ADVANCES in APPLIED and PURE MATHEMATICS

Proceedings of the 2014 International Conference on Pure Mathematics, Applied Mathematics, Computational Methods (PMAMCM 2014)

Santorini Island, Greece
July 17-21, 2014

ADVANCES in APPLIED and PURE MATHEMATICS

Proceedings of the 2014 International Conference on Pure Mathematics, Applied Mathematics, Computational Methods (PMAMCM 2014)

Santorini Island, Greece
July 17-21, 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 | 29

ISSN: 2227-4588
ISBN: 978-1-61804-240-8

ADVANCES in APPLIED and PURE MATHEMATICS

Proceedings of the 2014 International Conference on Pure Mathematics, Applied Mathematics, Computational Methods (PMAMCM 2014)

Santorini Island, Greece
July 17-21, 2014

Organizing Committee

General Chairs (EDITORS)

- Prof. Nikos E. Mastorakis
  Industrial Eng.Department
  Technical University of Sofia,
  Bulgaria
- Prof. Panos M. Pardalos,
  Distinguished Prof. of Industrial and
  Systems Engineering,
  University of Florida, USA
- Professor Ravi P. Agarwal
  Department of Mathematics
  Texas A&M University – Kingsville
  700 University Blvd.
  Kingsville, TX 78363-8202, USA
- Prof. Ljubiša Kočinac, University of Nis,
  Nis, Serbia

Senior Program Chair

- Prof. Valery Y. Glizer,
  Ort Braude College, Karmiel, Israel

Program Chairs

- Prof. Filippo Neri
  Dipartimento di Informatica e Sistemistica
  University of Naples "Federico II"
  Naples, Italy
- 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. Pradip Majumdar
  Department of Mechanical Engineering
  Northern Illinois University
  Dekalb, Illinois, USA

Special Session Chair

- Prof. Pavel Varacha
  Tomas Bata University in Zlin
  Faculty of Applied Informatics
  Department of Informatics and Artificial Intelligence
  Zlin, Czech Republic

Workshops Chair

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

Local Organizing Chair

- Prof. Klimis Ntalianis,
  Tech. Educ. Inst. of Athens (TEI), Athens, Greece

Publication Chair

- Prof. Gen Qi Xu
  Department of Mathematics
  Tianjin University
  Tianjin, China

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. Stefan Siegmund, Technische Universitaet Dresden, Germany
- Prof. Zoran Bojkovic, Univ. of Belgrade, Serbia
- Prof. Metin Demiralp, Istanbul Technical University, Turkey
- Prof. Imre Rudas, Obuda University, Budapest, Hungary

Program Committee for PURE MATHEMATICS

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. Nasser-Eddine Mohamed Ali Tatar, King Fahd University of Petroleum and Mineral, 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. 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
Prof. Chun-Gang Zhu, Dalian University of Technology, Dalian, China
Prof. Mohamed Kamal Aouf, Mansoura University, Mansoura City, Egypt
Prof. Yansheng Liu, Shandong Normal University, Jinan, Shandong, China
Prof. Naseer Shahzad, King Abdulaziz University, Jeddah, Saudi Arabia
Prof. Janusz Brzdek, Pedagogical University of Cracow, Poland
Prof. Mohammad T. Darvishi, Razi University, Kermanshah, Iran
Prof. Ahmed El-Sayed, Alexandria University, Alexandria, Egypt

Program Committee for APPLIED MATHEMATICS and COMPUTATIONAL METHODS

Prof. Martin Bohner, Missouri University of Science and Technology, Rolla, Missouri, USA
Prof. Martin Schechter, University of California, Irvine, USA
Prof. Ivan G. Avramidi, New Mexico Tech, Socorro, New Mexico, USA
Prof. Michel Chipot, University of Zurich, Zurich, Switzerland
Prof. Xiaodong Yan, University of Connecticut, Connecticut USA
Prof. Ravi P. Agarwal, Texas A&M University - Kingsville, Kingsville, TX, USA
Prof. Yushun Wang, Nanjing Normal university, Nanjing, China
Prof. Detlev Buchholz, Universitaet Goettingen, Goettingen, Germany
Prof. Patricia J. Y. Wong, Nanyang Technological University, Singapore
Prof. Andrei Korobeinikov, Centre de Recerca Matematica, Barcelona, Spain
Prof. Jim Zhu, Western Michigan University, Kalamazoo, MI, USA
Prof. Ferhan M. Atici, Department of Mathematics, Western Kentucky University, USA
Prof. Gerd Teschke, Institute for Computational Mathematics in Science and Technology, Germany
Prof. Meirong Zhang, Tsinghua University, Beijing, China
Prof. Lucio Boccardo, Universita degli Studi di Roma "La Sapienza", Roma, Italy
Prof. Shanhe Wu, Longyan University, Longyan, Fujian, China
Prof. Natig M. Atakishiyev, National Autonomous University of Mexico, Mexico
Prof. Jianming Zhan, Hubei University for Nationalities, Enshi, Hubei Province, China
Prof. Narcisa C. Apreutesei, Technical University of Iasi, Iasi, Romania
Prof. Chun-Gang Zhu, Dalian University of Technology, Dalian, China
Prof. Abdelghani Bellouquid, University Cadi Ayyad, Morocco
Prof. Jinde Cao, Southeast University/ King Abdulaziz University, China
Prof. Josef Diblik, Brno University of Technology, Brno, Czech Republic
Prof. Jianqing Chen, Fujian Normal University, Fuzhou, Fujian, China
Prof. Naseer Shahzad, King Abdulaziz University, Jeddah, Saudi Arabia
Prof. Sining Zheng, Dalian University of Technology, Dalian, China
Prof. Leszek Gasinski, Uniwersytet Jagielloński, Krakowie, Poland
Prof. Satit Saejung, Khon Kaen University, Muang District, Khon Kaen, Thailand
Prof. Juan J. Trujillo, Universidad de La Laguna, La Laguna, Tenerife, Spain
Prof. Tiecheng Xia, Department of Mathematics, Shanghai University, China
Prof. Stevo Stevic, Mathematical Institute Serbian Academy of Sciences and Arts, Beograd, Serbia
Prof. Lucas Jodar, Universitat Politecnica de Valencia, Valencia, Spain
Prof. Noemi Wolanski, Universidad de Buenos Aires, Buenos Aires, Argentina
Prof. Zhenya Yan, Chinese Academy of Sciences, Beijing, China
Prof. Juan Carlos Cortes Lopez, Universidad Politecnica de Valencia, Spain
Prof. Wei-Shih Du, National Kaohsiung Normal University, Kaohsiung City, Taiwan
Prof. Kailash C. Patidar, University of the Western Cape, Cape Town, South Africa
Prof. Hossein Jafari, University of Mazandaran, Babolsar, Iran
Prof. Abdel-Maksoud A Soliman, Suez Canal University, Egypt
Prof. Janusz Brzdek, Pedagogical University of Cracow, Cracow, Poland
Dr. Fasma Diele, Italian National Research Council (C.N.R.), Bari, Italy

Additional Reviewers

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

Table of Contents

ISBN: 978-1-61804-240-8

ISBN: 978-1-61804-240-8

Keynote Lecture 1

New Developments in Clifford Fourier Transforms

Sen. Ass. Prof. Dr. rer. nat. Eckhard Hitzer
Department of Material Science
International Christian University
Tokyo, Japan
E-mail: hitzer@icu.ac.jp

Abstract: We show how real and complex Fourier transforms are extended to W.R. Hamilton's algebra of quaternions and to W.K. Clifford's geometric algebras. This was initially motivated by applications in nuclear magnetic resonance and electric engineering. Followed by an ever wider range of applications in color image and signal processing. Clifford's geometric algebras are complete algebras, algebraically encoding a vector space and all its subspace elements, including Grassmannians (a vector space and all its subspaces of given dimension k). Applications include electromagnetism, and the processing of images, color images, vector field and climate data. Further developments of Clifford Fourier Transforms include operator exponential representations, and extensions to wider classes of integral transforms, like Clifford algebra versions of linear canonical transforms and wavelets.

Brief Biography of the Speaker: http://erkenntnis.icu.ac.jp/

ISBN: 978-1-61804-240-8

Keynote Lecture 2

Robust Adaptive Control of Linear Infinite Dimensional Symmetric Hyperbolic Systems with Application to Quantum Information Systems

Prof. Mark J. Balas
Distinguished Faculty
Aerospace Engineering Department &
Electrical Engineering Department
Embry-Riddle Aeronautical University
Daytona Beach, Florida
USA
E-mail: balasm@erau.edu

Abstract: Symmetric Hyperbolic Systems of partial differential equations describe many physical phenomena such as wave behavior, electromagnetic fields, and quantum fields. To illustrate the utility of the adaptive control law, we apply the results to control of symmetric hyperbolic systems with coercive boundary conditions.
Given a Symmetric Hyperbolic continuous-time infinite-dimensional plant on a Hilbert space and disturbances of known and unknown waveform, we show that there exists a stabilizing direct model reference adaptive control law with certain disturbance rejection and robustness properties. The closed loop system is shown to be exponentially convergent to a neighborhood with radius proportional to bounds on the size of the disturbance. The plant is described by a closed densely defined linear operator that generates a continuous semigroup of bounded operators on the Hilbert space of states. We will discuss the need and use of this kind of direct adaptive control in quantum information systems.

Brief Biography of the Speaker: Mark Balas is presently distinguished faculty in Aerospace Engineering at Embry-Riddle Aeronautical University. He was the Guthrie Nicholson Professor of Electrical Engineering and Head of the Electrical and Computer Engineering Department at the University of Wyoming. He has the following technical degrees: PhD in Mathematics, MS Electrical Engineering, MA Mathematics, and BS Electrical Engineering. He has held various positions in industry, academia, and government. Among his careers, he has been a university professor for over 35 years with RPI, MIT, University of Colorado-Boulder, and University of Wyoming, and has mentored 42 doctoral students. He has over 300 publications in archive journals, refereed conference proceedings and technical book chapters. He has been visiting faculty with the Institute for Quantum Information and the Control and Dynamics Division at the California Institute of Technology, the US Air Force Research Laboratory-Kirtland AFB, the NASA-Jet Propulsion Laboratory, the NASA Ames Research Center, and was the Associate Director of the University of Wyoming Wind Energy Research Center and adjunct faculty with the School of Energy Resources. He is a life fellow of the AIAA and a life fellow of the IEEE.

ISBN: 978-1-61804-240-8

Keynote Lecture 3

Multidimensional Optimization Methods with Fewer Steps Than the Dimension: A Case of "Insider Trading" in Chemical Physics

Prof. Paul G. Mezey
Canada Research Chair in Scientific Modeling and Simulation
Department of Chemistry and Department of Physics and Physical Oceanography
Memorial University of Newfoundland
Canada
E-mail: pmezey@mun.ca

Abstract: "Insider trading" in commerce takes advantage of information that is not commonly available, and a somewhat similar advantage plays a role in some specific, very high-dimensional optimization problems of chemical physics, in particular, molecular quantum mechanics. Using a specific application of the Variational Theorem for the expectation value of molecular Hamiltonians, an optimization problem of thousands of unknowns does often converge in fewer than hundred steps. The search for optimum, however, is typically starting from highly specific initial choices for the values of these unknowns, where the conditions imposed by physics, not formally included in the optimization algorithms, are taken into account in an implicit way. This rapid convergence also provides compatible choices for "hybrid optimization strategies", such as those applied in macromolecular quantum chemistry [1]. The efficiency of these approaches, although highly specific for the given problems, nevertheless, provides motivation for a similar, implicit use of side conditions for a better choice of approximate initial values of the unknowns to be determined.

[1]. P.G. Mezey, "On the Inherited "Purity" of Certain Extrapolated Density Matrices", Computational and Theoretical Chemistry, 1003, 130-133 (2013).

Brief Biography of the Speaker: http://www.mun.ca/research/explore/chairs/mezey.php

ISBN: 978-1-61804-240-8

Keynote Lecture 4

MvStudium_Group: A Family of Tools for Modeling and Simulation of Complex Dynamical Systems

Professor Yuri B. Senichenkov
co-author: Professor Yu. B. Kolesov
Distributed Computing and Networking Department
St. Petersburg State Polytechnical University
Russia
E-mail: sen@dcn.icc.spbstu.ru

Abstract: Designing of new version of Rand Model Designing under the name RMD 7 is coming to an end. It will be possible using dynamic objects, dynamic connections (bonds), and arrays of objects in the new version. These types are used for Simulation Modeling, and Agent Based Modeling. The first trial version will be available at year-end.
The tools developed by MvStudium_Group are considered by authors as universal tools for automation modeling and simulation of complex dynamical systems. We are feeling strongly that at least nitty-gritty real system is multi-component, hierarchical, and event-driven system. Modeling of such systems requires using object-oriented technologies, expressive graphical languages and various mathematical models for event-driven systems. The last versions of Model Vision Modeling Language are intended for multi-component models with variable structure and event-driven behavior.

Brief Biography of the Speaker: PhD degree in Numerical Analysis from St. Petersburg State University (1984).
Dr. Sci. degree (Computer Science) from St. Petersburg Polytechnic University (2005).
Author of 125 scientific publications-conference papers, articles, monographs and textbooks.
A board member of National Simulation Society - NSS (http://simulation.su/en.html), and Federation of European Simulation Societies- EuroSim (http://www.eurosim.info/).
A member of Scientific Editorial Board of "Simulation Notes Europe" Journal (http://www.sne-journal.org/), and "Computer Tools in Education" Journal(http://ipo.spb.ru/journal/).
Chairman and Chief-Editor of COMOD 2001-2014 conferences (https://dcn.icc.spbstu.ru/).

ISBN: 978-1-61804-240-8

ISBN: 978-1-61804-240-8

New Developments in Clifford Fourier Transforms

Eckhard Hitzer

Abstract—We show how real and complex Fourier transforms are extended to W.R. Hamiltons algebra of quaternions and to W.K. Clifford's geometric algebras. This was initially motivated by applications in nuclear magnetic resonance and electric engineering. Followed by an ever wider range of applications in color image and signal processing. Cliffords geometric algebras are complete algebras, algebraically encoding a vector space and all its subspace elements. Applications include electromagnetism, and the processing of images, color images, vector field and climate data. Further developments of Clifford Fourier Transforms include operator exponential representations, and extensions to wider classes of integral transforms, like Clifford algebra versions of linear canonical transforms and wavelets.

Keywords—Fourier transform, Clifford algebra, geometric algebra, quaternions, signal processing, linear canonical transform

I. Introduction

We begin by introducing Clifford Fourier transforms, including the important class of quaternion Fourier transforms mainly along the lines of [31] and [5], adding further detail, emphasize and new developments.

There is the alternative operator exponential Clifford Fourier transform (CFT) approach, mainly pursued by the Clifford Analysis Group at the university of Ghent (Belgium) [5]. New work in this direction closely related to the roots of -1 approach explained below is in preparation [11].

We mainly provide an overview of research based on the holistic investigation [28] of real geometric square roots of -1 in Clifford algebras $Cl(p,q)$ over real vector spaces $\mathbb{R}^{p,q}$. These algebras include real and complex numbers, quaternions, Pauli- and Dirac algebra, space time algebra, spinor algebra, Lie algebras, conformal geometric algebra and many more. The resulting CFTs are therefore perfectly tailored to work on functions valued in these algebras. In general the continuous manifolds of $\sqrt{-1}$ in $Cl(p,q)$ consist of several conjugacy classes and their connected components. Simple examples are shown in Fig. 1.

A CFT analyzes scalar, vector and multivector signals in terms of sine and cosine waves with multivector coefficients. Basically, the imaginary unit $i \in \mathbb{C}$ in the transformation kernel $e^{i}\phi = \cos\phi + i\sin\phi$ is replaced by a $\sqrt{-1}$ in $Cl(p,q)$. This produces a host of CFTs, an incomplete brief overview is sketched in Fig. 2, see also the historical overview in [5]. Additionally the $\sqrt{-1}$ in $Cl(p,q)$ allow to construct further types of integral transformations, notably Clifford wavelets [21], [37].

E. Hitzer is with the Department of Material Science, International Christian University, Mitaka, Tokyo, 181-8585 Japan e-mail: hitzer@icu.ac.jp.

thanks

Manuscript received May 31, 2014; revised ...

II. Clifford's geometric algebra

Definition 1 (Clifford's geometric algebra [15], [36])**.** *Let* $\{e_1, e_2, \ldots, e_p, e_{p+1}, \ldots, e_n\}$*, with* $n = p + q$*,* $e_k^2 = \varepsilon_k$*,* $\varepsilon_k = +1$ *for* $k = 1, \ldots, p$*,* $\varepsilon_k = -1$ *for* $k = p+1, \ldots, n$*, be an orthonormal base of the inner product vector space* $\mathbb{R}^{p,q}$ *with a geometric product according to the multiplication rules*

$$e_k e_l + e_l e_k = 2\varepsilon_k \delta_{k,l}, \qquad k, l = 1, \ldots n, \tag{1}$$

where $\delta_{k,l}$ *is the Kronecker symbol with* $\delta_{k,l} = 1$ *for* $k = l$*, and* $\delta_{k,l} = 0$ *for* $k \neq l$*. This non-commutative product and the additional axiom of associativity generate the* 2^n*-dimensional Clifford geometric algebra* $Cl(p,q) = Cl(\mathbb{R}^{p,q}) = Cl_{p,q} = \mathcal{G}_{p,q} = \mathbb{R}_{p,q}$ *over* $\mathbb{R}$*. The set* $\{e_A : A \subseteq \{1, \ldots, n\}\}$ *with* $e_A = e_{h_1} e_{h_2} \ldots e_{h_k}$*,* $1 \leq h_1 < \ldots < h_k \leq n$*,* $e_\emptyset = 1$*, forms a graded (blade) basis of* $Cl(p,q)$*. The grades* k *range from* 0 *for scalars,* 1 *for vectors,* 2 *for bivectors,* s *for* s*-vectors, up to* n *for pseudoscalars. The vector space* $\mathbb{R}^{p,q}$ *is included in* $Cl(p,q)$ *as the subset of 1-vectors. The general elements of* $Cl(p,q)$ *are real linear combinations of basis blades* e_A*, called Clifford numbers, multivectors or hypercomplex numbers.*

In general $\langle A \rangle_k$ denotes the grade k part of $A \in Cl(p,q)$. The parts of grade 0 and $k+s$, respectively, of the geometric product of a k-vector $A_k \in Cl(p,q)$ with an s-vector $B_s \in Cl(p,q)$

$$A_k * B_s := \langle A_k B_s \rangle_0, \qquad A_k \wedge B_s := \langle A_k B_s \rangle_{k+s}, \tag{2}$$

are called *scalar product* and *outer product*, respectively.

For Euclidean vector spaces $(n = p)$ we use $\mathbb{R}^n = \mathbb{R}^{n,0}$ and $Cl(n) = Cl(n,0)$. Every k-vector B that can be written as the outer product $B = \boldsymbol{b}_1 \wedge \boldsymbol{b}_2 \wedge \ldots \wedge \boldsymbol{b}_k$ of k vectors $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_k \in \mathbb{R}^{p,q}$ is called a *simple* k-vector or *blade*.

Multivectors $M \in Cl(p,q)$ have k-vector parts $(0 \leq k \leq n)$: scalar part $Sc(M) = \langle M \rangle = \langle M \rangle_0 = M_0 \in \mathbb{R}$, vector part $\langle M \rangle_1 \in \mathbb{R}^{p,q}$, bi-vector part $\langle M \rangle_2$, ..., and pseudoscalar part $\langle M \rangle_n \in \bigwedge^n \mathbb{R}^{p,q}$

$$M = \sum_A M_A \boldsymbol{e}_A = \langle M \rangle + \langle M \rangle_1 + \langle M \rangle_2 + \ldots + \langle M \rangle_n \,. \tag{3}$$

The *principal reverse* of $M \in Cl(p,q)$ defined as

$$\widetilde{M} = \sum_{k=0}^{n} (-1)^{\frac{k(k-1)}{2}} \langle \overline{M} \rangle_k, \tag{4}$$

often replaces complex conjugation and quaternion conjugation. Taking the *reverse* is equivalent to reversing the order of products of basis vectors in the basis blades e_A. The operation $\overline{M}$ means to change in the basis decomposition of M the sign of every vector of negative square $\overline{e_A} = \varepsilon_{h_1} e_{h_1} \varepsilon_{h_2} e_{h_2} \ldots \varepsilon_{h_k} e_{h_k}$, $1 \leq h_1 < \ldots < h_k \leq n$. Reversion, $\overline{M}$, and principal reversion are all involutions.

