial^2 t_3}{\partial r^2}+\frac{1}{r}\frac{\partial t_3}{\partial r}+\frac{\partial^2 t_3}{\partial z^2}\right) \quad (3)$$

$t>0$, $0<r<r_3$, $z_1<z<z_2$; $r_2<r<r_3$, $z_2<z<H$;

$$c_4\rho_4\frac{\partial t_4}{\partial t}=\lambda_4\left(\frac{\partial^2 t_4}{\partial r^2}+\frac{1}{r}\frac{\partial t_4}{\partial r}+\frac{\partial^2 t_4}{\partial z^2}\right) \quad (4)$$

$t>0$, $0<r<L$, $0<z<z_1$; $r_3<r<r_4$, $z_1<z<H$.

Here r – radial coordinate, m; z – axial coordinate, m; c – specific heat capacity, J/(kg·°C); ρ – density, kg/m^3; λ – coefficient of heat conduction, W/(m·°C); indexes: 1 – thermocouple junction, 2 – powder of an aluminum oxide, 3 – protective cover, 4 – air.

On boundaries "a thermocouple–powder", "powder–a protective cover", "a protective cover–air" the following boundary conditions were accepted:

$$T_1(r_1,z)=T_2(r_1,z);$$

$$-\lambda_1\left.\frac{\partial T_1}{\partial r}\right|_{r=r_1}=-\lambda_2\left.\frac{\partial T_2}{\partial r}\right|_{r=r_1}; \quad (5)$$

$$T_2(r_2,z)=T_3(r_2,z);$$

$$-\lambda_2\left.\frac{\partial T_2}{\partial r}\right|_{r=r_2}=-\lambda_3\left.\frac{\partial T_3}{\partial r}\right|_{r=r_2}; \quad (6)$$

$$T_3(r_3,z)=T_4(r_3,z);$$

$$-\lambda_3\left.\frac{\partial T_3}{\partial r}\right|_{r=r_3}=-\lambda_4\left.\frac{\partial T_4}{\partial r}\right|_{r=r_3}; \quad (7)$$

$$T_1(r,z_3)=T_2(r,z_3);$$

$$-\lambda_1\left.\frac{\partial T_3}{\partial z}\right|_{z=z_3}=-\lambda_2\left.\frac{\partial T_2}{\partial z}\right|_{z=z_3}; \quad (8)$$

$$T_2(r,z_2)=T_3(r,z_2);$$

$$-\lambda_2\left.\frac{\partial T_2}{\partial z}\right|_{z=z_2}=-\lambda_3\left.\frac{\partial T_3}{\partial z}\right|_{z=z_2} \quad (9)$$

$$T_3(r,z_2)=T_4(r,z_2);$$

$$-\lambda_3\left.\frac{\partial T_2}{\partial z}\right|_{z=z_1}=-\lambda_4\left.\frac{\partial T_3}{\partial z}\right|_{z=z_1} \quad (10)$$

Initial conditions define the temperature distribution in the thermocouple's sensitive element in an initial time point:

$$t=0;\ t=t_0,\ 0<r<R, \quad (11)$$

$$t=0;\ t=t_0,\ 0<z<H, \quad (12)$$

where t_0=20 °C – temperature corresponding to reference conditions.

Boundary conditions of heat transfer problem solution domain are defined as follows.

Boundary conditions of the first kind are set on $r=R$ boundary:

$r=R$, $t=t_p$, where t_r – temperature of a heating element.

Boundary conditions on r=0 symmetry axis:

$$r=0,\ \frac{\partial t}{\partial r}=0 \quad (13)$$

Boundary conditions of the first kind are set on z=0 boundary:

$$z=0;\ t=t_r \quad (14)$$

Boundary conditions on $z=H$ boundary:

$$z=H;\ \frac{\partial t}{\partial r}=0 \quad (15)$$

IV. Solution Procedures

The area of the solution of the task (fig. 1) is broken into the uniform grid consisting of 240 nodes. The slot pitch on radial and axial coordinates is equal $2{,}5\cdot10^{-2}$ mm. The step on a temporal grid changed in the range from 10^{-4} to 10^{-2} s for reduction of volume of computation and increase of accuracy of the decision.

System of equations 1–4 with the appropriate initial and boundary conditions decided using a method of finite differences. The solution of the difference analogs of the differential equations representing the linear algebraic equations was carried out by a local and one-dimensional method. The pro-race method was applied to the decision of system of the difference equations on the basis of the implicit four-point diagram [8].

V. Results and Discussion

Mathematical modeling were carried out at parameters [9–11]: λ_1=33,1 W/(m·°C), C_1=768 J/(kg·°C), ρ_1=8825 kg/m^3; thermocouple junction (type *S*): λ_1=50,4 W/(m·°C), C_1=139 J/(kg·°C), ρ_1=20710 kg/m^3; thermocouple junction (type *L*) λ_1=24,75 W/(m·°C), C_1=713 J/(kg·°C), ρ_1=8920 kg/m^3; powder Al_2O_3: λ_2=6,57 W/(m·°C), C_2=850 J/(kg·°C), ρ_2=1250 kg/m^3; protector case steel: λ_3=15 W/(m·°C), C_3=462 J/(kg·°C), ρ_3 =7900 kg/m^3; air: λ_4=0,026 W/(m·°C), C_4=1190 J/(kg·°C), ρ_4=1,161 kg/m^3.

Dependence of time of heating up of the thermocouple on value of air gap for different values of the taken temperature for the thermocouple of K type is given in fig. 2.

ISBN: 978-1-61804-251-4

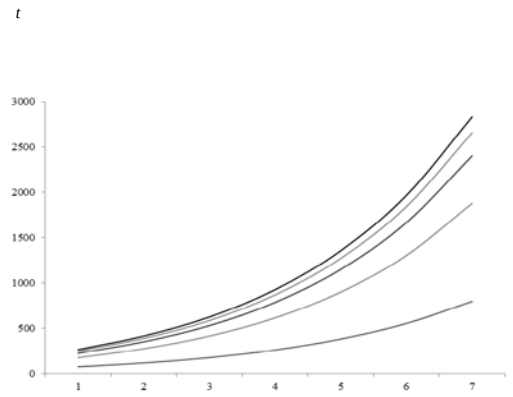


Fig. 2. The *K* type thermocouple sensitive element heating-up duration dependences on value of air gap between a sensitive element and a heating element: 1: *T*=577 °C; 2: *T*=277 °C; 3: *T*=177 °C; 4: *T*=77 °C; 5: *T*=27 °C.

Dependences of minimum necessary time of heating up in the conditions of air gap for *L* and *S* thermocouples are similar given in fig. 2 and have non-linear character that is confirmed by temperature distribution in the field of a sensitive element upon termination of heating up (fig. 3).

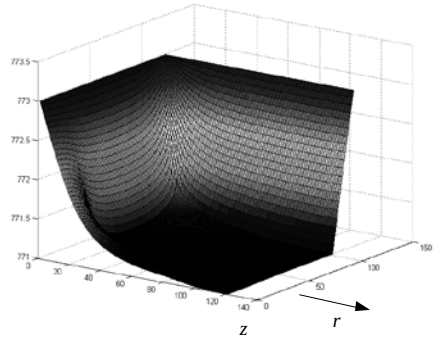


Fig. 3. Temperature field in the thermocouple of *K* type after the termination of heating to 500 °C

The analysis of fig. 2 says that the increase in thickness air between the thermocouple and object of measurement for surface thermocouples significantly increases duration of heating up to the necessary temperatures. Based on it is possible to draw an output that non-compliance with time of heating in the conditions of not full contact of the thermocouple with a surface of object of measurement will lead to big errors of results of measurements that is confirmed by the data provided in tab. 1–2.

TABLE I. VALUES OF THE RELATIVE ERROR OF MEASUREMENTS OF DIFFERENT VALUES OF TEMPERATURE THE THERMOCOUPLE OF *K* TYPE IN CASE OF VALUE OF AIR GAP OF 1 MM, %

t, s	300 °C	400 °C	500 °C
50	50,439	51,806	52,638
100	13,513	13,787	13,953
150	4,413	4,495	4,545
200	1,523	1,551	1,568
250	0,536	0,545	0,551
300	0,190	0,193	0,195

TABLE II. VALUES OF THE RELATIVE ERROR OF TEMPERATURE MEASUREMENTS 500 °C THE THERMOCOUPLE K TYPE IN CASE OF DIFFERENT VALUES OF VALUE OF AIR GAP, %

t, s	1 mm	2 mm	3 mm
100	13,953	32,566	63,924
150	4,545	14,098	32,862
200	1,568	6,628	18,610
250	0,551	3,227	11,052
300	0,195	1,597	6,737
350	0,069	0,798	4,171
400	0,025	0,400	2,606
450	0,008	0,201	1,637
500	0,004	0,101	1,032
550	0,002	0,051	0,652
600	0,001	0,026	0,414
650	0,001	0,012	0,261
700	0,000	0,006	0,166

The analysis of tab. 1-2 testifies that temperature measurement error in the conditions of incomplete contact with object of measurement can be lowered by surface thermocouples in particular at the expense of increase in duration of heating up. Thus determination of minimum necessary time of heating up can be executed with use of the developed model (1)–(15).

VI. CONCLUSION

The developed mathematical model can be used for prediction of time of heating up of the thermocouple in actual practice executions of measurements. Numerical results of researches can be used for an assessment of a possible error of measurements in different conditions of thermal contact.

The received results allow to draw the following outputs:

1) Dependence of time of heating up on value of air gap isn't linear and considerably increases in case of value of air gap more than 3 mm;

2) The measurement error rather strongly depends on conditions of thermal contact and in case of non-compliance with necessary duration of measurement can have essential impact on reliability of results of measurement.

REFERENCES

[1] L. Tsikonis, J. Albrektsson, J.Van herle, and D. Favrat "The effect of bias in gas temperature measurements on the control of a Solid Oxide Fuel Cells system," *Journal of Power Sources*, vol. 245, pp. 19–26, 2014.

[2] N. Zhu, K. Shan, S. Wang, and Y. Sun "An optimal control strategy with enhanced robustness for air-conditioning system considering model and measurement uncertainties," *Energy and Buildings*, vol. 67, pp. 540–550, 2013.

[3] J. Sulciner "Choosing RTDS and thermocouples," *Control Engineering*, vol. 46, no. 2, p.152, 1999.

[4] T.V. Borovkova, V.N. Yeliseyev, and I.I. Lopukhov, "Mathematical Modeling of Contact Thermocouple", *Physics of Particles and Nuclei Letter*, vol. 5, no.3, pp.274–277, 2008.

[5] G.V. Kuznetsov and K.M. Mukhammadeev, "Numerical estimation of errors of temperature measurments by thermocouples using special glues and pastes", *Journal of engineering thermophysics*, vol. 19, no. 1, pp. 17–22, 2010.

[6] G. Beges, M. Rudman, J. Drnovsek "Evaluation of Flat Surface Temperature Probes, "*International Journal of Thermophysics*, no. 32, pp. 396–406, 2011.

[7] IEC 60584-2. *International standard. Thermocouples. Part 2: Tolerances*, 1989.

[8] A. A. Samarskii, *The Theory of Difference Schemes*, Marcel Dekker, Inc., USA, 2001.

[9] .B. Vargaftik, *Reference Book on Thermophysical Properties of Gases and Liquids*, Stars, Moscow, 2006 [in Russian].

[10] P.A. Kinzie, *Thermocouple Temperature Measurement*, USA, Wiley-Interscience Publ., 1973.

R. Hultgren, *Selected Values of the Thermodynamic Properties of Binary Alloys*, USA, American Society for Metals, 1973.

Game-Theoretical Model of Coordination of Interests of State-Private Partnership

VLADIMIR V. GLUKHOV, IGOR V. ILIN
Institute of Industrial Economics and Management
Saint-Petersburg State Polytechnical University
195251, Saint-Petersburg, Politechnicheskaya, 29
RUSSIA
vicerector.me@spbstu.ru ilyin@fem.spbstu.ru http://www.spbstu.ru/

Abstract: In modern conditions the social and economic problems, associated with development of regions, lie at the intersection of State responsibility and entrepreneurial activities of private business. The paper deals with coordination of interests of business and administrations of the constituents of the Russian Federation to address the socio-economic development. A complex of game-theoretic models of involving companies, which have been prequalificated to complete the socio-economic tasks, is developed.
A theorem on the existence of situations of strict Nash equilibrium for the considered classes of games with the possibility of adjusting the institutional environment is established.

Key-Words: Game theory, Nash equilibrium, effective cooperation, state-private partnership, State administration.

1 Introduction

A broad interpretation of the state-private partnership involves the constructive interaction between business and government, not only in the economy but also in politics, culture, science, etc.

Among the basic features of state-private partnerships in a narrow (economic) treatment are the following: parties of state-private partnerships are the state and the private business; interaction between the parties has an official, legal basis; interaction between the parties is equal; state-private partnership has a clearly expressed public and social orientation; in projects of public-private partnerships resources and contributions of the parties are consolidated; financial risks and costs, as well as the results achieved by the parties are divided between them in predetermined proportions. As a rule, a public-private partnership requires that the state does not connect to the projects of business, but on the contrary, the state encourages companies to participate in the realization of socially significant projects [8].

The participants of state-private partnership are government organizations and business representatives. A large number of projects have a regional character. Among stakeholders, implementing the project, act local administration and business representatives. Illustrations of such projects are projects of development of transport and social infrastructure, business development etc. The key issues of successful projects are the formation of a set of participants and coordinating the interests of the parties involved.

2 Problem Formulation

High level of socio-economic development of regions is one of the main indicators of the successful activities of the public authorities [1].

This paper discusses the issues of harmonization of interests while realizing projects of state-private partnership aimed at the socio-economic development. The basis of most of these projects is cooperation of state and entrepreneurs [12]. Participants of these projects are the regional administrative authorities and involved companies. Each participant has their own interests and, accordingly, their behavioral strategies. Reconciliation of interests and interaction based on it is critically important issue for realization of such projects. On the other hand, effective coordination of interests from the project point of view can not be achieved without taking into account the features and the possible development of the institutional environment [11]. Using existing institutional instruments, their correction and the formation of new ones allows selecting the project participants efficiently. The possibility of coordinating the interests presupposes the existence of equilibrium situations in relation to the totality of stakeholders' strategies.

ISBN: 978-1-61804-251-4

One of the tools of organizing state-private partnership are electronic trading platforms [9]. They have several advantages:

• Increase of profitability by reducing costs and better supply conditions

• Increase of selling market, reduce of labor costs for implementation of procurement

• Reduce of dependence on suppliers and contractors, resistance of corruption schemes

• Increased transparency of the process

• Improving the public image of the company

• Guarantees of the contracts concluded and executed.

There are several procedures for the implementation of trading: auction, competition, request for proposals (RFP), request for quotation (RFQ) [2]. The most suitable procedure to achieve high-quality results is competition that consists of two parts: the admission of companies that submitted bids before the competition procedure starts and the election of winner among the companies that have passed the primary selection. The municipal administration countermarks the level of qualification, experience and assets of the potential winner so that he can realize the project. In the second part of the competition the winner is selected and the contract is concluded. It is difficult for one company to realize a large project without partners. Additional project participants (other companies, municipal administration) are needed. The purpose of companies' activities is to make profit at the expense of their own projects, the purpose of municipal administrations is to develop regional socio-economic infrastructure. To guarantee realization of the project, it is necessary to reconcile the interests of all participants. "In the current situation the most serious problem is the procedure of approval of individual strategies among individual participants" [13], if there are a lot of the participants (for example, about twenty), the coordination of their interests becomes rather problematic. It is necessary to form minimally sufficient number of participants for the implementation of socio-economic project in terms of management problems. This requires a pre-selection process based on a set of criteria to define a set of companies, among which the winner will be selected, responsible for implementing the project and the companies that may be involved in the project, as they have the necessary assets to achieve the project objectives. Pre-selection does not solve all problems of interaction between participants. Therefore, in addition to pre-selection it is necessary to create such conditions between prequalified companies, under which these companies can act in cooperation with municipal administration to address both social objectives and business goals. In the opinion of the authors, game theory methods can be used to solve this kind of tasks. In particular non-antagonistic games and a strict Nash equilibrium concept can help to develop decision-making mechanism for the implementation of socio-economic problems in this area.

In such a manner, it seems urgent to solve following problems. First, the formation of competitive selection procedure, which should have as a result a set of companies with total assets allowing them to solve the socio-economic problems. Secondly, modelling the process of interaction between companies using non-antagonistic games language. Third, the formation of such an institutional environment in which the interaction described in the form of game models has a strict Nash equilibrium.

3 Problem Solution

In this paper to solve the formulated task authors draw on the experience of municipal administrations of the Russian Federation on realization of social infrastructure development projects.

Social infrastructure development project is a project contributing to the development of educational, health, social welfare, culture, sports, recreation and other social facilities services.

Such projects have all the main features of state-private partnership. Modelling the interaction of stakeholders allows achieving general results.

Such project can be motivated by the need for development of social infrastructure, for instance the need to build or renovate kindergartens and schools, construct roads etc. Formal initiator of such project is the municipal administration. It can not realize the project on its own and needs to involve construction, engineering and other companies, ie implementing such a large project requires the efforts of several commercial companies and the municipal administration. Models of cooperation of the state and businesses are considered in scientific work of V.G.Varnavski, V.V.Gluhov, M.M. Safonov, T.M. Bogolib and others [3], [4], [6], [10], [14].

As was stated above it is necessary to form a set of participants, the potential of which allows achieving the project objectives. This provides an opportunity to establish different forms of interaction. For the purposes of this paper the specific form of interaction is not so essential.

In this paper, we focus on one of the important methods of interaction - electronic trading platforms.

Electronic trading platforms provide a few types of procedures: auction, competition, RFP, RFQ. In the auction procedure winner is the one who offered the lowest price (the main goal is cash savings), but for social projects the main goal is feasibility, the most suitable for this is the competition procedure.

The number of companies willing to participate in competition is, as a rule, large and exceeds twenty. However, the municipal administration needs a set of participants, which makes it possible to realize the project. From the viewpoint of manageability it is convenient to work with a small number of companies. Practice shows that it is enough to have three or four companies. Selection should be organized in such manner as to choose companies that have the appropriate assets, experience, are financially successful and will be able to implement the project together.

Before starting the competition procedure, it is advisable to hold an initial selection of participants to allow for participation only those who have a sufficient set of assets. As a result of the competition the winner is unveiled, it should be a company that is able to implement the project to address socio-economic problems. To improve the effectiveness of project realization, the municipal administration may also involve companies that have passed the preliminary selection, but have not won the competition to join the winner in realizing the project. Combining the efforts of the winner and companies passed the initial selection, the degree of feasibility of the project becomes very high [7].

Pre-selection should be based on financial, professional and legal criteria. They can be divided into two groups: the prohibitive criteria and permissive criteria. These criteria are not intended to be exhaustive information about the participant, but acting in accordance with them is enough to pass for those companies that are able to realize the project. The first group includes the criteria designed to prevent companies that just will not be able to implement the project from winning the competition, going into mathematical language this are the necessary conditions. The second group are those criteria accordance to which allows company realizing the project (sufficient conditions). Prohibitive criteria are divided into economic and legal. These are the economic components of prohibitive criteria:

1. There are rent arrears to the municipal budget coming from previously signed investment agreements.

2. There are arrears to other economic agents (eg, the pledged assets).

3. The aggregate value of assets (unencumbered by obligations) is less than the cost of the investment project with the possibility of obtaining a loan.

The legal aspects of prohibitive criteria include:

1. Violation the mode of use of urban areas and other real estate (buildings, structures) within previously implemented investment projects.

2. Companies that are in the process of bankruptcy.

3. Demonstrated criminal activities in the economic environment.

Criteria constitute a set of institutional characteristics of the environment in which the pre-selection of competition participants is held.

After a few firms will are pre-selected and the winner is determined, their further interaction should be analyzed. The analysis is based on what assets the company possesses. Depending on the assets of the winner, it is advisable to consider several possible situations. For modeling interactions let us use the language of game theory. Such interaction has properties that are typical for non-antagonistic games with a small number of players [5], [15]. The players are following: construction, engineering and other companies and the municipal administration (the player able to adjust the institutional environment). Each player is pursuing its own goals: the municipal administration aims to construct social infrastructure objects, construction and engineering companies aim to generate revenue through selling commercial and non-commercial premises. To achieve the goal a player is taking a sequence of actions, which forms his strategy. A set of selected players' strategies defines the situation. There appears a game in which it is needed to analyze the existence of a strict Nash equilibrium.

Let us describe players who have the ability to pass pre-selection, these include not possessing lands, but having enough amount of money for the project (we denote such firm as G_1), companies that possess the land on which the construction of social objects such as kindergartens, schools, etc. can be realized (denoted as G_2) and companies that own social facilities buildings (G_3). One of the players is the municipal administration, it is the player with the right of adjustment of the institutional environment. Each player has a finite set of strategies, which we denote by S_{ij}, where i - number of the player, j - number of strategy from the set of strategies of this player: $G_1 \longleftrightarrow S_{11}, \dots, S_{1n1}$; $G_2 \longleftrightarrow S_{21}, \dots, S_{2n2}$; $G_3 \longleftrightarrow S_{31}, \dots, S_{3n3}$; $G_4 \longleftrightarrow S_{41}, \dots, S_{4n4}$. To implement the chosen strategy the

player performs a sequence of steps that can be divided into two types: *T* - Transaction, *B* - Building. Every action has its indices ($T_{1.1}$, $T_{1.2}$, $T_{1.3}$, B_1, B_2, B_3, etc.).

Situations are formally designated as follows:

(S_{1i}, S_{2j}, S_{3k}, S_{4f}) = C_N(i, j, k, f), numbered as a matter of convenience,

Depending on the situation the gain is calculated for each player (H_i - the payoff function of i-th player):

H_1(C(i, j, k, f)), H_2(C(i, j, k, f)), H_3(C(i, j, k, f)), H_4(C(i, j, k, f)).

In order to analyze the situations for the presence of a strict Nash equilibrium, we consider several types of games. The type is defined by the number of players and a set of players' assets.

The first type. Two players: companies that do not have land ownership, but there is money for the lease of land and St. Petersburg, which owns the land. Construction companies (G_1) look to get land for the construction of commercial and non-commercial property for subsequent sale, the purpose of St. Petersburg is the construction of social infrastructure to meet the needs of the city population. We introduce the following notation:

B_1 – realty construction HS_1 on the land property FLD_1;

B_2 – kindergarten construction Y_1 on the land property FLD_1;

$T_{1.1}$ – lease of land FLD_1 by St. Petersburg;

$T_{1.2}$ – transmission to St. Petersburg in the rent for the land FLD_1 the kindergarten Y_1;

$T_{1.3}$ – realty selling HS_1;

$T_{1.4}$ – organizing of the private kindergarten;

$T_{1.5}$ – transmission of kindergarten Y_1 offset debs to St. Petersburg;

$T_{1.6}$ – publication in the media about the willingness to sell kindergarten Y_1 to other companies;

$T_{1.7}$ – lend-lease of the kindergarten Y_1 to other companies;

$T_{4.1}$ – lend-lease of land property FLD_1 to the player G_1;

$T_{4.2}$ – taking the kindergarten Y_1 into ownership of St. Petersburg towards rent for land property FLD_1;

$T_{4.3}$ – taking the kindergarten Y1 into ownership of St. Petersburg towards debt to municipal budget.

Let us consider following situations for players G_1 and G_4:

$S_{1.3}$: ($T_{1.1}$) -> (B_1) -> (B_2) -> ($T_{1.2}$) -> ($T_{1.5}$)

$S_{2.1}$: ($T_{2.1}$) -> ($T_{2.1.1}$)

$S_{1.3}$: ($T_{3.4}$)

Table 1. Situations and Gains

Situation	**Gain**
$C_1(S_{1.1}; S_{4.1})$ $S_{1.1}$: $T_{1.1} \rightarrow B_1 \rightarrow B_2 \rightarrow T_{1.3} \rightarrow T_{1.2}$ $S_{4.1}$: $T_{4.1} \rightarrow T_{4.2}$	$H_1(C_1)$ = – the price of land rental - real property construction costs - kindergarten construction costs + real property sale $H_4(C_1)$ = land rental – content of kindergarten + monthly income from kindergarten
$C_2(S_{1.2};S_{4.2})$ $S_{1.2}$: $T_{1.1} \rightarrow B_1 \rightarrow B_2 \rightarrow T_{1.3} \rightarrow T_{1.4}$ $S_{4.2}$: $T_{4.1}$	$H_1(C_2)$ = – the price of land rental - real property construction costs - kindergarten construction costs + real property sale + monthly income from kindergarten $H_4(C_2)$ = land rental
$C_3(S_{1.3};S_{4.3})$ $S_{1.3}$: $T_{1.1} \rightarrow B_1 \rightarrow B_2 \rightarrow T_{1.3} \rightarrow T_{1.5}$ $S_{4.3}$: $T_{4.1} \rightarrow T_{4.3}$	$H_1(C_3)$ = – the price of land rental - real property construction costs - kindergarten construction costs + real property sale + debt repayment $H_4(C_3)$ = land rental – content of kindergarten + monthly income from kindergarten
$C_4(S_{1.4};S_{4.2})$ $S_{1.4}$: $T_{1.1} \rightarrow B_1 \rightarrow B_2 \rightarrow T_{1.3} \rightarrow T_{1.6}$ $S_{4.2}$: $T_{4.1}$	$H_1(C_4)$ = – the price of land rental - real property construction costs - kindergarten construction costs + real property sale + kindergarten building costs $H_4(C_4)$ = land rental – «time spent on looking for company to manage a kindergarten»
$C_5(S_{1.5};S_{4.2})$ $S_{1.5}$: $T_{1.1} \rightarrow B_1 \rightarrow B_2 \rightarrow T_{1.3} \rightarrow T_{1.7}$ $S_{4.2}$: $T_{4.1}$	$H_1(C_5)$ = – the price of land rental - real property construction costs - kindergarten construction costs + real property sale + monthly income from kindergarten $H_4(C_5)$ = land rental

ISBN: 978-1-61804-251-4

In this game gains satisfy following constraints:

$H_1(C_2) + H_4(C_2) > H_1(C_1) + H_4(C_1)$

$H_1(C_3) + H_4(C_3) > H_1(C_1) + H_4(C_1)$

$H_1(C_2) + H_4(C_2) > H_1(C_4) + H_4(C_4)$

$H_1(C_3) + H_4(C_3) > H_1(C_4) + H_4(C_4)$

To compare the payoffs to the players in situations C_2 and C_3 there is a need to compare the debt of player G_1 and content of kindergarten. If debt to the municipal administration is more, then

$H_1(C_3) + H_4(C_3) > H_1(C_2) + H_4(C_2)$;

otherwise $H_1(C_3) + H_4(C_3) < H_1(C_2) + H_4(C_2)$.

Thus within this game there is a situation of strict Nash equilibrium.

Similarly we can prove that there are situations of strict Nash equilibrium in the following types of games.

The second type. Three players: companies that do not own land, but have money for the lease of land, companies that have land in private ownership and St. Petersburg without having its own land.

The third type. Two players: companies that have land in private ownership, but it is used as collateral by the credit organization and St. Petersburg, which owns the land for construction.

The fourth type. Two players: companies that have a kindergarten building on the right of private property, and St. Petersburg.

The fifth type (trivial game). One player: St. Petersburg, which owns the land and budget for construction.

The sixth type (trivial game). One player: companies that have a building on the right of private property and money for the organization of a kindergarten in it.

The seventh type (combined game). Three players: companies that do not own land, but have funds, companies that have kindergarten building on the right of private property, and St. Petersburg, owning the land.

The eighth type (combined game). Three players: companies that do not own land, but have funds, companies that have land in private ownership and St. Petersburg, owning the land.

From these statements, we may formulate the theorem on the existence of situations of strict Nash equilibrium for the considered classes of games among the possibility of adjusting the institutional environment, which forms the basis for effective cooperation in solving problems associated with the development of social infrastructure of the region.

Theorem: there is a situation of strict Nash equilibrium for the considered types of games among upon condition that the institutional environment can be adjusted.

4 Conclusion

The following results are obtained in the paper.

The mechanism of the interaction between business sector and municipal administration within the service of electronic trading platforms is analyzed.

Criteria complementing the existing procedure of competition between companies that may realize the project of development of social infrastructure have been developed.

The game-theoretic model of interaction between companies and municipal administration of joint social infrastructure projects is described.

It is proved that there are such corrections of the institutional environment that a strict Nash equilibrium exists within game-theoretic models of the described types.

The results can be implemented in the existing concept of electronic trading for effective interaction of municipal administrations with engineering, construction and other companies in the implementation of social infrastructure development projects.

The same kind of reasoning can be carried out for different types of state-private partnerships and a corresponding theorem can be proved.

References:

[1] Medvedev D.A. speech on the Tenth Economic Forum in Krasnoyarsk, http://www.ntv.ru/novosti/463416

[2] *On the State-Private Partnership*. The Project of Federal Law, Ministry of Economic Development of Russian Federation, www.economics.gov.ru

[3] Varnavskiy, V.G., Klimenko, A.V., Korolev, V.A., *State-Private Partnership: Theory and Application*, Gos. Univ. Vyashei Shkoly Ekonomiki, 2010

[4] Glukhov, V.V., Safonov, M.M., Typical models of state-private partnership, *St. Petersburg State Polytechnical University Journal. Economics*, No.6 (112), 2010, pp. 170-174

[5] Yurev, V.N., *Optimization methods in economics and management*, , Izd-vo Politechn. Un-ta, 2006

[6] Bogolib, T.M., International experience and world tendencies of developing State-Private

Partnership, *World Applied Sciences Journal*, Vol. 23, No. 3, 2013, pp. 360-369

[7] Varoufakis, Y., Modern and postmodern challenges to game theory, *Erkenntnis*, Vol. 38, No. 3, 1993, pp. 371-404

[8] Kuzniecova, T.E., Lebedev, N.A., Nikiforov, L.V. Conditions and prospects of development of modern Russia, *World Applied Sciences Journal*, Vol. 24, No. 8, 2013, pp. 1059-1064

[9] Ilin, I.E., New changes in federal law on the state-private partnership, *Auditor*, No. 5 (219), 2013, pp. 12-15

[10] Makarov, I.N., State-private partnership today. Modern economics: regulation and partnership, *Russian Entrepreneurship*, No. 8-2, 2009, pp. 22-28

[11] Cheberko, E.F., New trends in state and business, *St. Petersburg State University Journal*, Vol. 5, No. 4, 2008, pp. 22-31

[12] Lvov, D.S., *Institutional Economics*, Infra-M, 2001.

[13] Anisiforov, A.B., Ilin, I.V., Silkina, G.Y., Yurev, V.N., *Innovative development of industrial cluster*, Izd-vo Politechn. Un-ta, 2012

[14] Varnavskiy, V.G. *State-private partnership*, IMEMO RAN, 2009

[15] Petrosyan, L.A., Senkevich, N.A., Semina, E.A., *Game theory*, Knizhny dom "Universitet", 1998

Modeling the Unreliability and Condition Evolution of Engine Room Equipment with Respect to Maintenance and Overhaul Effect

Lenka Jirsová and Libor Jelínek

Abstract—The paper deals with mathematical modeling of failure rate of steam turbine hall mechanical equipment during its designed lifetime. There is described a connection of the failure rate model and equipment condition model with respect to wear that can be verified on the fly. Particularly the modeling is focused on typically unit manufactured components with low number of failures, long lifetime, various operation modes and a significant influence of preventive maintenance and repairs. The aim is to provide relevant information about expected failure rate and equipment condition evolution over time. Primarily, the prediction of these characteristics in the subsequent maintenance period is needed for the purpose of preventive maintenance and overhaul efficient scheduling. A conceptual framework for the model creation and its parameterization is presented, with a more detailed analysis of how to deal with the specific modeling tasks in the area where the commonly used methods and procedures fail.

Keywords—failure rate model, equipment condition , engine room equipment, preventive maintenance planing support

I. Introduction

ONE of the key activities of the company's management is the planning of investments in maintenance and related decision-making. The concern of managers is to make an optimal decision that has to be backed up by a thorough analysis and calculations. In the field of energetics, the planning is even more important due to its far-reaching consequences. So a suitable maintenance plan should be based on the current state of the monitored equipment and prediction of the state during the next planning period. This characteristics can be obtained from the corresponding model [1].

The equipment for electricity production is a complex mechanical mechanism. Its parts gradually wear out during operation and this leads to a failure of their normal function. The equipment condition is thus determined by the degree of wear and is closely related to the failure rate expressed as the risk of failures occurrences. A very important part of the operation of power plant equipment is its scheduled maintenance and repairs. It allows achieving of long-term reliability parameters during normal operation. The extent and effects of performed repairs are so significant that it is necessary to take into account their impact even during the failure rate model construction[2].

The starting point for failure rate modeling is the analysis of the failures occurrence probability during its lifetime [3]. This is determined by the equipment technical parameters given by the manufacturer, by the statistical evaluation of the failures occurrence history or by the history of the similar type of equipment.

An appropriate design of failure rate model requires to take into account all available information, i.e. statistical and expert estimations, in order to achieve the best fit of model and reality - all under the given operation conditions and maintenance. However, for systems with high reliability is not easy to parameterize model accurately due to the lack of relevant data. Also there is often problem with continuous updates (model refinement). For these reasons is suitable to use analogue model interpretation based on the equipment condition in term of wear. For the mechanical components the wear is very closely related to the failure rate. Wear is also moreover well-interpretable by a technical service, which can provide valuable subjective or objective findings for current state determination. Due to interactions between these models more precise estimation of risk of failures occurrence in the next period can be made, and a better justification for decision-making can be obtained.

Often it is necessary to estimate the risk of failures occurrence in the future for certain functional part or for the whole set of elements. For the model based on a hierarchical structure of elements the mentioned problem can be effectively solved through simulation Monte-Carlo. Then, the results can be processed in the form of the required prediction of temporal development of equipment failure.

A crucial event in terms of preventive maintenance is so-called depletion of life which is usually accompanied by a sharp increase of failure rate caused by exceeding the limit of wear.

For effective decision-making on priorities in the maintenance activities, the evaluation based on the model results should be focused on the prediction of this event. Beyond assessing the risks of failures occurrence and following downtime it is also important to take into account the economical aspect and the related implications, which needs to be solved simultaneously.

II. Modeling of equipment failure rate evolution

A. Difficulties of failure rate modeling in energetics

For power plant equipment the failure rate is mostly expressed by the failure rate function, so it is appropriate to describe its determination. Each elementary component of monitored system is assigned a sub-model in the form respecting the character of the failure source and development in time. These sub-models are associated in parallel or in series. The type of connection depends on the functional arrangement of the components. The generalization to higher functional units is modeled by using a hierarchical model structure.

This work was supported by the European Regional Development Found (ERDF), project NTIS New Technologies for the Information Society, European Centre of Excellence, CZ.1.05/1.1.00/02.0090.

L. Jirsová and L. Jelínek are with the European Research Centre of Excellence NTIS - New Technology for Information Society, University of West Bohemia, Pilsen, 30614 Czech Republic, e-mail: lenty@ntis.zcu.cz.

ISBN: 978-1-61804-251-4

For the modeling of the failure rate of the mechanical equipment a bathtub function is used. This function reflects the failure rate development over time, i.e. the development over time of the mean time between failures [4]. The commonly used method for determining the parameters is the statistical analysis of time between failures from the large amounts of similar type components operated and maintained under the similar conditions. However, for given equipment it is very difficult to use this approach due to its uniqueness or specific operating conditions.

Thus, the difficulty of the task is given mainly by the following facts

1) **Sparse data** the lack of statistical significance (few samples, poor evidence of history).
2) **Absence of data** representing the **end of life**.
3) **Data inhomogeneity** - unsuitable data association from the different operational conditions.
4) **Difficult estimation of preventive maintenance effect**.

From these facts it is obvious that in this case a specific modification or extension of the basic statistical estimation methods, based on a priori information and expert estimates must be used. Specifically, it regards the principles of pooling data from similar components by using statistical methods though that are not commonly used for this problem. Impact of few failure data is f.e. discussed in [5].

Primarily the weighted histograms of failures in the specific time intervals are used instead of time between failures. Further the identification methods based on the minimization of the relative error are used, and finally the different kinds of expert-oriented methods for smoothing and interpolation of failure rate characteristics are used. These methods are described in more details in [6].

As mentioned before, due to the practical unavailability of a precise failure rate model for that kind of equipment it is advisable to use the principle of continuous refinement of the model using all available current data from real operation. For this reason, it makes sense to switch to alternative forms of failure model representation - to the wear state model. Using this approach gives significantly better chance to obtain an appropriate assessment of the actual situation. The equipment condition can be obtained either by subjective visual detection or by objective measurement of subcomponents condition. The assessment of very low instantaneous values of immediate failure rate is practically realizable. That is why the proposed new approach is better.

B. Reliability modeling using failure rate function

In terms of operation reliability the modeling can be performed for part of the equipment system, called a logical unit as a block diagram where each combination of parts or elements represents the dependence of reliability of the part on the whole system. Each monitored element of the system is usually replaced by one element of the diagram. Each of the basic elements is expressed by its reliability, the bathtub curve representing the failure rate evolution over time of a given particular component.

In energetics, the failure rate is most often measured as the number of events per time unit or inversely in the mean time between failures [4]. Equipment failure rate is modeled mainly using the exponential or Weibull distribution or their modifications. Very interesting overview of the Weibull distribution modification can be found in [7]. Some modifications have also been successfully implemented in the model, which was developed at Department of Cybernetics, UWB.

To facilitate the implementation of some of the modifications, the failure rate function has following forms

$$\lambda(x) = \frac{\beta_1 x^{\beta_1 - 1}}{\eta_1^{\beta_1}} + \frac{\beta_2 x^{\beta_2 - 1}}{\eta_2^{\beta_2}} \tag{1}$$

or

$$\lambda(x) = \frac{\beta_1 x^{\beta_1 - 1}}{\eta_1^{\beta_1}} + \frac{\beta_2 x^{\beta_2 - 1}}{\eta_2^{\beta_2}} + \frac{\beta_3 x^{\beta_3 - 1}}{\eta_3^{\beta_3}}. \tag{2}$$

where β and η are the parameters of the chosen probability distribution.

C. Failure rate estimation beyond the horizon of available data

The available historical data describes only part of the failure rate evolution over time which mostly does not show the effects of aging. Given that it is necessary to estimate the future states that, despite repeatedly performed preventive repairs (considered as a partial recovery), will lead to extreme increase of the failure rate as a result of end of the technical life of the individual components. To calculate this part of the failure rate evolution only the available expert estimates of service life may be used in compliance with operation and maintenance given by manufacturer.

The estimate of the slope of the failure rate increase between the time of the given service life and the time of reaching the ultimate limit state specified by exceeding 50% of this life limit is fundamental for the new approach. The estimate of this slope will be discussed further.

Determination of prediction points of the future failure rate (i.e. beyond the horizon of real data) can then be realized using the following analytical form

$$\widehat{\lambda_f} = \lambda_k \cdot 10^{\frac{T_f - T_k}{T_l - T_k}}, \tag{3}$$

where λ_k is a mean failure rate during the last period of failure data evidence after previous overhaul, T_k is the time localization of the failure rate value placed in the middle of the interval between last general overhaul and time of the last data record, T_f represents the time localization of the desired point of computed failure rate value placed in the middle of the interval between time of the last data record and future overhaul (statistically calculated from history) and T_l is an expert estimation of the time of ultimate limit state.

D. Impact of scheduled maintenance elimination

If the given data from real operation are used, it is also necessary to take into account the fact, that the data are influenced by carrying out the scheduled maintenance especially repair with renovation when overhaul is performed. The main effect

ISBN: 978-1-61804-251-4

of the scheduled maintenance is repeated reduction of the effective aging of the components due to their partial recovery during maintenance.

Elimination of this effect is based on the time transformation after each carried out maintenance. It is assumed that the result of the partial recovery of the individual components is the effective age T_a reduction according to the relation

$$T_{E_k} = (T_{E_{k-1}} + T_{IM_k}) \cdot (1 - arf_k) \tag{4}$$

where T_{IM_k} is the overhaul interval and the arf (age reduction factor) characterizes the degree of recovery of components.

The parameter arf values have no direct physical meaning and they cannot be obtained even for short-term measurements. Moreover, the actual parameter cannot be even properly estimated by the experts. It is more suitable to use expert estimations of two other parameters the expected system components service life without carrying out the scheduled maintenance T_{L_C} with carrying out the scheduled maintenance T_{L_M}. Then the desired parameter arf can be identified. The principle of conversion is based on the above assumption about life extension by scheduled maintenance.

It is assumed that the sequence of time points given by subsequently added intervals between scheduled maintenances T_{IM} in a finite number N. The last scheduled maintenance reduces the age to $T_{L_C} - T_{IM}$, and after a subsequent interval T_{IM} is achieved the limit state of the component. Then the reduced age can be calculated by the cumulative application of relation (4) as follows

$$\begin{aligned} T_{L_C} = ((T_{IM_1} \cdot (1 - arf) + T_{IM_2}) \cdot (1 - arf) + \\ + T_{IM_3}) \cdot (1 - arf) + T_{IM_4} \cdots \end{aligned} \tag{5}$$

This relation can be modified to the form

$$T_{L_C} = \sum_{i=1}^{N} T_{IM_k} \cdot (1 - arf)^{i-1}, \qquad 0 < arf < 1. \tag{6}$$

For constant value of T_{IM} and given $N = round(T_{L_C}/T_{IM})$ the equation 6 can be also expressed analytically as the sum of the first N members of the geometric progression

$$T_{L_C} = T_{IM} \cdot \frac{(1 - arf)^N - 1}{(1 - arf) - 1}. \tag{7}$$

Parameter arf is the argument of this function and it is needed to perform the inversion. The analytical solution is not applicable for its complexity, but there is an approximate numerical solution of the parameter arf estimated in the range $(0, 1)$ using the successive approximations by halving the interval with adjustable maximum estimation error.

This relation may be used ti determine the arf parameter. A demonstration of the results of provided by relations (4) - (7) can be seen in Figure 1

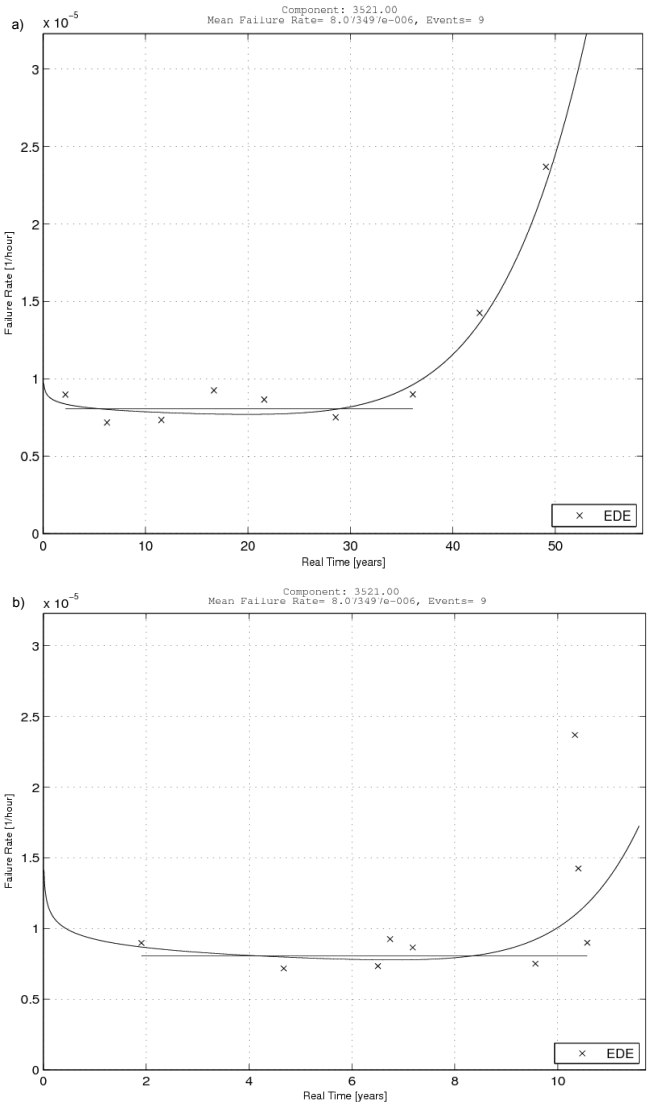


Fig. 1. Failure rate evolution over time a] without scheduled maintenance providing, b] with scheduled maintenance providing

E. Timeline scaling of the failure rate function

Another problem of the engine room equipment failure rate modeling is a normalization of failure rate characteristics of the same types of equipment obtained under the different operation conditions. The normalization principle is based on a time transformation of failure rate function. In order to combine data from different operation conditions, it is necessary to perform correction (scaling) of time between failures values based on current operation conditions (operation hours, starts, etc.). From the normalized data it is possible to deduce the reliability model independent of specific conditions. However, if a reliability model of a specific equipment is available then in order to be able to compare it with the corresponding equipment in another facility all it is necessary to recalculate are the model parameters but not the data. This can be done on the principle of time scale transformation of failure characteristics function in accordance to the effective

ISBN: 978-1-61804-251-4

time consumption criteria. In order to comply with certain assumptions, the transformation of a random variable allows the probability distribution change and for example simplify a complex model for easier implementation and calculations.

For further use let denote the system reliability time as τ, while calendar time will be denoted as t. τ and t are diffeomorphic. Suppose that there is a clear relation $t \to \tau$ for which functional dependencies $t = G(\tau)$ and $\tau = G^{-1}(t)$ can be defined that are continuous, increasing and positive. Further it is assumed that the transformation is linear function, i.e.

$$G(\tau) = k \cdot \tau. \tag{8}$$

If the failure model is represented by the exponential distribution of random variable t, the transformed random variable τ has distribution function $F(\tau)$ with a constant parameter λ defined as

$$F(\tau) = 1 - e^{-\lambda(k\tau)} = 1 - e^{-\widehat{\lambda}\tau} \tag{9}$$

with $\widehat{\lambda}$ representing the linear transformation of λ given as

$$\widehat{\lambda} = \lambda \cdot k. \tag{10}$$

Let consider the Weibull distribution function with constant parameters β and η given as

$$H(t) = \left(\frac{t}{\eta}\right)^{\beta}. \tag{11}$$

Linearly transformed random variable will have the distribution function related on 9. Substituting into $t = G(\tau)$ and modification of $H(t)$ to its original form we obtain the following formula

$$\begin{aligned} \widehat{H}(\lambda) = H\left(G(\tau)\right) = \left(\frac{G(\tau)}{\eta}\right)^{\beta} = \left(\frac{k \cdot \tau}{\eta}\right)^{\beta} = \\ = \left(\frac{\tau}{\frac{\eta}{k}}\right) = \left(\frac{\tau}{\widehat{\eta}}\right)^{\widehat{\beta}}. \end{aligned} \tag{12}$$

It follows that the linear time scale transformation causes a change in the parameter β and η values hereby

$$\widehat{\beta} = \beta, \tag{13}$$

$$\widehat{\eta} = \frac{\eta}{k}. \tag{14}$$

The failure model of a engine room equipment is typically composed of sub-components, each of which is represented in the form of the Bi-Wibull or Tri-Weibull distribution function. The failure rate has a typical bathtub curve. In accordance to the above described procedure applied to all the parts of the distribution function of each component the failure rate curve transformation is determined as shown in Figure 2.

A change of the distribution function causes modification of the corresponding random variable time scale.

Using the above-mentioned procedure, it is possible to obtain the bathtub curve of the failure rate time progress

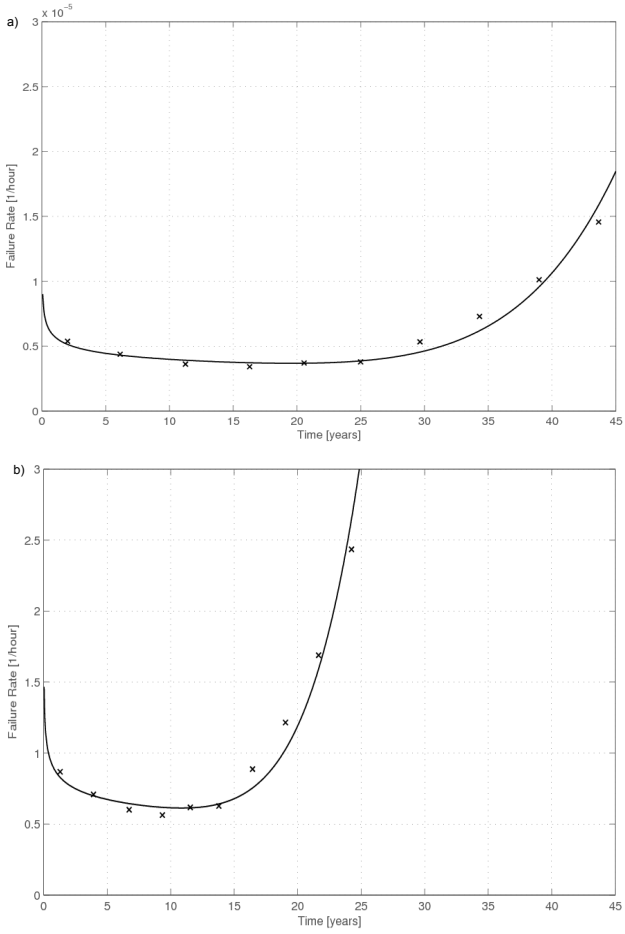


Fig. 2. Failure rate evolution over time a] with original parameters, b] after the transformation

including end of lifetime period, whole expressed in the effective operational time, clear from influence of operational conditions and renewal repairing effects.

Finally, the expert intervention to the failure rate function can be incorporated to the beginning of the curve, because in the beginning of operation the failure occurs rarely for this type of equipment.

III. Modeling of equipment condition evolution

Throughout the world, the information systems are developed with purpose of decision making and maintenance planning support. They are usually based on data given failure risk prediction for the next preventive maintenance planning period. It can be done based on development over time of the state of wear. However, for the proper operation of the information system it is useful to rely on a wear evolution over time rather then failure model, because it is more transparent, verifiable and useful in the maintenance planning process. For that purpose is necessary to transfer basic failure rate model to the dynamic wear state model, which would be able to take into account the corrections based on an actual findings from maintenance.

ISBN: 978-1-61804-251-4

Dynamic state model in terms of wear is directly derived from the failure rate model. State of the wear is given by failure probability after commissioning. This value is a function of the relative lifetime and is directly related to the integral of the failure rate function $\lambda(t)$. This function indicates what extent the wear probably occurred.

The failure rate function $\lambda(t)$ is related to the relative life of the equipment, which takes values $0 - 100\%$ of the total estimated useful life. Failure rate function values are also normalized so that the area under the curve is unit, i.e.

$$\int_0^{100} \lambda(t)dt = 100\%. \tag{15}$$

If we consider, that the new (not worn) equipment has 100% condition, then the state of wear denoted as s_M is given by

$$s_M(t) = 100 - \int_0^t \lambda(\tau)d\tau. \tag{16}$$

If it is need to express a condition change in term of wear after a certain time period, which is recalculated to the relative lifetime segment expressed by interval t_{k-1}, t_k then it is possible to express the state in the form

$$s_M(t_k) = s_M(t_{k-1}) - \int_{t_{k-1}}^{t_k} \lambda(\tau)d\tau. \tag{17}$$

The following relationship determine the value of the corrected state s_C based on a model state s_M and the wear state specified by the s_U and their credibilities

$$s_C = s_M \cdot \frac{c_U}{c_M + c_U} + s_U \cdot \frac{c_M}{c_M + c_U}, \tag{18}$$

where s_M and s_U is the state of equipment given by the model and the user, respectively. The credibilities are than denoted c_M and c_U for credibility of the model state s_M and user state s_U, respectively.

More about the development of state model of equipment condition you can find in [8]

IV. Conclusion

IS deployed and verified in real in Czech power plant. It deals with the failure rate progress model construction in relation to the dynamic wear state model. There are discussed difficulties and solving methods of the insufficient historical failure data moreover affected by variable operating and maintenance conditions. Finally is proposed the transfer of failure model to the wear state model that offers the better possibilities to verify and refine of the results in form of wear out prediction connected with the failure risk consequences. The proposed model offers a significant support in maintenance optimization with respect to limited resources.

References

[1] A. C. Marquez and A. S. Heguedas, "Models for maintenance optimization: a study for repairable systems and finite time periods," *Reliability Engineering & System Safety*, vol. 75, no. 3, pp. 367–377, 2002.

[2] G. Weckman, R. Shell, and J. Marvel, "Modeling the reliability of repairable systems in the aviation industry," *Computers & Industrial Engineering*, vol. 40, no. 1-2, pp. 51–63, 2001.

[3] K. Xie and W. Li, "Analytical model for unavailability due to aging failures in power systems," *International Journal of Electrical Power & Energy Systems*, vol. 31, no. 7-8, pp. 345–350, 2009.

[4] B. Šedivá, E. Wagnerová, F. Vávra, T. Ťoupal, and P. Marek, "Statistical monitoring of failures - methods and use," *Proceedings of the 11th International Scientific Conference Electric Power Engineering 2010*, pp. 611–615, 2010.

[5] R. Laggoune, A. Chateauneuf, and D. Aissani, "Impact of few failure data on the opportunistic replacement policy for multi-component systems," *Reliability Engineering & System Safety*, vol. 95, no. 2, pp. 108–119, 2010.

[6] L. Houdova, L. Houdova, L. Jelinek, and E. Janecek, "Approach to solving the task of availability prediction and cost optimization of a steam turbine," *Proceedings of the International Conference on Information Technology Interfaces, ITI*, pp. 629–634, 2010.

[7] H. Pham and C.-D. Lai, "On recent generalizations of the weibull distribution," *Reliability, IEEE Transactions on*, vol. 56, no. 3, pp. 454–458, Sept 2007.

[8] L. Jirsová and M.Flídr, "Dynamic state model of steam turbine hall equipment condition for maintenance planning and decision-making support," *MMMAS 2014*, 2014, submitted for publiccation.

On improvement of fault-tolerance in distributed hardware-software multi-agent systems and assessment of assured reliability

Alexei V. Igumnov, Sergey E. Saradgishvili

***Abstract*—**The problem of fault-tolerance in industry systems that are based on agent technology and consist of an agent model, hosts required for execution of agents and actuators required for interaction of agents with an environment is considered. New model of distributed hardware-software multi-agent system is presented. The reorganization technique is developed based on replication of tasks and actuators and on an introduction of redundant sets of agents and hosts. The presented fault-recovery methodology is developed to deal with failures of tasks, agents, hosts and actuators. Conditions required for success of the developed fault-recovery methodology are determined and the methodology is validated by the stated and proved theorem on fault-tolerance property of redundant distributed hardware-software multi-agent system. The developed methodology for formation of an operability function enables synthesis of analytic reliability function in accordance with logical-and-probabilistic methods. The level of fault-tolerance achieved by the presented fault-recovery methodology is equal to the theoretical assessment of probability of no-failure in accordance with described experiments.

***Keywords*—**fault-recovery, fault-tolerance, multi-agent system, operability function, redundancy, reliability.

I. Introduction

Multi-agent system (MAS) is usually considered as a system in which several agents operate and interact with one another [1]. The lack in fault-tolerance is recognized as one of root causes of small amount of deployed real world multi-agent systems [2], [3].

The problem considered in the article is defined by the following contradiction. On the one hand there is an interest in application of multi-agent technologies to development of various industry systems. We consider an industry system that is based on agent technology as the object of our research and name it distributed hardware-software (DHS) multi-agent system. We state that DHS MAS could not be represented by only a set of agents and shall include hosts required for execution of an agent model and actuators required for interaction of agents with an environment. Thus DHS MAS is prone to faults of each of its components. Whereas it is important to improve fault-tolerance of DHS MAS it is also required to determine a level of the improvement or an assured level of fault-tolerance. On the other hand existing fault-tolerance methods for MAS such as DarX [4], AAA for broker teams [5], Sentinels approach [6], FATMAS [7] consider MAS a set of agents and thus improve fault-tolerance only in cases of occurred faults of individual agents or faults of hosts on that an agent model is deployed for execution. Existing approaches do not provide any assessment of a level of fault-tolerance assured by their utilization and their effectiveness is proved only experimentally.

The objective of the research is to improve fault-tolerance of DHS MAS in cases of failures of agents, tasks of agents, hosts and actuators and to enable a theoretical assessment of assured fault-tolerance level during a design phase.

A. V. Igumnov is with the Information and Control Systems Department, Institute of Computing and Control, St. Petersburg State Polytechnical University, St. Petersburg, Russia (phone: 8-921-767-10-17; e-mail: Alexei.Igumnov@gmail.com).

S. E. Saradgishvili is with the Information and Control Systems Department, Institute of Computing and Control, St. Petersburg State Polytechnical University, St. Petersburg, Russia (e-mail: SSaradg@yandex.ru).

II. Reorganization technique

Incorporating redundant copies of system components has been recognized as one of the best methods for improvement of fault-tolerance. To define our own reorganization technique we shall at first develop a new model of DHS MAS. We have introduced a term of an agent platform (AP) considered as a component intended for execution of agents and a term of an actuator (ACT) considered as a component intended for interaction with an environment in which MAS is situated. Using such elements of MAS model introduced by FATMAS methodology [7] as tasks and agents we define DHS MAS as an ordered set $MAS = <T, A, HWP, HWR>$, where T is a set of tasks, A – a set of agents, HWP – a set of APs, HWR – a set of ACTs. The configuration of MAS is defined via following set of predicates:

- *confTaskAgent(t, a)* determines whether a particular task *t* belongs to a particular agent *a*;
- *confAgentHwp(a, hwp)* determines whether a particular agent *a* is deployed in a particular AP *hwp*;
- *reqHwrTask(hwr, t)* determines whether a particular ACT *hwr* is required for performing of a task *t*;
- *confHwrHwp(hwr, hwp)* determines whether a particular ACT *hwr* is accessible for a particular AP *hwp*, i.e. for all tasks of all agents deployed in AP *hwp*.

We consider a task of DHS MAS as a minimal functional

ISBN: 978-1-61804-251-4

instance that represents some functionality in accordance with system specification and requirements. We suppose that DHS MAS is in failure if it is not able to perform at least one of its tasks. Failures of all other components lead to inability to perform one or more of system tasks, e.g. tasks deployed in the agent that is in failure state, tasks deployed in all agents situated in AP that is in failure state or tasks that require utilization of the actuator that is in failure state.

It's suggested to distinguish components of DHS MAS that are unique in terms of implemented functionality such as tasks and ACTs and components that act as universal executive containers. We suppose that each AP shall be able to execute any agent and similarly each agent is supposed to be able to execute any of systems tasks.

Our reorganization technique is based on replication (i.e. creation of a full copy) of functional components as well as on introduction of redundant sets of universal executive containers. Terms of task type and actuator type were introduced based on sets of tasks *T* and actuators *HWR* of an existing DHS MAS to deal with identification of equivalent components in such manner that a univocal correspondence exists between a set of types and a set of corresponding components of the existing system. Having introduced sets of types we shall turn the necessity of utilization of a particular ACT for performing of a particular tasks to the necessity of utilization of ACT of a particular type for performing of a task of a particular type.

The developed reorganization technique comprises of following steps:

- define a set of task types and a set of actuator types;
- define necessity of utilization of ACT of a particular type by a task of a particular type;
- introduce replication of tasks and ACTs and define a set of tasks and a set of actuators;
- introduce a redundant set of agents and a redundant set of agent platforms;
- define configuration of DHS MAS in terms of existence of communication links between agent platforms and actuators, deployment of tasks on agents and deployment of agents on APs.

Application of our reorganization technique turns an existing DHS MAS to a redundant one that is defined as an ordered set *RMAS = <TT, RT, RA, RHWP, THWR, RHWR>*, where *TT* is a set of task types, *RT* – a set of tasks, *RA* – a set of agents, *RHWP* – a set of APs, *THWR* – a set of actuator types and *RHWR* is a set of actuators. The configuration of the MAS is defined via predicates as follows:

- *confTypeTask(tt, t)* is true if a type of a task *t* is *tt*;
- *confTypeHwr(thwr, hwr)* determines whether a type of an actuator *hwr* is *thwr*;
- *confTaskAgent(t, a)* determines deployment of tasks on a set of agents;
- *confAgentHwp(a, hwp)* determines deployment of agents on a set of agent platforms;
- *reqTHwrTTask(thwr, tt)* determines whether an actuator of a type *thwr* is required for performing of a task of a type *tt*;
- *confHwrHwp(hwr, hwp)* determines whether a particular actuator *hwr* is accessible for AP *hwp*.

We suppose that there shall be one and only one task in a replication group of equivalent tasks of the same type that exert an influence on an environment. Such task is called an active replica of a particular type. It's worth noting that one and only one active replica for each task type shall exist in a redundant DHS MAS. The model of a redundant DHS MAS is described in details in [8], [9].

III. Fault-Recovery Methodology

A. Database Schemas

To enable interaction of components in a redundant DHS MAS we have introduced a message-based communication protocol. We suppose that communication may occur only between an executive container and components that are deployed in this container and between agent platforms. It's also assumed that each AP shall be able to interact with all remote agent platforms via sending broadcast messaging.

We state that agents and APs of DHS MAS shall have limited knowledge regarding system configuration that is enough for successful operation. Thus we have defined database (DB) schemas for agents and APs. An agent *a* shall manage a DB *ADB(a) = {ATBT(a), ATBR(a), ATBAT(a)}* wherein:

- deployment table *ATBT(a)* contains records *(t, tt)* that links the task *t* deployed in the agent and its type *tt*;
- required actuators table *ATBR(a)* contains records *(tt, {thwr$_i$})* that links a type of task *tt* with required types of actuators for all task types from *ATBT(a)*;
- active replicas table *ATBAT(a)* contains records *(tt, t)* that links a type of task *tt* and an active replica *t* deployed in the agent in accordance with *ATBT(a)*.

An agent platform *hwp* shall manage a DB *HDB(hwp) = {HTBD(hwp), HTBR(hwp), HTBAR(hwp), HTBAT(hwp)}* wherein:

- deployment table *HTBD(hwp)* contains records *(a, t, tt)* that links an agent *a* deployed in AP, a task *t* deployed in the agent *a* and its type *tt*;
- required actuators table *HTBR(hwp)* contains records *(tt, {thwr$_i$})* and acts as an aggregator of *ATBR(a)* tables of all agents deployed in AP;
- available actuators table *HTBAR(hwp)* contains records *(hwr, thwr)* that links an available ACT *hwr* with its type;
- active replicas table *HTBAT(hwp)* includes record *(tt, a, rhwp)* for each task type *tt* that links the task type either with an agent *a* deployed in this AP if an active replica of type *tt* is deployed in this agent or with remote agent platform *rhwp* if an active replica of type *tt* is deployed in one of agents of remote AP.

It's worth noting that in a redundant DHS MAS a request to execute a particular task shall be turned to a request to execute a task of a particular type. To enable transparency of our

reorganization technique we suggest handling such requests iteratively by the agent that has received the request from one of its tasks, AP in which the agent is deployed and all remote APs. Let's consider processing of a request *pfm(tt)* to execute an active replica of type *tt* received by the agent *a* located in AP *hwp* from one of its tasks. The request is handled in following manner:

- if there is a task *t'* that is an active replica of type *tt* in accordance with *(tt, t')* record in table *ATBAT(a)* then the agent *a* shall execute it, otherwise the agent *a* shall escalate the processing to AP *hwp* through the *pfm(tt)* message;
- if an active replica of type *tt* is located in the agent *a'* of AP *hwp* in accordance with *(tt, a', 0)* record of table *HTBAT(hwp)* then AP *hwp* shall request the agent *a'* to execute an active replica through the *pfm(tt)* message;
- if an active replica of type *tt* is located in one of agents of one of remote APs *rhwp* in accordance with *(tt, 0, rhwp)* record of table *HTBAT(hwp)* then AP *hwp* shall escalate the processing to AP *rhwp* though the *pfm(tt)* message.

B. Fault-Recovery Procedures

We state that a redundant DHS MAS is in failure if it is not able to perform an active replica of at least one type. In turn failure of each component of a redundant DHS MAS leads to inability to perform one or more of system tasks. It's worth noting that some of these tasks are active replicas and some are not. We define fault-recovery task as a problem of search and activation of new replicas of all required types. All fault-recovery procedures are supposed to be performed iteratively in accordance with components hierarchy, i.e. the search procedure starts from the component responsible for failure detection, is escalated though a set of executive containers and may finish in one of remote APs.

We introduce a term of a consistent configuration of a redundant DHS MAS as follows:

- deployment table *ATBT(a)* of each agent *a* contains *(t, tt)* record if and only if the task *t* is deployed in the agent *a* (i.e. *confTaskAgent(t, a) = true*) and the task *t* could be executed;
- deployment table *HTBD(hwp)* of each AP *hwp* contains *(a, t, tt)* record if and only if the task *t* is deployed in the agent *a* that is deployed in AP *hwp* (i.e. *confAgentHwp(a, hwp) = true* & *confTaskAgent(t, a) = true*), the agent *a* is not in failure state and the task *t* could be executed;
- active replicas table *ATBAT(a)* of each agent *a* contains record *(tt, t)* if and only if the task *t* is an active replica of type *tt* and there is a record *(t, tt)* in *ATBT(a)* (i.e. the task *t* could be executed);
- available actuators table *HTBAR(hwp)* of each AP *hwp* contains record *(hwr, thwr)* if and only if ACT *hwr* is not in failure and is accessible (i.e. *confHwrHwp(hwr, hwp) = true*);
- active replicas table *HTBAT(hwp)* of each AP *hwp* contains record *(tt, a, 0)* if and only if the agent *a* is deployed in AP *hwp* (i.e. *confAgentHwp(a, hwp) = true*), the agent *a* is not in failure and there is a record *(tt, t)* in active replicas table *ATBAT(a)*;
- active replicas table *HTBAT(hwp)* of each AP *hwp* contains record *(tt, 0, rhwp)* if and only if AP *rhwp* is not in failure and there is a record *(tt, a', 0)* in active replicas table *HTBAT(rhwp)*;
- active replicas table *HTBAT(hwp)* of each AP *hwp* contains record *(tt, a, rhwp)* for each task type *tt*.

An occurred failure of each component of a redundant DHS MAS breaks a consistent configuration. Thus an appropriate fault-recovery procedure shall restore a consistent configuration by excluding records related to failed components and updating records related to locations of new active replicas in databases.

The *a_r_tpfailt(t)* procedure shall be performed by an agent *a* that has detected a failure of its particular task *t* and comprises of following steps:

- determine a type *tt* of the task *t* in accordance with the deployment table *ATBT(a)*;
- remove a record *(t, tt)* from *ATBT(a)*;
- remove a record *(tt, t)* from the active replicas table *ATBAT(a)*;
- send the *a_cmd_rm_task(t)* message to AP *hwp* in that the agent is deployed (processing of the received message by AP will result in excluding *(a, t, tt)* record from the deployment table *HTBD(hwp)*);
- if the task *t* is not an active replica of type *tt* than finish processing;
- if there is a record *(t', tt)* in the deployment table *ATBT(a)* then select *t'* as new active replica, add a record *(tt, t')* to the active replicas table *ATBAT(a)* and finish processing;
- send the *req_atask(tt)* message to AP *hwp* in that the agent is deployed to escalate the task of fault-recovery.

On reception of the *req_atask(tt)* message from one of its agents AP *hwp* shall perform the *h_r_tfail(tt)* procedure that comprises of following steps:

- if there is a record *(a', t', tt)* in the deployment table *HTBD(hwp)* then select the task *t'* deployed in the agent *a'* as new active replica, update the active replicas table *HTBAT(hwp)* to contain a record *(tt, a', 0)* for a type of task *tt*, send the *h_cmd_atask(t')* message to the agent *a'* and finish processing;
- send the broadcast *req_atask(tt)* message to all APs to escalate the task of fault-recovery.

On reception of the *h_cmd_atask(t')* message the agent *a'* shall add a record *(tt, t')* to the active replicas table *ATBAT(a).*

The *recv_req_atask(tt)* procedure is performed by each AP *rhwp* on reception of the broadcast *req_atask(tt)* message from one of APs and comprises of following steps:

- if there is a record *(a', t', tt)* in the deployment table *HTBD(rhwp)* then select the task *t'* deployed in the agent *a'* as new active replica, update the active replicas table *HTBAT* to contain a record *(tt, a', 0)* for a type of task *tt*, send the *h_cmd_atask(t')* message to the agent *a'*;
- if there is a record *(tt, a', 0)* in the active replicas table

HTBAT(rhwp) then send the broadcast *atask_announce(tt, rhwp)* message to all APs.

On reception of the broadcast *atask_announce(tt, rhwp)* message each AP *hwp* shall update an active replicas table *HTBAT(hwp)* to contain a record *(tt, 0, rhwp)* for a task type *tt*.

The *h_r_afailt(a)* procedure shall be performed by AP *hwp* that has detected a failure of its particular agent *a* and comprises of following steps:

- let *TF* be a set of tasks deployed in the agent *a*, *TTF* be a set of task types such that the active replica of each type from the said set *TTF* is located in the agent *a*;
- add each task *t* such that there is a record *(a, t, tt)* in the deployment table *HTBD (hwp)* to the set *TF*;
- for each task *t* from the set *TF*:
  - if there is a record *(tt, a, 0)* in the active replicas table *HTBAT(hwp)* then add the task type *tt* to the set *TTF*;
  - remove a record *(a, t, tt)* from the deployment table *HTBD(hwp)*;
- perform the *h_r_tfail(tt)* procedure for each task type *tt* from the set *TTF*.

The *h_r_hwrfail(hwr)* procedure shall be performed by each AP *hwp* that has detected a failure of the accessible actuator *hwr* and comprises of following steps:

- determine a type *thwr* of the actuator *hwr* in accordance with the available actuators table *HTBAR(hwp)*;
- remove a record *(hwr, thwr)* from *HTBAR(hwp)*;
- if there is a record *(hwr', thwr)* in the available actuators table *HTBAR(hwp)* then finish processing as another actuator of the same type is available;
- let *TF* be a set of tasks that are in failure caused by a failure of the actuator *hwr*, *TTF* be a set of task types such that an active replica of each type from the said set *TTF* requires utilization of an actuator of type *thwr*;
- add each task *t* of type *tt* such that there is a record *(tt, {thwr$_1$, …, thwr, … thwr$_n$})* in the required actuators table *HTBR(hwp)* to the set *TF*;
- for each task *t* from the set *TF*:
  - if there is a record *(tt, a, 0)* in the active replicas table *HTBAT(hwp)* then add the task type *tt* to the set *TTF*;
  - remove a record *(a, t, tt)* from the deployment table *HTBD(hwp)*;
  - send the *h_cmd_rm_task(t)* message to the agent *a* in that the task *t* is deployed (processing of the received message by the agent *a* will result in excluding *(t, tt)* record from the deployment table *ATBT(a)* and in excluding of *(tt, t)* record from the active replicas table *ATBAT(a)*);
- synchronize all APs;
- perform the *h_r_tfail(tt)* procedure for each task type *tt* from the set *TTF*;

The *h_r_hwpfail(rhwp)* procedure shall be performed by one of APs *hwp* that has detected a failure of AP *rhwp* and comprises of following steps:

- let *TTF* be a set of task types such that the active replica of each type from the said set *TTF* is located in AP *rhwp*;
- add each task type *tt* such that there is a record *(tt, 0, rhwp)* in the active replicas table *HTBAT(hwp)* to the set *TTF*;
- for each task type *tt* from the set *TTF*:
  - remove a record *(tt, 0, rhwp)* from the active replicas table *HTBAT(hwp)*;
  - perform the *h_r_tfail(tt)* procedure;
  - if there is a record *(tt, a, 0)* in the active replicas table *HTBAT(hwp)* (i.e. the *h_r_tfail(tt)* procedure has resulted in activation of new replica in the agent *a* of AP *hwp*) then send the broadcast *atask_announce(tt, hwp)* message to all APs.

IV. Theorem on Fault-Tolerance Property of Redundant Multi-Agent System

To validate the developed fault-recovery methodology we have stated and proved the theorem on fault-tolerance property of a redundant DHS MAS. To define conditions of the theorem we have introduced following additional functions and predicates:

- function *agentOfT(t)* determines the agent in that the task *t* is deployed:

$$agentOfT(t) = a \Leftrightarrow confTaskAgent(t,a) = true, \quad (1)$$

where *t* is a task, *a* – an agent;

- function *hwpOfT(t)* determines AP such that the task *t* is deployed in one of agents of this AP:

$$\begin{aligned} &hwpOfT(t) = hwp \Leftrightarrow (\exists a \in RA: \\ &confTaskAgent(t,a) = true \wedge \\ &confAgentHwp(a,hwp) = true), \end{aligned} \quad (2)$$

where *t* is a task, *hwp* is AP, *a* – an agent, *RA* – a set of agents;

- function *reqTHwr(tt)* determines a set of actuator types required for performing of a task of type *tt*:

$$\begin{aligned} &reqTHwr(tt) = \{thwr \in THWR \mid \\ &reqTHwrTTask(tt,thwr) = true\}, \end{aligned} \quad (3)$$

where *tt* is a task type, *thwr* – an actuator type, *THWR* – a set of actuator types;

- function *tOfTask(t)* determines a type of the task *t*:

$$tOfTask(t) = tt \Leftrightarrow confTypeTask(tt,t) = true, \quad (4)$$

where *t* is a task, *tt* is a task type;

- function *tOfHwr(hwr)* determines a type of the actuator *hwr*:

$$tOfHwr(hwr) = thwr \Leftrightarrow confTypeHwr(thwr,hwr) = true, \quad (5)$$

where *hwr* is an actuator, *thwr* is an actuator type;

- predicates *failT(t)*, *failA(a)*, *failP(hwp)*, *failR(hwr)* determines respectively states of failure of the task *t*, the agent *a*, AP *hwp* and ACT *hwr*;

- predicate *atType(tt, t)* determines whether the task *t* is an active replica of type *tt*.

As was noted above a failure of any component of a redundant DHS MAS leads to inability to perform one or more of system tasks. Thus we introduce the following predicate to define a global state of task failure:

$$\begin{aligned} &failTG(t) = false \Leftrightarrow (failT(t) = false) \wedge \\ &(failA(agentOfT(t)) = false) \wedge (failP(hwpOfT(t)) = false) \wedge \\ &(\forall thwr \in reqTHwr(tOfTask(t)), \exists hwr \in RHWR: \\ &tOfHwr(hwr) = thwr \wedge failR(hwr) = false \wedge \\ &confHwrHwp(hwr, hwpOfT(t)) = true), \end{aligned} \quad (6)$$

where *t* is a task, *thwr* – an actuator type, *hwr* – an actuator, *RHWR* – a set of actuators.

Theorem. The redundant DHS MAS with a consistent configuration will recover from detected failures of agents, tasks, agent platforms and actuators if following conditions are met:

- if failure of a task *t* which is the active replica of type *tt* is detected then there exists another task *t'* of type *tt* which is in an operable state:

$$\begin{aligned} &(failtT(t) = true \wedge atType(tOfTask(t), t) = true) \Rightarrow \\ &(\exists t' \in RT : (tOfTask(t') = tOfTask(t) \wedge failTG(t') = false)). \end{aligned} \quad (7)$$

- if failure of an agent *a* is detected then for each task t_i which is deployed in the agent *a* and is the active replica of type tt_i there exists another task t_i' of type tt_i from a set of tasks *RT* which is in an operable state:

$$\begin{aligned} &\forall t_i \in RT : (failA(a) = true \wedge confTaskAgent(t_i, a) = true \wedge \\ &atType(tOfTask(t_i), t_i) = true) \Rightarrow \\ &(\exists t_i' \in RT : (tOfTask(t_i') = tOfTask(t_i) \wedge failTG(t_i') = false)). \end{aligned} \quad (8)$$

- if failure of an agent platform *hwp* is detected then for each task t_i that is deployed in one of agents of AP *hwp* and is the active replica of type tt_i there exists another task t_i' of type tt_i from a set of tasks *RT* which is in an operable state:

$$\begin{aligned} &\forall t_i \in RT : (failP(hwp) = true \wedge hwpOfT(t_i) = hwp \wedge \\ &atType(tOfTask(t_i), t_i) = true) \Rightarrow \\ &(\exists t_i' \in RT : (tOfTask(t_i') = tOfTask(t_i) \wedge failTG(t_i') = false)). \end{aligned} \quad (9)$$

- if failure of an actuator *hwr* of type *thwr* is detected then for each task type tt_i which requires utilization of an actuator of type *thwr* there exists a task t_i' of type tt_i from a set of tasks *RT* which is in an operable state:

$$\begin{aligned} &\forall t_i \in RT : (tOfHwr(hwr) \in reqTHwr(tOfTask(t_i)) \wedge \\ &atType(tOfTask(t_i), t_i) = true \wedge failR(hwr) = true) \Rightarrow \\ &(\exists t_i' \in RT : (tOfTask(t_i') = tOfTask(t_i) \wedge \\ &failTG(t_i') = false)). \end{aligned} \quad (10)$$

Proof. Let's consider a processing of a request to execute an active replica of a particular type by redundant DHS MAS with a consistent configuration. Let the *pfm(tt)* request be received by an agent a_r deployed in AP hwp_r from its task t_r. As configuration is consistent there is a task t_a that is an active replica of type *tt*. If the task t_a is located in the agent a_r then due to consistent configuration there is a record (tt, t_a) in $ATBAT(a_r)$ and processing of the *pfm(tt)* request by the agent a_r will result in execution of the task t_a. Otherwise the agent a_r will send the *pfm(tt)* message to AP hwp_r. If the task t_a is located in an agent a_1 of the AP hwp_r then there is a record $(tt, a_1, 0)$ in $HTBAT(hwp_r)$ and a record (tt, t_a) in $ATBAT(a_1)$. Thus the processing of the *pfm(tt)* message by AP hwp_r will result in execution of the task t_a by the agent a_1. Otherwise AP hwp_r will send a broadcast *pfm(tt)* message to other APs. If the task t_a is located in some agent a_2 of remote AP hwp_2 then the processing of the *pfm(tt)* message by AP hwp_2 will result in execution of task t_a by the agent a_2 similarly to the case of deployment of the task t_a in the agent a_1 of AP hwp_r.

As the requested active replica is executed independently of its deployment while DHS MAS configuration is consistent and DHS MAS is supposed to be in non-failure state if it is able to execute an active replica of each type we shall proof that fault-recovery procedures restore a consistent configuration.

Let's consider a failure of the task *t* of type *tt*. Let *a* be an agent in that the task *t* is deployed, *hwp* be AP in that the agent *a* is deployed. In accordance with the conditions of theorem there is a task *t'* that is not in failure state. In accordance with *a_r_tpfail(t)* procedure records *(t, tt)* of *ATBT(a)* and *(a, t, tt)* of *HTBD(hwp)* will be removed. At this step a consistent configuration is partially restored in terms of conditions related to deployment tables. If the task *t* is not an active replica then processing will be finished, otherwise a record *(tt, t)* of *ATBAT(a)* will be removed.

If the task *t'* is deployed in the agent *a* (i.e. a record *(t', tt)* exists in *ATBT(a)*) then it will be selected as new active replica. In this case a record *(tt, a, 0)* of *HTBAT(hwp)* remains valid as well as a record *(tt, 0, hwp)* of *HTBAT(rhwp)* of each remote AP $rhwp \neq hwp$. Thus including a record *(tt, t')* in *ATBAT(a)* will restore a consistent configuration.

If the task *t'* is deployed in the agent $a_1 \neq a$ of the same AP *hwp* then it will be selected as new active replica. In accordance with the *h_r_fail(tt)* procedure a record *(tt, a, 0)* will be changed to a record $(tt, a_1, 0)$ in *HTBAT(hwp)* and a record *(tt, t')* will be added to $ATBAT(a_1)$. A record *(tt, 0, hwp)* of *HTBAT(rhwp)* of each remote AP $rhwp \neq hwp$ remains valid. Thus a consistent configuration is restored.

If the task *t'* is deployed in the agent $a_2 \neq a$ of the remote AP $rhwp \neq hwp$ then it will be selected as new active replica. In accordance with *recv_req_atask(tt)* procedure a record $(tt, a_2, 0)$ will be added to *HTBAT(rhwp)* and a record *(tt, t')* will be added to $ATBAT(a_2)$. A record *(tt, 0, hwp)* of $HTBAT(rhwp_i)$ of each remote AP $rhwp_i \neq rhwp$ will be changed to a record *(tt, 0, rhwp)* in accordance with processing of the broadcast *atask_announce(tt, rhwp)* message. Thus a

consistent configuration is restored. Consequently consistent configuration of redundant DHS MAS will be restored independently of deployment of the task *t'*.

In case of a failure of the agent *a* deployed in AP *hwp* each record *(a, t, tt)* will be removed from *HTBD(hwp)* in accordance with the *h_r_afail(a)* procedure and a consistent configuration will be partially restored in terms of conditions related to deployment tables. As was proved for a case of task failure the *h_r_tfail(tt)* procedure will activate new replica of the required type and will restore a consistent configuration in terms of conditions related to active replicas tables. The *h_r_tfail(tt)* procedure will be performed by AP *hwp* for each task type *tt* such than an active replica of this type were deployed in the agent *a*. In accordance with conditions of the theorem all *h_r_tfail(tt)* procedures will succeed. Thus a consistent configuration of DHS MAS will be restored.