ISBN: 978-1-61804-240-8

For $M, N \in Cl(p,q)$ we get $M * \widetilde{N} = \sum_A M_A N_A$. Two multivectors $M, N \in Cl(p,q)$ are *orthogonal* if and only if $M * \widetilde{N} = 0$. The modulus $|M|$ of a multivector $M \in Cl(p,q)$ is defined as

$$|M|^2 = M * \widetilde{M} = \sum_A M_A^2. \tag{5}$$

A. Multivector signal functions

A multivector valued function $f : \mathbb{R}^{p,q} \to Cl(p,q)$, has 2^n blade components $(f_A : \mathbb{R}^{p,q} \to \mathbb{R})$

$$f(\boldsymbol{x}) = \sum_A f_A(\boldsymbol{x}) \boldsymbol{e}_A. \tag{6}$$

We define the *inner product* of two functions $f, g : \mathbb{R}^{p,q} \to Cl(p,q)$ by

$$\begin{aligned}(f,g) &= \int_{\mathbb{R}^{p,q}} f(\boldsymbol{x}) \widetilde{g(\boldsymbol{x})}\, d^n\boldsymbol{x} \\ &= \sum_{A,B} \boldsymbol{e}_A \widetilde{\boldsymbol{e}_B} \int_{\mathbb{R}^{p,q}} f_A(\boldsymbol{x}) g_B(\boldsymbol{x})\, d^n\boldsymbol{x},\end{aligned} \tag{7}$$

with the *symmetric scalar part*

$$\langle f, g \rangle = \int_{\mathbb{R}^{p,q}} f(\boldsymbol{x}) * \widetilde{g(\boldsymbol{x})}\, d^n\boldsymbol{x} = \sum_A \int_{\mathbb{R}^{p,q}} f_A(\boldsymbol{x}) g_A(\boldsymbol{x})\, d^n\boldsymbol{x}, \tag{8}$$

and the $L^2(\mathbb{R}^{p,q}; Cl(p,q))$*-norm*

$$\|f\|^2 = \langle (f,f) \rangle = \int_{\mathbb{R}^{p,q}} |f(\boldsymbol{x})|^2 d^n\boldsymbol{x} = \sum_A \int_{\mathbb{R}^{p,q}} f_A^2(\boldsymbol{x})\, d^n\boldsymbol{x}, \tag{9}$$

$$L^2(\mathbb{R}^{p,q}; Cl(p,q)) = \{f : \mathbb{R}^{p,q} \to Cl(p,q) \mid \|f\| < \infty\}. \tag{10}$$

B. Square roots of -1 in Clifford algebras

Every Clifford algebra $Cl(p,q)$, $s_8 = (p-q) \bmod 8$, is isomorphic to one of the following (square) matrix algebras[1] $\mathcal{M}(2d, \mathbb{R})$, $\mathcal{M}(d, \mathbb{H})$, $\mathcal{M}(2d, \mathbb{R}^2)$, $\mathcal{M}(d, \mathbb{H}^2)$ or $\mathcal{M}(2d, \mathbb{C})$. The first argument of $\mathcal{M}$ is the dimension, the second the associated ring[2] $\mathbb{R}$ for $s_8 = 0, 2$, $\mathbb{R}^2$ for $s_8 = 1$, $\mathbb{C}$ for $s_8 = 3, 7$, $\mathbb{H}$ for $s_8 = 4, 6$, and $\mathbb{H}^2$ for $s_8 = 5$. For even n: $d = 2^{(n-2)/2}$, for odd n: $d = 2^{(n-3)/2}$.

It has been shown [27], [28] that $Sc(f) = 0$ for every square root of $-\mathbf{1}$ in every matrix algebra $\mathcal{A}$ isomorphic to $Cl(p,q)$. One can distinguish *ordinary* square roots of $-\mathbf{1}$, and *exceptional* ones. All square roots of $-\mathbf{1}$ in $Cl(p,q)$ can be computed using the package CLIFFORD for Maple [1], [3], [29], [38].

In all cases the *ordinary* square roots f of $-\mathbf{1}$ constitute a *unique conjugacy class* of dimension $\dim(\mathcal{A})/2$, which has *as many connected components as the group* $\mathbb{G}(\mathcal{A})$ of invertible elements in $\mathcal{A}$. Furthermore, we have $\mathrm{Spec}(f) = 0$ (zero pseudoscalar part) if the associated ring is $\mathbb{R}^2$, $\mathbb{H}^2$, or $\mathbb{C}$. The exceptional square roots of $-\mathbf{1}$ *only* exist if $\mathcal{A} \cong \mathcal{M}(2d, \mathbb{C})$.

For $\mathcal{A} = \mathcal{M}(2d, \mathbb{R})$, the centralizer (set of all elements in $Cl(p,q)$ commuting with f) and the conjugacy class of a square root f of $-\mathbf{1}$ both have $\mathbb{R}$-dimension $2d^2$ with *two connected components*. For the simplest case $d = 1$ we have the algebra $Cl(2,0)$ isomorphic to $\mathcal{M}(2, \mathbb{R})$.

For $\mathcal{A} = \mathcal{M}(2d, \mathbb{R}^2) = \mathcal{M}(2d, \mathbb{R}) \times \mathcal{M}(2d, \mathbb{R})$, the square roots of $(-\mathbf{1}, -\mathbf{1})$ are pairs of two square roots of $-\mathbf{1}$ in $\mathcal{M}(2d, \mathbb{R})$. They constitute a unique conjugacy class with *four connected components*, each of dimension $4d^2$. Regarding the four connected components, the group of inner automorphisms $\mathrm{Inn}(\mathcal{A})$ induces the permutations of the Klein group, whereas the quotient group $\mathrm{Aut}(\mathcal{A})/\mathrm{Inn}(\mathcal{A})$ is isomorphic to the group of isometries of a Euclidean square in 2D. The simplest example with $d = 1$ is $Cl(2,1)$ isomorphic to $M(2, \mathbb{R}^2) = \mathcal{M}(2, \mathbb{R}) \times \mathcal{M}(2, \mathbb{R})$.

For $\mathcal{A} = \mathcal{M}(d, \mathbb{H})$, the submanifold of the square roots f of $-\mathbf{1}$ is a *single connected conjugacy class* of $\mathbb{R}$-dimension $2d^2$ equal to the $\mathbb{R}$-dimension of the centralizer of every f. The easiest example is $\mathbb{H}$ itself for $d = 1$.

For $\mathcal{A} = \mathcal{M}(d, \mathbb{H}^2) = \mathcal{M}(d, \mathbb{H}) \times \mathcal{M}(d, \mathbb{H})$, the square roots of $(-\mathbf{1}, -\mathbf{1})$ are pairs of two square roots (f, f') of $-\mathbf{1}$ in $\mathcal{M}(d, \mathbb{H})$ and constitute a *unique connected conjugacy class* of $\mathbb{R}$-dimension $4d^2$. The group $\mathrm{Aut}(\mathcal{A})$ has two connected components: the neutral component $\mathrm{Inn}(\mathcal{A})$ connected to the identity and the second component containing the swap automorphism $(f, f') \mapsto (f', f)$. The simplest case for $d = 1$ is $\mathbb{H}^2$ isomorphic to $Cl(0,3)$.

For $\mathcal{A} = \mathcal{M}(2d, \mathbb{C})$, the square roots of $-\mathbf{1}$ are in *bijection to the idempotents* [2]. First, the *ordinary* square roots of $-\mathbf{1}$ (with $k = 0$) constitute a conjugacy class of $\mathbb{R}$-dimension $4d^2$ of a *single connected component* which is invariant under $\mathrm{Aut}(\mathcal{A})$. Second, there are $2d$ *conjugacy classes* of *exceptional* square roots of $-\mathbf{1}$, each composed of a *single connected component*, characterized by the equality $\mathrm{Spec}(f) = k/d$ (the pseudoscalar coefficient) with $\pm k \in \{1, 2, \ldots, d\}$, and their $\mathbb{R}$-dimensions are $4(d^2 - k^2)$. The group $\mathrm{Aut}(\mathcal{A})$ includes conjugation of the pseudoscalar $\omega \mapsto -\omega$ which maps the conjugacy class associated with k to the class associated with $-k$. The simplest case for $d = 1$ is the Pauli matrix algebra isomorphic to the geometric algebra $Cl(3,0)$ of 3D Euclidean space $\mathbb{R}^3$, and to complex biquaternions [42].

C. Quaternions

Quaternions are a special Clifford algebra, because the algebra of quaternions $\mathbb{H}$ is isomorphic to $Cl(0,2)$, and to the even grade subalgebra of the Clifford algebra of three-dimensional Euclidean space $Cl^+(3,0)$. But quaternions were initially known independently of Clifford algebras and have their own specific notation, which we briefly introduce here.

Gauss, Rodrigues and Hamilton's four-dimensional (4D) quaternion algebra $\mathbb{H}$ is defined over $\mathbb{R}$ with three imaginary units:

$$\begin{aligned}\boldsymbol{ij} = -\boldsymbol{ji} = \boldsymbol{k}, \qquad \boldsymbol{jk} = -\boldsymbol{kj} = \boldsymbol{i}, \qquad \boldsymbol{ki} = -\boldsymbol{ik} = \boldsymbol{j}, \\ \boldsymbol{i}^2 = \boldsymbol{j}^2 = \boldsymbol{k}^2 = \boldsymbol{ijk} = -1.\end{aligned} \tag{11}$$

[1] Compare chapter 16 on *matrix representations and periodicity of 8*, as well as Table 1 on p. 217 of [36].

[2] Associated ring means, that the matrix elements are from the respective ring $\mathbb{R}$, $\mathbb{R}^2$, $\mathbb{C}$, $\mathbb{H}$ or $\mathbb{H}^2$.

Every quaternion can be written explicitly as

$$q = q_r + q_i\boldsymbol{i} + q_j\boldsymbol{j} + q_k\boldsymbol{k} \in \mathbb{H}, \quad q_r, q_i, q_j, q_k \in \mathbb{R}, \quad (12)$$

and has a *quaternion conjugate* (equivalent[3] to Clifford conjugation in $Cl^+(3,0)$ and $Cl(0,2)$)

$$\overline{q} = q_r - q_i\boldsymbol{i} - q_j\boldsymbol{j} - q_k\boldsymbol{k}, \quad \overline{pq} = \overline{q}\,\overline{p}, \quad (13)$$

which leaves the scalar part q_r unchanged. This leads to the *norm* of $q \in \mathbb{H}$

$$|q| = \sqrt{q\overline{q}} = \sqrt{q_r^2 + q_i^2 + q_j^2 + q_k^2}, \qquad |pq| = |p||q|. \quad (14)$$

The part $V(q) = q - q_r = \frac{1}{2}(q - \overline{q}) = q_i\boldsymbol{i} + q_j\boldsymbol{j} + q_k\boldsymbol{k}$ is called a *pure* quaternion, and it squares to the negative number $-(q_i^2 + q_j^2 + q_k^2)$. Every unit quaternion (i.e. $|q| = 1$) can be written as:

$$\begin{aligned} q &= q_r + q_i\boldsymbol{i} + q_j\boldsymbol{j} + q_k\boldsymbol{k} = q_r + \sqrt{q_i^2 + q_j^2 + q_k^2}\,\boldsymbol{\mu}(q) \\ &= \cos\alpha + \boldsymbol{\mu}(q)\sin\alpha = e^{\alpha\,\boldsymbol{\mu}(q)}, \end{aligned} \quad (15)$$

where

$$\begin{aligned} &\cos\alpha = q_r, \qquad \sin\alpha = \sqrt{q_i^2 + q_j^2 + q_k^2}, \\ &\boldsymbol{\mu}(q) = \frac{V(q)}{|q|} = \frac{q_i\boldsymbol{i} + q_j\boldsymbol{j} + q_k\boldsymbol{k}}{\sqrt{q_i^2 + q_j^2 + q_k^2}}, \qquad \text{and} \qquad \boldsymbol{\mu}(q)^2 = -1. \end{aligned} \quad (16)$$

The *inverse* of a non-zero quaternion is

$$q^{-1} = \frac{\overline{q}}{|q|^2} = \frac{\overline{q}}{q\overline{q}}. \quad (17)$$

The *scalar part* of a quaternion is defined as

$$Sc(q) = q_r = \frac{1}{2}(q + \overline{q}), \quad (18)$$

with *symmetries*

$$\begin{aligned} Sc(pq) = Sc(qp) &= p_r q_r - p_i q_i - p_j q_j - p_k q_k, \\ Sc(q) = Sc(\overline{q}), &\quad \forall p, q \in \mathbb{H}, \end{aligned} \quad (19)$$

and *linearity*

$$\begin{aligned} Sc(\alpha p + \beta q) = \alpha Sc(p) + \beta Sc(q) = \alpha p_r + \beta q_r, \\ \forall p, q \in \mathbb{H}, \;\; \alpha, \beta \in \mathbb{R}. \end{aligned} \quad (20)$$

The scalar part and the quaternion conjugate allow the definition of the $\mathbb{R}^4$ *inner product*[4] of two quaternions p, q as

$$Sc(p\overline{q}) = p_r q_r + p_i q_i + p_j q_j + p_k q_k \in \mathbb{R}. \quad (21)$$

Definition 2 (Orthogonality of quaternions)**.** *Two quaternions $p, q \in \mathbb{H}$ are* orthogonal *$p \perp q$, if and only if the inner product $Sc(p\overline{q}) = 0$.*

[3]This may be important in generalisations of the QFT, such as to a space-time Fourier transform in [19], or a general two-sided Clifford Fourier transform in [24].

[4]Note that we do not use the notation $p \cdot q$, which is unconventional for full quaternions.

III. Inventory of Clifford Fourier Transforms

A. General geometric Fourier transform

Recently a rigorous effort was made in [8] to design a *general geometric Fourier transform*, that incorporates most of the previously known CFTs with the help of very general sets of left and right kernel factor products

$$\begin{aligned} \mathcal{F}_{GFT}\{h\}(\boldsymbol{\omega}) &= \int_{\mathbb{R}^{p',q'}} L(\mathbf{x},\omega)h(\mathbf{x})R(\mathbf{x},\omega)d^{n'}\mathbf{x}, \\ L(\mathbf{x},\omega) &= \prod_{s\in F_L} e^{-s(\mathbf{x},\omega)}, \end{aligned} \quad (22)$$

with $p' + q' = n'$, $F_L = \{s_1(\mathbf{x},\omega), \ldots, s_L(\mathbf{x},\omega)\}$ a set of mappings $\mathbb{R}^{p',q'} \times \mathbb{R}^{p',q'} \to \mathcal{I}^{p,q}$ into the manifold of real multiples of $\sqrt{-1}$ in $Cl(p,q)$. $R(\mathbf{x},\omega)$ is defined similarly, and $h : \mathbb{R}^{p',q'} \to Cl(p,q)$ is the multivector signal function.

B. CFT due to Sommen and Buelow

This clearly subsumes the *CFT due to Sommen and Buelow* [7]

$$\mathcal{F}_{SB}\{h\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^n} h(\mathbf{x}) \prod_{k=1}^{n} e^{-2\pi x_k \omega_k e_k} d^n\mathbf{x}, \quad (23)$$

where $\mathbf{x}, \omega \in \mathbb{R}^n$ with components x_k, ω_k, and $\{e_1, \ldots e_k\}$ is an orthonormal basis of $\mathbb{R}^{0,n}$, $h : \mathbb{R}^n \to Cl(0,n)$.

C. Color image CFT

It is further possible [16] to only pick strictly mutually commuting sets of $\sqrt{-1}$ in $Cl(p,q)$, e.g. e_1e_2, $e_3e_4 \in Cl(4,0)$ and construct CFTs with therefore commuting kernel factors in analogy to (23). Also contained in (22) is the *color image CFT* of [40]

$$\begin{aligned} \mathcal{F}_{CI}\{h\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^2} & e^{\frac{1}{2}\boldsymbol{\omega}\cdot\mathbf{x}I_4B} e^{\frac{1}{2}\boldsymbol{\omega}\cdot\mathbf{x}B} h(\mathbf{x}) \\ & e^{-\frac{1}{2}\boldsymbol{\omega}\cdot\mathbf{x}B} e^{-\frac{1}{2}\boldsymbol{\omega}\cdot\mathbf{x}I_4B} d^2\mathbf{x}, \end{aligned} \quad (24)$$

where $B \in Cl(4,0)$ is a bivector and $I_4B \in Cl(4,0)$ its dual complementary bivector. It is especially useful for the introduction of efficient non-marginal generalized color image Fourier descriptors.

D. Two-sided CFT

The main type of CFT, which we will review here is the general *two sided CFT* [24] with only one kernel factor on each side

$$\mathcal{F}^{f,g}\{h\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^{p',q'}} e^{-fu(\mathbf{x},\omega)} h(\mathbf{x}) e^{-gv(\mathbf{x},\omega)} d^{n'}\mathbf{x}, \quad (25)$$

with f, g two $\sqrt{-1}$ in $Cl(p,q)$, $u, v : \mathbb{R}^{p',q'} \times \mathbb{R}^{p',q'} \to \mathbb{R}$ and often $\mathbb{R}^{p',q'} = \mathbb{R}^{p,q}$. In the following we will discuss a family of transforms, which belong to this class of CFTs, see the lower half of Fig. 2.

ISBN: 978-1-61804-240-8

E. Quaternion Fourier Transform (QFT)

One of the nowadays most widely applied CFTs is the *quaternion Fourier transform* (QFT) [19], [26]

$$\mathcal{F}^{f,g}\{h\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^2} e^{-f x_1 \omega_1} h(\mathbf{x}) e^{-g x_2 \omega_2} d^2\mathbf{x}, \quad (26)$$

which also has variants were one of the left or right kernel factors is dropped, or both are placed together at the right or left side. It was first described by Ernst, et al, [14, pp. 307-308] (with $f = \boldsymbol{i}, g = \boldsymbol{j}$) for spectral analysis in two-dimensional nuclear magnetic resonance, suggesting to use the QFT as a method to independently adjust phase angles with respect to two frequency variables in two-dimensional spectroscopy. Later Ell [12] independently formulated and explored the QFT for the analysis of linear time-invariant systems of PDEs. The QFT was further applied by Buelow, et al [6] for image, video and texture analysis, by Sangwine et al [43], [5] for color image analysis and analysis of non-stationary improper complex signals, vector image processing, and quaternion polar signal representations. It is possible to split every quaternion-valued signal and its QFT into two quasi-complex components [26], which allow the application of complex discretization and fast FT methods. The split can be generalized to the general CFT (25) [24] in the form

$$x_\pm = \frac{1}{2}(x \pm fxg), \quad x \in Cl(p,q). \quad (27)$$

In the case of quaternions the quaternion coefficient space $\mathbb{R}^4$ is thereby split into two steerable (by the choice of two pure quaternions f, g) orthogonal two-dimensional planes [26]. The geometry of this split appears closely related to the quaternion geometry of rotations [39]. For colors expressed by quaternions, these two planes become chrominance and luminance when $f = g =$ gray line [13].

F. Quaternion Fourier Stieltjes transform

Georgiev and Morais have modified the QFT to a *quaternion Fourier Stieltjes transform* [18].

$$\mathcal{F}_{Stj}(\sigma^1, \sigma^2) = \int_{\mathbb{R}^2} e^{-f x_1 \omega_1} d\sigma^1(x_1) d\sigma^2(x_2) e^{-g x_2 \omega_2}, \quad (28)$$

with $f = -\boldsymbol{i}, g = -\boldsymbol{j}$, $\sigma^k : \mathbb{R} \to \mathbb{H}$, $|\sigma^k| \leq \delta_k$ for real numbers $0 < \delta_k < \infty$, $k = 1, 2$.

G. Quaternion Fourier Mellin transform, Clifford Fourier Mellin transform

Introducing polar coordinates in $\mathbb{R}^2$ allows to establish a *quaternion Fourier Mellin transform* (QFMT) [30]

$$\mathcal{F}_{QM}\{h\}(\nu, k) = \frac{1}{2\pi} \int_0^\infty \int_0^{2\pi} r^{-f\nu} h(r,\theta) e^{-gk\theta} d\theta dr/r, \quad \forall (\nu, k) \in \mathbb{R} \times \mathbb{Z}, \quad (29)$$

which can characterize 2D shapes rotation, translation and scale invariant, possibly including color encoded in the quaternion valued signal $h : \mathbb{R}^2 \to \mathbb{H}$ such that $|h|$ is summable over $\mathbb{R}_+^* \times \mathbb{S}^1$ under the measure $d\theta dr/r$, $\mathbb{R}^*$ the multiplicative group of positive non-zero numbers, and $f, g \in \mathbb{H}$ two $\sqrt{-1}$. The QFMT can be generalized straightforward to a *Clifford Fourier Mellin transform* applied to signals $h : \mathbb{R}^2 \to Cl(p,q), p+q=2$ [23], with $f, g \in Cl(p,q), p+q=2$.