Cases of failures of actuators and APs are considered in similar manner and were described in details in [8], [10].

V. Assessment of Assured Fault-Tolerance Level

Our reorganization technique and fault-recovery procedures improve fault-tolerance of DHS MAS. However it is also important to provide an assessment of an achieved level of fault-tolerance. Existing approaches described in [11], [12] enable calculation of probability of survival only in case of failures of hosts of a network on which MAS is deployed. So we have decided to use logical-and-probabilistic methods that enable synthesis of a reliability function of a system in an analytic form [13]. These methods are based on transformation of a logical operability function that determines a state of the system based on states of its components to such form which allows replacement of logical operators with arithmetical operators and logical variables with corresponding probabilities of no-failure [13]. Therefore there is a need for methodology for formation of an operability function of redundant DHS MAS.

As we have stated that redundant DHS MAS is in failure if it is not able to perform an active replica of at least one type we have defined the criteria of serviceability as follows:

$$\forall tt \in TT, \exists t \in RT : failTG(t) = false, \tag{11}$$

where *tt* is a task type, *t* – a task, *RT* – a set of tasks.

Our methodology for formation of an operability function is based on introduced terms of minimal functional configuration (MFC) and minimal operable configuration (MOC). We define MFC as an ordered set <*MT, MA, MHWP*>, where *MT* is a set of tasks, *MA* – a set of agents, *MHWP* – a set of APs. Each MFC shall represent one of the shortest paths of successful operation [13] without considering the necessity of actuators utilization. Thus MFC shall contain a minimal set of tasks required for successful operation of the redundant DHS MAS as well as sets of agents and APs required for execution of these tasks. We define MOC as an ordered set <*MFC, MWR*>, where *MFC* is a minimal functional configuration, *MWR* – a set of actuators of MOC. MOC shall represent the shortest path of successful operation in accordance with logical-and-probabilistic method. It's well known that an operability function may be formed as a disjunction of all conjunction defined by the shortest ways of successful operation [13]. Thus the operability function of the redundant MAS shall be determined in following manner:

$$\bigvee_{\forall MOC}\Big[\Big(\bigwedge_{\forall t\in MT} w(t)\Big)\wedge\Big(\bigwedge_{\forall a\in MA} w(a)\Big)\wedge \Big(\bigwedge_{\forall hwp\in MHWP} w(hwp)\Big)\wedge\Big(\bigwedge_{\forall hwr\in MWR} w(hwr)\Big)\Big]. \tag{12}$$

In (12) *MOC* is a minimal operable configuration, *MT*, *MA*, *MHWP* are respectively sets of tasks, agents and APs of MFC used for construction of MOC, *MWR* is a set of actuators of MOC, *t*, *a*, *hwp*, *hwr* are respectively a task, an agent, AP and an actuator, *w()* is a logical function representing a state of particular component.

Determination of all MOCs is required for formation of an operability function in form (12). The methodology for formation a set of all MOCs was developed. To define the methodology we have introduced a set of additional functions:

- *tasksOfTT(tt)* determines a set of tasks of a particular type:

$$tasksOfTT(tt) = \{t \in RT \mid confTypeTask(tt, t) = true\}, \tag{13}$$

where *tt* is a task type, *t* – a task, *RT* – a set of tasks;

- *rTHwrOfP(MFC, hwp)* determines a set of actuator types required for performing tasks of MFC that are deployed in agents of AP *hwp*:

$$\begin{aligned} &rTHwrOfP(MFC, hwp) = \{thwr \in THWR \mid \\ &\exists t \in MT \in MFC, \exists a \in MA \in MFC : \\ &confTaskAgent(t, a) = true \wedge confAgentHwp(a, hwp) = true \wedge \\ &thwr \in reqTHwr(tOfTask(t))\}, \end{aligned} \tag{14}$$

where *hwp* is AP, *thwr* – ACT type, *THWR* – a set of ACT types, *t* – a task, *MT* – a set of tasks of MFC, *a* – an agent, *MA* – a set of agents of MFC;

- *avTHwrOfP(MFC, hwp, thwr)* determines a set of actuators of a type *thwr* that are accessible for AP *hwp*:

$$\begin{aligned} &avTHwrOfP(MFC, hwp, thwr) = \{hwr \in RHWR \mid \\ &confHwrHwp(hwr, hwp) = true \wedge tOfHwr(hwr) = thwr\}, \end{aligned} \tag{15}$$

where *hwp* is AP, *thwr* – ACT type, *hwr* – ACT, *RHWR* is a set of ACTs.

First of all a set *MTA* of all sets of tasks *MT* that could act as a base of MFC is formed as follows:

$$\begin{aligned} &MTA = \prod_{\forall tt\in TT} tasksOfTT(tt) = \{MT_k = (t_1, ..., t_n) \mid \\ &n = \mid TT \mid, \forall i, \forall j \neq i : tOfTask(t_i), \neq tOfTask(t_j)\}, \end{aligned} \tag{16}$$

where *tt* is a task type, *TT* is a set of task types, t_i is a task.

A set of all MFCs is formed as follows based on a set *MTA* (16):

$$MFCA = \{MFC_k = (MT_k, MA_k, MHWP_k) \mid MT_k \in MTA, MA_k = \{a \in RA \mid \exists t \in MT_k : confTaskAgent(t,a) = true\}, MHWP_k = \{hwp \in RHWP \mid \exists a \in MA_k : confAgentHwp(a,hwp) = true\}\}, \quad (17)$$

where *a* is an agent, *RA* – a set of agents, *t* – a task, *hwp* is AP, *RHWP* – a set of APs.

For each MFC from a set *MFCA* (17) following steps shall be performed to determine a set *MRA* of all sets of ACTs *MR* that could be used for formation of MOC based on this MFC:

- determine whether MFC could be used for formation of MOC in terms of availability of actuators of all required types based on following condition:

$$\forall hwp \in MHWP \in MFC, \forall thwr \in rTHwrOfP(MFC,hwp) : avTHwrOfP(MFC,hwp,thwr) \neq \varnothing, \quad (18)$$

  where *hwp* is AP, *MHWP* is a set of APs of MFC, *thwr* is ACT type;
- for each AP of MFC determine a set of all sets of ACTs *MRH* that include one and only one ACT of each ACT type required for performing of tasks of MFC deployed in agents of this AP:

$$MRHA(MFC,hwp) = \{MRH = \{hwr\}\} = \prod_{\forall thwr \in rTHwrOfP(MFC,hwp)} avTHwrOfP(MFC,hwp,thwr), \quad (19)$$

  where *hwp* is AP, *hwr* is ACT, *thwr* is ACT type;
- determine a set *MRAP* of all sets of ACTs *MR* that could be used to build MOC based on this MFC:

$$MRAP(MFC) = \{MR_i = \bigcup_{\forall MRH_k \in MRU_i} MRH_k \mid \forall MRU_i \in MRAU(MFC)\}, MRAU(MFC) = \prod_{\forall hwp \in MHWP \in MFC} MRHA(MFC,hwp) = \{MRU_i = \{MRH_k\}\}, \quad (20)$$

  where *hwp* is AP, *MHWP* is a set of APs of MFC, *MRHA* is determined in accordance with (19);
- determine a desired set *MRA* by excluding from a set *MRAP* (20) such sets MR_i that do not meet the condition required for treatment of MOC as the shortest way of successful operation:

$$MRA(MFC) = \{MR \in MRAP(MFC) \mid \forall hwr \in MR, \exists hwp \in MHWP \in MFC : [\exists a \in MA, \exists t \in MT : agentOfT(t) = a \wedge confAgentHwp(a,hwp) = true \wedge tOfHwr(hwr) \in reqTHwr(tOfTask(t))] \wedge [\forall hwr' \in MR : (hwr' \neq hwr \wedge confHwrHwp(hwr',hwp) = true) \Rightarrow (tOfHwr(hwr') \neq tOfHwr(hwr))]\}, \quad (21)$$

  where *hwr*, *hwr'* are ACTs, *hwp* is AP, *a* is an agent, *t* is a task, *MHWP*, *MA*, *MT* are respectively sets of APs, agents and tasks of MFC.

A determined set *MOCA* of all MOCs of redundant DHS MAS is as follows:

$$MOCA = \{MOC = (MFC_k, MR_{ki}) \mid \forall MFC_k \in MFCA, \forall MR_{ki} \in MRA(MFC_k), \quad (22)$$

where *MOC* is minimal operable configuration, MFC_k is minimal functional configuration, MR_{ki} is a set of ACTs, *MFCA* is determined in accordance with (17), *MRA* is determined in accordance with (21).

Based on a set of all MOCs (22) the logical operability function of a redundant DHS MAS could be formed in accordance with (12). The formation of the operability function of the redundant DHS MAS in the form (12) enables application of logical-and-probabilistic methods for its transformation and synthesis of the reliability function in an analytic form.

VI. Experiment

The hypothesis to be verified though a set of experiments is as follows: the level of fault-tolerance achieved though utilization of the developed fault-recovery methodology is equal to the theoretical assessment. Experiments are based on statistical modelling of time to failure of all components and further processing of the obtained sequence of failures on the imitation model of test DHS MAS.

The first test redundant DHS MAS ($RMAS_1$) is presented on Fig. 1 and is defined by following sets: a set of task types $TT = \{tt_{19}, tt_{29}, tt_{39}\}$, a set of tasks $RT = \{t_{110}, t_{120}, t_{210}, t_{220}, t_{310}, t_{320}\}$, a set of agents $RA = \{a_{12}, a_{22}, a_{32}, a_{42}\}$, a set of APs $RHWP = \{h_{13}, h_{23}, h_{33}\}$, a set of ACT types $THWR = \{rt_{18}, rt_{28}, rt_{38}\}$ and a set of ACTs $RHWR = \{r_{111}, r_{121}, r_{211}, r_{221}, r_{311}, r_{321}\}$.

The configuration of DHS MAS $RMAS_1$ is as follows:

- *confTypeTask* is true on a set $\{(tt_{19}, t_{110}), (tt_{19}, t_{120}), (tt_{29}, t_{210}), (tt_{29}, t_{220}), (tt_{39}, t_{310}), (tt_{39}, t_{320})\}$;
- *confTypeHwr* is true on a set $\{(rt_{18}, r_{111}), (rt_{18}, r_{121}), (rt_{28}, r_{211}), (rt_{28}, r_{221}), (rt_{38}, r_{311}), (rt_{38}, r_{321})\}$;

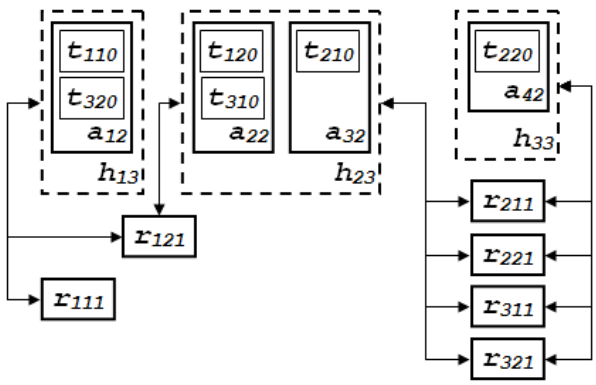


Fig. 1. Redundant DHS MAS configuration ($RMAS_1$)

- *confTaskAgent* is true on a set $\{(t_{110}, a_{12}), (t_{120}, a_{22}), (t_{210}, a_{32}), (t_{220}, a_{42}), (t_{310}, a_{22}), (t_{320}, a_{12})\}$;
- *confAgentHwp* is true on a set $\{(a_{12}, h_{13}), (a_{22}, h_{23}), (a_{32}, h_{23}), (a_{42}, h_{33})\}$;
- *reqTHwrTTask* is true on a set $\{(rt_{18}, tt_{19}), (rt_{18}, tt_{39}), (rt_{28}, tt_{29}), (rt_{38}, tt_{29})\}$;
- *confHwrHwp* is true on a set $\{(r_{111}, h_{13}), (r_{121}, h_{13}), (r_{121}, h_{23}), (r_{211}, h_{23}), (r_{211}, h_{33}), (r_{221}, h_{23}), (r_{221}, h_{33}), (r_{311}, h_{23}), (r_{311}, h_{33}), (r_{321}, h_{23}), (r_{321}, h_{33})\}$.

The second test redundant DHS MAS ($RMAS_2$) is characterized by an increased level of redundancy and its configuration is presented on Fig. 2.

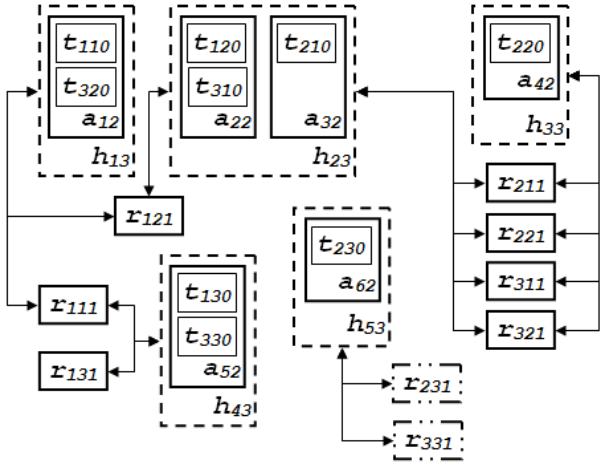


Fig. 2. Redundant DHS MAS configuration ($RMAS_2$)

Experimental assessments of probability of no-failure *PS(t)* as well as analytic reliability functions *PA(t)* for test DHS MASs $RMAS_1$ and $RMAS_2$ are presented on Fig. 3.

VII. Conclusion

A distributed hardware-software multi-agent system has been considered as the object of the research. The developed model of DHS MAS enables determination of new fault-tolerance approaches. The presented reorganization technique is based on replication of tasks that are considered as minimal functional instances and actuators as well as on introduction of redundant sets of agent platforms and agents that are treated as universal executive container for agents and tasks respectively. The developed fault-recovery methodology defines a set of fault-recovery procedures required to deal with failures of individual agents, tasks, agent platforms and actuators.

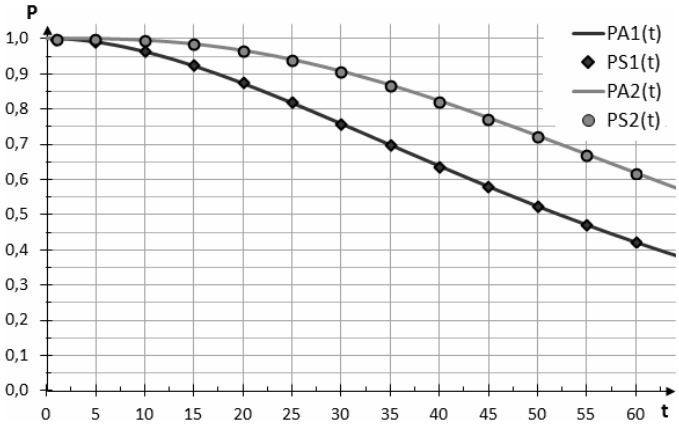


Fig. 3. Reliability functions and experimental assessments of probability of no-failure

The set of conditions required for success of fault-recovery procedures was determined and the theorem on fault-tolerance property of redundant DHS MAS was stated and proved to validate new fault-recovery methodology.

The developed methodology for construction of a logical operability function of a DHS MAS enables utilization of logical-and-probabilistic methods for theoretical assessment of an assured level of fault-tolerance during a design phase. In accordance with performed experiments the level of fault-tolerance achieved by utilization of developed fault-recovery procedures is equal to the theoretical assessment.

References

[1] L. C. Lee, H. S. Nwana, D. T. Ndumu, P. De Wilde, "The stability, scalability and performance of multi-agent systems," *BT Technology Journal*, vol. 16(3), pp. 94–103, July 1998.

[2] H. F. Ahmad, A. Ali, Z. A. Khan, S. Shahid, H. Suguri, "Decentralized architecture for fault tolerant multi agent system," in *Proceedings of the ISADS'2005 Autonomous Decentralized Systems*, pp. 167–174, 2005.

[3] R. Deters, A. Fedoruk, "Improving fault-tolerance by replicating agents," in *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, pp. 737–744, 2002.

[4] M. Bertier, O. Marin, P. Sens, "DARX – a framework for the fault-tolerant support of agent software," in *ISSRE'03 Proceedings of the 14th International Symposium on Software Reliability Engineering*, pp. 406–416, 2003.

[5] P. R. Cohen, S. Kumar, H. J. Levesque, "The adaptive agent architecture: achieving fault-tolerance using persistent broker teams," in *Proceedings of Fourth International Conference on MultiAgent Systems*, pp. 159–166, 2000.

[6] S. Haegg, "A sentinel approach to fault handling in multi-agent systems," *Multi-Agent Systems Methodologies and Applications*, pp. 181–195, 1997.

[7] S. Mellouli, "A reorganization strategy to build fault-tolerant multi-agent systems," *Advances in Artificial Intelligence : Lecture Notes in Computer Science*, vol. 4509, pp. 61–72, 2007.

[8] A. V. Igumnov, S. E. Saradgishvili, "Fault recovery in redundant multiagent systems," *St. Petersburg State Polytechnical University Journal. Computer Science. Telecommunication and Control Systems*, vol. 193(2), 2014, pp. 99–109.

[9] A. V. Igumnov, S. E. Saradgishvili, "Reliability assessment for redundant multi-agent systems," *Electronic Journal "Science and Education: Electronic Scientific and Technical Periodical"*, vol. 1, 2014. Available: http://dx.doi.org/10.7463/0114.0696290).

[10] A. V. Igumnov, "Fault-tolerance in redundant distributed hardware-software multi-agent systems," in *Proceedings of COMOD-2014 International Conference on Computer Modeling and Simulation*, pp. 155–161, 2014.

[11] S. Kraus, V. S. Subrahmanian, N. C. Tas, "Probabilistically survivable mass," in *Proceedings of IJCAI'2003 International Joint Conference on Artificial Intelligence*, vol. 3, pp. 789–795, 2003.

[12] S. Kraus, E. Neisterski, V. S. Subrahmanian, D. Peleg, Y. Zhang, "Computing the fault tolerance of multi-agent deployment," *Artificial Intelligence*, vol. 173(3), pp. 437–465, 2009.

[13] I. A. Ryabinin, G. N. Cherkesov, *Logiko-veroyatnostnye metody issledovaniya nadezhnosti strukturno-slozhnykh system [The logic-probabilistic research methods of structure-complex systems reliability]*. Moscow: Radio i svyaz' Publ., 1981. 264 p. (in Russian).

General Theory for Reproducible Data Processing: Apparatus Function and Reduction to an "Ideal" Experiment

R.R. Nigmatullin[a] , D. Striccoli[b] and W. Zhang[c]

***Abstract*—** The authors suggest a general theory for consideration of *all* experiments associated with measurements of reproducible data in one unified scheme. The suggested algorithm does *not* contain unjustified suppositions and the final function that is extracted from these measurements can be compared with hypotheses that are suggested by the theory adopted for the explanation of the object studied. This true function is free from the influence of the apparatus (instrumental) function (AF) and when the "best fit", or the most acceptable hypothesis, is *absent*, can be presented by a segment of the Fourier series. This segment is used as the *fitting function* and contains the number of the fitting parameters ($2K$+2) that are much less in comparison with initial data points N (N>>$2K$+2) The discrete set of the decomposition coefficients describes the final function *quantitatively* and can serve as an intermediate model (IM), that coincides with the conventional definition of the amplitude-frequency response (AFR) of the object studied. It can be used by theoreticians also for comparison of the suggested theory with experimental observations.

***Keywords*—** Ideal experiment, Reproducible measurements, Prony and Fourier decompositions.

INTRODUCTION

The measurements and different data processing form the foundation stone of all natural science and any attempt to push this stone over the hump seems useless. Many excellent books, reviews, pile of papers written by outstanding mathematicians, statisticians, experimentalists and theoreticians form a stable trend in the region of science as the data processing. All these well-known publications represent the great effort in the last decades in dealing with the data processing, fitting and forecasting in several application fields, and through different approaches. The question that we want to formulate in this paper will sound paradoxical for many researches: *Is it possible to create general and the unified theory for all reproducible data processing?* In this short paper we want to demonstrate in brief the basic ideas and justify the positive answer for this unexpected question. This theory can lead to *reconsideration* of the conventional point of view on treatment of reproducible data and create a new trend in the science as the data/signal processing.

[a]*Theoretical Physics Department, Institute of Physics, Kazan Federal university Kremlevskaya str. 18, Kazan, Tatarstan, Russian Federation (e-mail: renigmat@gmail.com)*
[b] *Department of Electrical and Information Engineering (DEI) Politecnico di Bari,, Via E. Orabona, 4, Bari, Italy (e-mail: domystric@gmail.com)*
[c]*Jinan University, College of Information Science and Technology, Department of Electronic Engineering, 510632, Shi-Pai, Guangzhou, Guangdong, China (e-mail: niuwei377@qq.com)*

I. DESCRIPTION OF THE TREATMENT PROCEDURE

Let us remind the definition of an *ideal* experiment that will be used in this paper. Let us suppose an object under study, a deterministic (or control) variable x that interacts with the object, and forms a desired response Pr(x). In an ideal reproducible experiment, this response is reproduced by each current measurement m:

$$y_m(x) \cong \Pr(x + m \cdot T_x) = \Pr(x + (m-1) \cdot T_x), \quad m = 1, 2, ..., M. \tag{1}$$

Here x – is the external (control) variable, T_x is a "period" of experiment expressed in terms of the control variable x. In expression (1) we make only one *supposition* that the properties of the object studied during the period of "time" T_x is *not* changed. As one can notice from (1) each current measurement in ideal experiment is *independent* from the previous responses and in this sense it does not have a memory (strong correlations between neighboring measurements). If $x = t$ coincides with temporal variable then $T_x = T$ coincides with the conventional definition of a period. The solution of this functional equation is well-known and (in case of discrete distribution of the given data points $x = x_j$; j=1, 2,…, N) coincides with the segment of the Fourier series

$$Pr(x) = A_0 + \sum_{k=1}^{K>>1} \left[Ac_k \cos\left(2\pi k \frac{x}{T_x}\right) + As_k \sin\left(2\pi k \frac{x}{T_x}\right) \right] \tag{2}$$

We *deliberately* show only the segment of the Fourier series used in this approach *as a fitting function* because in reality all

data points are always *discrete* and the number of "modes" k = 1,2,…, K (coinciding with the coefficients of the Fourier decomposition) is limited. We define here and below by the capital letter K the *finite* number of modes. The value of K can be calculated by supposing that the relative error is located in the interval [1%-10%]. This interval provides the desired fit of the measured function $y(x)$ to $Pr(x)$ with initially chosen number of modes k = 1,2,…, K figuring in (2). From relationships (1), (2) one important conclusion follows. For *ideally* reproducible experiment, which satisfies to condition (1) the segment of the F-transform (2) can be used as *intermediate model* (IM) and the number $2K+2$ of decomposition coefficients (A_0, Ac_k, As_k) (we should calculate the unknown value of T_x as additional nonlinear fitting parameter also) can be used as a set of the fitting parameters belonging to the IM. The meaning of these coefficients is well-known and actually this set defines the amplitude-"frequency" response (AFR) associated with the recorded "signal" $y(x) \approx Pr(x)$ coinciding with the measured function. Here we increase only the limits of interpretation of the conventional F-transform with respect to *any* deterministic/control variable x (including frequency also, if the control variable x coincides with some current ω) and show that the segment of this transformation can be used as a fitting function for quantitative description of an *ideal* experiment.

But, as it has been shown in papers [1-3] the real measurements are the strongly-correlated and have a memory. In this case they are described in terms of the Prony's decomposition that follows as the general solution from the functional equation

$$F(x+LT_x)=\sum_{l=0}^{L-1}a_l F(x+lT_x)+b \qquad (3)$$

In reality, as we will see below, one can calculate easily the set of the parameters $\{a_l, b\}$ by the linear least-square method (LLSM) if we *suppose* that $L = M$, where M coincides with the last measurement. But, up to now, we do *not* know how to calculate the true value of L. This true value of L is associated probably with deep physical reasons and finding of the value L merits a special research. The functional equation (11) describes mathematically a wide class of the QP processes and can be interpreted as follows. The measurement process that takes place during the interval $[(L\text{-}1)T_x, LT_x]$ partly depends on the measurements that have been happened on the previous temporal intervals $[lT_x, (l+1)T_x]$ with l = 0,1,…, L-1. The set of the constants $[a_l]$ (l = 0,1,…,L-1) can be *quantitatively* interpreted as the influence of a memory (strong correlations) between the successive measurements.

But many tests realized with available data [1-3] show that high-frequency fluctuations destroy a memory and in the case of their small influence the value of L accepts the values one or two only (L=1,2). The functional equation (3) describes mathematically a wide class of the quasi-periodic (QP) processes and the general solution of this equation is written in the form

$$(A)\ \sum_{l=0}^{L-1}a_l\neq 1:\ F(x)=\sum_{l=1}^{L}\left(\kappa_l\right)^{x/T_x}\mathrm{Pr}_l(x)+c_0,$$
$$c_0=\frac{b}{1-\sum_{l=0}^{L-1}a_l},$$
$$(B)\ \sum_{l=0}^{L-1}a_l=1:\ F(x)=\sum_{l=1}^{L}\left(\kappa_l\right)^{x/T_x}\mathrm{Pr}_l(x)+c_1\frac{x}{T_x}, \qquad (4)$$
$$c_1=\frac{b}{L-\sum_{l=0}^{L-1}l\cdot a_l}.$$

Here the functions $\mathrm{Pr}_l(x)$ define a set of periodic functions (l = 1, 2,…, L) from expression (3), the values κ_l coincide with the roots of the characteristic polynomial

$$P(\kappa)=\kappa^L-\sum_{l=0}^{L-1}a_l\kappa^l=0. \qquad (5)$$

In general, these roots can be positive, negative, g-fold degenerated (with the value of the degeneracy g) and complex-conjugated. This peculiarity distinguishes this decomposition from the conventional Prony decomposition considered in papers [4-5]. We should note also that for the case B in (4) one of the roots κ_l coincides with the unit value (κ_1=1) that leads to the pure periodic solution. As before, the finite set of the unknown periodic functions $\mathrm{Pr}_l(x,T_x)$ (l=1,2,…,L) is determined by expression (2). Now one can formulate the basic question: if we accept the verified and justified hypothesis (3) then is it become possible to eliminate the roots κ_l from solutions (4) and reduce the real measurements to the situation of ideal experiment (1) without memory? The positive answer on this key question is contained in expression below. For this case we have the following system of linear equations

$$F(x)=\sum_{l=1}^{L}EP_l(x)+c_0,$$
$$F\left(x+T\right)=\sum_{l=1}^{L}\kappa_l EP_l(x)+c_0,$$
$$\cdot \qquad (6)$$
$$F\left(x+\left(L-1\right)T\right)=\sum_{l=1}^{L}\kappa_l^{L-1}EP_l(x)+c_0,$$
$$where\ \ EP_l(x)=\left(\kappa_l\right)^{x/T_x}\mathrm{Pr}_l(x),\ l=0,1,...,L-1,$$
$$or\ \mathrm{Pr}_l(x)=EP_l(x)\cdot\left(\kappa_l\right)^{-x/T_x}$$

From this linear system one can find easily the unknown functions $EP_l(x)$ and then restore the unknown set of periodic functions $\mathrm{Pr}_l(x)$. It means that becomes *possible* to realize the reduction of a wide class of reproducible measurements presented initially in the frame of the desired IM and corresponding to the Prony's decomposition to an *ideal* experiment where only one periodic function figures. We note that the L-th order determinant of system (6) coincides with well-known Vandermonde determinant. It does *not* equal to

zero if all roots of equation (5) are *different*. So, finally we obtain only one *ideal* periodic function that corresponds to the reduction of the real set of measurements to an *ideal* (perfect) experiment

$$Pf(x) = \sum_{l=0}^{L-1} \mathrm{Pr}_l(x) \qquad (7)$$

The transition from real experiment (6) to an ideal one (7) eliminates the influence of the apparatus (instrumental) function and helps to reduce the IM that is expressed by the AFR of the function (7). The decomposition coefficients (A_0, T_x, Ac_k, As_k, k=1,2,…,K) entering into expression (7) of the last expression describe the total reproducible experiment studied *quantitatively*.

Unfortunately, the lack of space does not allow us to include the detailed description of original experiment related to quantitative analysis of exchange by EM packets between two wireless sensor nodes. This accurate experiment was successfully realized by one of the coauthors (D. Striccoli in DEI, Bari, Italy). We have also other examples based on available data and related to reproducible experiments performed in analytical chemistry and EPR (electronic paramagnetic resonance). All of them can be described in the frame of this general theory suggested above.

In the end of this short paper we should mark that the acceptable algorithm used for description of real data was described preliminary in paper [6] and some original results that are associated with consideration of other experiments have been submitted recently for publication.

II. RESULTS AND REMAINING PROBLEMS

To conclude this final section it is necessary to formulate a couple of problems that can merit an interest for the further research:

1. The memory problem that is appeared between neighboring measurements is *not* solved. Based on analysis and processing of many available data we proved only that the high-frequency fluctuations destroy a memory but the deep physical reasons that create this phenomenon are *not* known. In other words, in spite of the fact that long memory during the period of "time" T_x between all successive measurements exits ($L = M$) the reasons of appearance of a *partial* memory when $L < M$ are *not* clear. In this and other papers [6] we show only *how* to reduce approximately this true memory to three basic mean functions *ymn*(x), *yup*(x), *ydn*(x) that divide all measurements on three clusters covering the measurements close to ideal group (*ymn*(x)) and "marginal" measurements having the slopes exceeding the unit value (*yup*(x)) and less the unit value (*ydn*(x)), correspondingly. These three functions help to solve a problem of elimination of the apparatus function (AF). The explanation of this general phenomenon will be interesting for many researches working in different branches of natural science.

2. We found the key point that *conciliates* theory and experiment. All competing hypothesis should be presented in the form of the segment of the F-transform and compared with the function (7) that is obtained from reproducible measurements. This specific check-point can be sometimes crucial for experimentalists and theoreticians trying to understand the natural phenomenon studied from two opposite sides. But the justified "logic" of the paper and [6] prompts that the coincidence of the arguments from both sides should be focused on the attentive analysis of expression (7). The illustrative examples taken from the available experiments lead to the same conclusions. So, one can formulate a problem of creation of the unified metrological standard (UMS) that should be accepted by many experimentalists in order to supply their reliable data to theoreticians that want to understand the phenomenon studied from the opposite side.

ACKNOWLEDGMENTS

This basic idea of this paper is stimulated by the R&D project realized in the frame of the Jinan University-Kazan Federal University Joint Laboratory of "Information Science and Fractal Signal Processing".

REFERENCES

[1] R.R. Nigmatullin, A.A. Khamzin and J. T. Machado, Detection of quasi-periodic processes in complex systems: how do we quantitatively describe their properties? *Physica Scripta* **89** 015201,2014.

[2] R. R. Nigmatullin, S. I. Osokin, D. Baleanu, S. Al-Amri, A. Azam, A. Memic, The First Observation of Memory Effects in the InfraRed (FT-IR) Measurements: Do Successive Measurements Remember Each Other? *PLoS ONE*, Open access journal, April 9 (4) e94305,2014 .

[3] R. Nigmatullin, R. Rakhmatullin. Detection of quasi-periodic processes in repeated measurements: New approach for the fitting and clusterization of different data. *Journal of the CNSNS* **19** 4080-4093, 2014.

[4] Kahn M, Mackisack M, Osborne M, Smyth G. On the consistency of Prony's method and related algorithms. *Journal of Computational and Graphical Statistics* 1: 329–349,1992.

[5] Osborne M, Smyth GK (1995) A modified Prony algorithm for exponential function fitting. *Journal of Scientic and Statistical Computing* 16, 119–138, 1995.

[6]- R.R. Nigmatullin, R.M. Rakhmatullin, S.I. Osokin, How to reduce reproducible measurements to an ideal experiment? Magnetic Resonance in Solids, Electronic Journal, 16 (2) 1-19., 2014). http://mrsej.kpfu.ru.

Eddy Currents Computation by an Integral Equation Method

A. Kalimov, S. Shimansky
St. Petersburg State Polytechnic University,
Polytechnicheskaya 29, St. Petersburg, 195251, RUSSIA.

***Abstract*—** A formulation of the integral method for calculating time dependent electromagnetic fields and the eddy currents in conducting objects is considered in this paper. The current density vector is approximated by facet vector functions. Independent variables are associated with the co-tree branches of the graph built at the basis of triangular or tetrahedral mesh covering the space filled with conducting material. A system of algebraic equations for the unknown current density values is formed for the closed loops corresponding to the co-tree branches. The proposed method has been verified by solving a test problem for which an analytical solution is available.

***Keywords*—** magnetic field, eddy currents, integral equation, finite element method, facet elements.

I. Introduction

INTEGRAL methods are used successfully for modeling time dependent electromagnetic fields in conducting objects. Some formulations exploit different electromagnetic potentials or their combinations [1] – [3]. The current density vector together with the scalar electric potential are used to calculate eddy currents in conducting objects in frame of the Partial Element Equivalent Circuit formulation, originally proposed in [4]. This approach is implemented to the eddy currents modeling in 2D systems [5] – [7]. The current density vector is also used as a main variable in integral formulations for modeling surface currents in good conductors [8] – [9].

In this paper we propose a general formulation of such problem with the current density vector being the only main variable. Let us consider a system of equations governing the electromagnetic field inside conducting media:

$$\begin{aligned} rot\vec{H} &= \vec{J}, \\ div\vec{B} &= 0, \\ rot\vec{E} &= -\frac{d\vec{B}}{dt}. \end{aligned} \qquad (1)$$

A. Kalimov is with the Saint-Petersburg State Polytechnic University, 195251, Saint Petersburg, RUSSIA (phone: +7-950-045-6060; e-mail: alexanderkalimov@gmail.com).
S. Shimansky is with CADFEM-CIS – Saint-Petersburg, 195197, RUSSIA (e-mail: s.a.shimanskiy@gmail.com).

Here we suppose that the field characteristics change in time slowly enough, and the displacement currents may be neglected. To complete formulation of the problem it is necessary to add equation for the current density continuity

$$div\vec{J} = 0, \qquad (2)$$

and the standard constitutive relations for the conducting non-magnetic media:

$$\vec{J} = \gamma\vec{E},$$

$$\vec{B} = \mu_0\vec{H}.$$

The field intensity at any point of the considered space may be expressed as a superposition of the external field $\vec{H}_m$ and the field $\vec{H}_c$ induced by the eddy currents circulating in the conducting object:

$$\vec{H} = \vec{H}_c + \vec{H}_m$$

Both components of the magnetic field may be calculated using the Biot-Savart law:

$$\vec{H}_c(\vec{r}) = \frac{1}{4\pi}\int \frac{\vec{J}(\vec{r}')\times(\vec{r}-\vec{r}')}{|\vec{r}-\vec{r}'|^3}dV' \qquad (3)$$

Such presentation of the field intensity automatically fulfills the first and the second equations of the system (1) for the currents satisfying (2). Combining (3) and (1) we can derive an integro-differential equation:

$$rot\vec{J} = -\gamma\mu_0\frac{d}{dt}\left[\frac{1}{4\pi}\int\frac{\vec{J}(\vec{r}')\times(\vec{r}-\vec{r}')}{|\vec{r}-\vec{r}'|^3}dV' + \vec{H}_m\right]. \qquad (4)$$

Together with the relation (2) and appropriate boundary conditions for the normal component of the current density vector on the conductor surface the last equation defines a unique solution for the current density vector. This form of the integro-differential equation for the current density vector does not require introducing the scalar electric potential.

II. Approximation of the Integro-Differential Equation

A. A Choice of Independent Variables.

To solve numerically the equation (4) we split space filled with the conductor into a set of triangular or tetrahedral

elements depending on the space dimension. The unknown current density distribution is approximated by a superposition of the vector finite functions associated with the facets of the generated mesh:

$$\vec{J}(r) = \sum_{i=1}^{I} J_i \vec{\varphi}_i(\vec{r}) \qquad (5)$$

with $\vec{\varphi}(\vec{r})$ being the facet functions and J_i the components of the current density vector normal to the facets. Such choice of the current density approximation provides the continuity of the normal components of this vector by default and gives a simple way to set appropriate boundary conditions.

Evidently some of the values J_i are not independent because they should satisfy restriction (2). So a proper choice of the independent variables and corresponding functions should be undertaken. For this purpose we build a graph corresponding to the tetrahedral (or triangular in the case of a 2D problem) mesh. The nodes of the graph are associated with the central points of the elements. Each branch of the graph connects the centers of neighboring elements and so crosses one facet. Example of such graph corresponding to 2-dimensional object is shown in Fig. 1.

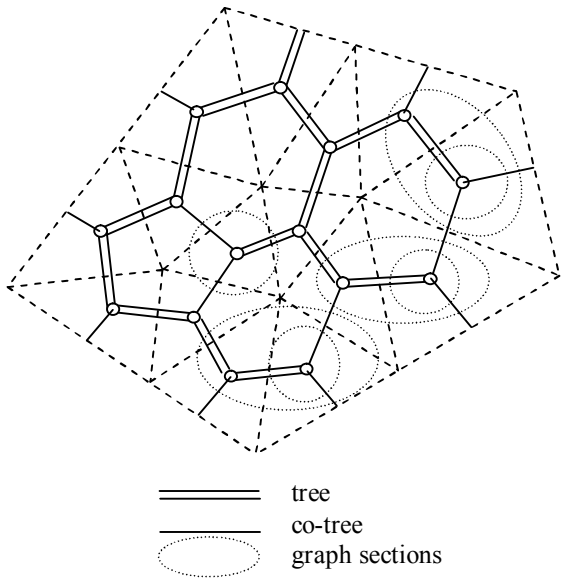


Fig.1. A graph corresponding to the triangular mesh.

The full graph of the mesh is separated into the main tree and the co-tree. Every section of the graph (closed line crossing only one tree branch) forms a closed surface (or loop in 2D) and so may be used to set a constitutive relation, equivalent to the equation (2):

$$\sum_{i=1}^{I} w_i s_i J_i = 0,$$

s_i is the area of the corresponding facet (length of the edge in 2D), $w_i = \pm 1$ is a multiplier depending on the direction of the current density with respect to the normal to the facet vector. Consequently only the variables corresponding to the co-tree branches may be regarded as independent.

Similar approach to this problem for the case of thin conducting plates when the surface charges and scalar electric potentials induced by the eddy currents may be neglected was described in [10].

B. Forming a System of Algebraic Equations.

In the general case the algebraic equations for the unknown current density values may be formed for the closed loops consisting of the graph branches. Such loop should include only one co-tree branch, while all others should belong to the main tree. Integration of the equation (4) along such closed lines gives:

$$\oint_l \vec{J}(\vec{r}) \cdot d\vec{l} + \frac{\mu\gamma}{4\pi} \frac{d}{dt} \oint_l d\vec{l} \int \frac{\vec{J}(r')}{|\vec{r} - \vec{r}'|} dV' = -\frac{\mu\gamma}{\mu_0} \oint_l \frac{d\vec{A}(\vec{r})}{dt} d\vec{l}, \qquad (6)$$

where $\vec{A}(\vec{r})$ is the magnetic vector potential induced by external sources. To derive the last equation we transformed integrals over the area to the integrals over the bounding closed lines using the Stokes' theorem. Evidently the number of the equations is equal to the number of unknowns and all these equations are linearly independent. Presentation of the current density vector $\vec{J}(\vec{r})$ in a form of (5) gives a system of algebraic equations with the unknown values J_i.

For the first order facet functions used for approximation the current density vector is constant inside each element because of additional restriction (2):

$$\vec{J}(r) = \sum_{i=1}^{K} J_i \vec{\varphi}_i(\vec{r}) = const \qquad (7)$$

For such approximation of the unknown function we used analytical expressions for the vector potential induced by the eddy currents flowing in the tetrahedral elements [11] - [12]:

$$\vec{A}(\vec{r}) = \int \frac{\vec{J}(r')}{|\vec{r} - \vec{r}'|} dV'.$$

C. Choice of the Central Points of the Elements.

Transformation of the integral equation (6) to the system of algebraic equations depends on the position of the central point of the elements. The most convenient position of this point was found to be a center of the circumscribed sphere (circle in the 2D case). Such choice provides relatively simple presentation for the integrals in (6).

Let us consider a fragment of a loop connecting a node inside an element with the center of the *n*-th facet (Fig.2).

ISBN: 978-1-61804-251-4

An integral from the current density vector along such line may be presented, taking into account (5), as:

$$\int_O^{C_n} \vec{J}(\vec{r})d\vec{l} = \sum_{k=1}^{K} J_k \cdot \int_O^{C_n} \vec{\psi}_k(\vec{r})d\vec{l} \qquad (8)$$

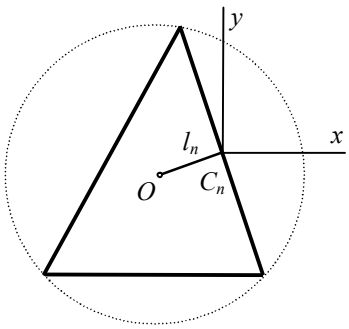


Fig.2. Position of the central node inside an element.

A total number of the facet functions inside the element $K = 4$ for the 3D problem. For the first order approximation of unknowns the facet function $\vec{\psi}_k(\vec{r})$ may by expressed directly by a polynomial:

$$\vec{\psi}_k(x,y) = \vec{\psi}_{k0} + \vec{\psi}_{kx} \cdot x + \vec{\psi}_{ky} \cdot y + \vec{\psi}_{kz} \cdot z . \qquad (9)$$

where $\vec{\psi}_{k0}$, $\vec{\psi}_{kx}$, $\vec{\psi}_{ky}$, $\vec{\psi}_{kz}$ are the vector constants. Taking into account a property of the current density distribution inside each element (7) we can conclude that only the first terms in (9) give a real contribution in the expression for the field integral (8):

$$\int_O^{C_n} \vec{J}(\vec{r})d\vec{l} = \sum_{k=1}^{K} J_k \cdot \int_O^{C_n} \vec{\psi}_{k0}d\vec{l} .$$

Evidently the last relation does not depend on the choice of the origin and orientation of the Cartesian coordinate system. It is convenient to choose such system with the origin in the center of the *n*-th facet (see Fig. 2). Normal to this facet components of all facet functions except $\vec{\psi}_n(\vec{r})$ are equal to zero at this point, while the same component of the $\vec{\psi}_n(\vec{r})$ has the unit value. That is why the scalar productions $\vec{\psi}_k \cdot d\vec{l}$ are equal to:

$$\vec{\psi}_k d\vec{l} = \vec{\psi}_{k0} d\vec{l} = 0, \quad k \neq n$$

$$\vec{\psi}_n d\vec{l} = \vec{\psi}_{n0} d\vec{l} = dl$$

Taking into account these relations we can reduce the integral (8) to a simple production:

$$\int_O^{C_n} \vec{J}(\vec{r})d\vec{l} = J_n \cdot l_n .$$

Similar expressions may be derived for the second term in the left part of the integral equation (6).

It is worthy to note that the considered choice of the central point is valid even in the case when this point lies outside the element (but inside the conducting object).

III. Test Problem

To verify the proposed algorithm we considered a long hollow conducting cylinder in the uniform external magnetic field with the harmonic time dependence:

$$B(t) = B_m \cdot \sin(\omega t)$$

To investigate this problem we applied a standard procedure:

- introduced complex amplitudes for the main time dependent variables $\vec{\mathbf{J}}(\vec{r})$ and $\vec{\mathbf{A}}(\vec{r})$;
- transformed the integral equation (6) to the corresponding complex form:

$$\oint_l \vec{\mathbf{J}}(\vec{r}) \cdot d\vec{l} + \frac{k^2}{4\pi} \oint_l d\vec{l} \int \frac{\vec{\mathbf{J}}(r')}{|\vec{r} - \vec{r}'|} dV' = -\frac{k^2}{\mu_0} \oint_l \vec{\mathbf{A}}(\vec{r})d\vec{l} .$$

The factor k^2 here is defined by the relation

$$k^2 = j\omega\mu_0\gamma ,$$

where j is the imaginary unit, ω is the angular frequency and γ is the material conductivity.

The direction of the field intensity vector is parallel to the cylinder axis. The inner and outer radii are equal to 0.5 and 1.0 m. This problem was chosen for analysis because the corresponding analytical solution may be easily derived. Typical triangular mesh after the discretization of the problem space is shown in Fig. 3.

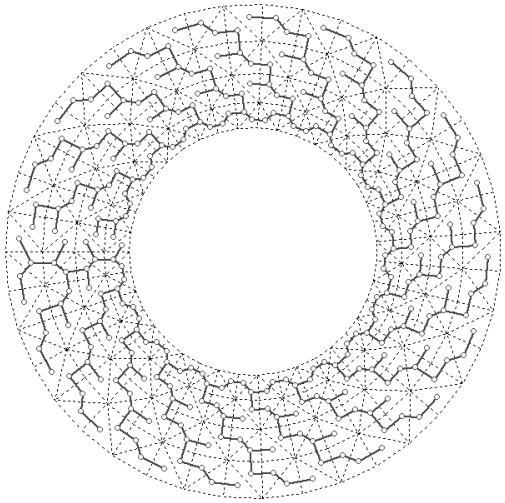


Fig.3. Typical discretization of the problem space and a configuration of the graph tree.

ISBN: 978-1-61804-251-4

The main series of calculations was performed for the mesh with the total number of the nodes $N = 2511$ and the elements $I = 4800$ which corresponds to 30 radial and 80 axial layers in the discretized problem space. The total number of unknowns in such a case was equal to $K = 2429$.

The current density and field intensity distributions depend strongly on the penetration depth of the alternating electromagnetic field into the conducting material

$$\delta = \sqrt{\frac{2}{\omega \mu_0 \gamma}}.$$

For this parameter we chose a value of δ = 0.1 m. Such choice corresponds to the strong radial dependence of the current density (Fig. 4).

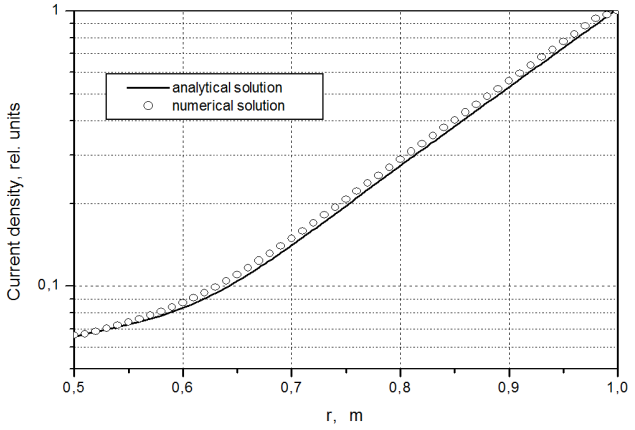


Fig.4. Dependence of the current density absolute value on the radius for δ = 0.1 m.

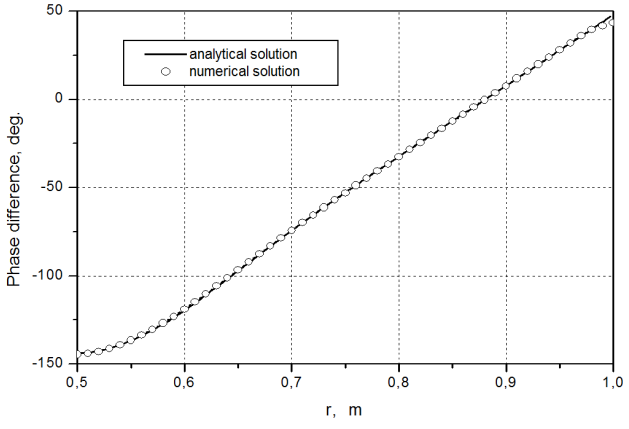


Fig.5. Dependence of the current density phase on the radius for δ = 0.1 m

The results of the eddy current computation (the amplitude and the phase) are shown in Fig.4 – Fig.5. These distributions demonstrate a reasonably good agreement between the numerically and analytically derived data.

IV. Conclusions

In this paper we describe a universal integral formulation for the eddy current computation. The proposed method may be applied for both 2D and 3D problems and does not require any gauging conditions for the main variable and gives a simple and natural way of involving boundary conditions. It is important that the developed algorithm is valid for the simply connected and multiply connected problem spaces without any modifications.

Approximation of the unknown current density vector is done on the basis of the facet vector functions. A proper choice of the independent unknowns provides a solution of the problem which satisfies one of the Maxwell equations (2) by default. For this purpose we build a graph with the nodes in the centers of the elements and branches crossing the element facets. A graph co-tree defines a set of independent variables. The algebraic equations approximating the integral equation (6) are formed for the closed loops corresponding to the co-tree branches. The proposed algorithm is verified by calculating eddy currents in a hollow conducting cylinder immersed in the uniform alternating magnetic field.

References

[1] S.J.Salon, B.Mathewson, S.Uda "An Integro-Differential Approach to Eddy Currents in Thin Plates," *IEEE Trans. Magn.,* vol. MAG-19, pp. 2405 – 2408, Nov. 1983.

[2] D. Zheng, K.R. Davey, "A Boundary Element Formulation for Thin Shell Problems," *IEEE Trans. Magn.,* vol. 32, pp. 675 – 677, May 1996.

[3] R. Albanese, G. Rubinacci "Integral formulation for 3D eddy-current computation using edge elements," *IEE Proceedings,* vol. 135, Pt. A, No 7, pp. 457– 462, Sep. 1988.

[4] A. Ruehli, "Equivalent circuit models for three-dimensional multiconductor systems," *IEEE Trans. Microw. Theory Tech*., vol. MTT-22, pp. 216-221, Feb. 1974.

[5] F. Freschi, G. Gruosso, M. Repetto, "Unstructured PEEC Formulation by Dual Discretization," *IEEE Microwave and Components Letters*, vol. 16, pp. 531-533, Oct. 2006.

[6] F. Freschi and M. Repetto, "A General Framework for Mixed Structured/Unstructured PEEC Modeling," *ACES Journal*, Vol. 23, pp. 200-206, Sep. 2008.

[7] P. Alotto, F. Desideri, F. Freschi, A. Maschio, M. Repetto, "Dual-PEEC modeling of a two-port TEM cell for VHF applications," *IEEE Trans. Magn.*, vol.47, pp.1486-1489, May 2011.

[8] G. Miano and F. Villone, "A surface integral formulation of Maxwell equations for topologically complex conducting domains*," IEEE Trans. Antennas and Propag*., vol. 53 pp. 4001– 4014, Dec. 2005.

[9] J-F. Lee, R. Lee, R. J. Burkholder, "Loop Star Basis Functions and a Robust Preconditioner for EFIE Scattering Problems," *IEEE Trans. Antennas and Propag.*, vol. 11, pp.1855–1863, Aug. 2003.

[10] H. Tsuboi, T.Asahara, F.Kobayashi, T.Misaki "Eddy Current Analysis on Thin Conducting Plate by an Integral Equation Method Using Edge Elements," *IEEE Trans. Magn.,* vol. 33, pp. 1346 – 1349, Mar. 1998.

[11] D. R. Wilton, S. M. Rao, A. W. Glisson, D. H. Schaubert, O. M. Al-Bundak, and C. M. Butler, "Potential Integrals for Uniform and Linear Source Distributions on Polygonal and Polyhedral Domains," *IEEE Trans. Antennas and Propag.*, vol. AP-32, pp. 276–281, Mar. 1984.

[12] R. D. Graglia, "On the Numerical Integration of the Linear Shape Functions Times the 3-D Green's Function or its Gradient on a Plane Triangle," *IEEE Trans. Antennas and Propag.*, vol. 41, pp. 1448–1455, Oct. 1993.

A Soft Clustering Approached with Feature Reduction using Principal Component Analysis

Phichete Julrode

***Abstract*—**This paper, investigates on using feature reduction based method, principal component analysis (PCA) with soft clustering approaches for medical diagnosis applications.The problems of high dimensional, noisy data and hidden outliers, which usually occur in the field of medical diagnosis, can seriously spoil the computation of several of types of learning, including medicalrelated clustering. The two particular soft clustering approaches, fuzzy c-means (FCM) and fuzzy ant based clustering (FAC) are applied in this paper as their soft clustering feature, support the increase of sensitivity for medical diagnosis. PCA is employed as preprocess of FCM and FAC for relieving the curse of high-dimensional, noisy data. Comparison tests among related methods, PCA-FCM, PCA-FAC with FCM and FAC alone are evaluated in terms of clustering objective function and adjusted rand index. In addition, dimension reduction, resulted from PCA is determined. Five significant medical data sets are employed in the experiments. Within the scope of this study, the superiority as well as the importance of using PCA as preprocessing of the efficient traditional FCM and FAC are pointed out.

***Keywords*—**Feature reduction, Principal component analysis, Fuzzy c-means, Fuzzy ant based clustering.

I. Introduction

Early detection of medical diagnosis problems such as breast cancer and diabetes, etc. is important to increase the chance of successful treatment. Clustering is a popular data analysis method and plays animportant role in data mining. It is used in several medical applications, for instance grouping related information from a patient's health history, physical examination, and laboratory results as part of the process of making a diagnosis.Soft clustering method, fuzzy c-means (FCM)algorithm [1], [2] has been used extensively. It yields the degree of membership value in each cluster; thus allows one piece of data to belong to two or more clusters. The other soft clustering that would be mentioned here refers to fuzzy ant based clustering (FAC) algorithm [3], [4]. Instead of using an arithmetic mean, fuzzy ant based clustering uses the arithmetic mean with euclidean distanceaverage. Moreover, FAC associates the influence weight of a multi-data on the cluster center in the following iterations [5]. FCM as well as FAC retain more information from the original data than those of crisp or hard. The feature, related to the degree of membership provides the increase of sensitivity for medical diagnosis.

In several real-world especially medical applications, therehave usually been the severe problems of high dimensional, noisy data and outliers.Such problems seriously spoil the computation of several types of learning, including clustering. In addition, irrelevant dimensional features deteriorate the generalizing performance of clustering. A linear feature reduction method, principal component analysis (PCA) is one of the important tools for coping with such dimensionality problems [6], [7]. In order to perform dimension reduction or features reduction, PCA maps the original predicting features into smaller numbers of features.

Thereby, this paper applies PCA as preprocessof the soft clustering method, FCM and FAC. This would lead to the improvement of the clustering efficiency. The performance of the clustering approaches, with and without PCAis evaluated in terms of unsupervised and supervised measurement, the clustering objective function and adjusted rand index (ARI) [8]. Furthermore, dimension reduction, resulted from PCA is determined. The results of PCA-FCM and PCA-FAC are compared with FCM as well as FAC. The rest of the paper is organized as follows. Section II introduces FCM and FAC clustering. In addition, the soft clustering with PCA preprocessing is described in this section. Then, experimental results are considered in section III. Finally, conclusions are made in section IV.

II. Using Principal Component Analysis with Soft Clustering Approaches

A. Fuzzy c-means (FCM)

Fuzzy C-Means (FCM) is a clustering method that allows a data point to belong to two or more clusters with different degrees of membership; unlike k-means (KM), the traditional clustering method that assigns a pattern to only a single cluster. FCM is widely used in pattern recognition. It is based on minimization of the following objective function:

$$\sum_{n=1}^{N}\sum_{c=1}^{C} \boldsymbol{u}_{nc}^{m}\|\boldsymbol{x}_n - \overline{\boldsymbol{x}}_c\|^2 \qquad (1)$$

where, $\mathbf{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n, ..., \boldsymbol{x}_N\}$, $\boldsymbol{x}_n$ is the n^{th} of d-dimensional measured data; is a set of data to be clustered $\boldsymbol{x}_c$ is a c^{th}cluster centers, where c = 1, 2, …, C. m, fuzziness degree controls the extent of membership sharing between fuzzy clusters; here it equals 2, $\boldsymbol{u}_{nc}$ is the degree of membership of input $\boldsymbol{x}_n$ in the

This work was supported in part by the Informatics Department, Faculty of Science and Technology, Phuket Rajabhat University, Thailand.

P. Julrode is with the Faculty of Science and Technology, Phuket Rajabhat University, Phuket, 83000 Thailand (corresponding author phone: +66 (076)-211-959; fax: +66 (076)-211-778; mobile: +66 (086)-115-6338; e-mail: phichete@pkru.ac.th).

cluster *c*. ‖*‖ is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown in Eq. (1). The update of membership $\boldsymbol{u}_{nj}$ and the cluster centers $\boldsymbol{x}_c$ follow Eq. (2) and Eq. (3) consecutively:

$$\boldsymbol{u}_{nj} = \left(\sum_{c=1}^{C} \frac{\left\| \boldsymbol{x}_n - \overline{\boldsymbol{x}}_j \right\|^{\frac{2}{m-1}}}{\| \boldsymbol{x}_n - \overline{\boldsymbol{x}}_c \|} \right)^{-1} \quad (2)$$

$$\overline{\boldsymbol{x}}_j = \frac{\sum_{c=1}^{C} \boldsymbol{u}_{nj}^{m} \boldsymbol{x}_n}{\sum_{n=1}^{N} \boldsymbol{u}_{nj}}. \quad (3)$$

This iteration will stop when: $max\left\{\left|\boldsymbol{u}_{nc}^{iter+1} - \boldsymbol{u}_{nc}^{iter}\right|\right\} < \varepsilon$where, ε is a termination criterion ranged between 0 and 1 and superscript *iter*is the iteration number. However, the problem of getting into local optima still exists in FCM learning.

B. Ant based clustering (ANT)

The ant systems, developed with concept of simple multi-agent principles emphasize distributiveness, flexibility, and robustness. Ant-based algorithm has been developed using swarm intelligence principles that emphasize distributiveness, direct or indirect interactions among relatively simple agents, flexibility, and robustness [9], [10]. By such competent characteristics, ant-based clustering more relieves the fast convergence during searching process than several other evolutionary approaches. Groups of ants cooperate to move cluster centers in feature space to search for optimal clusters partition.

Initially, the feature values are normalized between 0 and 1. Each ant is assigned to a particular feature of a cluster in a partition. The ants never change the feature, cluster or partition assigned to them. After randomly moving the cluster centers for a fixed number of iterations, called an epoch, the quality of the partition is evaluated. According to the former ant colony optimization, particular clusters partition is selected as the best one based on a certain cascading rules of probability. A certain additional memories are exploited for storing some numbers of good clusters. However, the ant colony optimization used in this paper selection does not need the additional memories to storemore than one good cluster centers; instead, it applies the heuristic technique, based on Boltzmann probability [11]. Such technique supports the exploration of the good clusters. Similar to the former ant colony optimization, there are two directions for the random movement of the ant. The positive direction is when the ant is moving in the feature space from 0 to 1, and the negative direction is when the ant is moving in the feature space from 1 to 0. If during the random movement the ant reaches the end of the feature space the ant reverses the direction. After a fixed number of epochs the ants stop. The cluster centers, finally obtained are then used as the clustering approaches.

C. Fuzzy ant based clustering (FAC)

The fuzzy ant-based clustering with cluster centroids positioning (FAC) was originally proposed by [12]. It is a combination between ant-based clustering (ANT) and FCM aiming to search for optimal partition of cluster centers. Initially, the feature values are normalized between 0 and 1. An ant is assigned to a particular feature of a cluster center in a partition. To search for the new partition of clusters, ants randomly move the clusters in a corporative manner. Two directions are defined for the random movement of the ant. The positive direction is when the ant is moving in the feature space from 0 to 1, and the negative direction is when the ant is moving in the feature space from 1 to 0. If during the random movement the ant reaches the end of the feature space, the ant reverses the direction. After moving the cluster centers for a fixed number of iterations, the quality of the partition is evaluated, using FCM objective function specified in Eq. (4):

$$FCM\_ObjectiveFunction = \sum_{i=1}^{N} \sum_{k=1}^{K} \boldsymbol{\mu}_{ik} \left\| \mathbf{x}_i - \mathbf{c}^k \right\|^2 \quad (4)$$

where$\boldsymbol{\mu}_{ik}$ represents membership of $\mathbf{x}_i$, which is sample *i*in cluster $\mathbf{c}^k$. For crisp data, $\boldsymbol{\mu}_{ik}$ is zero if $\mathbf{x}_i$ is in cluster $\mathbf{c}^k$, and is one if not. After the ant process is terminated, the best partition achieved is submitted to FCM to carrying on the clustering and attain the better result

D. Using principal component analysis with fuzzy c-means and fuzzy ant based clustering

Principal component analysis (PCA) is an orthogonal basis transformation [13]. Given a data set:$\{\mathbf{X}_n \in \boldsymbol{R}^D | n = 1, \ldots N\}$, where *D* is the number of dimensions, *N* refers to the samples size.$\mathbf{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \ldots \boldsymbol{y}_N\}$is given as a centered matrix;$\boldsymbol{y}_n = \mathbf{x}_n - \overline{\mathbf{x}}$, where$\overline{\mathbf{x}} = \sum_{n=1}^{N} \mathbf{x}_n / N$. The basis is found by diagonalizing the centered *N*×*N*covariance matrix, defined by Eq. (5).

$$\mathbf{M} = \sum_{n=1}^{N} (\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^T = \mathbf{Y}\mathbf{Y}^T \quad (5)$$

The coordinates in the eigenvector basis are called principal components. In PCA, one has to find eigenvalues λ and eigenvectors associated respectively within diagonal matrixλ and **V**of**M**, satisfyingEq. (6).

$$\lambda\ \mathbf{V} = \mathbf{M}\mathbf{V} \quad (6)$$

The size of each eigenvalue λ equals the amount of variance in the direction of the corresponding eigenvectors**V**. Thedirections of the first eigenvectors corresponding to the biggesteigenvalues cover as much variance as possible by *P* orthogonal directions, where$P \leq D$The principal eigenvectors**V'** are the principal directions of **Y**;and are the principal components. Entries of eachdimension in the new space, **W**are the projected values of data points on the principal direction**V'**; and is calculated by Eq. (7).

$$\mathbf{W} = \mathbf{Y}^T \mathbf{V} / \sqrt{\lambda} \quad (7)$$

Through such a PCA method, the main *P* of *D* dimensions is extracted; whilst noisy and irrelevant dimensional features that could seriously deteriorate the generalization performance of clustering are eliminated. The PCA is utilized in this paper as a preprocess that generates dimension reduction data with

noisy decreased.Such refined data is later employed in FCM and FAC learning process.

III. Experimental and Results

PCA-FCM and PCA-FAC are tested on seven benchmark medical data sets obtained from the URL, "http://archive.ics uci.edu/ml/datasets.html". The characteristics of the tested data sets are summarized in Table I.

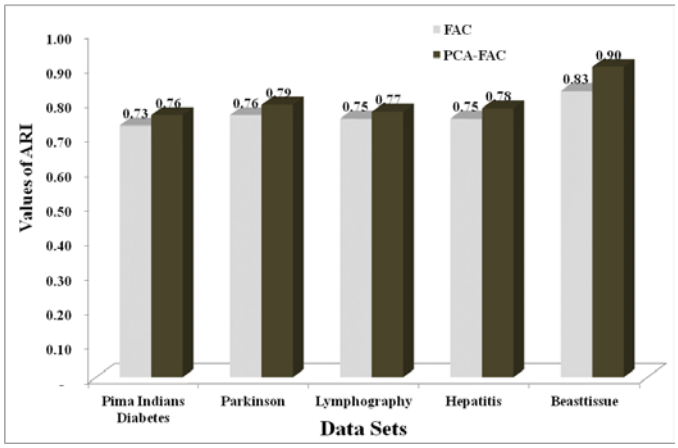
TABLE I. Characteristics of Data Sets

Name of data set	No. of classes	No. of features	Size of data set *(size of classes in parentheses)*
Pima Indians Diabetes	2	8	768 *(500, 268)*
Parkinson	2	22	195 *(48, 147)*
Lymphography	4	18	148 *(2, 67, 46, 33)*
Hepatitis	2	19	155 *(32, 123)*
Breast Tissue	6	9	106 *(21, 15, 18, 16, 14, 22)*

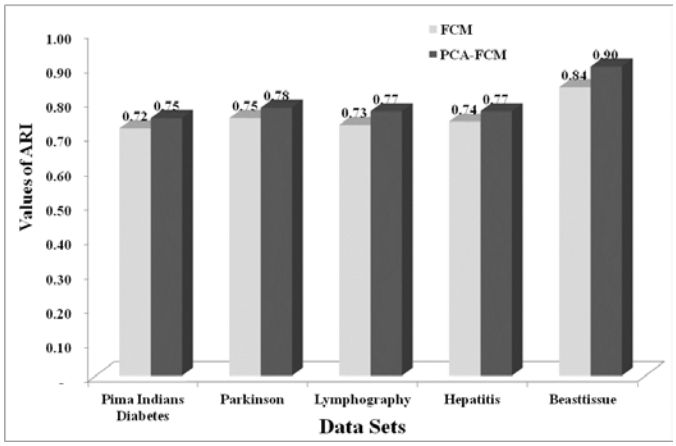
The performance comparison is performed on such two algorithms and PCA-FCM and PCA-FAC as well as other related: FCM and FAC alone. According to PCA feature reduction, the number of orthonormal eigenvectors, corresponding to the first 90% (threshold of selection) of accumulation of the largest eigenvalues of the covariance matrix would be used in the feature space. Such a selection criterion of eigenvector is applied for all data sets and related methods. The quality of the respective clustering are compared, where the quality is measured by two criteria: one refers to adjusted rand index (ARI). It is calculated based on the following procedures: suppose T is the true clustering of a data set based on domain knowledge and R a clustering result given by a clustering algorithm. Let a, b, c, and d, respectively, denote the number of pairs belonging to the same cluster in both T and R, the number of pairs belonging to the same cluster in T but to different clusters in R, the number of pairs belonging to different clusters in T but to the same cluster in R and the number of pairs belonging to different clusters in both T and R. The ARI(T,R) is then defined in Eq. (8).

$$ARI(T,R) = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)} \quad (8)$$

The value of ARI(T,R) lies between zero and one and higher value indicates that R is more similar to T. In addition, ARI(T,T) = 1.In Fig. 1 (a) and (b), with respect to most data sets PCA-FAC(dark bars) and FAC (light bars) as well as PCA-FCM (dark bars) and FCM (light bars) seems to generate competitive results, although for some other data sets the clustering with dimension reduction produce a particular higher degree in terms of ARI-based measure. Nevertheless, ARI measurement is perceived as a supervised evaluation method, since the true clustering results are known. By this reason, the conspicuous results may not be achieved.



(a)



(b)

Figure 1. The measurement (a) ARI value of FAC and PCA-FAC, (b) ARI value of FCM and PCA-FCM

The other quality measurement criterion here is relevant to FCM and FAC objective functions. Both of these measurements can be mentioned as unsupervised evaluation of clustering methods because of the inexistence of true clustering results. In Table II, it is explicitly seen theremarkable better clustering quality of PCA-FCM and PCA-FAC over FCM and FAC, in terms of the average of objective functions degree. The bold digits point out the better degree of objective function. The standard deviations are specified in bracket.

Other than such measurement, the time complexities of the related clustering methods are reported The complexities of FCM and FAC in the worst case are $O(nc^2)$; whilst PCA's is $O(n)$, where n and c consecutively represent the number of data and the number of clusters. Therefore, the use of PCA with either FCM or FAC does not give much effect.

TABLE II. CLUSTERING QUALITY OF FAC,PCA-FAC,FCM AND PCA-FCM IN TERMS OF OBJECTIVE FUNCTIONS

Data sets	FAC	PCA-FAC	FCM	PCA-FCM
Pima Indians Diabetes [8x768]	3.24E+06(1.79E+03)	**3.28E+01**(1.45E+00)	0.67E+06(1.19E+05)	**1.62E+02**(1.38E+00)
Parkinson [22x195]	0.94E+06(1.11E+02)	**1.79E+01**(1.63E+00)	2.78E+05(1.13E+04)	**3.11E+01**(3.16E+00)
Lymphography [18x148]	0.85E+03(1.55E+01)	**1.71E+01**(1.55E-02)	4.12E+01(1.94E-02)	**3.22E+00**(1.52E-03)
Hepatitis [19x155]	1.03E+06(1.04E+02)	**0.40E+00**(0.00E+00)	2.33E+05(1.20E+03)	**0.89E+01**(1.80E+00)
Breast Tissue [9x106]	1.53E+10(1.92E+03)	**0.12E+00**(1.66E-01)	1.16E+08(1.96E+07)	**0.65E+00**(4.72E-02)

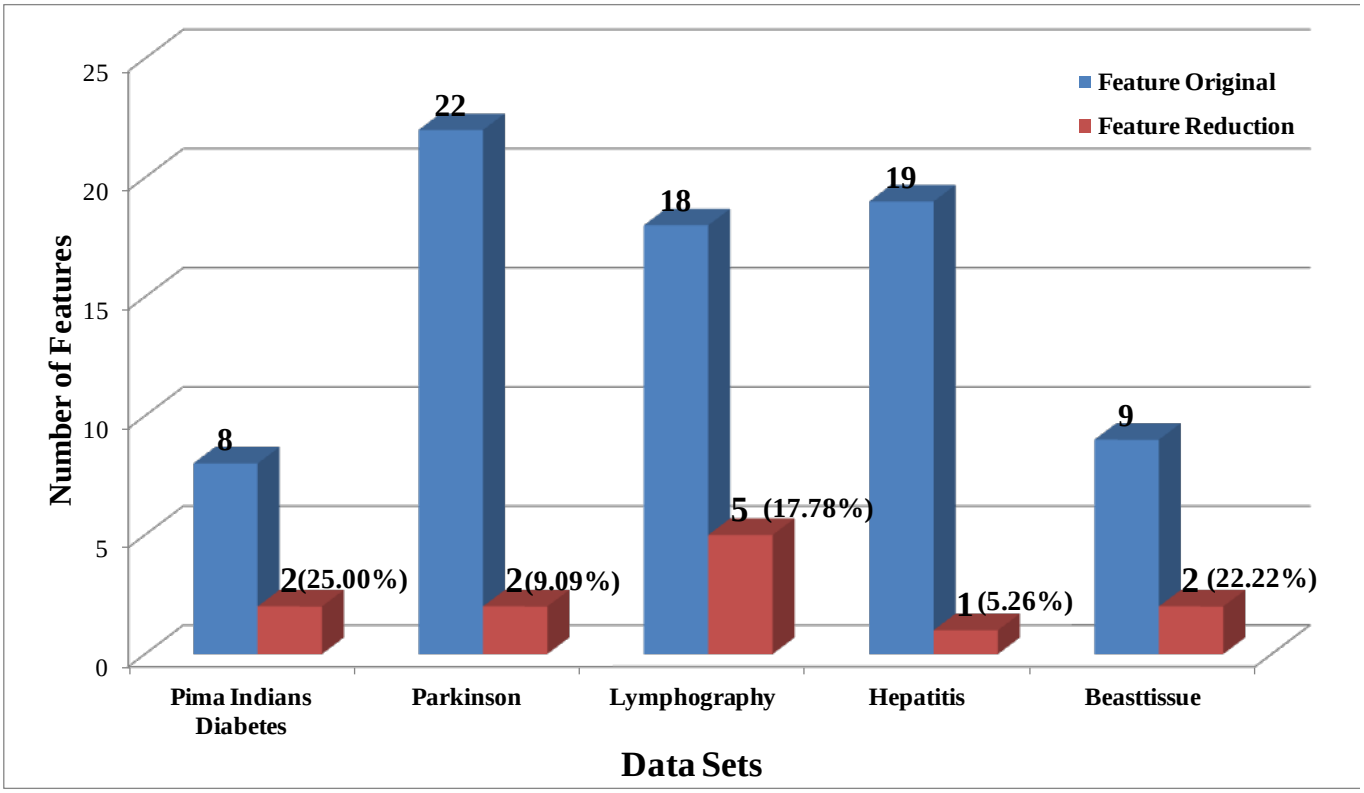


Figure 2.The dimension reduction on seven medical data sets, yielded from PCA

In Fig. 2, the dimension reduction is described in a form of bar graph. The light and dark bars respectively represent the original number of data dimensions and those of the reduced. The percentages of the after-reduced dimensions for each data set are pointed above the gray bars

IV. CONCLUSIONS

This paper investigates on using the principal component analysis, (PCA) as feature reduction preprocess for efficient soft clustering approaches, fuzzy c-means and K-harmonic means. The combined clustering approaches are named here, PCA-FCM and PCA-FAC. Both of the approaches are applied with medical diagnosis applications where the high dimensional, noisy data are related; and the sensitivity is needed in decision making. The performance of PCA-FCM, PCA-FAC along with FCM and FAC alone are put into comparison. Comparison tests are performed on 7 significant medical data sets. The performance measurements for each method are based on unsupervised and supervised criteria, FCM / FAC objective functions and the adjusted rand index (ARI).Such criteria are calculated over 10 independent runs. In terms of unsupervised measurement, the results indicate the significance and superiority of using the feature reduction, PCA on the soft clustering approaches, FCM and FAC; whilst the time complexity in the worst case regarding PCA, *O(n)* barely effects the time performance of the clustering process. Nevertheless, there still exist some drawbacks relating to PCA-FCM and PCA-FAC. One is that they require a priori known number of clusters. It is not applicablewhen the number of clusters is unknown. Some other ways ofclustering such as agglomerative or divisive may be required. Another interesting future work may concerns with using non-linear feature reduction instead of the linear for data preprocessing purpose.

ACKNOWLEDGMENT

We gratefully acknowledge the financial support from Science and Technology faculty, Phuket Rajabhat University, Thailand.

REFERENCES

[1] J. C. Bezdek, R. Ehrlich and W. Full, "FCM: The fuzzy c-means clustering algorithm,"*Computers & Geosciences*,vol. 3no. 10, pp. 191-203, 1984.

[2] W. Chen, M. L. Giger and U. Bick, "A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images," *Academic Radiology*,vol. 13, no. 1, pp. 63-72, 2006.

[3] S. Schockaert, M. D. Cock, C. Cornelis, and E. E. Kerre, "Fuzzy Ant Based Clustering,"*Springer Berlin Heidelberg*,vol. 3172, pp. 342-349, 2004.

[4] P. M. Kanade and L. O. Hall, "Fuzzy ants and clustering," *IEEE Trans. System Man Cybernetics, A Syst. Humans*, Vol. 37, pp. 758-769, 2007.

[5] P. Julrode and S. Supratid, "Improved Fuzzy Ant-Based Clustering: A Nonparametric Balance Between Exploitation and Exploration," *Research Journal of Applied Sciences*, vol. 8, no. 9, pp. 425-434, 2013.

[6] M. Andrzej and R. Waldemar, "Principal Components Analysis (PCA),"*Comput. Geosci*, vol. 19, no. 3, pp. 303-342, 1993.

[7] I. T. Jolliffe, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37-52,1986.

[8] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering,"*IEEE Trans. Patter Anal. Mach. Intell.*,vol. 13, no. 8, pp. 841-847, 1991

[9] Bonabeau E, Dorigo M, Theraulaz G. *Swarm Intelligence: from natural to artificial systems*. New York, USA: Oxford University Press, Inc., 1999.

[10] M. Dorigo , E. Bonabeau and G. Theraulaz, "Ant algorithms and stigmergy", *Future Generation Computer Systems*, 16(8), 2000, pp. 851-871.

[11] C. Y. Lee, "Entropy-Boltzmann selection in the genetic algorithms," *IEEE transactions on systems man and cybernetics Part B Cybernetics a publication of the IEEE Systems Man and Cybernetics Society*, vol. 33, no. 1, pp. 138-149, 2003.

[12] P. M. Kanade and L. O. Hall, "Fuzzy ants as a clustering concepts,"*In Proc. of the 22nd International Conference of the North American Fuzzy Information Processing Society (NAFIPS)*, pp. 227-232, 2003.

[13] S. Chen, S. A. Billings and W. Luo, "Orthogonal least squares methods and their application to non-lenear system identification," *International Journal of Control*, Vol. 56, No. 2, pp. 1013-1032, 1989.

Phichete Julrode received the B.Sc. in computer science from Phetburi Rajabhat University, Thailand in 1995. He received his master's degree in computer science from the Chiangmai University, Thailand in 2005and doctoral degree in information technology from Rangsit University, Thailand, in 2013. He is currently a lecturer in the informatics department, faculty of science and technology at Phuket Rajabhat University, Phuket, Thailand. His research interests include artificial intelligence, data mining, machine learning, and evolutionary algorithms.

Social return valuation by means of linear and nonlinear transformation methods in income taxation

KALININA OLGA
Department “Strategic Management”
Saint-Petersburg State Polytechnical University
Address: 195251, St. Petersburg, ul. Polytechnique, 29, III Academic Building, Rm. 409
RUSSIAN FEDERATION
olgakalinina@bk.ru, http://www.sm.spbstu.ru

Abstract: In this paper we propose methods for building progressive scales of income taxation. On the basis of our calculations, we identify the social return of introducing a progressive scale at the same time with changing the values of tax rates for the first group of population with the lowest income. A generalised model of transition to progressive taxation is formed.

Key-Words: progressive tax scale, income taxation, social return, social orientation, method, mathematical model.

1 Introduction. Retrospective analysis of the effects of introducing a flat income tax scale in Russia

Currently in Russia personal income tax (PIT) is one of the major taxes and ranks third after Value added tax (VAT) and corporate income tax. Along with the corporate income tax, it is central to regional and local budgets. In 2013, consolidated budget revenue of federal subjects of Russia accounted for about 5.9 trillion rubles, more than 1/3 is taken by income tax on individuals [20].

However, despite the third largest share of fiscal performance, the share of personal income tax in consolidated revenue is about 13%. According to experts, this is due to low tax potential of most of the population and the use of a flat income tax scale.

At the moment, the most common form of social orientation of taxation system in international practice is using progressive taxation. In 2001 progressive income taxation scale was declined in the Russian Federation, a flat tax rate of 13% was introduced instead. The main argument justifying the introduction of a flat tax scale was the idea that large revenues will be withdrawn from the shadow economy. This is proved by Table 1, showing personal income tax trends for the period 2000-2007.

The introduction of the flat rate in 2002 increased fiscal performance efficiency by almost 50%, in 2001-2002 personal income tax collections increased almost three times [10]. In the period from 2000 to 2007 monthly monetary income per capita increased from 2 502 rub. to 14 940 rub., i.e. almost six times, due to legitimate pay [3].

Other positive effects of introducing a flat tax scale include the following:

- Simplifying the procedure for determining the taxation base and the procedure for calculation and payment;
- Reducing taxpayers' costs related to reporting;
- Reducing the cost of the state administration.

However, in practice this reform justified itself only partially, as in the period from 2000 to the present time, the socio-economic inequality has been only reinforced – the gap between the rich and the poor as of 2013 according to official statistics was 16.2 times. Compared with the beginning of perestroika, at that time the gap between the poorest 10% and the richest was 5-7 times [19].

Analysis of existing opinions and positions of Russian scientists, economists and government officials revealed the following negative aspects of introducing a flat tax scale:

1. *Flat income tax is socially unfair*. This reform has worsened the situation of the lowest earning taxpayers:

- Table 2 shows PIT rates before and after introducing a flat tax scale in 2001:

Table 2

PIT rates before and after introducing a flat tax scale in 2001, %

RF in 2000	12	17,3	25,8
RF from 2001	13		

Table 1

Trends of remuneration and personal income tax, 2000-2007. *

Indicators	2000	2001	2002	2003	2004	2005	2006	2007
PIT, bn rubles	174,8	255,8	358,1	455,7	574,5	707,1	930,4	1266,6
% of GDP	2,4	2,9	3,3	3,4	3,4	3,3	3,5	3,8

* Source: Federal State Statistics Service of the Russian Federation

As you can see from Table 2, Russian government increased PIT rate by 1%, which had a negative effect on the population with low and middle income, whereas for rich and super-rich population groups the tax rate became significantly lower. Moreover, after the reform, taxes on income from equity investments in companies (dividends) were imposed at the rate of 6%, i.e. 24% lower than before.

- The amount of tax returns for the lowest earning taxpayers increased by 9% at a constant taxable income, despite the fact that the rate increased by 1% (from 12% to 13%). So, if before the introduction of the Tax Code with taxable income at a minimum subsistence level (2112 rub. in 2003) personal income tax amounted to 253 rub. (2 112 rub. × 0.12), after the amendments it was 275 rub. (2 112 rub. × 0.13).
- The tax rate on personal income is onerous for the majority of citizens with income below or at the subsistence level. At the PIT rate of 13%, regardless of the size of the revenue, the higher the income, the greater the received benefit. Thus, the quality of living has been polarized [7].
- Before the introduction of the Tax Code of the Russian Federation (TC RF), tax-free amount was equal to two minimum wage rates, whereas after its introduction tax-free amount was 400 rub. per month (a standard tax deduction). Furthermore, this amount remained unchanged – without taking into account inflation and the minimum wage. It means that for an employee who received income at the subsistence level (2 112 rub. in 2003) the annual amount of tax was 2 676 rub. [(2 112 rub. - 400 rub.) × 0.13 ×12], which is 27% more than his monthly income. In addition, each ruble for such an employee is worth more than for an employee receiving a larger income. From 1 January, 2009 the limit of applying a standard tax deduction was raised only up to 40 000 rub. Accordingly, for low-income citizens PIT is applied to the amount of income required for simple reproduction (currently minimum wage accounts for 5 554 rubles monthly or 66 648 rubles annually, since 01.01.2014); subsistence minimum per capita as for IV quarter of 2013 is 7 326 rub. monthly or 87 912 annually).
- According to State Statistics Committee of the Russian Federation, in 2001 nearly half of the income of the population was concentrated in the group with the highest incomes. As V.G. Panskov says [13], PIT rate reduction for this group has further enhanced the stratification of the population in terms of prosperity.

2. The effect of the introduction of a flat rate was generally not as significant:

- The introduction of a flat PIT rate of 13% has had no significant impact on the growth rate of wages and has not led to the disclosure of shadow incomes. This was largely due to the lack of proper control over the formation of income and tax payment [9]. According to B.H. Aliyev [1], the introduction of a flat rate has not solved the problem of withdrawing incomes from shadow economy as "businessmen do not care how much tax their employees pay, it is important for them how much they pay themselves."
- As for small businesses, formed as sole proprietorships, they were not interested in legalizing their income and disclosing real wage funds, because at that period the unified social tax rate was quite high and amounted to 35%.
- The introduction of a flat scale on personal income was premature, as the use of this scale requires certain conditions, in particular, broad middle class with relatively high wages and substantially lower differentiation of various population groups by income.
- Income tax collections have increased not due to the use of the unified reduced flat rate, but as a result of rising incomes. This is proved by Table 3, showing the trends of remuneration and personal income tax in 2000-2007.

ISBN: 978-1-61804-251-4

As Table 3 shows, PIT collections for the first year after the introduction of a flat rate accounted for 0.5% of GDP (from 2.5% in 2000 to 2.9% in 2001) and 0.4% for the second year, because remuneration of employees in percentage to GDP increased by 2.8% (from 29.1% in 2000 to 31.9% in 2001) and 3.3% accordingly.

Table 3

Trends of remuneration and personal income tax, 2000-2007. *

Indicators	2000	2001	2002	2003	2004	2005	2006	2007
PIT, bn rub	174,8	255,8	358,1	455,7	574,5	707,1	930,4	1266,6
% of GDP	2,4	2,9	3,3	3,4	3,4	3,3	3,5	3,8
Remuneration of employees, % of GDP	29,1	31,9	35,2	35,8	34,3	32,0	32,1	33,7
Hidden remuneration of employees, % of GDP	11,1	11,1	11,5	11,3	11,7	11,8	11,9	11,9
Companies' arrears of wages, bn rub.	43,7	31,7	29,9	30,6	24,4	14,3	5,8	4,2

* Source: Federal State Statistics Service of the Russian Federation + [6]

Retrospective analysis of the effects of introducing flat income taxation has led to the conclusion that Russian government officials, businessmen and scientists do not have a clear opinion about the suitability of flat taxation, almost immediately after the introduction of the flat tax scale the reform created its supporters and opponents.

Supporters of the existing flat tax scale believe that the current income taxation system is simple and transparent, and the rate is one of the lowest in Europe.

According to Russian economists, adverse effects of introducing a progressive tax include increasing fiscal differentiation of subjects of the Russian Federation, which leads to political tensions between regions, an increase of administrative and legal costs, and tax evasion [11].

Legislators, in their conclusion that the unified flat rate of taxation is fair, refer to Article 3 of the Tax Code of the Russian Federation, which declares the "principle of universality and equality of taxation" and that "taxes and levies can not be discriminatory" [21]. At the same time, advocates of the flat scale also appeal to the fundamental law of Laffer as the basis for tax policy – about the existence of a unified optimal tax rate [2,4], but it should be noted that the effect of this law does not prove the use of the tax rate of 13%.

Supporters of progressive taxation believe that the flat rate would lead to moral degradation of the population, growth of social aggression, lower investment activity of households and private businesses, reduced capacity of the domestic market and worsening competitiveness of the Russian economy as a whole.

In 2014, a meeting of the Expert Group on Strategies of socio-economic development of Russia for the period up to 2020 took place. One of the main issues of discussion was the introduction of a progressive income tax in Russia [15]. Participants of the discussion emphasized that progressive tax scale is a prerequisite for balance, social sustainability, equity and investment capacity of the economy. The introduction of progressive taxation will reduce social stratification, increase tax revenues and boost consumer demand.

Summing up this brief overview of various approaches of Russian economists and researchers to identifying the advantages and disadvantages of flat and progressive income taxation, we should note that Russian President Vladimir Putin does not disregard the possibility of transition to a differentiated taxation in the future [16].

Regarding international practice of taxation of the population, in most developed countries of the world progressive scale is used. For example, the range of tax rates in the United States varies from 0% for low earning workers to 39.6% for wealthy society sectors; in Great Britain this range is from 0 to 45%; in France it is 5÷41%; in Germany – 14÷45%; in Japan – 5÷50%. It should be noted that not only in developed countries but also in some of the BRIC countries progressive scale with a high maximum tax rate is applied: in China – 3÷45%, in Brazil – 7.5÷27.5%, in India – 10÷30% [14,22].

However, we should emphasize that some developed countries use flat taxation. For instance, in certain states of the United States (in particular, in Illinois, Massachusetts, Pennsylvania), in the Canadian province of Alberta, in Hong Kong and in Iceland there is a single tax rate on income. Flat income tax also exists in the countries of the former Soviet Union, in particular, the size of the tax rate in Estonia is 21%, in Georgia - 20%, Ukraine - 15%, Belarus - 12%; in the former CMEA: in the Czech Republic - 22%, Romania - 16%, Bulgaria – 10%. Flat scale is applied in taxation of most African countries [14,22].

ISBN: 978-1-61804-251-4

2 Problem Formulation

The identified features of income taxation in developed countries and in the Russian Federation, as well as cross-country analysis allow us to outline the following features of the income taxation system in Russia:

– Linear scale of taxation is used, whereas developed countries apply progressive scale;

– The lowest maximum rate of income tax is used;

– The share of tax revenues from personal income tax in the consolidated budget of the country and in the GDP is significantly lower compared to developed countries;

– There is no concept of non-taxable income, it is replaced by a nonequivalent standard tax deduction; in developed countries non-taxable income is generally equal to the minimum consumer basket;

– The gap between the wealthiest and poorest sectors of the population is significantly greater than in developed countries;

– Specific features of taxation lead to a greater stratification of society;

– Antisocial nature of taxation manifests itself when providing social and property-related tax deductions.

All this suggests the necessity of introducing a progressive income taxation system in the Russian Federation, which will replenish national budgets and achieve social equity by income redistribution.

When analyzing various economists' proposals to develop a progressive scale, attention is drawn to great numerical divergence of these recommendations. As an example will be given tax rates of a progressive scale offered by the Institute of Socio-Economic Studies and Population of the Russian Academy of Sciences [17] and the State Duma fraction Fair Russia [12].

Table 4

Example of various tax rates of a progressive scale offered by Russian economists

Institute of Socio-Economic Studies RAS		Fair Russia			
		Deputy O.G. Dmitrieva		Deputy O.A. Nilov	
		15.01.2013		21.03.2014	
million/year	%	million/year	%	million/year	%
under 13,2	0	under 3	13	under 5	13
13,2 – 30	16	3 – 15	25	5 – 50	18
30 – 60	30	15 – 30	35	50 – 500	23
60 – 100	43	> 30	50	> 500	28
100 – 150	50				
> 150	55				

From this table we can see that in proposals from one fraction, maximum taxable amount differs more than 15 times, whereas maximum rate differs almost 2 times.

The study showed that the reform proposals of domestic politicians and economists basically lack a comprehensive methodological framework, are fragmentary and are often simply declared without the development of methods and tools for introducing progressive income taxation and evaluating the efficiency of the proposed activities.

Therefore it is necessary to develop a “social” algorithm and methods of building tax scales in transition from the existing flat rate to progressive. The major questions that must be answered in the transformation of the tax scale are:

1) Whether the ongoing reform is aimed at increasing total fiscal revenue collection, or levy should remain the same as before the reform;

2) What tax rate for the low-income segment of the population should be considered acceptable in the process of redistributing the tax burden to the richer part of taxpayers.

Answers to other questions essential for reforming the tax scale are: the breakdown of taxpayers into groups by taxable income; choosing a linear or nonlinear progressive scale; taking into account the problem of tax evasion and so on within the proposed algorithm are regarded as the optimization of parameters selected as the basis for building the tax scale.

The experimental and statistical base for the study is a unified system of indicators of the Federal State Statistics Service.

ISBN: 978-1-61804-251-4

3 Problem Solution. Algorithm for building and optimization of a tax scale

3.1 Initial assumptions

In the economic literature we can find various terms to refer to tax scales, depending on the tax rate and the income subject to taxation. In this paper, we use the following terms:

Taxation scale is called *flat* if the value of the tax rate is the same for all taxpayers regardless of their individual income. In case when the tax rate rises with the increase of taxpayer's income the scale is called *progressive*. With proportional increase of tax rates and income, the scale is defined as *linearly progressive*, in other cases - as *nonlinear*. If the value of the tax rate decreases with increasing income of a taxpayer - the scale is *regressive*.

For building progressive scales of income taxation it is necessary to break all taxpayers into groups by the size of their income. We will assume that the number of such groups is *m* (the number of levels). We will arrange these groups with the increase of average income of taxpayers in the group and give them their respective number *i* = 1; 2; …*m*.

Currently, according to the Federal State Statistics Service of the Russian Federation, all Russian taxpayers are divided into 5 levels by 20 per cent (quintile) groups of distribution of total monetary income (*m* = 5), and 8 levels by average income per capita (*m* = 8) [18]; the first group is with the lowest income, the fifth and eighth with the highest (the Federal State Statistics Service also forms 10 percent (decile) groups, *m* = 10).

Each group has its own taxable base S_i:

$\sum_{i=1}^{m} S_i = S_0$ in which S_0 is existing taxable base.

With a flat scale of taxation (flat tax rate n_0) the total tax on personal income will be equal to:

$$C_0 = S_0 \cdot n_0 = \sum_{i=1}^{m} S_i \cdot n_0$$

For a progressive tax scale we introduce tax rates for each group, indicating them n_i accordingly.

Then the total tax base will be $S = \sum_{i=1}^{m} S_i$, and the general total tax:

$$C = \sum_{i=1}^{m} C_i = \sum_{i=1}^{m} S_i \cdot n_i \qquad (1)$$

3.2 Linear progressive tax scale. The case of constant tax collections

We will consider the case of transformation of a flat scale into progressive, in which the reform is not aimed at increasing the total income tax, i.e. with equal value of the tax base and the total income tax before and after the introduction of a progressive tax scale:

$$S = \sum_{i=1}^{m} S_i = S_0 \text{ and } C = \sum_{i=1}^{m} n_i \cdot S_i = C_0 \qquad (2)$$

We assume that the tax rate increases linearly with the taxpayer's income; such tax scales are used in a number of developed countries, such as USA, Canada, UK, France, Germany, etc.

The condition for linear increase of the tax rate is a constant value of the tax rate increment Δ from group to group $\Delta = n_{i+1} - n_i = c \quad o$ (3),
and the rate in the group will be:

$$n_i = n_1 + \Delta \cdot (i-1),$$

where n_1 is the tax rate on personal income for the first lowest earning group of taxpayers.

From formula (2) follows:

$$S_1 n_1 + S_2 n_2 + . \ S_m n_m - S_0 n_0 = 0 \qquad (4)$$

Inserting formula (3) into (4), we obtain:

$$S_1 n_1 + S_2 (n_1 + \Delta) + S_3 (n_1 + 2\Delta) + S_m [n_1 + (m-1)\Delta] - S_0 n_0 = 0$$

or

$$n_1 (S_1 + S_2 + . \ S_m) + \Delta [S_2 + 2S_3 + . \ (m.-1) S_m] - S_0 n_0 = 0$$

Given the condition of preservation of the tax base (2), we have:

$$(n_1 - n_0) S_0 + \Delta [S_2 + 2S_3 + . \ (m-1) S_m] = 0$$

hence the increase in the tax rate will be determined by the formula:

$$\Delta = \frac{(n_0 - n_1) S_0}{S_2 + 2S_3 + . \ (m.-1) S_m} \qquad (5)$$

Inserting formula (5) into (3), we find the values of all the rates for groups for a linear progressive tax scale:

$$n_i = n_1 + \frac{S_0 (n_0 - n_1)(i-1)}{\sum_{j=2}^{m} (j-1) S_j} =$$

$$n_1 + \frac{(n_0 - n_1)(i-1)}{\sum_{j=2}^{m} (j-1) \eta_j} \quad , \qquad (6)$$

in which $\eta_j = \frac{S_j}{S_0}$.

The results of calculation of tax rates by formula (6) are shown in Table 5. The values of income distribution η_i are taken from the Federal State Statistics Service of the RF for 2014. The range of tax rates for low income groups η_i fluctuated from

10% to 0%. Note that many countries practice a complete exemption from tax for the poor, so considering option $n_1 = 0$ is relevant.

The table shows redistribution of the tax burden from low earning groups to the wealthy. We should point out that the tax burden for the forth group of taxpayers has not changed, whereas for the fifth group it has increased by 35%.

Table 5

Linear progressive tax scale: the results of calculations, 2014

i	**1**	**2**	**3**	**4**	**5**	
η_i **n_1 %**	0,052	0,099	0,149	0,225	0,475	
13	13	13	13	13	13	**n_i %**
10	10	11	12	13	14	
5	5	7,7	10,4	13	15,8	
0	0	4,4	8,7	13	17,5	
13	1,000	1,000	1,000	1,000	1,000	**n_i /n_0**
10	0,769	0,847	0,925	1,000	1,080	
5	0,385	0,592	0,799	1,000	1,213	
0	0,000	0,336	0,673	1,000	1,346	

We can show that if in transforming the flat scale into linear progressive we take as a parameter the tax rate for "the wealthy" n_m, and not the rate for the poor n_1, then the formula for calculating all the other rates will be

$$n_i = n_m - \frac{(n_m - n_0)(m-i)}{\sum_{j=1}^{m-1}(m-j)\eta_j} \qquad (6a)$$

3.3 Linear progressive tax scale. The case of an increase in tax collections

Let us consider the case when the transformation of a flat tax should be accompanied by an increase in total income tax.

We introduce the *coefficient of the planned increase in tax revenue* by increasing the total tax on the income of individuals in α times: $C = \alpha \cdot C_0$

With a flat rate tax, additional tax burden is equally imposed on all people; with a linear progressive scale, tax rates will be calculated for the tax base $\alpha \cdot S_0$ by the formula:

$$n_i = n_1 + \frac{(\alpha \cdot n_0 - n_1)(i-1)}{\sum_{j=1}^{m}(j-1)\eta_i} \qquad (7)$$

The results of calculations of tax rates by formula (7) required for increasing tax revenues by 20% are presented in Table 6. Gradations of tax rates n_1 and the distribution of monetary income in groups η_i are similar to those contained in Table 5.

Table 6

Linear progressive tax scale, increasing in tax collections: the results of calculations, 2014

α = 1,2						
i	**1**	**2**	**3**	**4**	**5**	
η_i **n_1 %**	0,052	0,099	0,149	0,225	0,475	
13	13	13,9	14,7	15,6	16,5	**n_i %**
10	10	11,9	13,8	15,7	17,5	
5	5	8,6	12,1	15,7	19,3	
0	0	5,2	10,5	15,7	21	
13	1.000	1,067	1,135	1,202	1,269	

ISBN: 978-1-61804-251-4

10	0.769	0,914	1,059	1,204	1,349	$\mathbf{n_i / n_0}$
5	0,385	0,659	0,933	1,208	1,482	
0	0,000	0,404	0,808	1,211	1,615	

The obtained results shown in Tables 5 and 6, allow to visually compare the tax rates with an increase in the total tax revenue from income tax by 20%.

3.4 Nonlinear progressive tax scale

Let us consider a transformation mechanism of a nonlinear progressive tax scale. Due to the ambiguity of the use of terms in the economic literature, we will focus on the terms used in this paper for describing different taxation scales.

We will consider the case when the tax scale is a continuous function of the income W, $n = f(W)$. Figure 1 shows examples of three types of progressive taxation $n = f(W)$:

1 **–** flat scale (the tax rate does not depend on the size of income);

2 – linear progressive tax scale, whose specific features are described above;

3 – nonlinear progressive scale with increasing dynamics of tax rate growth;

4 – nonlinear progressive scale with decreasing dynamics of tax rate growth.

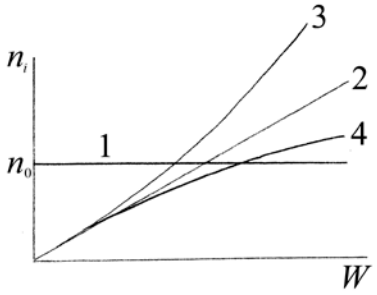


Fig. 1. Various types of tax scales

These terms will be also applied in case of tax distribution by groups to a step change of tax rate n_i from group tp group. Consideration of such taxation scales is of interest both to the study of shifting the tax burden on the highest income group, and in addressing the problem of tax evasion.

Exponential distribution and parameters distribution as a geometric progression are among the most common in describing various relations in the economy, and the corresponding mathematical models match physical nature of real processes and lead to constructive solutions in analytical and in algorithmic form [5]. In contrast to traditional methods, we will consider a nonlinear scale, changing according to the law of "double arithmetic progression", as a more convenient when describing a step tax scale.

For building a linear progressive scale the following ratios were used (3):

$$n_i = n_1 + \Delta \cdot (i-1);\ \Delta = n_i - n_{i-1} = c \quad o \ .$$

For a nonlinear scale, the value of tax rate increment Δ does not remain constant; we will assume that Δ itself changes in an arith metic progression:

$$\Delta_i = \Delta_1 + \delta \cdot (i-1);\ \delta = \Delta_i - \Delta_{i-1} = c \quad o \qquad (8)$$

Inserting formula (8) into (3) we obtain:

$$n_i = n_1 + [\Delta_1 + \delta(i-1)](i-1) \qquad (9)$$

Value δ will be defined as a part of Δ_1 :

$$\delta = k \cdot \Delta_1,\ -1 < k < 1,$$

where k is a coefficient of “nonlinearity”.

Then

$$n_i = n_1 + \Delta_1[1 + k(i-1)] \cdot (i-1) \qquad (10)$$

The proposed method of forming a non-linear scale is illustrated by Figure 2.

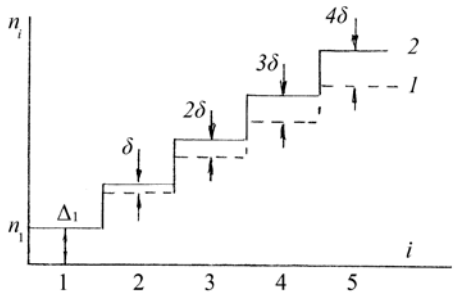


Fig. 2. Nonlinear progressive tax scale, where **1** (dotted) is a linear scale, **2** is a nonlinear scale ($\delta = k \cdot \Delta_1,\ k > 0$)

Inserting (10) into the equation (4) we will obtain:

$$S_1 n_1 + S_2[n_1 + \Delta_1(1+k)] + S_3[n_1 + \Delta_1(1+2k)] \cdot 2 + \ldots$$
$$+ S_m[n_1 + \Delta_1(1 + (m-1)k)](m-1) - S_0 n_0 = 0$$

or

$$n_1(S_1 + S_2 + \ldots S_m) + \Delta_1\{S_2(1+k) + 2S_3(1+2k) + \ldots\}$$
$$\{+ (m-1)S_m[1 + (m-1)k]\} - S_0 n_0 = 0$$

. Taking into account that $\sum_{i=1}^{m} S_i = S_0$, we have:

$$\Delta_1 = \frac{(n_0 - n_1)S_0}{\sum_{j=2}^{m} S_j(j-1)\left[1+(j-1)\cdot k\right]} \quad (11)$$

Inserting formula (11) into formula (10) and shifting to non-dimensional coefficients η_i, we will eventually get:

$$n_i = n_1 + \frac{(n_0 - n_1)\cdot S_0\left[1+k(i-1)\right](i-1)}{\sum_{j=2}^{m} S_j(j-1)\left[1+(j-1)\cdot k\right]} =$$

$$= n_1 + \frac{(n_0 - n_1)\cdot\left[1+k(i-1)\right]\cdot(i-1)}{\sum_{j=2}^{m} \eta_j(j-1)\left[1+(j-1)\cdot k\right]} \quad (12)$$

The calculation results of tax rates by formula (12) are shown in Tables 7 and 8 (for k = 0.2 and k = - 0.12); initial parameters correspond to Table 5.

Table 7

Nonlinear progressive tax scale, *k*>0: the results of calculations, 2014

***k* = 0,2**						
i	**1**	**2**	**3**	**4**	**5**	
η_i **n_1 %**	0,052	0,099	0,149	0,225	0,475	
13	13	13	13	13	13	**n_i %**
10	10	10,7	11,7	12,9	14,3	
5	5	6.9	9.4	12,6	16,4	
0	0	3.1	7.2	12,4	18,6	
13	1,000	1,000	1,000	1,000	1,000	**n_i /n_0**
10	0,769	0,824	0,898	0,989	1,099	
5	0,385	0,531	0,727	0,971	1,264	
0	0,000	0,238	0,556	0,953	1,430	

Table 8

Nonlinear progressive tax scale, *k*<0: the results of calculations, 2014

***k* = -0,12**						
i	**1**	**2**	**3**	**4**	**5**	
η_i **n_1 %**	0,052	0,099	0,149	0,225	0,475	
13	13	13	13	13	13	**n_i %**
10	10	11,5	12,6	13,3	13,6	
5	5	9,1	12,0	13,9	14,6	
0	0	6,6	11,4	14,4	15,6	
13	1,000	1,000	1,000	1,000	1,000	**n_i /n_0**
10	0,769	0,886	0,972	1,025	1,046	
5	0,385	0,697	0,924	1,066	1,123	
0	0,000	0,508	0,877	1,108	1,200	

3.5 Social return and the generalised model of transition to progressive taxation

Transition from flat to linear and nonlinear progressive taxation scales has a social nature of redistribution of tax burden among population. Social return means the difference between the values of tax rates for the most and least wealthy groups of taxpayers. In other words, for instance,

under the conditions of Table 7 social return (d) can be calculated according to the formula $d = n_5 - n_1$.

Comparing the obtained values of building a nonlinear progressive scale with the tax rates values of a linear progressive scale, we can conclude that:

- the greatest social return is observed when building a nonlinear scale at $k > 0$;
- the value of social return when using a linear income taxation scale ranks second;
- minimum, yet existing, social return is present when building a nonlinear progressive scale at $k < 0$.

Summarizing the results, we obtain a generalised mathematical model of transition to progressive income taxation, adaptive to changing political, economic and social conditions, which allows to calculate the coefficients of taxation for various tax scales simultaneously:

$$n_i = n_1 + \frac{(\alpha \cdot n_0 - n_1) \cdot [1 + k(i-1)] \cdot (i-1)}{\sum_{j=2}^{m} \eta_j (j-1) \cdot [1 + (j-1) \cdot k]}, \qquad (13)$$

where:

n_0 – tax rate of a flat taxation scale, $n_0 = 13\%$;

n_1 – tax rate for the first, lowest-income group (selected as a "basic" socially significant parameter);

n_i – tax rate in group i;

$\alpha = \frac{C}{C_0}$ - coefficient of a planned tax revenue increase;

m – the number of taxation groups;

k – coefficient of "nonlinearity" of the scale, which takes into account the pace of progressive scale changes;

$\eta_i = \frac{S_i}{S_0}$ - coefficients of monetary income distribution by groups in relation to total income S_0;

j – index used for summing the shares of coefficient n_i .

The generalised model allows us to estimate all possible options for the reform of the proposed transformation of a flat scale into progressive [8]. In this case the above mentioned coefficients vary depending on external conditions, as well as targets and challenges which face public authorities developing tax policy for the short and medium term.

It must be emphasized that this model remains valid for different taxpayers breakdowns, such as their division by average income per capita, when the entire population is divided into 8 levels ($m = 8$), and the corresponding values of monetary income are given in percentages [18].

An important tax scale optimization problem is to ensure the highest total income. A possible solution is connected with optimizing taxpayers breakdown by the value of coefficients n_i and selecting the number of levels m. It can be shown that since the values of coefficients η_j are included in formula (13) with their weight coefficients $(j-1)[1+(j-1)k]$, the possible maximum value of $\alpha_{\max}$ depends on dividing the values n_i into groups.

The practical value of the proposed mechanism of transition to progressive taxation scales and social return valuation is that taxation coefficients estimation can be used by state authorities when selecting a strategy of progressive income taxation.

4 Conclusion

The retrospective analysis of the effects of introducing a flat tax scale has shown that this reform has both advantages and disadvantages. The overview of the international practice demonstrates that most developed countries use progressive taxation. This paper is focused on two methods of building linear and nonlinear progressive tax scales, on the basis of which we calculate the social return emerging from redistribution of tax burden among the population. In conclusion we introduce a generalised mathematical model which allows to take into account various external factors win the process of transition to progressive taxation.

References:

1. Aliyev B.H., Kagirgadzhiyeva Z.K., Regarding the state regulation of the tax rate on personal income, *Finance and Credit*, Vol.26, 2010, pp. 10-14.

2. Ananishvilli Y., Patava V. *Taxes and macroeconomic balance: Laffer-Keynesian Synthesis*, Stockholm: CA&CC Press® Publishing, 2010, 142 p.

3. Antonova M.E. About the tax mechanism of solving social problems, *Finance*, Vol.1, 2009, pp. 38-41.

4. Balatsky E.O. Concerning the nature of inconsistence of the Russian fiscal system, *Society and Economy*, Vol.11/12, 2004, pp. 127-136.

5. Belolipetsky A.A., Gorelik V.A. *Economic and mathematical methods*, Moscow: Academy, 2010, 362 p.

6. Dadashev A.Z. About stimulating potential of tax policy, *Economist*, Vol.8, 2009, pp. 45-50.

7. Davydova L.V., Fokina O.G. The role of taxes in the formation economic growth strategy, *Finance and Credit*, Vol.28, 2004, pp. 7-9.

8. Kalinina O.V. Building a tax scale of income taxation with due regard for social orientation, *Proceedings of the Institutions of Higher Education*. Series "Economics, Finance and Management of Organization", ISUCT Publishing, Vol.02(08), 2011, pp. 27-34.

9. Lyutova I.I., Zaitseva S.S. *Fundamentals of theory and organization of state finance and taxation*, Moscow: NSHTU, 2007, 249 p.

10. Mishustin M.V. *Income tax will stay*, Interview in "Vesti 24", radio station "VestiFM", 20 November 2010.

11. Nazarov V. Five arguments against introducing a progressive income tax, *Forbes Russia*, 9 March, 2011.

12. Official website of political party "Fair Russia": http://www.spravedlivo.ru.

13. Panskov V.G. Budget-2001 is fraught with economic and social consequences, *Russian Economic Journal*, Vol. 10, 2000, pp. 3-11.

14. Popova L.V. and others. *Taxation systems of different countries*, Moscow: Delo i servis, 2010, 432 p.

15. Proceedings of the meeting of the expert group on fiscal policy in the framework of socio-economic development of the country up to 2020, official website: http://www.2020strategy.ru, available on June 2014.

16. Putin V. Discussion about the introduction of a progressive scale of personal income tax in Russia is possible, Interview to EIA "Prime" on 14.06.2013, http://www.1prime.ru, available on June 2014.

17. Regarding progressive income scale, "Modernizatsiya", official website: http://www.modern-rf.ru, available on June 2014.

18. Social status and standard of living of the Russian population in 2013, Moscow: Rosstat, 2013, 327 p.

19. Statistics data of Federal State Statistics Service, official website: http://www.gks.ru, available on June 2014.

20. Statistics data of Federal Tax Service, official website: http://www.nalog.ru/rn78, available on June 2014.

21. Tax Code of the Russian Federation. Parts 1 and 2. - 31.07.1998 № 146 – FL.

22. Worldwide tax statistics: analysis findings, official website: http://www.worldwide-tax.com, available on June 2014.

Multivariate k-Nearest Neighbors Distribution Function Estimates in Dose-effect Relationship

Mikhail Tikhov, Maxim Ivkin
Department of Applied Probability Theory,
Nizhny Novgorod State University, Nizhny Novgorod, Russia
tikhovm@mail.ru, ivkin_max@mail.ru

***Abstract*—** Asymptotical normality of the *k*-nearest neighbors estimates (*kNN*-estimates) are proved by method of characteristic functions. The behavior of *kNN*-estimates is compared with that of kernel estimates.

***Keywords*—** dose-effect model, multivariate data, nonparametric kernel estimates.

I. Introduction

Let $(\boldsymbol{U}_1,W_1),(\boldsymbol{U}_2,W_2),...,(\boldsymbol{U}_n,W_n)$ be independent, identically distributed random (*d*+1)-vectors where $\{\boldsymbol{U}_i\}, 1\le i\le n,$ is *d*-vectors (we will consider a case $d\ge 2$) with bounded continuous density $f(\boldsymbol{x})$, $W_i = I(\boldsymbol{X}_i < \boldsymbol{U}_i)$ is the indicator of an event $(\boldsymbol{X}_i < \boldsymbol{U}_i)$, *d*-vectors $\boldsymbol{X}_i$ has distribution function $Q(\boldsymbol{x})$ and continuous density $q(\boldsymbol{x})>0$. The problem is to estimate the distribution function $Q(\boldsymbol{x})$ from the sample $\mathsf{U}^{(n)} = \{(\boldsymbol{U}_i,W_i),\ \ 1\le i\le n\}$.

Usually as an estimate of $Q(\boldsymbol{x})$ nonparametric estimators are used.

In a case $d=1$ kernel estimators

$$\hat{Q}_n(x) = S_{2n}(x)/S_{1n}(x), \tag{1}$$

are applied, where $S_{jn}(x) = \frac{1}{nh}\sum_{i=1}^{n} W_i^{\,j-1} K\left(\frac{u_i - x}{h}\right), \quad j=1,2$ and the kernel $K(x)$ is nonnegative even function, and $\int K(x)\,dx = 1$. We have $\hat{F}_n(x) = 0$ if $S_{1n}(x) = 0$. As h we take $h = n^{-1/5}$.

S.S.Yang [1] proposed as an of the regression function $Q(x) = \mathbf{E}(W \mid U = x)$ the statistic $Q_n^*(x)$ defined

$$Q_n^*(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{i/n - F_n(x)}{h}\right) W_n^{[i]}, \tag{2}$$

where $F_n(x) = n^{-1}\sum_{i=1}^{n} I(U_i < x)$ is the empirical distribution function random variable (rv) U, identically distributed with rv's $U_1,U_2,...,U_n$. Let $U_n^{(1)} < ... < U_n^{(i)} < ... < U_n^{(n)}$ be order statistics, and $W_n^{[i]}$ pared with $U_n^{(i)}$ is called concomitant of the *i*-th order statistics in the sample $\mathsf{U}^{(n)}$.

Let $k = k(n)$ be a sequence of positive integers, and $\rho = \rho_n$ be the Euclidean distance between x and its *k*th nearest neighbor. The nearest neighbor estimate is

$$\tilde{Q}_n(x) = \tilde{S}_{2n}(x)/\tilde{S}_{1n}(x), \tag{3}$$

where $S_{jn}(x) = \frac{1}{n\rho}\sum_{i=1}^{n} W_i^{\,j-1} K\left(\frac{u_i - x}{\rho}\right)$.

Let's notice that the estimate (2) is also a nearest neighbor estimate, but now neighbors are defined in terms of distance based on the empirical distribution function.

The estimate $\hat{Q}_n(x)$ has variance

$\sigma_1^2 = Q(x)(1-Q(x))\|K\|^2 / f(x)(1+o(1/(nh)))$ (see, [2], [3]), where $\|K\|^2 = \int K^2(x)\,dx$, therefore, if the density $f(x) = 0$, then this case is better to use the estimate (2).

In this paper we study behavior of *multivariate k-nearest neighbors distribution function estimates*. Multivariate *kNN-estimates of density* have been considered in works ([4], [5]).

In considered work for an estimation of distribution function we use *kNN*-estimators and we show that they are consistent, asymptotically normal and asymptotically unbiased estimators.

Consider an estimate of $Q(\boldsymbol{x})$ given by

$$\hat{Q}_n(\boldsymbol{x}) = \frac{T_{2n}(\boldsymbol{x})}{T_{1n}(\boldsymbol{x})},$$

where

$$T_{2n}(\boldsymbol{x}) = \frac{1}{n\rho^d}\sum_{j=1}^{n} W_j K\left(\frac{\boldsymbol{U}_j - \boldsymbol{x}}{\rho}\right) = \frac{1}{n}\sum_{j=1}^{n} W_j K_\rho(\boldsymbol{U}_j - \boldsymbol{x}),$$

$$T_{1n}(\boldsymbol{x}) = \frac{1}{n\rho^d}\sum_{j=1}^{n} K\left(\frac{\boldsymbol{U}_j - \boldsymbol{x}}{\rho}\right) = \frac{1}{n}\sum_{j=1}^{n} K_\rho(\boldsymbol{U}_j - \boldsymbol{x}),$$

$\rho = \rho_n(\boldsymbol{x})$ is the Euclidean distance between $\boldsymbol{x}$ and *k*th nearest neighbor of $\boldsymbol{x}$ among the $\boldsymbol{U}_j$'s, $K(\boldsymbol{x})$ is a bounded integrable weight function with

$$\int K(\boldsymbol{u})\,d\boldsymbol{u} = 1, \tag{4}$$

$k = k(n)$ is a sequence of positive integer such that $k\to\infty$, $k/n\to 0$ as $n\to\infty$.

ISBN: 978-1-61804-251-4

II. Preliminary Results

Let us first consider the probability density $p(\boldsymbol{x})$ of the distance ρ between $\boldsymbol{x}$ and the kth nearest neighbor $\boldsymbol{x}$. Let $S_r = \{\boldsymbol{z} : \|\boldsymbol{z}-\boldsymbol{x}\| < r\}$, $G(r) = \mathbf{P}(S_r)$, and

$$G'(r) = \lim_{\delta\to 0}\frac{1}{\delta}\left[\int_{S_{r+\delta}} f(\boldsymbol{t})\,d\boldsymbol{t} - \int_{S_r} f(\boldsymbol{t})\,d\boldsymbol{t}\right] = \int_{\|\boldsymbol{x}-\boldsymbol{t}\|=r} f(\boldsymbol{t})\,d\sigma(\boldsymbol{t}),$$

where $\mathbf{P}$ is the probability measure with density f, and σ is the surface area of the sphere $\|\boldsymbol{x}-\boldsymbol{t}\| = r$, $c_d = \dfrac{\pi^{d/2} r^d}{\Gamma((d+2)/2)}$, $\beta_d = d \cdot c_d$.

Thus the density of ρ is

$$p_n(r) = \frac{n!}{(k-1)!(n-k)!} G^{k-1}(r)(1-G(r))^{n-k} G'(r).$$

The conditional $k-1$ observations $\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_{k-1}$ falling within the sphere about $\boldsymbol{x}$ whose radius is determined by $\boldsymbol{x}$ and $\boldsymbol{h}$, the remaining $(n-k)$ observations $\boldsymbol{V}_1, \boldsymbol{V}_2, ..., \boldsymbol{V}_{n-k}$ falling outside this sphere under the condition $\boldsymbol{h} = r$, is given as follows (see, [4])

$$p(\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{k-1}; \boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_{n-k}; \boldsymbol{h} | r) =$$
$$= \prod_{j=1}^{k-1}\left(\frac{f(\boldsymbol{y}_j)}{G(r)}\right)\prod_{l=1}^{n-k}\left(\frac{f(\boldsymbol{v}_l)}{1-G(r)}\right)\frac{f(\boldsymbol{h})}{G'(r)}, \qquad (5)$$

so that the $\boldsymbol{Y}_j$'s, $\boldsymbol{V}_l$'s and $\boldsymbol{h}$ are conditionally independent given $\boldsymbol{h} = r$ with respective marginal densities

$$\frac{f(\boldsymbol{y}_j)}{G(r)}, \ \frac{f(\boldsymbol{v}_l)}{1-G(r)}, \text{ and } \frac{f(\boldsymbol{h})}{G'(r)},$$

$\{\boldsymbol{y} : \|\boldsymbol{x}-\boldsymbol{y}\| < r\}$, $\{\boldsymbol{v} : \|\boldsymbol{x}-\boldsymbol{v}\| > r\}$, $\{\boldsymbol{h} : \|\boldsymbol{x}-\boldsymbol{h}\| = r\}$ where the conditional density of $\boldsymbol{h}$ given ρ is to be integrated with respect to the surface measure on the sphere of radius r about $\boldsymbol{x}$.

We are interested in computing moments of various functions of ρ. It is clear from what has been stated above that ρ has the same distribution as $G^{-1}(\xi)$, where ξ is the kth order statistic from an i.i.d. uniform (0,1) sample of size n. If we just assume f is bounded and continuous we have

$$G(r) = \int_{B_r} f(\boldsymbol{u})\,d\boldsymbol{u} = c_d f(\boldsymbol{x}) r^d +$$
$$+\int_{B_r}(f(\boldsymbol{u}) - f(\boldsymbol{x}))\,d\boldsymbol{u} = c_d f(\boldsymbol{x}) r^d + o(r^d)$$

as $r \downarrow 0$.

Then $t = G(r)$ it follows that when $f(\boldsymbol{x}) > 0$

$$(G^{-1}(t))^{\lambda} = r^{\lambda} = \left(\frac{t}{c_d f(\boldsymbol{x})}\right)^{\lambda/d} + o(t^{\lambda/d}).$$

In addition, from Theorem 1 [5] we have

$$G^{-1}(z) = \left[\frac{z}{c_0 f(\boldsymbol{x})}\right]^{1/d} -$$
$$-\left\{\frac{c_2 \nabla^2 f(\boldsymbol{x})}{2pc_0 f(\boldsymbol{x})}\right\}\cdot\left\{\frac{z}{c_0 f(\boldsymbol{x})}\right\}^{3/d} + o(z^{3/d})$$

under conditions

$$\int \|\boldsymbol{y}\|^2 K(\boldsymbol{y})\,d\boldsymbol{y} < \infty, \int K(\boldsymbol{y})\,d\boldsymbol{y} = 1,$$
$$\int y_i K(\boldsymbol{y})\,d\boldsymbol{y} = \int y_i y_j K(\boldsymbol{y})\,d\boldsymbol{y} = 0, i \neq j,$$
$$\int y_i^2 K(\boldsymbol{y})\,d\boldsymbol{y} > 0 \text{ for any } i.$$

III. Main Results

Let's consider a difference $\hat{Q}_n(\boldsymbol{x}) - Q(\boldsymbol{x})$. We have

$$\tau_n(\boldsymbol{x}) = \frac{T_{2n}(\boldsymbol{x})}{T_{1n}(\boldsymbol{x})} - Q(\boldsymbol{x}) = \frac{T_{2n}(\boldsymbol{x})}{T_{1n}(\boldsymbol{x})} - \frac{Q(\boldsymbol{x}) f(\boldsymbol{x})}{f(\boldsymbol{x})} =$$
$$= \frac{T_{2n}(\boldsymbol{x})(f(\boldsymbol{x}) - T_{1n}(\boldsymbol{x})) + T_{1n}(\boldsymbol{x})(T_{2n}(\boldsymbol{x}) - Q(\boldsymbol{x}) f(\boldsymbol{x}))}{T_{1n}(\boldsymbol{x}) f(\boldsymbol{x})} =$$
$$= \frac{T_{2n}(\boldsymbol{x})(f(\boldsymbol{x}) - T_{1n}(\boldsymbol{x}))}{T_{1n}(\boldsymbol{x}) f(\boldsymbol{x})} + \frac{(T_{2n}(\boldsymbol{x}) - Q(\boldsymbol{x}) f(\boldsymbol{x}))}{f(\boldsymbol{x})}.$$

If we show that $T_{1n}(\boldsymbol{x}) - f(\boldsymbol{x}) \xrightarrow[n\to\infty]{p} 0$ and

$T_{2n}(\boldsymbol{x}) - Q(\boldsymbol{x}) f(\boldsymbol{x}) \xrightarrow[n\to\infty]{p} 0$, then from Slutsky's theorem (see [11], p.388, Theorem A.102), owing to boundedness of $Q(\boldsymbol{x}) f(\boldsymbol{x})$ and using that $f(\boldsymbol{x}) \geq c_0 > 0$ we obtain the convergence $\tau_n(\boldsymbol{x})$ in probability to zero as $n \to \infty$: $\tau_n(\boldsymbol{x}) \xrightarrow[n\to\infty]{p} 0$ for every fixed $\boldsymbol{x}$. Besides, from this relation we receive the limiting distribution of $\tau_n(\boldsymbol{x})$.

Let's consider the characteristic function $\varphi_{1n}(\theta)$ of the statistic $T_{1n}(\boldsymbol{x})$. Let

$$\varphi_{1n}(\theta) = \mathbf{E}\left(\exp\left(i\theta T_{1n}(\boldsymbol{x})\right)\right).$$

From (3) we derive

$$\varphi_{1n}(\theta) = \int \left(\psi_{1n}(\theta, r)\right)^{k-1} \psi_{2n}(\theta, r)\left(\psi_{3n}(\theta, r)\right)^{n-k} p_n(r)dr,$$

where

$$\psi_{1n}(\theta, r) = \int_{\|\boldsymbol{y}-\boldsymbol{x}\|<r} \exp\left(\frac{i\theta}{n} Q(\boldsymbol{y}) K_r(\boldsymbol{y}-\boldsymbol{x})\right)\frac{f(\boldsymbol{y})}{G(r)}d\boldsymbol{y},$$

$$\psi_{2n}(\theta, r) = \int_{\|\boldsymbol{t}-\boldsymbol{x}\|=r} \exp\left(\frac{i\theta}{n} Q(\boldsymbol{t}) K_r(\boldsymbol{t}-\boldsymbol{x})\right)\frac{f(\boldsymbol{t})}{1-G(r)}d\boldsymbol{t}$$

$$\psi_{3n}(\theta, r) = \int_{\|\boldsymbol{v}-\boldsymbol{x}\|>r} \exp\left(\frac{i\theta}{n} Q(\boldsymbol{v}) K_r(\boldsymbol{v}-\boldsymbol{x})\right)\frac{f(\boldsymbol{v})}{1-G(r)}d\boldsymbol{v}. \text{By} \quad (5),$$

the first term is equals

$$\psi_{1n}(\theta, r) = \frac{1}{G(r)}\int_{\|\boldsymbol{u}\|<r} \exp\left(\frac{i\theta}{n} K_r(\boldsymbol{u})\right) f(\boldsymbol{x}-\boldsymbol{u}r)\,d\boldsymbol{u}.$$

Lemma 1. *For every* $\alpha \in \mathbf{R}^1$ *and* $n \geq 0$,

$$\left|e^{i\alpha} - \sum_{k=0}^{n}\frac{(i\alpha)^k}{k!}\right| \leq \min\left\{\frac{|\alpha|^{n+1}}{(n+1)!}, \frac{2|\alpha|^n}{n!}\right\}.$$

Proof. If $\alpha > 0$, then integrating by pats a n times of the integral $\int_0^{\alpha}(\alpha - t)^n e^{it} dt$ we receive the inequality

$$\left| e^{i\alpha} - \sum_{k=0}^{n} \frac{(i\alpha)^k}{k!} \right| \le \frac{|\alpha|^{n+1}}{(n+1)!}.$$

Using the inequality $|e^{it}-1| \le 2$ and replacing in the previous inequality n by $n-1$ we obtain the second part of the inequality.

Now, in virtue of the condition (4) on the kernel $K(x)$, using the fact that $K(x) \le M_1$, $f(x) \le M_2$, applying the results of the Lemma 1, we conclude that

$$\left| \int_{\|u\|<r} \left(\exp\left(\frac{i\theta}{n} K_r(x-u) \right) - 1 - \frac{1}{n}\left(i\theta K_r(x-u)\right) - \right.\right.$$

$$\left.\left. - \frac{1}{2n^2}\left(i\theta K_r(x-u)\right)^2 \right) f(u)\,du \right| \le$$

$$\le \frac{1}{6n^3} \int_{\|u\|<r} |\theta|^3 \left(K_r(x-u)\right)^3 f(u)\,du =$$

$$= \frac{|\theta|^3}{6n^3 r^{2p}} \int_{\|u\|<1} K^3(u) f(x-ur)\,du \le$$

$$\le \frac{|\theta|^3 M_1^3 M_2}{6n^3 r^{2p}} = \frac{|\theta|^3 M_1^3 M_2 c_p^2 f^2(x)}{6nk^2}.$$

Hence,

$$\int_{\|u\|<r} \left(\exp\left(\frac{i\theta}{n} K_r(x-u) \right) - 1 - \frac{1}{n}\left(i\theta K_r(x-u)\right) - \right.$$

$$\left. - \frac{1}{2n^2}\left(i\theta K_r(x-u)\right)^2 \right) f(u)\,du = O(k^{-3}).$$

Further,

$$\int_{\|t\|=r} \left| \exp\left(\frac{i\theta}{n} K_r(t) \right) - 1 \right| \frac{f(x+t)}{G'(r)}\,dt \le$$

$$\le \frac{|\theta|}{n} \int_{\|u\|=1} K(u) \frac{f(x+ur)}{G'(r)}\,du = O\left(\frac{1}{n}\right),$$

thus $\psi_{2n}(\theta, r) \to 1$ as $n \to \infty$.

Similarly,

$$\left| \int_{\|u\|>r} \left(\exp\left(\frac{i\theta}{n} K_r(x-u) \right) - 1 - \right.\right.$$

$$\left.\left. - \frac{1}{n}\left(i\theta K_r(x-u)\right) \right) f(u)\,du \right| \le$$

$$\le \frac{\theta^2}{2n^2} \int_{\|u\|>r} \left(K_r(u-x)\right)^2 f(u)\,du =$$

$$= \frac{\theta^2}{2n^2 r^p} \int_{\|u\|>1} \left(K(u)\right)^2 f(x-ur)\,du \le$$

$$\le \frac{\theta^2 M_1^2 M_2}{2n^2 r^d} = \frac{\theta^2 M_1^2 M_2 c_d f(x)}{2nk} = O\left(\frac{1}{nk}\right) = o\left(\frac{1}{n}\right).$$

For further we use the results of works [4], [5] under the conditions of Theorem 1

$$\mathbf{E}(T_{1n}(x)) = f(x) + \frac{a_d}{(f(x))^{2/d}} P(f)(x) \left(\frac{k}{n}\right)^{2/d} +$$

$$+ \frac{c_d f(x)}{k} \int_{\|u\|=1} K(u)\,d\Sigma_0 + o\left(\left(\frac{k}{n}\right)^{2/p} + \frac{1}{k} \right),$$

where $P(f)(x) = \sum_{i,j} \int u_i u_j K(u)\,du\, D_i D_j f(x)$,

$H_f(x) = (D_i D_j f(x))$ – Hessian matrix,

$$\mathbf{D}(T_{1n}(x)) = \frac{c_d f^2(x)}{k} \int K^2(u)\,du + o\left(\frac{1}{k}\right),$$

$a_d = (\Gamma((d+2)/2))^{2/d} / (2\pi)$,

and, accordingly,

$$\frac{1}{n} \int_{\|v\|>r} K_r^2(v-x) \frac{f(v)}{1-G(r)}\,dv =$$

$$= \frac{c_d f^2(x)}{k} \int K^2(u)\,du + o\left(\frac{1}{k}\right).$$

Therefore uniformly in $\theta \in [-T, T]$, where T is a real number

$$\int \exp\left(i\theta n^{-1} K_r(u-x)\right) f(u)\,du =$$

$$= 1 + i\theta f(x) k^{-1} - (1/2)\theta^2 c_d f^2(x) \| K \|^2 k^{-1} +$$

$$+ o\left(k^{-1}\right), (n \to \infty); \ \| K \|^2 = \int K^2(u)\,du.$$

Decompose $\ln \varphi_{1n}(t)$ in a series on exponents

$i\alpha - \beta = i\theta \frac{f(x)}{k} - \frac{\theta^2}{2} \frac{c_d f^2(x) \| K \|^2}{k}$ to the second term and use the fact that for any real α, $\left|\ln(1+i\alpha) - i\alpha\right| \le \frac{\alpha^2}{2}$.

Really,

$$\left|\ln(1+i\alpha) - i\alpha\right| = \left| i \int_0^{\alpha} \left(\frac{1}{1+ix} - 1 \right) dx \right| \le$$

$$\le \int_0^{\alpha} \frac{x}{|1+ix|}\,dx \le \int_0^{\alpha} x\,dx = \frac{\alpha^2}{2},$$

since $|1+ix| = \sqrt{1+x^2} \ge 1$ and $\frac{1}{|1+ix|} \le 1$.

Let $\alpha = \theta \frac{f(x)}{k}$ and $\beta = \frac{\theta^2}{2} \frac{c_d f^2(x) \| K \|^2}{k}$.

Then

$$\ln(1+i\alpha-\beta) = \ln(1-\beta) + \ln\left(1 + \frac{i\alpha}{1-\beta}\right) =$$

$$= -\beta + \frac{i\alpha}{1-\beta} + O\left(k^{-2}\right) = -\beta + i\alpha + O\left(k^{-2}\right),$$

since $\beta^2 = O\left(k^{-2}\right)$, $\frac{\alpha^2}{(1-\beta)^2} \le \alpha^2 = O\left(k^{-2}\right)$, as $|\beta| < \frac{1}{2}$.

Hence

$$\left|\ln(1+i\alpha-\beta) - i\alpha + \beta\right| = O\left(k^{-2}\right).$$

Therefore, as $n \to \infty$,

$$\left| \ln\left(1 + i\theta f(x) k^{-1} - (1/2)\theta^2 c_d f^2(x) \| K \|^2 k^{-1}\right) - \right.$$

$$\left. - \left(i\theta f(x) k^{-1} - (1/2)\theta^2 c_d f^2(x) \| K \|^2 k^{-1}\right) \right| =$$

$$= O\left(k^{-2}\right) = o\left(k^{-1}\right),$$

ISBN: 978-1-61804-251-4

since a function $f(\boldsymbol{x})$ for a fixed $\boldsymbol{x}$ is bounded, and $\|K\|^2<\infty$, $\theta\in[-T,T]$, then $o(1)$ converges uniformly to zero.

Since $b_{1n}\,p_n(b_{1n}r+b_{0n})$, where b_{1n},b_{0n} – appropriate normalizing multipliers, converges uniformly on any bounded interval $(-C,C)$, $C>0$ to the density of the limit distribution (see. [15]), and, given that the probability of hitting $b_{1n}^{-1}(\rho-b_{0n})$ into intervals $(-\infty,-C],[C,\infty)$ tend to zero (see. [15])), and the function $\exp\left(i\theta f(\boldsymbol{x})k^{-1}-(1/2)\theta^2c_d f^2(\boldsymbol{x})\|K\|^2k^{-1}\right)$ is bounded, we have that

$$\mathbf{E}\left(\exp\left(it\sqrt{k}\left(T_{1n}(\boldsymbol{x})-f(\boldsymbol{x})\right)\right)\right)\to$$

$$\to\exp\left(-(1/2)\theta^2c_d f^2(\boldsymbol{x})\|K\|^2\right)$$

as $n\to\infty$. Now from convergence of characteristic functions follows, that

$$\sqrt{k}(T_{1n}(\boldsymbol{x})-f(\boldsymbol{x}))\underset{n\to\infty}{\overset{d}{\to}}N(0,c_d f^2(\boldsymbol{x})\|K\|^2),$$

Whence also follows that

$$T_{1n}(\boldsymbol{x})-f(\boldsymbol{x})\underset{n\to\infty}{\overset{p}{\to}}0. \tag{6}$$

We now show that

$$\sqrt{k}(T_{2n}(\boldsymbol{x})-Q(\boldsymbol{x})f(\boldsymbol{x}))\underset{n\to\infty}{\overset{d}{\to}}$$

$$\to N(0,c_d Q^2(\boldsymbol{x})f^2(\boldsymbol{x})\|K\|^2).$$

Let

$$\varphi_{2n}(\theta)=\mathbf{E}(\exp(i\theta T_{2n}(\boldsymbol{x})))=$$

$$=\mathbf{E}\left(\exp\left(i\theta n^{-1}\sum_{j=1}^{n}I(\boldsymbol{X}_j<\boldsymbol{U}_j)K_\rho(\boldsymbol{U}_j-\boldsymbol{x})\right)\right).$$

Passing on first to the conditional expectation provided $\boldsymbol{U}_j$, and then arguing as above with respect to the characteristic function of the statistics $T_{1n}(\boldsymbol{x})$, we obtain the following representation

$$\varphi_{2n}(\theta)=\int\left(\lambda_{1n}(\theta,r)\right)^{k-1}\lambda_{2n}(\theta,r)\left(\lambda_{3n}(\theta,r)\right)^{n-k}p_n(r)\,dr,$$

where

$$\lambda_{1n}(\theta,r)=\int_{\|\boldsymbol{y}-\boldsymbol{x}\|<r}\exp\left(\frac{i\theta}{n}Q(\boldsymbol{y})K_r(\boldsymbol{y}-\boldsymbol{x})\right)\frac{f(\boldsymbol{y})}{G(r)}d\boldsymbol{y},$$

$$\lambda_{2n}(\theta,r)=\int_{\|\boldsymbol{t}-\boldsymbol{x}\|=r}\exp\left(\frac{i\theta}{n}F(\boldsymbol{t})K_r(\boldsymbol{t}-\boldsymbol{x})\right)\frac{f(\boldsymbol{t})}{G'(r)}d\boldsymbol{t},$$

$$\lambda_{3n}(\theta,r)=\int_{\|\boldsymbol{v}-\boldsymbol{x}\|>r}\exp\left(\frac{i\theta}{n}Q(\boldsymbol{v})K_r(\boldsymbol{v}-\boldsymbol{x})\right)\frac{f(\boldsymbol{v})}{1-G(r)}d\boldsymbol{v}.$$

Repeating the above arguments, but for functions $\lambda_{1n}(\theta,r)$, $\lambda_{2n}(\theta,r)$, $\lambda_{3n}(\theta,r)$ and $\varphi_{2n}(\theta)$, we find that

$$\sqrt{k}(T_{2n}(\boldsymbol{x})-Q(\boldsymbol{x})f(\boldsymbol{x}))\underset{n\to\infty}{\overset{d}{\to}}N(0,c_d Q^2(\boldsymbol{x})f^2(\boldsymbol{x})\|K\|^2).$$

Thus we have the following result.

Theorem 1. *Let the density be bounded and there is a third continuous bounded partial derivatives $f(\boldsymbol{x})$ and $Q(\boldsymbol{x})$, $\int\|\boldsymbol{u}\|^2K(\boldsymbol{u})\,d\boldsymbol{u}<\infty$, $\int\boldsymbol{u}K(\boldsymbol{u})\,d\boldsymbol{u}=\mathbf{0}$.*

Then

(i) $\sqrt{k}\left(T_{1n}(\boldsymbol{x})-f(\boldsymbol{x})\right)\underset{n\to\infty}{\overset{d}{\to}}N(0,c_d f^2(\boldsymbol{x})\|K\|^2)$,

(ii) $\sqrt{k}\left(T_{2n}(\boldsymbol{x})-Q(\boldsymbol{x})f(\boldsymbol{x})\right)\underset{n\to\infty}{\overset{d}{\to}}$

$$\to N(0,c_p f^2(\boldsymbol{x})Q^2(\boldsymbol{x})\|K\|^2).$$

The following theorem establishes the asymptotic normality of the estimator $\tilde{F}_n(x)$ of the distribution function $Q(\boldsymbol{x})$.

Theorem 2. *Let the conditions of Theorem 1 hold.*

Then

$$\sqrt{k}\left(\hat{Q}_n(\boldsymbol{x})-Q(\boldsymbol{x})\right)\underset{n\to\infty}{\overset{d}{\to}}N(0,Q(\boldsymbol{x})(1-Q(\boldsymbol{x}))\|K\|^2).$$

Proof. Let $T_1=T_{1n}(\boldsymbol{x})$, $T_2=T_{2n}(\boldsymbol{x})$,

$$Qf=Qf(\boldsymbol{x})=Q(\boldsymbol{x})f(\boldsymbol{x}),\ f=f(\boldsymbol{x}).$$

We have:

$$\frac{T_2}{T_1}=\frac{T_2-Qf+Qf}{T_1-f+f}=\frac{(T_2-Qf)+Qf}{f\left(1+\dfrac{T_1-f}{f}\right)}=$$

$$=\frac{T_2-Qf}{f}\left(1-\frac{T_1-f}{f}+O_p\left(\frac{(T_1-f)^2}{f^2}\right)\right)+$$

$$+\frac{Qf}{f}\left(1-\frac{T_1-f}{f}+O_p\left(\frac{(T_1-f)^2}{f^2}\right)\right),$$

since

$$\left|\frac{1}{1+x}-1+x\right|=\left|\frac{x^2}{1+x}\right|\le 2x^2=O\left(x^2\right),\ |x|\le\frac{1}{2}$$

and $T_1-f\underset{n\to\infty}{\overset{p}{\to}}0$, $T_2-Qf\underset{n\to\infty}{\overset{p}{\to}}0$.

Hence

$$\frac{T_2}{T_1}-\frac{Qf}{f}=\frac{T_2-Qf}{f}-\frac{Qf}{f^2}(T_1-f)+$$

$$+O_p\left(\frac{(T_2-Qf)(T_1-f)}{f^2}\right)+O_p\left(\frac{Qf\,(T_1-f)^2}{f^3}\right).$$

Arguing as in [16], [17] with respect to the statistics $T_{1n}(\boldsymbol{x})$ and $T_{2n}(\boldsymbol{x})$, it can be shown that with probability 1

$$\overline{\lim_{n\to\infty}}\sup_{\frac{c\ln n}{n}\le h\le 1}\frac{\sqrt{nh}\left\|T_1-\mathbf{E}(f)\right\|_\infty}{\sqrt{\max\left(\ln(1/h),\ln\ln n\right)}}=k_1(c)<\infty,$$

$$\overline{\lim_{n\to\infty}}\sup_{\frac{c\ln n}{n}\le h\le 1}\frac{\sqrt{nh}\left\|T_2-\mathbf{E}(Qf)\right\|_\infty}{\sqrt{\max\left(\ln(1/h),\ln\ln n\right)}}=k_2(c)<\infty.$$

where for sufficiently large n,

$$\left\|\left(\frac{T_2}{T_1}-F\right)-\frac{T_2-Qf}{f}+\frac{Qf}{f^2}(T_1-f)\right\|_\infty\le C_1\frac{\ln n}{k}.$$

ISBN: 978-1-61804-251-4

Thus,

$$\sqrt{k}\left\|\left(\frac{T_2}{T_1}-F\right)-\frac{T_2-Qf}{f}+\frac{Qf}{f^2}(T_1-f)\right\|_\infty \xrightarrow[n\to\infty]{p} 0.$$

Further,

$$\mathbf{D}\left(\frac{T_2}{T_1}-Q\right)=\left(\frac{1}{f^2}\mathbf{D}(T_2-Qf)+\frac{(Qf)^2}{f^4}\mathbf{D}(T_1-f)-\right.$$

$$\left.-2\frac{Qf}{f^3}\mathbf{Cov}(T_1-f,T_2-Qf)\right)\left(1+O_p\left(\frac{\ln^2 n}{k^2}\right)\right)=$$

$$=\frac{1}{f^2}\mathbf{D}(T_2)+\frac{(Qf)^2}{f^4}\mathbf{D}(T_1)-$$

$$-2\frac{Qf}{f^3}\mathbf{Cov}(T_1,T_2)\left(1+O_p\left(\frac{\ln^2 n}{k^2}\right)\right)$$

as $n\to\infty$.

Consider the expectation $\mathbf{E}(T_1\cdot T_2)$.

We have

$$\mathbf{E}(T_1\cdot T_2)=\mathbf{E}\left(\frac{1}{n}\sum_{i=1}^n K_\rho(\boldsymbol{U}_i-\boldsymbol{x})\cdot\frac{1}{n}\sum_{i=1}^n W_i K_\rho(\boldsymbol{U}_i-\boldsymbol{x})\right)=$$

$$=\frac{1}{n^2}\mathbf{E}\left(\sum_{i=j=1}^n W_i\left(K_\rho(\boldsymbol{U}_i-\boldsymbol{x})\right)^2+\right.$$

$$\left.+\sum_{i\neq j}K_\rho(\boldsymbol{U}_i-\boldsymbol{x})W_j K_\rho(\boldsymbol{U}_j-\boldsymbol{x})\right).$$

By virtue of independent and identically distributed pairs $(\boldsymbol{U}_1,W_1),...,(\boldsymbol{U}_n,W_n)$, we conclude that

$$\mathbf{E}(T_1\cdot T_2)=\frac{1}{n}\mathbf{E}\left(W_1\left(K_\rho(\boldsymbol{U}_1-\boldsymbol{x})\right)^2\right)+$$

$$+\frac{n-1}{n}\mathbf{E}\left(W_1K_\rho(\boldsymbol{U}_1-\boldsymbol{x})\right)\mathbf{E}\left(K_\rho(\boldsymbol{U}_2-\boldsymbol{x})\right)=$$

$$=\frac{1}{n}\int\mathbf{E}\left(I(\boldsymbol{U}_1>\boldsymbol{X}_1)K_\rho^2(\boldsymbol{U}_i-\boldsymbol{x})\middle|\boldsymbol{U}_1=\boldsymbol{u}\right)f(\boldsymbol{u})\,d\boldsymbol{u}+$$

$$+\frac{n-1}{n}\int\mathbf{E}\left(I(\boldsymbol{X}_1<\boldsymbol{U}_1)K_\rho(\boldsymbol{U}_1-\boldsymbol{x})\middle|\boldsymbol{U}_1=\boldsymbol{u}\right)f(\boldsymbol{u})\,d\boldsymbol{u}\times$$

$$\times\int K_\rho(\boldsymbol{u}-\boldsymbol{x})\,f(\boldsymbol{u})\,d\boldsymbol{u}.$$

Making the replacement $\boldsymbol{z}=(\boldsymbol{u}-\boldsymbol{x})r^{-1}$, we have

$$\mathbf{E}(T_1\cdot T_2)=(nr^d)^{-1}\int K^2(\boldsymbol{u}-\boldsymbol{x})Q(\boldsymbol{z}r+\boldsymbol{x})\times$$

$$\times f(\boldsymbol{z}r+\boldsymbol{x})\,d\boldsymbol{z}+(r^{2d})^{-1}(1-n^{-1})\times$$

$$\times\left(\int K(\boldsymbol{z})Q(\boldsymbol{z}r+\boldsymbol{x})f(\boldsymbol{z}r+\boldsymbol{x})\,d\boldsymbol{z}\right)\times$$

$$\times\left(\int K(\boldsymbol{z})f(\boldsymbol{z}r+\boldsymbol{x})\,d\boldsymbol{z}\right),$$

$$(nr^d)^{-1}\int K^2(\boldsymbol{u}-\boldsymbol{x})Q(\boldsymbol{z}r+\boldsymbol{x})f(\boldsymbol{z}r+\boldsymbol{x})d\boldsymbol{z}=$$

$$=k^{-1}Q(\boldsymbol{x})f^2(\boldsymbol{x})\|K\|^2+o(k^{-1}).$$

From the conditions on the kernel $K(\boldsymbol{x})$ and the conditions on the functions $Q(\boldsymbol{x})$, $f(\boldsymbol{x})$, we have

$$\int K(\boldsymbol{z})Q(\boldsymbol{z}r+\boldsymbol{x})f(\boldsymbol{z}r+\boldsymbol{x})\,d\boldsymbol{z}=Q(\boldsymbol{x})f(\boldsymbol{x})+o(k^{-1}).$$

Thus,

$$\mathbf{E}(T_1\cdot T_2)=Q(\boldsymbol{x})f^2(\boldsymbol{x})\|K\|^2k^{-1}+$$

$$+(1-n^{-1})\left(Q(\boldsymbol{x})f(\boldsymbol{x})+f(\boldsymbol{x})\right)+o(k^{-2}),$$

$$\mathbf{Cov}(T_1,T_2)=Q(\boldsymbol{x})f^2(\boldsymbol{x})\|K\|^2k^{-1}+O(k^{-2}).$$

So, as $n\to\infty$,

$$\mathbf{D}(\hat{Q}_n(\boldsymbol{x})-Q(\boldsymbol{x}))=Q(\boldsymbol{x})(1-Q(\boldsymbol{x}))\|K\|^2k^{-1}(1+o(1)).$$

Hence we conclude that

$$\sqrt{k}(\tilde{Q}_n(\boldsymbol{x})-Q(\boldsymbol{x}))\xrightarrow[n\to\infty]{d}\varsigma\in N(0,Q(\boldsymbol{x})(1-Q(\boldsymbol{x}))\|K\|^2).$$

REFERENCES

[1] Yang S. "Linear function of concomitants of order statistics with application to nonparametric estimation of a regression function." *Journal Amer. Statist. Assoc*, vol. 76, pp. 658-662, 1981.

[2] Tikhov M.S. "Statistical Estimation based on Interval Censored Data." Param. and Semiparam. Models with Appl. to Rel., Surv. Analisys, and Qual. of Life: Springer-Verlag: Theor.& Meth, pp.209-215, 2004.

[3] Tikhov M.S. "Statistical Estimation on the Basis of Interval-Censored Data." *Journal Math. Sciences*, vol. 119, no.3, pp. 321-335, 2004.

[4] Mack Y.P., Rosenblatt M. "Multivariate *k*-Nearest Neighbor Density Estimates." *Journal of Multivar. Analysis*, vol. 9, pp. 1-15, 1979.

[5] Hall P. "On Near Neighbour Estimates of a Multivariate Density," *Journal of Multivar. Analysis*, vol.13, pp. 24-39, 1983.

[6] Zinde-Walsh V. "Consequences of lack of smoothness in nonparametric estimation." *Quantil*, no 4, pp. 57-69, 2008.

[7] Kotlyarova Y., Zinde-Walsh V. "Non and semi-parametric estimation in models with unknown smoothness." *Economic Letters*, vol. 93, pp. 379-386, 2006.

[8] Kotlyarova Y., Zinde-Walsh V. "Robust kernel estimator for densities of unknown smoothness." *Journal of Nonparametric Statistics*, vol. 19, pp. 89-101, 2006.

[9] Krishtopenko S.V., Tikhov M.S. "Statistical estimation of the effective dose in dose-effect dependence using both direct and indirect observations." *2th All-Russian School-Colloquim on Stochastic Method*: M.: TVP, pp. 81-82, 1995.

[10] Krishtopenko S.V., Tikhov M.S. , Popova E.B. "Dose-effect." M.: Medicina, 2008, 228 p.

[11] Tikhov M.S. "Statistical estimation in dose-effect dependence with the help of the *kNN*-estimates." *Surveys in applied and industrial mathematics*. M.: TVP, vol.12, no 3, pp. 683-684, 2005.

[12] Tikhov M.S. "Asymptotical normality of *kNN*-estimates in dose-effect dependence." *Nizhny Novgorod State University Bulletin. Series: Mathematika*, no.1(4), pp. 129-137, 2006.

[13] Tikhov M.S. "Linear functions of the induced order statistics and nonparametric estimation of distributions in dose-effect dependence." *Surveys in applied and industrial mathematics*, vol. 6, no 1, pp. 234, 1999.

[14] Rao C.R. "Linear statistical inference and its applications." New York: John Wiley & Sons, 1965, 620 p.

[15] Smirnov N.V. "Theory Probability and Mathematical Statistics: selected works." M.: Nauka, 1970, 289 p.

[16] Einmahl U., Mason D.M. "Uniform in Bandwidth Consistency of Kernel-type Function Estimators." *Ann. Statist.*, vol. 33, no.3, pp.1380-1403, 2005.

[17] Tikhov M.S. "Nonparametric estimation of effective doses at quantal response." *Ufa math. journal*, vol. 5, no. 2, pp. 94-108, 2013.

A Fast Heuristics for Inferring Approximately Minimal Diagnostic Tests

Xenia Naidenova, Vladimir Parkhomenko, Alexander Rudenko

Abstract—A class of machine learning algorithms based on mining approximate classification tests is considered. A system called "DEFINE" of analogical reasoning based of mining these classification (diagnostic) tests is described. DEFINE is developing specially for machine learning problems in agriculture and motor industry. Some examples of the system application are given.

Keywords—Approximate classification test, reasoning by analogy, machine learning.

I. Introduction

THIS paper provides a method of inferring approximate diagnostic (classification) tests. Considering the sets of approximately minimal diagnostic tests as a characteristics portraits of object classes we have developed a model of reasoning by analogy. This model is implemented in the system called DEFINE. Approximately-minimal tests are designed to classify as many examples as it is possible according to the algorithm. They also negate all contradictory examples. The results of this systems application for predicting the type of tree species with the use of aerial photographs is described. The prediction of defects in rotating equipment is also considered.

II. A model of reasoning by analogy based on the sets of AMDTs

Approximately minimal or quasi minimal diagnostic test (AMDT) distinguishing an example e from all examples of class Q_x is a collection of attributes $\{A_1, A_2, \ldots, A_k\}$ such that e differs from any example of Q_x by value of at least one attribute of this collection. There are a plethora of algorithms for searching for tests, however if a certain algorithm is chosen then it is possible to consider it as a function $\phi(e, Q_x) = A_1, A_2, \ldots, A_k$. This function possesses the property that for familiar examples it will return the familiar or the same tests. Let T_{ij} be the set of tests such that any example $e \in Q_i$ is different from all examples of Q_j by at least one test of T_{ij} and for every test $t \in T_{ij}$ there is an example e such that it is different from all examples of Q_j only by this test. In other words, T_{ij} is the necessary and sufficient set of tests for distinguishing Q_i and Q_j. The set T_{ij} is also a function $f(Q_i, Q_j)$ determined by a certain test construction algorithm. The set T_{ij} is considered to be stable or changeable insignificantly with respect to different collections of examples from the same class Q_i. Let T_{ij} be the set of tests distinguishing the sets Q_i, Q_j of examples. Let T_{xj} be the set of tests distinguishing the sets Q_x, Q_j of examples, where Q_x, Q_i are taken from the same sampling (class) of examples. We assume that tests of T_{ij} and T_{xj}, completely coincide or at least greatly intersect. Analogical reasoning is defined as follows [1], [2]. Assume that the sets T_{ij} for all training sets Q_i, Q_j of examples, $i, j \in \{1, 2, \ldots, nk\}$, where nk is the number of classes, have been obtained by a certain algorithm. Let Q_x be a subset of examples belonging to one and the same but unknown class $x \in \{1, 2, \ldots, nk\}$. Construct sets of tests, $T_{xj} = f(Q_x, Q_j)$, $j \in \{1, 2, \ldots, nk\}$. If examples of Q_x belong to class $k \in \{1, 2, \ldots, nk\}$, then, in accordance with our assumption, the set of tests T_{xj} must be more similar to T_{kj}, than to T_{ij} for all $i \neq k$, $i \in \{1, 2, \ldots, nk\}$. This method can be considered as inference by analogy because we use the assumption of analogical properties of tests for similar examples constructed with the use of one and the same functional transformation (algorithm). The main problem of this method is related to the choice of the criterion or the measure of similarity between sets of tests. It is more reliable to use several criteria and to make decision based on the rule of voting between these criteria. We give an example of Inference by Analogy. Here Q is a set of training examples partitioned into 5 disjoint classes. The examples are described by 9 attributes (see, please, Tab.I).

TABLE I
The Set of Training Examples

No\Attr	1	2	3	4	5	6	7	8	9	Q
1	1	1	1	1	1	1	1	1	1	Q_1
2	1	1	2	2	1	1	2	1	1	Q_1
3	2	1	2	2	1	1	3	2	4	Q_1
4	2	1	2	4	1	1	3	2	4	Q_1
5	1	2	1	1	2	1	1	2	1	Q_2
6	1	2	3	3	2	1	3	1	1	Q_2
7	2	2	1	2	1	2	4	2	2	Q_2
8	2	3	4	4	1	1	3	2	2	Q_3
9	2	1	4	4	1	1	4	2	3	Q_3
10	3	4	5	3	3	3	1	3	4	Q_4
11	4	5	5	5	2	2	4	2	1	Q_4
12	3	5	5	1	2	1	5	2	5	Q_4
13	3	4	3	5	4	4	2	4	1	Q_5
14	4	2	3	3	2	4	2	2	1	Q_5

Tab.II contains a list of quasi-minimal tests for all pairs Q_i, Q_j, $i, j \in \{1, 2, 3, 4, 5\}$of classes. Let two examples be represented for predicting the class to which they belong to: $Q_x = \{(3, 4, 3, 5, 3, 4, 2, 1, 1),(4, 4, 3, 3, 2, 4, 2, 3, 5)\}$. Tab.III contains quasi-minimal tests distinguishing collection of examples Q_x from the sets of examples Q_1, Q_2, Q_3, Q_4, Q_5.

Compare the sets of tests for every pair $(Q_x Q_j)$, $j \in \{1, 2, 3, 4, 5\}$ with the sets of tests T_{ij} for all training sets Q_i, Q_j of examples, $i, j \in 1, 2, \ldots, 5$ (Tab.II). One of the possible decision rules says that if Q_x and Q_y are taken

ISBN: 978-1-61804-251-4

TABLE II
THE SETS OF TESTS FOR GIVEN CLASSES OF OBJECTS

Pairs of Q	Tests T_{ij}	No of tests
$Q_1 - Q_1$	4,7	1
$Q_1 - Q_2$	2	1
$Q_1 - Q_3$	3, 9	2
$Q_1 - Q_4$	1, 2, 3, 5	4
$Q_1 - Q_5$	1, 2, 3, 4, 5, 6	6
$Q_2 - Q_1$	2	1
$Q_2 - Q_2$	4, 7	2
$Q_2 - Q_3$	2, 3, 4	3
$Q_2 - Q_4$	1, 2, 3	3
$Q_2 - Q_5$	1, 6, 7	3
$Q_3 - Q_1$	3, 9	2
$Q_3 - Q_2$	2, 3, 4	3
$Q_3 - Q_3$	2, 7, 9	3
$Q_3 - Q_4$	1, 2, 3, 4,5, 9	6
$Q_3 - Q_5$	1, 2, 3, 4,5, 6, 7, 9	8
$Q_4 - Q_1$	1, 2, 3, 5	4
$Q_4 - Q_2$	1, 2, 3	3
$Q_4 - Q_3$	1, 2, 3, 4,5, 9	6
$Q_4 - Q_4$	4, 6, 7, 9	4
$Q_4 - Q_5$	6, 7	2
$Q_5 - Q_1$	1, 2, 3, 4, 5, 6	6
$Q_5 - Q_2$	1, 6, 7	3
$Q_5 - Q_3$	1, 2, 3, 4,5, 6, 7, 9	8
$Q_5 - Q_4$	6, 7	2
$Q_5 - Q_5$	8	1

TABLE III
THE SETS OF TESTS FOR EXAMPLES OF UNKNOWN CLASS

Pairs of Q	Tests T_{ij}	No of tests
$Q_x - Q_1$	1, 2, 3, 4, 5, 6	6
$Q_x - Q_2$	1, 2, 3, 6, 7	5
$Q_x - Q_3$	1, 2, 3, 4, 5, 6,7, 8,9	9
$Q_x - Q_4$	6,7	2
$Q_x - Q_5$	8,9	2

from the same class of examples, then the intersection of corresponding sets of tests $\{Q_xQ_j, Q_yQ_j\}$, $j \in \{1, 2, 3, 4, 5\}$ must be greater than for Q_xQ_z, for all $z \neq y$. The intersection of two sets of tests is referred to as the subset of coincident tests in these sets. The illustration of this decision rule is given in Tab.IV. Using this decision rule, we can conclude that Q_x and Q_5 are included in the same class. Let us remark that this example is taken from the real task, and the set Q_x of objects is predicted correctly.

III. DEFINE: THE SYSTEM FOR ANALOGICAL REASONING

The system DEFINE [3], [2] is based on the machine learning method described above. Principal structure of DEFINE is shown in Fig.1.

Describe the role of each program of the system. The block RDING is intended for inputting initial data and its transformation into attribute-value representation. In particular, if features of objects are continuous, then their discretization

TABLE IV
THE APPLICATION OF DECISION RULE TO THE SET Q_x

No	Q_1-Q_1	Q_2-Q_1	Q_3-Q_1	Q_4-Q_1	Q_5-Q_1
Q_x-Q_1	0	1	1	4	6
	Q_1-Q_2	Q_2-Q_2	Q_3-Q_2	Q_4-Q_2	Q_5-Q_2
Q_x-Q_2	1	1	2	3	3
	Q_1-Q_3	Q_2-Q_3	Q_3-Q_3	Q_4-Q_3	Q_5-Q_3
Q_x-Q_3	2	3	3	6	8
	Q_1-Q_4	Q_2-Q_4	Q_3-Q_4	Q_4-Q_4	Q_5-Q_4
Q_x-Q_4	0	0	0	2	2
	Q_1-Q_5	Q_2-Q_5	Q_3-Q_5	Q_4-Q_5	Q_5-Q_5
Q_x-Q_5	0	0	1	0	1
$\sum$	3	4	7	15	20

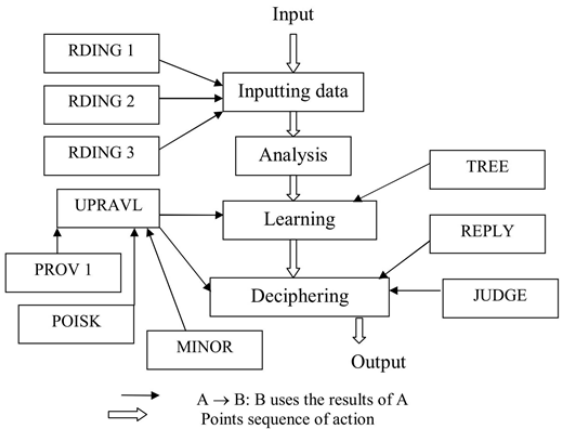


Fig. 1. The Structure of the System DEFINE

is required. In the case of using images of objects, the special complex of programs is used for calculating the values of some characteristics of images, extracting objects, calculating features of objects, and transforming them into attribute-value representation. Program SLOT forms the training and controlling sets of objects based on a given rule or a random choice. The programs UPRAVL, PROV1, POISK and MINOR serve for constructing the sets of tests T_{ij} distinguishing the sets Q_i, Q_j of object examples. Program TREE serves for compact representation of the set T_{ij} in the form of special structure vector-tree. This structure allows quickly checking whether a test t is contained in T_{ij}. Tests in the form of trees require less volume of memory space. Block Learning constructs the set of tests $T = \{T_{ji}\}$ and transforms them into the form of trees. Block Deciphering constructs the set $T = \{T_{xi}\}$ for unknown or control collection of examples. Program REPLY realizes several decision rules for estimating the degree of similarity between the sets Txi, T_{ji}, $j \in \{1, 2, \ldots, nk\}$, $i = 1, 2, \ldots, nk$. Program JUDGE performs the final decision. Block Analysis investigates initial data and gives the information about the degree of similarity and distinction between given classes of objects and some others informative characteristics. Algorithm TREE serves for transforming the test matrix into the structure of vector-tree or Decision Tree Matrix. Tests are lexicographically ordered and they are represented as the branches of an ordered decision

ISBN: 978-1-61804-251-4

TABLE V
DECISION TREE MATRIX

Decisions	Ordered decisions	Decision tree
2, 8	1, 3, 2	1 — 3 — 2
1, 3, 2	2, 8	2 — 8 —
2, 8, 15	2, 8, 15	— — 15
7, 5	3, 4, 5	3 — 4 — 5
7, 2, 1	7, 2, 1	7 — 2 — 1
3, 4, 5	7, 5	5

TABLE VI
THE VECTOR-TREE REPRESENTATION FOR THE TREE OF TAB.V

No	1	2	3	4	5	6	7	8	9	10	11
Node	1	8	3	0	2	0	0	2	16	8	0
No	12	13	14	15	16	17	18	19	20	21	22
Node	-1	15	0	0	3	23	4	0	5	0	0
No	23	24	25	26	27	28	29	30	31	32	
Node	7	0	2	30	1	0	0	5	0	0	

tree, an example of which is given in Tab.V.

The structure of tree is represented in the form of vector, the example of which, for the tree of Tab.V, is given in Tab.VI.

Generally, the structure of tree is determined as follows: If i-th component of vector-tree contains the value of a node, then $(i+1)$-th component:

1) contains the index of component containing the value of next node of the same level of tree if such a node exists;
2) is equal to 0 if such a node is absent;

(i + 2)-th component of vector-tree contains:

1) the value of next node of the same branch if such an element exists;
2) 0, if the next node of the same branch is absent and there is not an offshoot of the considered node;
3) -1 if the next node of the same branch is absent but the offshoots of considered node are present.