H. Volume-time CFT and spacetime CFT

The spacetime algebra $Cl(3,1)$ of Minkowski space with orthonormal vector basis $\{\mathbf{e}_t, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, $-\mathbf{e}_t^2 = \mathbf{e}_1^2 = \mathbf{e}_2^2 = \mathbf{e}_3^3$, has three blades $\mathbf{e}_t, i_3, i_{st}$ of time vector, unit space volume 3-vector and unit hyperspace volume 4-vector, which are isomorphic to Hamilton's three quaternion units

$$\mathbf{e}_t^2 = -1, \quad i_3 = \mathbf{e}_1\mathbf{e}_2\mathbf{e}_3 = \mathbf{e}_t^* = \mathbf{e}_t i_3^{-1}, i_3^2 = -1, \quad i_{st} = \mathbf{e}_t i_3, i_{st}^2 = -1. \quad (30)$$

The $Cl(3,1)$ subalgebra with basis $\{1, \mathbf{e}_t, i_3, i_{st}\}$ is therefore isomorphic to quaternions and allows to generalize the two-sided QFT to a *volume-time Fourier transform*

$$\mathcal{F}_{VT}\{h\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^{3,1}} e^{-\mathbf{e}_t \omega_t} h(\mathbf{x}) e^{-\vec{x}\cdot\vec{\omega}} d^4\mathbf{x}, \quad (31)$$

with $\mathbf{x} = t\mathbf{e}_t + \vec{x} \in \mathbb{R}^{3,1}$, $\vec{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + x_3\mathbf{e}_3$, $\boldsymbol{\omega} = \omega_t\mathbf{e}_t + \vec{\omega} \in \mathbb{R}^{3,1}$, $\vec{\omega} = \omega_1\mathbf{e}_1 + \omega_2\mathbf{e}_2 + \omega_3\mathbf{e}_3$. The split (27) with $f = \mathbf{e}_t$, $g = i_3 = \mathbf{e}_t^*$ becomes the spacetime split of special relativity

$$h_\pm = \frac{1}{2}(1 \pm \mathbf{e}_t h \mathbf{e}_t^*). \quad (32)$$

It is most interesting to observe, that the volume-time Fourier transform can indeed be applied to multivector signal functions valued in the whole spacetime algebra $h : \mathbb{R}^{3,1} \to Cl(3,1)$ without changing its form [19], [22]

$$\mathcal{F}_{ST}\{h\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^{3,1}} e^{-\mathbf{e}_t \omega_t} h(\mathbf{x}) e^{-i_3 \vec{x}\cdot\vec{\omega}} d^4\mathbf{x}. \quad (33)$$

The split (32) applied to *spacetime Fourier transform* (33) leads to a *multivector wavepacket analysis*

$$\mathcal{F}_{ST}\{h\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^{3,1}} h_+(\mathbf{x}) e^{-i_3(\vec{x}\cdot\vec{\omega} - t\omega_t)} d^4\mathbf{x} + \int_{\mathbb{R}^{3,1}} h_-(\mathbf{x}) e^{-i_3(\vec{x}\cdot\vec{\omega} + t\omega_t)} d^4\mathbf{x}, \quad (34)$$

in terms of right and left propagating spacetime multivector wave packets.

I. One-sided CFTs

Finally, we turn to *one-sided CFTs* [25], which are obtained by setting the phase function $u = 0$ in (25). A recent discrete *spinor CFT* used for edge and texture detection is given in [4], where the signal is represented as a spinor and the $\sqrt{-1}$ is a local tangent bivector $B \in Cl(3,0)$ to the image intensity surface ($\mathbf{e}_3$ is the intensity axis).

J. Pseudoscalar kernel CFTs

The following class of *one-sided CFTs which uses a single pseudoscalar* $\sqrt{-1}$ has been well studied and applied [20]

$$\mathcal{F}_{PS}\{h\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^n} h(\mathbf{x}) e^{-i_n \mathbf{x}\cdot\boldsymbol{\omega}} d^n\mathbf{x}, \quad i_n = \mathbf{e}_1\mathbf{e}_2\ldots\mathbf{e}_n, \quad n = 2, 3 (\mathrm{mod}\, 4), \tag{35}$$

where $h : \mathbb{R}^n \to Cl(n,0)$, and $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ is the orthonormal basis of $\mathbb{R}^n$. Historically the special case of (35), $n = 3$, was already introduced in 1990 [32] for the processing of electromagnetic fields. This same transform was later applied [17] to two-dimensional images embedded in $Cl(3,0)$ to yield a two-dimensional analytic signal, and in image structure processing. Moreover, the *pseudoscalar CFT* (35), $n = 3$, was successfully applied to three-dimensional vector field processing in [10], [9] with vector signal convolution based on Clifford's full geometric product of vectors. The theory of the transform has been thoroughly studied in [20].

For embedding one-dimensional signals in $\mathbb{R}^2$, [17] considered in (35) the special case of $n = 2$, and in [10], [9] this was also applied to the processing of two-dimensional vector fields.

Recent applications of (35) with $n = 2, 3$, to geographic information systems and climate data can be found in [47], [46], [35].

K. Quaternion and Clifford linear canonical transforms

Real and complex linear canonical transforms parametrize a continuum of transforms, which include the Fourier, fractional Fourier, Laplace, fractional Laplace, Gauss-Weierstrass, Bargmann, Fresnel, and Lorentz transforms, as well as scaling operations. A Fourier transform transforms multiplication with the space argument x into differentiation with respect to the frequency argument ω. In Schroedinger quantum mechancis this constitutes a rotation in position-momentum phase space. A linear canonical transform transforms the position and momentum operators into linear combinations (with a two-by-two real or complex parameter matrix), preserving the fundamental position-momentum commutator relationship, at the core of the uncertainty principle. The transform operator can be made to act on appropriate spaces of functions, and can be realized in the form of integral transforms, parametrized in terms of the four real (or complex) matrix parameters [44].

KitIan Kou et al [34] introduce the quaternionic linear canonical transform (QLCT). They consider a pair of unit determinant two-by-two matrices

$$A_1 = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix}, \tag{36}$$

with entries $a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2 \in \mathbb{R}$, $a_1d_1 - c_1b_1 = 1$, $a_2d_2 - c_2b_2 = 1$, where they disregard the cases $b_1 = 0$, $b_2 = 0$, for which the LCT is essentially a chirp multiplication.

We now *generalize* the definitions of [34] using the following two kernel functions with two pure unit quaternions $f, g \in \mathbb{H}$, $f^2 = g^2 = -1$, including the cases $f = \pm g$,

$$K^f_{A_1}(x_1, \omega_1) = \frac{1}{\sqrt{f2\pi b_1}} e^{f(a_1x_1^2 - 2x_1\omega_1 + d_1\omega_1^2)/(2b_1)},$$
$$K^g_{A_2}(x_2, \omega_2) = \frac{1}{\sqrt{g2\pi b_2}} e^{g(a_2x_2^2 - 2x_2\omega_1 + d_2\omega_2^2)/(2b_2)}. \tag{37}$$

The *two-sided QLCT* of signals $h \in L^1(\mathbb{R}^2, \mathbb{H})$ can now generally be defined as

$$\mathcal{L}^{f,g}(\boldsymbol{\omega}) = \int_{\mathbb{R}^2} K^f_{A_1}(x_1, \omega_1) h(\mathbf{x}) K^g_{A_2}(x_2, \omega_2) d^2\mathbf{x}. \tag{38}$$

The *left-sided* and *right-sided QLCTs* can be defined correspondingly by placing the two kernel factors both on the left or on the right[5], respectively. For $a_1 = d_1 = a_2 = d_2 = 0$, $b_1 = b_2 = 1$, the conventional two-sided (left-sided, right-sided) QFT is recovered. We note that it will be of interest to "complexify" the matrices A_1 and A_2, by including replacing $a_1 \to a_{1r} + fa_{1f}$, $a_2 \to a_{2r} + ga_{2g}$, etc. In [34] for $f = \boldsymbol{i}$ and $g = \boldsymbol{j}$ the right-sided QLCT and its properties, including an uncertainty principle are studied in some detail.

In [45] a complex Clifford linear canonical transform is defined and studied for signals $f \in L^1(\mathbb{R}^m, C^{m+1})$, where $C^{m+1} = \mathrm{span}\{1, e_1, \ldots, e_m\} \subset Cl(0,m)$ is the subspace of paravectors in $Cl(0,m)$. This includes uncertainty principles. Motivated by Remark 2.2 in [45], we now modify this definition to generalize the one-sided CFT of [25] for real Clifford algebras $Cl(n,0)$ to a *general real Clifford linear canonical transform (CLNT)*. We define the parameter matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad a, b, c, d \in \mathbb{R}, \quad ad - cb = 1. \tag{39}$$

We again omit the case $b = 0$ and define the kernel

$$K^f(\mathbf{x}, \boldsymbol{\omega}) = \frac{1}{\sqrt{f(2\pi)^n b}} e^{f(a\mathbf{x}^2 - 2\mathbf{x}\cdot\boldsymbol{\omega} + d\boldsymbol{\omega}^2)/(2b)}, \tag{40}$$

with the general square root of -1: $f \in Cl(n,0)$, $f^2 = -1$. Then the general real CLNT can be defined for signals $h \in L^1(\mathbb{R}^n; Cl(n,0))$ as

$$\mathcal{L}^f\{h\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^n} h(\mathbf{x}) K^f(\mathbf{x}, \boldsymbol{\omega}) d^n\mathbf{x}. \tag{41}$$

For $a = d = 0$, $b = 1$, the conventional one-sided CFT of [25] in $Cl(n,0)$ is recovered. It is again of interest to modify the entries of the parameter matrix to $a \to a_0 + fa_f, b \to b_0 + fb_f$, etc.

Similarly in [33] a *Clifford version of a linear canonical transform (CLCT)* for signals $h \in L^1(\mathbb{R}^m; \mathbb{R}^{m+1})$ is formulated using two-by-two parameter matrices $A_1, \ldots, A_m$, which maps $\mathbb{R}^m \to Cl(0,m)$. The Sommen Bülow CFT (23) is recovered for parameter matrix entries $a_k = d_k = 0$, $b_k = 1$, $1 \le k \le m$.

[5] In [34] the possibility of a more general pair of unit quaternions $f, g \in \mathbb{H}$, $f^2 = g^2 = -1$, is only indicated for the case of the right-sided QLCT, but with the restriction that f, g should be an *orthonormal pair* of pure quaternions, i.e. $Sc(f\overline{g}) = 0$. Otherwise [34] always strictly sets $f = \boldsymbol{i}$ and $g = \boldsymbol{j}$.

ISBN: 978-1-61804-240-8

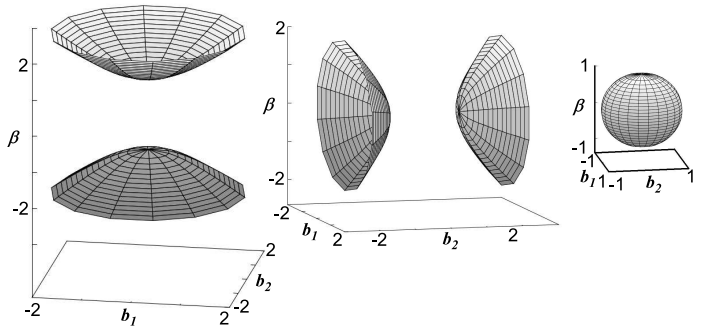


Fig. 1. Manifolds [28] of square roots f of -1 in $Cl(2,0)$ (left), $Cl(1,1)$ (center), and $Cl(0,2) \cong \mathbb{H}$ (right). The square roots are $f = \alpha + b_1 e_1 + b_2 e_2 + \beta e_{12}$, with $\alpha, b_1, b_2, \beta \in \mathbb{R}$, $\alpha = 0$, and $\beta^2 = b_1^2 e_2^2 + b_2^2 e_1^2 + e_1^2 e_2^2$.

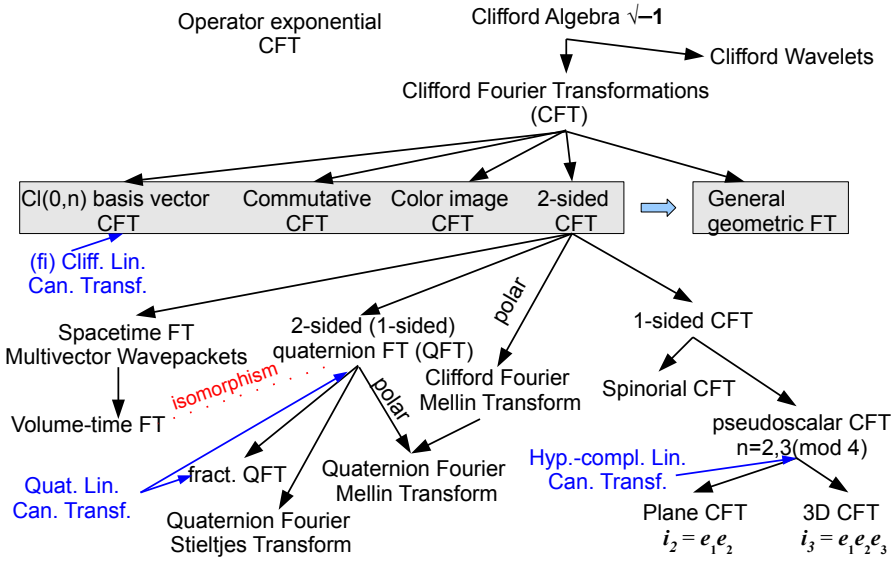


Fig. 2. Family tree of Clifford Fourier transformations.

IV. Conclusion

We have reviewed Clifford Fourier transforms which apply the manifolds of $\sqrt{-1} \in Cl(p,q)$ in order to create a rich variety of new Clifford valued Fourier transformations. The history of these transforms spans just over 30 years. Major steps in the development were: $Cl(0,n)$ CFTs, then pseudoscalar CFTs, and Quaternion FTs. In the 1990ies especially applications to electromagnetic fields/electronics and in signal/image processing dominated. This was followed by by color image processing and most recently applications in Geographic Information Systems (GIS). This paper could only feature a part of the approaches in CFT research, and only a part of the applications. Omitted were details on operator exponential CFT approach [5], and CFT for conformal geometric algebra. Regarding applications, e.g. CFT Fourier descriptor representations of shape [41] of B. Rosenhahn, et al was omitted. Note that there are further types of Clifford algebra/analysis related integral transforms: Clifford wavelets, Clifford radon transforms, Clifford Hilbert transforms, ... which we did not discuss.

Acknowledgment

Soli deo gloria. I do thank my dear family, and the organizers of the PMAMCM 2014 conference in Santorini, Greece.

References

[1] R. Abłamowicz, *Computations with Clifford and Grassmann Algebras*, Adv. Appl. Clifford Algebras **19**, No. 3–4 (2009), 499–545.

[2] R. Abłamowicz, B. Fauser, K. Podlaski, J. Rembieliński, *Idempotents of Clifford Algebras*. Czechoslovak Journal of Physics, **53** (11) (2003), 949–954.

[3] R. Abłamowicz and B. Fauser, CLIFFORD with Bigebra – A Maple Package for Computations with Clifford and Grassmann Algebras, http://math.tntech.edu/rafal/ (©1996-2012).

[4] T. Batard, M. Berthier, *CliffordFourier Transf. and Spinor Repr. of Images*, in: E. Hitzer, S.J. Sangwine (eds.), "Quaternion and Clifford Fourier Transf. and Wavelets", TIM **27**, Birkhauser, Basel, 2013, 177–195.

[5] F. Brackx, et al, *History of Quaternion and Clifford-Fourier Transf.*, in: E. Hitzer, S.J. Sangwine (eds.), "Quaternion and Clifford Fourier Transf. and Wavelets", TIM **27**, Birkhauser, Basel, 2013, xi–xxvii.

[6] T. Buelow, *Hypercomplex Spectral Signal Repr. for the Proc. and Analysis of Images*, PhD thesis, Univ. of Kiel, Germany, Inst. fuer Informatik und Prakt. Math., Aug. 1999.

[7] T. Buelow, et al, *Non-comm. Hypercomplex Fourier Transf. of Multidim. Signals*, in G. Sommer (ed.), "Geom. Comp. with Cliff. Algebras", Springer 2001, 187–207.

[8] R. Bujack, et al, *A General Geom. Fourier Transf.*, in: E. Hitzer, S.J. Sangwine (eds.), "Quaternion and Clifford Fourier Transf. and Wavelets", TIM **27**, Birkhauser, Basel, 2013, 155–176.

[9] J. Ebling, G. Scheuermann, *Clifford convolution and pattern matching on vector fields*, In Proc. IEEE Vis., **3**, IEEE Computer Society, Los Alamitos, 2003. 193–200,

[10] J. Ebling, G. Scheuermann, *Cliff. Four. transf. on vector fields*, IEEE Trans. on Vis. and Comp. Graph., **11**(4), (2005), 469–479.

[11] D. Eelbode, E. Hitzer, *Operator Exponentials for the Clifford Fourier Transform on Multivector Fields*, in preparation, 18 pages. Preprint: http://vixra.org/abs/1403.0310.

[12] T. A. Ell, *Quaternionic Fourier Transform for Analysis of Two-dimensional Linear Time-Invariant Partial Differential Systems.* in Proceedings of the 32nd IEEE Conference on Decision and Control, December 15-17, **2** (1993), 1830–1841.

[13] T.A. Ell, S.J. Sangwine, *Hypercomplex Fourier transforms of color images*, IEEE Trans. Image Process., **16**(1) (2007), 22–35.

[14] R. R. Ernst, et al, *Princ. of NMR in One and Two Dim.*, Int. Ser. of Monogr. on Chem., Oxford Univ. Press, 1987.

[15] M.I. Falcao, H.R. Malonek, *Generalized Exponentials through Appell sets in* $\mathbb{R}^{n+1}$ *and Bessel functions*, AIP Conference Proceedings, Vol. 936, pp. 738–741 (2007).

[16] M. Felsberg, et al, *Comm. Hypercomplex Fourier Transf. of Multidim. Signals*, in G. Sommer (ed.), "Geom. Comp. with Cliff. Algebras", Springer 2001, 209–229.

[17] M. Felsberg, *Low-Level Img. Proc. with the Struct. Multivec.*, PhD thesis, Univ. of Kiel, Inst. fuer Inf. & Prakt. Math., 2002.

[18] S. Georgiev, J. Morais, *Bochner's Theorems in the Framework of Quaternion Analysis* in: E. Hitzer, S.J. Sangwine (eds.), "Quaternion and Clifford Fourier Transf. and Wavelets", TIM **27**, Birkhauser, Basel, 2013, 85–104.

[19] E. Hitzer, *Quaternion Fourier Transform on Quaternion Fields and Generalizations*, AACA, **17** (2007), 497–517.

[20] E. Hitzer, B. Mawardi, *Clifford Fourier Transf. on Multivector Fields and Unc. Princ. for Dim.* $n = 2$ *(mod* 4*) and* $n = 3$ *(mod* 4*)*, P. Angles (ed.), AACA, **18**(S3,4), (2008), 715–736.

[21] E. Hitzer, *Cliff. (Geom.) Alg. Wavel. Transf.*, in V. Skala, D. Hildenbrand (eds.), Proc. GraVisMa 2009, Plzen, 2009, 94–101.

[22] E. Hitzer, *Directional Uncertainty Principle for Quaternion Fourier Transforms*, AACA, **20**(2) (2010), 271–284.

[23] E. Hitzer, *Clifford Fourier-Mellin transform with two real square roots of* -1 *in* $Cl(p,q)$*,* $p+q=2$, 9th ICNPAA 2012, AIP Conf. Proc., **1493**, (2012), 480–485.

[24] E. Hitzer, *Two-sided Clifford Fourier transf. with two square roots of* -1 *in* $Cl(p,q)$, Advances in Applied Clifford Algebras, 2014, 20 pages, DOI: 10.1007/s00006-014-0441-9. First published in M. Berthier, L. Fuchs, C. Saint-Jean (eds.) electronic Proceedings of AGACSE 2012, La Rochelle, France, 2-4 July 2012. Preprint: http://arxiv.org/abs/1306.2092 .

[25] E. Hitzer, *The Clifford Fourier transform in real Clifford algebras*, in E. H., K. Tachibana (eds.), "Session on Geometric Algebra and Applications, IKM 2012", Special Issue of Clifford Analysis, Clifford Algebras and their Applications, Vol. 2, No. 3, pp. 227-240, (2013). First published in K. Guerlebeck, T. Lahmer and F. Werner (eds.), electronic Proc. of 19th International Conference on the Application of Computer Science and Mathematics in Architecture and Civil Engineering, IKM 2012, Weimar, Germany, 0406 July 2012. Preprint: http://vixra.org/abs/1306.0130 .

[26] E. Hitzer, S. J. Sangwine, *The Orthogonal 2D Planes Split of Quaternions and Steerable Quaternion Fourier Transf.*, in: E. Hitzer, S.J. Sangwine (eds.), "Quaternion and Clifford Fourier Transf. and Wavelets", TIM **27**, Birkhauser, Basel, 2013, 15–39.