If j-th component of vector-tree is equal -1, then the value of the next node of offshoot is contained in $(j+1)$-th component of vector-tree. The first element of vector-tree contains the value of the first node of decision tree at the first level.

The first version of DEFINE has been implemented in FORTRAN for running on EC. The second version of DEFINE has been realized in Turbo C 2.0 DOS 3.0 on computers PC AT/XT and compatible ones with Video adapter CGA or emulating regime CGA. The module Define.exe has been 52 kb executable module. The current DEFINE version is developing for the prediction of defects in rotating equipment. Let us briefly discuss the problem.

A. Software for predicting defects in rotating equipment

We use the spectral information from a vibration detector to predict the defects of rotating equipment. The number of spectral parameters is determined based on the frequency domain and step by frequency.

Let the resulting spectral information be an object. Then the type of defect is referred to a class. The set of objects without defects can also be regarded as a class. The number of objects of each class in the training sample is determined by the required precision of the control deciphering. The number of objects increases until the necessary quality of deciphering is achieved.

The main part of program is implemented in the language PHP. This script language uses HTML, JavaScript and AJAX to process the input data. Inside the program, the data is stored in MySQL database. To make a desktop application, all scripts are packed in the shell written in Delphi.

We plan to make a web-oriented application to get a feedback from the potential users. The program will be independent from the operation system platform, i.e. the internet and web-browser are required. There is another advantage of PHP implementation. It is associated with the high speed of development. The program consists of four main components called Preferences, Data, Analysis and Learning.

Preferences component helps to set the frequency domain and step by frequency. The minimal number of learning steps is calculated here.

In this part, the frequency range and the step by frequency are selected. It is necessary in order to determine the number of parameters which will be used for predicting the defects. Here the prediction accuracy is calculated on the basis of which the minimum number of learning stages is determined.

Data component has a tool for editing the data of defects, i.e. one can add and delete the items from the list of defects. There is also a tool for editing the learning sets within the base of spectral signals. The component is closely related with a MySQL database.

Learning component takes the information from the data base, mines the logical rules (AMDTs) to predict the classes of objects.

Analysis component realizes the process of machine learning. It uses the results from the previous components and has the following functions:

- logical rules (AMDTs) mining between objects of unknown class and objects of learning classes;
- the rules from the previous step are compared with the rules from Learning component;
- predicting the class of unknown objects.

IV. THE RESULTS OF DEFINES APPLICATION

The system DEFINE has been used for deciphering the predominant species of trees based on aero photographs with the scale 1 : 3000 [3]. The forest parts have been picked out in the Chagodotchenskij forestry of Vologda region. The following types of trees have been chosen: pine-tree, aspen, birch, and fir-tree. The class of pine-trees has been partitioned into two subclasses: pine-tree 1 the trees of 70 years old, and pine-tree 2 the trees of 115 years old. For training and controlling sets of samples, the trees that are well predicted through stereoscope have been picked out with space distribution approximately equal to 15 -20 trees par 4-5 hectares. Images of trees have been analyzed by using the stereoscope. An operator has estimated visually the following set of photometrical and texture properties of trees: color of illuminated part of crown (1), color of shaded part of crown (2), form of projection

ISBN: 978-1-61804-251-4

TABLE VII
DECIPHERING THE SPECIES OF TREES (METHOD 2)

Tree type	Birch	Pine-tree1	Pine-tree2	Aspen	Fir-tree
Birch	100%				
Pine-tree1		100%			
Pine-tree2		22%	78%		
Aspen				100%	
Fir-tree					100%

of crown (3), form of the edge of projection (4), form of illuminated part of crown (5), form of shaded part of crown (6), structure of crown (7), texture of crown (8), passage from the illuminated to the shaded part crown (9), density of crown (10), closeness of crown (11), form of branches (12), size of branches (13), form of apex (14), and convexity of crown. For evaluating the color, the scale of color standards has been used. For coding the other properties, the semantic scales have been developed. The number of gradations of properties on the semantic scales was within the limits from 3 to 12. 500 images of trees (100 trees for each species) have been analyzed independently by two operators. The training set of samples has contained 60 descriptions for each species of trees, the set of control samples has contained 40 descriptions for each species of trees. Two methods have been used for deciphering. The first method (Method 1) deals with predicting the species to which belongs a subset of control trees taken from an unknown class. The decision rule is based on predicting the number of completely coincident tests in the conformable matrixes of tests T_{xi}, T_{ji}, $j, i, x \in \{$birch, pine-tree 1, pine-tree 2, aspen, fir-tree$\}$. The totalities of control examples are considered belonging to the species of trees for which the sum of agreements is the greatest, i.e. the result is i-th species for which $\sum ||T_{xi} \cap T_{ji}||$, $j \in \{$birch, pine-tree 1, pine-tree 2, aspen, fir-tree$\}$ is maximal among all $i \in \{$birch, pine-tree 1, pine-tree 2, aspen, fir-tree$\}$. Deciphering the species of trees by Method 1 has given 100% true reply for each species. The second method (Method 2) has been implemented for predicting the species to which belongs a single sample of tree not belonging to training sets of trees. This methods is detailed in [3]. In Tab.VII the percentage of correct answers obtained with the use of Method 2 is given. In this case, the part of 22% of the pine trees of 70 years old has been predicted not correctly.

The analysis of the stability of tests has been also carried out. Tab.VIII contains the results of experiments according to the data of one of the operators. We observe the disappearance of some tests with decreasing the volume of training set of examples. The number of unique tests proves to be not great. Results demonstrate the possibility to decrease the volume of training set in subsequent experiments. The frequency of occurring attributes in tests shows the usefulness or their informative power. Attributes 10 and 15 did not enter any test, so they are the least informative. Attribute 2 possesses the greatest informative power. To the informative attributes belong also attributes 1, 3, 4, 5, 6, 7, 12, and 14.

The system DEFINE has been applied very successfully for processing spectral information [4]. We have made also some

TABLE VIII
ESTIMATION OF TEST STABILITY (THE SCALE 1 : 3000, OPERATOR 1)

The volume of training set	100%	60%		40%	
Repeated\Unique	R	R	U	R	U
Tree pairs to be deciphered					
Birch-Pine-tree1	4	3	-	2	-
Birch-Pine-tree2	3	3	-	1	-
Birch-aspen	12	10	-	10	1
Birch-fir-tree	2	1	-	2	-
Pine-tree1-Pine-tree2	11	8	1	8	-
Pine-tree1-aspen	3	3	-	2	-
Pine-tree1-fir-tree	5	5	-	3	-
Pine-tree2-aspen	7	6	-	4	-
Pine-tree2-fir-tree	4	4	2	2	-
Aspen-fir-tree	2	2	-	2	-

TABLE IX
RESULTS OF PREDICTING THE DEFECTS IN ROTATING EQUIPMENT

Spectral signals	Q_1	Q_2	Q_3
Defects in rolling bearings	83%	89%	81%

experiments for predicting the defects in rotating equipment. There were 100 spectral signals both for each class of object defect and for the normal object working without defects. The program has been tuned as follows: the frequency domain was from 0 to 100 Hz and the step by frequency was 1 Hz. The prediction accuracy was 85%. Three different spectral signals of defects have been given for predicting their classes. The results of the prediction see, please, in Tab.IX.

V. CONCLUSION

The method of classification reasoning presented in this paper provides a framework for solving diverse and very important problems of constructing machine learning algorithms based on a unified logical model in which it is used a mode of analogical commonsense reasoning [5]. The peculiarities of this model include the fast algorithms for constructing approximately minimal diagnostic tests and the use of training examples of objects to be identified during machine learning process.

REFERENCES

[1] K. Najdenova, "A relational model of the analysis of experimental data," *Engineering Cybernetics*, vol. 20, no. 4, pp. 99–115, 1982.

[2] X. A. Naidenova and J. G. Polegaeva, "DEFINE - the system for generating hypotheses and reasoning by analogy on the basis of inferring functional dependencies," in *The Problem of Expert System Creation. Preprint of the Leningrad Institute for Informatics and Automation of the USSR Academy of Sciences (LIIAN)*, V. Ponomarev, Ed. Leningrad, USSR: LIIAN, 1989, vol. 111, pp. 20–21, (in Russian).

[3] X. Naidenova, J. Krilova, and I. Gnedash, "Deciphering objects based on relations of distinction and identity of objects descriptiosn in multivalued feature spaces," in *"Applying distant data and computers for investigating the natural resources of the Earth". Preprint*, V. Ponomarev, Ed. LS-ICC (Leningrad Scientific Investigative Computer Centre) of cademy of Sciences of USSR, Leningrad, 1983, vol. 111, pp. 29–46, (in Russian).

[4] X. A. Naidenova and J. G. Polegaeva, "Application of similarity-distinction relations for processing multi-spectral information," in *Theses of papers of All Union Conference "Image processing and remote investigations"*, V. P. Pyatkin, Ed., vol. 3, Novosibirsk, USSR, 1983, pp. 67–68, (in Russian).

[5] X. A. Naidenova and J. G. Polegaeva, “Model of human reasoning for deciphering forest’s images and its implementation on computer,” in *Theses of papers and reports of school-seminar ”Semiotic aspects of the intellectual activity formalization”*, Kutaisy, Georgia Soviet Socialist Republic, 1985, (in Russian).

Xenia Naidenova obtained Ph.D. in Computer Science from the St.Petersburg Electrotechnical University. Xenia is a senior researcher of the Group of Psycho Diagnostic Systems Automation at the Military Medical Academy, St.Petersburg, Russia. Email: ksennaid@gmail.com

Vladimir Parkhomenko is a software engineer in the St.Petersburg State Polytechnical University, St.Petersburg, Russia. Email: parhomenko.v@gmail.com

Alexander Rudenko is a student in software engineering in the St.Petersburg Electrotechnical University, St.Petersburg, Russia. Email: sanek1_91@mail.ru

ISBN: 978-1-61804-251-4

New Approach for Learning Process Evaluation in Neurodegenerative Diseases Research

Lucie Houdová and Eduard Janeček

***Abstract*—**The paper introduces a new approach of evaluation for performed neurodegenerative research experiments conducted on mice models. This approach is based on determination of learning process defined according to commonly used experimental methods for inbred mouse strains' motor and cognitive function measurement. The system representation of the learning process and the derivation of new evaluation methods are described. The attribute success and the attribute time constitute foundation of this evaluation. Novelty of this approach lies in the evaluation of the learning process utilizing a probabilistic approach and evaluation in physical units of random variable using Wasserstein (pseudo)metric. In addition, the employment of new approach for choosing the best approximation of the measured data and its suitability for performed experiments evaluation is described.

***Keywords*—**evaluation of biological experiments, motor and spatial learning, statistical distance, Wasserstein pseudometrics

I. Introduction

THE use of animal (biological) models is an important part of the biomedical research. Its purpose is to collect information for the understanding of biological processes in living organisms and the principles of the human body, examining the impact of various diseases, the development of safe and effective ways of preventing and treating diseases. Many animal models of neurodegenerative diseases have been recently developed due to advances in molecular genetics and the development of gene transfer technologies (e.g., Parkinson's disease, Alzheimer's disease, Huntington's disease). The main goal of animal models utilization is to prove hypotheses about the impact of any physiological or external factor to diseases which can be used in the treatment, in searching for causes or for verification of the impact of external influences.

In the case of the neurodegenerative diseases the mice, rat, and nonhuman primate models are especially used (for overview see [1]). Generally, mice are still the most commonly used animal models in medical research. Mouse genome has already been mapped, see [2], and due to its genetic similarity with human genome the employment of mice models allows the study of human genetic disorders and diseases with far greater accuracy and less risk.

One of the problems also studied by the Institute of Pathological Physiology, Faculty of Medicine in Pilsen, is the issue of cerebellar dysfunction, affecting motor and cognitive functions of the individual. The tools for determining the level of these dysfunctions are several methods of motor, and spatial learning, related with CNS excitability. Different kinds of experiments using mice models are performed under laboratory conditions and then evaluated primarily off-line. For this kind of research the inbred animals (in described case inbred strains of mice) are commonly used, especially for the preservation of desirable characteristics, such as the genetic mutation [3].

The research of neurodegenerative diseases associated with evaluation of the learning process can be from a technical point of view tackled in two ways. The first one is the creation of cybernetic models developed according to biological basis. So far, many computer models of classical conditioning, motor, and spatial learning were created (see e.g., [4], [5], [6], respectively). The second one is the task of the creation of evaluation models, which are constructed specifically for experimental animal models. This task is not usually solved. It should be emphasized that the computer models are also evaluated by performing experiments conducted on biological (animal) models. It should be noted that in principle it is not always possible to design them. Therefore, it is appropriate to focus on the second approach.

II. Commonly Used Methods Of Learning Process Measurement And Evaluation

The understanding of some natural phenomena is necessary for general examination of processes in all substantial aspects. Thus it is necessary to mention the methods of mice abilities measurement and commonly used methods for experimental data evaluation. The learning process is seen as motor or cognitive ability change.

The motor and spatial learning are evaluated differently, mainly due to the different methodology of testing (e.g. see [7]). Motor testing is generally performed at given time intervals (days/weeks/months), usually four times (one series) in set of three or four tests (test on horizontal bar, on ladder, on cylinder and not always performed “fall” test) in each testing period. The evaluation of experiment is performed for testing on each tool for every series based on mean percentage

This work was supported by the European Regional Development Found (ERDF), project NTIS New Technologies for the Information Society, European Centre of Excellence, CZ.1.05/1.1.00/02.0090.

Authors are with the NTIS, University of West Bohemia, Pilsen, 30100 Czech Republic (corresponding author to provide phone: +420-377-63-2587; e-mail: houdina@ntis.zcu.cz).

value of successful in testing group and corresponding standard error of the mean ($\mu_{u_p} \pm \mathrm{SEM}_{u_p}$). The time of remaining on the testing tool is not used for learning process evaluation.

The cognitive testing (spatial learning) is based on the mice placement into the pool (Morris Water Maze) consecutively at each cardinal point on the pool edge (N-S-W-E) during each test at given time intervals (usually days). Their task is to find the platform and to climb onto it. The mean latency (t_L) of platform finding for four start positions is measured. The latency is used for learning process evaluation. So, an evaluation of experiment is performed for each test by mean time value of mentioned mean latency in testing group and corresponding standard error of the mean ($\mu_{t_L} \pm \mathrm{SEM}_{t_L}$).

From afore mentioned it is obvious that all the parameters that are needed for commonly used learning process evaluation are only the mean and covariance of the normal distribution. This form of evaluation is clear for physician expert. However, sometimes they have problems with misuse of standard error of the mean and standard deviation [8].

However, physician expert would limit the possible outcomes which could be obtained from results of performed experiment on animal models and can be misleading in case of different abilities distribution than normal in examine group.

III. The System Model Defining The Learning Process

The system which defines the learning process is stochastic. Thus in order to present a new approach of experiments evaluation it is necessary to describe the investigated system as a stochastic system. For the above mentioned case it means that in order to describe the learning process of mice cognitive and motor function abilities measurement it is necessary to use the stochastic description.

The observed dynamic attributes are represented by a variable attributes u (success) and τ (time transition, generally latency). Particularly, in the case of cognitive testing (spatial learning) the examined attributes are the success of finding the platform in pool in given time (Morris Water Maze) and the time taken to reach the platform, in the case of motor learning the success rate is given by remaining on the testing tool for given time or jumping down actively and τ indicates the moment of animal fell off the tool (horizontal bar, ladder and rotating cylinder).

In term of performed laboratory experiments, the learning process is a discrete time random process (i.e. with discrete set T) and the change of attributes value can be described as an event. So the system can be described by a set of attributes y_k in the particular moment in time $t_k \in T$. I.e. *k*-th event can be described as an ordered pair $[y_k, t_k]$ where the event is considered to be the result of the test defined as $y_k = [u_k, \tau_k]$ (required in terms of exploring the learning process). The set T is a finite non-empty set of time instants t_k for $k = 1, 2, \ldots, M$, where M is the number of performed tests, which is strictly totally ordered ($t_1 < t_2 < \cdots < t_M$). The attribute u takes values 0, or 1 and attribute $\tau \le t_{\lim}$ where $t_{\lim}$ value depends on the chosen measurement method (for described tests usually 60 s).

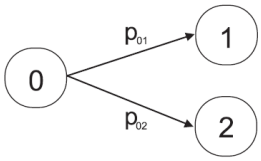


Fig. 1 three-state model of learning system given by motor or cognitive testing

Each motor or cognitive test can be described by three-state model of learning system shown in Fig. 1. Attribute success u indicates whether the mouse swims in the maze in the case of a cognitive test or whether it holds on the tool in the case of the motor test ($u = 0$). Change of the attribute value corresponds to the situation when an individual reaches the platform or falls from the tool, i.e. there is a transition to state $u = 1$, or leaves the initial (null) state after the time expiration $t_{\lim}$ ($u = 2$) by removing the mouse from the maze or the tool. It is a transition from null state to one of the final states at a certain time τ. Probabilities of transitions between states are given by the matrix

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{bmatrix}, \qquad (1)$$

where p_{ij} is probability of transitions between states i and j. If an event is defined as the change of attribute u then $p_{ii} = 0$ for $i \in \{0, 1, 2\}$. From the principle of the learning system model shown in Fig. 1 and testing methodology it is obvious that $p_{10} = p_{12} = p_{20} = p_{21} = 0$ and it applies that $p_{01} + p_{02} = 1$ where the probabilities p_{01} and p_{02} denote the probability of cognitive test success / motor test failure and the probability of cognitive test failure / motor test success, respectively.

The trajectory of the learning process for one individual mouse (one realization of the process) can be described by events, whose causality is determined by former events, i.e.

$$f\left([y_k, t_k] : [y_{k-1}, t_{k-1}], \ldots, [y_1, t_1]\right). \qquad (2)$$

From the perspective of analyzing and evaluating the learning process according to some factors given by performed experiment hypothesis, e.g. such as drugs affecting, is necessary the information about the entire measured group of animal individuals with the same characteristics (e.g. mice group affected by drugs / control group without affecting). It means physician experts are interested in the success evaluation and temporal distribution of exiting the initial state $\forall n \in \{1, \ldots, N\} : [y_{k,n}, t_k]$ in measured groups, where N is

the number of individuals in given group in time $t_k = t_1,\ldots,t_M$ for M the number of performed tests. The overall evaluated learning process in defined group is then described as

$$f\left(\left[y_{k,n},t_k\right]:\left[y_{k-1,n},t_{k-1}\right],\ldots,\left[y_{1,n},t_1\right]\right) \tag{3}$$

for $n=1,\ldots,N$ and $k=1,\ldots,M$.

IV. New Approach Of Learning Process Evaluation

The fundamental change of the new approach compared to that commonly used is the use of both attributes values (a success and a time transition) for each motor or spatial learning test evaluation. Moreover instead of normal distribution parameters the empirical distribution function and its approximation by suitable theoretical distribution is used.

The learning process is evaluated with success rate and then according to latency data which consists of few steps. The first step is to find suitable theoretical distribution for empirical distribution approximation, the second one is to use the method of learning process evaluation by using probabilistic approach and the last in physical units.

Utilizing an approximation of empirical data by the theoretical distribution is recommended due to small number of measured data sets. As mentioned above, the inbred mice individuals are genetically nearly identical, their characteristics are very close. This dramatically reduces the number of individuals needed for laboratory testing to obtain statistically significant results. Nevertheless, there still can be behavior exceptions.

A. The Evaluation of Learning Process According to the Attribut u (Success)

First of all, it is necessary to mention that the interpretation of measurement results for motor tests is performed for each test of experiment and not for a single test series as in the standard evaluation.

The success rate for each performed test (for each time t_k) is conclusively given for the cognitive tests

$$p = P[\tau < t_{\lim}], \tag{4}$$

and for motor tests

$$p = P[\tau = t_{\lim}], \tag{5}$$

where p represents the success rate, τ is the time transition realized by measured latency values, t_L in given group and $t_{\lim}$ is the methodically given time limitation of performed test. The subscript L in the variable t_L does not labelled index, it only points out that this is a time data latency.

The part of learning process evaluation using success rate is then given by the evolution of success rate p_{01}, respectively p_{02} (see Section III.) in specific time instants t_k for $k=1,\ldots,M$ where M is the number of performed tests.

B. The Evaluation of Learning Process According to the Attribute τ (Time Transition) using probability approach

For the empirical data reconstruction, the identification of parameter τ and its variations, the statistical characteristics can be used, such as the distribution function $F(x)$ (CDF) generally defined for every $x\in\langle-\infty,\infty\rangle$ for the random variable X as [9]

$$F(x) = P[X \le x]. \tag{6}$$

One of the methods for a suitable approximation selection according to the distribution function is the **assessment of statistical distances** of empirical distribution function and the chosen approximations employing a **probabilistic approach** (with respect to the y-coordinate). Fig. 2 illustrates possible approximation and the approach of suitability assessment.

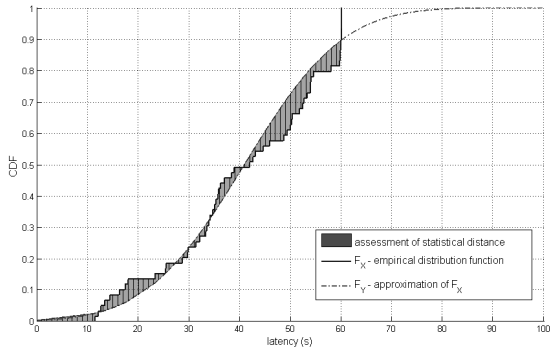


Fig. 2 approximation of empirical distribution function with the demonstration of using statistical distance employing a probabilistic approach

Denote the general random variable X as the observed random variable τ (whose realization coincides with the time of finding the platform or fall from the tool according to the testing method) and the variable x then will refer to latency t_L. The statistical distance between empirical distribution function $F_X(t_L)$, given as

$$F_X(t_L) = P[\tau \le t_L] = \sum_{x_i \le t_L} P[\tau = x_i] \quad \text{for } \forall i, \\ i\in I \text{ and } t_L \in \langle 0, t_{\lim}\rangle \tag{7}$$

where I represents the set of quantile indices, and cumulative distribution function $F_Y(t_L)$ of continuous random variable Y (chosen approximation) is then given as

$$D_p(F_X,F_Y) = \sum_{i\in I}\left(F_X(x_i) - F_Y(x_i)\right)^2, \tag{8}$$

where I represents the set of indices for partition the distribution function $F_Y(t_L)$ to determine the statistical distance $D_p(F_X, F_Y)$. The suitable choice of the empirical data approximation should have minimal value of $D_p(F_X, F_Y)$ for each time instant t_k.

Hereinafter, if the distribution function is mentioned, the approximation of empirical distribution function is meant unless otherwise specified.

The use of assessment of statistical distance employing a probabilistic approach for finding suitable approximation is not the only area of method application while evaluating the experiment results. Similarly, this approach can be used for determining the statistical distance of distribution functions in terms of exploration the learning process.

The **evaluation of learning process** employing a **probabilistic approach** means to determine the statistical distance of investigated $F_Y(t_L)$ and the reference cumulative distribution function $F_{ref}(t_L)$ as the area delimited by the given functions (curves), i.e.

$$D_p(F_k, F_{ref}) = \int_0^{t_{\lim}} \left| F_k(z) - F_{ref}(z) \right| dz, \qquad (9)$$

where $F_k(z)$ represents investigated distribution function in time t_k for which it is requested to determine the statistical distance from the reference distribution function $F_{ref}(z)$. $F_{ref}(z)$ most often corresponds to the distribution function of the first performed test (i.e. $F_{ref}(z) = F_1(z)$), in such a case the relation (9) is valid for $k = 1, \ldots, M$ where M is the number of performed tests or a priori defined distribution function (relation (9) is used for $k = a+1, \ldots, M$).

Given that $D_p(F_k, F_{ref})$ from relation (9) corresponds to the area between distribution functions, it is necessary for learning process evaluation to implement the learning-forgetting parameter $g(t_k)$, such as

$$g(t_k) = \begin{cases} 1 & \text{for} \quad \int_0^{t_{\lim}} F_k(z)dz < \int_0^{t_{\lim}} F_{ref}(z)dz \\ 0 & \text{for} \quad \int_0^{t_{\lim}} F_k(z)dz = \int_0^{t_{\lim}} F_{ref}(z)dz, \\ -1 & \text{for} \quad \int_0^{t_{\lim}} F_k(z)dz > \int_0^{t_{\lim}} F_{ref}(z)dz \end{cases} \qquad (10)$$

where values $g(t_k) = 1$ and $g(t_k) = -1$ correspond to learning for motor functions testing, to forgetting, respectively. For cognitive functions testing (spatial learning) the parameter interpretation is exactly the opposite, i.e. $g(t_k) = 1$ and $g(t_k) = -1$ correspond to forgetting, to learning, respectively. In this case the learning is defined as increase of t_L and not decrease as for motor learning.

C. The Evaluation of Learning Process According to the Attribute τ (Time Transition) using Wasserstein pseudometric

Suitable empirical distribution function approximation could be also selected according to **assessment of statistical distances** of empirical distribution function and the chosen approximations (as in the case described in Section IV.-B) but **in physical units** (in x-coordinate). In the Fig. 3 possible approximation and mentioned approach of suitability assessment is illustrated.

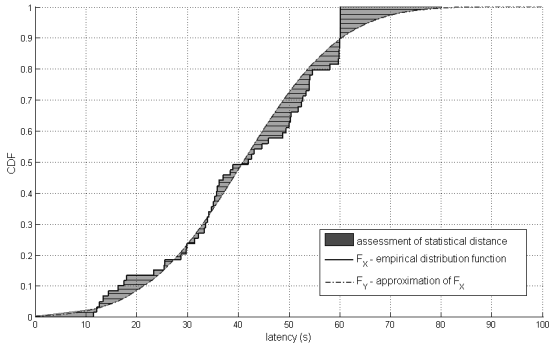


Fig. 3 approximation of empirical distribution function with the demonstration of approach using statistical distance in physical units by Wasserstein pseudometric

This approach is not so suitable for approximation selection due to performed experiment time limitation. However, the main area of employing of this approach is **the evaluation of learning process in physical units**.

The proposed evaluation method is to use certain metric defined by theories of the distances between trajectories in probability space Ω, i.e. using a specific probability measure.

For quantifying the distance between two probability measures P_1 and P_2 in the set of probability measures $\mathbf{P}_p(\Omega, d)$ on Ω was chosen the Wasserstein pseudometric (as a valid choice for the (pseudo)metric $d : \Omega \times \Omega \to \mathbb{R}^{\geq 0}$ p-norm was defined, consider any p such that $1 \leq p < \infty$) given by following definition, taken from [10].

Definition 1 *The Wasserstein pseudometric W_d^p between two probability measures P_1 and P_2 on a sample space Ω equipped with a pseudometric d is defined as*

$$W_d^p(P_1, P_2) = \left(\inf_{Q \in J(P_1, P_2)} \int_{\Omega \times \Omega} d(\omega, \eta)^p \, dQ(\omega, \eta) \right)^{1/p}, \qquad (11)$$

where $J(P_1, P_2)$ is the set of all possible joint distribution of P_1 and P_2.

$\omega, \eta \in \Omega : \mathbb{R}^{\geq 0} \to Y$, Y is the set of all allowable outputs; Q is the element of possible joint distribution of P_1 and P_2, and p means the p-norm. The required property of the set Ω is that there exists a $\eta \in \Omega$ such that $\int_{\Omega} d(\omega, \eta)^p \, dP_i(\omega)$ is finite and required property of metric d is $d(\omega, \eta) \geq 0, d(\omega, \phi) + d(\phi, \eta) \geq d(\omega, \eta)$ and $d(\omega, \eta) = 0$ if $\omega = \eta$.

For the purpose of the new evaluation approach it is sufficient to use one-dimensional metric d with p-norm equal to 1. Then the Wasserstein pseudometric W_d^p between two probability measures P_1 and P_2 on $\mathbf{P}_p(\Omega, d)$ for $p = 1$ is

$$W_d^{p=1}(P_1, P_2) = \left(\int_0^1 \left| F_{P_1}^{-1}(y) - F_{P_2}^{-1}(y) \right|^{p=1} dy \right)^{1/p=1}. \quad (12)$$

The proof can be found in [11]. Therefore, it is clear that the Wasserstein pseudometric is the area between the inverse distribution functions generated by the two probability measures. The statistical distance in physical units between the reference distribution function $F_{ref}(z)$ and measured experiment $F_k(z)$ is then given as

$$W_d(F_k, F_{ref}) = \int_0^1 \left| F_k^{-1}(z) - F_{ref}^{-1}(z) \right| dz, \quad (13)$$

where $F_k^{-1}(z)$ is examined inverse distribution function in the time t_k, for which it is required to determine the distance from the reference inverse distribution function $F_{ref}^{-1}(z)$. The choice of $F_{ref}^{-1}(z)$ corresponds to the choice of $F_{ref}(z)$ for $D_p(F_k, F_{ref})$ according to the relation (9). It is also necessary for learning process evaluation to implement the learning-forgetting parameter $g(t_k)$, see (10), as it is needed for the evaluation of learning process employing a probabilistic approach.

V. Discussion

In article described new approach based on statistical distance can be used almost from the beginning of performed experiment evaluation connected with learning process while choosing an appropriate approximation of the empirical data. The most suitable approximation is chosen as to minimize the criterion function (determining the distance between chosen approximations and the empirical distribution function) and its' suitability is not depended on the diversity of data as in the commonly used case of just using normal distribution.

The assessment of evaluation it is then built on the combination of evaluation according to success and time transition, i.e. with using success rate p and metrics $D_p(\cdot)$ and $W_d(\cdot)$ defined via statistical distance. The value of metric $W_d(\cdot)$, normalized according to the physical units, indicates the time increase of remaining on the testing tool for motor testing and time decrease of finding the platform in the case of cognitive testing, both from the reference time (in seconds). $D_p(\cdot)$ indicates the increase in the probability of achieving the limit value for motor testing or immediate platform finding for cognitive testing. While using these two methods the results obtained from experimental measurement always must be presented with learning-forgetting parameter, i.e. as $g(t_k) \cdot D_p(F_k, F_{ref})$ and $g(t_k) \cdot W_d(F_k, F_{ref})$.

One of the advantages of new approach is then the clear results interpretation unlike the commonly used evaluation by $\mu_{u_p} \pm \mathrm{SEM}_{u_p}$ or $\mu_{t_L} \pm \mathrm{SEM}_{t_L}$. While using presented approach the experiments results are interpreted by learning process description, i.e. medical expert knows how the latency changes (how many seconds), how it changes the probability of achieving required test results and how large part of the testing animal group achieved methodically required limitation during testing.

Thereby, this approach directly reflects the difference in individuals' abilities in the compared groups before or at the very beginning of experiment. The learning process is defined within that particular group. Thus, if at the beginning there is a different level of ability in each testing group, it is clear which group learns faster/slower, how is the difference, which is also important in the case of testing hypotheses about achieving the same/different results of performed experiment (during and the most often at the end of it). That is why it is necessary to know the distribution, i.e. to have a priori knowledge.

An important advantage of the proposed approach is also including motor learning during experiment performing even not individuals are not achieving methodically defined success (time limitation) during testing on tools.

VI. Conclusion

Generally the basic benefit of the new approach is a significant extension of the knowledge and information gained from the experiments. The presented methods are used more to describe a learning process, which is a cornerstone of experiment related to neurodegenerative diseases evaluation, than to describe the ability of testing individuals. The use of the proposed approach is not limited to the described kind of experiment but it can also be used for each experiment satisfying the conditions of described stochastic learning process.

References

[1] A. W. S. Chan, and Y. Agca, "Transgenic animal models of neurodegenerative diseases," in *Sourcebook of Models for Biomedical Research*, M. P. Conn, Ed. Totowa: Springer (Humana Press), 2008, pp. 323–330.

[2] Mouse Genome Sequencing Consortium, "Initial sequencing and comparative analysis of the mouse genome," *Nature*, vol. 420, no. 6915, pp. 520–562, Dec. 2002.

[3] L. C. Strong, "Inbred Mice in Science," in *Origins of Inbred Mice*, H. C. Morse III., ed. New York: Academic Press, 1978, pp. 45–67.
[4] C. Balkenius, and J. Moren, "Computational Models of Classical Conditioning: A Comparative Study," in *From Animals to Animats 5: Proc. of the 5th Int. Conf. on Simulation of Adaptive Behaviour*, Cambridge, MA: MIT Press, 1998, pp. 348-353.
[5] J. S. Albus, "A New Approach to Manipulator Control: The Cerebellar Model Articulation Controller," in *A Century of Excellence in Measurements, Standards, and Technology*, D. R. Lide, Ed. CRC Press, 2002, pp. 237-240.
[6] A. Arleo, and W. Gerstner, "Spatial Condition and Neuro-Mimetic Navigation: A Model of Hippocampal Place Cell Activity," *Biol. Cybern*., vol. 83, no. 3, pp. 287-299, Aug. 2000.
[7] E. Porras-García,J. Cendelín, E. Domínguez-del-Toro, F. Vožeh and J. M. Delgado-García, "Purkinje cell loss affects differentially the execution, acquisition and prepulse inhibition of skeletal and facial motor responses in Lurcher mice," *Europ. J. of Neuroscience*, vol. 21, no. 4, pp. 979–988, Feb. 2005.
[8] P. Nagele, "Misuse of standard error of the mean (SEM) when reporting variable of a sample. A critical evaluation of four anaesthesia journals.," *Brit. J. of Anaesthesia*, vol. 90, no. 4, pp. 514–516, Apr. 2003.
[9] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. New York: McGraw-Hill, 2002, ch. 4.
[10] D. Thorsley and E. Klavins, "Approximating Stochastic Biochemical Processes with Wasserstein Pseudometrics," *IET Systems Biology*, vol. 4, no. 3, pp. 193–211, May 2010.
[11] S. S. Vallander, "Calculation of the Wasserstein Distance Between Probability Distributions on the Line," *Theory Probab. Appl.*, vol. 18, no. 4, Sep. 1974, pp. 784–786 [Transl. Teor. Veroyatnost. i Primenen., vol. 18, no. 4, 1973, pp. 824-827]

Concentration transfer for the problem of two-phase flow of a fluid and multicomponent gas mixture in anisotropic medium

D. O. Dill, A. M. Bubenchikov

***Abstract*—** The applicability of one MUSCL type scheme for solving problem of two-phase flow of a fluid and gas mixture in anisotropic permeability medium has been investigated. For the numerical solution the finite volume method (FVM) with a nonlinear two-point approximation of flows on the volume boundaries has been used. Calculations have been carried out for various parameters of the medium. The obtained results allow us to conclude on the effectiveness of the investigated schemes to solving this problem.

***Keywords*—** convective concentration transfer, two-phase flow, MUSCL type scheme, FVM, anisotropic medium.

I. Introduction

Solving tasks of two-phase flow in porous media has a wide range of practical applications, both in oil and gas production and in dealing with ecological problems. In recent years there has been observed an increasing interest in studying the process of methane extraction from underdeveloped coal seams saturated with water [1, 2]. In addition, carbon dioxide is often pumped into the coal-bed for enhancing of methane recovery. For the purpose of successful managing and predicting a gas and fluid mixture movement, mathematical modelling techniques are widely used.

One of the ecological problems, in dealing with which the model of two-phase flow of a fluid and gas mixture is also used, is monitoring methane accumulations while flooding exhausted coal mines. The peculiarity of this model is the convective concentration transfer for each of the components, the numerical calculation of which usually causes difficulties. Besides, the numerical implementation of these models requires schemes applicable for unstructured meshes, as well as in the case of anisotropy of permeability.

The aim of this work is to study the effectiveness of one MUSCL type scheme [3] for modelling the convective concentration transfer of the gas mixture components in the case of displacement fluid and one component from an anisotropic permeability medium by the another component.

D. O. Dill is with the Theoretical Mechanics Department, Tomsk State University, Tomsk, Russia (corresponding author to provide phone: +7-952-179-4060; e-mail: gradpower@list.ru).

A. M. Bubenchikov is with the Theoretical Mechanics Department, Tomsk State University, Tomsk, Russia (corresponding author to provide phone: +7-913-850-0937; e-mail: alexy121@mail.ru).

II. Mathematical model

The model of two-phase flow of a fluid (we will consider water) and gas mixture, which is studied in this paper, is based on the law of mass conservation (Darcy's law) and has the following form:

$$\frac{\partial c^i m(1-s)}{\partial t} - div(\frac{\mathrm{K}k_r^g(s)}{\mu^i} c^i \vec{\nabla} p^g) = q^i, \quad i = \overline{1,nc}$$

$$\frac{\partial \rho^f ms}{\partial t} - div(\frac{\mathrm{K}k_r^f(s)}{\mu^f} \rho^f (\vec{\nabla}(p^g - p^c(s)) - \rho^f \vec{g})) = q^f,$$

where c^i, μ^i are concentration and dynamic viscosity of the *i-th* component of a gas mixture, m is medium porosity, *s* is water saturation, K is permeability tensor of a medium, p^g is pressure of a gas mixture, *nc* is the number of mixture components, ρ^f, μ^f are fluid density and viscosity, q^i, q^f are corresponding source components. The dependence of relative permeabilities for each phase, as well as capillary pressure being dependent on moisture saturation, is determined by closing relations of Van Genuchten - Mualem [4] as follows:

$$k_r^g(s) = \sqrt{1-s_e}\left[1 - s_e^{1/m}\right]^{2m},$$

$$k_r^f(s) = \sqrt{s_e}\left[1 - \left(1 - s_e^{1/m}\right)^m\right]^2,$$

$$p^c(s) = \frac{\rho^f g}{\alpha}\left[\frac{1 - s_e^{1/m}}{s_e^{1/m}}\right]^{1-m},$$

where $s_e = \frac{s - s_{res}}{1 - s_{res}}$ - effective water saturation, s_{res} - residual water content, α and m – parameters of porous medium.

We will assume gas pressure, water saturation and concentration of gas mixture components to be the main independent variables. Thus, we get an *nc*+1 equation for *nc*+2 unknowns. The missing relation is obtained from the gas mixture law. In the considered problem, we use the ideal gas law for two-component gas mixture. In a case of higher pressures, it can be easily replaced by the real gas mixture law, which takes into account the component composition of this mixture.

ISBN: 978-1-61804-251-4

The flow is considered in a model two-dimensional domain with impermeable walls, an inlet and an outlet, where the Dirichlet boundary conditions are established (Fig.1). The no-flow conditions are established on the walls. Fig.1 also presents the location of the anisotropic permeability area. Along the dashed lines permeability is 10 times greater than that across, as well as in the rest of the considered area.

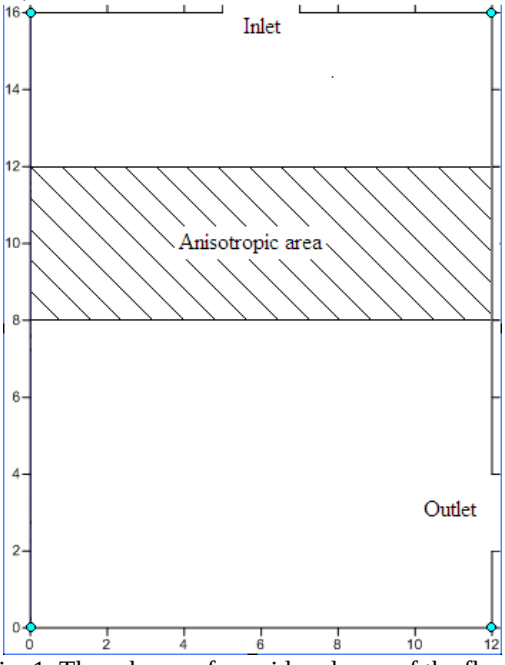


Fig. 1. The scheme of considered area of the flow.

III. Numerical solution

For the purpose of constructing the discrete analogues for the original equations in partial derivatives we use the FVM with a nonlinear two-point approximation of flows on the volume sides [5], which allows to take into account the anisotropy permeability of medium and to summarize the numerical procedure for irregular meshes. The main unknown variables are determined in the centres of finite volumes. The non-linear approximation is different from the standard because it determines the coefficients which, in the case of a nonlinear scheme, contain weighted value contributions of the unknowns in the neighbouring volumes.

To approximate the concentration values on the finite volume boundaries for the convective transfer we used the MUSCL type scheme with local gradient constraints for each finite volume. Concentration gradient, calculated at one side, is constrained in such a way that values, obtained by interpolation with it's using at other sides, do not exceed the values, obtained by interpolation with using the gradients, calculated at these sides.

The final discrete equations are linearized using Newton's method. The resulting sparse matrix is solved by the block SOR method [6]. To control the computation, the balance relations between the fluid and the gases ingoing, outgoing and contained in the area under consideration are checked at each time step.

IV. Results and discussion

The parameters, which were used for calculating the flow, are given in Table 1. For the computation structured mesh was used. The calculations carried out for different values of parameters in the closing relations of Van Genuchten - Mualem showed their influence only on the water saturation frontage, but not on the nature of the flow.

The Figures 2-5 give concentration distribution of the 1st component at different times for cases of presence area with anisotropic permeability (2, 3) and its absence (4, 5). In an anisotropic region the typical acceleration of flow and change its direction can be observed. In this case uncharacteristic fluctuations, which could be due to possible incompatibility of two schemes: a nonlinear two-point scheme for calculating flows on the volume sides and MUSCL type schemes for calculating concentrations, isn't observed. Near the inlet can be seen a sharp increase in concentration due to the displacement fluid by gas and a higher gas pressure.

Table 1. Parameters used in the study.

Parameter	Value
Area size	12x16 m
Area thickness	0.5 m
Finite volume size	0.5x0.5x0.5 m
Porosity	0.07
Permeability	9.869233×10^{-15} m^2 (10 mD)
Water viscosity	8.9×10^{-4} Pa s
Water density	1000 kg/m^3
Gas mixture viscosity	1.78×10^{-5} Pa s
Gas mixture density	1.07 kg/m^3
Boundary conditions at the inlet	
Pressure	120 kPa
Water saturation	0.16
1st comp. concentration	50 mol/m^3
2nd comp. concentration	0 mol/m^3
Boundary conditions at the outlet and initial conditions	
Pressure	100 kPa
Water saturation	0.6
1st comp. concentration	0 mol/m^3
2nd comp. concentration	41.5 mol/m^3

ISBN: 978-1-61804-251-4

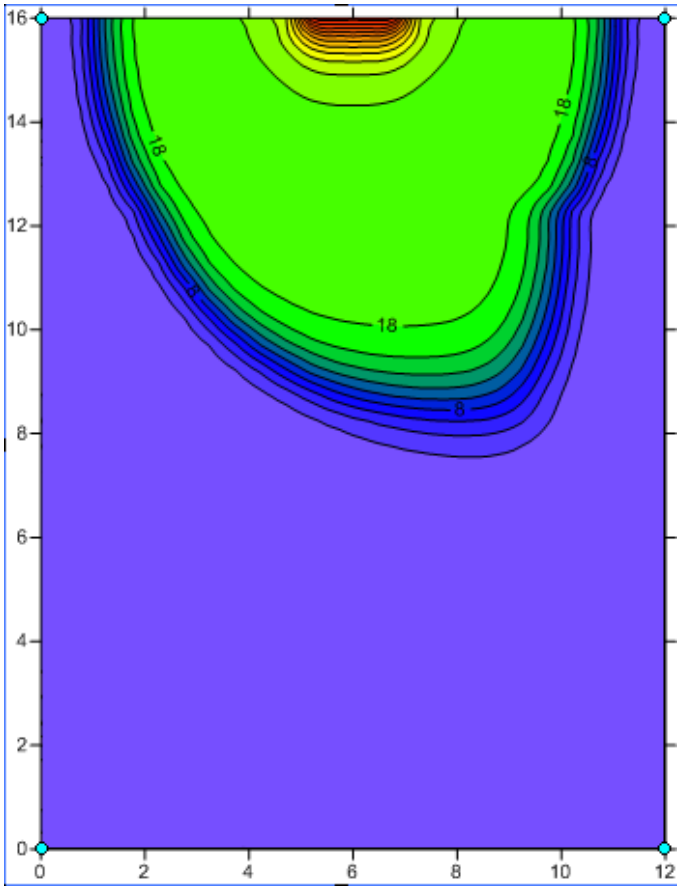


Fig. 2. Concentration distribution at time 200 hours.

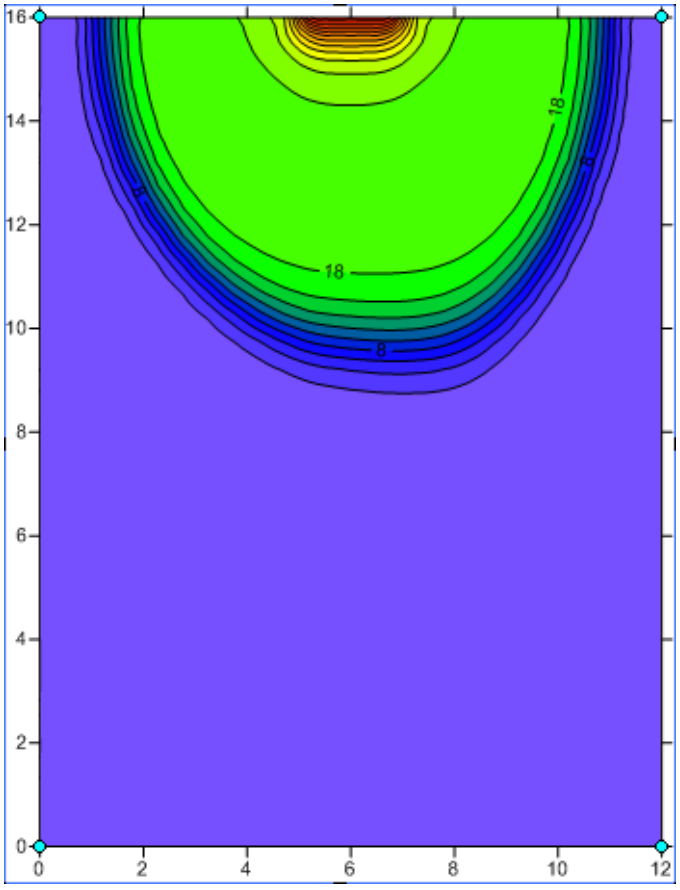


Fig. 4. Concentration distribution at time 200 hours.

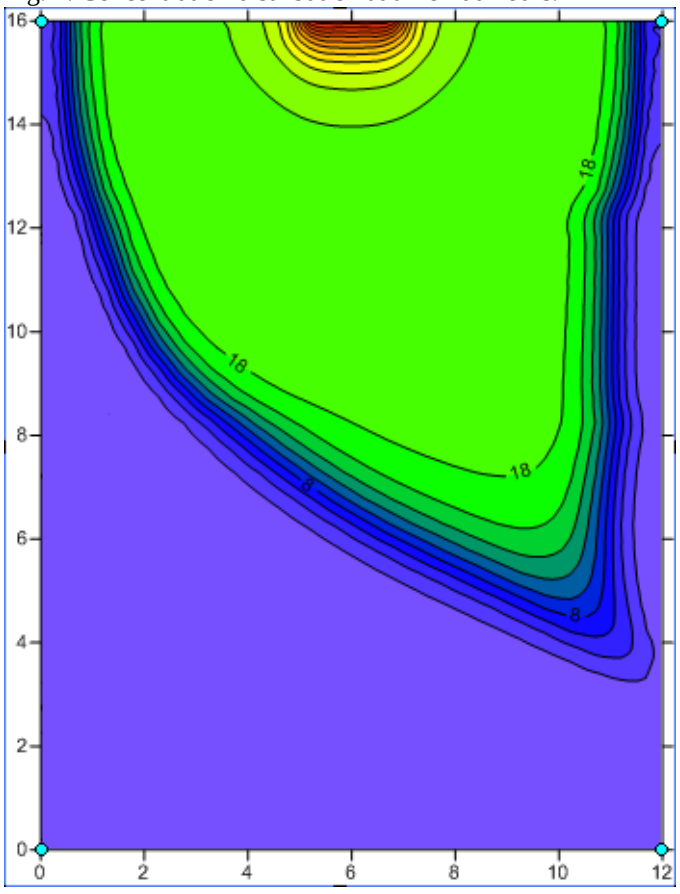


Fig. 3. Concentration distribution at time 350 hours.

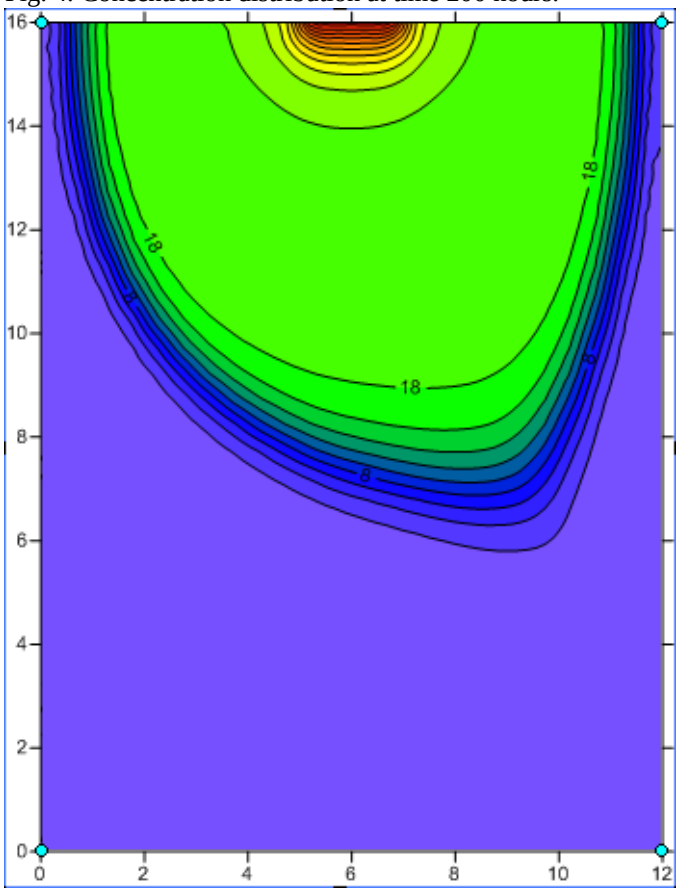


Fig. 5. Concentration distribution at time 350 hours.

ISBN: 978-1-61804-251-4

V. Conclusion

The results of numeric experiment allow us to conclude on the applicability of the investigated scheme for cases of two-phase flow of a fluid and gas mixture in anisotropic permeability medium and recommend it for such calculations. However it should be noted that with increasing size of the finite volume the numerical diffusion could significantly impact on the accuracy of the results. It is necessary to take into account for the majority of the schemes used with the FVM.

References

[1] M. Jamshidi, K.Jessen, "Water production in enhanced coalbed methane operations," *J. of Petroleum. Science and Engineering*, vol. 92-93, 2012, pp. 56–64.

[2] P. Thararoop, Z. T. Karpyn and T. Ertekin, "Development of a multi-mechanistic, dual-porosity, dual-permeability, numerical flow model for coalbed methane reservoirs," *J. of Natural Gas Science and Engineering*, vol. 8, 2012, pp. 121–131.

[3] M. E. Hubbard, "Multidimensional slope limiters for MUSCL-type finite volume schemes on unstructured grids," *J. of Computational Physics*, vol. 155, 1999, pp. 54–74.

[4] M. Th. Van Genuchten, "A closed-form equation for predicting the hydraulic conductivity of unsaturated soils," *Soil Science Society American J.*, vol. 44, 1980, pp. 892–898.

[5] K. D. Nikitin, "Finite volumes method for the convection-diffusion problem and two-phase flow models," Ph.D. dissertation, Moscow, 2010.

[6] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., Philadelphia: SIAM, 2003, ch. 3.

The mathematical model of the dynamics of bounded Cartesian plumes

Khaled S. Al Mashrafi

***Abstract*—**The mathematical model of the dynamics of a column of buoyant fluid rising in a less buoyant fluid bounded by two fixed vertical walls a finite distance apart is investigated, as an example of a compositional plume in a bounded domain. This is an extension to the mathematical model of Cartesian plume in the absence of the sidewalls by Eltayeb and Loper (1994). The problem is governed by five dimensionless parameters: (i) the Grashof number, R , which is the ratio of the buoyancy force due to the difference in concentration of light material of the plume and that of the surrounding fluid to the viscous force, (ii) the Prandtl number, σ , which is the ratio of viscosity to thermal diffusivity, (iii) the thickness of the plume, (iv) the distance between the two vertical walls and (v) the distance between the plume and the nearest sidewall made dimensionless using the salt-finger length scale as a unit of length. The main objective of this study is to investigate the influence of the boundary on the solution obtained when the surrounding fluid is unbounded. The mean (basic) state solution is independent of R and σ . The symmetry of the solution present in the absence of the boundaries is here lost unless the plume lies half-way between the two sidewalls. In the absence of the boundaries, the plume is always unstable with growth rate $O(R)$. The instability takes the form of one of two uncoupled categories of solutions: the varicose (V) mode, and the sinuous (S) mode. The introduction of the boundaries introduces dramatic changes to the stability of the plume. A region of instability with a growth rate of $O(1)$ appears when the plume is thin and lies close to the boundary. For other regions of the plane, the plume has a growth rate of the same order of magnitude as in the absence of the boundaries but its magnitude is reduced as the distance between the sidewalls is decreased. Instability again takes the form of one of two categories of solutions which related to the S and V modes but here modified by the presence of the boundaries and the position of the plume relative to the nearest boundary.

***Keywords*—**Bounded domain, Compositional plumes, Growth rate, Stability

I. Introduction

If a fluid of two components of different densities is cooled from below, the component with the higher melting temperature solidifies first and settles at the bottom of the container to form a layer of mixed solid crystals and light fluid (Copley *et al.* 1970). The layer so produces at the bottom is known as a mushy layer (Huppert 1990). If solid crystals so formed belong to the heavier component, they will settle at the bottom to produce a solid layer (Chen and Chen 1991). As the solidification process continues, the mushy layer becomes thicker and eventually becomes unstable (Worster 1992). Due to its instability, compositional plumes of the light component rise from the mushy layer through the melt layer to the top (Eltayeb and Loper 1991). These types of plumes represent directional fluid flow rising in another fluid of different properties and composition. Such a channel flow is defined as a compositional plume (Al Mashrafi and Eltayeb 2014(a)).

The mathematical model of the dynamics of compositional plumes is important for the understanding the properties of the fluid alloys in the presence of heat and pressure. This understanding is essential for some real life applications including industrial (e.g. iron casting), geophysical (e.g. solidification at inner core boundary (ICB) of Earth, mantle plumes) and environmental (e.g. salt fingers, sea ice).

In industrial field, one of the problems the iron casting industry faces is the appearance of freckles in iron bars (Smeltzer 1959). When iron ore is poured into ingots, the trapped air escapes in the form of thin air channels which when the iron bar solidifies appear as very thin black strips along the iron bar. These pockets lead to a weakness in the iron bar.

In geophysics, it is believed that the outer core of Earth is an alloy consisting mainly of iron and nickel and some light materials such sulphur, oxygen, hydrogen, and helium, whereas the inner core is a solid layer consisting of iron and nickel (Loper 1978). Although the temperature increases with depth in the interior of Earth, some researchers suggested that the growth of the inner core is due to solidification occurring at the inner core boundary (ICB) of Earth because of the high pressure (Verhoogen 1961; Braginsky's 1963). The iron solidifies leaving the other elements in fluid form thus forming a mushy layer. The iron, which is the heavier element, settles on the ICB, and leads to growth of the inner core (Loper 1987).

The mushy layer can become unstable (Moffatt 1989) and its instability can take the form of thin filaments of light material rising through the outer fluid core in the form of compositional plumes (see **figure 1**), and probably interacts with the magnetic field of Earth and contributes to the regeneration of the geodynamo (Loper and Roberts 1978).

In environment, compositional plumes appear in some situations like salt fingers phenomenon (see **figure 2**). In very

ISBN: 978-1-61804-251-4

cold regions where water at the bottom of the sea is very cold, the water on the surface of the sea becomes hot and relatively fresh. In the case when diffusion of heat is much faster than salt diffusion, the water at the top loses its heat before it loses its concentration of salt (Stern 1960). As a result, the water on the surface becomes cold and dense with salt. It sinks as compositional plumes to produce salt fingers (Howard and Veronis 1992). This phenomenon plays an important role in the climate system in the very cold regions. The indication that the appearance of the compositional plumes is harmful in iron casting and useful in the geophysics and environment has motivated studies on the dynamics of the compositional plumes.

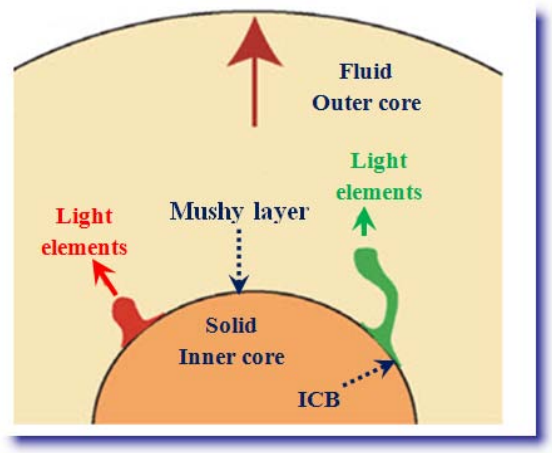


Fig. 1 schematic diagram of the solidification at inner core boundary of Earth. A mushy layer of thickness nearly 1 kilometer appears at ICB (and it is very thin so that it cannot be seen at this level of resolution). Note that the temperature and pressure are very high at ICB. (Source: Buffett, B. A. 2007).

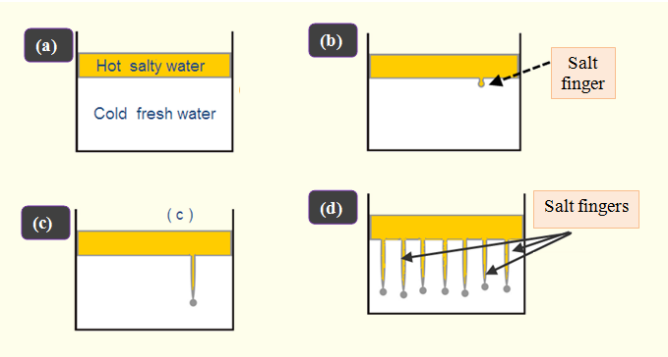


Fig. 2 schematic diagram of salt finger development. Note that the diffusion of heat must be faster than the diffusion of salt for salt fingers to form.

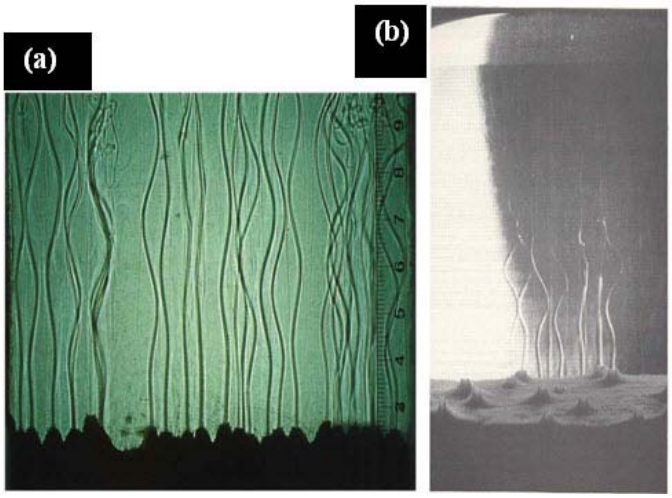


Fig.3 aqueous ammonium chloride solution (a. the experiment by Huppert ,1990 ; b. the experiment by Eltayeb and Loper, 1991).

The dynamics of compositional plumes has been investigated theoretically and experimentally. The experimental studies (see **figure 3**) showed that the plume flow seems to stable (Sample and Hellawell 1984) although some experiments observed that it can be unstable (Classen *et al.* 1999). In the addition, the experiments showed that the behavior of the plume depends on its position inside the container whether it is close to the wall or away from it (Hellawell *et al.* 1993). On the other hand, the dynamics of the plume has been modeled theoretically to investigate the behavior of the compositional plumes. The study by Eltayeb and Loper (1991) was the first on the stability of compositional plumes. They examined the stability of vertical interfaces, across which there is an imposed jump in composition, in the presence of a vertically oriented stabilizing temperature gradient. They also made other studies on more realistic forms of plumes (Eltayeb and Loper 1994; 1997). All these studies concluded that the plume is always unstable. It is also shown that the instability of the plume occur in the presence of rotation (Eltayeb and Hamza 1998) or magnetic fields (Eltayeb *et al.* 2005) or by the simultaneous action of rotation and magnetic field (Eltayeb 2006). All these studies on the plumes are conducted on unbounded domains and the material diffusion is negligible.

The comparison between the theoretical models and experimental studies is marred by the difference on the domain of the fluid that surrounding the plume. So, the current study is an attempt to investigate theoretically the influence of boundaries on the dynamics of plumes. A simple model is developed taking into account the main factors that are expected to influence the plume rising in a bounded domain. The model is illustrated in **figure 4**. A column of fluid of finite thickness, $2x_0$, is rising vertically upwards in a fluid of different concentration which is bounded on either side by rigid vertical walls, a distance d apart. We choose the origin

ISBN: 978-1-61804-251-4

such that the plume interfaces are situated at $x = \pm x_0$ and the walls at $x = -a_2$ and $x = a_1$ where $a_1 + a_2 = d$.

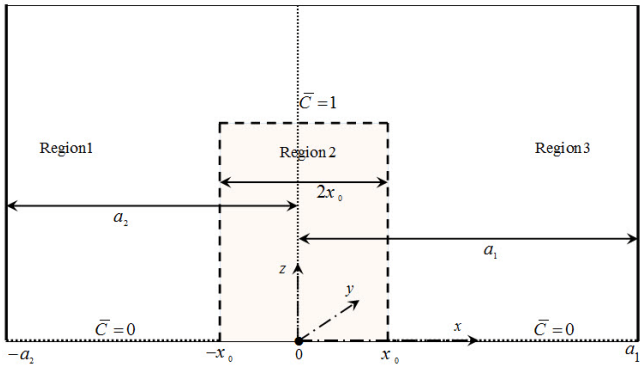


Fig. 4 the geometry of the problem showing the profile of the basic state concentration of light material representing a plume of width, $2x_0$, and concentration, 1 , rising vertically in a rotating finite fluid of width, $d = a_1 + a_2$, and concentration , 0. The plume is bounded by two rigid vertical planes on either side such that the centre of the plume is a distance a_1 from the wall on the right and a_2 from the wall on the left.

II. Formulation of the Model:

We consider a two-component incompressible fluid where the concentration of the solvent component (light material) in the fluid is C , and the temperature of the fluid is T . The two fluids have the same kinematic viscosity, ν , thermal diffusivity, κ , and material diffusion is negligible.The system of the flow is governed by the equations of motion, mass, heat, concentration, and state (Eltayeb and Loper 1991). These equations are

$$\rho_0 \left[\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u}.\nabla)\boldsymbol{u} \right] = -\nabla p + \rho_0 \nu \nabla^2 \boldsymbol{u} - \rho g \hat{\mathbf{z}}, \quad (1)$$

$$\nabla . \boldsymbol{u} = 0 , \quad (2)$$

$$\frac{\partial T}{\partial t} + \boldsymbol{u} . \nabla T = \kappa \nabla^2 T , \quad (3)$$

$$\frac{\partial C}{\partial t} + \boldsymbol{u} . \nabla C = 0, \quad (4)$$

$$\frac{\rho}{\rho_0} = 1 - \alpha \left(T - T_0 \right) - \beta \left(C - C_0 \right) , \quad (5)$$

where $\boldsymbol{u}$ is the velocity vector, p the pressure, g the uniform acceleration of gravity, $\hat{\mathbf{z}}$ is the upward unit vector, t the time, α the coefficient of thermal expansion, β the coefficient of compositional expansion, ρ the density, and (ρ_0, T_0, C_0) the constant reference values. Here we have assumed that the fluid is Boussinesq which neglects density variations in the equation of motion except when they occur in the gravity term.

The equations (1) - (5) allow a hydrostatic balance governed by

$$\frac{dp_h}{dz} + \rho g = 0, \quad \frac{d^2 T_h}{dz^2} = 0, \quad \mathbf{u}_h = \mathbf{0}, \quad C_h = C_0 . \quad (6)$$

Motivated by the experimental work on plumes rising from mushy layers, we take a temperature profile

$$T_h = \gamma z + T_0 , \quad (7)$$

where γ is a positive constant so that the temperature increases with height making the fluid stably stratified thermally and any instabilities will be due to transport of material.

We now cast the equations (1) - (5) into dimensionless form. Since our interest is the instabilities effected by the rising buoyant fluid, we take the maximum amplitude, $\tilde{C}$, of the mean state concentration as a unit of concentration of light material. In order to maintain the effects of both concentration and temperature stratification, we take the unit of temperature, T_u, as

$$T_u = \frac{\beta}{\alpha} \tilde{C} . \quad (8)$$

The viscous forces are important because the rising column, as observed in experiments, is very thin. We must then maintain the viscous forces at the same order of magnitude as the buoyancy forces. This suggests that the length scale, L, and unit of velocity, U, can be chosen as

$$L = \left(\frac{\nu \kappa}{\alpha \gamma g} \right)^{1/4}, \qquad U = \beta \tilde{C} \left(\frac{g \kappa}{\alpha \gamma \nu} \right)^{1/2} . \quad (9)$$

This length scale is the usual salt-finger length scale (Turner 1973; Eltayeb and Loper 1991). It then follows that the convective time scale, t_c, defined by

$$t_c = \frac{L}{U}, \quad (10)$$

is the relevant time scale for the problem for the growth rate of the instabilities (Al Mashrafi and Eltayeb 2014(a)). The unit of the pressure is

$$\tilde{p} = \rho_0 \beta \tilde{C} \left(\frac{\nu g^3 \kappa}{\alpha \gamma} \right)^{1/4} . \quad (11)$$

Then the equations (1) - (5) in dimensionless form are

$$R\left[\frac{\partial \boldsymbol{u}}{\partial t}+(\boldsymbol{u}.\nabla)\boldsymbol{u}\right]=-\nabla\left(p+\frac{z}{\beta\tilde{C}}\right)\\ +\nabla^2\boldsymbol{u}+\left(T-T_r+C-C_r\right)\hat{\mathbf{z}}\Bigg\}, \quad (12)$$

$$\nabla.\boldsymbol{u}=0\ , \quad (13)$$

$$R\,\sigma\left[\frac{\partial T}{\partial t}+\boldsymbol{u}.\nabla T\right]=\nabla^2 T\ , \quad (14)$$

$$\frac{\partial C}{\partial t}+\boldsymbol{u}.\nabla C=0\ , \quad (15)$$

where the dimensionless parameters R and σ are known as the Grashoff number and the Prandtl number defined by

$$R=\frac{UL}{\nu},\qquad \sigma=\frac{\nu}{\kappa}. \quad (16)$$

The Grashoff number is the ratio of the buoyancy force due to the difference in concentration of light material of the plume and that of the surrounding fluid to the viscous force, and the Prandtl number is the ratio of viscosity to thermal diffusivity.

We define a Cartesian coordinate system $O(x,y,z)$ in which Oz is vertically upwards and Ox, Oy are horizontal. We assume that the variables of the system (12) - (15) take the form

$$\boldsymbol{u}(x,y,z,t)=\mathbf{0}+\bar{w}(x)\,\hat{\mathbf{z}}+\varepsilon\boldsymbol{u}^{\dagger}(x,y,z,t)\ , \quad (17)$$

$$C(x,y,z,t)=C_0+\bar{C}(x)+\varepsilon C^{\dagger}(x,y,z,t)\ , \quad (18)$$

$$p(x,y,z,t)=p_h+\bar{p}(x)+\varepsilon\, p^{\dagger}(x,y,z,t)\ , \quad (19)$$

$$T(x,y,z,t)=T_h+\bar{T}(x)+\varepsilon T^{\dagger}(x,y,z,t)\ , \quad (20)$$

such that the variables with subscript h have hydrostatic contribution and given by

$$T_h=T_0+\frac{(z-z_0)}{\sigma R}\\ p_h=p_0-\frac{(z-z_0)}{\beta\tilde{C}}+\frac{(z-z_0)^2}{2\sigma R}\Bigg\}, \quad (21)$$

in which z_0 is a reference value.

The variables with an 'overbar' are basic state variables dependent only on a horizontal coordinate x, and the variables with 'dagger' indicate a perturbation of small amplitude $\varepsilon<<1$. The flow (17) is chosen to represent a vertical mean flow corresponding to a rising plume. The small amplitude perturbations are introduced in order to study the stability of the plume flow.

The boundary conditions of the system are: (i) the heat and momentum fluxes and the variables of the system except the concentration are continuous across the interfaces, (ii) the two walls are rigid and maintained at the hydrostatic temperature, and (iii) the two interfaces of the plume are material surfaces. The last condition indicates that no mass transfer across the two interfaces of the plume, and this condition applies when material diffusion is negligible (Al Mashrafi and Eltayeb 2014(a)).

Substituting the expressions (17) - (20) into the system (12) - (15), the terms independent of ε give the basic state equations

$$-\frac{d\,\bar{p}}{dx}\hat{\mathbf{x}}+\left(\frac{d^2\bar{w}}{dx^2}+\bar{C}+\bar{T}\right)\hat{\mathbf{z}}=0, \quad (22)$$

$$\frac{d^2\bar{T}}{dx^2}=\bar{w}\,. \quad (23)$$

These equations are discussed in section 3 below.

The order ε terms in the equations provide the linearised perturbation equations

$$R\left[\frac{\partial \boldsymbol{u}^{\dagger}}{\partial t}+\bar{w}\,\hat{\mathbf{z}}.\nabla\boldsymbol{u}^{\dagger}+\left(\boldsymbol{u}^{\dagger}.\nabla\bar{w}\right)\hat{\mathbf{z}}\right]=-\nabla p^{\dagger}\\ +\ \nabla^2\boldsymbol{u}^{\dagger}+\left(T^{\dagger}+C^{\dagger}\right)\hat{\mathbf{z}}\Bigg\}, \quad (24)$$

$$\nabla.\boldsymbol{u}^{\dagger}=0\ , \quad (25)$$

$$\sigma R\left[\frac{\partial T^{\dagger}}{\partial t}+\bar{w}\frac{\partial T^{\dagger}}{\partial z}+\boldsymbol{u}^{\dagger}.\nabla\bar{T}\right]+\ \boldsymbol{u}^{\dagger}.\hat{\mathbf{z}}\\ =\ \nabla^2 T^{\dagger}\Bigg\}, \quad (26)$$

$$\frac{\partial C^{\dagger}}{\partial t}+\bar{w}\frac{\partial C^{\dagger}}{\partial z}+\boldsymbol{u}^{\dagger}.\nabla\bar{C}\ =0\,. \quad (27)$$

The perturbation equations are solved in section 4 below.

III. Basic State Solution

Equation (15) is automatically satisfied for the basic state and we are free to choose a concentration function $\bar{C}(x)$. Thus we choose a top-hat profile

$$\bar{C}(x)=\begin{cases}1\ , & |x|\le x_0\\ 0\ , & -a_2\le x<-x_0\ ,\ x_0<x\le a_1\end{cases} \quad (28)$$

If we define

$$F(x)=\bar{T}(x)-\mathrm{i}\bar{w}(x)\ , \quad (29)$$

then the equations (22) and (23) give

$$\bar{p} = 0, \qquad \frac{d^2F}{dx^2} - \mathrm{i}F = \mathrm{i}\bar{C}\ . \tag{30}$$

The equation (30) is solved subject to the boundary conditi

$$\left.\begin{array}{l}\text{(i) } F,\ \dfrac{dF}{dx} \text{ continuous across } x = \pm x_0 \\ \text{(ii) } F = 0 \text{ at } x = a_1, -a_2\end{array}\right\}. \tag{31}$$

The solution is given by (Al Mashrafi and Eltayeb 2014(a))

$$F(x) = \begin{cases} -A\ \sinh(ka_1)\ \sinh[k\,(x + a_2)]\ ; & -a_2 \le x < -x_0 \\ -A\ \sinh(ka_1)\ \sinh[k\,(x + a_2)] + \cosh[k\,(x + x_0)] - 1\ ; & \\ A\ \sinh(ka_2)\ \sinh[k\,(x - a_1)]\ ; & x_0 < x \le a_1 \end{cases} \tag{32}$$

where A and k are defined by

$$A = \frac{2\sinh(kx_0)}{\sinh(kd)},\ k = \frac{1}{\sqrt{2}}(1+\mathrm{i}), d = a_1 + a_2. \tag{33}$$

The basic state solution (32) has been evaluated numerically and some samples of the results are illustrated in **figures 5** and **6** (see Al Mashrafi and Eltayeb 2014(a)). The influence of the sidewalls on the basic state solution is illustrated in **figure 5**. For a plume with given thickness, a decrease in the distance d between the sidewalls enhances the plume flow. If the plume is nearer to one sidewall than to the other, the down flow outside the plume is reduced in the region between the plume and the nearest sidewall and enhanced on the other side leading to lack of symmetry.

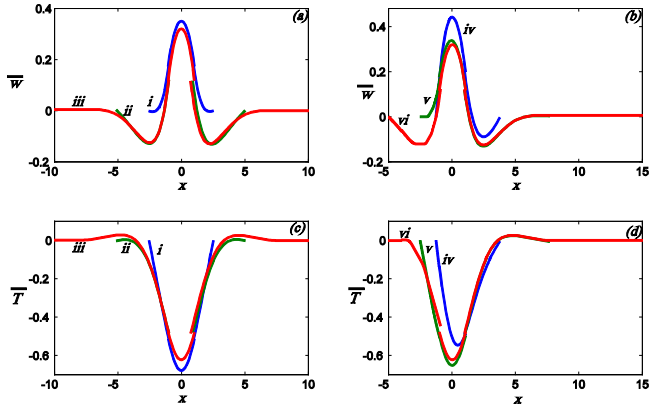


Fig. 5 the profiles of $\bar{w}(x)$ and $\bar{T}(x)$ for $x_0 = 1$ and different values of d and a_2. The subplots (a) and (c) refer to $\bar{w}$ and $\bar{T}$ when the plume is equidistant from the sidewalls and the labels *i*, *ii*, *iii* correspond to $d = 5, 10, 20$, respectively. The subplots (b) and (d) refer to $\bar{w}$ and $\bar{T}$ when $a_2 = d/4$, and labels *iv* , *v* , *vi* correspond to $d = 5, 10, 20$.

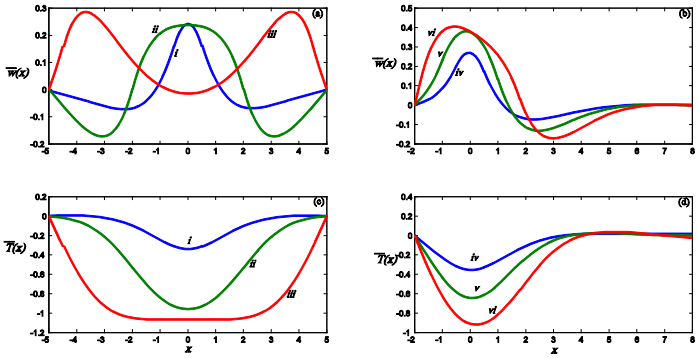


Fig. 6 the profiles of $\bar{w}(x)$ and $\bar{T}(x)$, for different values of plume thickness, $2x_0$, and distance, a_2 , from the sidewall on the left when $d = 10$. The subplots (a) and (c) refer to $\bar{w}$ and $\bar{T}$ when the plume is positioned half-way between the two sidewalls and the labels *i, ii, iii* correspond to $x_0 = 0.5, 2.0, 4.5$, respectively. The subplots (b) and (d) refer to $\bar{w}$ and $\bar{T}$ when $a_2 = 2$ and the labels *iv, v, vi* correspond to $x_0 = 0.5, 1, 1.8$, respectively.

Figure 6 shows a sample of the profiles of the solutions $\bar{T}$ and $\bar{w}$ for different values of the plume thickness $2x_0$ and distance a_2 from the nearest wall for $d = 10$. It is clear that the sidewalls have a strong influence on the solution. The profiles are symmetric when the plume is situated half-way between the sidewalls, but as the position of the plume moves towards a sidewall, symmetry is broken. The oscillatory nature of the velocity profile introduces negative flow (i.e., downwards flow) within the plume when it is wide, and this has an effect on the net transport of material by the plume.

IV. The Stability Analysis

We use the perturbation equations (24) - (27) to investigate the linear stability of the basic state solution given by (32). We assume that the interface at the plane $x = x_0$ is given a small harmonic disturbance of the form (see **figure 7**)

$$x = x_0 + \varepsilon f\ + c.c.;\ f = \exp\left(\Omega t + \mathrm{i}(my - nz)\right), \tag{34}$$

where m and n are the horizontal and vertical wavenumbers, respectively, $c.c.$ refers to the complex conjugate, and Ω is a constant which can conveniently be expressed as

$$\Omega = \Omega_r + \mathrm{i}\,\Omega_\mathrm{i}\ . \tag{35}$$

ISBN: 978-1-61804-251-4

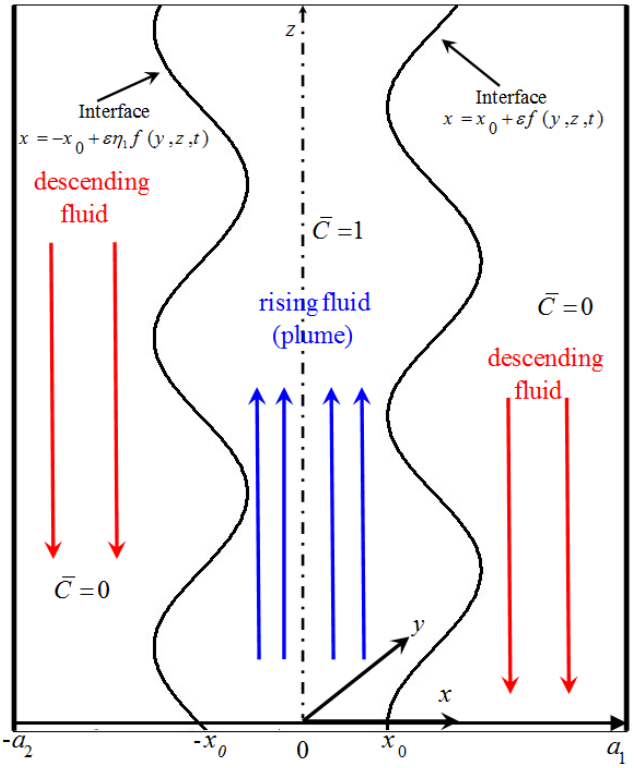


Fig. 7 illustration of the disturbances of the interfaces. Here $f\left(y,z,t\right) = \exp\left(\Omega t + \mathrm{i}my - \mathrm{i}nz\right)$.

The real part Ω_r governs the variations of the amplitude of the disturbance with time, and hence it determines the stability of the disturbance. If it is negative for all possible values of the wavenumbers m and n, then the plume is stable, while if at least one pair of m and n gives a positive value, then the plume is unstable. If Ω_r vanishes for all values of the wavenumbers, the plume is neutrally stable. If the preferred mode occurs for m, n both non-zero, it is referred to as a 3-dimensional mode, and if $m = 0$, it is called 2-dimensional. The case $n = 0$ and $m \neq 0$ is found not to occur. The imaginary part Ω_i determines the phase speeds of the disturbance. The vertical phase speed u_z and the horizontal phase speed u_h are defined as

$$u_z = \frac{\Omega_i}{n}, \qquad u_h = \frac{\Omega_i}{m} = \frac{n u_z}{m}. \tag{36}$$

We note that u_h is defined only if $m \neq 0$.

The disturbance (34) will propagate into the fluid, and affect the second interface, and the variables of the system to produce the perturbations. Consequently, the interface at $x = -x_0$ can be written in the form

$$x = -x_0 + \varepsilon\eta_1 f + c.c., \tag{37}$$

where η_1 is a measure of the amplitude of the interface determined by the boundary conditions of the problem, and the perturbation variables take the form

$$\left\{\boldsymbol{u}^{\dagger}, C^{\dagger}, T^{\dagger}, p^{\dagger}\right\} = \left\{-\mathrm{i}n u, nm v, w, C, T, -\mathrm{i}n p\right\} f, \tag{38}$$

where the factors $-\mathrm{i}n$, nm, and $-\mathrm{i}n$ are introduced in the variables u, v and p, respectively for convenience.

Substituting the variables (38) into (24) – (27) gives the following system of ordinary differential equations in the variable x

$$Du - m^2 v + w = 0, \tag{39}$$

$$\Delta u - Dp = R\,\bar{\Omega}\,u, \tag{40}$$

$$\Delta v - p = R\,\bar{\Omega}\,v, \tag{41}$$

$$\Delta w + T + C + n^2 p = R\left(\bar{\Omega}\,w - \mathrm{i}n\,u\,D\bar{w}\right), \tag{42}$$

$$\Delta T - w = \sigma R\left(\bar{\Omega}\,T - \mathrm{i}n\,u\,D\bar{T}\right), \tag{43}$$

$$\bar{\Omega}\,C = 0, \tag{44}$$

where

$$\left.\begin{aligned} b^2 &= m^2 + n^2, \quad \bar{\Omega} = \Omega - \mathrm{i}n\,\bar{w} \\ D &\equiv \frac{d}{dx}, \quad \Delta \equiv D^2 - b^2 \end{aligned}\right\}. \tag{45}$$

It is found useful to derive the following three equations

$$\Delta\varsigma = R\left(-\mathrm{i}nv\,D\bar{w} + \bar{\Omega}\varsigma\right), \tag{46}$$

$$n^2 u = m^2\varsigma - D\left(w + p\right) - R\,\bar{\Omega}\,u, \tag{47}$$

$$\Delta p - T - C = 2\mathrm{i}n\,R\,u\,D\bar{w}, \tag{48}$$

where the equation (46) is driven by differentiate (41) with respect to x and subtract (40), the equation (47) is driven by differentiate (39) once and subtract (40), and the equation (48) is driven by apply the operator Δ to (39) and use (40) - (42). The variable ς related to the vertical component of the vorticity and given by

$$\varsigma = \left(Dv - u\right) = \frac{1}{nm}\hat{\mathbf{z}} \cdot \mathrm{curl}\boldsymbol{u}. \tag{49}$$

The equation (44) indicates that the perturbation concentration vanishes everywhere in the fluid. Thus

$$C = 0. \tag{50}$$

The boundary conditions of the system are given by

$$\left.\begin{array}{l} D\left(w+p\right), v, w, T, p, Dv, DT \\ \text{are continuous across } \; x=\pm x_0 \end{array}\right\}, \tag{51}$$

$$\left.\begin{array}{l} \left\langle Dw\left(x_0\right)\right\rangle = \left\langle \bar{C}\left(x_0\right)\right\rangle \\ \left\langle Dw\left(-x_0\right)\right\rangle = \eta_1 \left\langle \bar{C}\left(-x_0\right)\right\rangle \end{array}\right\}, \tag{52}$$

$$\left.\begin{array}{l} -\mathrm{i}n u\left(x_0\right) = \Omega - \mathrm{i}n \bar{w}\left(x_0\right) \\ -\mathrm{i}n u\left(-x_0\right) = \left[\Omega - \mathrm{i}n \bar{w}\left(-x_0\right)\right] \eta_1 \end{array}\right\}, \tag{53}$$

$$\left.\begin{array}{l} D\left(w+p\right) = v = w = T = \varsigma = 0 \\ \text{at } \; x = a_1 \, , \; x = -a_2 \end{array}\right\}, \tag{54}$$

where we have introduced the operator

$$\left\langle f\left(\alpha\right)\right\rangle = f\left(\alpha^-\right) - f\left(\alpha^+\right), \tag{55}$$

representing the jump across $x=\alpha$. Note that we have used the equation (47) to replace the continuity of u and its vanishing on the boundary by $D\left(w+p\right)$ since ς is also continuous at the interfaces and vanishes on the boundary. The jump conditions (52) represent the condition that the interfaces are material surfaces.

The previous studies on a compositional plume showed that the plume flow is unstable for small value of Grashoff number (Eltayeb and Loper 1994). This dimensionless number measures the strength of the plume, resulting from the maximum amplitude of the basic concentration. Thus if the plume is unstable for small R , it will be unstable for all possible values of R . We can then expand the variables and Ω in the small parameter R as

$$\left.\begin{array}{l} f\left(x, y, z, t\right) = \sum_{s=0}^{\infty} f_s\left(x, y, z, t\right) \; R^s \\ \Omega = \sum_{s=1}^{\infty} \Omega_s \, R^{s-1} \quad , \quad R << 1 \end{array}\right\}, \tag{56}$$

where $f\left(x, y, z, t\right)$ indicates any of the perturbation variables u , v , w , p and T .

Substituting the expressions (56) into the system (39) - (43), (46) - (48) and the associated boundary conditions (51) - (54) and equating the coefficients of R^s (s = 0 , 1 , 2 , ….) to zero, we get systems of ordinary differential equations which can be solved successively to find an expression for the growth rate. The two systems obtained for R^0 (referred to as Problem 0) and R^1 (referred to as problem 1) are sufficient to determine the stability of the interfaces, to leading order.

4.1. Problem 0

The coefficients of R^0 in the system (41) – (43), (46) - (48) are given by

$$Du_0 - m^2 v_0 + w_0 = 0 \; , \tag{57}$$

$$\Delta w_0 + T_0 + n^2 p_0 = 0 \quad , \tag{58}$$

$$\Delta T_0 - w_0 = 0 \; , \tag{59}$$

$$\Delta \varsigma_0 = 0 \; , \tag{60}$$

$$n^2 u_0 = m^2 \varsigma_0 - D\left(w_0 + p_0\right), \tag{61}$$

$$\Delta p_0 - T_0 = 0 \, , \tag{62}$$

associated to the boundary conditions

$$\left.\begin{array}{l} D\left(p_0 + w_0\right) = w_0 = \varsigma_0 = T_0 = 0 \\ \text{at } \; x = a_1 \, , \; x = -a_2 \end{array}\right\}, \tag{63}$$

$$\left.\begin{array}{l} \varsigma_0, w_0, T_0, \, p_0 \, , DT_0 \, , D\left(p_0 + w_0\right) \\ \text{are continuous across } \; x = \pm x_0 \end{array}\right\}, \tag{64}$$

$$\left.\begin{array}{l} \left\langle Dw_0\left(x_0\right)\right\rangle = \left\langle \bar{C}\left(x_0\right)\right\rangle \\ \left\langle Dw_0\left(-x_0\right)\right\rangle = \eta_1 \left\langle \bar{C}\left(-x_0\right)\right\rangle \end{array}\right\}, \tag{65}$$

$$\left.\begin{array}{l} -\mathrm{i}n u_0\left(x_0\right) = \Omega_1 - \mathrm{i}n \bar{w}\left(x_0\right) \\ -\mathrm{i}n u_0\left(-x_0\right) = \left[\Omega_1 - \mathrm{i}n \bar{w}\left(-x_0\right)\right] \eta_1 \end{array}\right\}. \tag{66}$$

The system (57) - (62) subject to the boundary conditions (63) - (65) can be solved straightforward by elimination to obtain the variables of the system. After that, the application of the boundary conditions (66) gives an expression for the growth rate Ω_1 and the displacement of the interface η_1 (see Al Mashrafi and Eltayeb 2014(a)). This leads to

$$\Omega_1 = \frac{\mathrm{i}n}{2}\left(-S_1 \pm \sqrt{\Delta_p}\right), \qquad \Delta_p = S_1^2 - 4S_2 \, , \tag{67}$$

$$\eta_1 = \frac{-N_{j-}}{\left(\Omega_1 / \mathrm{i}n\right) - \bar{w}\left(-x_0\right) + M_{j+}}, \tag{68}$$

in which

$$S_1 = N_{j+} + M_{j+} - \bar{w}(x_0) - \bar{w}(-x_0), \tag{69}$$

$$S_2 = \left\{N_{j+} - \bar{w}(x_0)\right\}\left\{M_{j+} - \bar{w}(-x_0)\right\} - N_{j-}\, M_{j-}, \tag{70}$$

$$\left.\begin{aligned} N_{j\pm} &= \sum_{j=1}^{3} \frac{-\lambda_j F_j}{\sinh(\lambda_j d)}\; S^{(1)}\; C^{(2\pm)} \\ S^{(1)} &= \sinh\{\lambda_j (x_0 - a_1)\}, \\ C^{(2\pm)} &= \cosh\{\lambda_j (x_0 \pm a_2)\} \end{aligned}\right\}, \tag{71}$$

$$\left.\begin{aligned} M_{j\pm} &= \sum_{j=1}^{3} \frac{-\lambda_j F_j}{\sinh(\lambda_j d)}\, S^{(2)}\; C^{(1\pm)} \\ S^{(2)} &= \sinh\{\lambda_j (x_0 - a_2)\}, \\ C^{(1\pm)} &= \cosh\{\lambda_j (x_0 \pm a_1)\} \end{aligned}\right\}, \tag{72}$$

with

$$F_j = \frac{\mu_j^2}{\lambda_j (2\mu_j + 3n^2)}, \tag{73}$$

and μ_j $(j = 1,2,3)$ are the roots of the cubic equation

$$\mu_j^3 + \mu_j + n^2 = 0, \quad \text{where} \quad \lambda_j = \sqrt{\mu_j + b^2}. \tag{74}$$

The properties of the roots of the cubic equation (74) make the expressions $N_{j\pm}$ and $M_{j\pm}$ real. Thus the quantities S_1 and S_2 are real, and then the discriminant Δ_p in the equation (67) is real. If the discriminant is positive, the two roots are distinct and real, while if it is zero, the two roots are real and equal. In either case, Ω_1 is imaginary, and hence the system is neutrally stable and we should investigate the next order of approximation (Problem 1). Here Ω_1 is given by

$$\left.\begin{aligned} \Omega_1^{(a)} &= \mathrm{i}\frac{n}{2}\left(-S_1 + \sqrt{\Delta_p}\right) \\ \Omega_1^{(b)} &= \mathrm{i}\frac{n}{2}\left(-S_1 - \sqrt{\Delta_p}\right) \end{aligned}\right\}, \tag{75}$$

where the superscript a and b indicate two modes.

If the discriminant is negative, then the two roots are complex conjugate numbers. The two values of Ω_1 possess real parts of opposite sign. The root with the positive real part defines the unstable mode. The two roots are given by

$$\left.\begin{aligned} \Omega_1^{(a)} &= \frac{n}{2}\left\{-\sqrt{-\Delta_p} - \mathrm{i}S_1\right\} \\ \Omega_1^{(b)} &= \frac{n}{2}\left\{\sqrt{-\Delta_p} - \mathrm{i}S_1\right\} \end{aligned}\right\}. \tag{76}$$

The amplitudes $\eta_1^{(a)}$ and $\eta_1^{(b)}$ can be evaluated by using the expression (68) to find that

$$\eta_1^{(k)} = \frac{-N_{j-}}{\left(\Omega_1^{(k)}/\mathrm{i}n\right) - \bar{w}(-x_0) + M_{j+}}, k = \{a, b\}. \tag{77}$$

In the absence of the sidewalls, the two modes are such that the two interfaces of the plume are either in phase giving a sinuous solution or out-of-phase giving a varicose solution (see **figure 8**). In both cases, Ω_1 is imaginary and the disturbances are neutral at this level of approximation of the growth rate. The introduction of the boundaries has destroyed the symmetry unless the plumes are situated halfway between the sidewalls. However, if the discriminant is real, the two modes are neutrally stable and propagate with different phase speeds but remain either in-phase or out-of-phase. If the discriminant is negative and the plume is unstable, then $\eta_1^{(k)}$ is complex and hence it has a phase shift so that the two interfaces are neither in-phase nor out-of-phase.

It follows from equations (76) that if one of the two modes has a positive growth rate, the other must have a negative growth rate. This means that one of the two modes can be unstable. Since $n > 0$, then the MS mode is the unstable one. In **figure 9**, we illustrate the behaviour of the MS mode in the wavenumber plane for sample values of x_0, a_1, a_2. It is noticeable that as the plume gets thinner, the area of the wavenumber plane for which instability exists becomes larger, although the maximum growth rate becomes smaller. If the, on the other hand, the thickness of the plume increases, the area of instability shrinks and eventually disappears when x_0 exceeds 0.6. The mode of instability is 2-dimensional and the vertical wavenumber decreases with increasing x_0.

The preferred mode of the instability is identified as the maximum possible value of the growth rates (76) as a function of the wavenumbers m and n for fixed values of the

parameters x_0 , a_1 and a_2 . In other words, we solve the differential equations

$$\frac{\partial}{\partial m}\Omega_1^{(k)}(x_0,a_1,a_2)=0\ ,\ \frac{\partial}{\partial n}\Omega_1^{(k)}(x_0,a_1,a_2)=0\ . \quad (78)$$

The maximisation of the growth rates (76) when $\Delta_p<0$ shows that instability is possible only for the MS mode when a_2-x_0 take values not exceeding 0.25 and the unstable mode is 2-dimensional $(m_c=0)$ and propagates vertically upwards $(U_c>0)$.

A sample of the profiles of the preferred mode is shown in **figure 10**. The growth rate increases from 0 at the boundary to a maximum at a distance, a_0, before it starts to decrease. The distance a_0 depends on the thickness of the plume and on a_2, d . The vertical wavenumber, n_c, decreases as the plume moves away from the wall if the plume is very thin but it increases if the plume is relatively thick.

The region of instability in the (x_0, a_2-x_0) plane is exhibited in **figure 11**. We note that region of instability increases as the thickness of the plume increases, reaching a maximum before it starts to decrease as the thickness of the plume approaches 0. The plume is neutral in the whole space except when the plume is very thin and close to the wall. We investigate the stability of the problem 1.

4.2. Problem 1

The coefficients of R^1 in the perturbation equations give the set

$$\Delta u_1 = M_u\ , \quad (79)$$

$$\Delta v_1 - p_1 = M_V\ , \quad (80)$$

$$\Delta w_1 + T_1 + n^2 p_1 = M_w\ , \quad (81)$$

$$\Delta T_1 - w_1 = M_T\ , \quad (82)$$

$$\Delta p_1 - T_1 = M_p\ , \quad (83)$$

$$n^2 u_1 = m^2 \varsigma_1 - D\left(w_1 + p_1\right) - \bar{\Omega}_1 u_0\ , \quad (84)$$

$$\Delta\varsigma_1 = -\mathrm{i}m\left(w_0 - n^2 v_0\right)D\bar{w}\ , \quad (85)$$

in which

$$\left.\begin{aligned} M_u &= Dp_1 + \left(\Omega_1 - \mathrm{i}n\,\bar{w}\right)u_0 \\ M_v &= \left(\Omega_1 - \mathrm{i}n\,\bar{w}\right)v_0 \\ M_w &= \left(\Omega_1 - \mathrm{i}n\,\bar{w}\right)w_0 - \mathrm{i}n\,u_0 D\bar{w} \\ M_p &= 2\mathrm{i}n\,u_0 D\bar{w} \\ M_T &= \sigma\left(\Omega_1 - \mathrm{i}n\bar{w}\right)T_0 - \mathrm{i}n\sigma\,u_0 D\bar{T} \end{aligned}\right\}\ . \quad (86)$$

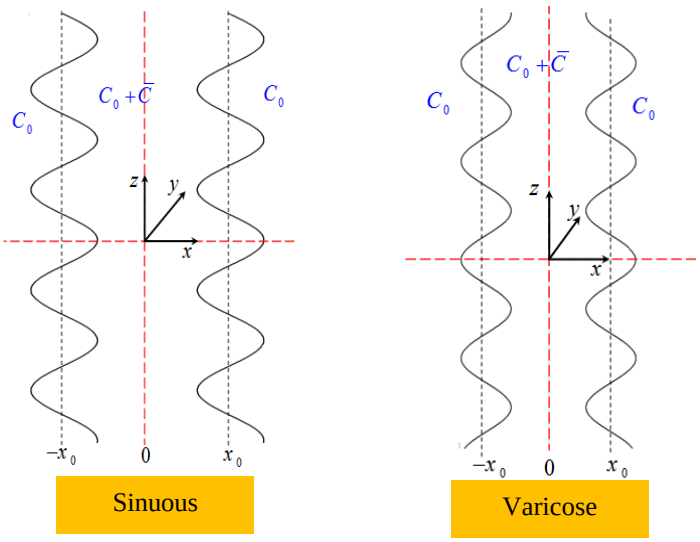


Fig. 8 illustration of the two modes of the interfaces in the absence of boundaries. Note that the presence of boundaries destroyed this symmetry unless the plume half way between the two sidewalls. The new modes refer to modified sinuous mode (MS) and modified varicose mode (MV) in the presence of boundaries.

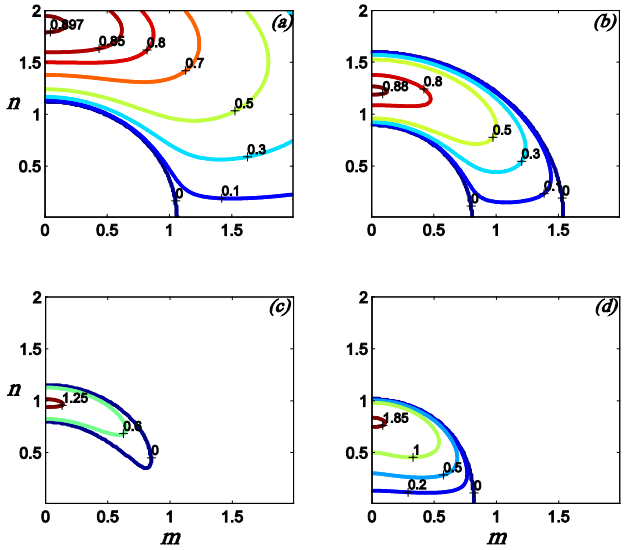


Fig. 9 contours of the growth rate of the MS mode Ω_1 in the (m,n) plane. Here $d=10$, for the pair (x_0,a_2) = (a) $(0.1,0.25)$, (b) $(0.2,0.4)$, (c) $(0.3,0.5)$, and (d) $(0.4,0.55)$. The region outside the curve $\Omega_1=0$ is neutral.

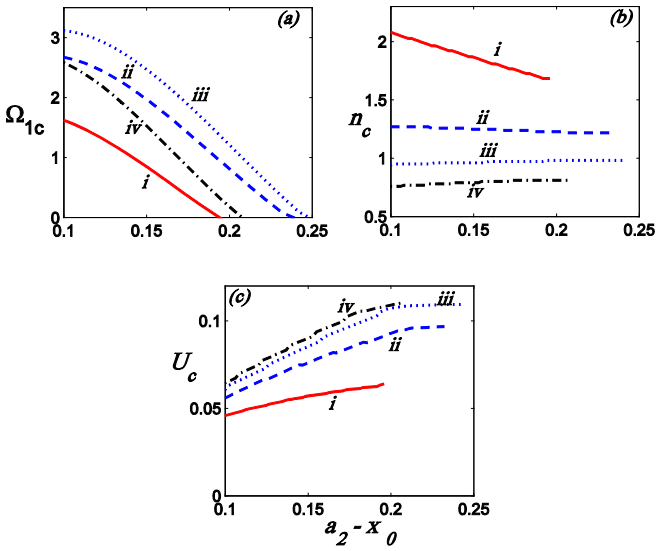


Fig. 10 the preferred mode of instability, in the form of the MS mode, with growth rate of order $O(1)$ as a function of $a_2 - x_0$, for four different values of x_0 ; (i) 0.1 , (ii) 0.2 , (iii) 0.3 and (iv) 0.4 , when $d = 10$.

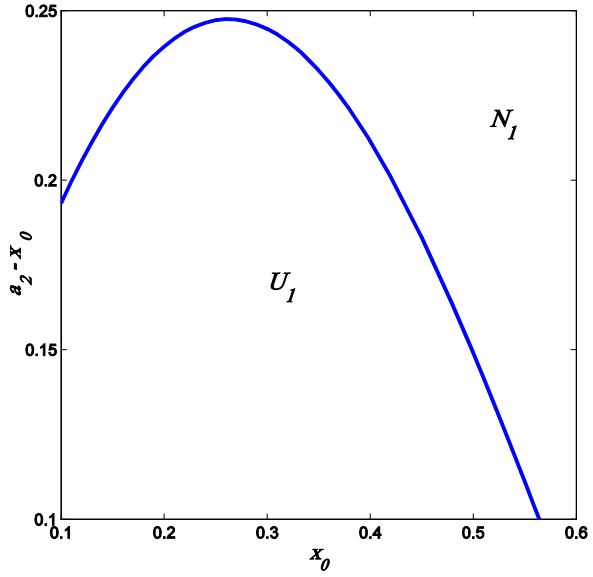


Fig. 11 the regime diagram of the bounded Cartesian plume in the plane $(x_0, a_2 - x_0)$ for the growth rate of order $O(1)$. The region labelled N_1 refers to neutral disturbance, while the region labelled U_1 refers to instability of the mode. The preferred mode here is MS and the mode MV is stable.

The associated boundary conditions are

$$\left.\begin{aligned} v_1 = w_1 = T_1 = \zeta_1 = D(p_1 + w_1) = 0 \\ \text{at } x = a_1 \,, x = -a_2 \end{aligned}\right\}, \tag{87}$$

$$\left.\begin{aligned} v_1, w_1, T_1, p_1 \,, \zeta_1, D\zeta_1, Dv_1, DT_1, Dw_1 \\ , D(p_1 + w_1) \text{ are continuous across } x = \pm x_0 \end{aligned}\right\}, \tag{88}$$

$$\Omega_2 = -in\, u_1(x_0) \;, \quad \Omega_2 \eta_1 = -in\, u_1(-x_0). \tag{89}$$

The equations and boundary conditions (79) - (89) can be solved by deriving the solvability condition for the non-homogenous system (see Al Mashrafi and Eltayeb 2014(a)). The derivation leads to the expression of the growth rate Ω_2 which can be written as

$$\left.\begin{aligned} \Omega_2^{(k)} &= \frac{-1}{1+\left(\eta_1^{(k)}\right)^2} W^{(k)} \\ W^{(k)} &= \Omega_{21}^{(k)} + \Omega_{22}^{(k)} + \Omega_{23}^{(k)} + \Omega_{24}^{(k)} + \frac{\mathrm{i}}{n}\hat{g}^{(k)} \end{aligned}\right\}, \tag{90}$$

and k takes the symbols MV or MS. The expressions $\Omega_{21}^{(k)}$, $\Omega_{22}^{(k)}$, $\Omega_{23}^{(k)}$, $\Omega_{24}^{(k)}$ and $\hat{g}^{(k)}$ are given by following integrals

$$\left.\begin{aligned} \Omega_{21}^{(k)} &= -\,\mathrm{i}\,\Omega_1^{(k)} \int_{-a_2}^{a_1} P_1^{(k)}\, dx \\ P_1^{(k)} &= n\, u_0^{(k)} G^{(k)} - \frac{1}{n} w_0^{(k)}\, H^{(k)} \end{aligned}\right\}, \tag{91}$$

$$\left.\begin{aligned} \Omega_{22}^{(k)} &= \frac{\mathrm{i}\,\Omega_1^{(k)}}{n} \sum_{j=1}^{3} C_j \int_{-a_2}^{a_1} P_2^{(k)}\, dx \\ P_2^{(k)} &= \left\{ w_0^{(k)} + \sigma\, \mu_j\, T_0^{(k)} \right\} H_j^{(k)} \end{aligned}\right\}, \tag{92}$$

$$\left.\begin{aligned} \Omega_{23}^{(k)} &= \int_{-a_2}^{a_1} \left(P_3^{(k)} + P_{32}^{(k)} \right) dx \;;\; P_{32}^{(k)} = \bar{w}\, w_0^{(k)} H^{(k)} \\ P_{31}^{(k)} &= -u_0^{(k)} \left\{ n^2 \bar{w} G^{(k)} + H^{(k)} D\bar{w} \right\} \end{aligned}\right\}, \tag{93}$$

$$\left.\begin{aligned} \Omega_{24}^{(k)} &= \sum_{j=1}^{3} -C_j \int_{-a_2}^{a_1} H_j^{(k)} \left(P_{41}^{(k)} + P_{42}^{(k)} \right) dx \\ P_{41}^{(k)} &= -\bar{w} \left\{ w_0^{(k)} + \sigma\, \mu_j\, T_0^{(k)} \right\} \\ P_{42}^{(k)} &= \left\{ \left(2\mu_j^2 + 1 \right) D\bar{w} - \sigma\, \mu_j\, D\bar{T} \right\} u_0^{(k)} \end{aligned}\right\}, \tag{94}$$

ISBN: 978-1-61804-251-4

$$\hat{g}^{(k)} = g^{(k)}(-a_2) + g^{(k)}(a_1)\,, \tag{95}$$

in which $G^{(k)}$, $H^{(k)}$, $H_j^{(k)}$, C_j , $g^{(k)}(-a_2)$ and $g^{(k)}(a_1)$ are defined by

$$G = \frac{1}{b}\begin{cases} -F_{1b}^{(k)} \sinh[b(x+a_2)]\,; & -a_2 \le x < -x_0 \\ \left(-F_{1b}^{(k)} \sinh[b(x+a_2)] - \eta_1^{(k)} \sinh[b(x+x_0)]\right) & \\ F_{2b}^{(k)} \sinh[b(x-a_1)]\,; & x_0 < x \le a_1 \end{cases} \tag{96}$$

$$H^{(k)} = \begin{cases} -F_{1b}^{(k)} \cosh[b(x+a_2)] & ; \; -a_2 \le x < -x_0 \\ -F_{1b}^{(k)} \cosh[b(x+a_2)] - \eta_1^{(k)} \cosh[b(x+x_0)] & \\ F_{2b}^{(k)} \cosh[b(x-a_1)] & ; \; x_0 < x \le a_1 \end{cases} \tag{97}$$

$$H_j^{(k)} = \begin{cases} -F_{1j}^{(k)} \cosh[\lambda_j(x+a_2)] & ; \; -a_2 \le x < -x_0 \\ -F_{1j}^{(k)} \cosh[\lambda_j(x+a_2)] - \eta_1^{(k)} \cosh[\lambda_j(x+x_0)] & \\ F_{2j}^{(k)} \cosh[\lambda_j(x-a_1)] & ; \; x_0 < x \le a_1 \end{cases} \tag{98}$$

$$C_j = \frac{-n^2}{3n^2 + 2\mu_j}\,, \tag{99}$$

$$g^{(k)}(a_1) = \sum_{j=1}^{3} -C_j\,\mu_j\,F_{2j}^{(k)}\left(DT_1^{(k)}(a_1) + \mu_j\,p_1^{(k)}(a_1)\right), \tag{100}$$

$$g^{(k)}(-a_2) = \sum_{j=1}^{3} -C_j\,\mu_j F_{1j}^{(k)}\left(DT_1^{(k)}(-a_2) + \mu_j\,p_1^{(k)}(-a_2)\right), \tag{101}$$

and $F_{1b}^{(k)}$, $F_{2b}^{(k)}$, $F_{1j}^{(k)}$ and $F_{2j}^{(k)}$ are defined by

$$F_{1b}^{(k)} = \frac{-1}{\sinh(bd)}\left\{\eta_1^{(k)} \sinh(b(a_1+x_0)) + \sinh(b(a_1-x_0))\right\}, \tag{102}$$

$$F_{2b}^{(k)} = \frac{-1}{\sinh(bd)}\left\{\eta_1^{(k)} \sinh(b(a_2-x_0)) + \sinh(b(a_2+x_0))\right\}, \tag{103}$$

$$F_{1j}^{(k)} = \frac{-1}{\sinh(\lambda_j d)}\left\{\eta_1^{(k)} \sinh(\lambda_j(a_1+x_0)) + \sinh(\lambda_j(a_1-x_0))\right\}, \tag{104}$$

$$F_{2j}^{(k)} = \frac{-1}{\sinh(\lambda_j d)}\left\{\eta_1^{(k)} \sinh(\lambda_j(a_2-x_0)) + \sinh(\lambda_j(a_2+x_0))\right\}. \tag{105}$$

The growth rate $\Omega_2^{(k)}$, defined by (90), is the sum of the five terms according to the influence of the parameters and variables of the system. The growth rates $\Omega_{21}^{(k)}$ and $\Omega_{23}^{(k)}$ corresponds to the interactions between the variables and growth rate of the zero order system, while $\Omega_{23}^{(k)}$ and $\Omega_{24}^{(k)}$ corresponds to the interactions between zero order variables and basic state variables. The growth rates $\Omega_{21}^{(k)}$ and $\Omega_{23}^{(k)}$ are independent of Prandtl number, but $\Omega_{22}^{(k)}$ and $\Omega_{24}^{(k)}$ are linearly dependent on σ . The growth rate term $\hat{g}^{(k)}$ is brought about by the boundaries. Since the zero order variables are real and the first order variables are imaginary, then the growth rate $\Omega_2^{(k)}$ is real and hence it will determine the growth rate of the disturbance.

The numerical computations of the growth rate (90) showed that the plume is always unstable at a growth rate of $O(R)$. The maximum growth rate at any particular point in the parameter space (x_0, a_2, σ) can belong to the MS or the MV mode depending in a complicated way on the relative magnitudes of the parameters. As any one parameter is varied keeping the other two fixed, the preferred mode of one type can change to the other mode when the parameter reaches a certain value. Moreover, variations of a parameter can also lead to a mode of particular type (i.e., MS or MV) changing from two-dimensional to three-dimensional, or the reverse, when the parameter increases through a certain value. This is due to the fact that the expression (90) can possess more than one local maximum and as the parameter is increased, the larger of the two maxima decreases and the smaller increases until a value is reached when the smaller one overtakes the originally larger one and becomes preferred. **Figure 12** illustrates such behaviour for a sample of the parameters.

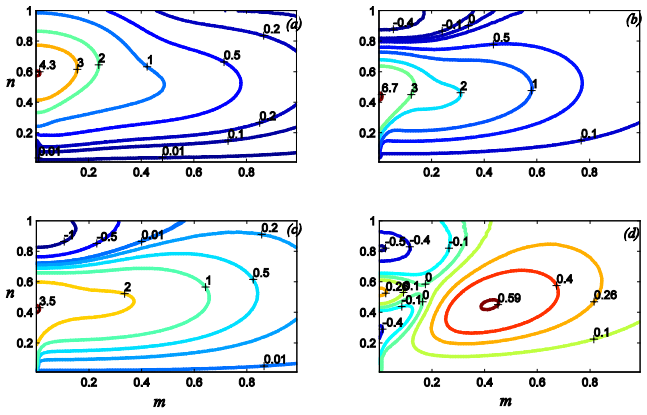


Fig. 12 contours of the growth rate of the modes MV, as in (a) and (c), and MS, as in (b) and (d). Here $x_0 = 2$, $\sigma = 10$, and $d = 10$, and $a_2 = 3$ for (a), (b) and $a_2 = 5$ for (c), (d). (a), (c) refer to the MV mode and (b) , (d) refer to the MS mode. Note that the MS mode is preferred for $a_2 = 3$ and the MV mode is preferred when $a_2 = 5$.

ISBN: 978-1-61804-251-4

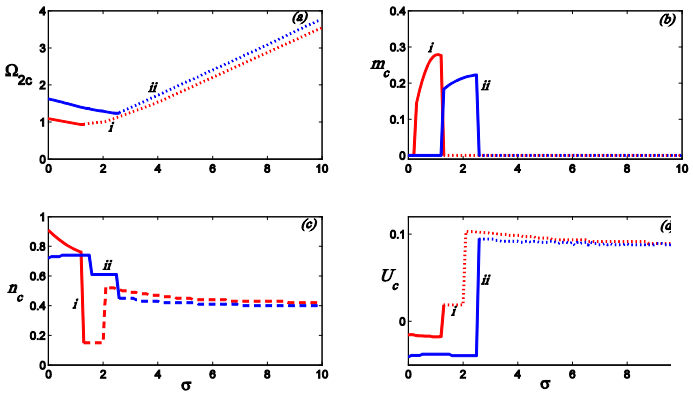


Fig. 13 illustration of the influence of the sidewalls on the stability the Cartesian plume. The preferred mode as a function of Pranc number, σ, when $x_0 = 2$ and the plume is situated halfwa between the sidewalls (i.e., $a_1 = a_2$). The curves i and ii refer to tw different distances between the sidewalls: (i) $d = 10$, ar (ii) $d = 20$. The solid curve refers to the MS mode while the brok one refers to the MV mode.

In **figure 13** we illustrate the dependence of the preferred mode of instability on the Prandtl number, σ, in a way that allows comparison with the limiting case of no sidewalls. For small values of the Prandtl number the MS mode is preferred while the MV mode is preferred for large Prandtl numbers. This agrees well with the case of no sidewalls (Eltayeb and Loper 1994). The value, σ_0 of the Prandtl number at which the mode changes from MS to MV depends on the distance between the plume and the nearest wall. As the sidewall gets closer, σ_0 increases indicating that the presence of the boundaries tends to suppress the MV mode. The presence of the boundaries also tends to stabilise the plume as the growth rate is reduced in magnitude with the decrease in d. It is noteworthy that whatever the values of d or σ, the MS mode is three-dimensional and the MV is two-dimensional when the plume is equidistant from the sidewalls.

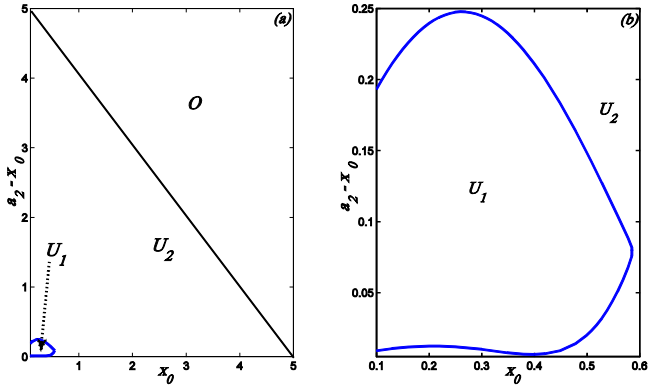


Fig. 14 the regime diagram of the bounded plume in the plane $(x_0, a_2 - x_0)$ for $d = 10$. In (a), the regions labelled U_1 and U_2 refer to instabilities with growth rate $O(1)$ and $O(R)$, respectively. The area O is outside the domain since x_0 cannot exceeds a_2. Subfigure (b) shows a magnification of the area for $a_2 - x_0 \le 0.25$ and $x_0 \le 0.6$ of figure (a).

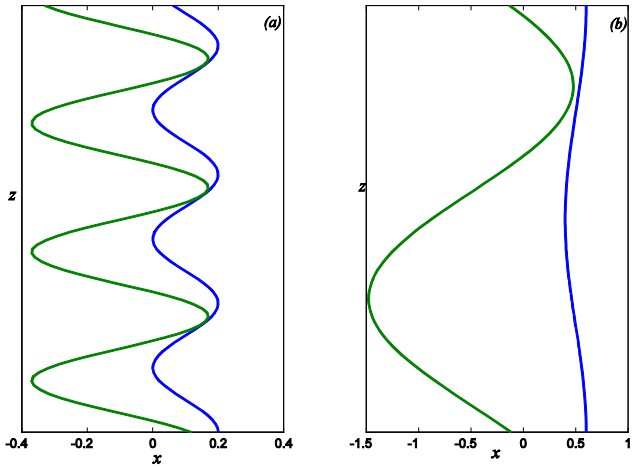


Fig. 15 a sample of the profiles of the interfaces of the unstable mode for two values of the pair (x_0, a_2) when $d = 10$. The profiles are magnified for clarity by the same factor $\varepsilon\,(=0.1)$. (a) $x_0 = 0.1$, $a_2 = 0.2$, $n_c = 2.08$, $m_c = 0$ and (b) $x_0 = 0.5$, $a_2 = 0.6$, $n_c = 0.63$, $m_c = 0$. Note that the interface profiles are very close at regular intervals.

In contrast with the unbounded plume where instability is $O(R)$ everywhere in the parameter space, the instability of the bounded Cartesian plume has instabilities with growth rates of $O(1)$ and $O(R)$. The region in the parameter space where there is instability with the larger growth rate (i.e., $O(1)$), is small and depends on the distance between the walls. In **figure 14**, the regime diagram for the two instabilities is shown when $d = 10$. This instability with growth rate $O(1)$ occurs only if the plume is relatively thin (of thickness not more than about half the salt-finger length scale) and its distance from the sidewall does not exceed about 0.25. We also note that when the plume is very close to the wall the growth rate becomes smaller.

The preferred mode is associated with plume interfaces that are determined by (34) and (37). The amplitude at $x = x_0$ is fixed at the value 1 while the amplitude at the interface at $x = -x_0$ is determined by η_1, which is determined by the parameters of the preferred mode for any prescribed values of

x_0, a_2, σ, d . In **figure 15** we give samples of the profiles of the interfaces relating to some preferred modes. It is noteworthy that the interfaces are very close at regular points across the length of the plume and this may indicate a tendency to break into blobs.

5. Conclusion

The mathematical model of the dynamics of bounded Cartesian plumes has investigated. This work is an extension to previous works on the dynamics of compositional plumes in fluids of infinite extent. In order to get insight into the general influence of the boundaries, we use a simple model in Cartesian geometry, which was investigated in the absence of boundaries by Eltayeb and Loper (1994). Such an arrangement allowed us to compare the two cases with and without boundaries.

The influence of the boundaries on the Eltayeb-Loper model is examined by introducing two vertical walls on either side of the rising column of buoyant fluid. The presence of the sidewalls introduced two more dimensionless parameters to the problem. These are the distance between the two sidewalls, d , and the distance between the plume and the nearest sidewall, a_2. In the absence of the sidewalls, the system possesses a basic state that is even in the distance, x, normal to the interfaces of the plume and perturbations that fall into two uncoupled categories of even (referred to as the varicose or V mode) and odd (referred to as the sinuous or S mode) in x. The presence of walls destroys this symmetry for all plumes that are not equidistant from the walls, but the two modes remain, and are termed the modified varicose (or MV) and the modified sinuous (or MS) .

The basic state of the bounded plume is modified by the presence of the walls and its behaviour depends on the distance between the plume and the nearest wall. The consequence of these is that the stability of the bounded plume is drastically different from that of the unbounded plume. Indeed the analysis of the stability showed that a new mode of instability appears when the plume is close to a sidewall. In contrast to the case of no boundaries where the growth rate is $O(R)$, here the order of magnitude of the growth rate depends on the distance of the plume from the nearest boundary as well as its thickness. If the distance from the boundary is small and the plume is thin, then the instability has a growth rate $O(1)$. For other values of plume thickness and distance from the nearest sidewall, the plume is unstable with a growth rate $O(R)$ similar to that in the absence of the boundaries. Although the order of magnitude of the growth rate in this last situation is the same as that in the absence of the walls, the magnitude of the growth rate is reduced as the distance between the two walls decreases.

References

[1] K. S. Al Mashrafi, and I. A. Eltayeb, " The influence of boundaries on the stability of compositional plumes". Open Journal of Fluid Dynamics, vol. 4, 2014(a), 83-102.

[2] S. I. Braginsky, "Structure of the F-layer and reasons for convection in the Earth's core", Doklady Akademiya Nauk SSSR , 149, 1963, 8-10.

[3] B. A. Buffett, "Geophysics: Taking earth's temperature". Science. 315(5820), 2007, 1801 – 1802, doi:10.1126/science.1140470.

[4] C. F. Chen, and F. Chen, "Experimental study of directional solidification of aqueous ammonium chloride solution". J. Fluid Mech. vol. 227, 1991, 567-586.

[5] S. Classen, M. Heimpel, and U. Christensen, "Blob instability in rotating compositional convection". Geophys. Res. Lett., 26:1, 1999, 135 - 138.

[6] S. M. Copley, A. F. Giamei, S. M. Johnson, and M. F. Hornbecker, "The origin of freckles in unidirectionally solidified castings". Metall. Trans., vol. 1, 1970, 2193 - 2204.

[7] I. A. Eltayeb, "The stability of a compositional plume rotating in the presence of a magnetic field". Geophys. Astrophys. Fluid Dynam., vol. 100, 2006, 429 – 455.

[8] I. A. Eltayeb and E. A. Hamza, "Compositional convection in the presence of rotation". J. Fluid Mech., vol. 354, 1998, 277-299.

[9] I. A. Eltayeb and D. E. Loper, "On the stability of vertical double-diffusive interfaces. Part1. A single plane interface". J. Fluid Mech., vol. 228, 1991, 149 - 181.

[10] I. A. Eltayeb and D. E. Loper, "On the stability of vertical double-diffusive interfaces. Part2. Two parallel interfaces". J. Fluid Mech., vol. 267, 1994, 251 - 271.

[11] I. A. Eltayeb and D. E. Loper, "On the stability of vertical double-diffusive interfaces. Part3. Cylindrical interfaces". J. Fluid Mech., vol. 353, 199745 - 66.

[12] I. A. Eltayeb, E. A Hamza, J. A. Jervase, E. A. Krishnan and D. E. Loper, "Compositional convection in the presence of a magnetic field. II. A Cartesian plume". Proc. R. Soc. Lond A, vol. 461, 2005 2605 - 2633.

[13] A. Hellawell, J. R. Sarazin and R. S. Steube, "Channel convection in partly solidified systems". Phil. Trans. R. Soc. Lond., A345 , 1993 , 507 – 544.

[14] L. N. Howard and G. Veronis, "Stability of salt fingers with negligible diffusivity". J. Fluid Mech., vol. 239, 1992, 511 – 522.

[15] H. E. Huppert, "The fluid mechanics of solidification". J. Fluid Mech., vol. , 1990, 209 - 240.

[16] D. E. Loper, "The gravitationally powered dynamo". Geophys. J. R. Astron. Soc., vol. 54 , 1978, 389 - 404.

[17] D. E. Loper and P.H. Roberts, "On the motion of an iron-alloy core containing a slurry. I". Geophys . Astrophys. Fluid Dyn., vol. 9 , 1978 , 289 – 321.

[18] D. E. Loper and P. H. Roberts, "A study of conditions at the inner core boundary of the Earth. Phys". Earth Planer. Int., vol. 24 , 1981, 302 – 307.

[19] H. K. Moffatt, "Liquid metal MHD and the geodynamo". In Liquid Metal Magneto-hydromagnetics (ed. Lielpeters J. and Moreau R.) , 1989, 403 - 412. Dordecht : Kluwer.

[20] A. K. Sample and A. Hellawell, "The mechanisms of formation and prevention of channel segregation during alloy solidification". Metall. Trans., A 15A, 1984, 2163-2173.

[21] C. E. Smeltzer, "Solve Steel " freckle" Mystery", Iron Age, vol. 184, 1959 , 188-189.

[22] M. E. Stern, "The "salt fountain" and thermohaline convection". Tellus, vol. 12, 1960, 172 – 175.

[23] J. S. Turner, "Buoyancy Effects in Fluids", Cambridge University Press, Cambridge, U.K, 1973.

[24] J. Verhoogen, "Heat balance of the Earth's core". Geophys. J. R. Astron. Soc., vol. 4, 1961, 276 - 281.

[25] M. G. Worster, "Instabilities of the liquid and mushy regions during solidification of alloys". J. Fluid Mech., vol. 237, 1992, 649 - 669.

Pole Shape Optimization in Multipole Magnets

A. Kalimov, P. Nalimov
St. Petersburg State Polytechnic University,
Polytechnicheskaya 29, St. Petersburg, 195251, RUSSIA.

Abstract*—** Efficient method of the pole shape optimization in multipole magnets is analyzed. The method is based on compensating of the high order harmonics in the good field area. The considered optimization strategy is applied for the round and elliptical apertures of the magnets. The developed algorithms are applied for the pole shape optimization in the magnet designing software ***MULTIMAG based on the second order finite element technology.

***Keywords*—** Accelerator magnets, magnetic fields, optimization, particle beam optics.

I. Introduction

MULTIPOLE magnets are widely used in the lines for transporting beams of charged particles. The very important elements of such lines are the focusing magnetic lenses consisting usually of 2 – 3 magnets with the quadrupole rotational symmetry of the magnetic field [1]. To achieve the best focusing effect and avoid undesired distortion of the beam, these magnets should induce the field in the round or elliptical aperture with the flux density magnitude proportional to the radius of the sample point. The requirements to the field quality in the magnet aperture depend strongly on the real application of the lens. For example in the charge particle accelerators the deviation of the magnetic flux density inside the "good field area" from the ideal dependence typically should not exceed several units of 10^{-4} relative units. To provide the desired field configuration a magnet designer should find appropriate position of the coils and the optimal shape of the iron constructions with a very high accuracy. It is convenient to describe the field distribution in the aperture in terms of the amplitudes of high-order harmonics. Moreover just these amplitudes are typically used for analyzing of the beam dynamics in the charge particle optics [1].

Let us consider the magnetic field inside the aperture of the quadrupole magnet. The scalar magnetic potential $U(r,\theta)$ may be expressed in polar coordinates with g being the radius of the magnet aperture by a harmonic series:

$$U(r,\theta) = U_1 \cdot \frac{r}{g}\cos(\theta) + U_2 \cdot \left(\frac{r}{g}\right)^2 \cos(2\theta) + \ldots . \quad (1)$$

It should be noted that strictly saying this expansion is valid only inside a circular area. Application of this formula to the magnetic field distribution inside an elliptical aperture requires additional discussion. For a fourfold symmetric system of a typical quadrupole only the coefficients U_2, U_6, U_{10}, are nonzero. To determine amplitudes of the potential harmonics we consider field properties along a circle of radius R bounding the "good field area". The maximum deviation of the magnetic flux density from the ideal distribution along this border radius is always bigger than similar deviation inside the aperture and so value of this quantity may be used as a criterion of the field quality. The magnetic potential along such a circle may be expanded in a Fourier series to match (1). In the case when:

- the magnet pole is very wide;
- the magnetic permeability of the yoke material is big enough;
- the influence of the coils inducing the primary field may be neglected,

the magnetic potential of the quadrupole magnet is described within the round aperture as:

$$U(r,\theta) = U_2 \cdot (r/g)^2 \cos(2\theta) \quad (2)$$

provided the shape of the pole is hyperbolic defined by the relation in Cartesian coordinate system as:

$$x \cdot y = 2g . \quad (3)$$

In most cases however the ideal quadrupole field distribution (2) is disturbed by the finite dimensions of the pole, the presence of coils, nonlinear magnetic properties of the steel etc. which all cause high-order harmonics.

II. Optimization strategy

The main idea of the proposed optimization strategy is compensation of the high order harmonics by appropriate

A. Kalimov is with the Saint-Petersburg State Polytechnic University, 195251, Saint Petersburg, RUSSIA (phone: +7-950-045-6060; e-mail: alexanderkalimov@gmail.com).
P. Nalimov is with the Saint-Petersburg State Polytechnic University, 195251, Saint Petersburg, RUSSIA (e-mail: pavel_nalimov@mail.ru).

ISBN: 978-1-61804-251-4

modification of the pole shape [3, 4]. The first step of the optimization procedure is approximating the pole tip surface by a curve, corresponding to the desired field Fourier-expansion in the aperture. For this purpose we generate a line of the constant magnetic potential which is described as a superposition of several field harmonics:

$$U(r,\theta) = \sum_{j=1}^{J} P_j \cdot (r/g)^j \cdot \cos(j \cdot \theta) = const \ . \quad (4)$$

Each of the parameters P_j corresponds to the amplitude of one harmonic in the multipole description of the pole surface. The radial coordinates of the points at this line are defined by solving non-linear equation (3) with fixed angular coordinates. Additional restriction on the pole profile may be a fixed width of the pole or the fixed Cartesian coordinates of its extreme points depending on the specific requirements to the magnet design principles. Corresponding examples of the pole tip shape are shown in Fig. 1 – Fig. 2.

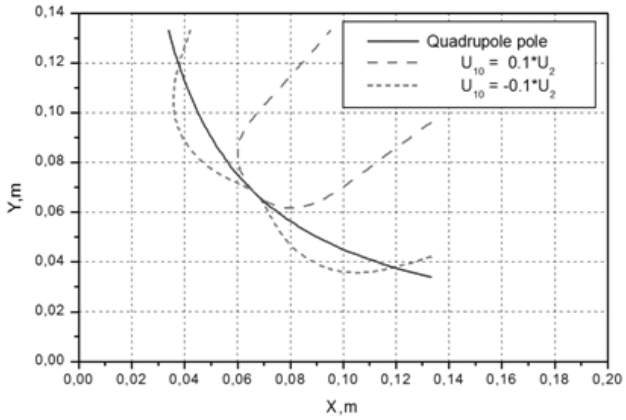


Fig. 1. The shape of the pole tip for the different contributions of the 10-th harmonic. Extreme coordinates of the pole profile are fixed.

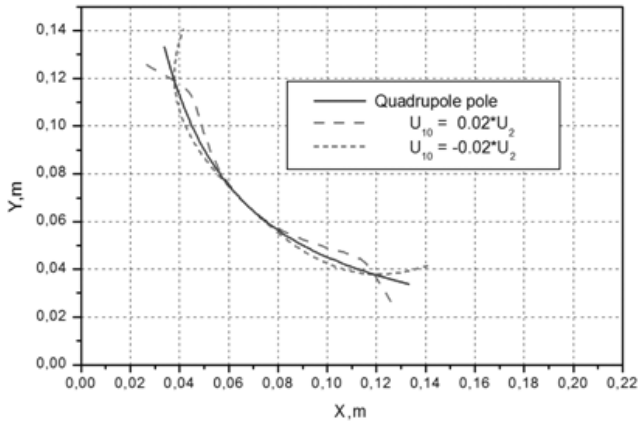


Fig. 2. The shape of the pole tip for the different contributions of the 10-th harmonic. Width of the pole is fixed.

We can notice that keeping constant the pole width restricts strongly a probable diapason of the higher harmonics variation. This circumstance must be taken into account during the procedures of the pole shape variations.

A goal function for the optimization procedure is constructed as a sum of the high order harmonic amplitudes of the magnetic potential:

$$G(P_1, P_2, ...P_M) = \sum_{n=1}^{N} \left| \left(U_n(P_1, P_2, ...P_M) - F_n \right) / U_2 \right| . \quad (5)$$

For a fourfold symmetric system of a typical quadrupole only the coefficients U_2, U_6, U_{10}, U_{14} ... and corresponding field harmonic amplitudes F_2, F_6, F_{10}, F_{14} … are nonzero. The number of the controlled harmonics N is not necessarily equal to the number of optimized parameters M and may be bigger.

Such choice of the goal function gives possibility to exclude high order harmonics almost independently. This property of the goal function parameters is demonstrated in Fig. 3 showing the spectrum of the field components exited by different pole shape harmonics.

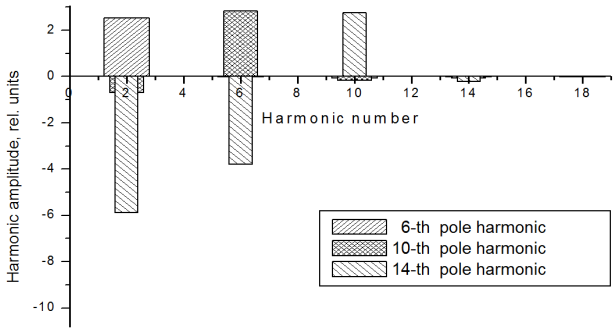


Fig. 3. Field harmonics in the aperture induced by the pole shape harmonics.

We can see that higher pole shape harmonics excite mainly the same and lower field harmonics in the aperture. So starting with the highest pole shape harmonic we can suppress consequently all undesired field components in the aperture. Even better convergence and robustness demonstrates the Newton - Raphson descent method with the adaptive step applied to the solution of this problem. It is important to note that all derivatives necessary for the Jacoby matrix formation are calculated numerically by solving the second order finite element problem for the quadrupole with deformed pole shapes.

III. Optimization of the Pole Shape in the Magnets with a Circular Aperture

The best convergence of the considered algorithm is observed for the magnets with the circular “good field area” –

ISBN: 978-1-61804-251-4

a part of the aperture where the beam of charged particles is supposed to pass. Here we describe main properties of this algorithm. For this purpose we use results obtained for the magnets which are supposed to be installed in the Collector Ring (CR) of the Facility for Anti-proton and Ion Research (FAIR) – a part of a challenging international project started in Darmstadt, Germany [5]. One of the magnet types of this Collector Ring is a so called "narrow quadrupole". This magnet has the aperture of 95 mm radius and a circular "good field area" with the radius of 90 mm. According to the ion optics requirements the deviation of the magnetic field intensity from the ideal quadrupole dependence inside the "good field area" should not exceed margins of $\pm 5 \cdot 10^{-4}$ relative units. To find the desired pole tip shape we used specially developed software ***MULTIMAG*** [2] with the embedded optimization module based on the algorithm described here.

Investigation of the optimization procedure shows that the convergence rate and the final field quality depend strongly on the number of optimized parameters. Our experience shows that for obtaining satisfactory field distribution in the magnet aperture minimization of 3 – 4 amplitudes of the first lower harmonics is usually enough. A number of iterations necessary for the procedure convergence in these cases is typically less than 10. Dependencies of the convergence rate on the number of optimized parameters for the considered magnet are shown in Fig. 4.

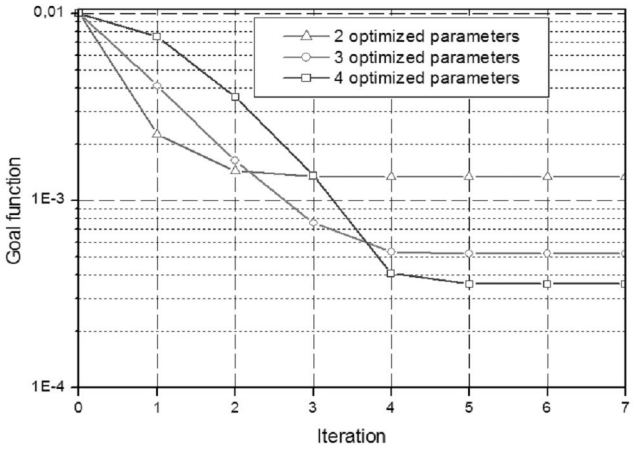


Fig. 4. Convergence of the optimization procedure for the different number of the optimized parameters.

Stabilization of the goal function (4) after several iterations in all dependencies shown in the last plot may be explained by the basic property of the approximation functions formulated above – the optimized pole shape harmonics can not influence on the amplitudes of the higher field components. So the asymptotic value of the goal function is mainly defined by initial contribution of the field harmonics with the order exceeding the highest optimized one.

The best results for the pole tip shape were obtained for the case when the number of the optimized parameters was equal to 6. The achieved field quality in this situation corresponds to the maximum field deviation inside the "good field area" less than $\pm 1.5 \cdot 10^{-4}$ relative units while the 4 optimized parameters allows to reach the maximum field deviation of about $\pm 4 \cdot 10^{-4}$ relative units. The results of the pole optimization are demonstrated in Fig, 5 – Fig, 6.

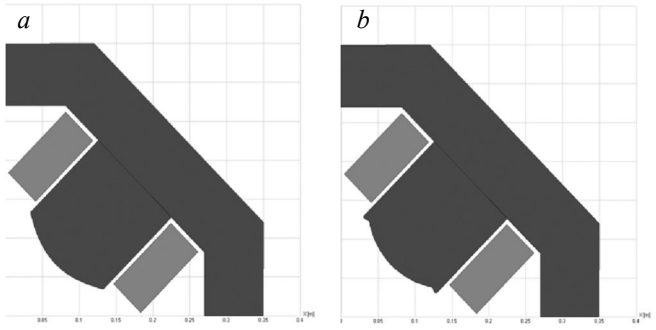


Fig. 5. Cross section of the quadrupole magnet with hyperbolic (*a*) and optimized (*b*) pole profile.

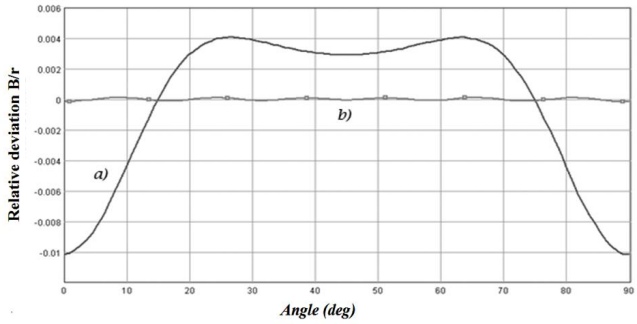


Fig. 6. Field gradient along the circular border of the "good field area" for the hyperbolic (*a*) and optimized (*b*) pole profiles. The maximum flux density deviation for the optimized pole shape is less than $\pm 4 \cdot 10^{-4}$.

We can notice that the field distribution after the pole optimization corresponds to a quickly oscillating high order harmonic with the low amplitude.

IV. Optimization of the Pole Shape in the Magnets with Elliptical Aperture

Sometimes the beam profile does not fit the circular area between the magnet poles. In such a case the "good field area" has a more complicated shape, very often – elliptical. Generally saying the magnetic potential and the field characteristics can not be expanded into series (1) in such situation. Nevertheless the expansion of this form is very desirable because the majority of the widely used methods of the charged particle tracing (especially in accelerators and storage rings) imply possibility of such field presentation [1].

ISBN: 978-1-61804-251-4

So inside the elliptical aperture we can consider only approximation of the magnetic field by harmonic series. Such approximation may be performed in different ways. In [6] the author propose to use special elliptical functions for such procedure. Here we use a direct approximation of the magnetic flux density by a series (1) with finite number of terms. To find the best approximation the harmonic amplitudes are fitted to minimize the difference between the calculated and approximated field values along the border of the "good field area":

$$\sum_i \left[\sum_{n=1} H_i \sin(n\theta_i) r_i^{n-1} - \widetilde{H}_i \right]^2 \rightarrow \min .$$

This problem is solved using least-squares technique. For this purpose the radial component of the magnetic field $\widetilde{H}_{ri}$ was calculated in the points located at the border of the aperture. One of consequences of such field approximation is a noticeable influence of the pole shape harmonics on the higher field harmonics, not only the lower as in the case of the round aperture. Another disadvantage of the used field approximation procedure is relatively strong contribution of the high order harmonics, especially in the case of the big difference between the lengths of the ellipsis axes. These circumstances lead to a worth convergence of the optimization procedure and requires more optimization parameters and several times more iterations to achieve stabilization of the goal function (4) compared to the system with the round "good field area". Nevertheless finally the developed algorithm allows to reach reasonably high field quality in the elliptical aperture as well.

Series of the magnets to be installed in the Collector Ring of the FAIR project [5] has an elliptical "good field area" with the half-axes 200 mm and 90 mm The pole tip radius of the magnet is 145 mm. The required field quality corresponds to the maximum possible field deviation from the ideal quadrupole dependence of $\pm 5 \cdot 10^{-4}$ relative units. An example of the pole shape optimization is shown in Fig,7 – Fig, 8. The convergence was achieved after 20 iterations for 6 optimized pole shape harmonics

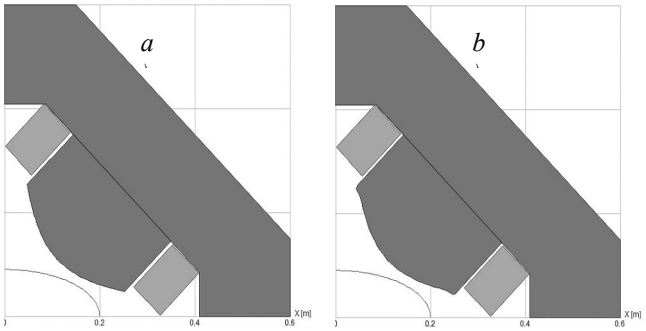


Fig. 7. Cross section of the quadrupole magnet with elliptical aperture with the hyperbolic (*a*) and optimized (*b*) pole profile.

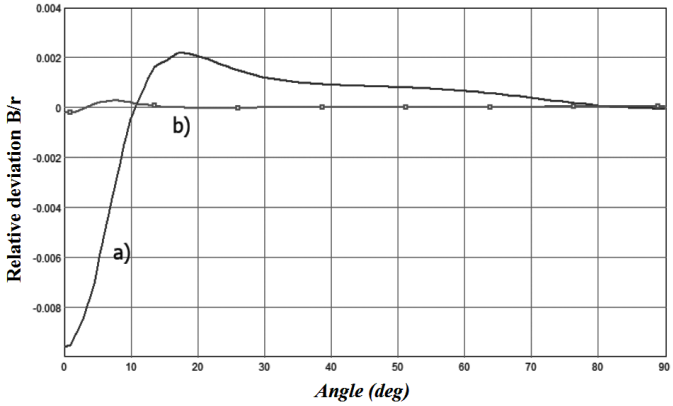


Fig. 8. Field gradient along the elliptical border of the "good field area" for the hyperbolic (a) and optimized (b) pole profiles.

The achieved uniformity of the field gradient along the border of the "good field area" is better than $\pm 4 \cdot 10^{-4}$ relative units. This value is more than 30 times less than initial deviation of this parameter from the average level obtained for the hyperbolic pole.

V. Conclusions

The considered strategy of the pole shape optimization in multipole accelerator magnets with the round and elliptical apertures demonstrates fast and reliable convergence and allows to achieve the high field quality in the aperture. Corresponding algorithm implemented in the software ***MULTIMAG*** helps the user to reduce considerably the time necessary for the design of high performance quadrupole magnets.

References

[1] H. Wollnik, "Optics of Charged Particles," Acad. Press, Orlando, 1987.
[2] A.Chernosvitov, A. Kalimov, H. Wollnik, "Design of a Iron Dominated Quadrupole Magnet with a High Pole-Tip Flux Density", *IEEE Trans. On Applied Superconductivity,* vol.12, No.1, pp. 1430-1433, 2002.
[3] P.R. Sarma, R.K. Bhandari "A new method of finding the pole profile in quadrupole magnets for obtaining high field quality," Review of Scientific Instruments, vol. 69, No 8, 1998, pp. 2909 - 2911.
[4] A. Kalimov, A.Potienko, H. Wollnik, "Optimization of the Pole Shape of Quadrupole Magnets by MULTIMAG", *IEEE Trans. On Applied Superconductivity,* vol. 16, pp. 1282-1286, 2006.
[5] A. Dolinskii, O. Gorda, S. Litvinov, F. Nolden, C. Peschke, I. Schurig, M. Steck, "The CR-RESR Storage Ring Complex of the FAIR Project," Proc. of EPAC-08, Genoa, Italy, pp. 2996 – 2998, 2008.
[6] P.Schnizer, B.Schnizer, P.Akishin, and E.Fischer, "Theory and application of plane elliptic multipoles for static magnetic fields," Nucl. Instr. Meth. A, vol. 607, no. 3, pp. 505-516, 2009.

Implementation of ECDH through Software Code Scheduling with Minimum number of Point Computations

Sakthivel Arumugam.

***Abstract*—**Elliptic Curve Diffie-Hellamn shortly known as ECDH is one of the key exchange which provides a more secure environment for wireless network. The implementation of ECDH has a set of point operations such as point addition, subtraction, multiplication, division and squaring. All these operations are implemented based on two operations. They are point addition and multiplication. But, the time complexity of point addition is lesser than the time complexity of point multiplication. So, it is necessary to find out suitable implementations for point multiplication to take minimum number of clock cycles. Then, it will automatically reduce the power consummation, and also, it is necessary to support for parallel processing. Based on these constraints, the proposed implementations is more suitable for exchanging secret keys between two communication parties in software application areas.

***Keywords*—** Clock Cycles, Elliptic Curve Diffie-Hellman, KeyEexchange, Time Complexity, Parallel Computation and Wireless Network

I. Introduction

Main challenges in the wireless communication are to provide a high security services, more power consume and usage of low bandwidth for communication media [5][2]. Hence, it is very essential to find out suitable algorithm for compromising the same. Considering this, the better implementation of Diffie Hellman using Elliptic Curve over finite field is suggested for key exchange [14]. It is known as Elliptic Curve Diffie Hellman (ECDH), and it is used for Elliptic Curve Cryptography (ECC) and Elliptic Curve Digital Signature Algorithm (ECDSA). An implantation of ECDH is classified into two main categories. They are: ECDH over prime field and ECDH over binary field. In these types, ECDH over prime field (p) is a software approach for key exchange, and ECDH over binary field (2n) for hardware approach [9]. This paper is mainly focused on ECDH over prime field, and it analyzes the point operations for parallel computation based on software scheduling.

The ECDH over prime field is implemented with two point computations. They are: point addition and point multiplication [4]. Here, the execution time of point multiplication is more than the point addition. The point multiplication is defined with two operations called by point addition and point doubling [3]. These operations are computed based on the scalar value (k). For example, the point addition with point doubling is computed sometimes, and some other times the point doubling is only performed. So it is purely based on the probability value of the value of scalar bits. It is an essential to optimize point multiplication for utilization hardware units at minimum level. Some of the available implementations for point multiplication do not support for parallel computation [6]. But, the proposed technique supports for parallel computation through software scheduling to use resources in minimum level.

For this reason, the proposed paper is organized into six sections. The section 2 explains the basics of Elliptic Curve over binary field, and followed by, the section gives the detail survey of literature in the section 3. Then, the proposed work is discussed in section 4. Next, the section 5 analyzes the experimental results briefly and it is compared with other techniques in the section 6. Finally the paper is concluded in the section 7 with future enhancement and applications.

II. Survey of Literature

The survey of literature is conducted for the point multiplication of EC, and the different implementations are analyzed to identify the demerits for parallel computations. Here, the surveys of literature is conducted for four different methods such as linear scalar point multiplication, double and add point multiplication, monotgomery point multiplication and Jacobian point multiplication [10][11][13][9].

The equation (8) is known as linear point multiplication and its procedure is as follows:

This work was supported in part by the U.S. Department of Commerce under Grant BS123456 (sponsor and financial support acknowledgment goes here). Paper titles should be written in uppercase and lowercase letters, not all uppercase. Avoid writing long formulas with subscripts in the title; short formulas that identify the elements are fine (e.g., "Nd–Fe–B"). Do not write "(Invited)" in the title. Full names of authors are preferred in the author field, but are not required. Put a space between authors' initials.

F. A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (corresponding author to provide phone: 303-555-5555; fax: 303-555-5555; e-mail: author@ boulder.nist.gov).

S. B. Author, Jr., was with Rice University, Houston, TX 77005 USA. He is now with the Department of Physics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: author@lamar. colostate.edu).

T. C. Author is with the Electrical Engineering Department, University of Colorado, Boulder, CO 80309 USA, on leave from the National Research Institute for Metals, Tsukuba, Japan (e-mail: author@nrim.go.jp).

procedure linearmultiplication(Point P, Integer k)
1. Integer I, $Q_0 \leftarrow (0,0)$
2. $I \leftarrow 1$
3. if(I = k)
   (a) compare P with Q_0 .
   (b) compute and update P and Q_0 by using equation(4),(5),(6) or (7)
   (c) I←I+1 goto step 3.
4. return P← kP.

The left to right or right to left binary methodologies process a loop scanning bits of the scalar, and it is performing a point doubling followed by a point addition based on the scalar bit value[10]. It is also known as double and add or add and double methodology and its procedure is as follows:

Procedure RL(Point P, Integer k(k_{n-1} to k_0)$_2$)
1. $R_0 \leftarrow P$, $R_1 \leftarrow P$ and $i \leftarrow n-2$.
2. $R_0 \leftarrow 2R_0$.
3. if $k_i = 1$ then
(a) $R_0 \leftarrow R_0 + R_1$.
(b) I ← one time shift right of i go to step 3.
4. return R_0.

Another method of double and add is Montgomery point multiplication. It avoids the main drawback of double and add known as power analysis attack. It means, that the point addition or point multiplication is identified based on the power analysis in each time of iteration [13]. The procedure of this methodology is as follows:

Procedure Mont(Point P, Integer k(k_{n-1} to k_0)$_2$)
1. $R_0 \leftarrow 0$
2. $R_1 \leftarrow P$
3. i← n-2
4. if $k_i = 0$ then
   4.1 $R_1 \leftarrow R_0 + R_1$
   4.2 $R_0 \leftarrow 2R_0$
5. else
   5.1 $R_0 \leftarrow R_0 + R_1$
   5.2 $R_1 \leftarrow 2R_1$
6. I ← one time shift right of i go to step 4.
7. return R_0

Another point multiplication called as Jacobian point multiplication is required more field registers to store intermediate results [9][11]. In this case, the numbers of inversion operations are reduced.

Finally, the survey is concluded and it is given in the following Table 1

Table 1: shows the procedure for ECDH protocol

Methodologies	**Linear**	**Montgomery**	**Double and Add**
Binary length of k value in kP denoted by n	Best Case	2^n -1 times of Point addition with a Point Doubling	n times of point doubling and point addition	n times of point doubling with probability of point addition
	Worst case			
Number of Dependences	Data	More	More	More
	Control	More	More	Less
	Loop carried	More	More	Less
	Register	Less	Less	More
Occurrence of Hazards and Stalls	RAW	YES	YES	YES
	WAR	YES	YES	YES
	WAW	NO	NO	YES
Parallel Processing		NA*	NA*	NA*

N → number of bits

The point addition and point doubling are occurred based on the scalar bits value. It is purely based on the probability of scalar bit value, whether it is one or zero. Based on this, the point multiplication will be calculated. The more number of data dependencies, register dependencies, control dependencies are also identified based on this relationship, which are going to affect the parallel processing. So the speed will not be improved and the number of times will not be reduced. Automatically the lifetime of hardware will be effected [8].

III. Related Work

A. Finite Field over Zp

Theory of Finite Field is a main part of number theory, which is going to use in cryptology and digital signature concepts. It is a combination of Abelain group, commutative ring and field concepts [3]. It is defined with p-1 elements. They are determined based on a prime number [4].

It means that
- *GF(P^n)* →prime field of order n,
- *GF(P^n)* contains P^n elements,
- *GF(P^{n-1})* elements are residue classes modulo P^n.

B. ECDH Over Prime Field GF (p)

En exponential key agreement is also known as the diffie Hellman key agreement, which is used to allow users for exchanging a secret key over an insecure media without any prior secrets [7]. This paper proposes the Diffie-Hellman protocol using Elliptic Curve over prime field with modified

ISBN: 978-1-61804-251-4

point computations for rapid processing. For this purpose, the following general Weierstrass equation [15] is considered:

$$y^2 + a_1 xy + a^3 y = x^3 + a_2 x^2 + a_4 x + a_5 \quad (1)$$

where $a_1, a_2, a_3, a_4, a_5 \in$ coefficient and $x, y \in$ variables.

The equation (1) is simplified as the following for cubic operations to compute points on elliptic curve[15]:

$$y^2 = x^3 + ax^2 + b \quad (2)$$

Based on a and b, the different values of x and y are calculated to plot points on Elliptic Curve. The point may be positive or negative co-ordinate values. So the set of positive points are identified by using the following equation (3) [15][12].

$$\Delta = 4a^3 + 27b^2 \text{ (modulo) } p = 0 \text{ (modulo) } p \quad (3)$$

where p is a prime

The importance of EC is to compute the point multiplication Q = kP, where k is a scalar value and P is a point on the elliptic curve. The point multiplication is computed through two operations such as point addition and point doubling. The point addition is a basic operation of EC between two different points and the point doubling is a fundamental operation of two same points.

P_1 and P_2 are points on Elliptic Curve and the result of $P_1 + P_2$ is also on EC as shown equations (3),(4),(5) and (6) [12].

$P_1=(x_1,y_1)$ and $P_2=(0,0)\in E$ then

$R=P_1+P_2=P_1+O=P_1$ (4)

$P_1=(x_1, y_1)$, $P_2=(x_1,-y_1)\in E$ then

$R=P_1+P_2=P_1+(-P_1)=O$ (5)

$P_1=(x_1, y_1)$, $P_2=(x_2,y_2)\in E$ and $P_1=P_2$ then

$R=P_1+P_2=(x_3, y_3)$

where $x_3=\lambda-x_1-x_2$ and $y_3=\lambda(x_1-x_3)-y_1$

$\lambda=(y_2 - y_1)/(x_2 - x_1)$ where $P_1 \neq P_2$ (6)

$\lambda=(3x_1^2)+a)/2y_1$ where $P_1 = P_2$ (7)

Based on the equations (4),(5),(6) and (7), the point multiplication is performed, and it is denoted by[15][12]:

$$Q = P + 2P + \ldots\ldots + (k-1)P + kP \quad (8)$$

Using these operations, the EC over Diffie Hellman is implemented and it is shown in clearly in the following Table 2 [1].

Table 2: shows the procedure for ECDH protocol

I Common Parameters:	
• Define Ep(a,b) where a,b∈variables, p∈ prime.	
• Find out P=n×p, where n is scalar value.	
II Sender	**III Receiver**
1. Assume private key $(k_A) \rightarrow$ (Secret key)	1. Assume private key $(k_B) \rightarrow$ (Secret key)
2. Compute public key $P_A = k_A \times P$	2.Compute public key $P_B = k_B \times P$
3. Send P_A to Receiver	3.Send P_B to Sender
4. Find out $K=k_A \times P_B$ $= k_A \times k_B \times P$	4. Find out $K= k_B \times P_A$ $= k_B \times k_A \times P$
5. Therefore, $k_B = K / k_A \times P$	5. Therefore, $k_A = K / k_B \times P$

The secret keys are exchanged for ECC and ECDSA applications [3]. In these applications, the public keys are assumed as secret keys [14][12].

C. Code Scheduling for Parallel Computation

The strength of ECDH is based on the difficulty of solving elliptic curve over point operations. These operations are analyzed, based on two criteria. They are: mathematical or computationally secured testing. It is mathematically better than other algorithms [2]. At the same time, it cannot assure to confirm the computationally secured way because of its implementation. when the point multiplication is implemented by using software code scheduling for parallel computation, it will increase execution speed. So the time complexity becomes better then other algorithms. This software code scheduling reduces the different dependencies existing in the point multiplication. They are: data, name, control and loop carried dependencies. All these dependence will affect computations of point multiplications [1].

IV. INNOVATIVE METHODOLOGY

For analyzing point multiplication in ECDH, the scalar value of 'k' is used to create binary and skew trees, and each node has a point value. Based on tree values, the kP is computed. This case, k value is assumed as 15. When the k value is divided by 2 in every time, quotient values 7,3,1 and remainder values 1,1,1 are obtained. Then, the binary tree is created based on the quotient values. The points are used to compute point doubling operation based on the equation (6). Subsequently the skew tree is also formed based on the value of remainders. The points are used to compute point addition on the equation (7). Finally, these two trees are summed by using equations (4), (5), (6) or (7) to compute point multiplication.

Procedure PointCompute(Point P,Integer k)

Point $P_1 \leftarrow (0,0), P_2 \leftarrow (0,0)$

1. If k=1 then One time of P
2. If k>1 then
   (a) $Q \leftarrow k/2$
   (b) if(Q>0)then

P_1←call PointDoubleBinary(Point P) and
P← P_1
(c) R ← k (modulo) 2
(d)if(R=1)then
P_2 ← call PointDoubleSkew(Point P) and
P← P_2
3. k ← k/2 and goto step 2.
4. call Procedure for PointSummazation(Point P_1 , Point P_2)

Procedure PointDoubleBinary(Point P, Quotient Q)

Point Sum ← (0,0)
1. Find out SUM ← P+SUM based on equation (4),(5),(6) or (7)
2. return SUM and this computation is diagrammatically shown in 1.c.

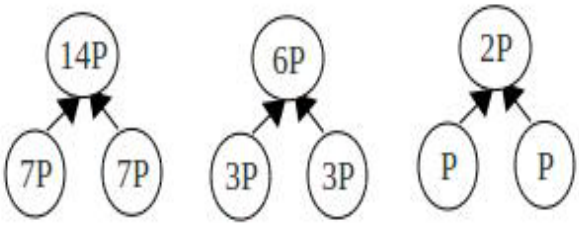


Figure 1.c. Binary Tree Computation of kP

Procedure PointDoubleSkew(Point P, Remainder R)

Point Sum ← (0,0)
1. Find out SUM ← P+SUM based on equation (4),(5),(6) or (7)
2. return SUM and this computation is diagrammatically shown in 1.d.

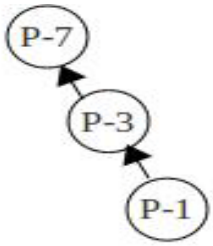


Figure 1.d. Skew Tree Computation of kP

Procedure for PointSummazation(Point 1 , Point P2)

Point Sum ←(0,0)
1. Find out SUM ← P_1 +P_2 based on equation(4),(5),(6) or (7).
2. Display point doubling value for kP.

This innovative tree multiplication methodology minimizes the number of loop carried, data and control dependencies through software code scheduling. These dependencies are going to avoid the different types of hazards and stalls at the time of execution based on its availabilities. The example for hazards and stalls are Write after Read(WAR), Read after Write(RAW) and Write After Write(WAW) to affect loop level parallelism.

V. Experimental Result

The mathematical problem describes the resources, that required by a computer to solve the problem is known as computational complexity theory. This computational complexity is important for many branches in computer science, especially cryptography, digital signature and key exchanges. In this case, the computational complexity of point multiplication is analyzed in the form of time complexity.

For experimental, the $y^2=x^3+ax+b$ Equation is considered to generate a set of points for Elliptic Curve based on the equation 3. The $E_p(a,b)$ is assumed as $E_{11111}(1,1)$. The point (x=0,y=1) is chosen for point multiplication. The simulation environment is given in the Table 3 for both linear scalar and the proposed binary tree multiplication. To measure the number of clock pulses for the proposed work, the system with Intel(R) Core(TM)2 Quad CPU Q800 at 1.33 GHz speed and 1 cache space configuration are considered under the Ubuntu 10.04 version OS.

Table 3: shows the EC parameters for Point multiplication of kP

Parameter	Type	Value
E_p	Input	$E_{(11111)}$
a	Input	1
b	Input	1
x	Input	0
y	Input	1
P	Input	(x,y)
K	Input	Number of times(N)
k_i	Input	0<N<p
kP	Output	E(11111)
kP execution time	Output	Number of clock pulses

The proposed method and linear point multiplication are simulated, and their execution times are measured with clock pulses as given in the Table-4.

Table 4: shows the number of clock cycles needed to compute point multiplication for both linear and proposed cases based on the different value of k and P in the unit of seconds.

calculates kP		Clock pulses		
i	2^i	Best	Worst	Linear
2	4	1	1	3

3	8	2	3	4
4	16	3	5	6
5	32	4	6	13
6	64	5	7	24
7	128	6	9	46
8	256	6	10	91
9	512	7	11	191
10	1024	7	11	380
11	2048	8	12	748
12	4096	9	13	1503
13	8192	9	14	3054

VI. Result Analysis and Comparison

The proposed method is analyzed in three cases where as the linear point multiplication is analyzed in one time, because the difference between worst and best cases in linear point multiplication is one time of computation.

First Case of proposed method is known as best case, and its time complexity is analyzed under optimal conditions. It means that, there is no remainder of k value for all iterations ($k = 2^n$ where $n>0$. Because it takes only $\log_2 n$ times to compute kP based on quotient point computation and no need to compute skew tree computation. It is denoted by $O(\log_2 N)$.
The second case is called by worst case and its time complexity is examined under all possible conditions. It means that, there is a remainder and quotient values of k for all iteration ($k=2^n-1$, where $n>0$). The k value takes $\log_2 n$ times to compute kP based on quotient point computation with $\log_2 n$ times based on remainder point computation. So the time complexity of this case is $O(\log_2 N)+ O(\log_2 N)$ times.

Final case is an average case, in which the way algorithm acts under the probability of execution. It means, that there is a remainder for k values in some iterations and no remainder for some other iterations ($k=2^n$ or 2^n-1, where $n>0$). Its time complexity is defined as $O(\log_2 N)+O(Prob\{\log_2 N\})$.

All time complexities are redefined by $\log_2 N+1$, $2\log_2 N+1$ and $\log_2 N+Prob\{\log_2 N\}+1$. The value of '1' denotes the final computation of combining quotient and remainder point computations to compute kP.

Then the proposed method is analyzed with graph shown in Figure 3(a) and Figure 3(b). In this graph, the y-axis denotes k values in the form of 2^i and x axis total number of execution time for kP.

The proposed strategy is also to reduce the number of dependencies (n) into $\log_2 N+1$ for best case and $2\log_2 N+1$ for worst case. Beside the proposed technique is useful to minimize the number of data dependences, control dependences and loop carried dependences.

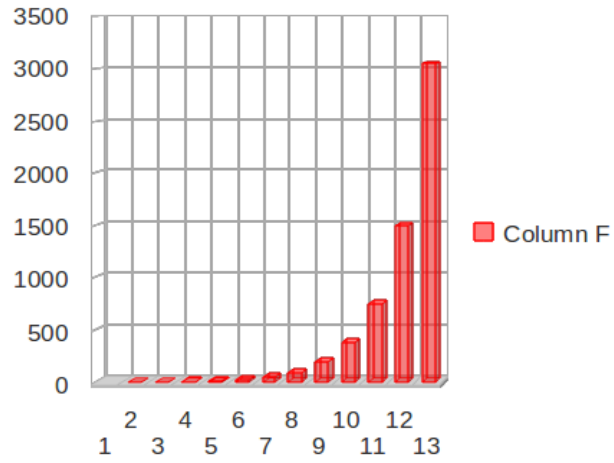


Figure. 3(a). This Graph shows that the total amount of execution time needed to compute kP for Liner point multiplication.

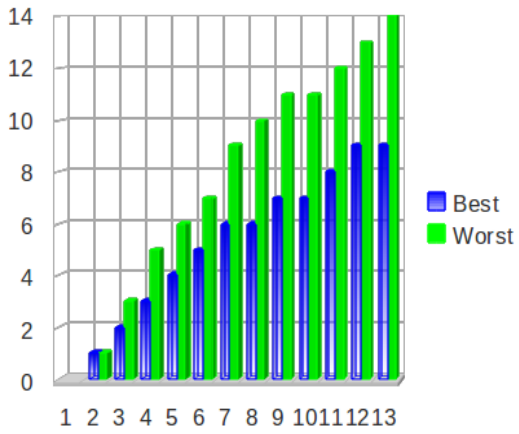


Figure. 3(b). This graph compares that the total amount time needed to compute kp based on best & worst cases of point multiplication.