[27] E. Hitzer, R. Abłamowicz, *Geometric Roots of* -1 *in Clifford Algebras* $Cl(p,q)$ *with* $p+q \leq 4$. Adv. Appl. Clifford Algebras, **21**(1), (2011) 121–144, DOI: 10.1007/s00006-010-0240-x.

[28] E. Hitzer, J. Helmstetter, and R. Abłamowicz, *Square roots of -1 in real Clifford algebras*, in: E. Hitzer, S.J. Sangwine (eds.), "Quaternion and Clifford Fourier Transf. and Wavelets", TIM **27**, Birkhauser, Basel, 2013, 123–153.

[29] E. Hitzer, J. Helmstetter, and R. Abłamowicz, Maple worksheets created with `CLIFFORD` for a verification of results in [28], http://math.tntech.edu/rafal/publications.html (©2012).

[30] E. Hitzer, *Quaternionic Fourier-Mellin Transf.*, in T. Sugawa (ed.), Proc. of ICFIDCAA 2011, Hiroshima, Japan, Tohoku Univ. Press, Sendai (2013), ii, 123–131.

[31] E. Hitzer, *Extending Fourier transformations to Hamiltons quaternions and Cliffords geometric algebras*, In T. Simos, G. Psihoyios and C. Tsitouras (eds.), Numerical Analysis and Applied Mathematics ICNAAM 2013, AIP Conf. Proc. 1558, pp. 529–532 (2013). DOI: 10.1063/1.4825544, Preprint: http://viXra.org/abs/1310.0249

[32] B. Jancewicz. *Trivector Fourier transf. and electromag. Field*, J. of Math. Phys., **31**(8), (1990), 1847–1852.

[33] K. Kou, J. Morais, Y. Zhang, *Generalized prolate spheroidal wave functions for offset linear canonical transform in Clifford Analysis*, Math. Meth. Appl. Sci. 2013, **36** pp. 1028-1041. DOI: 10.1002/mma.2657.

[34] K. Kou, J-Y. Ou, J. Morais, *On Uncertainty Principle for Quaternionic Linear Canonical Transform*, Abs. and App. Anal., Vol. 2013, IC 72592, 14 pp.

[35] Linwang Yuan, et al, *Geom. Alg. for Multidim.-Unified Geogr. Inf. System*, AACA, **23** (2013), 497–518.

[36] P. Lounesto, *Clifford Algebras and Spinors*, CUP, Cambridge (UK), 2001.

[37] B. Mawardi, E. Hitzer, *Clifford Algebra Cl(3,0)-valued Wavelet Transf., Clifford Wavelet Uncertainty Inequality and Clifford Gabor Wavelets*, Int. J. of Wavelets, Multiresolution and Inf. Proc., **5**(6) (2007), 997–1019.

[38] Waterloo Maple Incorporated, *Maple, a general purpose computer algebra system*. Waterloo, `http://www.maplesoft.com` (©2012).

[39] L. Meister, H. Schaeben, *A concise quaternion geometry of rotations*, Mathematical Methods in the Applied Sciences 2005; **28**(1): pp. 101–126.

[40] J. Mennesson, et al, *Color Obj. Recogn. Based on a Clifford Fourier Transf.*, in L. Dorst, J. Lasenby, "Guide to Geom. Algebra in Pract.", Springer 2011, 175–191.

[41] B. Rosenhahn, G. Sommer *Pose estimation of free-form objects*, European Conference on Computer Vision, Springer-Verlag, Berlin, **127**, pp. 414–427, Prague, 2004, edited by Pajdla, T.; Matas, J.

[42] S. J. Sangwine, *Biquaternion (Complexified Quaternion) Roots of* -1. Adv. Appl. Clifford Algebras **16**(1) (2006), 63–68.

[43] S. J. Sangwine, *Fourier transforms of colour images using quaternion, or hypercomplex, numbers*, Electronics Letters, **32**(21) (1996), 1979–1980.

[44] K.B. Wolf, *Integral Transforms in Science and Engineering*, Chapters 9&10, New York, Plenum Press, 1979. http://www.fis.unam.mx/~bwolf/integral.html

[45] Y. Yang, K. Kou, *Uncertainty principles for hypercomplex signals in the linear canoncial transform domains*, Signal Proc., Vol. 95 (2014), pp. 67–75.

[46] Yuan Linwang, et al, *Pattern Forced Geophys. Vec. Field Segm. based on Clifford FFT*, To appear in Computer & Geoscience.

[47] Yu Zhaoyuan, et al, *Clifford Algebra for Geophys. Vector Fields*, To appear in Nonlinear Process in Geophysics.

Eckhard Hitzer Eckhard Hitzer is Senior Associate Professor at the Department of Material Science at the International Christian University in Mitaka/Tokyo, Japan. His special interests are theoretical and applied Clifford geometric algebras and Clifford analysis, including applications to crystal symmetry visualization, neural networks, signal and image processing. Additionally he is interested in environmental radiation measurements, renewable energy and energy efficient buildings.

Computing the distribution function via adaptive multilevel splitting

Ioannis Phinikettos, Ioannis Demetriou and Axel Gandy

Abstract—The method of adaptive multilevel splitting is extended by estimating the whole distribution function and not only the probability over some point. An asymptotic result is proved that is used to construct confidenc bands inside a closed interval in the tail of the distribution function. A simulation study is performed giving empirical evidence to the theory. Confidenc bands are constructed using as an example the infinit norm of a multivariate normal random vector.

I. Introduction

A powerful method for rare event simulation, is multilevel splitting. Multilevel splitting, introduced in [8] and [11], has become very popular the last decades. The main principle of multilevel splitting, is to split the state space into a sequence of sub-levels that decrease to a rarer set. Then the probability of this rarer set is estimated by using all the subsequent conditional probabilities of the sub-levels.

One of the main problems of multilevel splitting is how to fi the sub-levels in advance. Several adaptive methods were introduced by [3], [4], and [7]. The difference to this article is that we are reporting all the hitting probabilities up to a certain threshold i.e. constructing the distribution function.

[7] have shown that several properties for their probability estimator hold. They have shown that their estimator follows a certain discrete distribution. Using this distribution, they have constructed appropriate confidenc intervals, shown that the estimator is unbiased and have given an explicit form of the variance. We extend their results by showing that the estimated distribution function follows a certain stochastic process and given this process, we construct appropriate confidenc bands.

We will apply adaptive multilevel splitting to construct the confidenc bands for the infinit norm of a multivariate normal random vector. The multivariate normal distribution is an important tool in the statistical community. The most important algorithm that is used for the computation of multivariate normal probabilities is given in [5]. Another method, that is designed to exploit the infinit norm is given in [9].

The article is structured as follows. In Section II, we state and prove the main theorem and give the form of the confidenc bands. In Section III, we apply the method to the multivariate normal distribution and construct the appropriate confidenc bands. A discussion is contained in Section IV.

II. Methodology

Suppose we want to estimate the hitting probability $p_c = \mathbb{P}(\phi(X) > c)$, where X is a random element on some a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with distribution function μ, together with some measurable real function $\phi(X) : \Omega \to \mathbb{R}$ and $c \in \mathbb{R}$. We apply adaptive multilevel splitting for the estimation of the whole distribution function of $\phi(X)$ up to a certain threshold and not only of the probability over that point.

A. Algorithm

In this section, we present the algorithm for computing the CDF of $\phi(X)$ via adaptive multilevel splitting using the idea of [7]. The algorithm is given below:

Algorithm II.1 (Computing the CDF via adaptive multilevel splitting)**.**

1) *Input* $c \in \mathbb{R}$ *and* $N \in \mathbb{N}$. *Set* $p = 1-1/N$ *and* $c_0 = -\infty$.
2) *Generate* N *i.i.d. random elements* $X_1^{(1)}, \ldots, X_N^{(1)} \sim X$. *Set* $j := 1$.
3) *Let* $c_j = \min_i \phi(X_i^{(j)})$.
4) *If* $c_j \geq c$ *set* $c_j = c$ *and GOTO 7.*
5) *Let*

$$X_i^{(j+1)} = \begin{cases} X_i^{(j)} & \text{if } \phi(X_i^{(j)}) > c_j \\ \tilde{X}_i \sim \mathcal{L}(X|\phi(X) > c_j) & \text{if } \phi(X_i^{(j)}) \leq c_j, \end{cases}$$

   where the $\tilde{X}_i$ *are independent and also independent of* $\{X_1^{(j)}, \ldots, X_N^{(j)}\}$.
6) *Set* $j := j + 1$ *and GOTO 3.*
7) *The CDF is given by the step function*

$$\hat{F}(k) = \sum_{i=0}^{j-1} (1 - p^i) \mathcal{I}\{c_i \leq k < c_{i+1}\}, \quad k \leq c. \quad (1)$$

The critical part of this algorithm is the updating Step 5. At each iteration the paths have to be independent and drawn from the conditional law. Those particles that were over the threshold cause no problem. The ones that were below have to be killed and regenerate new particles, keeping the sample size at each iteration constant.

Efficien ways to regenerate the particles have been studied in [3], [4] and [7] . In [4] adaptive multilevel splitting was considered for the estimation of very small entrance probabilities in strong Markov processes using an almost surely finit stopping time. The particles that were below the threshold in Step 5 were killed. For each killed particle, a new particle from the conditional law was generated by choosing uniformly a survived particle and extending it from the firs point it crossed the threshold until the stopping time. [3] and [7] considered adaptive multilevel splitting in static distributions. In this setting the particles could not be extended and some transition kernel (e.g. a Metropolis kernel) was constructed to

ISBN: 978-1-61804-240-8

overcome this difficult . The new transition was accepted if its corresponding value was over the threshold.

Before we move to the next section, let us introduce some notation. We denote the random variable $M_c = \max\{m \in \mathbb{N} : c_m \leq c\}$, which depends on the number of particles N. The adaptive multilevel splitting estimator for $\mathbb{P}(\phi(X) > c) = 1 - F(c)$, using the idea of [7], is then given by $\hat{p}_c = (1-1/N)^{M_c}$. Since we are interested in estimating the distribution function F, we consider the stochastic process $\{M_c\}_{c\in I_F}$, where I_F denotes the support of F.

B. Construction of the confidence bands

We assume that the distribution function F of $\phi(X)$ is continuous and in addition that F is strictly increasing over some closed interval $I = [c_{\min}, c_{\max}]$, such that $0 \leq F(c_{\min}) < F(c_{\max}) < 1$. The reason for introducing the interval I, is that we want to construct confidenc bands over I in the tail of the distribution function F.

We denote the survivor function and the integrated hazard function of $\phi(X)$ by $S(x) = 1-F(x)$ and $\Lambda(x) = -\log S(x)$, respectively. We also defin the set $\tilde{I} = \{\Lambda(c) : c \in I\}$, which is also a closed interval given by $\tilde{I} = [\Lambda(c_{\min}), \Lambda(c_{\max})]$.

The next proposition proves that $\{M_c\}_{c\in I_F}$ is a Poisson process of rate N, subject to the time transformation $c \to \Lambda(c)$.

Proposition II.1. *$\{M_c\}_{c\in I_F} \overset{d}{=} \{\tilde{M}_{\Lambda(c)}\}_{c\in I_F}$, where $\{\tilde{M}_t\}_{t\in\mathbb{R}_{\geq 0}}$ is a Poisson process of rate N.*

Proof: The proof is exactly the same as Corollary 1 of [7], but with a different conclusion. The random variable M_c can be written as

$$\begin{aligned} M_c &= \max\{m : c_m \leq c\} \\ &= \max\{m : S(c_m) \geq S(c)\} \\ &= \max\{m : \Lambda(c_m) \leq \Lambda(c)\}. \end{aligned}$$

The random variables $\Lambda(c_1), \dots \Lambda(c_m), \dots$ can be viewed as the successive arrival times of a Poisson process of rate N, as it has been proved in Theorem 1 of [7]. If we consider the stochastic process $\{M_c\}_{c\in I_F}$, this is just the definitio of a Poisson process subject to the time transformation $c \to \Lambda(c)$. ■

Note that since $\Lambda : I_F \to \mathbb{R}_{\geq 0}$ might not be injective, the process $\{\tilde{M}_{\Lambda(c)}\}_{c\in I_F}$ might not be a Poisson process of standard form. But inside the interval $I \subset I_F$ it is injective and as $\Lambda(x)$ is continuous, $\{\tilde{M}_{\Lambda(c)}\}_{c\in I}$ is a Poisson process restricted to I subject to the time transformation $c \to \Lambda(c)$. First, let us defin the stochastic process $\{A_c\}_{c\in I}$, with

$$A_c = a_N(\log \hat{p}_c - \log \hat{p}_{c_{\min}} - N \log(1-1/N)[\Lambda(c) - \Lambda(c_{\min})]), \tag{2}$$

where $a_N = \frac{1}{\sqrt{N}\log(1-1/N)}$. Also let $\{B_t\}_{t\in\mathbb{R}_{\geq 0}}$ be a standard Brownian motion. Convergence in distribution ($\overset{d}{\to}$) always be as $N \to \infty$. The next theorem proves an asymptotic result for A_c, which can be used to construct confidenc bands for the distribution function of $\phi(X)|\phi(X) > c_{\min}$ over the interval I.

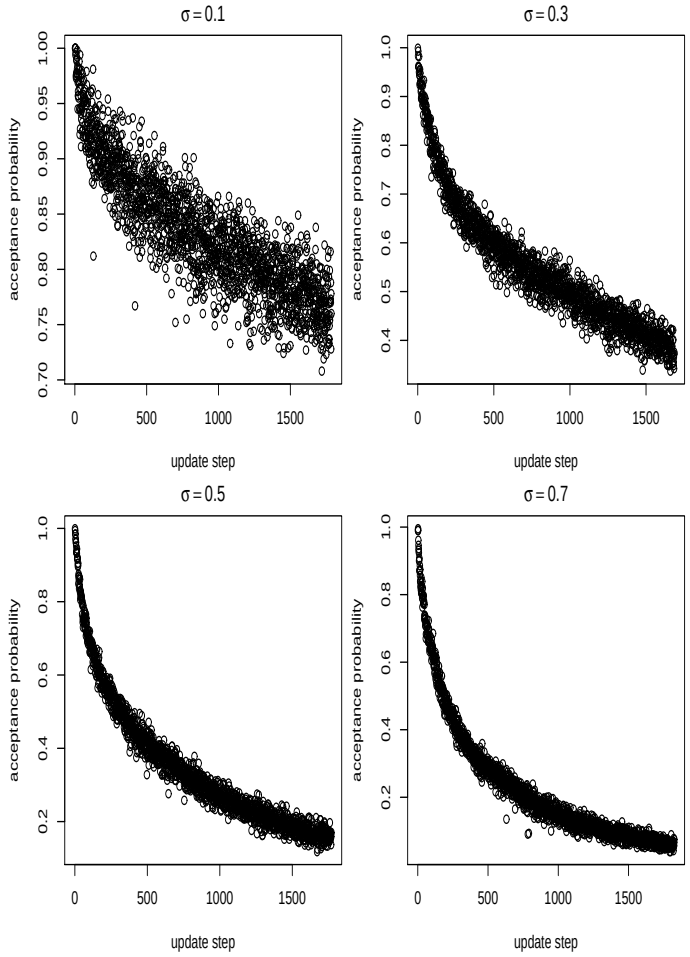


Fig. 1. Plots of the acceptance probability of the transition kernel for each step of the algorithm. We use a sample size $N = 100$, $t = 1000$ with $c = 20$ and for covariance matrix Σ_2.

Theorem II.1. *Let $b = \log(\hat{p}_{c_{\min}}) - \log(\hat{p}_{c_{\max}})$. Then*

$$\frac{\sup_{c\in I}|A_c|}{\sqrt{b}} \overset{d}{\to} \sup_{0\leq t\leq 1}|B_t|.$$

The proof of the theorem is based on the following lemma. For the next lemma, the symbol $\overset{d}{\to}$ denote convergence in distribution in the set of cádlág functions endowed with the Skorohod topology [**?**,]Chapters 1 and 3]billingsley1999convergence.

Lemma 1. *Define the process $\{Z_t\}_{t\in\mathbb{R}_{\geq 0}}$ by $Z_t = \frac{1}{\sqrt{N}}(\tilde{M}_t - Nt)$, where $\tilde{M}_t$ is a Poisson process of rate N. The following is true*

$$\{Z_t\}_{t\in\mathbb{R}_{\geq 0}} \overset{d}{\to} \{B_t\}_{t\in\mathbb{R}_{\geq 0}}.$$

Proof: To prove this lemma, we need to show the 3 suffi cient conditions of Proposition 1 of [10]. A brief introduction to Poisson process and martingale theory can be found in [1, Section 2.2].

Firstly, we need to show that Z_t is a local martingale. As Nt is the compensator of $\tilde{M}_t$, then Z_t is a martingale and therefore a local martingale.

Secondly, we have to show that the maximum of the jumps of Z_t converges to 0 as $N \to \infty$. As a Poisson process attains

jumps of size 1, then Z_t attains jumps of size $1/\sqrt{N}$ which turns to 0 as $N \to \infty$.

Finally, we need the predictable variation process $< . >$ of Z_t to converge in probability to an increasing function H. In our case we must have $H(t) = t$. We have $< \tilde{M}_t - Nt >= Nt$ and as $< \beta D_t >= \beta^2 < D_t >$ for $\beta \in \mathbb{R}$ and any stochastic process D_t, the third property holds. As the three sufficien conditions hold, the result follows.

■

Using this lemma, we are now able to prove the theorem.

Proof of Theorem II.1: As $\log \hat{p}_c = M_c \log(1 - 1/N)$, the process $\{A_c\}_{c \in I}$ is transformed to

$$\begin{aligned} A_c &= \frac{1}{\sqrt{N}}(M_c - M_{c_{\min}} - N(\Lambda(c) - \Lambda(c_{\min}))) \\ &\stackrel{d}{=} \frac{1}{\sqrt{N}}(\tilde{M}_{\Lambda(c)} - \tilde{M}_{\Lambda(c_{\min})} - N(\Lambda(c) - \Lambda(c_{\min}))),\ c \in I. \end{aligned}$$

Consider the continuous time transformation $c \to \Lambda^{-1}(t + \Lambda(c_{\min}))$. We get the transformed process $\{\tilde{A}_t\}_{t \in \hat{I}}$, with

$$\tilde{A}_t = \frac{1}{\sqrt{N}}(\tilde{M}_{t+\Lambda(c_{\min})} - \tilde{M}_{\Lambda(c_{\min})} - Nt),$$

where $\hat{I} = [0, \Lambda(c_{\max}) - \Lambda(c_{\min})]$. Since a Poisson process is a Levy process, the process $\tilde{M}_{t+\Lambda(c_{\min})} - \tilde{M}_{\Lambda(c_{\min})}$ is also a Poisson process of rate N. So we can use Lemma 1 to say $\{\tilde{A}_t\}_{t \in \hat{I}} \stackrel{d}{\to} \{B_t\}_{t \in \hat{I}}$.

As the map $\sup|.|$, where the supremum runs over some closed interval, is continuous between the set of cádlág functions to the set of real numbers, by the continuous mapping theorem [2, Theorem 2.7], we get the convergence

$$\sup_{t \in \hat{I}} |\tilde{A}_t| \stackrel{d}{\to} \sup_{t \in \hat{I}} |B_t|. \tag{3}$$

Next, we use the result that $\{B_t\}_{0 \le t \le \tilde{t}}/\sqrt{\tilde{t}}$ is the same as a standard Brownian motion on $[0, 1]$ and also use Slutsky's lemma. [7] have proved that $\hat{p}_c$ is an unbiased estimator of $S(c)$ for all c and also proved that $\text{var}(\hat{p}_c) = p_c^2(p_c^{-1/N} - 1)$. As $\text{var}(\hat{p}_c) \to 0$ as $N \to \infty$, by a standard result, we get consistency i.e. $\hat{p}_c \stackrel{d}{\to} S(c)$. As the function $\log(x)$ is continuous for all $0 < x < 1$, we get $-\log(\hat{p}_c) \stackrel{d}{\to} \Lambda(c)$ for all c. We have that

$$\begin{aligned} \frac{\sup_{t \in \hat{I}} |B_t|}{\sqrt{b}} &= \frac{\sup_{t \in \hat{I}} |B_t|}{\sqrt{\Lambda(c_{\max}) - \Lambda(c_{\min})}} \frac{\sqrt{\Lambda(c_{\max}) - \Lambda(c_{\min})}}{\sqrt{-\log(\hat{p}_{c_{\max}}) + \log(\hat{p}_{c_{\min}})}} \\ &\stackrel{d}{\to} \sup_{0 \le t \le 1} |B_t|. \end{aligned}$$

As $\sup_{c \in I} |A_c| \stackrel{d}{=} \sup_{t \in \hat{I}} |\tilde{A}_t|$, we get the required result

$$\frac{\sup_{c \in I} |A_c|}{\sqrt{b}} \stackrel{d}{\to} \sup_{0 \le t \le 1} |B_t|.$$

■

Next, we construct the confidenc bands for the conditional integrated hazard function $Y(c) = \Lambda(c) - \Lambda(c_{\min})$ and the conditional survivor function $W(c) = S(c)/S(c_{\min})$. Their corresponding estimators are given by $\hat{Y}(c) = \hat{\Lambda}(c) - \hat{\Lambda}(c_{\min})$ and $\hat{W}(c) = \hat{S}(c)/\hat{S}(c_{\min})$, where $\hat{S}(k) = \hat{p}_k$ and $\hat{\Lambda}(k) = -\log(\hat{p}_k)$.