VII. Conclusion

This proposed point multiplication over prime field is used to exchange key values through insecure channels. It is used for transmitting a secret key of symmetric key cryptography, a public key of asymmetric key cryptography or one time usage of keys in public environment. Because this point multiplication takes less computation time to find out its result. It reduces number of clock pulses, power consumption and increases the performance of EC in wireless environment. It also reduces the number of dependent operations into independent operations to eliminate data hazards and stalls for parallel computation. So the life time of the resources will be increased.

ISBN: 978-1-61804-251-4

Acknowledgments

The author sincerely thanks esteemed reviewers for their valuable comments which have made solving point multiplication significantly using different methodologies based on the probabilities.

REFERENCES

[1] Christian Lederer, Roland Mader, Manuel Koschuch, Johann Großschädl, Alexander Szekely, Stefan Tillich, "Energy-Efficient Implementation of ECDH Key Exchange for Wireless Sensor Networks", Information Security Theory and Practice. Smart Devices, Pervasive Systems, and Ubiquitous Networks-Lecture Notes in Computer Science Volume 5746, 2009, pp 112-127

[2] Kristin Lauter, "The Advantages Of Elliptic Curve Cryptography For Wireless Security," in Proc. IEEE Wireless Communications, Feb 2004, pp 62-67.

[3] Daniyal M. Alghazzawi, "A Novel Password Based Multi Party Key Agreement Protocol On Elliptic Curve," International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 1, Feb 2012, pp 75-83.

[4]. Ram Ratan Ahirwal and Manoj Ahke, "Elliptic Curve Diffie-Hellman Key Exchange Algorithm for Securing Hypertext Information on Wide Area Network," International Journal of Computer Science and Information Technologies, Vol. 4 (2) , 2013, 363 - 368

[5].Moshaddique Al Ameen and Kyung-sup Kwak, "Social Issues in Wireless Sensor Networks with Healthcare Perspective," The International Arab Journal of Information Technology, Vol. 8, No. 1, January 2011, pp. 52-58

[6].Patrick Longa and Ali Miri, "Fast and Flexible Elliptic Curve Cryptography point arithmetic over Prime fields," *IEEE Transactions on computers* vol.57,No.3,pp.289-302. May.2008.

[7]. Ranjan Bose, Information Theory, Coding and Cryptography,TMCH Edition, 2008.

[8].L.Hennessy and David A. Pattersom, Computer Architecture a Quantitative Approach,*Elsevier*,4th Edition.2007

[9].Adnan Abdul-Aziz, "Area Flexible GF(2^k) Elliptic Curve Cryptography Coprocessor," The International Arab Journal of Information Technology, Vol. 4, No. 1, Jan 2007, pp. 1-10.

[10].Adnan Abdul-Aziz Gutub, "Fast 160-Bits GF (P) Elliptic Curve Crypto Hardware of High-Radix Scalable Multipliers," The International Arab Journal of Information Technology, Vol. 3, No. 4, Oct 2006, pp. 342-349.

[11].Essam Al-Daoud, "An Improved Implementation of Elliptic Curve Digital Signature by Using Sparse Elements," The International Arab Journal of Information Technology, Vol. 1, No. 2, July 2004, pp. 203-208

[12].Arockia Jansirani, Rengansivagurunathan Rajesh, Ramasamy Balasubramanian, and Perumal Eswaran, "Hi-Tech Authentication for Palette Images Using Digital Signature and Data Hiding," The International Arab Journal of Information Technology, Vol. 8, No. 2, April 2011, pp. 117-123.

[13].Lo'ai Tawalbeh, Yaser Jararweh, and Abidalrahman Moh'd, "An Integrated Radix-4 Modular Divider/Multiplier Hardware Architecture for Cryptographic Applications," The International Arab Journal of Information Technology, Vol. 9, No. 2, May 2012, pp. 284-290.

[14].Qiong Pu1 and Shuhua Wu3, "Secure and Efficient SIP Authentication Scheme for Converged VoIP Networks," The International Arab Journal of Information Technology, Vol. 9, No. 6, Nov 2012, pp. 553-561.

[15].Debiao He and Jianhua Chen, "An Efficient Certificateless Designated Verifier Signature Scheme," The International Arab Journal of Information Technology, Vol. 10, No. 4, July 2013, pp. 389-396.

Dr. A. Sakthivel is working as Professor and Head in Information Security and Cryptography Centre, Information Technology, Coimbatore for past four years. His date of birth is 01-08-1975 in Tamilnadu, India. He Completed Bachelor of Engineering from Bharathiyar University, Mater of Engineering from Manonmaniyam Sundaranar University and Ph.D from Anna University, Tamilnadu, India in the field of computer science and engineering.

The author has 15 years academic experience in different colleges such as SNS College of Technology and Sri Ramakrishna Engineering College. He has published 8 papers in international journals, 2 papers in international conference and 3 papers in national conferences. He is also edited a chapter named Exception handling Object oriented programming with C++ written by E. Balagurusamy and it is published by TMCH as 5th edition 2012. He is also a reviewer of International Arab Journal of Information Technology Indexed by ISI and reviewed more than 50 Articles. He got two times best reviewer ward from the same journal in 2012 and 2013.

The author is a life member of Indian Society for Technical Education, Advanced Computing and Communication Society and International Association for the engineers and Computer Scientist IAENG.

Computer modelling of hydropower-driven systems with thermal and electric energy sources

A. I. Ozerskij

***Abstract*—**In the report the methodology and results of computer modelling of the difficult hydropower-driven systems equipped with primary energy sources of two types are stated: explosive motors and the electric motors working under trying conditions of maintenance. Last together with hydropower-driven systems are modelled as uniform heat-electrohydromechanical systems, optimum control with which can be carried out on a basis mechanotronic the approach and a principle of a maximum of L.S. Pontrjagin.

***Keywords*—**computer modeling, hydropower-driven systems, thermal and electric energy sources

1. Introduction. Hydropower-driven systems (HDS) are highly effective and perspectiv power systems which are applied in power engineering, engineering industry, on transport, and also in other branches of economy practically everywhere. Complexity of conditions at which these systems are frequently maintained, creates the problems connected with their profitability, power- and of resource saving, and also reliability in-process, especially - under trying conditions maintenance: a dustiness and gassed conditions of air medium, at frequent and considerable overloadings, vibration, etc. 90 % of emergencies of the hydraulic drive of cars, result from pollution of their operating fluids [1].

The analysis of perspectiv directions of raise of reliability and profitability investigated HDS showed expediency “ampoulisition” their hydraulic systems (HS) by a complete retention of their operating fluids and gases from a circumambient and air replacement as working medium on nitrogen [1].

Researches showed that perfection HDS with “ampoulisition” hydraulic systems (AHS) is possible on a basis not only their full-scale tests, but also computer modelling together with sources of external loadings and energy sources: explosive motors (EM) and electric motors (EM) as uniform mechatronic warmly – electrichydromechanical systems [1]. The last allows to carry out the complex approach to raise of reliability and profitability of such systems not only a constructive way, but also by perfection of their computer models. It gives the chance to replace an essential part of expensive full-scale tests of systems investigated here with calculations with the computer, and will promote sampling of the optimal solutions of the basic design problems. Researches showed that perfection of computer models of such difficult systems demands creation of new methodology of their computer modelling answering to modern level of knowledge of features of teamwork thermal, electric and the hydraulic machiner (HM) under trying conditions of maintenance [1]– [6].

2. A statement of problem. Modern level of knowledge of features of work investigated here AHS shows that the physical processes accompanying their work on not settlement and dynamic regimes, essentially it is more difficult than processes, characteristic for work usual - not “ampoulisition” – systems. Complexity of processes of start AHS speaks in small relative volumes (nearby 1 %) gas in a gas cavity (a gas pillow) basic forecastle for operating fluid storage. This feature can lead to vacuum emersion in this forecastle and call break-down of work of pumping up and basic pumps not only at system start, but also on design conditions of its work. Complexity of processes of start of system speaks as well that it occurs at filling with operating fluid of hydraulic channels of pipelines and system hydrocars. Work of all system occurs at variable extent of filling of a forecastle operating fluid. Therefore, unlike method of calculation HDS based on the way of Euler traditionally used in hydrodynamics, allowing to solve problems of hydraulics with motionless boundary lines of continuous medium as problems with the algebraic nonlinear equations, here the new method of calculation based on the way Lagranzha which is the more general method is developed for modelling of hydrodynamic processes. This method allows to solve problems with mobile boundary lines of medium as problems for the ordinary nonlinear differential equations. In the capacity of unknown persons here we observe co-ordinates of these mobile boundary lines. Generally these problems are reduced to problems of a hydromechanics with mobile contact ruptures of two types: a liquid - gas and a liquid - a solid. Ruptures can move in channels of hydraulic highways, elements and assemblies of difficult geometrical forms with jet, blade and volume HM [2]. The new method of calculation matches to specificity of tasks in view and expands area of justice of their solution. The developed method allows to create new directions of perfection physical, mathematical, and also computer models HM and HDS. The specified models differ that characteristics not only static, but also dynamic processes of traffic of liquid

ISBN: 978-1-61804-251-4

medium with mobile boundary lines in channels of hydraulic highways, and also in channels volume, blade and jet HM allow to count. Models are built on the general integrated relationships for moving in space and changing in a time t ranges of integration: volume and $V(t)$ the square of a surface $\sigma(t)$ of the medium, consisting of the same corpuscles of a liquid. These relationships an essence: Law of conservation of mass, energy, a pulse (momentum) and an angular momentum (momentum) for liquid medium with mobile boundary lines [3]:

$$\frac{d}{dt}\int_{V(t)}\rho dV=\int_{V(t)}\frac{\partial\rho}{\partial t}dV+\int_{\sigma(t)}\rho\upsilon_n d\sigma=\sum_{(i,j)}\frac{d}{dt}M_{i,j}, \quad (1)$$

$$\frac{d}{dt}\int_{V(t)}\rho\bar{\upsilon}dV=\int_{V(t)}\frac{\partial}{\partial t}(\rho\bar{\upsilon})dV+\int_{\sigma(t)}(\rho\bar{\upsilon})\upsilon_n\cdot d\sigma=$$
$$=\int_{V(t)}\bar{f}\rho dV+\int_{\sigma(t)}\bar{p}_n d\sigma+\sum_{(i,j)}\bar{K}_{i,j}, \quad (2)$$

$$\frac{d}{dt}\int_{V(t)}\rho(\frac{1}{2}\upsilon^2+u)dV=\int_{V(t)}\frac{\partial}{\partial t}\left[\rho(\frac{1}{2}\upsilon^2+u)\right]dV+\int_{\sigma(t)}\rho(\frac{1}{2}\upsilon^2+u)\upsilon_n d\sigma=$$
$$=\int_{V(t)}\rho\bar{f}\cdot\bar{\upsilon}dV+\int_{\sigma(t)}\bar{p}\cdot\bar{\upsilon}_n d\sigma+\int_{V(t)}p\cdot div\bar{\upsilon}\cdot dV+\frac{dQ}{dt}+\sum_{i,j}N_{i,j}, \quad (3)$$

$$\frac{d}{dt}\int_{V(t)}(\bar{r}\times\rho\bar{\upsilon})dV=\int_{V(t)}\frac{\partial}{\partial t}(\bar{r}\times\rho\bar{\upsilon})dV+\int_{\sigma(t)}(\bar{r}\times\rho\bar{\upsilon})\upsilon_n d\sigma=$$
$$=\int_{V(t)}(\bar{r}\times\rho\bar{f})dV+\int_{\sigma(t)}(\bar{r}\times\bar{p}_n)d\sigma+\sum_{(i,j)}\bar{L}_{i,j}. \quad (4)$$

Here $\upsilon, m/s$ – a vector of absolute speed of traffic of a liquid in the given point; $\rho, kg/m^3$ – liquid density; $\bar{f}, m/s^2$, and $\bar{p}=\bar{p}_n+\bar{p}_\tau, N/m^2$ – accordingly, vectors tensions the mass and superficial forces acting on elements of mass and surfaces, restricting mobile liquid medium; u, Joule/kg – specific internal (thermal) energy of a liquid; $\frac{dQ}{dt}, W$ – power of thermal energy brought to a liquid from the outside; the $\sum_{i,j}\frac{dM_{i,j}}{dt}, kg/s, \sum_{i,j}N_{i,j}, W$ – sums of powers of additional sources and weight and energy flows, accordingly, had in the observed liquid medium; the, $\sum_{i,j}K_{i,j}, kg\cdot m/s, \sum_{i,j}L_{i,j}, N\cdot m$ – sums of additional sources and flows of a momentum and the angular momentum of a liquid which is in a liquid; $\bar{r}, m$ - radius - the vector spent to the given point of a liquid from point O* – the centre of twirl of medium (fig. 1.).

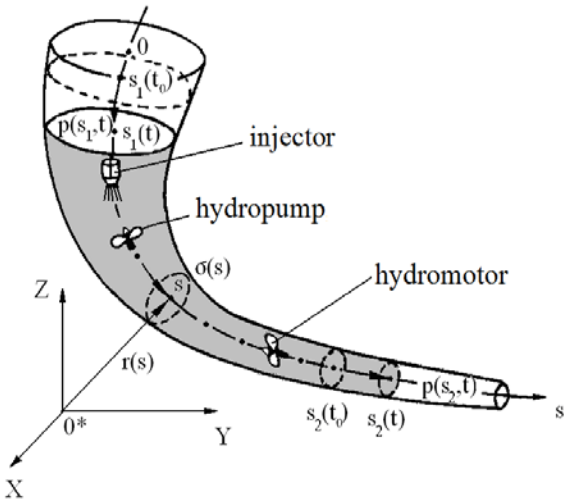


Fig. 1 To a statement of problem of calculation of one-dimensional traffic of liquid medium with contact ruptures in channels with the hydraulic machiner.

In an one-dimensional case observed here (fig. 1.) modelling of dynamic operating modes HDS is reduced to the consecutive solution of some Cauchy problems for systems of the ordinary nonlinear differential equations second aspect usages:

$$A_1(x)\frac{d^2x}{dt^2}+B_1(x)\frac{d^2\varphi}{dt^2}+C_1(x)\left(\frac{dx}{dt}\right)^2+Д_1(x)\left(\frac{d\varphi}{dt}\right)^2+$$
$$+E_1(x)\frac{dx}{dt}\frac{d\varphi}{dt}=P(x,t)-\Delta P_{LOSSES}$$

$$A_2(x)\frac{d^2x}{dt^2}+B_2(x)\frac{d^2\varphi}{dt^2}+C_2(x)\left(\frac{dx}{dt}\right)^2+Д_2(x)\left(\frac{d\varphi}{dt}\right)^2+$$
$$+E_2(x)\frac{dx}{dt}\frac{d\varphi}{dt}=M(t)-M_{\text{FRICTION}} \quad (5)$$

the contact ruptures allowed concerning mobile coordinates x (t) =s_1 (ξ_1, t) or x (t) =s_2 (ξ_2, t), and also $\varphi=\varphi(t)$ angles of rotation of shafts blade or volume: pumps, turbines, etc.; $\omega=\omega(t)=\frac{d\varphi}{dt}$ angular speed of the

shaft HM**.** Here in the equations (5): $P(x,t)$ – a difference of average values of the fluid pressure, reacting on surfaces of mobile borders of a stream; ΔP_{LOSSES} – pressure losses; $M(t)$ – the moment of awake forces; M_{FRICTION} – the moment of frictional forces.

Generally modelling HDS with energy sources in the form of electric synchronous and asynchronous motors, and also motor internal combustion MIC to system of the equations (5) the system of the differential equations allowed concerning electric currents of windings of a rotor and the stator of electric cars, and also the system of the algebraic and differential equations presenting dynamic operating modes MIC is added. Feature of the methods of dynamic calculation HDS resulted here and their difference from known methods consists that they allow to count difficult processes of filling and dump of channels of hydraulic highways and cars GPS. These processes accompany regimes of overloadings, start, stop, cavitation operating modes HDS, etc. Methods of calculation of piston hydraulic and thermal cars differ from known themes that features of work of each piston of the car here are considered. On the one hand it complicates calculations, but with another - allows to reveal agency of the separate piston on dynamics of the car as a whole that raises worth of the gained results and quality of research.

3. The problem solution. On the basis of presented above methods of calculation and systems of the equations the methodology of computer modelling HDS with thermal and electric energy sources has been developed. The created models have been realised in the licensed mathematical medium Mathcad. Adequacy of models is confirmed by the data of full-scale tests AHS, and also researches on the experimental complexes modelling the power hydraulic drive with specified systems (fig. 2,3).

Fig. 2 An aspect of an experimental complex.

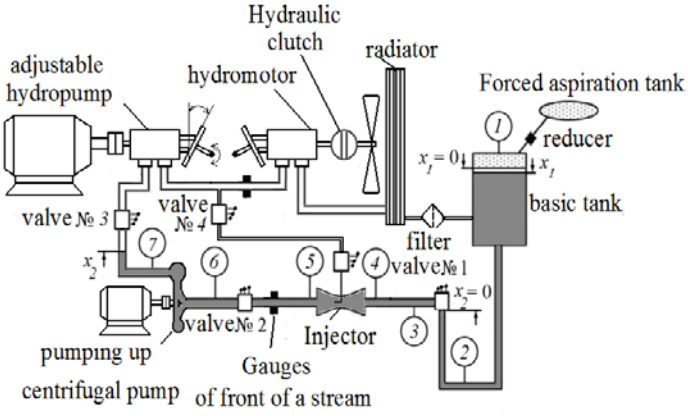


Fig. 3 The settlement circuit design of filling with operating fluid of channels of hydraulic highways and cars at start electrichydrodriven systems.

AHS consisted of a forecastle with nitrogen, the compressor (vacuum pump), an inspirator and pumping up lobed the pump with the electric motor. Into composition of the power drive also entered: the volume hydropump, the volume hydromotor, and also the hydroclutch with the ventilating fan of an integral cooling system of injection engine KamAz-740. In model HDS regulating by means of an inspirator having a nozzle with adjustable diameter is provided. Results of modelling and empirical data are resulted on fig. 4. Researches of processes of start of system on a rate of inflow were carried out by operating fluid of its hydraulic channels at consecutive start of an inspirator and loded the pump. On fig. 4,5 results of calculations (full lines) and the data of experience (dashed lines) are presented. On fig. 4 the data about change of pressure of gas in a gas cavity “ampoulisition” a forecastle is presented. An initial gauge pressure of gas in a tank: $P_{GAS} = 0,5MPa$. Figures here note the curves corresponding to various initial values of relative

volume (in %) a gas cavity of a tank. On fig. 5 the picture of change of pressure P_4 before an input in an injector is represented at different values of time of start of the centrifugal pump and an injector. Initial superfluous pressure of gas in a tank: $P_{GAS} = 0,1MPa$.

On fig. 6 the picture of computer modelling of dynamic loudspeaker of start and regulating HDS with volume HM by change of diameter of a jet orifice is shown. On fig. 7 it is shown a time history of speed of a stream of operating fluid in the channel of the delivery pipe after the aksialno-piston hydropump HDS at its start from asynchronous EM (electric motor) (fig. 3) [9].

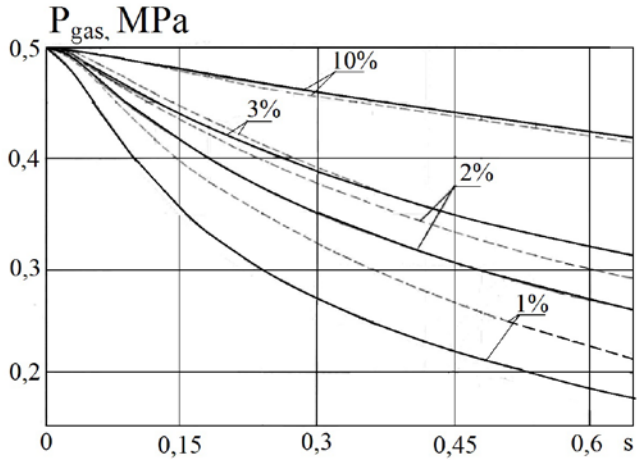


Fig. 4 Change of pressure in the isolated gas cavity of a tank at start-up.

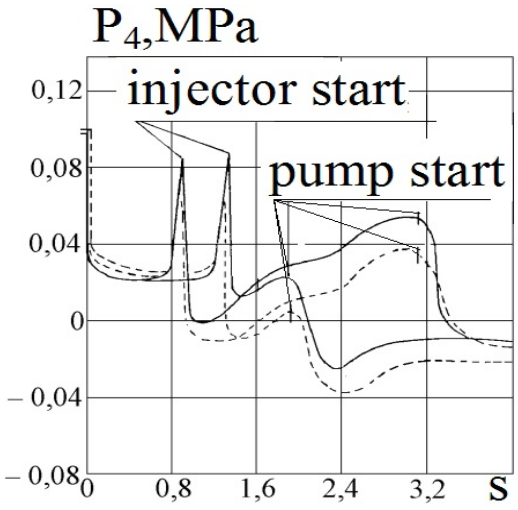


Fig. 5 The Inlet pressure in an inspirator.

Physical, mathematical and computer models of processes of traffic of operating fluids in the hydrodynamic drive are developed: hydroclutches with bled and filled at overloadings lobed the hydraulic machiner: pumps and turbines. Models are created on the basis of the approach of Lagranzh and the general integrated relationships (1)-(4).

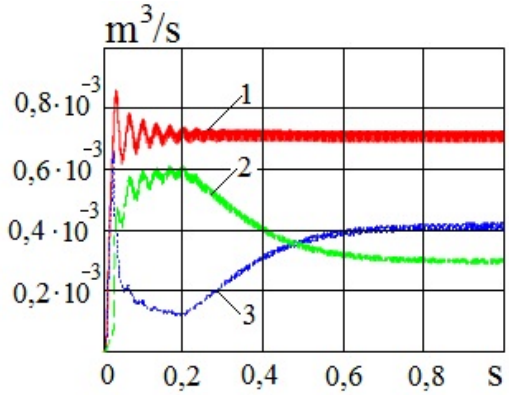


Fig. 6 The operating fluid charge: 1 – after the pump; 2 – and 3 – in injection and motor highways.

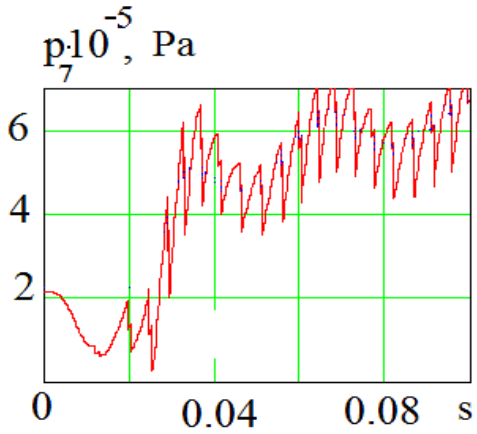


Fig. 7 Pressure upon an input in the volume pump at its start from asynchronous EM. Is shown influence of each piston.

Physical, mathematical and computer models of processes of traffic of operating fluids in the hydrodynamic drive are developed: hydroclutches with bled and filled at overloadings lobed the hydraulic machiner: pumps and turbines. Models are created on the basis of the approach of Lagranzh and the general integrated relationships (1). They allow to count to dynamics of traffic of liquid medium with mobile boundary lines. Mathematical models are shown to the solution of a Cauchy problem for system of the ordinary differential equations of the second order of an aspect (5) concerning working radiuses changing in a time r_1 (t) and r_2 (t) blade hydrocars (pumps and turbines), and also angles $\varphi_1(t)$ and $\varphi_2(t)$ turn of the power shaft by everyone blade hydroclutch cars. On their basis in mathematical medium Mathcad 13 computer models of the hydrodynamic drive are created. The closed fluid couplings of separate geometric series (fig. 8a), presented by V.N.Prokofevym were modelled. The analysis of models is resulted in [4]. Models of hydroclutches of the given row with static and dynamic emptying itself (traction and limiting), working together with MIC and EM under trying conditions maintenance [4, 6] are created. Models are based on the account (by means of empirical data) an energy loss of operating fluid

ISBN: 978-1-61804-251-4

of the hydroclutch on various static regimes of its work [4]. At construction of models of hydroclutches of the given row empirical data about the basic opeating characteristic of the hydroclutch were used: dependences of factor λ a torque of the hydroclutch from a slip ratio ε its vane wheel rotors: ($\lambda = \lambda(\varepsilon)$ fig. 8b). Thus skilled tabular function $\lambda = \lambda(\varepsilon)$ was approximated by means of splines in the environment of Mathcad 13 and in the further calculations was used as an analytic function.

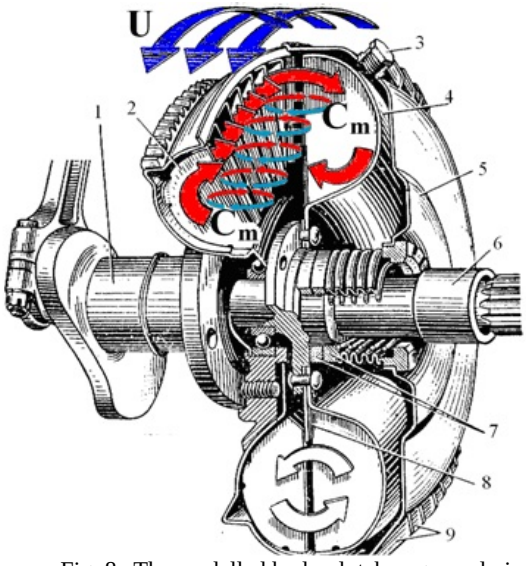


Fig. 8a The modelled hydroclutch: a general view.

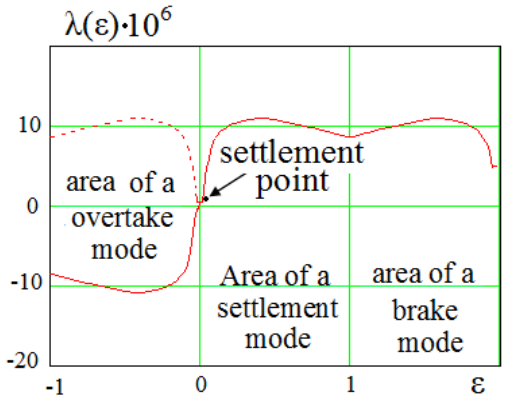


Fig. 8b The loading characteristic hydroclutch.

Adequacy of models EHMS is shown by comparison of results of numerical experiments with known empirical data about features of work of investigated hydroclutches on various operating conditions: settlement, stop, overtake and brake regimes (fig. 9a, 9b).

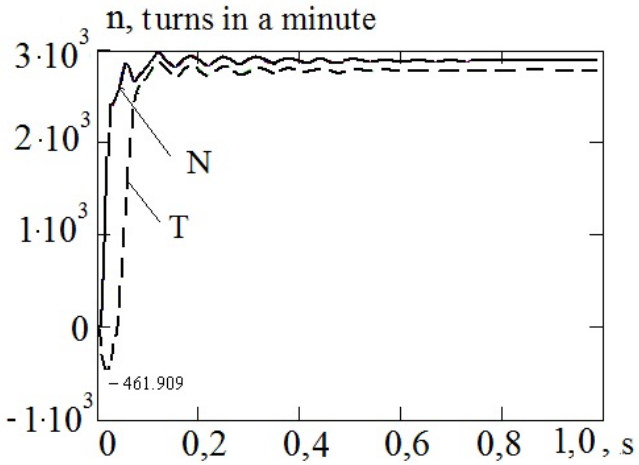


Fig. 9a Start EHMS with a steady load.

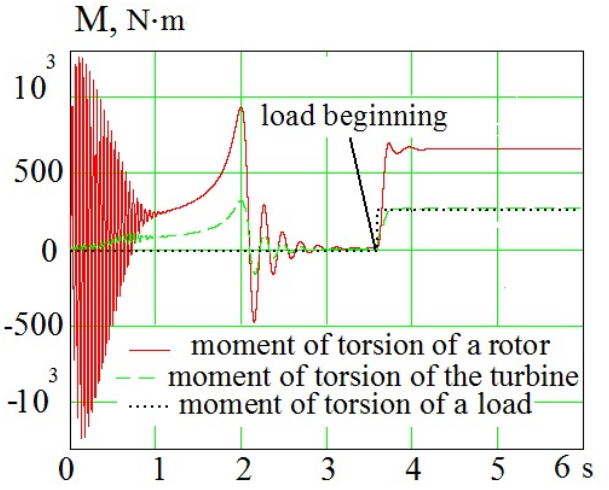


Fig.9b Exit on idling and load beginning.

On the basis of natural and numerical experiments the main signs of similarity of principles of act and operating modes of hydroclutches and electric asynchronous motors (EAM) are revealed: presence of sliding, etc. In particular, and others at an exit on idling pass those to a fading oscillatory regime. Here the basic regime periodically changes on overtake at which the turbine transfers power to the pump, and a rotor–on stator. The specified signs are manifested as well that, both in hydraulic clutches, and in EAM it is possible to gate out active, jet and exchange making powers, and also overtake (generating), stop and brake regimes. On the basis of the spent researches the method of calculation and designing of the electrohydrodynamic transfers working on various regimes under trying conditions of maintenance is developed. The method allows to size up dynamic parametres and characteristics of teamwork of hydroclutches with EAM in the specified conditions and to define conditions

ISBN: 978-1-61804-251-4

of reliable start and economic work of electrohydrodynamic energy transfers. Dynamic models of the power drive from HM volume type, are created by working together with electric synchronous and asynchronous cars. The main signs of similarity of principles of act and operating modes of the volume hydraulic drive are revealed: the hydropump and the hydromotor working in common, and an electric synchronous motor: presence of synchronism, etc. Those and others at start are retracted in synchronism, and at an exit on idling pass to a fading oscillatory regime. Here the basic regime periodically changes on overtake at which the hydromotor transfers power to the pump, and a rotor – to the stator.

On the basis of the researches executed in [7] – [8], the computer models MIC considering features of dynamic operating modes of each cylinder of the piston car, including on perspectiv water fuel emulsions are created. Adequacy of models is shown by comparison of results of modelling with the data of skilled tests of injection engines of type M100, including with a water additive in fuel (fig 10,11). Computer models of injection engines of the specified type with fluid couplings investigated above geometric series are developed.

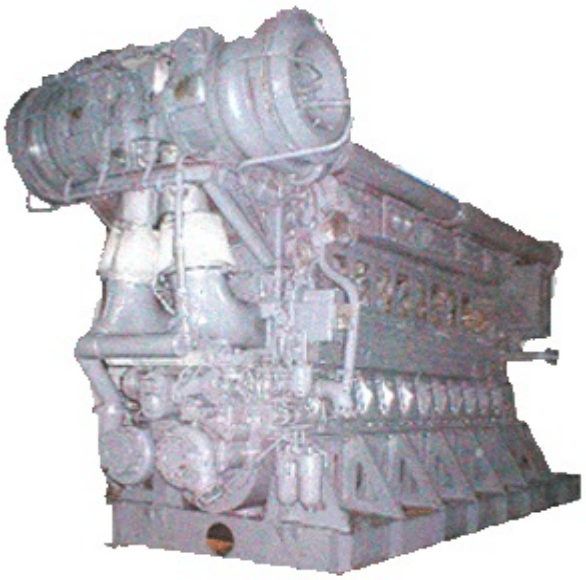


Fig. 10 The general view of the injection engine of type M100.

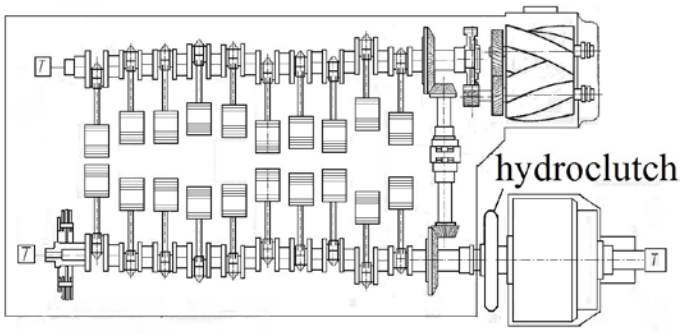


Fig. 11 The Kinematic scheme of the injection engine of type M100.

Results of full-scale tests and data of computer modelling of dynamics of start of the specified injection engine without the hydroclutch and with the hydroclutch of an investigated row (fig. 12, 13) are above resulted. Results of numerical experiments and the researches of dynamics of teamwork of cars executed on their basis under trying conditions are resulted. The built models thermal and the hydraulic machiner adequately reflect all main operational properties of their teamwork, including under trying conditions maintenance. Models reflect ability of hydroclutches: to protect the propeller from overloadings, to reduce amplitude of the torsion oscillations called on one of shafts of the hydroclutch, by "transfer" of these oscillations by other shaft of this hydrocar, ability of hydroclutches to "filtrate" not to pass, under certain conditions, the specified oscillations from one shaft on another, etc.

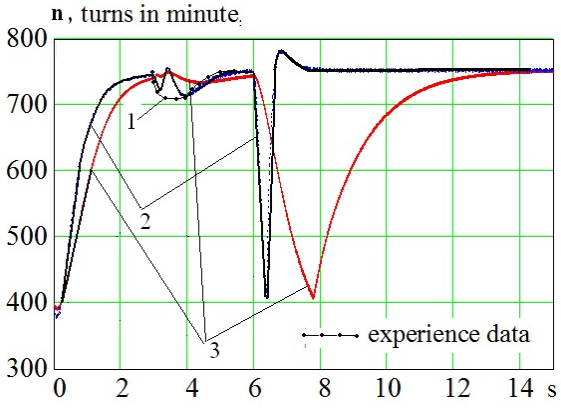


Fig. 12 Rotational speed n of the shaft of the injection engine without the hydroclutch and with the hydroclutch: 1 – natural experiment without the hydroclutch; 2 – numerical experiment without the hydroclutch; 3 – numerical experiment with the hydroclutch.

ISBN: 978-1-61804-251-4

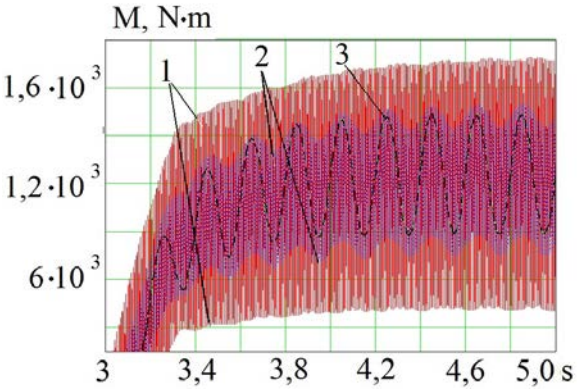


Fig. 13 Dynamics of torsion oscillations: the propeller 1-shaft the hydroclutch; 2-turbine shaft; 3-simple harmonic motions (with frequency of 5 Hz) a torque of external loading.

Numerical experiments show that at an oscilation frequency of the external loading close to own frequency of torsion oscillations of the shaft of the injection engine, there is resonance condition which define character of change of all parametres of system: heat power installation and hydromechanical transfer, including: torsion oscillations, the relative charge of operating fluid in hydroclutch and a rotational speed of shafts of the hydroclutch and the injection engine.

4. The conclusion and leading-outs. The methodology of computer modelling of the difficult hydropower-driven systems equipped with primary energy sources of two types is created: explosive motors and electric motors. Computer models of hydraulic, thermal and electric cars reflect their main operational properties at teamwork. These models can be used not only for detailed research of dynamic characteristics of the specified cars, but also in systems of a computer-aided design of these cars. Considering extreme complexity of the dynamic processes accompanying work of investigated cars, and also modern level of knowledge of character of the valid dynamic processes accompanying their work, the author considers that the computer models created by it can be used only for reception of the approached estimations of expected dynamic characteristics of again created cars and systems. The created models can be useful to reception of substantiated conclusions about possibility martempering of characteristics of hydraulic, thermal and electric cars and can be recommended for use in the design and exploratory organisations developing perspectiv warm - electrichydrodriven systems.

References

[1] Ozerskij A.I., Babenkov J.I., Shoshiashvili M.E. Perspectiv directions of development of the power hydraulic actuator. Magazine "News of higher educational institutions". North the Caucasian region. Engineering science. 2008. №6. p.55-61.

[2] Ozerskij A.I, Poluhin D.A, Sizonov V.S. Research of one-dimensional movements of liquid medium with contact ruptures in the highways containing pumps. News. AH the USSR. Power engineering and transport. 1979. № 2. p. 143-150.

[3] Ozerskij A.I. Application of the approach of Lagranzha to a problem solving of dynamics of hydraulic systems of hydropower-driven and heat power installations. The bulletin of the Don state technical university. 2010. A volume 10 №6 (49) p. 914- 924.

[4] Ozerskij A.I. Model of the hydroclutch with an asynchronous electric motor. Magazine "News of higher educational institutions". North the Caucasian region. Engineering science. 2011. №5 p.58-66.

[5] Ozerskij A.I. Bas of modelling of the hydroclutches working under trying conditions of maintenance. Magazine "News of higher educational institutions". North the Caucasian region. Engineering science. 2012. №1. p.105-113.

[6] Ozerskij A.I., Pustovetov M.Y, Shoshishvili E.M. Computer modelling of the electrohydrodynamic drive. Magazine "News of higher educational institutions". North the Caucasian region. Engineering science. 2012. №4. p. 48-55.

[7] Ozerskij A.I., Ivanov I.A., Babenkov Y.I. Model of working process of the injection engine on water fuel emulsions. Magazine "News of higher educational institutions". North the Caucasian region. Engineering science. 2011. №6. p. 79 − 85.

[8] Ozerskij A.I. Modelling of work of the hydroclutch with the injection engine under trying conditions maintenance. Magazine "News of higher educational institutions". North the Caucasian region. Engineering science. 2012. №2. p. 77-84.

[9] Ozerskij A.I., Shoshiashvili M. E. A method of calculation of dynamic operating modes electrichydrodriven with ampoulisition hydraulic system. Magazine "News of higher educational institutions". North the Caucasian region. Engineering science. 2014. №1. p.20– 26.

Dynamics of financial market stability factors in terms of financial globalization

Rustam R. Akhmetov

***Abstract*—** In the course of economic globalization a process of formation of the united financial market has evolved. Integratedness of different sectors of financial market into its global form is largely stipulated by development of financial innovations and of new financial instruments. Subjacent and timeless nature of modern financial transactions leads to blurring of distinction between money and capital markets. The united market means a gradual merge of financial risks and increase of volumes and unpredictability of aggregate risk in the result of synergetic effect contained in the nature of financial market as of a nonlinear dynamical system. Serious influence on stability of financial market has a behavioral psychology of its participants that is connected with reflexivity specific to financial capital. Financial market can be described only by nonlinear stochastic equations. In this article we made the efforts to set the parameters for such a model. Basing on non-stability of a global financial market and under influence of globalization factors it was concluded that it is necessary to check the hypotheses on formation of a financial cycle apart from general business cycle.

***Keywords*—** Financial globalization, financial cycle, financial market stability, nonlinear dynamical systems, securitization.

I. Introduction

Consequences of economic crisis, problems of financial state of several states including leading economics of the world, issues of Russia's entry to the World Trade Organization (WTO) and discussions connected therewith – all of it can be regarded as a reflection of world economics internationalization and globalization process. We single out two methodological principles of global financial market analysis: first - structural and functional, second – institutional (Akhmetov, 2011). The first principle is connected with study of functions of money and capital, their changes in a modern world, dynamics of offer and demand on money and capital. Structural analysis relates to structure and dynamics of financial markets. Institutional principle binds market analysis primarily with organization of the market and functioning of its institutions. Issues of institutional development of a financial market are highly important for study and specification of modern processes in the sphere of finances.

From this point of view globalization process is a factor that unites in itself both methodological approaches. The central component of world economic globalization is financial globalization. In the result of financial globalization the capital became much more mobile having moved all over the world to the most attractive and profitable capabilities of application. Nature of operations of the global market participants with diversification of assets and liabilities according to states and regions, availability of wide network of representative offices, branches and subsidiary organizations abroad so far does not allow identifying them only with the country of national identity. Financial globalization has strengthened influence of international markets on performance of credit and borrowing operations by residents of different countries that lead to growth of international network of financial institutions and corporations increase of business participation falling on foreign countries and to fundamental changes in their systems of organization of financial flows management.

Along with the benefits financial openness and integration are increasing risk. Consequence in advance of the development of financial globalization is that capital flows are poorly coordinated with the flow of technology. That's why globalization does not actually provide full of technological exchange (G.Mosey). This contradicts the traditional economic paradigm, according to which globalization, accompanied by the liberalization of capital movements, should lead to a reduction in systemic risk. In fact, the global financial crisis 2007-2009 refuted this assertion. Neither diversification of bank portfolios nor the policy of the monetary authorities do not reduce banking risks, but rather strengthens them (L.Laeven and F.Valencia).

Financial sphere pretends to be an absolute leader of economic globalization. The notion of globalization is often understood as a wide distribution and spread primarily of financial institutions and financial markets. According to Independent Strategy specialists (D.Roche, B.McKee, G.Manca and oth.) about 40% of global manufacturing of industrial products, 60% of global value product, 70-90% of world trade and international finances are influenced by globalization. This proves the fact that financial sphere is much more affected by globalization processes much more than other economic spheres (New Monetarism, 2006).

In our opinion all diversity of specific features and forms financial globalization can be generalized by four groups:

- General financial forms proving rapid growth of financial markets;

ISBN: 978-1-61804-251-4

- Organizational features reflecting transformation processes in content and structure of financial institutions;
- Management forms connected with the change of role of governments and international organizations;
- Informational and technological forms and ideological features specifying development and growth of information technologies and changes in the sphere of social consciousness (Akhmetov, 2009).

II. Methodology of financial markets globalization analysis

We would like to consider each of the form closer in order to determine the degree of their influence on stability of financial markets.

1. General financial features of globalization can be more vividly shown in correlation of real and financial sectors. Financial sector part in global manufacturing of products and services has grown and dominates both by its part and by its economic significance. If in 1975 the proportion of international operations with shares and stocks with GDP in developed countries was not more than 5% , by the beginning of XXI century it has grown up to 700%. What caused such a growth?

One of the main reasons was disinflation that is the process of deceleration of price growth rates. Thus, if the average income of 10-year's bonds in the markets of the OECD countries for 25 years (1981-2006) has amounted 3.7%, inflation during the same period has amounted only to1.7%.Deceleration of inflation rates during the last quarter of the century is stipulated by several fundamental factors. First reason was a strengthened purposeful anti-inflation policy of governments and of central banks of the states. Monetary powers of developed countries have transformed low inflation into prior purpose of their activity. Governments of the states - OECD members have limited interference into economics and state expenses ("Reaganomics" and "Thatcherism" in 1980-s). The result was decreasing of budgetary expenses and diminution of the need to financing budgetary deficits including at the expense of increase of money supply.

Secondly, in the result of globalization international trade barriers have been lowered, and a flow of correspondingly cheap goods has swept into markets of highly developed countries primarily from Asia (India, China etc.). The prices began to decrease due to of competitiveness of the manufacturers.

The third factor of disinflation was a serious increase of efficiency of manufacturing and trading of private companies all over the world. Development of new technologies, including internet service in the sphere of trading and of information exchange and also increase of management quality within the companies made a significant contribution to it.

Stable deceleration of inflation rates in a long-term perspective has gradually lead to the fact that real interest rates in Organization for Economic Cooperation and Development countries nowadays have become much lower than their long-term values. During disinflation period financial assets were increasing much quicker than material assets and GDP.

Table 1. Growth of global financial assets

	1980	**1990**	**2000**	**2005**	**2010**
Financial assets (money, shares, bonds) of countries of the Seven as the interest from GDP	150	210	370	400	530

ISBN: 978-1-61804-251-4

In general financial sense globalization is also characterized by the fact that rapid change of international financial development brought new directions and forms of internationalization of manufacturing, trade and finances that in its turn leads to convergence of capitals of many countries in different forms and modifications (figures 1).

One of such modifications has become synchronization of business cycle development during the last several years. Opinion about world cycle desynchronization that has appeared before crisis of 2007-09 has been based on the fact of general long-term recovery in global economics during 16 years – from 1992 to 2007. Data specified on graphics (figures 2) shows dynamics of industrial manufacturing in four leading OECD countries has been changing synchronically during the last 20 years. Unemployment situation has not been so clear: European labor market at times of its recovery has been significantly different from American and Japanese markets – that is most noticeable on the example of Germany. There is evidence that the integration of trade and industry processes have a stronger impact on the finance sphere than consumption (Kose, Prasad, Terrones, 2003).

2. Organization features of financial globalization, reflecting internal processes in the sphere of different sectors of financial market and transformations in the content and structure of financial institutions. Organizational peculiarities of globalization processes in 1990-2000 were expressed by:

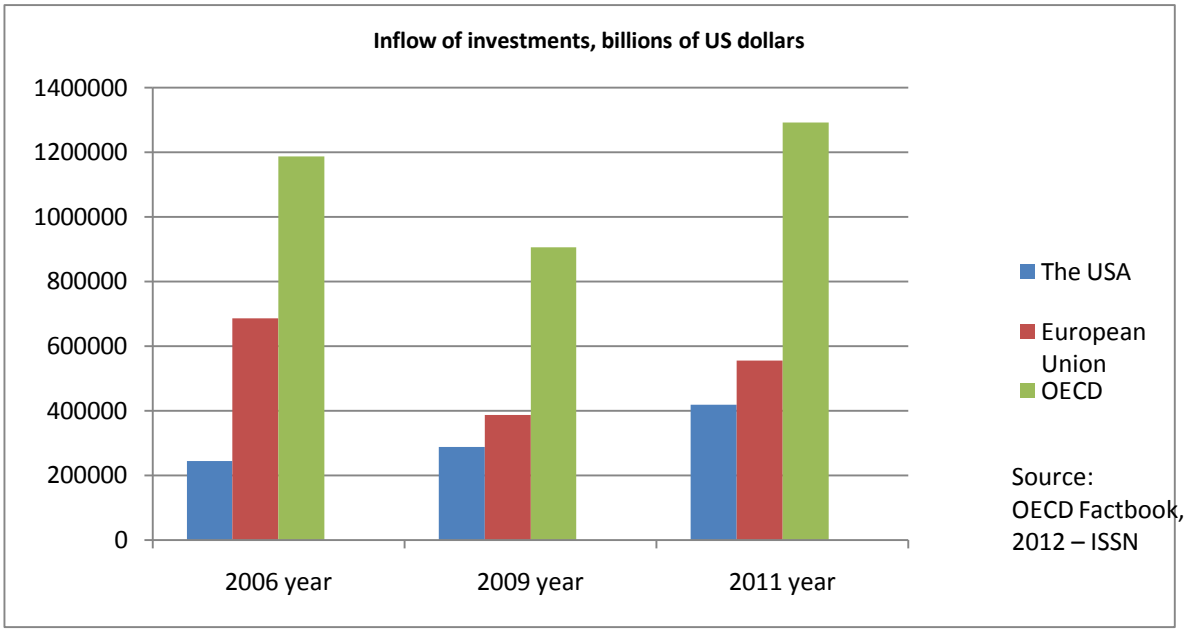
- Convergence of activities of banks with non-banking financial institutions (funds, insurance organizations and etc.);
- Securitization of assets and erasing of distinction between monetary and capital markets;
- Decrease of the role of transnational companies and transnational banks (TNC and TNB) in global economics.

On the one hand, there was *blurring* of the banks from traditional financial institutions into new more diversified structures. On the other hand, non-banking financial organizations have so much infiltrated into activity of banks at the pick of deregulation of 80-s that in deprived the banks from the leading role in big business financing. Broker firms offer their clients cash management services, transactions control service, insurance products sale. Insurance agents begin to register sale of securities, insurance companies sale share of mutual funds. Banks also trade with mutual funds, offering in addition to it discount broker services.

Appearance of new financial institutions reflecting features of different sectors of the market became the result of convergence of financial institutions. The brightest and the most stable example of it is the securitization process that is a transformation of its financial assets into securities.

The other important organizational feature of globalization is a process of gradual blurring of the distinction between monetary and capital markets. Securitization lays the basis for it. Let us consider securitizing mortgages emission as the example. Upon issuing them financial institution can save payment for initial mortgages. This allows them to take two advantages: firstly, they do a comparatively easy transfer of the interest rate and credit risks to the investor. Secondly, in fact issuing a new mortgage on the basis of a previous one the issuer procures a some kind of recycling of the capital. Bank or other financial institution uses a mortgage capital acquired in the result of emission in order to perform a further mortgage loan, and to reflect as income assets obtained from it (St.Veale, 1987).

Figure 1. Foreign direct investments in OECD countries.



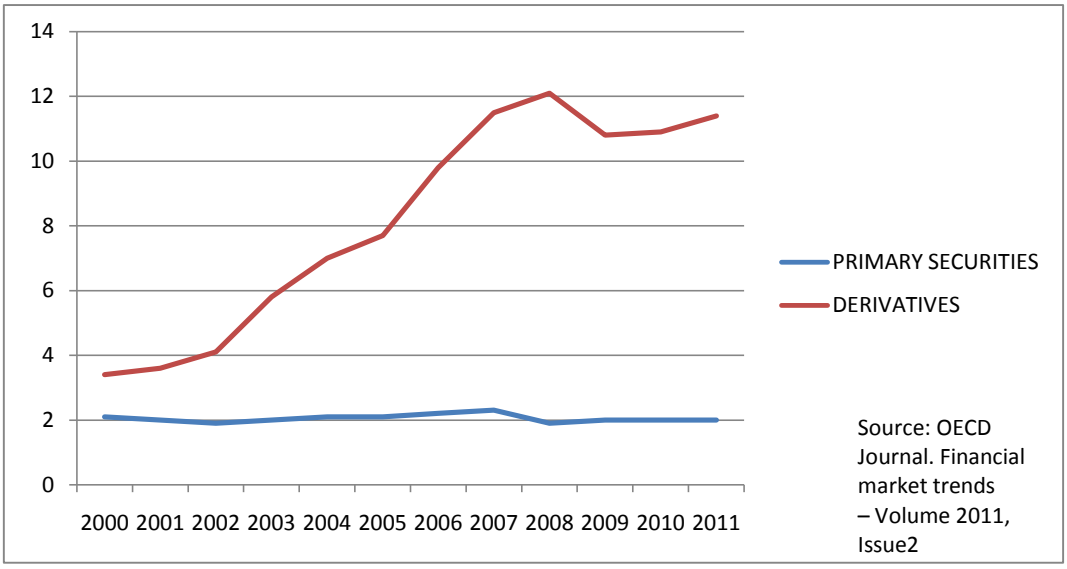
Emission of individual mortgages is connected with the increased risk. Firstly, because they require larger primary investments: most individual mortgages have a high minimal cost. That itself increases nominal size of risk capital. Secondly, such securities based on derivative instruments have a more complex credit risk

evaluation. Investor often lack of time, resources and expert capacities for evaluation of the property being mortgaged, employment verification, credit checks and other things necessary for adequate and complete evaluation of a credit risk of individual securitizing mortgage. Thirdly, there is no a stable market on such mortgages that indirectly causes liquidity problems.

For these reasons liquid market of individual mortgages has not been developed until institutes operating them began securitizing their mortgage portfolios. For the first glance it allowed to significantly simplify analysis and trading of these securities. As for mortgages securitization process widely began when financial institutions (banks etc.) began to sale their mortgages to more specialized sophisticated finance brokers. This concerns not only MBS but also a whole range of other similar instruments (participating certificates, collateralized mortgage obligations, adjustable rate mortgages, collateralized bond obligations, credit default swaps), that in its turn caused a rush of global growth of derivatives and risks connected therewith.

Figure 4 shows growth of primary securities (bonds and equities) and bank assets compares with off-market derivatives. The growth rate of derivatives volume was almost five times bigger than the primary securities during the period from 1998 to 2011. If in the middle of 1998 volume of global derivatives amounted to 2.5-3 worlds' GDP, on the eve of the crisis, in June 2008, it has been 12-13 times bigger than worlds' GDP. At the same time primary securities volume during the same period has remained stable that has been approximately two times bigger than GDP.

Figure 4. Dynamics of world volume of primary securities and nominal cost of derivatives



Earlier banks have explained mountains of derivatives by the fact that they were necessary to control risks and also for innovation and efficiency in economics (A.Blundell-Wignall, 2011). Some of derivatives performed social tasks connected with hedging of business risks (A.Blundell-Wignall). It should be noted the use of derivatives for tax arbitration (interest rates swaps for the differences tax treatment of products). Credit-Default Swaps (CDS) have been widely used for arbitrary regulation in order to minimize the required bank capital.

However, according to many researchers during the last decade this social use of derivatives has been minimal. Their general increasing trend is reported to make the decade the worse one from the point of view of financial risks from the times of the Great Depression. So called *social* function of derivatives allowed leverage to rise and this way to extremely increase the risk (Blundell-Wignall and Atkinson, 2011). The important fact in this relation is the continuous increase of a gap between revolving derivatives and primary securities. Due to the fact that primary securities lay the foundation for and largely ensure security of derivatives market, the said divergent trend between the volumes of them shows increasing overuse of the same security (rehypothecation), that multiplies joint risk through bank system.

3. Managerial features of globalization are connected with the change of roles of governments and international organizations. The current level of global market mechanisms is far from perfect and yet can not perform a function of a global regulator, because speculative methods of actions of its members is a feature both of the market in general and of actors taking part therein. Difficulties in national and international regulation of global processes occur.

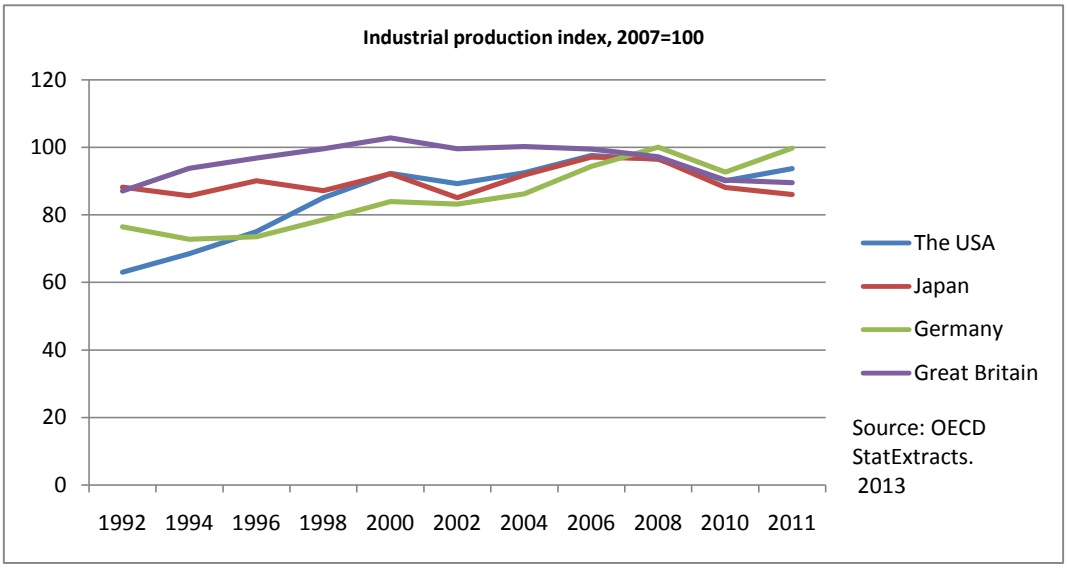
The important factor of globalization is that modern means of informatization allow to radically change the whole system of management from small firms up to global economic and financial system. Therefore, the role and significance of the state and of international organization in regulation of global financial system will be changed. Participation of both of them in affecting financial markets shifts to sector-specific, but stricter supervision and control (J.Sax, 1994).

ISBN: 978-1-61804-251-4

4. The other important feature of globalization is informational and ideological component. This feature is so important that some authors play the main emphasis on it. In our opinion informational component of globalization can be generalized by following factors. First of all, it is a growth of requirements to the market and its participants on revealing of information. Secondly, psychological effect of financial innovations, that is appearance and implementation of new instruments and technologies in financial markets, is an important aspect. The third form is informational inequality formed in the result of technological revolution.

Limited rationalizing and computing problems of information processing lead to substitution of a complete analysis of the whole information by precedent or analogue solution, by use of *a priori* or *herdlike* behavior. Rapid growth of information, its continuous updating and improvement is accompanied by a dynamic increase of bulk of excessive, repeating, inaccurate information – appearance of so called *noises*. Popularity of the idea of irrational control in modern management can be served as the example of practical activity in terms of informational inequity. The Companies act by trial and error not in a manner of maximal rational reckoning.

Figure 3. Dynamics of industrial manufacturing of developed countries



ISBN: 978-1-61804-251-4

III. Results

Development of main forms of financial globalization had a great impact on financial market stability. How the above mentioned tendencies can reflect on global financial market development rates? What effect do these tendencies have on financial market stability: do they contribute to its growth or strengthen destabilization?

We assume that financial markets correspond to nonlinear dynamical system. This statement is based on the following hypotheses. First of all, we agree that change of prices in financial markets happen in a random way. It means that subsequent changes of prices do not depend on each other. This statement has been proved by facts (L.Bachelier, M.Kendall, E.Fama, 1981). Bachelier argued that the expectation of the speculator is zero, and the process of price changes S=$(S_t)t \geq 0$ is a random process. Exploring the time series of prices with a time interval Δt, he noticed that the difference $(S_t - S_{t-\Delta})$ has zero mean and fluctuation order $\sqrt{\Delta}$. These properties have a random walk: $S_t = S_0 + \sum_{\Delta}^{k\Delta} x_\Delta$, where x_Δ - identically distributed independent variables, which can take two values, $\pm\sigma\sqrt{\Delta}$ with equal probabilities (1/2). Limit as aspiration $\Delta \to 0$ leads to a random process based on Brownian (Wiener) motion.

M.Kendall claimed not the prices themselves and their logarithms are subject to a random walk, that is, if $z_n = \ln\frac{S_n}{S_{n-1}}$, then $S_n = S_0 e^{z_n}$, n≥1, where Z_n –

Secondly, we assume that reflexivity is specific to global financial capital. As G.Soros has shown there is a bilateral connection between current decisions and future events in stock market. Forthcoming quotes depend on current expectations of investors just as prices themselves influence on expectations. For the reason of reflexivity in the market the balance is nearly to be unreachable (G.Soros, 1988).

Thirdly, the leading role in financial markets plays not rational behavior of the investors (that in our opinion does not exist at all), but a psychological factor of market members. The market is a correlation of psychologies of its participants lead by their individual motives that mostly do not have much common with reasonability. This gives a reason to think that change of market prices and market behavior in general can not be described by classical financial models with a sufficient degree of authenticity. Here works the mechanism of the reflexivity: prices influence on expectations, expectations form prices. In crucial moments during crises actions of market laws of competitiveness and pricing are being overlapped with panic, rumors, perverted expectations. It is difficult to predict investor's behavior as well as financial rates during such periods.

Such behavior is studied in theory of behavioral finance and in modern models of nonlinear systems and stochastic theories of crises and cycles. Nonlinearity brings numerous developmental pathways that are refracted in bifurcation points in an unpredictable manner. Nonlinear systems mathematically can be described with nonlinear differential equations binding range of the unknown function at a point and range of its derivatives of different orders at the same point:

$$F(t, x, x^1, x^{11}, ...x^{(n)}) = 0, \quad (1)$$

where x = x (t) is an unknown function depending on time variable t.

Provided that because market dynamics is of occasional and largely discrete nature, it can be subject primarily to stochastic differential equations that include occasional processes. Such equations have complex appearance and solution because solving is also a stochastic process. For this system of first-order stochastic differential equation in a form of Langevin equation are used:

$$x_i = \frac{dx_i}{dt} = f_i(x) + \sum_{m=1}^{n} g_i^m(x)\eta_m(t), \quad (2)$$

where $x = \{x_i | 1 \leq i \leq k\}$ is a set of unknowns, f_i and g_i are arbitrary functions, and η_m is random function from time.

Economic sense of stochastic differential equations for financial market can described as follows: financial assets price is a result of balance of profitability of the company-issuer and market risk functions on the one hand and functions of investors' expectations on the other hand.

In order to assign a dynamical system it is necessary to describe its phase field that is a variety of possible states at a fixed point of time. Moreover, it is necessary to assign a variety of timepoints t and (most importantly) the rule describing movement of points of the phase field during the time.

Let us assume that phase field X is a variety of all possible states in periods of time T, including variety t. If we know information appearance (processing) speed $\alpha_i(x)$, path described by point $x_0 \in X$ will be a solution to differential equation $\frac{dx}{dt} = \alpha(x)$. Equations system $\begin{cases} \frac{dx}{dt} = \alpha \\ \frac{d\alpha}{dt} = -\beta x \end{cases}$ establishes a continuous-time dynamical system (harmonic oscillator). Such system can shape various fluctuating motions, in this case time fluctuations of market prices towards shares depending on expectations of investors. However, having any assignment of dynamical system it is not always possible to find it and describe its pathways in an explicit form. That's why usually more simple issues about general behavior of the system are being considered.

Here appears the biggest difficulty formulating the financial market regularities: phase field of this system can not be accurately assigned by a set of numbers or end variety of the field in a multidimensional space, because in this case we deal with nonlinear system. The internal feature of nonlinear systems is a synergetic effect, dynamical chaos, lack of predictable pathways of development and stability.

ISBN: 978-1-61804-251-4

IV. Discussion

Financial stability is an integral part of the overall economic stability. We define economic sustainability as the ability of an object (the economic system) to resist cyclic phenomena in the economy and the impact of external factors beyond the system.

John Downes and Jordan Goodman distinguish several types of stabilization: the currency, economic, market trading (J.Downes, J.Goodman, 1995). The meaning of all treatments reduced to price and current market stability. According to the so-called "crisis" financial stability definition of stability is regarded as state of the financial system or the opposite unstable market, i.e. which shall not involve destabilizing the situation bearing a threat the financial crisis (O.Lakshina, H.Chekmareva, 2005).

In this context, mention should be made of the theory of financial stability H.P.Minsky . According to it the stock market passively reflects estimates of future returns on investments made by real investors. The very same financial system although does not effect on decision-making in the real sector, but because of their uncertainty makes the economy inherently unstable (H.Minsky, 1983). According to Minsky asset valuation is not objective process, as is done in the face of uncertainty. This essentially means recognizing emotions by integral part of market behavior. According to Minsky , the boom periods caused to a tendency to reduce expectations of risks and waiting for the value of assets. This causes growth of lending and, consequently, increased vulnerability to risk. Liquidity problems can cause a crisis of insolvency through a “domino effect” (S.Dow, 2010).

V. Conclusions

Let us return to main forms of financial market globalization. Within the framework of the reviewed model they can lead to double-side effect. Being the elements uniting and converging markets, they will contribute to their stability. At the same time being elements that unreasonably expand and complexify markets, they are factors increasing their risks and destabilization potential. In financial world are becoming less multiple repeated situations and events basing on which statistical regularities can be determined. Quantitative optimization allows answering the question which part of the risk can be really measured, but does not help to answer the question what is the real total risk value?

Financial market supervisions pay too little attention to systemic risks arising from leverage and potential implications of rapidly increasing financial globalization for the transmission of shocks across the borders (P.Padoan, 2012). In modern structure of financial market we can talk about immanent instability having not only momentary but in a cycle nature. Cyclicity of financial markets does not have a direct material basis, but is proved by a number of preconditions. Among those we rate:

- capacity of blowing financial and credit bubbles in the result of growth of nonproductive sector of economy;

- mass distribution of the latest financial instruments and technologies of increased risks;

- procycle behavior of some leading and coincident indicators in financial market;

- deregulation of financial markets within the framework of global economy and national economics of the leading countries.

Analysis of the said interrelations and assessment of financial cycle formation possibility is a further stage of financial market stability factors research.

References

[1] Akhmetov R., 2011. Methodological approaches to analysis of financial market conditions /Intelligence. Innovations. Investments, 2011, No.1: pp.104-109.

[2] Mosey G., 2002. Globalization and regionalization processes in the world economy/Economist, No. 9: pp.24-28.

[3] Systemic Banking Crises: A New Database. Laeven, L. and Valencia, F., 2008. IMF Working Paper, Nov. – WP/08/224.

[4] New Monetarism. Roche D., Manca G., Mckee B. and others, 2006. Independent Strategy. Date views: 13.10.2011 www.instrategy.com/books.php (pdf).

[5] Akhmetov, R., 2009. Development of financial markets within modern cycle /Finance and credit, 27(363):51-55.

[6] Kose, M., E.Prasad and M.Terrones, 2003. How does globalization affect the synchronization of business cycle? /IMF Working Paper, January, International Monetary Fund, 2003.

[7] Stocks, bonds, options, futures. Investments and their markets, 1987. Ed. by St.Veale. New-York Institute of Finance, Simon & Schuster Co., pp:332.

[8] Blundell-Wignall, A., 2011. Solving the Financial and Sovereign Debt Crises in Europe /OECD Journal: Financial market trends, Issue2: 1-23.

[9] Blundell-Wignall A. and P.Atkinson, 2011, Global SIFIs, Derivatives and Financial Stability, OECD Journal, Financial Market Trends, vol. 2011/1.

[10] Sax, J., 1994, Beautiful Renaissance. The Economist, October 1:27-28.

[11] Fama, E., 1981, Stock Returns, Real Activity, Inflation and Money. American Economic Review, September, Vol.71, No.4:545-565.

[12] Soros, G., 1988, The Alchemy of Finance. Date views 25.12.2013 www.polbu.ru

[13] Downes J., Goodman J.F., 1995. Dictionary of Finance and Investment Terms. New-York: Barron’s, pp.628.

[14] Lakshina O., Chekmareva H., 2005. Financial stability analysis: practice and methodology/Money and Credit, 10, p.25.

[15] Minsky H.P. The Financial Instability Hypothesis: an Interpretation of Keynes and an Alternative to “Standard Theory”/J.M.Keynes. Critical Assessements. Ed. By J.C.Wood. London. 1983, pp.:282-292.

[16] Dow S.C., 2010. The Psychology of Financial Markets: Keynes, Minsky and Emotional Finance/ Voprosy Economiki, 1, pp.: 199-213.

[17] Padoan, P.C., 2012. Economy: The evolving paradigm. OECD Yearbook, 18.

Social investments of Russian business: Problems and Prospects

ANNA B.TESLYA
Department «Global and regional economy»
Saint Petersburg State Polytechnical University
St. Petersburg, Severny 63-5-49
RUSSIA
anntes@list.ru

Abstract: The article discusses the impact of social investment on the firm value. The problem of the formation of social investment in Russia, obstacles to the development of social investment are considered. Specificity and social investment trends in the country are revealed. Criteria and indicators to harmonize the results of social investment and firm value are proposed.

Key-Words: social investment, firm value, stakeholders, social effect, corporate community involvement,

1 Introduction

Social investment is an objective process that is underway in Russia with the government playing the major role, companies developing in an oligarchical way, and the social institutions of the civil society being immature. The forms and ways of participation as well as the terms of collaboration between the companies in the process of social investment in Russia are being established. The questions of the investment's social efficiency and economic effectiveness become an important aspect of the corporate social policy as a part of the general business strategy.

That is why it is necessary to draw the attention of the officials, the business and the mainstream audience to the problem of social investment development in an effort to increase the quality of interaction between the business sector and the society. At the same time, each side mentioned must fully realize what kind of profit such interaction can bring and the way this profit is formed.

2 Problem Formulation

In order to activate the process of social investment development in Russia, it is crucial, above all for managers, to fully understand the mechanism of social investment influence on the results of the companies' financial and business activity. The main obstacles to this are the ambiguity of interpretation of the term "social investment", difficulties in estimating the economic effect of social investment, and its deferred character. The importance of this matter can be proved by the fact that, according to [15], the main reason why the companies participated in the implementation of social projects was moral compulsion and the wish to help, as stated by 45% of the respondents.

Social investment is often defined as an investment into the social sphere "in order to increase the quality of life by creating new technologies and mechanisms of funds allocation among different social groups in accordance with their needs". A positive social effect is, undoubtedly, one of the important characteristics of social investment. However, it necessary to distinguish between social investment and other kinds of the companies' social activity, such as corporate volunteering or charity.

We define social investment as material, technological, managerial, financial and other resources aimed at the implementation of social programs, developed in accordance with the interests of the major internal and external stakeholders. As a result of this implementation, the company is supposed to get both social and economic effect in a strategic outlook. The economic effect that the company gets is likely to be deferred and hard to estimate.

ISBN: 978-1-61804-251-4

Corporate decisions based on ethical values and aimed at meeting the requirements and expectations of the stakeholders are regarded as necessary (European approach) or legally unnecessary (American approach).

Shareholders, company's investors and staff are regarded as major stakeholders, then go the local community and non-profit institutions. The system of the company priorities will vary depending on the industry sector, the extent of authority of the different categories of external stakeholders (regional and local authorities, civil institutions, non-profit institutions), and the regional location of subsidiary companies.

A different structure of the company's social investments will be formed depending on the extent of authority of the different groups of stakeholders. Two kinds of social investments are defined: internal (investment in staff training, health care and safety of labor) and external (fair business policy towards consumers and business partners, environmental activity and efficient use of resources, investments in local communities development). In the long run, social investment is aimed at developing the human capital of the company itself or the local community.

Priority directions of social investment may vary depending on the subject of the investment, still they need to provide firm value increase in the long-term period. It is problematic to determine a static or econometric relationship between social investment and the financial results of the company's activity. Indeterminacy of effect, difficulty in data acquisition, secrecy of data, indeterminacy of the relation between figures that reflect the company's social activity and its financial results explain the quantitative nature of the most articles concerning this issue that emphasize certain benefits of social investment for the corporation.

In the world's practice there is a number of research that proves the existence of direct and feedback coupling, yet denies the existence of the relationship itself. For example, [7] states that the companies engaged in social investment have Return on Equity (ROE) 9.8% higher, Return on assets (ROA) 3,55% higher, and Return on Sales (ROS) 2.79% higher than those that neglect it. The profit of these companies' shareholders is estimated to be 63,5% higher.

Examining the relationship between social investment and the company's activity efficiency, a positive correlation between Tobin's *q* and the company's environmental activity was discovered [3]. The sampling features American manufacturing companies from the S&P 500 index that operate in the USA and the countries with an average income. The companies under examination were divided into three categories:

- first category - companies that operate across the world according to the American ecological standards; - second category - companies that operate according to the standards higher than American;
- third category - companies that pursue lower standards where possible.

A more thorough research that was conducted later with a broader sampling proved the existence of a relationship between the lower level of pollution and the higher financial results (Tobin's q) [4]. It was impossible, however, to determine the direction of the influence.

A Russian research that should be noted is the one conducted by [12], an econometric analysis of investments in the implementation of environmental and corporate social responsibility policy and the investment attractiveness of the companies by the means of correlation and regression analysis. The research proved the existence of a direct correlation between investments in regional development, environmental policy implementation and the investment attractiveness of a company.

The correlation between a company's social activity and its value is also proved by practices. According to the research conducted by McKinsey&Co, about 2/3 of financial directors and about 75% of investment managers agreed that social investment contributes to the increase of firm value, but as for the extent of the impact, their assessment varied [6]. A poll conducted among the representatives of Novgorod business showed that 40,9% agreed that social investment has a positive impact on the company's financial results, the same number of respondents denied such impact exists, and 18,2% didn't answer [10].

Another point of view is that social investment limits the company's ability to use resources efficiently, compelling it to include low-profit projects into their portfolio, which is bad for the company performance.

Nevertheless, the current worldwide trend is the increasing number of investors who are eager to invest in companies that operate in accordance with their ethical principles, which is proved by the development of socially responsible investment. The number of socially responsible investment assets in the USA was six times larger in 2012 that it was in 1995 - $3774 billion against $639 billion. The same is true for the EU. In Russia, the first socially

ISBN: 978-1-61804-251-4

responsible investment funds have appeared (Uralsib bank, Invest - Management company, Econica - Finance).

[2] shows that American consumers take into account the company's socially responsible policy when buying its products, still they would buy another company's products if the prices are lower.

The public reaction to the company's social activity isn't always univocal and predictable. Statistic processing of the results of the polls conducted shows that consumers' reaction depends mostly on the way they perceive the company's activity, not on the activity itself [1]. Among the factors that lead to a positive feedback, the research names the consistency between the direction of social investment and the company's mission as well as the time for the consumer feedback to become apparent. The research shows that not every social project implemented has a positive effect on the results of the company's activity [5]. Among the Russian publications of interest [10,11,12] consider different aspects of the development of social investment in Russia. In [15, 16] generalized the real experience of the most successful and socially responsible business and proposed a methodology for evaluating the effectiveness of social investment companies of various forms of ownership, and the profile of activities.

3 Problem Solution

In Russia, social investment is still forming and developing. Since the '90s up until the beginning of '00s companies appropriated funds mostly for the single charitable projects due to the peculiarities of Russian legislation, which allowed not to pay taxes if the sum was less than 3%.

Since 2000 social investment has become more systematic. The first non-financial environmental report was presented by OJSC Ryasanskaya GRES in 2000. OJSC Gazprom began producing reports in 2001. In 2012, 69 companies published their reports. As of May 26th, 2014, 134 companies and 472 reports were registered in the National register of non-financial reports from 2000 to 2014. That includes: 41 environmental report, 219 social reports, 150 sustainable development reports, 42 integrated reports, 20 industry reports.

The number of Russian companies that produce regular reports has been growing bit by bit since 2000, and the quality of information reported increases as well. The number of reports that presents information about the company's triune total, including economic, environmental and social component, is also growing.

The aforementioned research proves that positive feedback from external stakeholders to the company's social activity is a crucial condition for the efficiency of social investment. Since mass media show little interest to the company's social activity, presenting public reports becomes an important way of delivering information. In Russia, public confidence in company activity is deficient, so it is advisable to conduct an independent assessment of the information that a company provides. The efficiency of this approach is proven by the growing number of Russian companies that use professional or public verification. Today, less than a half of the companies that produce social or other public reports report using verification procedures. New forms of independent verification appear, such as public hearing. Nevertheless, according to RBC, 61,7% think there are no socially responsible companies in Russia. According to [15], more than 70% of Russian companies do charity work, yet 55% of respondents admitted not knowing about this.

A research conducted by the Association of Managers in 2008 showed that most Russian companies are compelled to do charity by the regional authority.

There are specific reasons for such compulsion: first, there are companies that got a vast social infrastructure during privatization. Second, there are many monotowns in Russia that were built around a township-forming enterprise. These enterprises are commonly associated with manufacturing and mining industries. Such companies have to establish mutually beneficial relations with the regional and local authorities on the matter of social policy and make considerable social investments including environmental protection.

The correlation between social investment in regional development, environmental policy implementation and investment attractiveness [2] is a crucial factor for Russian companies. The ability to implement social projects is supervised by federal and regional authorities that can create favorable conditions or hinder social investment by adjusting tax and other policies for the business sector (excessive tax burden, administrative racket).

Public interest towards the companies' social activities is low. Thus, 57,6% of respondents claimed that their attitude to the company that participated in solving local problems would get better, over 10% said their attitude wouldn't change, 12% didn't answer [10]. At the same time, only 11% of the respondents think that authorities should

ISBN: 978-1-61804-251-4

be solving social problems, while 80% think that companies should take active part in that too [10].

The demand growth for business involvement into solution of local communities' problems is forming gradually. Before 2008 the considerable portion consisted of internal social investments (investment in staff training, health care and safety of labor), and now, it should be noted, that large companies started to cooperate with local communities and it conforms to global tendencies for social investment development.

Russian companies are characterized with independent choice and realization of social investments projects. They interact poorly with non-profit organizations and funds while in Europe and the USA non-profit organizations are the active partners of social sphere which have a considerable influence on the tendencies and forms of corporate social programs. Non-transparent activity and lack of competent specialists in this sphere are among the reasons that prevents Russian companies from working with non-profit organizations.

In terms of absence of the feedback and poorly marked public demand for companies' social activity as well as inactivity of non-profit organizations, activation of interaction with governmental institutions, in particular, may become a condition that will make social investment attractive for companies.

At the same time, the companies are interested in the opportunity to determine the items of expenditure on social purposes independently and provide transparency of their social programs. The necessity of formulating the priority directions of regions social development is becoming the aim of regional and local authorities. As well as the provision of their attractiveness for companies through the set of measures to support the social investment, for instance, through the reduction of tax burden.

Since 2008, due to the economic crisis the majority of Russian companies have ceased to consider social investment to be only a tool for interacting with the government and it has become an element of company's strategy that is confirmed with corporate reports. The Russian business is in the stage of transition from image support of vulnerable groups to implementation of social projects at the turn of internal and external company's social policy, social investment and investments into the development of human capital asset.

Despite the growth of companies' social activity, there is no mechanism of formulating and articulation of community needs in Russia. Also, there is a low level of social recognition of company's business. In such conditions the advantages from informational transparency may not offset losses on social investment and risk increase (conflict between managers and shareholders, possible increase of taxation, investors discontent, raise of production prices due to the expenditure on social projects, increase of vulnerability to competitors, conflict between separate groups of stakeholders). The long-term aim of Russian business is to make a mechanism of showing up the society requests and feedback.

Another factor that prevents the development of social investment is lack of understanding of the consequences of a social investment impact on the mechanism of company value increase which is typical for both Russian and foreign managers. According to the research performed by McKinsey, approximately a quarter of interrogated specialists in social investment find in difficult to evaluate the impact of social investment on company's value [6]. Among the main tendencies of impact were stated consolidation of relationships with customers (70.4%) and the government (70.5%) as well as consolidation of political position of a company or its leader (70.4%). The last figure displays the typical tendency for Russia that is characterized with directors of large companies being simultaneously a deputy in city and regional legislative bodies. In such a case, company's social activity becomes the condition for saving and increasing of social base and political assets of a company's director. However, such an approach to social investment does not always fully correspond to company's long-term interests. Among the tendencies of impact "the improvement of reputation and company's image in community" is noted by 93.2% of business representatives, however only 75.6% of respondents stated that their relation towards the company that takes part in problem solution of local communities will improve.

Insufficient understanding of the necessity of social investment realization and the mechanism of its impact on financial performance of a company may lead to the conflict of interest between main stakeholders because of the diversity of their interests, different evaluation of allowed risk and profitability goal from investment. Conflict of interest can show up on account of incomplete list of stakeholders, deficient conformity of their interests during the process of project portfolio forming and also because of stakeholders striving for maximizing individual benefit in the short run. The achievement of interest balance is becoming a key task as such a proportion in allocation of short-

term and long-term benefits between the participants of different level will provide the opportunity to achieve the desired company value increase in the long-term.

The company should take into consideration not only traditional indices of project income and risk but also different indicators showing social effect received while making a decision about realization of investment projects. In order to evaluate the effectiveness of social investments one can use the same indices that were used for evaluation of mercantile projects as social investments have a lot of similar characteristics with mercantile investments. The necessity of accounting of mercantile project results as well as its social significance brings to the need of evaluation of social effect in monetary terms and the priority of social criteria for decision making above mercantile. According to the investor's aims the following types of choices (Fig. 1) are theoretically and practically possible.

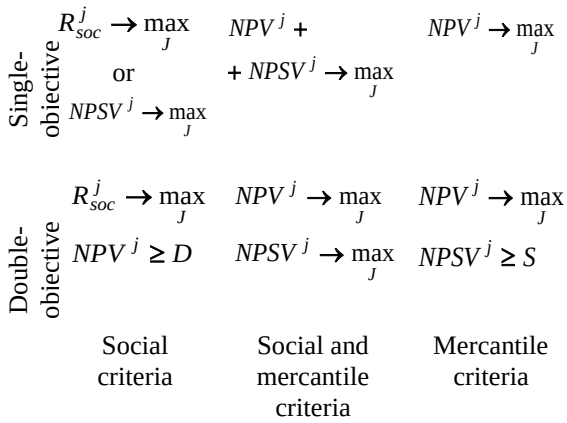


Fig. 1. Project selection criteria

Where: j - project, J = (j1, j2, …, jn) - portfolio, S, D - floor limits of profitability and social impact acceptable for investor, NPV^j - net present value of a mercantile project, $NPSV^j$ - net present social value of a project, Rsoc - social profitability (project's social effectiveness).

Conventional analysis of investment projects effectiveness assumes the possibility of project realization if:

- project's profitability is higher than the cost of called-up investments (WACC);
- net present value has positive value.

If social investment projects are evaluated in terms of conventional indices of investment attractiveness, they can be recognized as ineffective because they can have:

- negative net present value;
- project's profitability is lower than the cost of called-up investments (WACC);
- low internal rate of return (IRR);
- the profitability index of investment close to 1 with positive net present value.

The companies that do not perform social investment can reach a high value increase level in short-term when channeling resources on highly profitable projects realization. However, companies that do not share the concept of social responsibility in long-term will face significant social risks that must lead to cost increase of loan and owned capital and reduction of corporation value growth rate. At the same time, excessive volume of social investment can also lead to overestimated social liabilities that increase company's risks and decrease its value.

The projects connected with social investment are "conditionally ineffective". Their aim is to form or support company's competitive advantage. On the whole, the dependence of company's value from social investment can be displayed with the following indices:

- growth and strengthening of company's cash flows caused by the increase of the income from products quality growth, the quantity of loyal customers, loyalty of clients and income that exceeds expenses for improvement in the quality of products;
- reduction of production unit cost through working efficiency growth that compensate the expenses for improvement of working conditions;
- acceleration of working capital turnover through receiving comfortable conditions for interaction with debtors and creditors;
- decrease in cost of called-up capital through the reduction of systematic and nonsystematic risk;
- appearance of extra investment directions and extra opportunities for a company.

The exposure of company's rise in value factors in long-term in compliance with the benefits received through company's social activity and forming of the set of indices that allows to monitor the performance of chosen strategy for social investment becomes a company's aim. According to the aforementioned directions of the impact of social investment on company's value, it is possible to use the following system of indices:

-analysis of company's value growth rate that is determined on the basis of discounted cash flow or economic value added;

- comparison of forecasted *WACC* and actual *WACC* with constant capital structure or forecasted and actual cost of debt.
- comparison of present and forecasted indices ROS, ROE, ROA, ROI, ROCE or ROIC;

ISBN: 978-1-61804-251-4

- analysis of spread of average weighted ROIC (ROI) and WACC;
- analysis of ROI (ROIC) and WACC growth rates that reflect the dynamics of efficiency of investments and the change in company's financial risk;
- comparison of average weighted IRR by mercantile and social projects with the cost of source funding (WACC) for decision making considering the allowability of social investment projects realization;
- dispersion indices of incoming cash flows of a company;
- indices for labor efficiency and product costs.

4 Conclusion

The impact of social investment on a company is ambiguous. It means that there is a necessity of careful and cautious approach to decision making considering the project realization of social investment. The following factors which can provide positive economic impact from social investment should be noted: forming of long-term strategy for social investment and its acceptance with company's main development strategy; forming of positive feedback on social investment programme realization from stakeholders; appearance of the long-run results.

Social investment's ambiguous influence makes the following tasks of particular relevance: evaluation of economic viability of consequences from social investment for a particular period of time (planning period); determination of possible (marginal) amount of finance directed to financing of social investment projects in every certain moment of time within the framework of planning period; forming of the set of indices allowing to evaluate the economic consequences of social investment for companies.

The possibility to receive a positive economic effect from social investment is confirmed indirectly with the growth of companies' social activity including those from Russia. Analysis of foreign experience in social investments shows that there is a unique approach to stimulation of companies' social activity that depends on historical, institutional and cultural features of a country. The system of relationships between corporate sector, the government and local communities is forming in Russia as well taking into account complicated federal establishment of the country, specific character and development disparity of certain territories and regions, consciousness of corporate necessity of social investment as an extra source for self-development.

References:

[1] Becker-Olsen, Karen L., Andrew B. Cudmore, and Ronald P. Hill, The Impact of Perceived Corporate Social Responsibility on Consumer Behavior, *Journal of Business Research*, 59 (January), 2006, pp. 46-53.

[2] Cryer, E. and Ross, W. The Influence of Firm Behavior on purchase intention: do consumers really care about business ethics? *Journal of Consumer Marketing*, 14(6), 1997, pp. 421-433.

[3] Dowell G., Hart S., and Yeung B., Do Corporate Environmental Standards Create or Destroy Market Value?, *Management Science*, Vol. 46, 2000, pp. 1059-1074.

[4] King, Lennox, Does It Really Pay to Be Green? *Journal of Industrial Ecology*, Vol. 5. 2001, pp. 105-116

[5] Tom J. Brown and Peter A. Dacin, The company and the product: corporate associations and consumer product responses , *Journal of Marketing*, Vol. 61, No. 1 (Jan., 1997), pp. 68-84.

[6] Tracey Keys, Tomas Malnight, and Kees van der Graaf. *Making the most of Corporate social responsibility*. McKinsey Quarterly, 2009. http://mckinseyquarterly.com

[7] Weiser John, Simon Zadek, *Conversations with Disbelievers*, The Ford Foundation, 2000.

[8] Blagov Y. E., Litovchenko S. E., Ivanova E. A., *The report on social investments in Russia*, Association of Managers, 2008.

[9] Igoshina A. S. The policy of corporate social liability and company's investment attractiveness, *Economics, management, finance: materials of international scientific conference.* Perm: Mercurii, 2011, P. 40-44

[10] Vinnikov V.S. Management of social investment in corporations: theoretical and methodological aspects. A manuscript (diss. for the degree of Ph.D.), 2007.

[11] Danilova O.V. Business on the way to social investmtnt. Labour and Social Affairs, 12, 2011, pp. 26-32.

[12] Peskova O.S.. Boriskina T.B. . Development of processes of social investment in the development of corporate social responsibility. Economic sciences, Vol. 101, 2013, pp. 114-118.

[13] "Socially responsible business and social development" (research resutls), Veliky Novgorod, 2002.

[14] Tulchinskii G. L. *Corporate social investment and social partnership: technology and*

effectiveness evaluation,Department of operative polygraphy, National Research University, Higher School of Economics, Saint-Petersburg, 2012.

[15] Tulchinskii G. L., Oleinik O. V., Tulchinskaya L. E. et al. The Programme "Effective social investments and social partnership (ESISP)

Radial-Basis Functions Neural Network for Text Independent Speaker Recognition

A.A.Yakovenko, G.F.Malyhina,
Institute of Information Technology and Control Systems
St. Petersburg State Polytechnical University
St. Petersburg, Russia
e-mail: g_f_malychina@mail.ru
annother_@hotmail.com

Abstract — RBF neural network is proposed for solution the problem of text-independent speaker recognition. Recognition is based on estimation of sufficiently large set of acoustic features, construction of multidimensional histograms and approximation histograms with probability density functions with possibility of wide shape variation. Method allowed to reduce the probability of errors when decision was making.
Keywords—text independent identification, speaker recognition, radial-basis functions neural network.

1. INTRODUCTION

Biometric recognition systems allows us to find the right connection to authorize any person in information systems. In recent years the interest in voice biometrics has been increased [4, 5]. This is completely in demand in the areas of organization access permissions in information systems, biometric solving search and forensic accounting, voice verification of the driver and passengers, in the management elements of smart home, in banking systems, contact centers, etc. Identification of speaker's voice provides a unique opportunity to secure access to information, remote maintenance and examination to establish the identity.

Speaker identification and speaker verification problem is divided into two tasks a text-dependent identification and text-independent identification and can run using open set of speakers or closed set [4]. In the case of a closed set of speakers, phonogram will obviously belong to a particular individual, but if the phonogram does not belong to any candidate, then the problem is solved on an open set of speakers.

If identification system trained in advance to recognize universe passphrase delivered by announcer, then it is a text-dependent identification system. Phonemic dictionary and phrase structure in this case requires smaller amount of training speech data. The necessity of pronouncing passphrase during training and during the operation of the system limits the practical range of its application.

Identification system based on text-independent approach does not contain information about the uttered phrase. It is trained and then tested on arbitrary voice and speech data. The effectiveness of such identification systems is lower than in the text-dependent. But voice recognition in this case has broader application, since knowledge of uttered phrase is optional.

Voice identification reduces to the problem of deciding which of the plurality of speakers most likely belongs to the tested track. Since human speech is regarded as an acoustic signal, the analysis of the signal takes place by means of digital processing.

Develop a system of identification occurs in three stages [3]: on the first stage implementation of features extraction, on the second - modeling of speakers and on the third stage - decision-making is carried out. Thus, in general, a standard system for speaker voice recognition extracting unit comprise primary feature vectors of the speech signal and the simulation unit speaker's voice, which are divided according to the tasks. Since actual recordings made under conditions, there are many extraneous signals, various kinds of noise, impulse noise and congested areas of speech, preprocessing and noise removing stage can improve the efficiency further processing.

Special pre-processing algorithms of the entire signal, perform the selection of speech segments, and feature extraction for each segment [2]. Thus, the operation of the automatic

ISBN: 978-1-61804-251-4

text-independent announcer identification includes several stages:

1. Feature extraction.
2. Modeling of speaker.
3. Comparison of the speaker models.

This soundtrack is mapped to the reference speaker soundtrack, by comparing the decision, whether a voice recording belong to this person or different people.

2. FEATURE EXTRACTION

Feature extraction process inherently is not specific to the tasks of speaker identification, but rather is common to most areas of speech technology. For the analysis in the speech signal is assumed to use a set of features such as signal energy, linear prediction coefficients, coefficients of smoothed power spectrum, coefficients of real cepstrum, formant frequencies and pitch period for voiced phonemes. Present correlation between features can be reduced by applying principal component analysis to the vector features.

Energy of signal:

$$E(n) = \sum_{m=-N}^{N} x^2(m) \cdot w(n-m),$$

Where $w(n-m)$ - window function, for example, a Hamming window:

$$h = \left[0.45 - 0.46 \cdot \cos 2\pi n/(N-1)\right], \qquad 0 \le n \le N-1.$$

Linear prediction coefficients, which are the result of solving a system of linear equations Yule - Walker:

$$\sum_{k=1}^{p} a_k R_n(i-k) = R(i), \quad 1 < i <= p,$$

where p – prediction order, $R(i)$ - autocorrelation function, a_k - linear prediction coefficients $1 < k <= p$. Hamming window reduces the prediction error, as the first p samples of a rectangular window with linearly unpredictable.Autocorrelation function calculated with a window:

$$R_n(k) = \frac{1}{N-1-k} \cdot \sum_{m=0}^{N-1-k} \left[x(m)h(n-m)x\right] \cdot \left[x(m+k)h(m-k-m)\right].$$

Formant frequency is determined by the smoothed power spectrum:

$$|H(z)|^2 = \left|\frac{G}{A(z)}\right|^2,$$

where $z = e^{j\omega}$, $H(z)$ - the transfer function of the vocal tract, $A(z)$ - z-transform of linear prediction coefficient sequences.

Cepstral coefficients:

$$c(n) = \frac{1}{N}\sum_{k=0}^{N-1} \log|X(k)| \cdot e^{j\frac{2\pi kn}{N}},$$

where $X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j\frac{2\pi kn}{N}}$ - Fourier transform of signal frame.

Pitch period is determined using the window $l(n)$ for cepstrum:

$$T(n) = c(n) * l(n) \quad l(n) = \begin{cases} 0 \,|\, n| < n_0 \\ 1 \,|\, n| \ge n_0 \end{cases} \quad n_0 = \arg\max(T(n)) \ne 0,$$

where n_0 - pitch period.

Obtained characteristics form the feature vectors $\mathbf{X}$ for each speech segment.

3. MODELING OF SPEAKER

A set of multivariate probability density functions (PDS) describe the hidden acoustic classes of feature vectors. PDS is suitable to approximate arbitrary distributions of the components of acoustic features, making PDS quite convenient for applications in text independent speaker identification and verification.

Usually in the problem of text independent speaker recognition a Gaussian Mixture Models (GMM) are used. GMM is a speaker probabilistic model for multivariate probability density functions (PDS). This model has the obvious disadvantage is that the distribution of acoustic features of speech signals are non-Gaussian, distributions are more peaked. A family of PDS of various shapes, are characterized by three parameters: the expectation m_x, standard deviation σ_x^2 and shape parameter α.

$$f(x) = \frac{\alpha}{2\lambda\boldsymbol{\Sigma}_x \cdot \Gamma\left(\frac{1}{\alpha}\right)} \cdot \exp\left(-\left|\frac{\mathbf{x}-\mathbf{m}_x}{\lambda\boldsymbol{\Sigma}_x}\right|^{\alpha}\right), \quad (1)$$

where $\Gamma(a) \equiv \int_0^{\infty} x^{a-1}\exp(-x)dx$

The distribution function has the form

$$F(x) = \int_{-\infty}^{\left(\frac{x-m_x}{\lambda\sigma_x}\right)^{\alpha}} \frac{1}{2\tilde{A}\left(\frac{1}{\alpha}\right)} \exp^{-\zeta}\zeta^{\frac{1}{\alpha}-1}d\zeta,$$

Scale parameter ß = $\lambda\sigma_x$ of distribution depends on the multiplier λ, which is expressed in

ISBN: 978-1-61804-251-4

terms of the shape parameter α according to the relationship:

$$\lambda = \sqrt{\frac{\tilde{A}\left(\frac{1}{\alpha}\right)}{\tilde{A}\left(\frac{3}{\alpha}\right)}}$$

Centers of GMS are proposed to determine using Radial-Basis Function (RBF) network. Centers of RBF and other parameters of network undergo a supervised learning process. The most convenient for RBF network learning is a gradient descent algorithm that represents a generalization of the Least Mean Square (LMS) algorithm.

The family of RBF networks is broad enough to uniformly approximate any continuous function on a compact set.

Family of RBF networks consists of functions represented by:

$$F(\mathbf{x}) = \sum_{i=1}^{m} a_i \phi(\mathbf{w}_i^T \mathbf{x}) \qquad (2)$$

where m - the number of neurons in the first layer, a_i, $\mathbf{w}_i$ - coefficients of neural network, $\phi(.)$ - the activation function.

As the activation function $\phi(\mathbf{w}_i^T \mathbf{x})$ in the expression (2) a family of exponential distributions (1) with the shape parameter α is proposed.

Calculating the mean square error of approximation of the mixture of multidimensional sampling distributions:

$$\varepsilon = \frac{1}{2}\sum_{j=1}^{N} e_j^2$$

where N – is the size of the training sample .

where N - is the size of the training sample.

Error signal defined by:

$$e_j = d_j - \sum_{i=1}^{M} w_i f(\mathbf{x_j} - \mathbf{m}_i)$$

where d_j – data. The requirement is to find parameters w_i, $\mathbf{m}_i$, $\mathbf{\Sigma}$, α_i.

For better convergence of the algorithm initial values of parameters are selected. Clustering of the sample data is performed according to the method of *k*-means, which estimates initial value of the centers $\mathbf{m}_i$, the initial values of α_i are chosen close to the $\alpha_i = 2$, correlation matrix $\mathbf{\Sigma}$ is chosen close to diagonal, weights are initialized with random values.

Neural network training procedure is performed incrementally. Changing weights on the next step:

$$w_i(n+1) = w_i(n) - \eta_1 \frac{\partial \varepsilon(n)}{\partial w_i(n)} \qquad i = 1,..m_i$$

$$\frac{\partial \varepsilon(n)}{\partial w_i(n)} = \sum_{j=1}^{N} e_j(n) f(\mathbf{x_j} - \mathbf{m}_i(n))$$

Adjustment of the position of the centers:

$$t_i(n+1) = t_i(n) - \eta_2 \frac{\partial \varepsilon(n)}{\partial t_i(n)}, \qquad i = 1,..m_i$$

$$\frac{\partial \varepsilon(n)}{\partial t_i(n)} = \alpha w_i(n) \sum_{j=1}^{N} e_j(n) f'(\mathbf{x_j} - \mathbf{m}_i(n)) \mathbf{\Sigma}^{-1} (\mathbf{x_j} - \mathbf{t}_i(n))^{\alpha - 1}$$

Adjustment of distribution width:

$$\mathbf{\Sigma}_i^{-1}(n+1) = \mathbf{\Sigma}_i^{-1}(n) - \eta_3 \frac{\partial \varepsilon(n)}{\partial \mathbf{\Sigma}_i^{-1}(n)}, \qquad i = 1,..m_i$$

$$\frac{\partial \varepsilon(n)}{\partial \mathbf{\Sigma}_i^{-1}(n)} = -\alpha w_i(n) \sum_{j=1}^{N} e_j(n) f'(\mathbf{x_j} - \mathbf{m}_i(n)) \mathbf{Q}_{ij}(n)$$

$$\mathbf{Q}_{ij}(n) = (\mathbf{x_j} - \mathbf{m}_i(n))^{\alpha - 1} (\mathbf{x_j} - \mathbf{m}_i(n))^T$$

Adjustment of the PDS shape parameter:

$$\alpha_i(n+1) = \alpha_i(n) - \eta_2 \frac{\partial \varepsilon(n)}{\partial \alpha_i(n)}, \qquad i = 1,..m_i$$

$$\frac{\partial \varepsilon(n)}{\partial \alpha_i(n)} = 2 w_i(n) \cdot$$

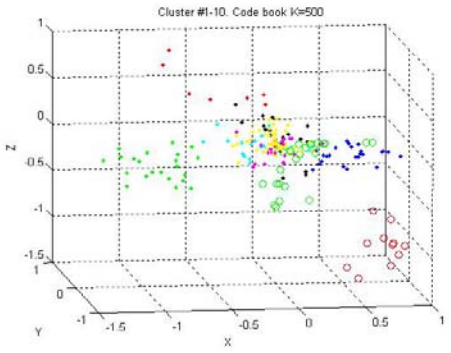
$$\sum_{j=1}^{N} e_j(n) f(\mathbf{x_j} - \mathbf{m}_i(n)) \alpha^{-1} + f'(\mathbf{x_j} - \mathbf{m}_i(n)) \left(\alpha \left| \frac{\mathbf{x} - \mathbf{m}(n)}{\lambda \mathbf{\Sigma}} \right|^{\alpha - 1} \right)$$

4. RESULTS OF EXPERIMENT

The experiment used 15 phonograms recording any text longer than 22000 samples. Analyzed male and female voices same and different speakers, for each phonogram obtained multidimensional histogram features. An example is shown in Figure 1.

ISBN: 978-1-61804-251-4

Figure 1. Projection ща histogram on the three main components of features



K-means obtained initial values of distributions centers. Number of PDS ranged from 10 to 500.

Estimation of errors of the first and second kind for different size of RBF neural network? for different sample sizes and for different speakers, in order to determine the optimal parameters for recording speaker identification.

REFERENCES

[1] A.N. Vasiliev, D.A. Tarkhov Neural Network Modeling: Principles. Algorithms. Applications: Scientific publication/STU. St. Petersburg: Publishing House of STU, 2009, 527 p.

[2] Kotov V.V. Automatic text speaker identification based on a telephone conversation. Science Week XXXIX STU: Proceedings of the International Scientific and Practical Conference. Charles VIII. - St. Petersburg. Univ Polytechnic. University Press, 2010, p. 122-124.

[3] Malykhina G.F. Engineering and technical protection of information: Speech Technology: Textbooks / STU. St. Petersburg: Publishing House of STU, 2004, 243 p.

[4] Pervushin E.A. Basic methods for speaker recognition / / Mathematical Structures and Modeling. - Omsk, 2011, NVyp. 24. - S. 41-54.

[5] Sorokin V.N., V'yugin V.V., Tankin A.A. Information technology in the technical and socio-economic systems. Individual voice recognition: analytical review / / Information Processes. - Moscow, 2012, Volume 12, N 1. - p. 1-30

ISBN: 978-1-61804-251-4

The application of discriminant analysis for estimation of the regional investment attractiveness

IZOTOV ALEKSANDR, ROSTOVA OLGA
Department "Information Systems in Economics and Management"
Saint-Petersburg State Polytechnical University
Address: 195251, St. Petersburg, ul. Polytechnique, 29, III Academic Building, Rm. 303
RUSSIAN FEDERATION
izotovs@gmail.com, o.rostova_isem@mail.ru, http://www.isem-fem.spb.ru

Abstract: The article is dedicated to the problem of selection of the most indicative indicators for estimation of the investment attractiveness and applicability of the discriminant analysis for operative classification of the regions.

Key-Words: discriminant analysis, investment climate, regional rating, factors of investment attractiveness.

1. Introduction

The changes induced by the globalizing processes in the economy make the competition between the federal subjects stronger, which is also valid on the regional level. Here one of the main problems of effective regional development is limited investment resources needed to achieve its strategic goals and tasks. The scarcity of resources makes its negative impact on the economic growth and won't let the desired structural changes in the economy. While competing for the resources positive credit (investment) ratings of various rating agencies play a significant role.

2. Problem Formulation

The investment decisions are one the most difficult when viewed from the selection procedure. They are based on the multivariate and multi-criteria estimations of many factors and trends, often having mixed dynamics. That's why the investment attractiveness estimation for the territory is the most important aspect of investment decision making. The accuracy of estimation affects the consequences for the investor and for targeted the economic system. The more difficult is the situation, the higher is the grade in which experience and intuition of the investor should rely on the results of professionally made estimations of the investment climate of the country or of the region.

Much research have been done and many research papers was dedicated to the investment climate appraisals [1, 4]. The performed analysis of the references allows to make up the comparative characteristics of the investment attractiveness appraisal methods.

2.1. Validity of ratings

The question arises: whether the existing ratings are valid? Recently in the Russian mass media the question of a not-unbiased ratings attributed by the leading rating agencies has widely been discussed. The initiative of creating of our own international rating agencies has become a matter of concern. The task of gaining confidence on the international level is a tough one. The international rating agencies have been earning their good name for decades.

One of the methods of testing the adequacy of the existing ratings and determining the factors exerting the major influence on the investment attractiveness of the region is discriminant analysis. It allows to find out the most significant factors of some given rating agency. The discriminant analysis can be applied only for factors that can be quantified. This poses some limitations, because one can't consider, say, the changes of political environment with the method. However, the quantifiable indicators as a rule make a substantial contribution to the final result.

2.2. Finding the set of significant factors

In the economic literature one can find many methods of estimation of the investment attractiveness of the region. This methods vary depending on the goals of analysis according to the number of analyzed indicators and their qualitative characteristics. Some researchers suggest using up

to 200 of various indicators [7]. Many of the indicators are interrelated and, consequently, are duplicating each other. Moreover, the significance level of each indicator is pretty much a subjective estimation. According to this, the task is to examine the possibilities of discriminant analysis application for classification of regions of the Russian Federation in the investment attractiveness aspect and to determine the most indicative factors.

3. Problem Solution

The first stage of research is determining the set of indicators influencing the investment attractiveness of the region. The analysis of various methods of investment attractiveness estimation has been performed [9]. It should be pointed out that the features of each method depend on different characteristics of investment climate given by the methods. The latter can be explained with the target of analysis as the guideline for the developers of the method. Each method has its own user, i.e. the investor that will make the calculations in order to make an investment decision. The goals of investment climate appraisals have stipulated the different approaches for information sources (statistical data, scientific research data or expert appraisals), for determining of the main factors and indicators, for the organization of the research itself, distinguishing between the directions and stages of research. As a result, the regional ratings from one research are not valid for decision making in other circumstances [5,6].

The accomplished research allows to identify the groups of factors influencing the investment climate of the region:

- The factors of economic development;
- The factors of economic growth dynamics;
- The factors of the social sphere;
- The technological factors;
- The factors of the institutional sphere.

For comparability of the indicators between the regions the factors were calculated in per capita or in percents of the total format.

The regional ratings for the Russian Federation attributed by the international rating agencies Standard & Poor's, Moody's, Fitch and by the Russian rating agency Expert were usen in the research.

The discriminant analysis was carried out on the basis of more than 50 indicators of socio-economic environment of the Russian regions [8,3]. The list of indicators included into the models are outlined below:

Production factors:

f_2 – production of the manufacturing sector, thsd. RUR/pers.;
f_4 – production of the agriculture, thsd. RUR/pers.;
f_5 – mining output, thsd. RUR/pers.;
f_6 – fixed assets investments, thsd. RUR/pers.;
f_7 – fixed assets in the economy, thsd. RUR/pers.;
f_8 – regional inflation, %;
f_9 – building, thsd. RUR/pers.;
f_{10} – production of electricity, gas and water, thsd. RUR/pers.;

The economic growth dynamics factors:

d_2 – industrial production index, %;
d_5 – retail sales turnover index, in % to the previous year;

Non-budget factors:

nb_1 – total financial result (profit minus loss) of the organizational economic activity, thsd. RUR/pers.;
nb_2 – the share of companies with losses, %;
nb_3 – the share of the outstanding accounts payable to the total, %;
nb_4 – the share of the outstanding accounts receivable to the total, %;

The budget factors:

b_1 – the income of the regional budgets, thsd. RUR/pers.;

The factors of labor potential:

l_4 – the number of students per 10000 of population;

The factors of consumption:

p_1 - the regional subsistence wage, RUR/mnth.;
p_2 – the ratio of the average income to the subsistence wage, %;
p_3 – the average wage, RUR;
p_7 – retail sales turnover, thsd. RUR/pers.;

The factors of social environment:

s_1 – the unemployment level, %;
s_2 – the housing level, sq.m./pers.;

The factors of infrastructure:

tr_1 – the density of the roads, km./10000sq.km.;
tr_2 – the automobile cargo transportation, thsd. tn./pers.;
tr_3 – the density of the railroads, km./10000sq.km.;
tr_4 – the railroad cargo transportation, thsd. tn./pers.;

Information and communication factors:

inf_2 – the share of companies using LAN, %;
inf_3 – the share of companies using special software, %;
inf_4 – the share of companies using WAN, %;

Innovative factors:

inn_5 – the innovation activity of companies, %

Institutional factors:

$inst_1$ – number of organizations per 100000 pers.;
$inst_2$ – number of credit organizations per 100000 pers.;

ISBN: 978-1-61804-251-4

$inst_4$ – number of organizations with foreign capital per 1000 pers.;
$inst_5$ – number of small enterprises per 10000 pers.

The analysis is being carried out with the help of STATISTICA using the stepwise discriminant analysis [2]. The method of stepwise inclusion of indicators was used (menu option – forward stepwise). The research of international ratings was performed with the figures of 2012. It should be stressed, that the discriminant analyses gives no opportunity to estimate the influence of factors on the final result. The results of the analysis help to discover the indicators enabling the correct classification of the objects between the groups.

The results below are obtained with the Standard & Poor's rating. Table 1 illustrates the results of the 5^{th} step of the discriminant analysis. The discrimination of the regions is highly significant (Wilks' Lambda =0,0008; F=34; p<0,0000). With the 5% error probability all the variables within the model are statistically significant (the p-level column). The percentage of correct forecasts equals 100% (table 2). Involvement of new variables on the further steps of the discriminant analysis has led to appearance of insignificant factors and didn't improve the quality of classification.

Table 1.

Discriminant Function Analysis Summary (Standard & Poor's) Step 5

N of vars in model: 5; Grouping: SP (4 grps) Wilks' Lambda: ,00008 approx. F (15,16)=33,969 p< ,0000						
	Wilks' - Lambda	Partial - Lambda	F-remove - (3,6)	p-level	Toler.	1-Toler. - (R-Sqr.)
tr_1	0.0972	0.0008	2550.4	0.000000	0.015	0.985
tr_3	0.0063	0.0121	163.9	0.000004	0.008	0.992
d_5	0.0024	0.0315	61.5	0.000067	0.033	0.967
nb_4	0.0013	0.0546	34.7	0.000348	0.097	0.903
p_1	0.0004	0.1981	8.1	0.015683	0.093	0.907

Table 2.

Classification Matrix (Standard & Poor's)

Rows: Observed classifications Columns: Predicted classifications					
	Percent - Correct	BBB p=.143	BB p=.357	BB+ p=.357	B+ p=.1429
BBB	100	2	0	0	0
BB	100	0	5	0	0
BB+	100	0	0	5	0
B+	100	0	0	0	2
Total	100	2	5	5	2

The results of discriminant analysis enable to estimate the contribution of each factor to the total discrimrnation of the regions (table 1, column Partial Lambda). The less the value of Wilks' partial statistics the greater is the contribution of the indicator. In the table 1 the indicators are sorted according to their descending significance for the correct classification.

The most contribution is given by the infrastructure factors. The application of only this factors allows to classify correctly 78% of the regions. The next significant factors are trade dynamics and the share of the outstanding accounts receivable. Adding up of these factors improves the classification up to 98%. Table 3 gives the coefficients of the built discrimination functions.

Table 3.

Classification Functions; grouping: Standard & Poor's

	BBB p=.1429	BB p=.3571	BB+ p=.3571	B+ p=.1429
tr_1	-4.8	-10.3	-10.9	-11.3
tr_3	9.1	18.9	20.0	20.7
d_5	198.9	391.7	414.9	428.9
nb_4	105.1	210.1	223.1	231.1
p_1	0.2	0.4	0.5	0.5
Constant	-6449	-23025	-25828	-27610

The following discriminant functions are obtained:

For the group BBB

$$d_1 = -6449 - 5 \cdot tr_1 + 9 \cdot tr_3 + 199 \cdot d_5 + 105 \cdot nb_4 + 0.2 \cdot p_1$$

For the group BB

$$d_2 = -23025 - 10 \cdot tr_1 + 19 \cdot tr_3 + 392 \cdot d_5 + 210 \cdot nb_4 + 0.4 \cdot p_1$$

For the group BB+

$$d_3 = -25828 - 11 \cdot tr_1 + 20 \cdot tr_3 + 415 \cdot d_5 + 223 \cdot nb_4 + 0.5 \cdot p_1$$

For the group B+

$$d_4 = -27610 - 11 \cdot tr_1 + 21 \cdot tr_3 + 429 \cdot d_5 + 231 \cdot nb_4 + 0.5 \cdot p_1$$

Into the obtained discriminant functions one substitutes the values of the indicators and the object is being attributed to the group with the maximum

ISBN: 978-1-61804-251-4

value of the discriminant function. Using the obtained functions one can classify the other regions or forecast the changes of the regional rating with the changes of any indicator.

The results below are obtained with the Moody's rating. The model that enables correct classification is obtained on the 9th step of discriminant analysis (table 4). The discrimination of the regions is highly significant (Wilks' Lambda =0,000; F=124; p<0,0000). With the 5% error probability all the variables within the model are statistically significant (the p-level column). The percentage of correct forecasts equals 100% (table 5). Involvement of new variables on the further steps of the discriminant analysis has led to appearance of insignificant factors and didn't improve the quality of classification.

Table 4.

Discriminant Function Analysis Summary (Moody's) Step 9

N of vars in model: 9; Grouping: Moodys (4 grps) Wilks' Lambda: .00000 approx. F (27,9)=123.58 p< .0000						
	Wilks' - Lambda	Partial - Lambda	F-remove - (3.3)	p-level	Toler.	1-Toler. - (R-Sqr.)
tr_1	0.001379	0.000026	38942.62	0.000000	0.000027	0.999973
tr_3	0.001057	0.000034	29841.20	0.000000	0.000031	0.999969
d_5	0.000079	0.000446	2242.59	0.000016	0.000371	0.999629
inf_4	0.000029	0.001217	820.89	0.000072	0.000080	0.999920
inf_3	0.000019	0.001841	542.26	0.000134	0.000184	0.999816
p_2	0.000024	0.001482	673.95	0.000097	0.000649	0.999351
f_4	0.000010	0.003642	273.59	0.000373	0.004134	0.995866
nb_2	0.000003	0.012259	80.57	0.002296	0.004204	0.995796
$inst_4$	0.000001	0.060448	15.54	0.024767	0.013582	0.986418

Table 5.

Classification Matrix (Moody's)

Rows: Observed classifications Columns: Predicted classifications					
	Percent – Correct	Ba3 p=.200	Baa1 p=.133	Ba1 p=.333	Ba2 p=.333
Ba3	100	3	0	0	0
Baa1	100	0	2	0	0
Ba1	100	0	0	5	0
Ba2	100	0	0	0	5
Total	100	3	2	5	5

Let's analyze the significance of the indicators for the correct classification. The most contribution is given by the infrastructure factors. The usage of only these factors allows to classify correctly 60% of the regions. The next significant factor is trade dynamics. Adding up of this factor improves the classification up to 80%. The next significant factors are information and communication factors. Adding up of these factors improves the classification up to 93%. Table 6 gives the coefficients of the built discrimination functions.

On the next stage of the research the analysis is made based on the Fitch rating. The model that enables correct classification is obtained on the 18th step of discriminant analysis (table 7).

Table 6.

Classification Functions; grouping: Moody's

	Ba3 p=.200	Baa1 p=.133	Ba1 p=.333	Ba2 p=.333
tr_1	-2145	201	-2101	-2053
tr_3	2788	-260	2731	2669
d_5	21824	-2023	21376	20892
inf_4	45218	-4256	44293	43289
inf_3	-36241	3427	-35501	-34696
p_2	1532	-143	1501	1466
f_4	3405	-320	3336	3260
nb_2	10748	-994	10532	10291
$inst_4$	-243439	22614	-238729	-233101
Constant	-2136971	-20025	-2050454	-1958424

The discrimination of the regions is highly significant (Wilks' Lambda =0,000; F=11,9; p<0,0000). The usage of 16 indicators in the model has led to statistical insignificance of two of them (inst2, S2). However the exclusion of them from the model deteriorated the quality of classification. The percentage of correct forecasts equals 100% (table 8). Table 9 gives the coefficients of the built discriminant functions.

ISBN: 978-1-61804-251-4

Table 7.

Discriminant Function Analysis Summary (Fitch) Step 18

N of vars in model: 16; Grouping: Fitch (5 grps) Wilks' Lambda: .00000 approx. F (72,33)=11.933 p< .0000						
	Wilks' - Lambda	Partial - Lambda	F-remove - (4,8)	p-level	Toler.	1-Toler. - (R-Sqr.)
tr_1	0.000220	0.011688	169.1221	0.000000	0.033896	0.966104
tr_3	0.000024	0.108522	16.4294	0.000633	0.032274	0.967726
p_3	0.000023	0.110187	16.1510	0.000672	0.015443	0.984557
p_1	0.000017	0.148986	11.4241	0.002170	0.013766	0.986234
d_5	0.000012	0.213440	7.3703	0.008605	0.151093	0.848907
s_3	0.000055	0.046400	41.1032	0.000022	0.008836	0.991164
$inst_2$	0.000004	0.593265	1.3712	0.325420	0.205785	0.794215
f_{10}	0.000009	0.299546	4.6768	0.030609	0.053066	0.946935
f_6	0.000017	0.149790	11.3521	0.002215	0.043783	0.956217
inn_5	0.000010	0.248302	6.0547	0.015231	0.147960	0.852040
f_3	0.000039	0.065292	28.6316	0.000086	0.006758	0.993242
s_2	0.000006	0.395216	3.0605	0.083417	0.180648	0.819352
tr_4	0.000013	0.195994	8.2044	0.006221	0.071740	0.928260
d_3	0.000040	0.064572	28.9730	0.000082	0.019700	0.980300
nb_4	0.000018	0.144697	11.8220	0.001938	0.052423	0.947577
f_8	0.000015	0.175261	9.4115	0.004056	0.081494	0.918505

Table 8.

Classification Matrix (Fitch)

Rows: Observed classifications Columns: Predicted classifications						
	Percent – Correct	B+	BB	BB-	BB+	BBB
B+	100	4	0	0	0	0
BB	100	0	7	0	0	0
BB-	100	0	0	9	0	0
BB+	100	0	0	0	8	0
BBB	100	0	0	0	0	2
Total	100	4	7	9	8	2

Table 9.

Classification Functions; grouping: Fitch

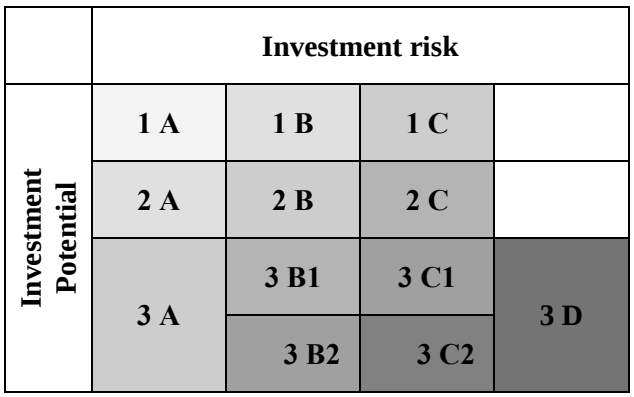
	B+	BB	BB-	BB+	BBB
tr_1	-13	-13	-13	-13	-8
tr_3	-14	-15	-14	-14	-17
p_3	0	0	0	0	0
p_1	-1	-1	-1	-1	-1
d_5	-114	-118	-112	-113	-141
s_3	-10762	-10805	-10533	-10530	-9183
$inst_2$	24846	24232	23524	22548	9866
f_{10}	104	104	101	101	87
f_6	-29	-30	-29	-28	-29
inn_5	-124	-128	-121	-120	-148
f_3	31894	31973	31224	31250	26673
s_2	-206	-202	-202	-204	-109
tr_4	10161	10269	9938	9936	9387
d_3	-241	-242	-236	-237	-205
nb_4	-467	-469	-457	-458	-401
f_8	2714	2725	2665	2668	2395
Constant	-136443	-137119	-131699	-131831	-107177

ISBN: 978-1-61804-251-4

Let's select the most significant indicators for correct classification. The most contribution is given by the infrastructure factors. The next significant factors are consumption potential regional economic growth dynamics. These factors enable classification of 70% of the regions.

In the method of Expert rating agency the investment climate is consisting of: the investment potential (the sum of the objective prerequisites for making effective investments, that are dependent on the existence and the variety of the investment objects); the investment risk (the probability of loosing either the investments or the interest on the investments). The rating actually is distribution of the regions into 12 groups (figure 1). According to the method, the regions, attributed to the 1A group, are the most preferable for investors, while the regions in the 3A group are the less desired investment targets. The classification enables the potential investors to focus their attention only on the regions that are satisfying the needs of the investor most of all when viewed from the expected risk and return [7].

The discriminant analysis was done with the application of the Russian agency Expert data for the period 2006-2012. The results for each year of this period were approximately the same. The differences were only in the percentage of the correctly predicted results, which were between 94% an 100%. Therefore we present only the results for the last step for the year 2012 (table 10):



1A	Maximal potential – minimal risk
2A	Medium potential – minimal risk
3A	Low potential – minimal risk
1B	High potential - medium risk
2B	Medium potential – medium risk
3B1	Reduced potential – medium risk
3B2	Small potential – medium risk
1C	High potential – high risk
2C	Medium potential – high risk
3C1	Reduced potential – high risk
3C2	Small potential – high risk
3D	Low potential – extreme risk

Figure 1. Distribution of the regions into groups

Table 10.

Discriminant Function Analysis Summary, Step 17

Step 17, N of vars in model: 17; Grouping: r (7 grps) Wilks' Lambda: .00422 approx. F (102,252)=3.9784 p<0.0000						
	Wilks' - Lambda	Partial - Lambda	F-remove - (6,43)	p-level	Toler.	1-Toler. - (R-Sqr.)
$inst_1$	0.013214	0.319418	15.26998	0.000000	0.491336	0.508664
f_4	0.009612	0.439082	9.15527	0.000002	0.321723	0.678277
tr_4	0.007620	0.553903	5.77182	0.000179	0.648170	0.351830
l_4	0.005335	0.791097	1.89248	0.104072	0.464392	0.535609
s_1	0.004945	0.853521	1.22993	0.310030	0.613314	0.386686
p_7	0.005740	0.735321	2.57965	0.031828	0.363983	0.636017
$inst_2$	0.006027	0.700337	3.06650	0.013770	0.449365	0.550635
nb_3	0.004876	0.865683	1.11196	0.371323	0.592620	0.407380
f_5	0.005986	0.705091	2.99751	0.015496	0.162681	0.837319
f_9	0.008316	0.507532	6.95397	0.000033	0.044788	0.955212
f_6	0.006729	0.627241	4.25903	0.001881	0.049192	0.950808
d_2	0.004885	0.863945	1.12861	0.362126	0.679495	0.320506
$inst_5$	0.005412	0.779895	2.02261	0.083286	0.512698	0.487302
nb_2	0.005403	0.781231	2.00689	0.085565	0.466530	0.533470
b_2	0.005011	0.842216	1.34263	0.259553	0.664757	0.335243
d_1	0.005434	0.776753	2.05978	0.078133	0.416595	0.583405
inn_1	0.005266	0.801504	1.77486	0.127107	0.540670	0.459330

ISBN: 978-1-61804-251-4

The discrimination of the regions is significant (Wilks' Lambda =0,00422; F=3,98; p<0,0000). Not all the indicators involved in the model appeared to be significant. However the reduction of number of indicators from the model deteriorated the quality of classification. The percentage of correct forecasts equals 97% (table 11).

Two regions (Amur and Pskov regions) are attributed to the group 3B1 by mistake. Let's illustrate the fragment of the table with the distances from these objects to the centers of the each group (table 12). We can see from the table that the distance from these objects to the centers of the groups 3B1 and 3B2 is almost the same. That leads to an incorrect classification. In the table 13 one can see the coefficients of the obtained discriminant functions.

Table 11.

Classification Matrix

	Percent	3B1	3B2	2A	3A1	3C2	2B	1A
3B1	100	34	0	0	0	0	0	0
3B2	80	2	8	0	0	0	0	0
2A	100	0	0	2	0	0	0	0
3A1	100	0	0	0	4	0	0	0
3C2	100	0	0	0	0	6	0	0
2B	100	0	0	0	0	0	9	0
1A	100	0	0	0	0	0	0	3
Total	97	36	8	2	4	6	9	3

Table 12.

Squared Mahalanobis Distances from Group Centroids

Incorrect classifications are marked with *								
	Observed	3B1	3B2	2A	3A1	3C2	2B	1A
Novgorod region	3B2	17.6999	14.4514	45.0583	22.3191	23.3459	51.2070	131.1173
* Amur region	3B2	25.0270	27.6361	69.1655	61.2027	37.7799	40.4197	146.0429
Kostroma region	3B2	19.6512	10.8463	76.7756	55.1213	22.3769	62.2044	170.5927
* Pskov region	3B2	11.8408	13.3354	72.7242	45.3653	42.5260	42.5257	151.5080

Table 13.

Classification Functions

	3B1	**3B2**	**2A**	**3A1**	**3C2**	**2B**	**1A**
$inst_1$	0.02	0.01	0.03	0.02	0.02	0.04	0.08
f_4	-3.64	-3.57	-2.89	-3.21	-3.34	-3.67	-3.08
tr_4	28.88	25.30	42.24	32.48	21.56	44.53	41.56
l_4	0.13	0.12	0.11	0.11	0.10	0.14	0.13
s_1	9.49	9.94	10.72	8.91	11.07	10.11	10.28
p_7	0.45	0.40	0.65	0.60	0.32	0.56	0.58
$inst_2$	-372.72	-194.98	-348.96	-709.21	191.98	-432.64	-682.59
nb_3	3.68	3.71	3.30	3.53	4.02	3.59	3.20
f_5	0.23	0.23	0.19	0.20	0.21	0.21	0.16
f_9	0.19	0.13	0.38	-0.05	0.37	0.40	1.36
f_6	-0.56	-0.55	-0.61	-0.40	-0.58	-0.64	-0.86
d_2	2.84	2.73	2.66	3.14	2.90	2.78	2.97
$inst_5$	0.17	0.16	0.23	0.22	0.13	0.17	0.07
nb_2	3.46	3.49	3.51	3.23	3.75	3.34	4.54
b_2	659.04	639.35	627.22	625.14	639.28	661.62	707.34
d_1	31.89	32.15	30.88	31.10	31.23	32.19	30.06
inn_1	-883.63	-897.63	-780.33	-811.92	-844.34	-871.05	-722.86
Constant	-2254.33	-2245.46	-2179.74	-2194.97	-2194.74	-2317.12	-2272.64

Let's enumerate the most significant indicators for correct classification (from high significance to the low): institutional, industrial, infrastructure factors and factors of labor and consumer potential. These factors enable classification of 81% of the regions.

4. Conclusion

The procedure of searching for vast information sets is a time-taking one because the data needed for

ISBN: 978-1-61804-251-4

rating calculations is usually published with significant delays. The obtained discriminant functions allows to make an operative estimations of regional rating with the usage of several indicators. The research shows that the discriminant analysis enables pretty much precise classifications of the regions.

The most informative indicators for regional classification according to the methods of the international rating agencies appeared to be the infrastructure factors, the economic growth dynamics and the consumer potential factors. As for the method used in the Russian agency Expert the most informative indicators are institutional, industrial and infrastructure factors and the factors of labor and consumer potential. It can be explained with, the fact that each of the analyzed ratings is oriented on their users that are supposed to apply the rating while making their investment decisions.

References:

[1] Burtseva T.A., Problems of statistics, *Indicative model of monitoring of investment attractiveness of the region*, No.6, 2009, pp. 37 - 45.

[2] Dubrov A.M., Mhitaryn B.C., Troshin L.I., *Multivariate statistical metods*, Finansy i statistika, 2011.

[3] *The regions of Russia.2013: Socio-economic indicator*, Rosstat, 2013.

[4] Gradov A.P., Kuzin B.I., Mednikov M.D., Sokolitsyn A.S., *Regional economy*, Peter, 2005.

[5] Guskova T.N., Ryabtsev V.N., *Estimation of investment attractiveness of objects by statistical methods*, Publishing house of RGUTS, 2009.

[6] Kuznetsov S.V., *Investment potential of the region: estimation and implementation mechanisms*, Publishing house of IPE RAS, 2003.

[7] Maksimov I.B., *Investment climate: estimation technique*, Publishing house of BSUEL, 2011.

[8] Surinov A.E., *Russia in figures. 2013: Summary statistical compilation*, Rosstat, 2013.

[9] Suloeva S.B, Rostova O.V., *Management of the investment process in the region (concepts, methods and tools)*, Publishing house of Polytechnical University, 2009.

Approach to Information Requirements Identification of Procurement Process of Custom Production

ANASTASIA I. LYOVINA, ALISSA S. DUBGORN
Department of Information Systems in Economics and Management
St. Petersburg State Polytechnical University
Polytekhnicheskaya str. 29, 195251
RUSSIA
alyovina@gmail.com, alissa.dubgorn@gmail.com
http://www.isem-fem.spb.ru

Abstract: Procurement process has become one of the key business processes of the enterprise in the last few years. This is especially true in regard to manufacturing enterprises focused on custom production, where it is vitally important to have exact amount of necessary production resources for the best price. To guarantee the stable execution of the procurement process the right identification of its information requirements is needed. The paper shows an approach to identify the information requirements of procurement process of manufacturing enterprise, using elements of the enterprise architecture concept and mathematical models of inventory management. The approbation of the given approach is made in the Case Company.

Key-Words: Procurement, purchasing and supply, enterprise architecture, information service, business process, mathematical tool.

1 Introduction

Manufacturing enterprise is a complex organizational and technical system that provides full cycle of production of outputs. The manufacturing process is supported by and depends on a number of support activities: infrastructure, procurement etc.

The complexity of managing a modern manufacturing enterprise is caused by a variety of assets involved and processes implemented that need proper coordination between themselves. In particular, the manufacturing enterprise puts forward special demands on the procurement system, which is caused by the wide range of produced outputs, the wide range of consumed material resources and necessity to comply with the terms of order execution. Effective procurement is one of the key factors of enterprise's competitiveness. Procurement must be organized in such a way as to ensure timely receipt of the necessary production resources and at the same time to avoid the inefficient use of funds for stocking.

Traditional management usually relies on accounting methods rather than optimization ones. As a result, information systems, supporting traditional management, reflect the movement of inventories, but do not contain built-in mechanisms to manage them. Mathematical models of inventory management allow not only to meet requirements for customer satisfaction (internal or external) but also to reduce the cost of resource procurement on one hand and to lower production costs by means of reduction of stocks and amount of work in progress. In this regard, it is reasonable to implement the system of mathematical models in procurement planning and implementation. The article describes the approach to modeling procurement processes of the manufacturing enterprise and to setting requirements for procurement information services involving the use of optimization mathematical models.

2 Problem Formulation

According to the Value Chain Model [1] all activities of the organization are split into 'primary activities' and 'support activities' – the first are facilitated by the latter ones. One of the support activities is "Procurement" – it is a function of purchasing of resources used in the value-creating activities (Figure 1). Value chains provide a high-level organization of the functions that an enterprise performs. To provide a more detailed view, these top-level business functions are broken down to functions of smaller granularity and, ultimately, to activities of operational business processes [2].

Today however, the expectations of Procurement are shifting. As referred to in "A global survey of Procurement functions" of KPMG, not so long ago

Procurement became an add-on service. Many executives are increasingly looking to Procurement to engage the business in strategic conversations about how the supply chain can be optimized to deliver the greatest returns [3].

According to [4] all firms (including Just-In-Time operations) keep the supply of inventory for the following reasons:

1. To maintain independence of operations;
2. To meet variation in product demand;
3. To allow flexibility in production scheduling;
4. To provide a safeguard for variation in raw material delivery time;
5. To take advantage of economic purchase order size;
6. Other domain specific reasons.

Procurement process of manufacturing enterprises is a very information demand process that requires precise, well-timed and reasonable data to be performed appropriately. In line with most of the manufacturing enterprise functions procurement management functions are supported by enterprise information system, in this case Enterprise Resource Planning (ERP) System, which is a part of enterprise IT architecture. The connection between business needs of an enterprise and its IT architecture is realized through information services. Within planning Information systems should be able to support and react promptly and precisely to the requests of business environment.

Procurement process deals with data concerning nomenclature of materials and component parts, suppliers, material consumption rate, scope and time of delivery, warehouse capacity etc. In order to provide IT services that meet the information requirements of users more completely and precisely, mathematical models of inventory management can be used. The approach to information requirements identification of procurement process should include the algorithm to be followed and the list of mathematical models to be used within the certain steps of the algorithm. Such an approach would help to provide a certain level of IT-support of the procurement process.

3 Problem Solution

3.1 Planning the amount of inventory

Inventory is the stock of any item or resource used in an organization. An inventory system is the set of policies and controls that monitor levels of inventory and determine what levels should be maintained, when stock should be replenished, and how large orders should be. [4]

The amount of different kinds of the resources stored in the stock must be monitored on regular basis (Fig.1). The current stock – is the main part of stocks that continuously provides production process before the next delivery. The amount of the current stock depends on the frequency and quantity of delivery and the resource demand from the production. Also for each resource item an insurance stock must be estimated. Insurance stock – is an amount of resource, which supply production in case of unexpected circumstances. It means that if, for example, the next delivery is late, and the resource requirements are covered by the insurance stock. The optimal size of this stock can reduce the costs of its storage and at the same time it has to meet the level of resource demand. Otherwise the supply shortage appeared. What can be the consequences? Of course, this can result in the suspension of production, which generally causes great losses, because time constraints are exceeded. Moreover necessity of urgent search of the supplier causes the increase of total cost of resource.

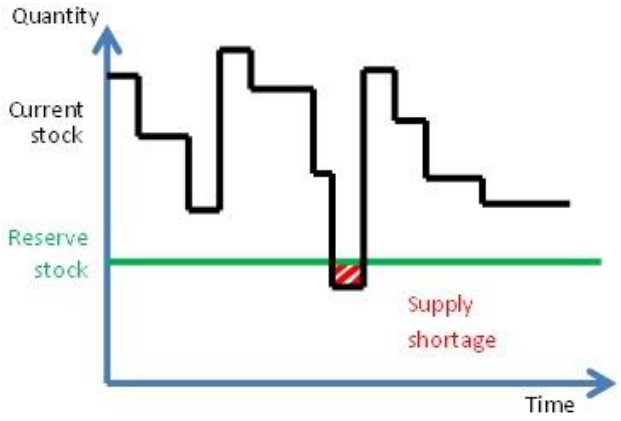


Fig. 1. Example of current stock dynamics

As one of the functions of supply management is reducing expenses on purchasing, transportation and storage of resources it is important to understand the structure of resources cost. It consists of:

1. Purchase cost – usually it is the largest part of the total cost. This price is stated in the document "Order to supplier".
2. Delivery costs – the costs for preparation and transportation of resources. The delivery can be held by the supplier, by Logistics Company or the company can deliver resources on its own. In the last case, the cost will be lower, but not every company have transportation department with necessary equipment. Moreover the cost of insurance is also must be taken in the account especially when we deliver resources by sea or on a long destination.

ISBN: 978-1-61804-251-4

3. Storage costs – costs, connected with warehousing and providing needed storage conditions. It consists of:

- Electricity, water and heating supply of the warehouse.
- The salary of warehouse personnel.
- Taxes and other expenses.

Almost all mathematical models of supply management use these 3 items mentioned above for estimating the optimal order size and period.

3.2 The role of Business and IT in identifying information requirements of the procurement process

Let's consider the Enterprise Architecture concept to understand the importance of identifying information requirements of Procurement process.

There are several standards, frameworks and methodologies of enterprise architecture management such as Zachman Framework for Enterprise Architecture, Extended Enterprise Architecture Framework, GERAM, ISO 19439-2006 and others [5]. All these methodologies have different point of views about how many layers the enterprise architecture model should extend.

The layers show the main elements of the enterprise. Authors of this paper rely on the TOGAF (The Open Group Architecture Framework), which declares following layers of the enterprise architecture [6]:

- Business
- Information or Data
- Application
- Technology

Showing the enterprise architecture model within layers allows specifying the relationships between enterprise core components. The idea is that each layer contains components that execute processes and offer services to the layer above. This concept is shown in Fig. 2.

The alignment between Business and Information Technology is a key issue in every organization and showing, how they can fit together is one of the key objectives of enterprise architecture [8, 9]. This is also true for the manufacturing enterprise and for its procurement process. The execution of procurement business process (Business layer of enterprise architecture) requires various information (Data layer), that can be received with the help of information system of the enterprise (Application layer) using computers, mobile devices or other technical resources (Technology layer).

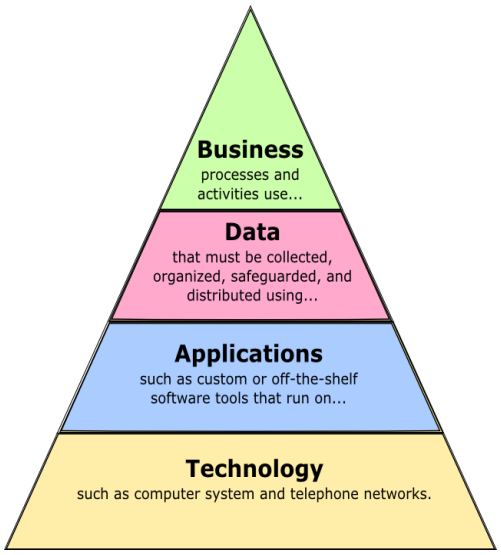


Fig. 2. Layers of Enterprise Architecture [7]

On the level of Business it is important for an enterprise to have the business process management in place. To find and analyze all information requirements of the procurement process, this process itself should be managed in a proper manner. Business process management includes several stages [10], the first and one of the most important is "Design, document and implement process". If the procurement process of the manufacturing enterprise is modeled with the use of appropriate notation, there will be shown functions and events of the process that require specific information. The model of the process will also image the resources of information needed and the possible forms of its presentation.

If talking about IT, it is reasonable to say that the value of IT for business is not in the IT per se, but in providing right IT services in the right way [11]. It is within the scope of the Information Technology Service Management (ITSM) to provide value through the services In the case of manufacturing enterprise and its procurement process, ITSM is focused on providing the process executers with the right information fast enough, in a convenient form of presentation and giving possibility to process this information (for example, if there are mathematical models of inventory management, the relevant calculating tool should be included in the procurement module of the information system). To manage IT services and IT operations in the best way, authors recommend using IT Infrastructure Library (ITIL), which is widely adopted as a framework for ITSM.

3.3 Approach to identify information requirements of procurement process

The following steps should be fulfilled in order to identify all the information requirements of the inventory management process:

1. Analyse and model the procurement process (preferable business process modelling notation – EPC);
2. Define all the needed information inputs and their sources and outputs of the process;
3. Define the type of all information inputs and outputs: primary (raw) data or processed data;
4. For all processes data define the tools and techniques of its processing (including mathematical models);
5. Define the document flow supporting the information flow of the process.

This process needs not only storing and retrieving the data, but requires complicated mathematical calculations as well. The efficient IT support of procurement process is impossible without modern mathematical tools and techniques.

3.4 Mathematical models of inventory management with deterministic demand

To improve the inventory policy of the company for when and how much to replenish the inventory the following steps are used (the 4th step of the approach presented above):

- Formulate a mathematical model describing the behaviour of the inventory system;
- Seek an optimal inventory policy with respect to this model;
- Use a computerized information processing system to maintain a record of the current inventory levels;
- Using this record of current inventory levels, apply the optimal inventory policy to signal when and how much to replenish inventory [12].

"The purpose of an inventory control system is to determine when and how much to order. This decision should be based on the stock situation, the anticipated demand, and different cost factors." [13] In order to support this activity properly, procurement module the enterprise IT system should realize the IT services of calculation and reporting of all the key data of the procurement process. Thus, the main objectives of the inventory management IT services are:

- Calculating optimal size of order (Q^*);
- Calculating optimal frequency of orders (T^*);
- Take into account different types of costs while calculating (TVC) and minimize them.

There is a vast variety of inventory models classification [14, 15]. The most common models used in practice are those based on the economic order quantity (EOQ) model – static model with deterministic demand [13, 14]. Main prerequisites of the models with deterministic demand are:

- Demand is known;
- Instant receipting of product;
- Discounts aren`t considered;
- Deficit isn`t admitted;
- Resources may be analyzed separately.

Basic EOQ model, also known as Wilson formula:

$$Q^* = \sqrt{\frac{2KD}{h}}, \quad T^* = \frac{Q^*}{D} = \sqrt{\frac{2K}{Dh}},$$

$$TVC = K\frac{D}{Q} + h\frac{Q}{2} \to min \qquad (1)$$

The EOQ model with planned shortage:

$$Q^* = \sqrt{\frac{2KD(1+{}^{h}/_{p})}{h}}, \quad T^* = \frac{Q^*}{D} = \sqrt{\frac{2K(1+{}^{h}/_{p})}{hD}},$$

$$S^* = Q^*\frac{h}{h+p},$$

$$TVC = K\frac{D}{Q} + h\frac{(Q-S)^2}{2Q} + p\frac{S^2}{2Q} \to min \qquad (2)$$

The EOQ model with quantity discounts:

$$TVC = cD + K\frac{D}{Q} + h\frac{Q}{2} \to min \qquad (3)$$

The EOQ model with gradual replenishment:

$$Q^* = \sqrt{\frac{2KD}{h(1-{}^{D}/_{R})}}, \quad T^* = \frac{Q^*}{D} = \sqrt{\frac{2K}{hD(1-{}^{D}/_{R})}},$$

$$TVC = K\frac{D}{Q} + h\frac{Q}{2}\left(1-\frac{D}{R}\right) \to min \qquad (4)$$

where P – purchase cost,
K – ordering (setup) costs,
h – holding costs per unit and time unit,
D – demand per time unit,
p – unit shortage cost,
S – maximum shortage,
c – unit acquisition cost,
R – production rate if producing continuously.

The other important inventory management question after the optimal order size is "when to order", i.e. re-order point:

$$ROP = D \times L \qquad (5)$$

where D – demand per time unit,
L – lead time (delivery time) – the time between the placing and receipt of an order.

This means, an order is placed when the inventory level reaches the ROP, and the

ISBN: 978-1-61804-251-4

new inventory arrives at the same moment the inventory is reaching zero. When a safety stock is maintained, then the reorder point is written as the following:

$$ROP = D \times L + SS \quad (6)$$

where SS – safety stock. [16]

3.5 Approbation of the approach in the Case Company

The company Lenpolygraphmash (hereinafter referred as a Case Company) is a manufacturing company that was founded in St. Petersburg in 1890. The core business of the company is developing and manufacturing printing machinery with special functionality for the Ministry of Defense of the Russian Federation, products of led and woodworking industry. [17] Like all companies with complicated production process the Case Company needs a certain level of IT-support for effective business performance. Among other IT-system implementation issues, one of the challenges within procurement module of IT-system was providing a consistent supply management support.

As the Case Company runs a custom production, which means the uniqueness and importance of each single custom order, it requires a smoothly running procurement process supported by the appropriate IT functionality. In order to provide it the clear requirements definition is needed.

The analysis of the Case Company allows modeling the process landscape (Fig. 3), which helps to identify the environment of the purchase and supply process and as a consequence – the sources and the recipients of information from this process. After that the process itself can be analyzed and modelled (Fig. 4).

On the basis of resource requirements from production planning and preproduction processes the overall resource requirements are calculated. After that, a document "Purchase Plan" is created which includes:

1. Nomenclature of resources, its serials and characteristics.
2. The amount of every position of required nomenclature.
3. The time constraints when every position is needed.

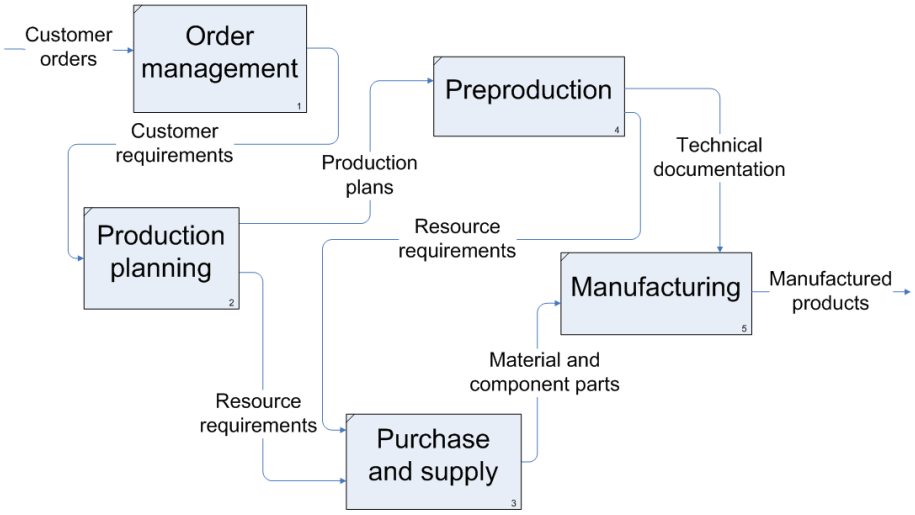


Fig. 3. Context diagram of purchase and supply process

The most complicated function of this process is "Estimating the optimal order size and period". By this function the Supply Department has a list of resources, which are needed for production, and has to decide when (estimating the optimal period) and how much (estimating the optimal order size) the resources will be bought, trying to minimize expenses. It is very important to use adequate mathematical tools, because every mathematical model of supply management has implementation conditions and limitations. Using the information about the prices of resources, average lead time, minimum and maximum batch size and the document "Purchase Plan" a document "Optimal purchase strategy" is created. The latter is supposed to content mathematical models of inventory management with deterministic demand (mentioned in the 3.3) for calculating the parameters of resource and component parts purchasing.

In simple terms, the main aim of the purchase and supply process is to deliver resources with right characteristics, by the right time and to the right place. The consequences of mistakes within the document "Optimal purchase strategy" can be of

ISBN: 978-1-61804-251-4

two 2 types and both can cause big losses:

1. Suspension of production – in case the needed resource have not been delivered in time or delivered in time, but has another characteristics;
2. Increase of carrying costs – in case it was purchased more resources than needed which means the need to store more, than it was planned.

After having created the "Purchase Plan" the suppliers of the resources must be chosen. Different factors are used as the criterion of selecting supplier being: price, approach of just in time delivery, known in industry, size of organization, geographical location, and quality, evaluation of environment, capacity, services, and delay in delivering good, packing, transportation and storing [18]. Traditionally, suppliers are selected among those whom have ability to represent concerned quality, time of delivery and suggestive price. Supplier selection techniques are analyzed in [19].

The supplier selection procedure of the Case company is described in internal documents. Generally, each position has its rating. While analyzing a particular supplier, every position is evaluated and by means of multiplying position ratings the total supplier's points are calculated. The supplier, who gained the best score, is concluded a contract. Traditionally this procedure is used only for new suppliers evaluating. The long-term partners do not need to go through this procedure.

Before the date of purchase comes, order for resource replenishment must have been made. This fact is registered by the document "Order to the supplier". This document must contain the following obligatory positions:

1. A list of ordered materials and component parts, with detail characteristics;
2. Delivery dates;
3. Quality requirements and methods of quality measuring;
4. Order price;
5. Responsibilities of the parties.

After having paid supplier's invoice the Supply Department controls the resource delivery and correctness of filling the forwarding documentation.

Resource arrival is registered by the document "Goods and services arrival". After having verified the quality of arrived resources (incoming inspection) and filling the "Report of inspection", where they register the results of inspection, the checked resources are put to the stock, using the document "Materials receipt ticket".

The outputs of the purchase and supply process are materials and component parts issued to the departments. Resource issuing begins with the processing of the received request for resources from departments. If the previously established limit of resource consumption for the particular department is not exceeded, the requested resource is issued using the document "Material requisition". If the limit is exceeded, the request is corrected and processed again. At the end of the year the total resource consumption are analyzed and the limits can be changed if needed.

It is easy to see that purchase and supply process is very information demanding. It requires a certain documents to be created as well as collecting and keeping of external information for next analysis. Mathematical models are supposed to be used in order to increase work capacity. After having analyzed the purchase and supply process of the Case Company the requirements for supply module of the information system can be set. This module should allow managers of Supply Department to:

1. Maintain all the necessary supply management functions and create appropriate documents – such as:
   - form orders to suppliers;
   - register payment for supplies;
   - register the resources arrival, movement and issue;
   - fix the inventory making, etc.
2. Monitor the execution of supply process by providing analytical information, presented in convenient format of automatically-made reports. For example, report "Inventory level diagram" – a diagram for every resource, which shows the inventory level changing in dynamic, report "Days till new order" – a table that indicates reaching the day of placing a new order, etc.
3. Facilitate a process of making summary resource requirements on the basis of "Production Plan" and specifications by automatic calculation using mathematical models.

The following information requirements for the particular functions of the purchase and supply process were found out:

1. Plan resource requirements:
   a. Annual production plans;
   b. Last year consumption;
2. Estimate the optimal order size and periodicity:
   a. Constrains:
      - Carrying costs, Shortage costs, Delivery costs;
      - Storage conditions;
      - Minimum order size;
      - Maximum order size;
      - Discounts for amount;

ISBN: 978-1-61804-251-4

- Possibility for joint replenishment (items with the same supplier / source city);

b. Supplier reliability (timely delivery, price rising);

c. Convenient values for periods and order size;

3. Check the necessity for making a new order, make an order:
   a. Current inventory level for every item;
   b. Optimal order size, reorder point and periodicity;
4. Register the resources arrival:
   a. Delivery time and costs;
   b. Quality and quantity of resources arrived;
5. Issue resources to the department:
   a. Requested amounts.

The description of the IT services provided all the requirements mentioned above are presented in the Appendix 1.

The effectiveness of process execution is evaluated by performance indicators. Performance indicators of "Purchasing and supply" process after implementing of supply module that includes mathematical models of inventory management can be the following (the list can be modified or expanded):

1. The level of provision departments with resources;
2. Optimization of order, delivery and storage costs;
3. Control of stock reserves limits;
4. Reduction of losses during transportation and storage.

4 Conclusion

In order to provide uninterrupted manufacturing process all the supportive activities should be appropriately organized and computerized where it is needed. Procurement process is very important for the manufacturing and it is very information demanding. To provide the efficient IT support of procurement process it should be analysed carefully from the information requirements point of view. Clearly defined information requirements form the foundation for further development of appropriate IT services.

ISBN: 978-1-61804-251-4

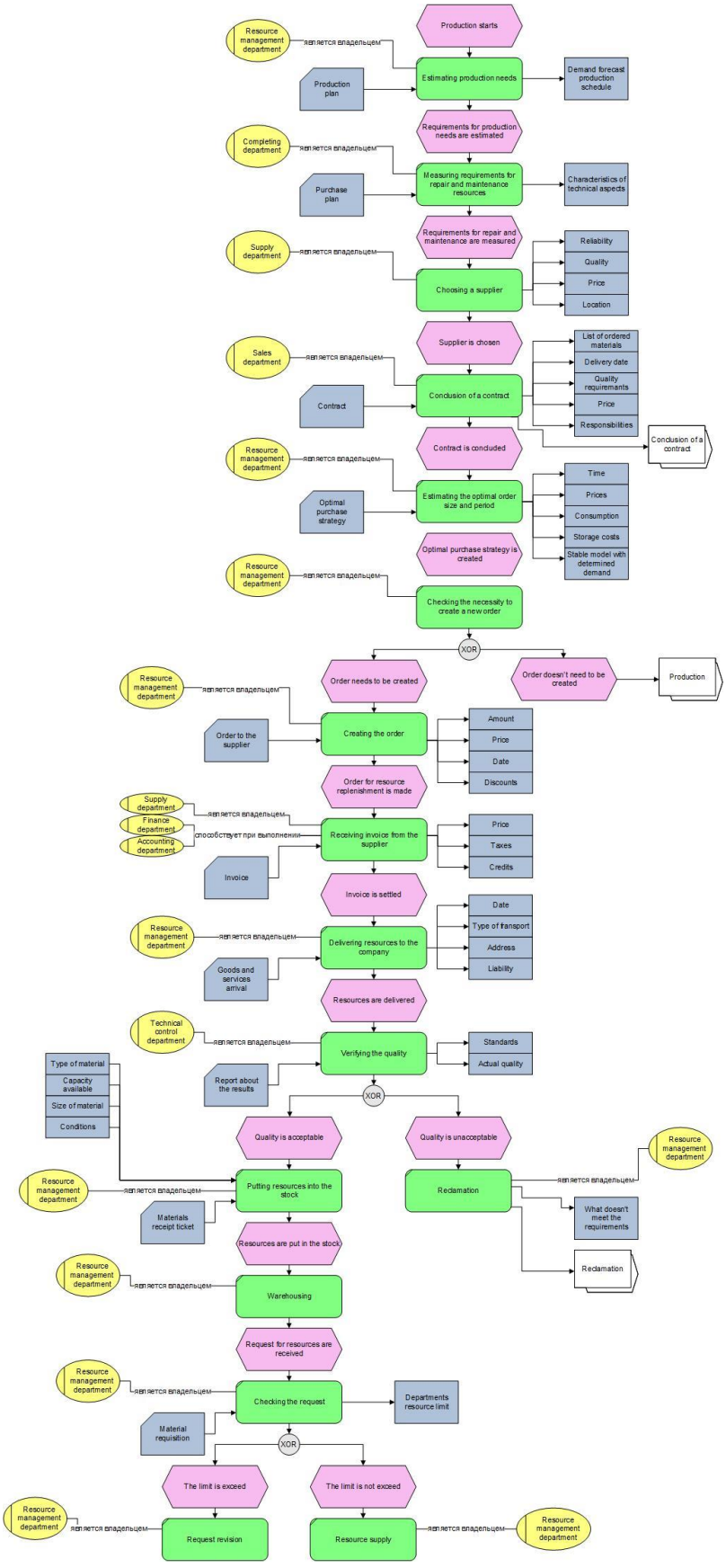


Fig. 4. Example of procurement process of custom production (Lenpolygraphmash manufacturing enterprise (St. Petersburg, Russia)

ISBN: 978-1-61804-251-4

References:

[1] Porter, M. E., *Competitive Advantage: Creating and Sustaining Superior Performance.* New York: Simon and Schuster, 1985

[2] Weske, M., *Business Process Management: Concepts, Languages, Architectures.* Berlin: Springer, 2007

[3] The Power of Procurement. A global survey on Procurement functions, KPMG, 2012 https://www.kpmg.com/US/en/IssuesAndInsights/ArticlesPublications/Documents/the-power-of-procurement-a-global-survey-of-procurement-functions.pdf

[4] Jacobs, R. F., Chase , R., *Operations and Supply Chain Management*, McGraw-Hill/Irwin, 2013

[5] Lankhorst, M., *Enterprise Architecture at Work. Modelling, Communication, Analysis*, Springer-Verlag, 2013

[6] *TOGAF Version 9. The Open Group Architecture Framework (TOGAF)*, TSO, 2009

[7] Hewlett, N.E., *The USDA Enterprise Architecture Program*, PMP CEA, Enterprise Architecture Team, USDA-OCIO, January 25, 2006

[8] Pereira, C.M., Sousa, P., *Enterprise Architecture: Business and IT Alignment*, ACM Symposium on Applied Computing, 2005

[9] Ilin, I. V., Antipin, A. R., & Lyovina, A. I. (2013). *Business architecture modeling for process- and project-oriented companies.* Economics and Management, 32-38.

[10] Becker, J., Kugeler, M., Rosemann, M., *Process Management. A guide for the Design of Business Processes*, Springer, 2011

[11] *IT Infrastructure Library (ITIL®)* V.3., The Open Group, 2007

[12] Hillier, F., Hillier, M., *Introduction to Management Science. A Modeling and Cases Studies Approach with Spreadsheets,* McGraw-Hill/Irwin, 2013

[13] Axsäter, S., *Inventory Control*. Springer, 2006

[14] Muckstadt, J. A., Sapra, A., *Principles of Inventory Management,* Springer, 2010

[15] Hadley, G., Whitin, T. M., *A Review of Alternative Approaches to Inventory Theory*, Santa Monica, California: Rand Corporation, 1964

[16] Gonzalez J.L., Gonzalez D., *Analysis of an Economic Order Quantity and Reorder Point Inventory Control Model for Company XYZ*. California Polytechnic State University, San Luis Obispo, 2010 http://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1006&context=imesp

[17] Lenpolygrafmash official web-site. http://www.lenpoligraphmash.spb.ru/

[18] Shekari H., Afshari M.A., Nikooparvar M.Z. *Developing a Mathematical Model for Optimizing Four Echelons Supply Chain Network Flexibility*. 13th Annual International Conference, New Delhi. Data Views 20.05.2014. http://www.internationalseminar.org/XIII_AIS/TS%205/17.%20Ms.%20Hamideh%20Shekari.pdf

[19] Yang, J.L., Chiu, H.N., Tzeng, G.H. and Yeh, R.H. (2008). *Vendor selection by integrated fuzzy MCDM techniques with independent and interdependent relationships*. Information Sciences 178, 4166–4183.

ISBN: 978-1-61804-251-4

Appendix 1

Description of the information services of supplying process and their content

The following IT-services are involved in the supply and procurement process:

1. Data Book "Nomenclature" – stores the information about all used by the company items of materials and component parts:
    a. Name
    b. Unit of measure
    c. Nomenclature group
    d. Nomenclature type (material / component part / semi-finished product)
    e. Storage conditions required
2. Data Book "Contractors" – stores information about suppliers
    a. Full organization name
    b. Contact information (phone number, e-mail, fax)
    c. Discounts for quantities
3. Document "Production plan" – sets the amounts for items that are planned to be produced in the certain period (year, month)
    a. Period
    b. The list of items with the amounts
4. Document "Purchasing plan" – sets the amounts for items that are planned to be purchased in the certain period (year, month)
    a. Period
    b. The list of items with the amounts
5. Document "Setting inventory models constraints" – for every item in the list sets the following constraints for inventory models:
    a. Carrying costs
    b. Maximum / minimum batch size from every supplier
    c. An average delivery time from every supplier
6. Document "Optimal purchasing strategy estimation" – calculates the optimal output parameters for inventory models:
    a. Optimal order periods for periodic models
    b. Optimal order quantity, Reorder point (ROP) for continuous models
7. Document "Order to the supplier" – fixes the preliminary agreement with the supplier to provide the company with resources till the certain date:
    a. Date
    b. Supplier
    c. Amount
    d. Price
8. Document "Goods and services arrival" – registers resources arrived from suppliers:
    a. Date
    b. Resource item
    c. Supplier
    d. Amount
    e. Price
    f. Order (link to the document)
9. Document "Material requisition" – registers the amount of resources that was delivered to the departments:
    a. Date
    b. Resource item
    c. Amount
    d. Department
10. Report "EOQ comparison" – a table that compares EOQ value with average order size:

Input parameters:

    a. User: Date interval
    b. User: Resource item or nomenclature group
    c. Arrived amount
    d. EOQ

Output parameters:

    a. Resource item or nomenclature group
    b. Avr. order size
    c. EOQ
    d. Delta (in units & %)

11. Report "Inventory level diagram" – a diagram that can be drawn for every resource. It shows the inventory level changing in dynamic:

Input & output parameters:

    a. User: Date interval
    b. User: Resource item
    c. Arrived amount
    d. Delivered amount
    e. ROP level
    f. Optimal periodicity

12. Report "Reorder point reaching" – a table that indicates reaching to the ROP level (for continuous models)

Input parameters:

    a. User: Date
    b. User: Resource item
    c. Current level (up to the Date)

ISBN: 978-1-61804-251-4

d. ROP

Output parameters:

a. Resource item
b. Current level
c. ROP
d. Delta (in units & %)
e. Explanation (order, not to order, prepare to order)

13. Report "Days till new order" – a table that indicates reaching the day of placing new order:

Input parameters:

a. User: Date
b. User: Resource item
c. Last order moment
d. Optimal periodicity

Output parameters:

a. Resource item
b. Optimal periodicity
c. Days from last order moment
d. Delta (in days & %)

ISBN: 978-1-61804-251-4

Economic and Mathematical Models and Statistical Models of Operational Planning

V. A. LEVENTSOV
Institute of Industrial Economics and Management
St. Petersburg State Polytechnical University
RUSSIA
Polytechnicheskaya Street 29, St. Petersburg, 195251
vleventsov@spbstu.ru

Abstract: The article considers the issues related to the use of mathematical tools in the form of methods and models in drawing up production schedules for production floors at industrial enterprises, which, to a large extent, depend on optimal employment of the resources available. The models that have been suggested take into account various technical, technological, economic and business features of production. This model is distinctive due to its multicriteriality. At the same time, in order to simplify the process of production schedules elaborating, the paper proposes a statistical model that enables to facilitate operational plans and to consider, without loss of quality, the entire variety of business and economic production parameters.

Key-Words: Operational planning; mathematical and statistical models; profitability of production; production plan; economic and mathematical model; production scheduling.

1 Introduction

Efficiency of an industrial enterprise depends directly on its production and economic management system and. to a large extent, on its in-house operational planning. Competitiveness of an enterprise is strongly dependent on how flexibly production planning reacts to a change in market conditions.

In the current market conditions, the following factors considerably increase the role and significance of in-house operational planning for many important management activities: focus of production on demand; quick reaction to change in demand through modifying the line of goods and production output volume; possible deviation of the actual production process from operational schedules. These factors are especially difficult to consider in recurring production.

Today we have accumulated a sufficient range of economic and mathematical models and methods of operational planning in recurring production, but this problematic field still needs to be studied.

In the mid 1950s systematic and very profound research was initiated in order to build and analyze mathematic scheduling models, to elaborate and use routine decision-making methods. First examples of successful research that was carried out at that time and is worth mentioning include network scheduling methods. Some interesting results were obtained in the field of queuing systems [1, 13]. At this very time the term "scheduling theory" appeared [4, 13].

Today, since industrial production in Russia is experiencing revival, there is growing interest towards scheduling problems. However, as many authors say [2, 6, 7, 13, 15, 17, 19], new real objectives in the scheduling theory have appeared and caused certain difficulties. So it is reasonable to expect a rise in relevancy of operational planning methods.

The analysis of papers dedicated to the scheduling theory lets us conclude as follows.

First, all papers can be divided into four groups:

- individual problem stating and solving in the scheduling theory;
- problem solving methods;
- applied problem solving in the conventional sector;
- new areas to apply the scheduling theory.

It has to be mentioned that publications, as a rule, address individual cases among general ones. We can make a note of the following problems: Johnson [3], Akers, Friedman, Lenstra & Reeg, Lenstra & Rinnooy Kan. In their turn, individual problems are marked off general ones in the mentioned class. For example, Bellman-Johnson problems for two machine tools, the same for three machine tools; two service problem; single route Johnson's problem; conveyer type problems. Thus, the literature sees into a wide range of

ISBN: 978-1-61804-251-4

individual problems of the scheduling theory and calendar planning quite in detail. However, the aforementioned publications do not have general solutions to problems of the scheduling theory; tough problems of the scheduling theory are approached separately and indeterminate problems are mentioned.

The number of models and degree of their similarity and versatility are gradually growing and grasping a larger scope of possible applications – scheduling of production, transport, military operations, teaching, IT processes, etc. As these models are getting more and more sophisticated the same is happening to routine decision-making methods that use these models.

The following may be said about the existing problem-solving methods. The scheduling theory objective is a specific optimization problem and practically all optimization methods known today are used to solve it. The scheduling theory objectives come down to the problem of mathematic programming: linear, non-linear [17], dynamic, integral-valued, and discrete ones [18]. Network setting of a problem [1, 13], the game theory method [5] are widely used. Asymptotic methods and methods of multi-extreme problem-solving [2, 14] are studied. The combining method, the graph theory method [1, 7], in particular, the mixed graph method [7]. Methods of multi-criterion problem-solving are presented, methods of stability analysis of problem-solving in the scheduling theory [16] are considered. A great number of research papers are dedicated to approximation methods: the Monte Carlo method [12], "bottleneck" method, etc.

Analysis of practical application of the scheduling theory methods proves their high efficiency.

The simplest and most commonly used operational planning method is a graphic method. Charts and graphs are easy to read and allow managers to analyze production capacity utilization and improve operational plans following their gut feeling and common sense.

The Gantt chart of production capacity utilization has several major limits of use. One of them is that it does not take into account the variety of production situations, such as breakdowns or human errors, which demand to do the same job again. The chart has to be regularly updated when new works appear or time assessments are revised. The drawback of simple graphic methods is that dependences between operations are not clearly seen. It is especially noticeable when the process is complicated, sub-divided and includes a lot of operations.

This drawback of graphic methods excludes network analysis. The complex of works to be planned is reflected as a network model. It gives more opportunities to calculate time characteristics and simulate situations. At the same time the network modeling of a work complex does not allow us to see the workplace capacity charts.