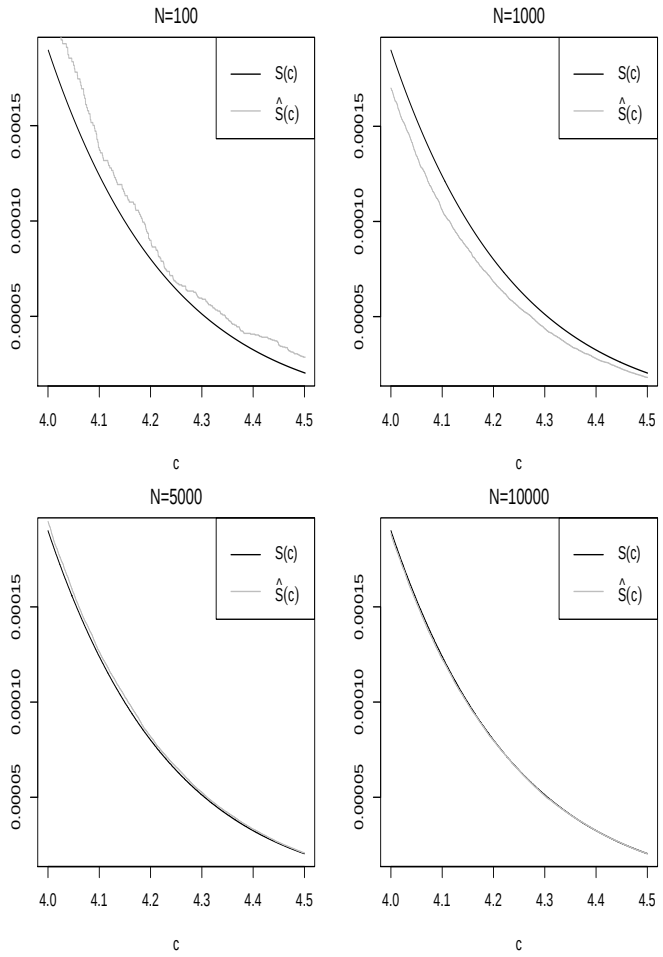


Fig. 2. Plots of the survivor function $S(c)$ and the its estimate $\hat{S}(c)$ using different sample sizes N and $c \in [4, 4.5]$ for covariance matrix Σ_1.

Corollary 1. *The conditional integrated hazard function $Y(c)$, with $c \in I$, has α - confidence bands given by*

$$Y^{\pm}(c) = \frac{\hat{Y}(c)}{z_N} \pm \frac{h_\alpha \sqrt{b}}{a_N z_N}, \quad c \in I, \tag{4}$$

where $z_N = -N \log(1 - 1/N)$ and h_α is the α -quantile for the distribution function of $\sup_{t \in [0,1]} |B(t)|$. Equivalently, we get confidence for the the conditional survivor distribution $W(c)$ by

$$W^{\pm}(c) = \exp(-Y^{\pm}(c)), \quad c \in I. \tag{5}$$

Proof: Use Theorem II.1 and solve appropriately ■

III. Computing the distribution function of the infinity norm of a multivariate normal random vector

In this section, we apply Algorithm II.1 and construct appropriate confidenc bands for both the integrated hazard function and the distribution function together with their conditional versions. As an example, we use the multivariate normal distribution.

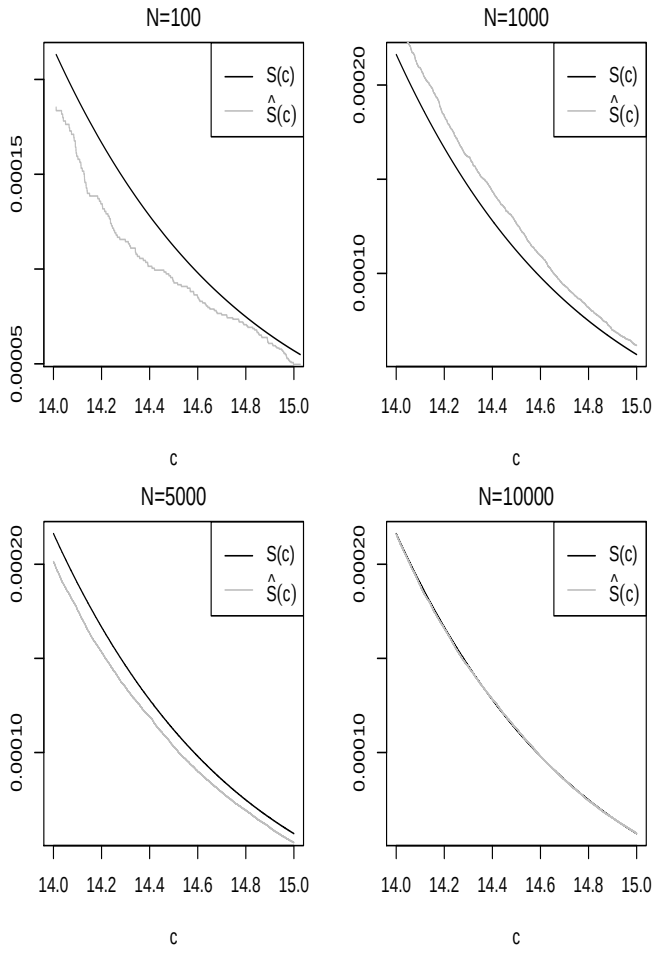


Fig. 3. Plots of the survivor function $S(c)$ and the its estimate $\hat{S}(c)$ using different sample sizes N and $c \in [14, 15]$ for covariance matrix Σ_2.

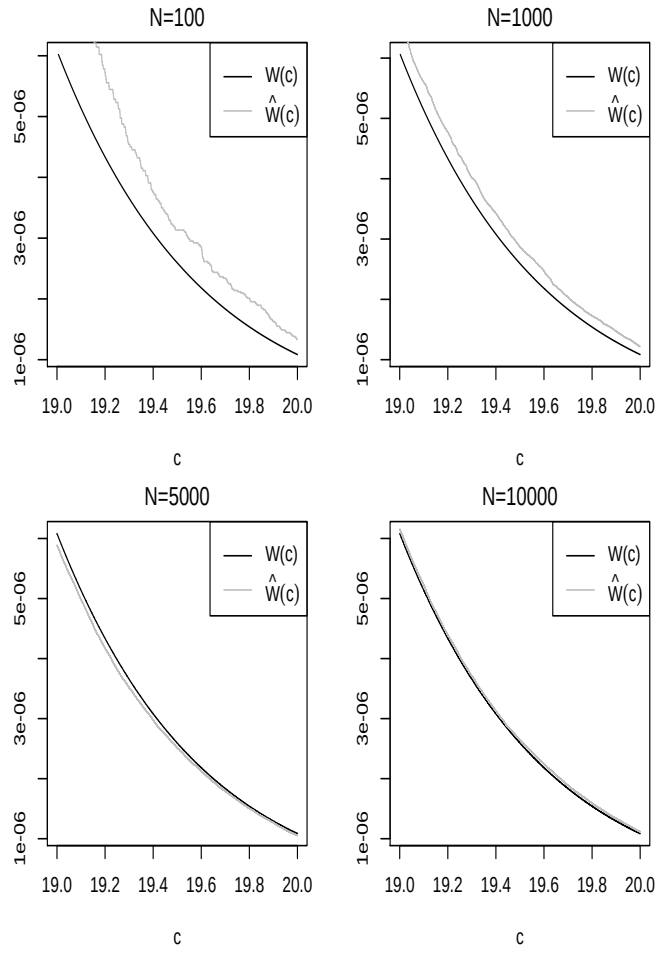


Fig. 4. Plots of the conditional survivor function $W(c)$ and the its estimate $\hat{W}(c)$ using different sample sizes N with $c_{\min} = 10$ and $c \in [19, 20]$ for covariance matrix Σ_2.

Let X denote a zero mean multivariate normal random vector with covariance matrix Σ. We want to estimate the distribution function of $\|X\|_\infty$, i.e. $\phi \equiv \|.\|_\infty$. This form satisfie all of our assumptions from Section II-B. Actually the distribution function is strictly increasing everywhere in $[0, \infty)$. Of course, we can write $X = BZ$ where $\Sigma = BB^t$ and Z is a standard multivariate normal random vector. There are several choices for the matrix B i.e. one can use the Cholesky decomposition or the singular value decomposition.

We have discussed in Section II-A several methods to regenerate new paths in Step 5 of Algorithm II.1. For the current example we use the transition kernel for the standard multivariate normal distribution from [7] with a slightly modification Suppose the chain is at state x_n. Then the proposed transition is given by

$$x_{n+1} = \frac{x_n + \sigma B Z_n}{\sqrt{1 + \sigma^2}},$$

where Z_n denotes a standard normal random vector. Then the transition kernel of the algorithm is completed by accepting the proposed transition if its infinit norm exceeds the threshold of the corresponding step.

We test this algorithm in situations where we know the explicit solution of the distribution function. We use covariance matrices Σ of the following diagonal form:

- $\Sigma_1 = \text{diag}(1, 1, 1)$,
- $\Sigma_2 = \text{diag}(7, 12, 11, 11, 12, 9, 11, 7, 10, 11)$.

Before we move to simulation results, we evaluate the ad-hoc choice for the transition parameter σ. Figure 1 plots the acceptance probability of the transition kernel for each subsequent updating step of the algorithm for different values of σ, using the covariance matrix Σ_2 and $t = 1000$ transitions. The usual rule of thumb for acceptance probabilities is around 0.3. Considering the plots, we notice that at the start, the parameter σ should be higher and gradually decreased with rarer events. We are not investigating this point further but rather choose $\sigma = 0.3$ and continue with this value onwards. We also use $t = 50$ kernel transitions.

We begin with Figures 2 - 4. These figure contain the plots of the survivor function $S(c)$ or the conditional survivor function $W(c)$ together with their corresponding estimates $\hat{S}(c)$ and $\hat{W}(c)$ respectively. For convenience, we have used certain closed intervals over their support. We have used both covariance matrices Σ_1 and Σ_2 with different samples sizes N. As the sample size increases, we notice the convergence of the estimate to the true quantity. For $N = 10000$, the estimate is mimicking the true quantity with high accuracy. But even

ISBN: 978-1-61804-240-8

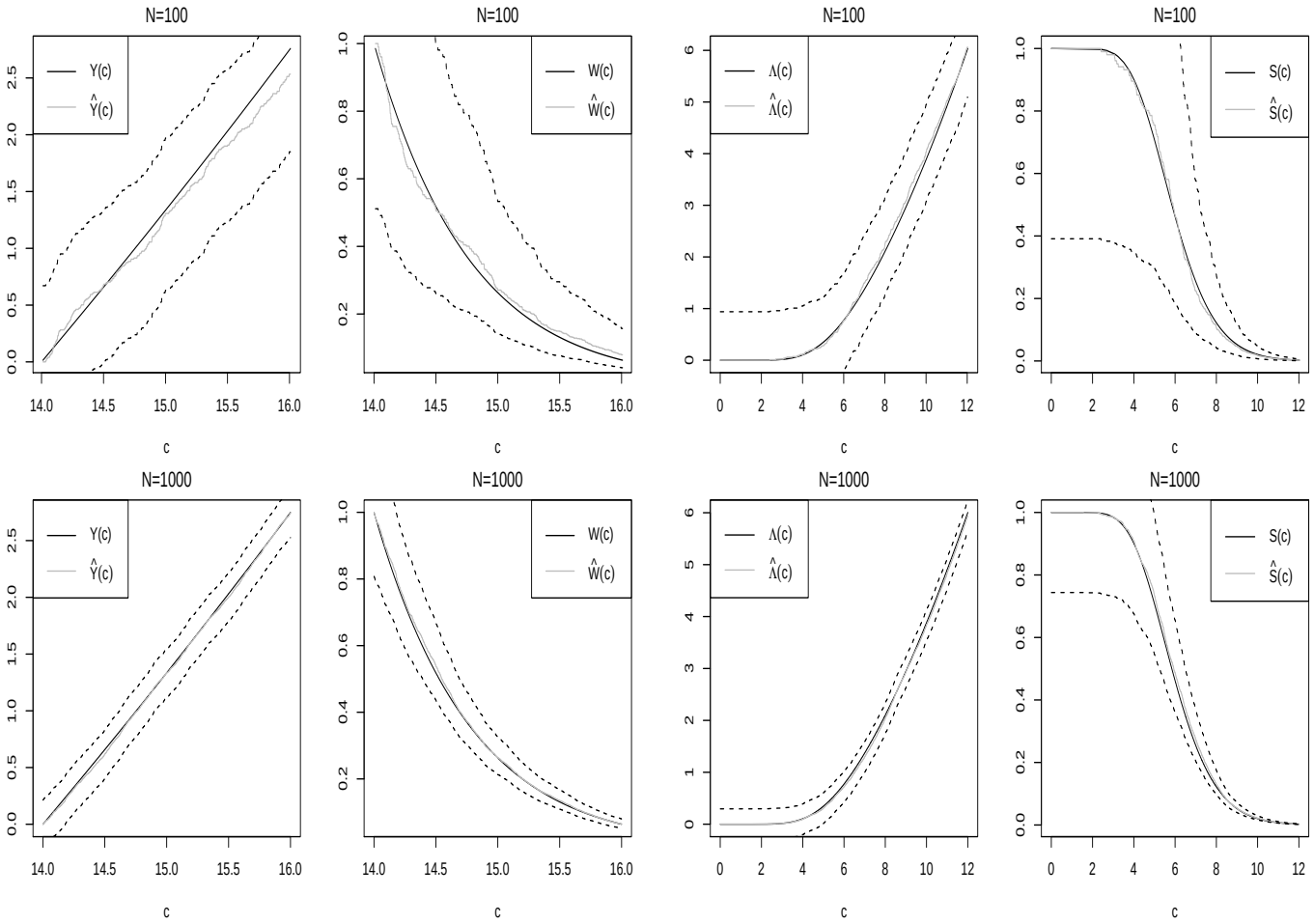


Fig. 5. Confidenc bands for the conditional integrated hazard function $Y(c)$ and the conditional survivor function $W(c)$ with covariance matrix Σ_2. We have different sample sizes N together with $c_{\min} = 14$ and $c_{\max} = 16$.

Fig. 6. Confidenc bands for the integrated hazard function $\Lambda(c)$ and the survivor function $S(c)$ for $c \in [0, 12]$ with covariance matrix Σ_2. We have used different sample sizes N.

with a choice of the very small sample size $N = 100$, the adaptive multilevel splitting estimator seems to outperform the usual empirical distribution estimator. We also notice that the estimator is most likely to be on one side of the true quantity and especially for lower sample sizes.

We continue by testing the coverage probability of the confi dence bands given in (4), i.e. we are estimating the probability that the conditional integrated hazard function $Y(c)$ lies inside the confidenc bands $Y^{\pm}(c)$, where c lies in the closed interval $I = [c_{\min}, c_{\max}]$. We are using $\alpha = 0.95$ - quantile of the distribution function of $\sup_{t\in[0,1]} |B(t)|$ which is $h_\alpha \approx 2.24$. Each estimation is based on 1000 replications using different sample sizes $N \in \{10^2, 10^3, 10^4\}$. For covariance matrix Σ_1, we have used $I = [4, 6]$ and got the coverage probabilities 0.946, 0.946 and 0.955 for $N = 10^2, 10^3, 10^4$ respectively. For the covariance matrix Σ_2, we have taken $I = [14, 20]$ and we got 0.894, 0.916 and 0.942 respectively. In both cases, the coverage probability seems to converge to the true value 0.95, giving empirical evidence to the Theorem II.1.

In the remaining Figures 5 - 8, we are plotting confidenc bands for the integrated hazard function and the survivor function or their conditional versions. Each plot is based on one run for sample sizes $N \in \{100, 1000\}$. As expected, the confidenc bands get narrower with increasing sample size. In all figures we notice that the true quantity always lies inside the confidenc bands. The confidenc bands also mimics the true quantity.

Remark 1. *One can apply Algorithm II.1 for the estimation of the multivariate t distribution. There are different versions of the t distribution. One such form is given in [12]. [6] describe this form and in their Section 2.1 construct a crude Monte Carlo estimator. Considering the form of the estimator, it can be seen, given some transition kernel for the variables, that our algorithm can be easily fit to this example.*

IV. Discussion

We have extended the results of [7] by applying adaptive multilevel splitting for the estimation of the whole distribution function and not only for the probability over some point.

A simulation study was performed, using as an example the infinit norm of a multivariate normal random vector. Confidenc bands were constructed for both the distribution function and the integrated hazard function together with their conditional versions. A test for the coverage probability of the

ISBN: 978-1-61804-240-8

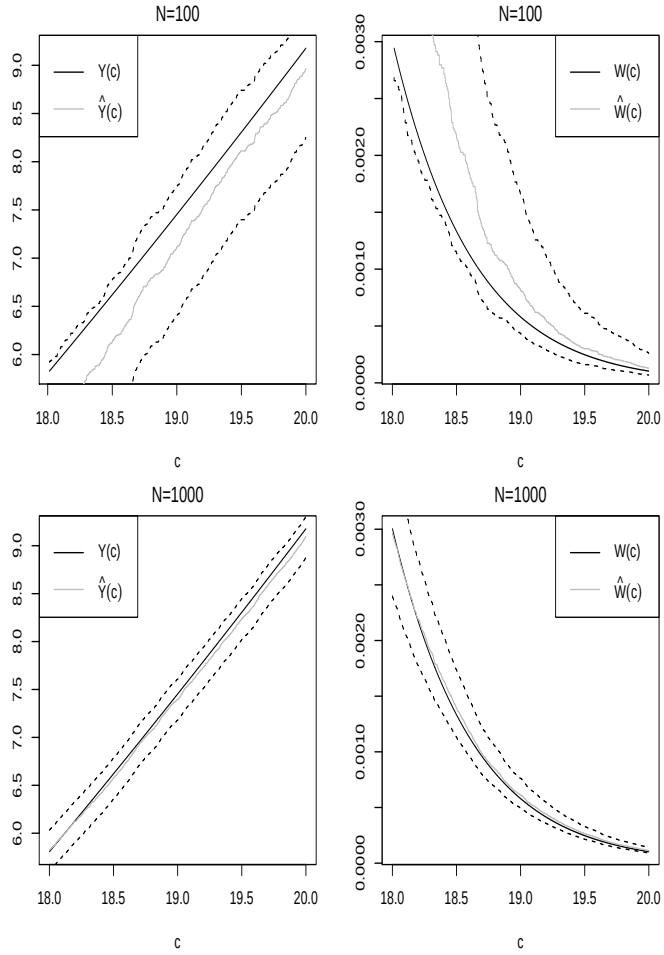


Fig. 7. Confidenc bands for the conditional integrated hazard function $Y(c)$ and the conditional survivor function $W(c)$ with covariance matrix Σ_2. We have different sample sizes N and $c_{\min} = 14$ with $c \in [18, 20]$.

conditional distribution function, showed that the theoretical results were consistent in practice.

In Theorem II.1 we got convergence in distribution of (3) using the continuous functional $\sup|.|$. Other continuous functionals can be used that result to known distributions and this is a topic for further research.