Different techniques, which we use today, treat the problem of optimal values calculation of the given norms in a different way. Table 1 includes comparison of the common scheduling techniques [13].

In the context of scheduling problem-solving on big amounts of information, the 2nd, 6th and 8th techniques can be of the biggest practical interest. If these techniques, which contain different logical rules and priority functions, are used, scheduling problems can be simplified and actually solved.

Table 1 – Main Scheduling Techniques

Scheduling technique	Brief characteristic of the technique
1. Reverse scheduling (Gantt charts)	Schedule of operations needed to satisfy demand is modeled as straight lines on time axis in reverse direction from the date of completion.
2. Direct scheduling (Gantt charts)	Modeling goes forward from the set date to the date when the complex of works will be completed.
3. Network analysis	It is used in a similar way as direct and reverse scheduling, but can reflect more sophisticated logical interrelations and interdependences between works that have to be completed as part of a project.
4. Modular retrievable packer (MRP); manufacturing resource planning (MRPII)	It is similar to reverse scheduling, but considers stock management and capacity management.

ISBN: 978-1-61804-251-4

5. "Just in time" concept (JIT)	Due to rational organization of production environment and with the help of the information system "KANBAN", operational scheduling presents no problems
6. Law of priorities	Determination of the best order to pass a given complex of works through a given sequence of workplaces.
7. "Travelling salesman problem"	It allows minimizing the total time needed to readjust the system; gives an accurate solution.
8. Work schedule and assignments	Schedules and assignments are set when certain work centers are available for those who want to use them. They are used in service systems, as a rule.
9. Optimization of production technologies	Scheduling of material flow through a "bottleneck" of the process

Thus, the aforementioned allows us to conclude that the problem of the scheduling theory and calendar planning with the use of new methods is an important and relevant scientific problem.

2 Problem Formulation

Today's approaches to operational scheduling of workplaces as an optimality criterion, as a rule, use various time characteristics that do not always properly reflect the dynamics of economic parameters of production.

The main scheduling problems are presented in Table 2.

Table 2 – Approaches to selection of the best scheduling variants

Year	**Authors**	**Criterion**
1966	J. Muth, J. Thompson, P. Winters.	Cost minimization (stocks, equipment adjustment, supply lags)
1975	Conway P.V., Maxwell, V.L., Miller, L.V.	Minimization of the beginning moments sum of final operations for all works
1979	Sokolitsyn, S.A.	V.A. Petrov's method is used as a basis: minimization of the production cycle length with the use of scheduling guidelines (priority sequence depends on whether production time of separate item batches increases or decreases).
1982, 1988	Kyzin, B.I.	Minimization of the total production cycle length with the use of priority laws and preference functions
1984	Tyutyukin V.K.	Minimization of the total production cycle length through finding successive locally optimal plans
2002	Tsarev V.V.	Minimization of planned employment variances from actual employment of workers. Minimization of production in progress

The main drawback of most scheduling models that are recommended is the use of a single criterion approach to a given problem. However, the scheduling problem is, in its own sense, a multi-purpose task: on the one hand, it is necessary to complete the production program to a full extent. On the other hand, it has to be done with maximal efficiency of production. Thus, in scheduling, costs related to allocation and completion of orders can change and the profit share that the workshop

obtains can vary due to different scopes of work or as a result of dynamics in the used time resources. Therefore, when choosing an index of production in progress as an optimality criterion for operational scheduling of workplaces, working assets of an enterprise are consciously and evidently understated, because scheduled work results in longer production cycle which is a major factor affecting the size of production in progress. According to the research of authors, when a workshop operates in accordance with the elaborated schedule, the size of production in progress is 1.5-2.0 times bigger in comparison to the value that is used to calculate rated working assets. If this condition is not considered, the value of rated working assets will be understated and, as a result, the production program of an enterprise may fail.

Moreover, results of the research have shown that the production plan cannot always be fully completed in the current planning horizon if the former one has been elaborated based on the common model with restrictions of resource facilities. Hence, the equipment gets less used in comparison to the normative level and there is a drop in stock at the supply warehouse with simultaneous increase in the stock at the finished products warehouse due to the line of goods made to cooperation.

In recurring production it is economically important to manufacture the order in bigger batches. If the batch size of parts put into production grows, less time is needed to readjust the equipment, the fund of fieldwork per shift increases, labor productivity and quality of produced goods improve. Furthermore, less amounts of materials are consumed and less working time is spent to manufacture additional samples that are used for customization technological purposes, for example, quality control of blanking operations.

Economic performance of a company also improves because of steadier output and more regular work of production divisions. Regular operation of workshops and sections is reached through more accurate planning of demand for production capacities and a more balanced utilization of separate groups of equipment and workplaces during a planning horizon. Utilization is usually balanced through changing calendar dates of goods production and transferring production of separate lines of goods between divisions.

With less overtime work and lost hours due to organizational reasons, cost of manufactured goods goes down and their quality improves.

Thus, operational planning of production in the existing market conditions is a useful tool to improve efficiency of in-house planning as a whole. Enhanced operational planning helps to obtain a positive financial result from company's activities, brings additional profits and improves profitability.

So, it is reasonable to develop such an optimization model that would be based on implementing a multi-purpose approach and consider the essential features of schedules. More appropriate optimality criteria of the scheduling problem, in our opinion, include maximization of production profitability and compliance with the scope and line of good of the production plan by means of internal resources of a workshop.

One of possible types of economic and mathematical scheduling models for machine shops can be an optimization model that uses a multi-criterion approach and is one of the structural elements in the operational production management scheme.

3 Problem-Solving

Let there be a machine shop (with subject or technological specialization) that has a certain number of machines, including doubling machines. Technological operations for production of batches of parts on the corresponding machines within a given time period have to be organized in such a way that parts are produced and delivered within certain deadlines. Technological routes for parts to pass machines are different. Each separate batch of parts is characterized with preparation time standard whereas each part is defined by the content and sequence of technological operations and standards of floor-to-floor time. Operation of workplaces of the machine shop has to be scheduled in such a way that a number of major requirements are met and certain objectives are reached, being represented as optimality criteria.

Most important requirements to be considered in the optimization model are the following:

- each machine can do no more than one part-operation;
- before parts have been processed in the previous operation of the technological process, beginning of their processing in the following operation cannot be planned;
- the same process operation to manufacture a part can be planned for any doubling machine that is off-duty;
- process of part-operation manufacturing on a machine cannot be interrupted;
- workplaces are loaded within the limits of the

ISBN: 978-1-61804-251-4

normative time reserve.

In the scheduling problem, beginning and end time of technological process operations is unknown for certain machines and different batches of time.

The analyzed literary sources have brought us to a conclusion that various criteria and algorithms are used in operational scheduling of machine sections workplaces whereas multi-criterion approach is mainly used in scheduling tasks. The economic and mathematical scheduling model of workplaces operation for subject and technological sections of machine shops that is proposed here includes two optimality criteria: percent completion of the production plan by means of the internal resources of a workshop and production profitability.

The scheduling problem can be presented in the following way [9,10]. It is necessary to choose such a variant of the standard-plan (schedule) of workplaces operation of a machine shop that

$$P(\mu_m^*) \to \max,\ \mu_m^* \in M_{\text{вп}}; \quad (1)$$

$$R(\mu_m^*) \to \max,\ \mu_m^* \in M_{\text{вп}}; \quad (2)$$

$$0 < R(\mu_m^*) \le R_{\text{и}}; \quad (3)$$

$$0 < P(\mu_m^*) \le 100; \quad (4)$$

$$M_{\text{вп}} = \{\mu_m : 1 \le \mu_m \le K_{\text{вп}}\}; \quad (5)$$

$$K_{\text{вп}} = (\pi!)^q, \quad (6)$$

where

μ_m – number of the schedule variant of workplaces operation of the workshop for m iteration of calendar scheduling;

μ_m^* – number of the best schedule variant for workplaces operation for m iteration of its development;

$M_{\text{вп}}$ – manifold of principally possible variants of schedules for workplaces operation;

$K_{\text{вп}}$ – total number of principally possible variants of schedules for workplaces operation;

π – total number of batches of parts (orders) that represent the line of goods in the production plan of the workshop;

q – number of pieces of production equipment in the workshop;

R – production profitability if the workshop operates by μ_m variant of schedules that reflects efficiency of workshop's contribution into the net profit value of the company in relation to the share of the company capital (fixed and working one) allocated to the workshop for operational management, %;

$R_{\text{и}}$ – production profitability mediated by the line of goods and scope of the production plan set for the workshop, established prices and structure of production prime costs, amount of fixed production assets of the workshop and working assets that depend on the accepted batching of products, %;

P – percent completion of the production plan by means of the workshop's internal resources if it operates in accordance with μ_m variant of the schedule.

Objective function (1) reflects the condition for search of such a variant of the schedule that would result in the biggest percentage of the workshop production plan implemented with the use of internal resources of the workshop.

Objective function (2) contributes to finding such a variant of the schedule that would provide maximal production profitability.

Restriction (3) allows considering only those variants of the schedule that would make it possible to get a positive production profitability value and would not exceed the value that the workshop could get if it operated in ideal conditions.

Restriction (4) reflects the need for complete or partial implementation of the workshop production plan by means of internal resources of the workshop.

Formulae 5 and 6 reflect the manifold and number of possible variants of the schedule at the set qualities of batches of parts (orders) $-\pi$ and number of pieces of technological equipment $-q$.

Practical experience shows that it takes a lot of time to elaborate one variant of the schedule if no computer is used. Thus, to develop one variant of the part-operation schedule for a large machine-tool section, which produces 120 or more types of parts per year, about 20 man-months are needed [20]. Such a big time consumption and need for schedules to operate efficiently justify application of modern personal computers to elaborate schedules for sections and workshops, which considerably diminish labor intensity of scheduling. Thus, it takes 1-4 hours for a computer to elaborate one variant of a part-operation schedule, depending on the dimension of the problem and complexity of the calculation system, which are conditioned by the optimization model applied and mathematic method of its implementation [20].

The two-criterion model elaborated for scheduling of a machine shop calls for lots of input business, economic and technological parameters to be implemented, which are not always available to the full extent for an operating enterprise [11, 21]. This drawback can be eliminated with a more compact economic and statistical scheduling model

for a machine shop. It just needs few technological and organizational parameters, such as operational technological routes, norms of floor-to-floor time by operations, structure of the technological equipment park, line of goods and scope of the production plan, time reserves, workplace capacity with one type of technological operation (work) within one month and some others [8, 10].

Study of cause and effect relations between indices of profitability, recurring production level, equipment utilization and others has allowed finding the factors with the biggest impact on production profitability variation.

This statistical study of economic processes at an enterprise has revealed connection between the phenomena that are reviewed, its qualitative assessment and analytical expression.

In our opinion, change of the effective condition or criterion index of the economic and mathematical model – production profitability is conditioned by change of such factors as coefficient of workplace capacity with one type of work within one month, an average size of a labor subject batch, capacity efficient of technological equipment or standard calculated floor-to-floor time. As a result, the only economic parameter of the model is production profitability in the form of functional dependence on the coefficient of workplace capacity with one technological operation (work) within one month.

The economic and statistical model [22] of scheduling for a machine shop is formed in the following way. It is necessary to choose such a variant of the schedule for workplaces of the machine shop so that can be identified as

$$P\left(\mu_m^*\right) \to \max, \quad \mu_m^* \in M_{\text{вп}} \tag{7}$$

$$r_{kc}\left(\mu_m\right) = \begin{cases} 0.1024 + 0.0084Ln(K_{\text{т}} - 0.023), \\ \text{if there is subject specialization} \\ \text{of the workshop operating structure;} \\ 0.0554 + 0.5793K_{\text{т}} - 2.9101K_{\text{т}}{}^2, \\ \text{if there is technological specialization} \\ \text{of the enterprise operating structure;} \end{cases} \tag{8}$$

$$0 < P\left(\mu_m^*\right) \le 100; \tag{9}$$

$$M_{\text{вп}} = \left\{\mu_m : 1 \le \mu_m \le K_{\text{вп}}\right\}; \tag{10}$$

$$K_{\text{вп}} = (\pi!)^q, \tag{11}$$

where $r_{\text{кс}}$ – production profitability in the form of functional dependence; $K_{\text{т}}, \overline{K}_{\text{т}}$ – coefficient of workplace capacity with one work within one month in the workshop that characterizes the capacity degree of one impersonal piece of equipment with one production work (part-operation) that belongs in the monthly scope of work of the workshop.

Expressions of the economic and statistical model (7), (9), (10), (11) do not require clarifying since they have not been reviewed in the economic and mathematical model.

The expression of the economic and statistical model (8) is analogous to the expression of the economic and mathematical model (2), the only difference being the fact that the production profitability value, which is obtained in a statistical way and which represents a functional dependence, is used in its restrictions.

4 Conclusion

Increased production efficiency is an important problem for mechanical engineering enterprises. One of the possible ways to increase production efficiency in market relations is an improved in-house operational planning and management system.

Improved operational planning in recurring production, due to its specific performance features, implies development of such operational planning models that would fully use the potential of information technology.

Theoretical and methodological basis of the research relies on papers of Russian and foreign scientists on mathematic and instrument economic methods of scheduling for machine shops of mechanical engineering enterprises. Problems have been solved through principles of comprehensive approach, theory of sets, economic and mathematical modeling, including simulation modeling, statistical and logical analysis.

Thus, the research results have allowed the author to propose a new two-criterion approach to finding an acceptable scheduling option for workplaces of machine sections that would consider both demand for goods of the enterprise and production profitability. This approach has been used to develop an optimization two-criterion economic and mechanical model and one-criterion economic and statistical model of operational scheduling for machine shop sections.

The optimization model that has been developed is based on implementation of multipurpose approach and considers the essential feature of operational scheduling for workplaces. The model uses two optimality criteria: maximization of production profitability and implementation of the

production plan by means of the internal resources of a workshop.

References:

1. [1] X1. Vagner G. *Fundamentals of Operation Research*, Moscow, 1972. – Vol.3: – P. 98-150.
2. [2] X2. Goncharov V.N. *Operational Production Control*. (Experience of system development and enhancement). – Moscow: Ekonomika, 1987. – 120 p.
3. [3] X3. Johnson S. *Optimal Two- and Three-Stage Production Schedules with Setup Times Included*. – In the book: Kibern. St. Petersburg. (New series). Issue 1. – Moscow: Ekonomika, 1965. – P.78-86
4. [4] X4. Kantsedal S.A. On Scheduling Classes / Kantsedal, S.A. Malykh, O.N. // Kibernetika. – 1981. – No. 6. – P. 66-74.
5. [5] X5. Kostevich, L.S. *Game Theory. Research of Operations* - Minsk: Vysshaya shkola, 1982. – 232 p.
6. [6] X6. Kuzhin, B.I. Yuriev, V.N. Shakhdinarov, G.M. *Methods and Models for Running a Company*. – St.Petersbug: Piter, 2001. – 432 p.
7. [7] X87 Labsker, L.G. Babeshko, L.O. *Game Methods in Managing Economy and Business*. – Moscow: Delo, 2001. – 464 p.
8. [8] X8. Leventsov, V.A. Valentik-Levitskay, E.G. Shnitin, Y.V. *Research of Economic Parameters Dynamics If Organizational Production Conditions Change (Object-Closed Sections of a Machine Shop: Case Study)*. In the book: XXX Anniversary Week of Science of St. Petersburg State Polytechnical University. Part IX: Materials of the inter-university scientific conference. St. Petersburg: St.Petersburg State Technical University, 2002. – P. 168-171
9. [9] X9. Leventsov, V.A. Shnitin, Y.V. *Simulated Model of Scheduling.* In the book: Scientific and technical news of St. Petersburg State Technical University, 4(46)/2006. St. Petersburg: Polytechnical University, 2006, p. 325–331
10.[10] X10. Leventsov, V.A. Shnitin, Y.V. *Analysis of Organizational and Economic Production Parameters in Drawing Up Schedules*. Works of St.Petersburg State Technical University, 503/2007. St. Petersburg: Polytechnical University, 2007. – p. 103-114
11.[11] X11. Leventsov, V.A. Shnitin, Y.V. *Two-criterion Model of Scheduling for Workplace Operation.* In the book: Economics and Management of a Contemporary Enterprise: Problems and Prospects: Papers of the 9th research and practice conference. 20-25 September 2007. St. Petersburg: Polytechnical University, 2007. – P. 203–211
12.[12] X12. Lunev, V.A. *Mathematical Modellin and Planning of an Experiment*. – St. Petersburg: Polytechnical University, 2006. – 164 p.
13.[13] X13. Makarov, V.M. *Diversification of a Production Management System under Conditions of a Dynamic Demand: Theory, Methods, Algorithms.* – St. Petersburg: Polytechnical University, 2002. – 351p.
14.[14] X14. Mischenko, A.V. Kovalev, M.I. *Credit Resource Management in an Enterprise Belonging in a Real Sector of Economy // Management in Russia and Abroad*. – 1999. – 14. – P.112–124
15.[15] X15. Petrov, Y.A., et al. *Enterprise Management Automation: Information Technology – Management Theory and Practice* / Petrov, Y.A. Shlimovich, E.L. Iryupin, YV.- Moscow: Finance and Statistics, 2001. – 160 p.
16.[16] X16. Salamatin, N.A. Panfilova, E.E. *Production Program Management in Mechanical Engineering Enterprises: Textbook.* Part 1-2. – Moscow: The State University of Management, 2000. – 266 p.
17.[17] X17. Salamatin, N.A. Fel, A.V. Shalamova, N.G. *New Information Technology in Production Management.* – Moscow: The State University of Management, 1996. – 130 p.
18.[18] X18. Sergienko, I.V. *Mathematical Models and Methods of Problem-Solving for Discrete Optimization*. – Kiev: Naukova Dumka, 1985.-384 p.
19.[19] X19. Faiengold, M.L. Kuznetsov, D.V. *Technique to Calculate Production Cycle and Calendar Plans of Production Output* / Under scientific edition of Faiengold, M.L. – Vladimir:

Vladimir State University, 2003. – 111 p.
20.[20] X20. Tsarev, V.V. *In-house Planning*. – St. Petersburg: Piter, 2002. – 496 p.
21.[21] X21. Shnitin, Y.V. Leventsov, V.A. *Milti-Criterion Optimization When Scheduling Workplace Operation in a Machine Shop.* In the book: Economics and Management: Theory and Practice. Management of Structural Transformations in the Economy of Russia: Papers of the 8th research and practice conference. 20-25 December 2006. St. Petersburg: Polytechnical University, 2006. – P. 655-659
22.[22] X22. Shnitin, Y.V. Leventsov, V.A. *Statistical Model of Operational Production Management (Machine Shop: Case Study)*. In the book: Economics and Management: Problems and Prospects: Papers of the International Research and Practice Conference. 6-11 June 2005. St. Petersburg: Polytechnical University, 2005. – P. 457-460
23.[23] X23. Yuriev, V.N. Kuzmenkov, V.A. *Optimization Methods in Economics and Management.* – St. Petersburg: Polytechnical University, 2006. – 804 c.
24.[24] X24. Heizer J.H., Render B. *Production and Operations Management: Strategies and Tactics.* 3 th ed. Boston, Allyn and Bacon, 19

Dynamic State Model of Steam Turbine Hall Equipment Condition for Maintenance Planning and Decision-making Support

Lenka Jirsová and Miroslav Flídr

Abstract—The article describes development of a dynamic state model of steam turbine hall mechanical equipment condition with respect to its wear. The model is developed as a part of an information system designed for maintenance planning and decision-making support. Beside of description of the state model of equipment condition the article also present the way in which the system combined the pure modeled state with the information provided by the power plant technical staff. This provided information is either based on subjectively observed data or objectively measured data of the monitored equipment actual state of wear that is collected within the information system. Further it is important that the uncertainty of both the modeled state and of the provided information is taken into account. Therefore the state model is extended with means dealing with state credibility.

Keywords—equipment condition, steam turbine hall equipment, state estimation, state credibility

I. Introduction

THE problem of complex equipment maintenance planning and optimization, especially in environments with a large risk of potential failure consequences (e.g. aerospace components, chemical industry, nuclear energy, etc.) is being solved for long time [1], [2], [3]. Unfortunately, due to the low number of failures, long durability, equipment uniqueness and strong dependency on many other factors, the standard procedures based on the modeling of failure rate often provide unsatisfactory results. They often lead to too pessimistic prognosis and expensive maintenance activities as well.

Maintenance planning of a power plant equipment is an intricate task. It depends on the technical and legislative regulations which determine mandatory periodic maintenance. Moreover, it depends on the available information on the equipment condition that could make it possible to effectively allocate the available financial resources for preventive repairs with the aim to minimize the risk of equipment failures. The equipment state evolution and the associated risk of failures in the subsequent planning period contains the relevant information for decision-making of maintenance activities. For this reason it is necessary to have a suitable model that can provide this information in an understandable form.

It would be advantageous to have a supporting tool that would provide means for consolidation of all the information necessary for maintenance planning and for decision-making process. The resulting system must support the following tasks:

- gather all the constrains on maintenance given by either the manufacturer or by mandatory requirements given by the regulator,
- serve the technical staff as point where to store and retrieve the most accurate and actual information about equipment maintenance history,
- model the equipment state in order to be able to make outlook necessary for planing maintenance in the subsequent planing period,
- support the task of maintenance planing and decision-making.

Thus, such a information system would integrate all the necessary information needed for completing all the tasks necessary for effective maintenance of the power plant equipment.

The goal of this paper is to present crucial part of such information system (IS) that fulfills these tasks. It describes the way how to model the equipment condition. The presentation will cover three topics. First, the mathematical model of the equipment state in terms of wear will be introduced, where as a base of the model serves the model of the failures rate. Second, it is also essential to take into account the uncertainty of all the available information and thus the credibility of the modeled state is discussed. Finally, the employment of the information on the equipment condition within the developed IS will be presented.

The structure of the paper is as follows. Section II is devoted to a presentation of failure rate model that will be in the subsequent section used to determine the equipment condition. Then in Section IV the way of dealing with uncertainty in the state model will be described. The Section V shows how to use the information about the state of wear and its credibility is used for categorization of the state of the equipment form the risk point of view. The last but one section finally present how the presented state model of equipment condition is employed in the IS.

II. Power Plant Equipment Failure Rate Model

Equipment condition state model is directly derived from the normalized failure rate model and depends on the target lifetime. Therefore, it is first necessary to discuss the failure rate model.

The failure rate model is determined on the basis of statistical evidence evaluation of equipment failures during real operation. Statistical evaluation usually identifies the failure rate given the number of events per time unit or inversely in the mean time between failures [4], [5]. For the purpose of the statistical evaluation it is logical to use the data records about equipment failures from general operation. However, the number of such records is very low for the engine room

This work was supported by the European Regional Development Found (ERDF), project NTIS New Technologies for the Information Society, European Centre of Excellence, CZ.1.05/1.1.00/02.0090.

L. Jirsová and M. Flídr are with the European Research Centre of Excellence NTIS - New Technology for Information Society, University of West Bohemia, Pilsen, 30614 Czech Republic, e-mail: lenty@ntis.zcu.cz.

ISBN: 978-1-61804-251-4

equipment of a power plant. These data are further affected by errors caused by various operation modes and various repair and maintenance strategies. Therefore, the failure rate cannot be evaluated strictly using the statistical methods and it is necessary to make some major adjustments. For the correct function of the developed IS few basic precautions concerning the failure rate model are made.

- A failure rate histogram is determined based on the data gathered in the time period between two consecutive overhauls performed on each component or group of components that are taken into account.
- These statistical calculations are further augmented by components age reduction recalculation based on knowledge of number of failures, intervals between overhauls and target lifetime.
- It is necessary to determine the components endpoints of the failure rate evolution over time (i.e. the end of life) on the basis of expert opinion on the maximum lifetime expectation with nad without taking into account the carrying out the planned maintenance.
- After that it is possible to interpolate the histogram points by exponential or Bi-Weibull (or Tri-Weibull) probability density function resulting in the so called bathtub curve. That task can be performed optimizing the function parameters using least squares.

More thorough description of this procedure can be found in [6].

The use of the Bi-Weibull distribution for description of the failure rate is overly practical due to possibly high computational demands that has be taken into account when developing the IS. On the other hand, the Bi-Weibull distribution has an advantage, that the corresponding probability density function is a smooth function. Thus, the function can be sufficiently accurately approximated by another function, which is simpler in terms of description and calculations. As a suitable approximation the piecewise monotone Hermite cubic spline function can be used. This polynomial function can be easily used to preserve the original shape of the bathtub curve using only few interpolation points. First, the resampling of the bathtub function representing Bi-Weibull probability density function is performed for each component. Second, it is necessary to cope with the fact that the function diverges to infinity in zero. However, this behavior does not reflect reality and it is advisable to omit the initial segment of the bathtub function. Subsequently, it is necessary to normalize the modified failure rate function so that the area under the function is equal to one and also the time scale of the function will be normalized. Then it is possible to approximate the function by piecewise monotone Hermite cubic spline function.

In order to proceed with the approximation it is now important to select N time points $x_k \in < 0, 100\% >$ with $x_1 = 0$ and $x_N = 100\%$. The points should be selected so as the function values y_k represent the failure rate function in most accurate way. Then the piecewise monotone Hermite cubic spline function denoted as $P_k(x)$ for $k = 1, \ldots, N-1$ is for the on the domain $x \in < x_k, x_{k+1} >$ defined by the following relation

$$P_k(x) = \frac{3h_k q_k^2 - 2q_k^3}{h_k^3} y_{k+1} + \frac{h_k^3 - 3h_k q_k^2 + 2q_k^3}{h_k^3} y_k + \frac{q_k^2(q_k - h_k)}{h_k^2} d_{k+1} + \frac{q_k(q_k - h)^2}{h_k^2} d_k, \tag{1}$$

where the following denomination is used

$$q_k = x - x_k, \tag{2}$$

$$h_k = x_{k+1} - x_k \tag{3}$$

with d_k representing the first derivation of $P_k(x)$ with respect to x_k. The value of the polynomial function $P_k(x)$ on the domain $x \in < x_k, x_{k+1} >$ can also be expressed in a more compact manner

$$P(x) = y_k + p_k d_k + p_k^2 c_k + p_k^3 b_k, \tag{4}$$

where are the coefficients of quadratic and cubic term in the polynomial are given by relations

$$c_k = \frac{3\delta_k - 2d_k - d_{k+1}}{h_k}, \tag{5}$$

$$b_k = \frac{d_k - 2\delta_k + d_{k+1}}{h_k^2}, \tag{6}$$

with denoting δ_k by relation

$$\delta_k = \frac{y_{k+1} - y_k}{h_k}. \tag{7}$$

It should be noted that for the purposes of the IS are those coefficients d_k, c_k a b_k calculated for the selected points (x_k,y_k) beforehand and the evaluation of such approximated failure rate function is obtained by simple polynomial interpolation.

An example of such failure rate function given by Bi-Weibull probability density function ans its approximation using the described cubic spline function is shown in Figure 1.

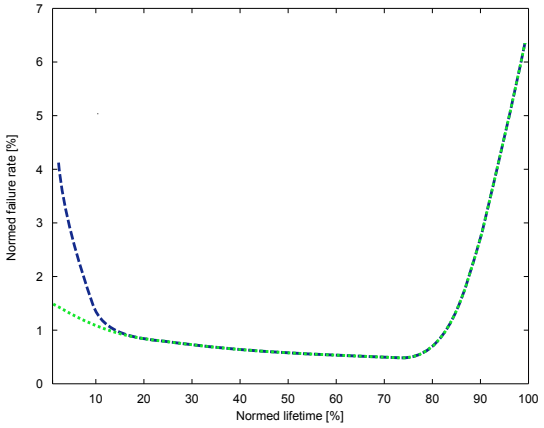


Fig. 1. The failure rate function (dashed line) and its approximation using piecewise monotone Hermite cubic spline function (dotted line)

ISBN: 978-1-61804-251-4

III. The Model of the State in Terms of the Wear

This section will present the state model of the equipment condition. The model is tightly coupled with the model of failures which is described by means of the failure function. The domain of this function is normalized so it is possible to use the term relative lifetime which takes values $0 - 100\%$. As it was previously mentioned this function is normalized so it must hold that

$$\int_0^{100} \lambda(x)dx = 1 \tag{8}$$

with lambda denoting the approximated failure function, i.e. given as composition of polynomial $P_k(x) \forall k$.

The basic idea is that the equipment at the time when it is put into operation is not worn and therefore has 100% state. The value of equipment state reduces over time. When the equipment will attain 100% of lifetime, the value of state is equal to zero.

If we consider that the new (not worn) equipment has 100% condition, then the state of wear is given as

$$s_M(x) = 100 - \int_0^x \lambda(\tau)d\tau \tag{9}$$

If it is necessary to express a condition change in term of wear after a certain time period, which is recalculated to the relative lifetime segment expressed by interval $x \in < x_t, x_{t-1} >$ then it is possible to express the state in the form

$$s_M(x_t) = s_M(x_{t-1}) - \int_{x_{t-1}}^{x_t} \lambda(\tau)d\tau \tag{10}$$

or alternatively in the form of t-variant linear differential equation

$$s_M(c_t) = a_{t,t-1} s_M(x_{t-1}), \tag{11}$$

where

$$a_{t,t-1} = \frac{s_M(x_{t-1}) - \int_{x_{t-1}}^{x_t} \lambda(\tau)d\tau}{s_M(x_{t-1})}. \tag{12}$$

In both cases the following initial condition is considered

$$s_M(0) = 100\%. \tag{13}$$

In the Figure 2 an example of the condition evolution is presented that corresponds to the failure rate model according to Figure 1 and that is obtained by an application of the relation (9).

IV. Credibility of State Estimation Determination

In the above model that describes the state in terms of failure rate, the state is function of the relative time. The relative time is given by the target lifetime and by the length of the modeled equipment operation. Both of these values are not precisely known and must be estimated. The condition value inevitably depends on the estimate of the random variable, the relative lifetime. Thus, it is always important to accompany the information about the equipment condition together with its credibility.

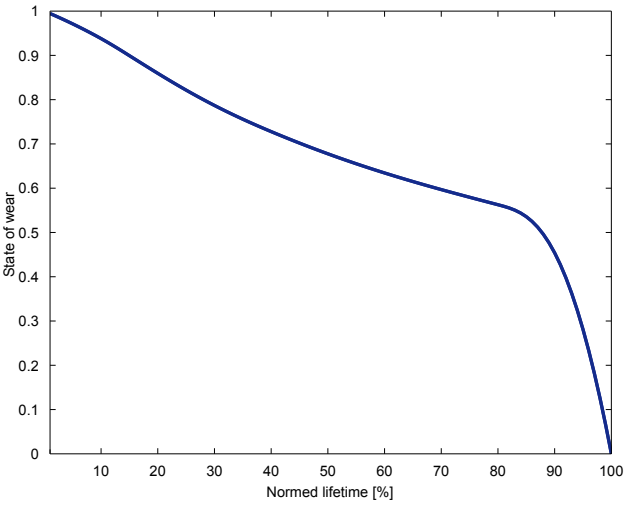


Fig. 2. The evolution of the equipment condition

Generally the credibility of the equipment condition evolves in time. More specifically, if the only information about the equipment condition was only a dynamic model, then the credibility should be reduced over time. Because the power plant is a system composed of a large number of equipment it can not be realistically assumed that the technical service staff personnel will update the state of wear on regular basis. This could after some time result in completely unrealistic information about the state without any credibility. For this reason, IS assumed that the state of wear credibility is piecewise constant. The change to the credibility of the state may occur only on the basis of information submitted by the user, i.e the technical staff personnel. The user provides this information to IS on the basis of objective and subjective findings, e.g. after inspection or repair of the equipment. The state of wear is then corrected accordingly and also its credibility is updated.

The state of wear determined only on the basis of the model is essentially only an estimate of the actual equipment condition. This is the reason why users are able to directly specify the value of the state and its credibility within the IS. The IS distinguishes between the data provided by the user and the data provided by the model, i.e. the user state s_U and its credibility v_U and model state s_M and its credibility v_M, respectively. Based on these to distinguished states and their credibilities the corrected state s_C and its credibility v_C are determined. The following relation is used for calculation of the corrected state

$$s_C = s_M \cdot \frac{v_U}{v_M + v_U} + s_U \cdot \frac{v_M}{v_M + v_U}. \tag{14}$$

and it credibility is finally determines as

$$v_C = \begin{cases} v_M + (v_U - v_M) \cdot \dfrac{v_M}{v_M + v_U}, & v_M \leq v_U, \\ v_U + (v_M - v_U) \cdot \dfrac{v_U}{v_M + v_U}, & v_M > v_U. \end{cases} \tag{15}$$

This corrected state s_C and its credibility then replaces the current modeled state s_M for all subsequent calculation until new correction after another user data update.

ISBN: 978-1-61804-251-4

It should be noted that this correction changes the equipment condition so it actually moves the relative lifetime of the equipment. Thus is influences the relative age of the equipment. E.g in particular case of an overhaul the technical staff provides the user state that could mean that the equipment is refurbished. In such case this would result in higher value of the equipment state (i.e. the equipment is less worn) and thus the longer actual lifetime.

V. The Categorization of the State

In addition to the state in terms of wear and its credibility the IS should present to users the risk category to which the selected equipment can be assigned. This categorization should help in deciding when to schedule maintenance.

It is assumed that for the categorization of the equipment it will suffice to distinguish the following four categories of the state

- running-in,
- normal operation,
- critical condition,
- end of life.

To what category the given state belongs is determined by the state limits, which are set individually by the type of equipment. The value ranges of defining categories of state are sharply defined by non-overlapping intervals.

The state can be generally associated into more than one category. The degree of membership in individual categories is expressed as a percentage. The sum of membership degrees of all classification classes must be equal to one hundred percent. The degree of membership is determined as follows.

1) On the basis of the corrected state value s_C and is credibility v_C will the following interval boundaries be determined as follows.
   - If the following condition is valid
   $$50 - \frac{v_C}{2} \leq s_C \leq 50 + \frac{v_K}{2}, \tag{16}$$
   then the interval is defined as
   $$< s_C + \frac{v_C}{2} - 50, s_C - \frac{v_C}{2} + 50 > . \tag{17}$$
   - If $s_C < 50 - \frac{v_C}{2}$, then the interval $< 0, 100 - v_C >$ is used.
   - In case if the condition $s_C > 50 + \frac{v_C}{2}$ holds, then the interval $< v_C, 100 >$ is used.
2) The degree of membership to individual categories of state is determined finding all the intersections of the above defined interval with intervals defining the categories. The degree of membership is given by the percentage by which the relevant intersection contributes to the whole interval defining the category.

VI. Information System for Decision-Making Support

All procedures described above are implemented in developed IS. The IS is built as the independent comprehensive methodological and technical solutions for the systematic relevant technical and economic data acquisition. The IS consists of several parts including the module for data collection, data processing, evaluation and long-term maintenance planning.

Computationally the most important part is module for data processing and evaluation, that perform an operations for the state of the wear model obtaining and its actualization. The module includes tools for displaying and statistical evaluation of the data in the form of comparative reports and graphs presented by location, logical components, specific models, technical jobs, types of equipment etc. In terms of the system function the most important is the evaluation of the risk analysis of individual technological equipment.

Most important for user is the module for long term planning that follows up this module for evaluation. It provides functionality to support the creation of a maintenance plan for a planning period (usually a year). The resulting plan includes three types of maintenance, which take into account all relevant aspects for maintenance planning, i.e. ongoing maintenance representing the periodic interventions, one-time maintenance or occasional maintenance planned for higher technology units maintenance, where the higher level of risk is expected.

The resulting decision making data are designed in the form of risk matrix that shows for each selected equipment the current condition. It clearly shows the overall risks and costs associated with different variants of the maintenance strategy at the current equipment state.

VII. Conclusion

The reliability oriented approach to state modeling of complex technological equipment condition is presented in the paper. This approach is implemented in IS for maintenance planning and decision-making support that is also presented. The IS stores and retrieves the most accurate and actual information about equipment maintenance history and model the equipment condition in order to be able to make outlook necessary for planing maintenance in the subsequent planing period. IS is deployed and verified in real in Czech power plant.

References

[1] P. L. Saldanha, E. A. de Simone, and P. F. e Melo, "An application of non-homogeneous poisson point processes to the reliability analysis of service water pumps," *Nuclear Engineering and Design*, vol. 210, no. 1-3, pp. 125–133, 2001.

[2] J. lu Sheng, Z. Lui, F. hui Xing, and D. mei Zhang, "Optimization of maintenance strategy for marine generators," in *Electrical Insulation Conference and Electrical Manufacturing Expo, 2007*, Oct 2007, pp. 72–75.

[3] A. C. Marquez and A. S. Heguedas, "Models for maintenance optimization: a study for repairable systems and finite time periods," *Reliability Engineering & System Safety*, vol. 75, no. 3, pp. 367–377, 2002.

[4] B. Šedivá, E. Wagnerová, F. Vávra, T. Ťoupal, and P.Marek, "Statistical monitoring of failures - methods and use," *Proceedings of the 11th International Scientific Conference Electric Power Engineering 2010*, pp. 611–615, 2010.

[5] L. Houdova, L. Houdova, L. Jelinek, and E. Janecek, "Approach to solving the task of availability prediction and cost optimization of a steam turbine," *Proceedings of the International Conference on Information Technology Interfaces, ITI*, pp. 629–634, 2010.

[6] L. Jirsová and L.Jelínek, "Modeling of unreliability and condition evolution of engine room equipment with respect to maintenance and overhaul effect," *MMMAS 2014*, 2014, submitted for publiccation.

ISBN: 978-1-61804-251-4

An economic and mathematical approach to determining key product quality parameters when placing a state defense order

E.S. ARTEMENKO
Engineering Economic Institute
Saint Petersburg State Polytechnic University
195251, Saint Petersburg, ul. Politekhnicheskaya 29
RUSSIA
Evg_art@mail.ru

Abstract: The necessity to determine key quality parameters of military products as the most important constituent for calculating their fair price is substantiated. Advantages of using the fuzzy sets theory apparatus when taking economic managerial solutions have been considered. A possible approach to determining key product quality parameters when placing a state defense order using the method of intersection of convex fuzzy sets is suggested. The main areas of application of the suggested approach for defense contractors have been highlighted.

Key-Words: Fuzzy sets theory. Expert assessment methods. State defense order. Key quality parameters of military products. Method of intersection of convex fuzzy sets. Optimization of a production program.

1 Introduction

Currently the necessity to work out a price forming mechanism for high-tech military products that must encourage military contractors not only to manufacture time-proven specimens of armaments but also to increase R&D costs, develop and create specimens of military machinery having no analogs abroad and, in addition, to reduce their production costs comprehensively is exceedingly relevant.

The existing approach to price formation does not encourage implementation with military contractors of organizational, engineering, scientific, and other innovations facilitating reduction of production costs and improvement of quality of products.

2 Problem Formulation

It is necessary to work out a new approach to efficient price formation the key objective of which must be to determine the fair price of military products by correlating their price and the product quality parameters key to the customer.

The substance of the approach consists in the unity of two aspects: the first one reflects the interests of the customer (state) in terms of using budgetary funds effectively from the military economic perspective, the second one reflects the interests of the contractor and consists in an economic appeal of the order.

Efficient price formation when placing a state defense order implies the meeting of the following obligatory principles:

- a created specimen of military products must have operational and physical characteristics not lower than the specified ones;
- an order must be fulfilled within the times specified by the contract and in required quantities.

Optimization of budgetary funds should be attained by choosing an option of creating military products ensuring achievement of the required effect with minimal expenses of financial resources along its entire life cycle – considering maintenance and repair costs during the entire life, disposal costs after writing-off, insurance costs and possible raising of additional debt funds.

3 Problem Solution

To solve a problem of determining key product quality parameters when placing a state defense order, it is deemed possible to use an assumption of vague perception of partners' requirements to product quality, i.e. of presentation of a preference given by the customer to this or that product as a convex fuzzy set.

Subjectivity of human perception was not subject to mathematical description for long. Mathematics is an exact science, whereas man can perceive this or

that mathematical dependency differently, treating it on the basis of one's own concepts.

All processes implemented or controlled by man should be referred to "fuzzy", "indistinct", or "blurred" processes, whereas man possesses, apart from an ability to reason and ratiocinate, an ability to take into account considerations of both general and associated nature at the same time.

A problem of uniting general consideration and logical reasoning was solved with an appearance of the fuzzy sets theory suggested by a Prof. of Berkley University (California, USA) L. Zadeh in the 1960s [1]. The fuzzy sets theory enabled operation of a mathematically fuzzy presentation of concepts having qualitative and subjective characteristics. Academic schools were established based on works by L. Zadeh, R. Bellman, R. Yager [2;3]. Works by domestic scientists D.A. Pospelov, A.P. Ryzhov, S.A. Orlovsky [4;5;6] had a noticeable effect on the scientific research in the sphere of fuzzy sets.

Unlike conventional mathematics requiring at each step simulation of exact and unambiguous formulation of regularities, fuzzy logic is at an entirely different level of thought due to which only a minimal set of regularities is required to be determined in the process of simulation. In the limit, when preciseness grows, fuzzy logic comes to standard, Boolean logic [7].

Values obtained as a result of fuzzy measurements are in many things similar to distributions of the probabilities theory but free from the drawbacks characteristic of them, such as a small number of analyzable distribution functions, the necessity of their forced normalization, meeting additivity requirements, difficult substantiation of adequacy of mathematical abstraction to describe behavior of actual values. Compared to a probabilistic method, a fuzzy method allows the volume of the performed calculations to be greatly reduced, which in its turn leads to an increased speed of fuzzy systems. The drawbacks of fuzzy systems should include:

- absence of standard methodology of their design;
- impossibility of mathematical analysis of fuzzy systems by respective methods;
- disadvantage in the precision of calculations compared to more conventional (such as probabilistic) methods with no influence in the process being simulated of any subjective factors.

Nevertheless, intuitive simplicity of fuzzy logic assures its successful application in various systems of control and analysis of economic information, and the set-theoretical approach allows political and economic variables of the market for high-tech military products to be considered.

The methods of the theory of fuzzy sets and structures lie at the intersection of mathematical simulation methods (whereas a formal mathematical apparatus is used) and expert assessment methods (whereas a membership function is built with the use of the latter in the fuzzy sets theory).

A factor dictating the necessity to use expert assessments in the practice of determining the fair price when placing a state defense order is that in the current economic realities a decision maker often has no data necessary to it to the full extent and relations between them, i.e. acts under the conditions of incomplete and unclear information.

Using expert methods as tools of scientific prevision becomes necessary under the conditions when one has to operate indices not directly measurable quantitatively, which is for instance fair when determining the membership function values. On the other hand, unavoidable errors of the expert assessment method imbedded at the initial stage of model building are streamlined in the course of calculations made on the basis of the theory of fuzzy sets and structures. Thus, the methods of the fuzzy sets theory turn out rather efficient when determining key product quality parameters when placing a state defense order.

The following basic classification signs of ways to formalize fuzziness are singled out:

- by type of presenting a fuzzy subjective assessment of any value (fuzzy set);
- by type of range of values of a membership function;
- by type of membership function domain;
- by type of correlation between the domain and the range of values (single-valued, many-valued);
- by sign of uniformity or non-uniformity of the range of value of the membership function.

The fuzzy sets theory is applied when solving a wide range of practical problems in various spheres of science and practice such as: control of manufacturing processes, defense complexes, database creation, design of computer technologies, and, finally, decision making simulation.

To determine key product quality parameters when placing a state defense order, a method of intersection of convex fuzzy sets is applicable, modified for solving this problem within one commodity profile with a fixed set of commodity characteristics [8;9].

Let the key quality parameters mean a set of properties which knowingly satisfy the customer of

ISBN: 978-1-61804-251-4

high-tech military products, i.e. the state. Ensuring presence of similar properties warrants to a military contractor a stable demand for its products in combination with a minimal risk of refusal to conclude a contract, helps to understand the necessary directions of the technical evolution of its products, to determine the key technological competences, specialization and scales of activities in a specific market.

Let the following be assigned:

$X = \{x_1, x_2, \dots, x_n\}$ – a set of products in the product range of a military contractor, $l \in 1:n$;

$Y = \{y_1, y_2, \dots, y_p\}$ – a set of product quality parameters, $i \in 1:p$;

$Z = \{z_1, z_2, \dots, z_m\}$ – a set of product characteristics required by the state costumer, $j \in 1:m$.

It is required to determine a set of key product quality parameters when placing a state defense order. The model is built on the following assumptions:

- manufacturers and the consumer, respectively military contractors and the state, operate in the market;
- the products $x_1, x_2, \dots, x_n$ are characterized by quality parameters;
- quality parameter membership degrees $y_1, x_2, \dots, y_p$ of products vary between individual products $x_1, x_2, \dots, x_n$;
- one commodity is preferred to another each time when its quality characters Y are closer to the customer's requirements z_j.

For each type of high-tech military products it is appropriate to pick its unique set of quality parameters among there might be: «reliability», «maintenance costs», «repair costs over the entire service life», «costs of disposal after writing-off», «insurance costs», etc.

Let $\xi_R: X \times Y \to [0; 1]$ be a membership function of a fuzzy relation R determined with the help of an expert. The relation R is presented as a matrix where elements of each line $\xi_R(x_i; y_i)$ express relative quality parameter membership degrees y_i of a certain product x_i. The higher the value, the more important is the character.

Similarly the function $\psi_s: Y \times Z \to [0; 1]$ is presented, the membership function of a fuzzy relation S. For all $y \in S$ and all $z \in Z$ the value $\psi_s(y_i; z_j)$ is equal to the degree of compatibility of the requirement z_j of the state customer with the quality parameters of a specific product y_i. The values of the matrix S reflect relative degrees of importance of quality parameters y_i by a criterion z_j when the state takes a decision to purchase this or that batch of military products.

The matrices R and S yield the matrix T the elements of which $\mu_T(x_i; z_j)$ are determined by the membership function:

$$\mu_T(x_i; z_j) = \frac{\sum_y \xi_R(x_i; y) * \psi_s(y; z_j)}{\sum_y \xi_R(x_i; y)} \quad (1)$$

for all $x \in X, y \in Y, z \in Z$.

The total of $\sum_y \xi_R(x_i; y)$ is equal to a fuzzy set degree indicating the number of most important quality parameters y, inherent in the products x_i from the point of view of the state customer.

Then, a matrix of pairwise minimums is built:

$$L = \begin{pmatrix} \mu_T(x_1; z_1) \wedge \mu_T(x_1; z_2) & \dots & \mu_T(x_1; z_{m-1}) \wedge \mu_T(x_1; z_m) \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots & \dots & \dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \mu_T(x_n; z_1) \wedge \mu_T(x_n; z_2) & \dots & \mu_T(x_n; z_{m-1}) \wedge \mu_T(x_n; z_m) \end{pmatrix}$$

The threshold of division of quality parameters b is limited with a stipulation:

$$b < \min_j \max_i \min\left(\mu_T(x_i; z_j), \mu_T(x_i; z_{j+1})\right) \quad (2)$$

After the threshold b is chosen, one can determine for any z_j a level subset M_j:

$$M_j = \left\{x_i | \mu_T(x_i; z_j) \geq \min_j \max_i \min\left(\mu_T(x_i; z_j), \mu_T(x_i; z_{j+1})\right)\right\} \quad (3)$$

A set of key quality parameters of the products of an enterprise is described by uniting level subsets:

$$M = \bigcup_j M_j \quad (4)$$

4 Conclusion

Calculating key product quality parameters when placing a state defense order enables the military contractors to determine:

- how to optimize the commodity line – products with what characteristics are required to be in the line with the retention of the existing structure of the state defense order;
- how to modify the product line with the prescribed variation of the customer's requirements, i.e. what strategic actions are to be taken in case of a change in the current parameters of the state defense order;
- how to optimize a set of product quality parameters in case of exclusion from the production program of the products the quality parameters of which are not satisfactory for the customer, or inclusion of the commodities the parameters of which are suitable for it.

Results of this problem can be used when taking a decision to include in the production program of this or that product. It is required therefore, by having determined the membership function of the assumed product x_{n+1}, to make calculations according to the given algorithm and determine in which degree it

belongs to a set of key product quality parameters, and if it does, whether it will not drive out any products from the set $x_1, x_2, \dots, x_n$ which are already in the production program of the enterprise. Based on this assessment, a person responsible for formation of the production program of an enterprise can take a positive, wait-and-see, or negative decision.

References:

[7] 1.Batyrshin I.Z., *Basic operations of fuzzy logic and their generalizations,* Otechestvo, 2001.

[2] 2.Bellman R., Zadeh L.A. *Decision-making under fuzzy conditions,* Mir, 1976.

[6] 3.Orlovsky S.A., *Problems of decision making with fuzzy initial information*, Nauka, 1981.

[4] 4.Pospelov I.G., *Simulation of economic structures*, Computing Center of Russian Academy of Sciences, 2003.

[5] 5.Ryzhov A.P., *Elements of the fuzzy sets theory and fuzziness measurements,* Dialog-Moscow State University, 1998.

[8] 6.Saati T.L., *Decision making. Hierarchy analysis method*, Radio I Svyaz, 1993.

[9] 7.Vashchekin A.N., *Mathematical simulation of commercial activities of a wholesaler*, Moscow State University of Culture, 2002.

[3] 8.Yager R.N., *Fuzzy sets and the theory of possibilities*, Radio I Svyaz, 1986.

[1] 9.Zadeh L.A., *The concept of a linguistic variable and its application to approximate reasoning*, Mir, 1976.

ISBN: 978-1-61804-251-4

The procedure of image identification as a method of raising consumer demand

YAKOVLEV ANDREY ANATOLYEVICH
Economics and Management of Technologies and Landed Property
St.Petersburg State Polytechnical University
195251, St.Petersburg, Polytechnicheskaya, 29
RUSSIA
ckf@bk.ru

Abstract: The article presents an approach designed by the author to solving the problem of establishing an optimal set of consumer properties of a product that meet consumers' needs best, which would make it possible to determine its place on the market rationally in view of demands of the target audience.
The solution of the problem is based on using the theory of linear filtration and image identification, and it is development of a decision rule, which enables minimizing the number of errors of consumer demand. As a result, it is possible to assess in advance the expediency of offering the market a product with exclusive features that would enable the company to raise its profits.

Key-words: Consumer demand, image identification problem, attribute classifier, generation of attributes, selection of attributes.

1 Introduction

One of possible ways of increase in profitability of the company in the conditions of market process - creation of an exclusive product. This product should most, in comparison with products already deduced on the market, to consider, and probably and to advance, needs of the consumer.

Provide a consumer demand only the full account, and in some cases, even anticipatory formation of consumer expectations can. Possibility of the solution of a similar task assumes existence in structure of governing body of the organization, a specialized element capable to organize collecting, processing, classification and studying of statistical data necessary for adoption of the operating decision on creation of the interesting consumer of a product, and also sufficient computing capacities and perfect computing algorithm. In this article use possibility in these purposes of the mathematical procedure known as «recognition of images» is considered.

2 Problem Formulation

Image identification is a process of object classification by individual categories or classes. For example, according to the Classification of the European Economic Committee modern passenger cars are customarily classified by attributes that are oriented towards the target market segmentation rather than a description of any specific features of the cars. For the same cars that pass crash tests, EuroNCAP applies its own classification (generates its own system of attributes) in order to differentiate between car categories that may be compared with each other by the parameters that are important for passive safety, such as the dimensions and the weight and type of the body. In the USSR, the united industrial standard OH 025270-66 of 1966 was applied, according to which cars were divided into classes depending on the engine volume and dry weight. In North America, cars have been traditionally classified based on the wheelbase length and (lately) the net volume of the passenger compartment.

In the framework of any classification mentioned here, specific cars (objects) are called images. The classification is based on cases. A case is an object that has been classified earlier and is taken as a reference for solving classification problems.

In solving classification problems it is customarily assumed that all the objects (phenomena) are divided into a finite number of classes, for which a finite number of objects (cases) is known and has been researched. In particular, according to the classification of the European Economic Committee,

the market of passenger cars is divided into the following segments:

A: Mini cars
B: Small cars
C: Medium cars
D: Larger cars
E: Executive cars
F: Luxury cars
S: Sport coupés
M: Multi purpose cars
J: Sports utility (SUV)

The objective of image identification is attributing a new identifiable object to a certain class.

These segments are used by manufacturers to determine the car's position on the market, while specific car concepts within the same segment may have completely different features and use different technologies and sets of options depending on the manufacturer. The measurements that are used for image classification are called attributes. An attribute is a quantitative measurement of an object of any nature. An aggregate of attributes relating to the same image is an attribute vector. Attribute vectors assume values in the attribute space. Every manufacturer is free to choose the means for meeting the demands of the target audience in a specific market segment and therefore free to create (generate) their own classification that meets their needs best.

In the framework of the identification problem it is considered that every image corresponds with one value of the attribute vector and, vice versa, each value of the attribute vector corresponds to a single image.

A classifier (a decision rule) is a rule of attributing an image to one of the classes based on its attribute vector. Generation of attributes is a process of identification of attributes that describe the object most completely (determining the necessary and sufficient suit of attributes). Selection of attributes involves identification of attributes that have the best classification properties for the specific objective.

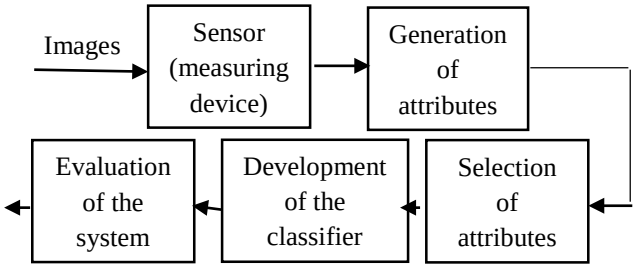


Fig. 1. The sequence of solving the attribute identification problem

2.1 Statement of the classification problem

The formal statement of the classification problem is based on the following assumptions:

Ω is the set of identification objects (the image space). $\omega : \varpi \in \Omega$ is the object of identification (the image). $g(\omega) : \Omega \to M$, $M\{1,2,\ldots,m\}$ is an indicator function that breaks the image space Ω into m non-overlapping classes $\Omega^1, \Omega^2, \ldots \Omega^m$. The indicator function is unknown to the observer.

X is the space of observations perceived by the observer (the attribute space).

$X(\omega) : \Omega \to X$ is a function that matches every object ω with a point $x(\omega)$ in the attribute space. The vector $x(\omega)$ is the image of the object that is perceived by the observer. In the attribute space, non-overlapping sets of points $K_i \subset X$, $i=1,2,\ldots,m$, which correspond to images of the same class, have been determined.

$\hat{g}(x) : X \to M$ is the decision rule, the evaluation for $g(\omega)$ is based on $x(\omega)$, i.e. $\hat{g}(x) = \hat{g}(x(\omega))$.

Let $x_j = x(\omega_j)$, $j=1,2,\ldots,N$ be the information about the functions $g(\omega)$ and $x(\omega)$ that is available to the observer; however, these functions themselves are unknown to the observer. Then (g_j, x_j), $j=1,2,\ldots,N$ is the set of cases.

The problem is developing the decision rule $\hat{g}(x)$, which enables identification with a minimum number of errors. In the event of a Euclidean space of attributes ($X=R^l$), the quality of the decision rule is measured by the frequency of generation of right decisions. It is evaluated by attributing some probability measure to the set of objects Ω. In this case, the problem is stated as $\min P\{\hat{g}(x(\omega)) \neq g(\omega)\}$. The practical goal of solving the problem is providing the best description of key trends of change of the target audience's needs.

In particular, as applied to the car market it is basically specification of parameters that determine the processes of providing automatic control of certain driving functions, raise fuel efficiency, and improve dynamics (control of the engine and transmission gear), active safety (control of the brake system), and comfort (control of suspension, etc.).

3 The stages of solving the image identification problem

1. There exists an ***Object*** that has a set of properties, which can be evaluated (measured) objectively using various sensors (measuring devices).

2. Generation of attributes based on linear transformations

ISBN: 978-1-61804-251-4

The purpose of such generation of attributes is reducing the information down to "significant" by transforming the initial set of measurements into a new set of attributes. Usually, the objective is isolation of components that contain the fundamental information.

Let $X \in R^m$ be the set of attributes, $Y \in R^l$, the set of attributes that must be chosen in the process of selection, while $l < m$. Then the selection problem is stated as follows: $X \to Y$.

3. Statement of the attribute selection problem

Let us set the vector of attributes $X \in R^m$. Among them one must select the most informative ones, i.e. obtain the new attribute vector $Y \in R^l$, while $l < m$. The idea of selecting attributes is isolation of attributes that lead to big distances between classes and small distances inside classes. The main motivation for reducing the number of attributes is reducing the computational complexity and increasing the generality of the classifier.

The selection of attributes is preceded by their background processing for the purpose of their scaling and performing additional improvements. The main operations of background processing are described by the following three items.

Removal of runouts, i.e. points that are located "very far away" from the average value. The distance is usually measured in average deviations, e.g. $2\sigma \sim 95\%$, $3\sigma \sim 99\%$ for a Gauss distribution.

Normalization. Attributes that have big values can influence the classifier stronger than others, which distorts the correctness of the classifier. Therefore, their influence must be reduced by way of normalization. Let x_i be the case and $x_i=(x_{i1},\ldots,x_{il})$ the attributes. Then

$$\bar{x}^{(k)} = \frac{1}{N}\sum_{i=1}^{N} x_{i_k}, \quad k = 1,2,\ldots,l$$

is the averaging of the attribute (basically, its mathematical expectation). Let us designate the scattering evaluation as

$$\left(\sigma^{(k)}\right)^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{i_k} - \bar{x}^{(k)})^2$$

Then the normalized attributes are set as follows:

$$\tilde{x}_{i_k} = \frac{x_{i_k} - \bar{x}^{(k)}}{\sigma^{(k)}}.$$

Data omissions (losses). For many cases some of the attributes may be unknown, and only the ones with the same suit of attributes may be selected. If attributes may not be discarded, they are complemented, e.g. using heuristics.

In particular, the method of selection based on verifying statistical hypotheses can be used. In such case, having obtained the values of the object attributes as implementation of random values in the course of measurement, we can find their distribution with methods of mathematical statistics. In the event of coincidence of the distribution for different classes, the attribute will not distinguish such classes; if the distributions are different, the attribute will distinguish them. Therefore, the problem of selection based on verifying statistical hypotheses is solved by evaluating the discriminant capacity of each individual attribute.

4. Construction of the classifier

Let Ω be the space of images, X the space of attributes, $g(\omega)$, $\omega \in \Omega$ the indicator function, and M the set of attributes. Then $g : \Omega \to M$.

Also, let $X = \langle x(\omega_i), g(\omega_i)\rangle$, $i=1,\ldots,N$ be the set of cases, and $\hat{g}(x)$ the decision rule.

Then $\hat{g} : X \to M$.

The selection of the decision rule is based on the minimization $d(g, \hat{g}) \to min$ where d is the metric, the measure of proximity of functions $g(\omega)$ and $\hat{g}(x(\omega))$. The construction of $\hat{g}$ is called the problem of teaching. The procedure of establishment is the teacher, and the cases are the teaching sequence.

5. Evaluation of the system (quality of training of the classifier)

The relative share of disagreements of the classification with the teacher for the decision rule is $= \frac{m}{N}$, where $m=|\{\omega_i : g(\omega_i) \neq \hat{g}(x(\omega_i)), i=1,2,\ldots,N\}|$. The reliability of teaching of the classifier is the probability of obtaining the decision rule with a preset quality.

Let $f(x,\alpha)$ be the class of discriminant functions where $\alpha \in A$ is the parameter. The number of degrees of freedom in the selection of a certain function in the class is determined by the number of parameters in the vector α, i.e. the dimension of A. The classifier's capacity of dividing is increased with the increase of the number of degrees of freedom.

The probability model

Let the cases be the result of implementation of random values. Let us consider the value of risk (i.e., error) related to the classification. Les us determine the notions of the average risk (mathematical expectation of the function of losses) and empirical risk (the average value of the error at the teaching selection).

ISBN: 978-1-61804-251-4

Let us assume that the algebra σ and measure P are set in Ω. Also, let x be the attribute vector, $\tilde{f}$ is the class of functions, from which the decision rule is selected, $f(x,\alpha)$ the decision rule (the result of classification), which assumes the value of 0 or 1 with a fixed parameter vector, χ the characteristic function of the set, and A the set of parameters describing various functions in $\tilde{f}$.
Then $\hat{g}=f(x,\alpha)$, where $f\in\tilde{f}$ и $f:X\times A\to M$, $y=g(\omega)$.
In these designations, the average risk is expressed as follows:

$$K(\alpha)=\int_X \chi\{y\neq f(x,\alpha)\}dP.$$

For the cases of two classes, when $M=\{0,1\}$, we have:

$$K(\alpha)=\int_\Omega (y-f(x,\alpha))^2 dP$$

$$K(\alpha)=\int_{(X,M)} (y-f(x,\alpha))^2 dP(x,y),$$

where dP is the probability measure on the space X.

The problem of finding the best classifier
Let us consider the minimization of the composite function: $K(\alpha)\to\min_{\alpha\in A}$.
The problem of finding the best classifier is about finding such an α^* that $K(\alpha^*)=\min_{\alpha\in A} K(\alpha)$.
If a minimum does not exist, one must find such an α^* that $\left|K(\alpha^*)-\inf_{\alpha\in A} K(\alpha)\right|<\delta$.
In other words, one must solve the problem of average risk minimization.
The problem of empirical risk minimization is stated as follows: $K_{emp}(\alpha)\to\min_{\alpha\in A}$,
where we minimize random values using the parameter α, which is any possible parameter.
Ideally, mutually related evaluations of the empirical and average risk must be obtained.
It must be noted that the less is l, the easier it is to construct $f(x,\alpha)$ so that $K_{emp}(\alpha)$ is equal or very close to zero. However, the true value of $K(\alpha)$ in such case may be very different from $K_{emp}(\alpha)$. One must choose such an $f(x,\alpha)$ that there would be uniform convergence by α of the expression:

$$P\left\{\sup_\alpha\left|K_{emp}(\alpha)-K(\alpha)\right|>\varepsilon\right\}\xrightarrow[l\to\infty]{}0.$$

In fact, it is the convergence of frequencies to the mathematical expectation.

6. Ultimately, a suit of classification attributes is established, which make it possible to distinguish various products by their consumer properties with confidence. As it is known, if products become impossible to distinguish by their design features, it means that their properties have become part of a set of standard characteristics for that product type and they are now in the consumer product category.

In practice, continuous solving of the classification problem makes it possible to promptly identify the moment when products with new unique and competitive attributes appear on the market, such attributes being capable of determining the development trends of the consumer demand of the target audience in every consumer segment, by way of rational generation of attributes.

A possible example is the one described in *Motor Trend Magazine* for 1986 which declared Ford Taurus the car of the year, describing the smallest design features such as the coffee cup holder.

The appearance of this classification attribute celebrated a new stage of the competitive battle about the design of coffee cup holders, which was ultimately won by Lexus that applied their 'commitment to perfection' approach in their design. As a result, the holder was decorated with walnut and had a hydraulic drive, forced blockage, and a soft rubber clamp that fit any cup size.

4 Conclusion

Therefore, this paper illustrates the process of solving the image identification problem, which enables determining the set of consumer properties of a product that meet consumers' needs best on the basis of development of an optimal attribute classifier.

As a result, it is possible to assess in advance the expediency of offering the market a product with exclusive features that would enable the company to raise its profits.

References:

1. Ventsel Y.S.. The probability theory: A textbook for higher education institutions. – 6th publ. ster. – M.: Vyssh. shk., 1999. - 576 p.
2. Glukhov V.V., Mednikov M.D., Korobko S.B.. Mathematical methods and models for management. 2nd publ., edited and supplemented – SPb.: 2005. - 528 p.
3. Milocevich D. M60 Program Management for Improved Business Results / Dragan Z. Milosevic; Translated from English by Y.V. Mamontov; edited by S.I. Neizvestny. - M.: Company IT; DMK Press, 2008. - 729 p.

Improvement of Strategic and Operational Efficiency of Clusters Based on Enterprise Architecture Model

IGOR V. ILIN, ALEKSEI B. ANISIFOROV
Institute of Industrial Economics and Management
Saint-Petersburg State Polytechnical University
195251, Saint-Petersburg, Politechnicheskaya, 29
RUSSIA
ilyin@fem.spbstu.ru http://www.spbstu.ru/

Abstract: The paper gives an analysis of the impact of architecture model of industrial cluster development on the growth of its strategic and operational efficiency through synergy of architecture.

Key-Words: Industrial cluster, cluster architecture, enterprise architecture, business architecture, IT architecture, synergistic effect, development projects, architecture model, architecture synergy.

1 Introduction

The main idea of the cluster concept in the industrial sector is the creation of cooperative ties between its members. They allow organizations to increase the efficiency of operations, deploy new technologies and products earlier than others, providing innovative development of all participants.

A peculiar kind of combination of within the boundaries of the cluster competition and cooperation ensure effective use of resources, innovation, simplifies management of innovative projects, increases the efficiency of production and business activities and promotes the formation of a synergistic effect [1]. Building a system of information and knowledge sharing with the use of modern means of communication, it contributes to the development of innovative and productive capacity and exchange of business models. The information infrastructure, based on modern methods and models of information systems development, data, and knowledge exchange, is particularly important for development and growth of the cluster. This infrastructure should support the organizational and economic mechanism of cluster management and mathematical methods and models, on which it relies. Construction and development of such infrastructure is only possible on the basis of an architecture approach.

Cluster architecture is an essential tool of organizational changes in the cluster, and these changes affect both the processes of knowledge management and reengineering activities. It includes two core elements: business architecture and IT architecture, which are closely interrelated.

2 Problems of assessing strategic and operational efficiency of the cluster

2.1 Architectural approach to the development of the cluster

IT- and business architecture also act as an important tool of strategic goals fulfillment and dealing with operational tasks through adequate information infrastructure of the cluster.

Business architecture identifies key assets associated with the information required by the business, describes the functionality required to implement the logic execution and optimization of business processes, affecting the information infrastructure management and support service issues. IT architecture through a set of its elements ensures the implementation (execution) of business processes. Logical models of IT services allow you to choose the specific technology, and application architecture can cope with the complexity of the using information resources and information systems by cluster members.

From the standpoint of GIGA GROUP [2] architecture development requires significant investments in "standards, processes, technologies and interfaces in order to improve the capability of the organization and reduce the cost of development and maintenance of information systems."

When it comes to industrial clusters, it is necessary to consider both the cluster architecture in general, and the architecture of individual enterprises and organizations which are members of the cluster.

Thus, the cluster architecture defines the general structure of its business and IT, including partners and participants and forms a model of managing its

ISBN: 978-1-61804-251-4

activities to achieve strategic goals and deal with current operational tasks, also forming a common business vision and ensuring interoperability and integration where necessary. An architecture approach allows companies to react quickly to economic, political, financial changes, conducting appropriate optimization of business models and information infrastructure to ensure the growth of its operating efficiency and innovation.

Business model is a set of interrelated strategic decisions that form the way how the company is doing business, which determines how the creation and assignment of values within the firm value network happens. It should explain how the cluster members generate the income flow [3]. Herewith this value proposition is not an abstract process, and aims to meet the needs of certain customers of the specific market segment. The process of creating value [4] is implemented within the framework of the value chain by cluster members. Building of a cluster architecture model will require several projects, forming a model of its development, determining funding sources, sequence of the geographic expansion of its activities that can significantly affect the economic model of the cluster, also affecting the structure of costs and sources of income of the organization. Selected characteristics of cluster architecture influence the results of its operations. A relationship between this model and the performance indicators of the company has been established in several studies [5].

Thus, cluster development and economic results of its operations have to be determined by the quality of its business model and the support of it, i.e. its architecture. Given the diversity of effects that make up the overall efficiency of the cluster, the most important of them can be selected - the *innovation efficiency*, *investment efficiency* and *synergistic effect of cluster activity*.

The architecture model of cluster development allows agreeing actions of the members and staff to achieve the main cluster goal. Its basic principle, which in many ways was the reason for the high efficiency of this model: **only those can be managed what can be described and measured** [6].

2.2 Synergistic effect of an industrial cluster

To assess the strategic and operational efficiency of cluster activities not only on the general efficiency performance indicators should be used, but also indicators of industrial, financial and investment efficiency. Indicators of fiscal efficiency, taking into account the social and financial implications of its activities for the budgets of all levels, are also important for this assessment.

Cluster as a sustainable partnership has the potential exceeding the simple sum of individual components of the potentials. This increment occurs as a result of cooperation, the combination of cooperation and competition and effective use of partners' possibilities. The architecture approach in company management is also essential for increase of this potential. In fact, we can speak about a certain synergetic effect of clusters.

I. Ansoff identifies five types of synergy: sales synergy, operational synergy, investment synergy, management synergy, information synergy [7]. Almost all of these kinds of synergies pose a consequence of the use of unified information infrastructure, information technology of data and knowledge exchange, integration of information resources, the use of common management models, business process integration, and econometric models. I.e. they base themselves on the cluster architecture.

Apart from above the cluster architecture allows building an integrated business model of the cluster, which will provide the support not only of management processes, but also cluster development processes. Therefore, in authors' opinion, it is viable to speak not about information synergy, but the architecture synergy.

Application of general points of synergetic economy to analysis of the clusters efficiency suggests that the important synergistic effects of the cluster are not limited to:

- effect of knowledge sharing in the cluster;
- effect of infrastructure sharing;
- effect of incremental cash flow due to the addition of cash flows of companies in the cluster;
- effect of reducing transaction costs [8],

but also include the effect of the implementation of architecture model of cluster development.

Achieving all of these effects is assured by architecture model, allowing cluster members to respond to the economic situation while making management decisions quickly and adequately.

ISBN: 978-1-61804-251-4

Building the architecture of cluster allows maintaining technological, informational, organizational and financial aspects of operational management, as well as management of investment and innovation projects.

Architecture model of cluster development allows the growth of strategic and operational efficiency and leads to the occurrence of synergy.

3 Main directions of evaluation of cluster activity efficiency based on enterprise architecture model

3.1 Evaluating the efficiency of operational activities

We can distinguish two main areas of cluster activity evaluation: the efficiency of operational activities and efficiency of innovation activities and investment.

While evaluating operational activities the following performance indicators are used: profitability, turnover, business activity, financial stability and others. A cluster form of production, based on an architecture model, contributes to the improvement of financial aspects of cluster members, especially in terms of cash flow management [9]. Cash flow is the movement of money into or out of a business, project, or financial product. The form of cash flow is:

$$CF = CG - CP \quad (1)$$

where CG - cash inflow;

CP - cash outflow;

CF - net cash flows.

The clusters synergetic effect is determined by the fact that the sum of the cash flows of individual enterprises will be less than the total cash flow of the cluster. Such an effect in a formalized manner looks as follows:

$$CF_c > \sum CF \quad (2)$$

where $\sum CF$ – amount of cash flows of enterprises in the cluster;

CF_c - the balance of the cash flows of the cluster.

The cluster has a significant impact on the size of inflow (*CG*) and outflow (*CP*) funds. Increase in cash flow in the cluster structures is aligned with the fact that the aggregate demand for the products in the conditions of cluster existence is much higher, because cluster contributes to the formation of so-called "determinants of demand" [10].

In any case, when evaluating the efficiency we have to rely on the basic concepts of financial management: the concept of cash flow, the concept of changes of the monetary unit value at different times, the concept of cost of capital, the concept of trade-off between risk and return, the concept of the cost of missed opportunities. In addition, the architecture model of cluster management can significantly reduce the cost of information infrastructure maintenance.

3.2 Evaluating the efficiency of development projects

3.2.1 Evaluating the efficiency of investment projects

Evaluating the efficiency of development projects is based on standard criteria for investment analysis, based on discounting: the net present value of the project (*NPV*); internal rate of return of the project (*IRR*); payback period based on discounted estimates (*DPP*); project profitability index (*PI*). Methods for calculating these indices are widely known, but the indicators themselves, with different nature, may contradict each other, i.e. a project adopted in accordance with one criterion may be not with regard to economic considerations of another criterion [11]. In addition, some factors allow obtaining absolute evaluation calculations, but others give relative calculations.

Because of the specificity of innovation projects, not all of the classical criteria can be used to analyze the efficiency of financing innovation. Analysis of numerous studies on the problem of evaluating the efficiency of innovations has shown that in most cases the authors' attention is focused on methodologies for assessing the economic efficiency of investment projects [12]. However, despite the common methodology for assessing the cost-efficiency of such projects, innovative projects have a number of specific characteristics that must be considered in the assessment of their efficiency. Most often, implementation of innovations is aimed at significant changes of indicators of economic activity of cluster or individual structures, which requires significant investments.

Some experts, such as O.N. Zemskova [13], offer other indicators of investment analysis, which are of some interest, because consider the fact that some costly innovative projects in the cluster can not always be financed from its own funds.

These indicators often include the debt capital and its costs; these are: the degree of internal

financing of the upfront investment and the annualized net present value (*ANPV*), as well as *IRR* and *DPP*, calculated on the basis of *ANPV*.

The upfront investment required at the initial stage of the innovation project is the most risky and less liquid one. In this regard, the share of project initiator in the upfront investment acts for other investors as an indicator of seriousness of his intentions and validity assessment of prospects for the development of commercial innovation. The second indicator is used to control the absence of the shortage of available funds and allows concluding about the financial viability of the project.

The net present value of the project is calculated for the whole period of its implementation, based on the annual net flows of the project, excluding its financing scheme. Traditionally, the current index value of the project is counted according to the following formula:

$$NPV = \sum_{t=1}^{T}(R_t - Z_t)\frac{1}{(1+r)} \quad (3)$$

where R_t - cash inflow in year *t*;

Z_t - cash outflow year *t*;

r - the discount rate;

T - project life in years.

The project is deemed efficient if *NPV > 0*.

This approach to the calculation of NPV does not account the likelihood of financing deficit and debt financing. For the control of financing deficit the *ANPV* indicator can be used. In contrast to the traditional *NPV*, calculated on the basis of cash flows from investing and current activities, the annualized net present value is based on three types of flows: investment, current and fiscal. So it is necessary to have in the project management documentation a plan for its financing, including the credit plan.

Annualized net present value is calculated by the formula:

$$ANPV = -(F_{int} + D) + \sum_{t=1}^{T}\frac{CF_t^a}{(1+r)^t} \quad (4)$$

where F_{int} – planned financing investments in the project from its internal funds in the starting year;

D – debt attracted to cover the deficit in upfront investments;

CF_t^a– annualized net cash flows in the year *t*, calculated not only on the basis of project cash flows, but also on the cost of servicing loans taken to finance the project.

If *NPV > ANPV* (with the proviso that *NPV* > 0), project may be called efficient; moreover, the efficiency can be increased in real terms of debt financing. Opposite inequality shows that debt funding may negatively affect the efficiency of the project. If both indicators are equal, the efficiency of the project does not depend on the funding scheme.

Possible differences between *NPV* and *ANPV* are associated with characteristics of the cash flow. In the initial stages of most projects cash flow is negative due to the lack of return, moreover the debt capital requires maintenance costs. As a result, the net cash flow at the end of the project is less than the cash flow into account when calculating *NPV*. Thus, the indicator *ANPV* may be a reliable tool for evaluating the effectiveness of debt funds.

3.2.2 Evaluating the efficiency of innovative projects

Analysis of the continuous innovation activities in the cluster leads to the conclusion that within the same time period innovative projects are at various stages of implementation and may involve several participants. In this case, the use of traditional methods of project analysis becomes impossible or insufficient - both positive and negative cash flows of individual projects and existing operations can mutually neutralize each other and distort the overall evaluation of the efficiency of using debt and internal resources.

It is therefore necessary, as rightly said by V.I. Barilenko [14] to distinguish the concept of "analysis of innovation projects" and "analysis of innovation activity during the reporting period". Many of the characteristics of innovation activity allow using them for analysis of indicators and standard analytical procedures used in the traditional analysis of economic activities, such as evaluating the efficiency of using debt and internal funds.

Features of the innovation activity of industrial cluster and its funding processes are so large that it requires the development and use of special techniques and indicators. Innovation development of modern industrial clusters, requiring significant costs of raising and maintenance of necessary funding, makes use of particular importance to assess the effectiveness of innovation indicators of economic value added *EVA*.

The essence of the *EVA* concept is logically linked to the objectives of the analysis of debt funds in innovation activity as it is that the whole company is regarded as a kind of investment project with an initial capital, which requires the involvement of certain costs. The difference between yields subjected to innovation and the cost of capital invested in it, determines the amount of economic value added.

M.P. Apin offers to calculate the economic value in different ways [15]:

ISBN: 978-1-61804-251-4

$$EVA = NP - (WACC * IC) = \left(\frac{NP}{IC} - WACC\right) * IC = (ROI - WACC) * IC \quad (5)$$

where *IC* - invested in venture capital;
WACC - weighted average cost of capital;
NP - net profit;
ROI - return on invested capital.

One of the ways to evaluate the market value of the company is the addition of the net assets of the balance sheet and the amount of *EVA* given to this point in time. In accordance to this the market value of the enterprise may exceed the carrying value of net assets or be lower than it, depending on the size of future amount of *EVA*.

This implies three possible alternatives of an average weighted price of the invested capital in innovative enterprise and its profitability, and as consequence in the *EVA* index:

1) *EVA* = 0 for *ROI* = *WACC*. Return on invested capital is equal to the cost of raising this capital and the market value of the company's net assets is identical to simply balance. In such a situation there is no market gain on investments in innovation development of the enterprise;

2) *EVA* > 0 for *ROI* > *WACC*. Excess margin investment cost of financing means increase in the market value of the enterprise over the carrying value of net assets at the expense of innovation. This situation stimulates additional investment in innovation development;

3) *EVA* < 0 for *ROI* < *WACC*. *EVA* negative value indicates a decrease in the market value of the enterprise, as the cost of capital is greater than its benefits. In this case, the owners lose their invested funds, which encourages them to make a decision about moving their capital to another, more efficient project.

In the context of existing industrial cluster the main way to increase *EVA* is the development and implementation of effective innovation projects. Since the implementation of such projects, as a rule, is based on debt financing, it is logical to propose a methodology to evaluate the efficiency of the debt funding using the relative index of debt payoff. The proposed indicator is advisable to rely on the base of created as a result of innovation enterprise economic value added, as it is *EVA* that reflects the value created over the cost of raising all the capital invested.

Profitability of innovation can not be ensured by increasing expenses on their implementation - it usually means a simple waste of resources for increasingly obsolescent projects. According to experts [16], the way out can be found not in increasing the expenses, but improving the efficiency of base costs. This will increase the return on investment in innovation, improving the *ROI* indicator.

4 Conclusion

Availability of the performance criteria of cluster development projects helps fund rationally and spend the available resources efficiently, as well as to provide high quality management decisions and their information support in a constantly changing economic environment within the cluster architecture.

Any innovation and investment project, in all economic significance must contain either social effect, for example, more jobs, clean production, etc. Therefore, forming a project portfolio and verification of its investment support, should be based on a balanced business model of the cluster.

Thus, an architecture approach to cluster management allows an integral representation of its strategy, business processes of the cluster members and the IT architecture, corresponding to strategic goals and tactical objectives.

References:

[1] Bychkova, G.M., The justification for applying a synergistic approach to the assessment of the efficiency of the cluster, *Izvestiya ITEA*, No. 6 (62), 2008
[2] rpp/nashauceba.ru
[3] Mahadevan, B., Business models for Internet-based e-commerce, *California Management Review*, Vol.42, Issue 4, 2000
[4] Zott, C., Amit, R., Business model design and the performance of entrepreneurial firms, *Organization Science*, Vol.18, 2007, pp. 181–199
[5] Malone, T.W., Weill, P., Lai, R.K., D'Urso, V.T., Herman, G., Apel, T.G., and Woerner, S.L. Do Some Business Models Perform Better than Others? *MIT Sloan Working Paper*, 2006
[6] Anisiforov, A.B., Ilin, I.V., Silkina, G.Y., Yurev, V.N., *Innovative development of industrial cluster*, Izd-vo Politechn. Un-ta, 2012
[7] Ansoff, I., *New corporative strategy*, Piter-Press, 1999
[8] Gutova, A.V., Cash Flow Management: theoretical aspects, *Financial Management*, No.4, 2004
[9] Khasanov, R.Kh., Economic difficulties of regional and industrial clusters, *Problems of modern economics*, No. 3 (31), 2009
[10] Porter, M., *On competition*, Harvard Business School, 1998

[11] Goldstein, G.Y., *Strategic innovation management*, Izd-vo TRTU, 2004
[12] Tumina, T.A., Methodology to assess the effectiveness of innovation, *Transportation in Russia*, 2009
[13] Zemskova, O.N., Analysis of the effectiveness of debt financing innovation enterprise, *Economics*, 2010, pp. 391-396
[14] Barilenko, V.I., Analytical assessment of the effectiveness of innovation enterprise, *Regional economics*, No. 42, 2009
[15] Vassilyev, I.A., *Analysis of effectiveness of innovation projects*, VEDI, 2001
[16] Semenova, T.Y., Innovative programs and projects in the system of regional development, *Problems of the modern economy. Eurasian International scientific-analytical journal*, No. 3 (23), 2007, pp. 523-528

Authors Index

ISBN: 978-1-61804-251-4