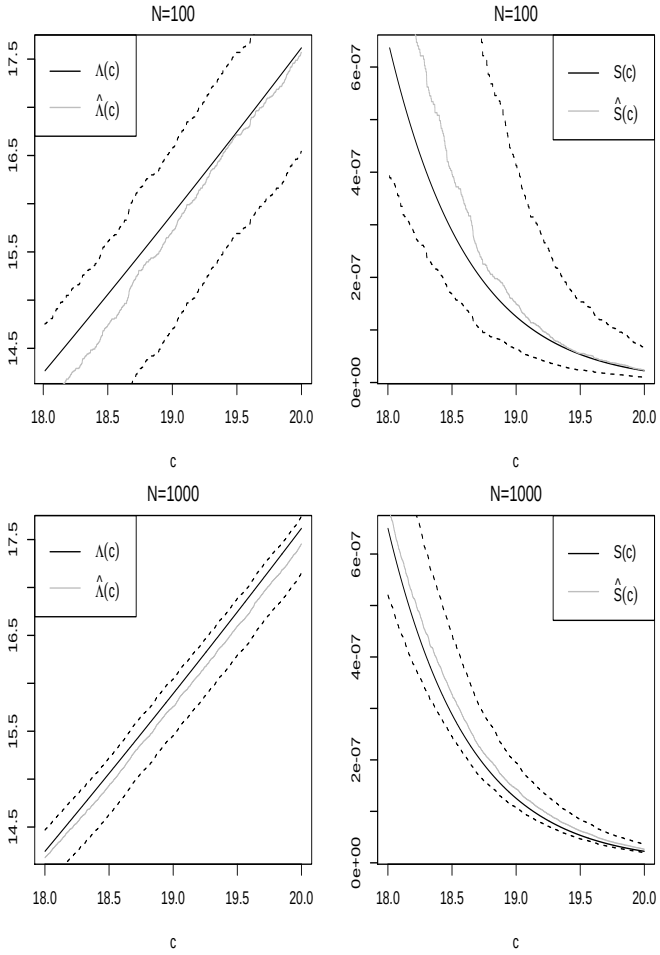


Fig. 8. Confidenc bands for the integrated hazard function $\Lambda(c)$ and the survivor function $S(c)$ for $c \in [18, 20]$ with covariance matrix Σ_2. We have used different sample sizes N.

REFERENCES

[1] Aalen, O., Borgan, Ø., and Gjessing, H.: *Survival and event history analysis: a process point of view*. Springer Verlag (2008).

[2] Billingsley, P.: *Convergence of probability measures*. Wiley New York (1999).

[3] Cérou, F., Del Moral, P., Furon, T., and Guyader, A.: Rare event simulation for a static distribution. *INRIA-00350762* (2009).

[4] Cérou, F. and Guyader, A.: Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, 25(2):417–443 (2007).

[5] Genz, A.: Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–149 (1992).

[6] Genz, A. and Bretz, F.: Comparison of methods for the computation of multivariate t probabilities. *Journal of Computational and Graphical Statistics*, 11(4):950–971 (2002).

[7] Guyader, A., Hengartner, N., and Matzner-Lber, E.: Simulation and estimation of extreme quantiles and extreme probabilities. *Preprint* (2011).

[8] Kahn, H. and Harris, T.: Estimation of particle transmission by random sampling. *National Bureau of Standards Applied Mathematics Series*, 12:27–30 (1951).

[9] Phinikettos, I. and Gandy, A.: Fast computation of high dimensional multivariate normal probabilities. *Computational Statistics and Data Analysis* (2010).

[10] Rebolledo, R.: Central limit theorems for local martingales. *Probability Theory and Related Fields*, 51(3):269–286 (1980).

[11] Rosenbluth, M. and Rosenbluth, A.: Monte Carlo calculation of the average extension of molecular chains. *The Journal of Chemical Physics*, 23(2):356–359 (1955).

[12] Tong, Y.: Multivariate normal distribution (1990).

Recovering of Secrets using the BCJR algorithm

Marcel Fernandez

Abstract—Chung, Graham and Leighton defined the guessing secrets game in [1]. In this game, player **B** has to guess a set of $c > 1$ secrets that player **A** has choosen from a set of N secrets. To unveil the secrets, player **B** is allowed to ask a series of boolean questions. For each question, **A** can adversarially select one of the secrets but once his choice is made he must answer truthfully. In this paper we present a solution to the $c = 2$ guessing secrets problem consisting in an error correcting code equipped with a tracing algorithm that, using the Bahl, Cocke, Jelinek and Raviv algorithm as its underlying routine, efficiently recovers the secrets.

Keywords—BCJR algorithm, coding theory, guessing secrets, separating codes.

I. Introduction

In the original "I've got a secret" TV game show [2] a contestant with a secret was questioned by four panelists. The questions were directed towards guessing the secret. A prize money was given to the contestant if the secret could not be guessed by the panel. In this paper we consider a variant of the game, as defined by Chung, Graham and Leighton [1]. In this variant, called "guessing secrets", there are two players **A** and **B**. Player **A** draws a subset of $c \geq 2$ secrets from a set Π of N secrets. Player **B** asks a series of questions in order discover the secrets. We will follow the approach of Alon, Guruswami, Kaufman and Sudan discussed in [3].

The game of guessing secrets is related to many different topics in communications and security such as separating systems [4], efficient delivery of Internet content [1] and the construction of schemes for the copyright protection of digital data [3]. As a matter of fact, our results can be used as a tracing algorithm for the fingerprinting code in [5].

A. Our contribution

We present a solution to the guessing secrets problem consisting in a (2,2)-separating linear block code. We also design a tracing algorithm that, from the trellis representation of the code, recovers the secrets using the Bahl, Cocke, Jelinek and Raviv (BCJR) algorithm [6] as its underlying routine. The algorithm discussed consists of several iterations of the BCJR algorithm that "corrects" (in a list decoding flavor) $\lfloor \frac{d-1}{2} \rfloor + 1$ errors, which is one more error than the error correcting bound of the code. This result might be of independent interest.

The paper is organized as follows. Section II gives an overview of the coding theory concepts used throughout the paper. Section III presents a formal description of the game of guessing secrets for the case of $c = 2$ secrets. In Section IV we show that dual binary Hamming codes give a solution to the game of guessing secrets. We present a new analysis of the Bahl, Cocke, Jelinek and Raviv algorithm in Section V. In Section VI, a tracing algorithm that allows to recover the secrets, using the BCJR algorithm as its underlying routine, is discussed. Finally, our conclusions are given in Section VII.

This work has been supported in part by the Spanish Government through project Consolider Ingenio 2007 CSD2007-00004 "ARES" and TEC2011-26491 "COPPI".

Marcel Fernandez is with the Department of Telematics Engineering. Universitat Politècnica de Catalunya. C/ Jordi Girona 1 i 3. Campus Nord, Mod C3, UPC. 08034 Barcelona. Spain.

II. Background on coding theory

A. Binary (2,2)-separating codes

In this section we give a description of binary (2,2)-separating codes.

Let $\mathbb{F}_2^n$ be the vector space over $\mathbb{F}_2$, then $C \subseteq \mathbb{F}_2^n$ is called a *code*. The field, $\mathbb{F}_2$ is called the *code alphabet*. A code C is called a *linear code* if it forms a subspace of $\mathbb{F}_2^n$. The number of nonzero coordinates in $\mathbf{x}$ is called the *weight* of $\mathbf{x}$ and is commonly denoted by $w(\mathbf{x})$. The *Hamming distance* $\mathbf{d}(\mathbf{a}, \mathbf{b})$ between two words $\mathbf{a}, \mathbf{b} \in \mathbb{F}_q^n$ is the number of positions where $\mathbf{a}$ and $\mathbf{b}$ differ. The *minimum distance* d of C, is defined as the smallest distance between two different code words. If the dimension of the subspace is k, and its minimum Hamming distance is d, then we call C an [*n,k,d*]-code. An error correcting code of minimum distance d can correct up to can correct $\lfloor \frac{d-1}{2} \rfloor$ errors.

A $(n-k) \times n$ matrix $\mathbf{H}$, is a *parity check matrix* for the code C, if C is the set of code words $\mathbf{c}$ for which $\mathbf{Hc} = \mathbf{0}$, where $\mathbf{0}$ is the all-zero $(n-k)$ tuple. Each row of the matrix is called a *parity check equation*. A code whose code words satisfy all the parity check equations of a parity check matrix is called a *parity check code*.

For any two words $\mathbf{a}$, $\mathbf{b}$ in $\mathbb{F}_q^n$ we define the *set* of *descendants* $D(\mathbf{a}, \mathbf{b})$ as $D(\mathbf{a}, \mathbf{b}) := \{x \in \mathbb{F}_q^n : x_i \in \{a_i, b_i\}, 1 \leq i \leq n\}$. For a code C, the *descendant code* C^* is defined as: $C^* := \bigcup_{\mathbf{a} \in C, \mathbf{b} \in C} D(\mathbf{a}, \mathbf{b})$.

If $\mathbf{c} \in C^*$ is a descendant of $\mathbf{a}$ and $\mathbf{b}$, then we call $\mathbf{a}$ and $\mathbf{b}$ *parents* of $\mathbf{c}$.

A code C is $(2, 2)$-*separating* [4], if for any two disjoint subsets of code words of size two, $\{\mathbf{a}, \mathbf{b}\}$ and $\{\mathbf{c}, \mathbf{d}\}$, where $\{\mathbf{a}, \mathbf{b}\} \cap \{\mathbf{c}, \mathbf{d}\} = \emptyset$, their respective sets of descendants are also disjoint, $D(\mathbf{a}, \mathbf{b}) \cap D(\mathbf{c}, \mathbf{d}) = \emptyset$.

Next corollary from [7] gives a sufficient condition for a linear code to be (2,2)-separating.

Corollary 1 ([7]): All linear, equidistant codes are (2,2)-separating.

ISBN: 978-1-61804-240-8

B. Dual binary Hamming codes

In this paper we will make extensive use of dual binary Hamming codes.

Dual binary Hammming codes are codes with parameters $[n = 2^k - 1, k, d = 2^{k-1}]$, where n represents the code length, k its dimension and d its minimum distance. Moreover, $N = 2^k$ denotes the number of code words. Dual binary Hamming codes are $(2, 2)$-separating, equidistant codes. All code words except the all zero code words have the same Hamming weight.

C. Trellis representation of block codes

The contents of this section are based on [8]. For a binary linear block code, a *trellis* is defined as a graph in which the nodes represent states, and the edges represent transitions between these states. The nodes are grouped into sets S_t, indexed by a "time" parameter t, $0 \leq t \leq n$. The parameter t indicates the *depth* of the node. The edges are unidirectional, with the direction of the edge going from the node at depth t, to the node at depth $t+1$. Each edge is labeled using an element of $\mathbb{F}_2$.

In any depth t, the number of states in the set S_t is at most $2^{(n-k)}$. The states at depth t are denoted by $\mathbf{s}_t^i$, for certain values of i, $i \in \{0, 1, \ldots, 2^{(n-k)} - 1\}$. The states will be identified by binary $(n-k)$-tuples.

In the trellis representation of a code C, each distinct path corresponds to a different code word, in which the labels of the edges in the path are precisely the code word symbols. The correspondence between paths and code words is one to one, and it is readily seen from the construction process of the trellis, that we now present.

The construction algorithm of the trellis of a linear block code, uses the fact that every code word of C must satisfy all the parity check equations imposed by the parity check matrix $\mathbf{H}$. In this case, the code words are precisely the coefficients $c_1, c_2, \ldots, c_n$ of the linear combinations of the columns $\mathbf{h}_i$ of $\mathbf{H}$, that satisfy

$$c_1\mathbf{h}_1 + c_2\mathbf{h}_2 + \cdots + c_n\mathbf{h}_n = \mathbf{0}, \tag{1}$$

where $\mathbf{0}$ is the all zero $(n-k)$-tuple.

Intuitively, the algorithm first constructs a graph, in which all linear combinations of the columns of $\mathbf{H}$ are represented by a distinct path. Then removes all paths corresponding to the linear combinations that do not not satisfy (1).

1) Initialization (depth $t = 0$):
   $S_0 = \{\mathbf{s}_0^0\}$, where $\mathbf{s}_0^0 = (0, \ldots, 0)$.
2) Iterate for each depth $t = 0, 1, \ldots, (n-1)$.
   a) Construct $S_{t+1} = \{\mathbf{s}_{t+1}^0, \ldots, \mathbf{s}_{t+1}^{|I_{t+1}|}\}$, using
   $\mathbf{s}_{t+1}^j = \mathbf{s}_t^i + c_l\mathbf{h}_{t+1}$
   $\forall i \in I_t$ and $l = 0, 1$.
   b) For every $i \in I_t$, according to 2a:
   - Draw a connecting edge between the node $\mathbf{s}_t^i$ and the 2 nodes it generates at depth $(t+1)$, according to 2a.
   - Label each edge $\theta_t^{i,j}$, with the value of $c_j \in \mathbb{F}_2$ that generated $\mathbf{s}_{t+1}^j$ from $\mathbf{s}_t^i$.

0	0	0	0	0	0	0	A
0	0	1	1	1	0	1	B
0	1	0	1	0	1	1	C
0	1	1	0	1	1	0	D
1	0	0	0	1	1	1	E
1	0	1	1	0	1	0	F
1	1	0	1	1	0	0	G
1	1	1	0	0	0	1	H

Fig. 1. The dual binary Hamming [7,3,4] code

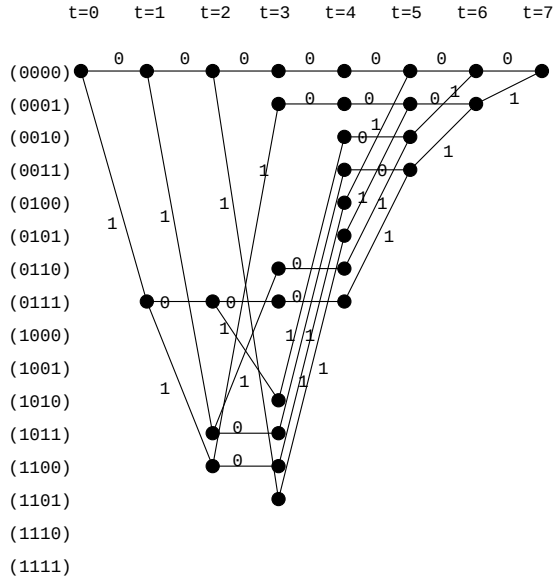


Fig. 2. Trellis for the dual binary Hamming [7,3,4] code

3) Remove all nodes that do not have a path to the all-zero state at depth n, and also remove all edges incident to these nodes.

According to the convention in 2b, for every edge $\theta_t^{i,j}$, we can define the function $\mathbf{label\_of}(\theta_t^{i,j})$ that, given a code word $\mathbf{c} = (c_1, c_2, \ldots, c_n)$, returns the c_j that generated $\mathbf{s}_{t+1}^j$ from $\mathbf{s}_t^i$

There are 2^k different paths in the trellis starting at depth 0 and ending at depth n, each path corresponding to a code word. Since the nodes (states) are generated by adding linear combinations of $(n-k)$-tuples of elements of $\mathbb{F}_2$, the number of nodes (states) at each depth is at most $2^{(n-k)}$. As an example, and because we will use it below, we take the dual binary Hamming [7,3,4] code. For this code, we show in Figures 1 and 2, the complete set of code words and the trellis representation respectively. Note that there is a one-to-one correspondence between both figures.

D. The Bahl, Cocke, Jelinek and Raviv Algorithm

We provide the basic facts of the Bahl, Cocke, Jelinek and Raviv algorithm.

Given the trellis of a code, the BCJR algorithm outputs the reliability of each symbol of the received word. More

precisely it helps to compute the *a posteriori probability* (APP) functions:

1) $P(S_t = m|\mathbf{r}_1^n)$ (**associated with each node in the trellis**) that indicates the conditional probability of being in state m at time instant t given that the received bit sequence is $\mathbf{r}_1^n$.
2) $P(S_{t-1} = m'; S_t = m|\mathbf{r}_1^n)$ (**associated with each branch in the trellis**) that indicates the joint probability of being in state m' at time $t-1$ and in state m at time t given that the received bit sequence is $\mathbf{r}_1^n$.

However it is simpler to obtain the joint probabilities

1) The function $\lambda_t(m)$ is defined as the joint probability of being in state m at time instant t and that the received bit sequence (word) is $\mathbf{r}_1^n$.

$$\lambda_t(m) = P(S_t = m; \mathbf{r}_1^n)$$

2) The function $\sigma_t(m', m)$ is defined as the joint probability of being in state m' at time instant $t-1$, and in state m at time instant t, and that the received bit sequence is $\mathbf{r}_1^n$.

$$\sigma_t(m', m) = P(S_{t-1} = m'; S_t = m; \mathbf{r}_1^n)$$

Note that since $P(\mathbf{r}_1^n) = \lambda_n(0)$ the APP probabilities are easily obtained through the following expressions

$$P(S_t = m|\mathbf{r}_1^n) = \frac{P(S_t = m; \mathbf{r}_1^n)}{P(\mathbf{r}_1^n)} = \frac{\lambda_t(m)}{\lambda_n(0)} \tag{2}$$

$$P(S_{t-1} = m'; S_t = m|\mathbf{r}_1^n) = \frac{P(S_{t-1} = m'; S_t = m; \mathbf{r}_1^n)}{P(\mathbf{r}_1^n)} = \frac{\sigma_t(m', m)}{\lambda_n(0)} \tag{3}$$

III. Guessing two secrets with binary answers

In this section we present a formal description of the game of guessing secrets for the case of $c = 2$ secrets.

In the first part of the game, player **A** draws exactly two secrets $S = \{\mathbf{s}_1, \mathbf{s}_2\}$, from a set Π of N secrets. Then, player **B** asks a series of boolean questions in order discover the secrets. For each question asked, **A** can adversarially choose a secret among the 2 secrets, but once the choice is made he must answer truthfully.

We first note that there's no way to guarantee that player **B** can learn both secrets, since if all replies are related to just one of the two secrets, then **B** cannot learn nothing about the other.

Note also, that **B** can never assert that a certain secret is one of **A**'s secrets, since **A** can always take three secrets $\{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3\}$ and answer using a majority strategy. In this case, the answer that **A** provides will be feasible for the three sets of secrets $\{\mathbf{s}_1, \mathbf{s}_2\}$, $\{\mathbf{s}_1, \mathbf{s}_3\}$ and $\{\mathbf{s}_2, \mathbf{s}_3\}$.

Using the above reasoning, we see that for a given answer we have the following possible configurations for the sets of secrets: A *star* configuration, when all pairs of secrets share a common element. A *degenerated star* configuration, when there is a single pair of secrets. And a *triangle* configuration, when there are three possible disjoint pairs secrets.

The solution for the $c = 2$ secrets problem will then consist, in finding the appropriate star or triangle configuration for a given sequence of answers. Also, we require the strategy to be *invertible* [1], which means that, given a sequence of answers, there exists an efficient algorithm capable of recovering the secrets.

A. Explicit construction of the strategy

Following the discussion in [3], we denote the questions in an oblivious strategy as a sequence $\mathcal{G}$ of n boolean functions $g_i : \{1, \ldots, N\} \to \{0, 1\}$. For a given secret $\mathbf{x}$ the sequence of answers to the questions g_i will then be $C(\mathbf{x}) = \langle g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_n(\mathbf{x})\rangle$.

Without loss of generality we suppose that $\log_2 N$ is an integer. In this case, using the binary representation for $\{1, \ldots, N\}$ we can redefine C as the mapping $C : \{0, 1\}^{\log_2 N} \to \{0, 1\}^n$. From this point of view C can be seen as an error-correcting code. From now on we will refer to a given strategy $\mathcal{G}$ using its associated code C, and to the sequence of answers to a given secret using its associated code word.

The question now is: which properties an error-correcting code must possess in order to solve the guessing secrets problem?. Depending on the sequence of answers, player **B** needs to recover a triangle or a star configuration. In either case, he can use the following strategy. Use the $N = |\Pi|$ secrets as vertices to construct a complete graph K_N. The pair of secrets $(\mathbf{s}_1, \mathbf{s}_2)$ can then be seen as an edge of K_N. Since we are considering each question as function $g_i : \{1, \ldots, N\} \to \{0, 1\}$, the answer induces a partition $\Pi = g_i^{-1}(0) \cup g_i^{-1}(1)$. If the answer of player **A** to question g_i is $a \in \{0, 1\}$ and the pair of secrets chosen by **A** is $(\mathbf{s}_1, \mathbf{s}_2)$, we have that $(\mathbf{s}_1, \mathbf{s}_2) \cap g_i^{-1}(a) \neq \emptyset$. Now player **B** can remove all edges within the subgraph of K_N spanned by $g_i^{-1}(1 - a)$. It follows that from the questions g_i $(1 \leq i \leq n)$, that **B** asks, he must be able to remove all edges until he is left with a subgraph "that contains no pair of disjoint edges" [3].

We now show how the strategy described in the previous paragraph can be accomplished using a certain code C. Let $C(\mathbf{s}_1), C(\mathbf{s}_2), C(\mathbf{s}_3)$ and $C(\mathbf{s}_4)$ be the sequence of answers associated with four distinct secrets $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$ and $\mathbf{s}_4$. Note that each sequence will correspond to a code word of C. The questions that **B** asks, should have the following property: for every two disjoint pairs of secrets, there is a question g_i that allows to rule out at least one of the pairs. This implies that there should exist at least one value i, $i \in \{1, \ldots, n\}$, called the *discriminating index* for which $C(\mathbf{s}_1)_i = C(\mathbf{s}_2)_i \neq C(\mathbf{s}_3)_i = C(\mathbf{s}_4)_i$. A code with a discriminating index for every two disjoint pairs of code words, is called a (2,2)-separating code [4] and was defined above in Section II. Moreover, such a code gives a strategy that solves the $c = 2$ guessing secrets game. Thus, we have proved the following lema.

Lemma 1: [3] A (2,2)-separating code solves the $c = 2$ guessing secrets game.

IV. Solution to the Guessing Secrets Game. Dual binary Hamming codes

We begin to discuss our work in this section.

Using Lemma 1 and the reasoning above we have the following lemma whose proof is immediate.

Lemma 2: Let N be the number of secrets in the set of secrets. Without loss of generality suppose that $N = 2^k$ for a given k. Then a $[n = 2^k - 1, k, d = 2^{k-1}]$ dual binary Hamming code solves the $c = 2$ guessing secrets game.

Remark 1: Note that for a (2,2)-separating $[n, k, d]$ code solving the $c = 2$ guessing secrets game, n is the number of questions, 2^k is the number of secrets in the set. Moreover, a descendant as defined in Section II is the sequence of answers given by player **B**, and the parents of this descendant are the secrets chosen by player **A**.

We now obtain some results on dual binary Hamming codes related to the guessing secrets problem. For lack of space we do not provide the proofs.

We first show that considering a dual binary Hamming code C, a 2-coalition Γ cannot generate any descendant that is closer (in the Hamming sense) to a code word $\mathbf{w} \in C - \Gamma$ than is to the coalition's own code words, that is

$$\min\{\mathbf{d}(\mathbf{x}, \mathbf{y}) | \mathbf{x} \in \Gamma\} \leq (\mathbf{w}, \mathbf{y}), \ \ \forall \mathbf{w} \in C - \Gamma.$$

According to Remark 1 this means that the code word of one of the secrets chosen by player **A**, will as close to the sequence of answers given by player **B** as any other code word.

Proposition 1: Let C be a dual binary Hamming $[n = 2^k - 1, k, d = 2^{k-1}]$ code. Let $\Gamma = \{\mathbf{u}, \mathbf{v}\} \subset C$ be a coalition, and let $\mathbf{y}$ be a descendant generated by Γ.

Then, we always have that

$$\mathbf{d}(\mathbf{w}, \mathbf{y}) \geq \frac{d}{2} \quad \text{and} \quad \min\{\mathbf{d}(\mathbf{x}, \mathbf{y}) | \mathbf{x} \in \Gamma\} \leq \frac{d}{2},$$

where $\mathbf{w}$ is any code word, $\mathbf{w} \in C - \Gamma$.

From Proposition 1 it follows that the worst situation is when

$$\mathbf{d}(\mathbf{y}, \mathbf{w}) = \mathbf{d}(\mathbf{y}, \mathbf{u}) = \mathbf{d}(\mathbf{y}, \mathbf{v}) = \frac{d}{2}, \tag{4}$$

for some $\mathbf{w} \in C - \Gamma$.

Note that the $(2, 2)$-separability of the dual binary Hamming codes, determines that there can only exist a single code word $\mathbf{w}$ with this property. Moreover, for this to happen, the descendant $\mathbf{y}$ must have exactly $d/2$ symbols from $\mathbf{u}$ and $d/2$ symbols from $\mathbf{v}$ where $\mathbf{u}$ and $\mathbf{v}$ are different.

Next proposition gives the necessary conditions, for (4) to be satisfied. This is precisely the triangle configuration as defined in Section III.

Proposition 2: Let C be a dual binary $[n = 2^k - 1, k, d = 2^{k-1}]$ Hamming code. Let $\Gamma = \{\mathbf{u}, \mathbf{v}\} \subset C$ be a coalition and let $\mathbf{y}$ be a descendant generated by the coalition Γ. Then

$$\mathbf{d}(\mathbf{y}, \mathbf{w}) = \mathbf{d}(\mathbf{y}, \mathbf{u}) = \mathbf{d}(\mathbf{y}, \mathbf{v}) = \frac{d}{2}, \tag{5}$$

only if

$$\mathbf{d}(\mathbf{y}, \mathbf{u}) = \mathbf{d}(\mathbf{y}, \mathbf{v}) = \frac{d}{2},$$

where $\mathbf{u}$ and $\mathbf{v}$ are different, and therefore the Hamming weight of the descendant $\mathbf{y}$, denoted by $w(\mathbf{y})$, satisfies $w(\mathbf{y}) \bmod 2 = 0$.

Note that in this case, in order to recover the secrets, we exceed the correcting capacity of the code and we will have to use tailor made decoding algorithms such as the ones discussed below in Section VI.

V. The Bahl, Cocke, Jelinek and Raviv Algorithm

We continue to present our work in this section where provide a new analysis of the Bahl, Cocke, Jelinek and Raviv algorithm described in Section II-D.

A. Computation of the joint probability function $\sigma_t(m', m)$

For our purposes we will need the $\sigma_t(m', m)$ function. From the definitions in Section II-D, and the analysis in the Appendix, the $\sigma_t(m', m)$ function *gives information about the transitions (symbols) in the trellis at time t.* This transition information can be obtained by checking what happens in the trellis **before**, **after** and **at** time t. Intuitively σ *gives us the probability of a given symbol in a given position.*

Therefore, if t represents each position in the code, $1 \leq t \leq n$, we have that

Symbol 0

$$\sum_{\text{branch}(m,m')=0} \sigma_t(m, m') = \frac{1}{2^k} \sum_{\mathbf{c}:c_t=0} P(r_1|c_1) \cdots P(r_t|c_t = 0) \cdots P(r_n|c_n) \tag{6}$$

Symbol 1

$$\sum_{\text{branch}(m,m')=1} \sigma_t(m, m') = \frac{1}{2^k} \sum_{\mathbf{c}:c_t=1} P(r_1|c_1) \cdots P(r_t|c_t = 1) \cdots P(r_n|c_n) \tag{7}$$

Remark 2: We also note, that a for a given position, say $t = j$, a code word indicates us to make a decision for a given symbol if this symbol is precisely the label of the edge of the path in the trellis corresponding to this code word in $t = j$.

ISBN: 978-1-61804-240-8

B. Correcting 1 error

As a warmup, first suppose we are transmitting the all-zero code word through a noisy channel, and that one error occurs in the first bit, so $\mathbf{r}_1^t = (1,0,0,0,0,0,0)$.

Since the code is equidistant $d = 4$, by comparing $\mathbf{r}_1^t = (1,0,0,0,0,0,0)$ with each code word (see Figure 1) we have that there is 1 code word at exactly distance 1 (A), 4 code words at distance 3 (E, F, G, H) and 3 code words at distance 5 (B, C, D).

We will use the following notation. Let $\mathbf{r}_1^n = (r_1, r_2, \ldots, r_7)$ be the '*received*' word. By using the labels in Figure 1 and Figure 2, we can express $P(\mathbf{r}_1^n|A)$ as

$$P(\mathbf{r}_1^n|A) := P(r_1|0)P(r_2|0)P(r_3|0)\cdots P(r_6|0)P(r_7|0)$$

Without loss of generality, we now turn our attention to the $4th$ position. Then, for $t = 4$, we can express (6) and (7) as:

Symbol 0

$$\sigma_4(0,0) + \sigma_4(1,1) + \sigma_4(6,6) + \sigma_4(7,7) = \frac{1}{8}(P(\mathbf{r}_1^n|A) + P(\mathbf{r}_1^n|D) + P(\mathbf{r}_1^n|E) + P(\mathbf{r}_1^n|H))$$

Symbol 1

$$\sigma_4(0,2) + \sigma_4(11,3) + \sigma_4(12,4) + \sigma_4(13,5) = \frac{1}{8}(P(\mathbf{r}_1^n|B) + P(\mathbf{r}_1^n|C) + P(\mathbf{r}_1^n|F) + P(\mathbf{r}_1^n|G))$$

In this position, we have that code words A, D, E, H point us towards making a decision in favor of symbol '0' whereas code words B, C, F, G lead us to decide in favor of symbol '1'. Since we are assuming that we have transmitted the all-zero code word and that we have received $\mathbf{r}_1^t = (1,0,0,0,0,0,0)$, we note that:

1) The closest code word to $\mathbf{r}_1^t = (1,0,0,0,0,0,0)$, which is code word A leads us to decide in favor of a '0', that intuitively says that there is NOT an error in this position. The rest of the code words confirm this intuition.
2) The code words that indicate us to go for a '0' are the ones that the label of the edges that the path of these code words pass through is a '0'.

Now let

$$P(r_i|s) = \begin{cases} \frac{1+\epsilon}{2} & \text{if } r_i = s \\ \frac{1-\epsilon}{2} & \text{if } r_i \neq s \end{cases} \quad (8)$$

This is a reasonable assumption, since in fact we are saying that there is a higher probability that the received symbol is the one that has been sent.

By taking $\mathbf{r}_1^t = (1,0,0,0,0,0,0)$ and using (8), the $\sigma_4(m', m)$ expressions are:

Symbol 0

$$\begin{aligned}&\sigma_4(0,0) + \sigma_4(1,1) + \sigma_4(6,6) + \sigma_4(7,7) = \\ &\frac{1}{8}P(r_1|0)P(r_2|0)P(r_3|0)\mathbf{P(r_4|0)}P(r_5|0)P(r_6|0)P(r_7|0) \\ &+ \frac{1}{8}P(r_1|1)P(r_2|0)P(r_3|0)\mathbf{P(r_4|0)}P(r_5|1)P(r_6|1)P(r_7|1) \\ &+ \frac{1}{8}P(r_1|0)P(r_2|1)P(r_3|1)\mathbf{P(r_4|0)}P(r_5|1)P(r_6|1)P(r_7|0) \\ &+ \frac{1}{8}P(r_1|1)P(r_2|1)P(r_3|1)\mathbf{P(r_4|0)}P(r_5|0)P(r_6|0)P(r_7|1) \\ &\overset{(8)}{=} \frac{1}{8}\frac{1}{2^7}((1+\epsilon)^6(1-\epsilon) + 2(1+\epsilon)^4(1-\epsilon)^3 + (1+\epsilon)^2(1-\epsilon)^5)\end{aligned}$$

Symbol 1

$$\begin{aligned}&\sigma_4(0,2) + \sigma_4(11,3) + \sigma_4(12,4) + \sigma_4(13,5) = \\ &\frac{1}{8}P(r_1|1)P(r_2|0)P(r_3|1)\mathbf{P(r_4|1)}P(r_5|0)P(r_6|1)P(r_7|0) \\ &+ \frac{1}{8}P(r_1|0)P(r_2|0)P(r_3|1)\mathbf{P(r_4|1)}P(r_5|1)P(r_6|0)P(r_7|1) \\ &+ \frac{1}{8}P(r_1|1)P(r_2|1)P(r_3|0)\mathbf{P(r_4|1)}P(r_5|1)P(r_6|0)P(r_7|0) \\ &+ \frac{1}{8}P(r_1|0)P(r_2|1)P(r_3|0)\mathbf{P(r_4|1)}P(r_5|0)P(r_6|1)P(r_7|1) \\ &\overset{(8)}{=} \frac{1}{8}\frac{1}{2^7}\left(2(1+\epsilon)^4(1-\epsilon)^3 + 2(1+\epsilon)^2(1-\epsilon)^5\right)\end{aligned}$$

Therefore,

$$\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=4} = \frac{\sigma_4(0,0) + \sigma_4(1,1) + \sigma_4(6,6) + \sigma_4(7,7)}{\sigma_4(0,2) + \sigma_4(11,3) + \sigma_4(12,4) + \sigma_4(13,5)} = \frac{1+\epsilon^2}{(1-\epsilon)^2} \geq 1 \quad \text{for} \quad 0 \leq \epsilon \leq 1 \quad (9)$$

So the algorithm points us to the **correct** decision of a '0' in the 4th position. Note that the same reasoning applies to positions $2, 3, 5, 6, 7$.

We now evaluate the $1st$ position, which is the position where the error has occurred.

Symbol 0

$$\sigma_1(0,0) = \frac{1}{8}(P(\mathbf{r}_1^n|A) + P(\mathbf{r}_1^n|B) + P(\mathbf{r}_1^n|C) + P(\mathbf{r}_1^n|D))$$

Symbol 1

$$\sigma_1(0,7) = \frac{1}{8}(P(\mathbf{r}_1^n|E) + P(\mathbf{r}_1^n|F) + P(\mathbf{r}_1^n|G) + P(\mathbf{r}_1^n|H))$$

Again, since we are assuming that we have transmitted the all-zero code word and that we have received $\mathbf{r}_1^t = (1,0,0,0,0,0,0)$, we note that A, the closest code word to $\mathbf{r}_1^t = (1,0,0,0,0,0,0)$, leads to a decision in favor of a '0'.

By taking $\mathbf{r}_1^t = (1,0,0,0,0,0,0)$ and using (8) again, the $\sigma_1(m', m)$ expressions are:

Symbol 0

$$\sigma_1(0,0) \overset{(8)}{=} \frac{1}{8}\frac{1}{2^7}\left((1+\epsilon)^6(1-\epsilon) + 3(1+\epsilon)^2(1-\epsilon)^5\right)$$

Symbol 1

$$\sigma_1(0,7) \overset{(8)}{=} \frac{1}{8}\frac{1}{2^7}4(1+\epsilon)^4(1-\epsilon)^3$$

Therefore,

$$\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=1} = \frac{\sigma_1(0,0)}{\sigma_1(0,7)} = \frac{1-2\epsilon+6\epsilon^2-2\epsilon^3+\epsilon^4}{(-1+\epsilon)^2(1+\epsilon)^2}$$
$$\geq 1 \quad \text{for} \quad 0.3 \leq \epsilon \leq 1 \tag{10}$$

So the algorithm points us to the **correct** decision of a '0' in the 1st position, and the error can be corrected.

C. Correcting beyond the error correcting bound. Identifying the parents of a descendant

In this section we discuss a key property of the BCJR algorithm. This property is essential for the results in this paper. Intuitively, this property consists of the following fact. Suppose we run the BCJR algorithm using as input a descendant $\mathbf{z}$ of a certain coalition $\{\mathbf{u}, \mathbf{v}\}$. In the symbols of the descendant $\mathbf{z}$ where the parents $\mathbf{u}$ and $\mathbf{v}$ agree the BCJR algorithm returns a higher reliability.

This is better illustrated with an example. Suppose now that we have the descendant $\mathbf{r}_2^t = (1,1,0,0,0,0,0)$. Note that $\mathbf{r}_2^t$ can only be generated by the coalitions of code words $\{A, G\}$, $\{A, H\}$ and $\{G, H\}$ (see Figure 1). Note also that in the 6th position all three code words A, G, H have a 0. On the other hand in the 4th position A and H have a 0 and G has a 1. We will see that for the 0 in the 6th position the BCJR algorithm outputs a higher reliability than for the 0 in the 4th position. In other words we will prove the following proposition.

Proposition 3: Let $\Gamma = \{\mathbf{u}, \mathbf{v}\}$ be two code words of a dual binary Hamming code. Let $\mathbf{z}$ be a descendant created by $\{\mathbf{u}, \mathbf{v}\}$. Then:

1) For a star configuration (see Section III), the output reliabilities of the symbols of $\mathbf{z}$ given by the BCJR algorithm will correspond to $\mathbf{u}$ if $\mathbf{d}(\mathbf{u}, \mathbf{z}) \leq \frac{d}{2} - 1$ and to $\mathbf{v}$ otherwise.
2) For a degenerated star configuration (see Section III), the output reliabilities of the symbols of $\mathbf{z}$ given by the BCJR algorithm will be larger for symbols in which $\mathbf{u}$ and $\mathbf{v}$ agree, than in symbols where they differ.
3) In case another code word say $\mathbf{w}$ forms a triangle configuration with $\mathbf{u}$ and $\mathbf{v}$ (see Section III) then the output reliabilities of the symbols of $\mathbf{z}$ given by the BCJR algorithm are largest in the positions where $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ agree.

For clarity, we prove Proposition 3 using an example.

1) Example. Proof of Proposition 3: Since the code is equidistant $d = 4$. By by comparing $\mathbf{r}_2^t = (1,1,0,0,0,0,0)$ with each code word we have that there are 3 code words at exactly distance 2 (A, G, H), 4 code words at distance 4 (C, D, E, F) and 1 code word at distance 6 (B)

Again without loss of generality, we turn our attention to the $4th$ position. We have seen that in this position

Symbol 0

$$\sigma_4(0,0) + \sigma_4(1,1) + \sigma_4(6,6) + \sigma_4(7,7) = \frac{1}{8}(P(\mathbf{r}_1^n|A) + P(\mathbf{r}_1^n|D) + P(\mathbf{r}_1^n|E) + P(\mathbf{r}_1^n|H))$$

Symbol 1

$$\sigma_4(0,2) + \sigma_4(11,3) + \sigma_4(12,4) + \sigma_4(13,5) = \frac{1}{8}(P(\mathbf{r}_1^n|B) + P(\mathbf{r}_1^n|C) + P(\mathbf{r}_1^n|F) + P(\mathbf{r}_1^n|G))$$

We have that code words A, D, E, H point us towards making a decision in favor of symbol '0' whereas code words B, C, F, G lead us to decide in favor of symbol '1'. Since we are assuming that the descendant is $\mathbf{r}_2^t = (1,1,0,0,0,0,0)$:

1) The closest code words to $\mathbf{r}_2^t = (1,1,0,0,0,0,0)$ are A, G, H. A and H lead us to decide in favor of a '0' while G leads us towards a '1'. This intuitively says that the symbol in this position should be a '0'.
2) The other 2 code words that indicate us to make a decision for a '0', are D, E. They are at distance 4 of $\mathbf{r}_2^t = (1,1,0,0,0,0,0)$, which is the minimum distance of the code.

By taking $\mathbf{r}_2^t = (1,1,0,0,0,0,0)$ and again using (8), the $\sigma_4(m', m)$ expressions are:

Symbol 0

$$\sigma_4(0,0) + \sigma_4(1,1) + \sigma_4(6,6) + \sigma_4(7,7) = \overset{(8)}{=} \frac{1}{8}\frac{1}{2^7}\left(2(1+\epsilon)^5(1-\epsilon)^2 + 2(1+\epsilon)^3(1-\epsilon)^4\right)$$

Symbol 1

$$\sigma_4(0,2) + \sigma_4(11,3) + \sigma_4(12,4) + \sigma_4(13,5) = \overset{(8)}{=} \frac{1}{8}\frac{1}{2^7}\left(2(1+\epsilon)^3(1-\epsilon)^4 + (1+\epsilon)(1-\epsilon)^6 + (1+\epsilon)^5(1-\epsilon)^2\right)$$

Therefore,

$$\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=4} = \frac{\sigma_4(0,0) + \sigma_4(1,1) + \sigma_4(6,6) + \sigma_4(7,7)}{\sigma_4(0,2) + \sigma_4(11,3) + \sigma_4(12,4) + \sigma_4(13,5)} = \frac{(1+\epsilon)^2}{1+\epsilon^2} > 1 \tag{11}$$

So the algorithm points us for a decision of a '0' in the 4th position. Note that the same reasoning applies to positions $1, 2, 3, 5, 7$.

We now evaluate the $6th$ position (which is the only position where the 3 parents A, G, H have the same symbol ('0')).

Symbol 0

$$\sigma_6(0,0) + \sigma_6(1,1) = \frac{1}{8}(P(\mathbf{r}_1^n|A) + P(\mathbf{r}_1^n|B) + P(\mathbf{r}_1^n|G) + P(\mathbf{r}_1^n|H))$$

Symbol 1

$$\sigma_6(2,0) + \sigma_6(3,1) = \frac{1}{8}(P(\mathbf{r}_1^n|C) + P(\mathbf{r}_1^n|D) + P(\mathbf{r}_1^n|E) + P(\mathbf{r}_1^n|F))$$

In this position, we have that code words A, B, G, H point us towards making a decision in favor of symbol '0' whereas code words C, D, E, F lead us to decide in favor of symbol '1'. Again, since we are assuming that the descendant is $\mathbf{r}_2^t = (1,1,0,0,0,0,0)$, we note that:

1) Now A, G, H, the closest code words to $\mathbf{r}_2^t = (1,1,0,0,0,0,0)$, lead to a decision in favor of a '0'. Intuitively this indicates that the symbol in this position is '*strong*'.
2) The other code word that indicates a '0' decision is B, which is the code word at a larger distance from $\mathbf{r}_2^t = (1,1,0,0,0,0,0)$. This is a consequence of the following:
We start with code words A and G

0	0	0	0	0	0	0	A
1	1	0	1	1	0	0	G
					0		(*)
					0		(**)

There exists a code word with a '0' in the 6th position at distance 4 from both A and G. Since A and G disagree in 4 positions and the 6th position is fixed, the only possibility is that this code word is different from A and G where A and G agree, and in the remaining 4 positions, agrees in two of them with A and in two of them with G. We take H to be this code word.

0	0	0	0	0	0	0	A
1	1	0	1	1	0	0	G
1	1	1	0	0	0	1	H
					0		(**)

Now the also must exist a code word with a '0' in the 6th position at distance 4 from A, G and H. Since again the 6th position is fixed, there 6 available positions and this code word must agree in two of them with A (and be different from A in the remaining 4), in two of them with B (and be different from B in the remaining 4) and in two of them with C(and again be different from C in the remaining 4). This is the same as saying that it must be equal to A in the positions in with the symbol of A is the minority of the symbols of A, G and H.

0	**0**	0	0	0	0	0	A
1	1	0	**1**	**1**	0	0	G
1	1	**1**	0	0	0	**1**	H
0	**0**	**1**	**1**	**1**	0	**1**	B
1	1	0	0	0	0	0	$\mathbf{r}_2^t$

Summarizing, for the $[7,3,4]$ dual binary Hamming code a descendant $\mathbf{r}_2^t = (1,1,0,0,0,0,0)$ contains an equal number of symbols from both A and G, and forms a triangle configuration with code word H. Moreover, if we take A, G and H then the descendant can be seen as constructed according to a *majority* decision. Since we found B according to a *minority* decision, $\mathbf{r}_2^t = (1,1,0,0,0,0,0)$ and B **only agree in one position**.

By taking $\mathbf{r}_2^t = (1,1,0,0,0,0,0)$ and using (8) again, the $\sigma_6(m',m)$ expressions are:

Symbol 0

$$\sigma_6(0,0)+\sigma_6(1,1) \overset{(8)}{=} \frac{1}{8}\frac{1}{2^7}\left((1+\epsilon)(1-\epsilon)^6+3(1+\epsilon)^5(1-\epsilon)^2\right)$$

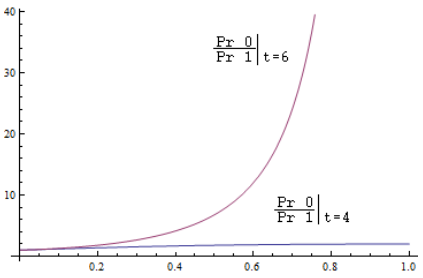


Fig. 3. Symbol reliabilities

Symbol 1

$$\sigma_6(2,0)+\sigma_6(3,1) \overset{(8)}{=} \frac{1}{8}\frac{1}{2^7}4(1+\epsilon)^3(1-\epsilon)^4$$

Therefore,

$$\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=6} = \frac{\sigma_1(0,0)}{\sigma_1(0,7)} = \frac{1+2\epsilon+6\epsilon^2+2\epsilon^3+\epsilon^4}{(-1+\epsilon)^2(1+\epsilon)^2} > 1 \quad \text{for } \ 0 \le \epsilon \le 1 \tag{12}$$

Finally, we compare (11) and(12). We note that always

$$\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=6} > \left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=4} \tag{13}$$

This can also be shown in Figure 3. Therefore, the BCJR algorithm returns a larger reliability for a symbol in a position in which the parents of a descendant agree. This proves Proposition 3.

VI. Efficient recovery of the secrets

We now approach the problem of how to efficiently recover the secrets, when the strategy used is a dual binary Hamming code. To recover the secrets we first need a way to relate the word associated to a sequence of answers given by $\mathbf{A}$, with the code words corresponding to these secrets. This was done in Remark 1. Now, if we denote by $\mathbf{z}$ the word corresponding to the sequence of answers given by player $\mathbf{A}$, then according to Section III, Proposition 1 and Proposition 2 we have that:

1) In a star configuration, for the common secret, say $\mathbf{u}$, we have that $\mathbf{d}(\mathbf{u},\mathbf{z}) \le \frac{d}{2} - 1$.
2) In a "degenerated" star configuration, for the single pair of secrets, say $\{\mathbf{u},\mathbf{v}\}$, we have that $\mathbf{d}(\mathbf{u},\mathbf{z}) = \mathbf{d}(\mathbf{v},\mathbf{z}) = \frac{d}{2}$.
3) In a triangle configuration, for the three possible pairs of secrets, say $\{\mathbf{u},\mathbf{v}\}$, $\{\mathbf{u},\mathbf{w}\}$ and $\{\mathbf{v},\mathbf{w}\}$, we have that $\mathbf{d}(\mathbf{u},\mathbf{z}) = \mathbf{d}(\mathbf{v},\mathbf{z}) = \mathbf{d}(\mathbf{w},\mathbf{z}) = \frac{d}{2}$.

Therefore, we need an algorithm that outputs all code words of a (2,2)-separating code within distance $d/2$ of $\mathbf{z}$. Since the error correcting bound of the code is $\lfloor \frac{d-1}{2} \rfloor$ we have that in both cases, "degenerated" star and triangle,

we need to correct one more than the error correcting bound of the code. As it is shown below, this can be done by modifying the Bahl, Cocke, Jelinek and Raviv algorithm.

A. Recovering secrets with the BCJR algorithm

In this section we use Proposition 3, to efficiently recover secrets using the BCJR algorithm.

We first give an intuitive description of the algorithm. Recall that given a sequence of answers $\mathbf{z}$ we need to find, either the unique code word at a distance less or equal than $\frac{d}{2}-1$ of $\mathbf{z}$, or the code word, or the two or three code words at a distance $\frac{d}{2}$ of $\mathbf{z}$.

Let $\mathbf{z}=(z_1,z_2,\ldots,z_n)$ be a descendant. We run the BCJR algorithm with input $\mathbf{z}$. Let $\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}$, $1\leq j\leq n$, the output of the BCJR algorithm. According to Proposition 3 we know that in the positions where the parents of $\mathbf{z}$ agree we will obtain a larger reliability towards a given symbol. Since we wish to obtain the parents of $\mathbf{z}$ we set the probability of these symbols to 1. In the remaining positions the symbols of the parents differ from each other. Note that if in one of these positions we set the probability of one of the symbols to 1, we will make the descendant 'closer' in Hamming sense to one of the parents. Running again the BCJR algorithm with these modified probabilities will yield this parent. Once we obtain a parent, we search for a position in which this parent and the descendant are different. By setting, in this position, the probability of the symbol of the descendant to 1, and running again the BCJR algorithm we will obtain another parent.

In the following algorithm we will have occasion to use the following rules.

Rule 1:

$$\text{Symbol}|_{t=j}=\begin{cases}0 & \text{if } \left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}>1\\ 1 & \text{if } \left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}<1\end{cases}$$

Rule 2:

1) $Pr(r_j|c_j=0)=1$ and $Pr(r_j|c_j=1)=0$ if $\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}>1$
2) $Pr(r_j|c_j=0)=0$ and $Pr(r_j|c_j=1)=1$ if $\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}<1$

Tracing BCJR Algorithm. (TBCJRA)

Input:

Dual binary Hammming code $[n=2^k-1,k,d=2^{k-1}]$

A descendant $\mathbf{z}=(z_1,z_2,\ldots,z_n)$.

Output:

A list L containing the parents of $\mathbf{z}$

Initialization: $L:=\{\emptyset\}$

1) *First Steps*: Run the BCJR algorithm using $\mathbf{z}$ and obtain $\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}$, $1\leq j\leq n$.
   a) $\mathbf{t}_1:=$ Apply *Rule 1* to $\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}$, $1\leq j\leq n$
   b) if $\mathrm{d}(\mathbf{t}_1,\mathbf{z})<d/2$ add $\mathbf{t}_1$ to L and exit.
   c) else go to *Iteration*
2) *Iteration*:
   a) Find the positions $\{j_1,\ldots,j_s\}$ in which the values of $\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}$, $1\leq j\leq n$ are maximum.
   b) Apply *Rule 2* to these positions and run the BCJR algoritm to obtain $\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}$, $1\leq j\leq n$, $j\notin\{j_1,\ldots,j_s\}$.
   c) $\mathbf{t}_2:=$ Apply *Rule 1* to $\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}$, $1\leq j\leq n$
   d) add $\mathbf{t}_2$ to L
   e) Find a position j_l, $j_l\notin\{j_1,\ldots,j_s\}$, in which $\mathbf{t}_2$ and $\mathbf{z}$ disagree and set $Pr(z_{j_l}|c_{j_l}=z_{j_l})=1$ and $Pr(z_{j_l}|c_{j_l}!=z_{j_l})=0$
   f) Run the BCJR algoritm to obtain $\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}$, $1\leq j\leq n$, $j\notin\{j_1,\ldots,j_s\}\cup j_l$.
   g) $\mathbf{t}_3:=$ Apply *Rule 1* to $\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}$, $1\leq j\leq n$
   h) add $\mathbf{t}_3$ to L
   i) Find a position j_m, $j_m\notin\{j_1,\ldots,j_s\}\cup j_l$, in which $\mathbf{t}_3$ and $\mathbf{z}$ disagree and set $Pr(z_{j_m}|c_{j_m}=z_{j_m})=1$ and $Pr(z_{j_m}|c_{j_m}!=z_{j_m})=0$
   j) Run the BCJR algoritm to obtain $\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}$, $1\leq j\leq n$, $j\notin\{j_1,\ldots,j_s\}\cup j_l$.
   k) $\mathbf{t}_4:=$ Apply *Rule 1* to $\left.\frac{Pr\ 0}{Pr\ 1}\right|_{t=j}$, $1\leq j\leq n$
   l) if $\mathbf{t}_4$ is different from both $\mathbf{t}_2$ and $\mathbf{t}_3$, then add $\mathbf{t}_4$ to L
   m) Output L and exit

B. Correctness of the algorithm

We have to show that the code words in the list L are the parents of the descendant $\mathbf{z}$. We first show the correctness of Step 1b. If the output at Step 1b is not empty is because all "errors" in the descendant have been corrected. This means that one of the parents was at a distance less than $d/2-1$ from the descendant. By Proposition 1 and Proposition 3 this is the only traceable parent, and this parent is precisely $\mathbf{t}_1$.

We now show that the output at Step 2m, contains all the parents of $\mathbf{z}$. In this case, this is true again from Proposition 3 and Proposition 2. If the configuration is a degenerated star configuration then $\mathbf{t}_2$ and $\mathbf{t}_3$ are the parents of $\mathbf{z}$. In the remaining case $\mathbf{t}_2$, $\mathbf{t}_3$ and $\mathbf{t}_4$ form a triangle.

VII. Conclusions

In this paper we present an explicit set of questions that solves the $c=2$ guessing secrets problem together with

an efficient algorithm to recover the secrets. The explicit set of questions is based on a dual binary Hamming code. The recovery of the secrets consists in the decoding of a block code beyond its error correction bound. In order to perform this decoding efficiently we present a modification of the BCJR algorithm, that passing through the trellis representing the block code, returns all the code words of the code within distance $d/2$ of a given word.

Appendix

Given a trellis, to compute the joint probability functions $\lambda_t(m)$ and $\sigma_t(m', m)$ the following auxiliary functions are used:

1) The *before* function $\alpha_t(m)$

$$\alpha_t(m) = P(S_t = m; \mathbf{r}_1^t)$$

that denotes the joint probability of being in state m at time instant t and that the received bit sequence *before* (up to) time t is $\mathbf{r}_1^t$.
2) The *after* function $\beta_t(m)$

$$\beta_t(m) = P(\mathbf{r}_{t+1}^n | S_t = m)$$

that denotes the probability of receiving the bit sequence $\mathbf{r}_{t+1}^n$ *after* time t conditioned on being in state m at time t.
3) The *transition* function $\gamma_t(m', m)$

$$\gamma_t(m', m) = P(S_t = m; r_t | S_{t-1} = m')$$

that denotes the joint probability of being at state m at time t and that the received bit at time t is r_t conditioned on being in state m' at time $t-1$.

A. Computation of the auxiliary functions $\alpha_t(m)$, $\beta_t(m)$ and $\gamma_t(m', m)$

1) The transition function $\gamma_t(m', m)$: In the most general case the *transition* function $\gamma_t(m', m)$, is computed according to the following expression

$$\gamma_t(m', m) = \sum_{x \in \mathbb{F}_2} P(S_t = m | S_{t-1} = m') \cdot P(x | S_{t-1} = m', S_t = m) \cdot P(r_t | x)$$

where

- $P(S_t = m | S_{t-1} = m')$ is the probability of being in state m at time t given that the state at time $t-1$ is m'.
- $P(x | S_{t-1} = m', S_t = m)$ is the probability that the *code word* symbol is x given that the transition is from state m' at time $t-1$ to state m at time t.
- $P(r_t | x)$ is the transition probability of the discrete memoryless channel, that is, the probability that the symbol at the symbol at the channel output is r_t given that the symbol at the input is x.

2) The before function $\alpha_t(m)$: The function $\alpha_t(m) = P(S_t = m; \mathbf{r}_1^t)$ denotes the joint probability of being in state m at time instant t and that the received bit sequence *before* (up to) time t is $\mathbf{r}_1^t$.

Therefore,

$$\alpha_t(m) = \sum_{m'} \alpha_{t-1}(m') \cdot \gamma_t(m', m)$$

Since we will always assume that at time $t = 0$ the state of the trellis is $m = 0$, that is $S_0 = 0$, then the boundary conditions on *alpha* are

$$\begin{aligned} \alpha_0(0) &= 1 \\ \alpha_0(m) &= 0 \quad \text{for } m \neq 0 \end{aligned} \tag{14}$$

3) The after function $\beta_t(m)$: The *after* function $\beta_t(m) = P(\mathbf{r}_{t+1}^n | S_t = m)$ denotes the probability of receiving the bit sequence $\mathbf{r}_{t+1}^n$ *after* time t conditioned on being in state m at time t.

Therefore,

$$\beta_t(m) = \sum_{m'} \beta_{t+1}(m') \cdot \gamma_{t+1}(m, m')$$

Since we will always assume that at time $t = n$ the state of the trellis is $m = 0$, that is $S_n = 0$, then the boundary conditions on *beta* are

$$\begin{aligned} \beta_n(0) &= 1 \\ \beta_n(m) &= 0 \quad \text{for } m \neq 0 \end{aligned} \tag{15}$$

B. Computation of the joint probability functions $\lambda_t(m)$ and $\sigma_t(m', m)$

1) Obtaining of $\lambda_t(m)$: We first recall that $\lambda_t(m) = P(S_t = m; \mathbf{r}_1^n)$ indicates the joint probability of being in state m at time instant t and that the received bit sequence (word) is $\mathbf{r}_1^n$.

Therefore,

$$P(S_t = m; \mathbf{r}_1^n) = P(S_t = m; \mathbf{r}_1^t) \cdot P(\mathbf{r}_{t+1}^n | S_t = m) \tag{16}$$

which is the same as

$$\lambda_t(m) = \alpha_t(m) \cdot \beta_t(m) \tag{17}$$

Intuitively this says that λ *gives information about the states (nodes in the trellis) at time t, and that this information can be obtanided by watching what happens in the trellis* ***before*** *and* ***after*** *time t.*

2) Obtaining of $\sigma_t(m', m)$: We recall that the $\sigma_t(m', m) = P(S_{t-1} = m'; S_t = m; \mathbf{r}_1^n)$ function is defined as the joint probability of being in state m' at time instant $t-1$, and in state m at time instant t, and that the received bit sequence is $\mathbf{r}_1^n$.

Therefore,

$$P(S_{t-1} = m'; S_t = m; \mathbf{r}_1^n) = P(S_{t-1} = m'; \mathbf{r}_1^{t-1}) \cdot P(S_t = m; r_t | S_{t-1} = m') \cdot P(\mathbf{r}_{t+1}^n | S_t = m)$$

that again is the same as

$$\sigma_t(m', m) = \alpha_{t-1}(m') \cdot \gamma_t(m', m) \cdot \beta_t(m) \tag{18}$$

ISBN: 978-1-61804-240-8

As an intuitive explanation we can say that $\sigma_t(m', m)$ *gives information about the transitions (symbols) in the trellis at time t, and that this transition information can be obtained by checking what happens in the trellis* ***before****,* ***after*** *and* ***at*** *time t.*

C. Graphical interpretation of the auxiliary functions $\alpha_t(m)$*,* $\beta_t(m)$ *and* $\gamma_t(m', m)$

In Section VII-A1 we discussed the computation of the probability function $\gamma_t(m', m) = P(S_t = m; r_t | S_{t-1} = m')$, obtaining the expression

$$\gamma_t(m', m) = \sum_{x \in \mathbb{F}_2} P(S_t = m | S_{t-1} = m') \cdot P(x | S_{t-1} = m', S_t = m) \cdot P(r_t | x)$$

In the case that the trellis corresponds to the dual Hamming binary code, we have that

- The probability $P(S_t = m | S_{t-1} = m')$ is

$$P(S_t = m | S_{t-1} = m') = \begin{cases} 1/2 & \text{if 2 edges depart m'} \\ 1 & \text{if 1 edge departs m'} \end{cases}$$

- The probability $P(x | S_{t-1} = m', S_t = m)$ is

$$P(x | S_{t-1} = m', S_t = m) \begin{cases} 1 & \text{if edge from } m' \text{ to } m = x \\ 0 & \text{otherwise} \end{cases}$$

This implies that

$$\gamma_t(m', m) = \begin{cases} \frac{1}{2} P(r_t | x) & \text{if edge from } m' \text{ to } m = x \\ & \text{and 2 edges depart from state m'} \\ P(r_t | x) & \text{if edge from } m' \text{ to } m = x \\ & \text{and 1 edges departs from state m'} \end{cases} \tag{19}$$

In Section VII-A2 we saw that the *before* function $\alpha_t(m) = P(S_t = m; \mathbf{r}_1^t)$ can be computed from the trellis recursively using

$$\alpha_t(m) = \sum_{m'} \alpha_{t-1}(m') \cdot \gamma_t(m', m)$$

with the boundary conditions

$$\begin{aligned} \alpha_0(0) &= 1 \\ \alpha_0(m) &= 0 \quad \text{for } m \neq 0 \end{aligned} \tag{20}$$

The value of $\gamma_t(m', m)$ is readily obtained from (19), and therefore to obtain $\alpha_t(m)$ we need $\alpha_{t-1}(m')$ so we will be moving through the trellis from left to right, i.e. in the **forward** direction.

In Section VII-A2 we saw that the *after* function $\beta_t(m) = P(\mathbf{r}_{t+1}^n | S_t = m)$ can be computed from the trellis recursively using

$$\beta_t(m) = \sum_{m'} \beta_{t+1}(m') \cdot \gamma_{t+1}(m, m')$$

with the boundary conditions

$$\begin{aligned} \beta_n(0) &= 1 \\ \beta_n(m) &= 0 \quad \text{for } m \neq 0 \end{aligned} \tag{21}$$

Again, the value of $\gamma_t(m', m)$ is readily obtained from (19), and therefore to obtain $\beta_t(m)$ we need $\beta_{t+1}(m')$ so we will be moving through the trellis from right to left, i.e. in the **backward** direction.

Acknowledgement

This work has been supported in part by the Spanish Government through project Consolider Ingenio 2007 CSD2007-00004 "ARES" and TEC2011-26491 "COPPI".

References

[1] F. Chung, R. Graham, and T. Leighton, "Guessing secrets," *The Electronic Journal of Combinatorics*, vol. 8, p. R13, 2001.

[2] "I've got a secret. A classic tv gameshow." *http://www.timvp.com/ivegotse.html*.

[3] N. Alon, V. Guruswami, T. Kaufman, and M. Sudan, "Guessing secrets efficiently via list-decoding," in *Proc. of the 13th Annual ACM-SIAM SODA*, 2002, pp. 254–262.

[4] Y. L. Sagalovich, "Separating systems," *Probl. Inform. Trans.*, vol. 30, no. 2, pp. 14–35, 1994.

[5] J. Domingo-Ferrer and J. Herrera-Joancomartí, "Simple collusion-secure fingerprinting schemes for images," in *Proceedings of the Information Technology: Coding and Computing–ITCC'00.* IEEE Computer Society, 2000, pp. 128–132.

[6] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate (corresp.)," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 284–287, 1974.

[7] G. Cohen, S. Encheva, and H. G. Schaathun, "On separating codes," ENST, Paris, Tech. Rep., 2001.

[8] J. K. Wolf, "Efficient maximum likelihood decoding of linear block codes using a trellis," *IEEE Trans. Inform. Theory*, vol. 24, pp. 76–80, 1978.

Efficient Numerical Method in the High-Frequency Anti-Plane Diffraction by an Interface Crack

Michael Remizov and Mezhlum Sumbatyan

***Abstract*—In the problem of high-frequency diffraction by interface cracks in linear elastic materials we propose a numerical method which is based on a separation of the oscillating solution and a certain slowly varying function. New effective approximate factorization of the symbolic function is offered while using of the Wiener-Hopf method for high-frequency asymptotic solution. This technique described in literature for regular (Fredholm) integral equations is applied here to hyper-singular equations arising in diffraction by thin cracks on the boarder between two different elastic media. The algorithm proposed is efficient for both high and moderate frequencies.**

***Index Terms*—diffraction, integral equation, high frequency, numerical method.**

I. Introduction

THE high-frequency regime is a field of the diffraction theory where standard numerical methods encounter significant obstacles since these require too huge discrete grids. Various approaches have been proposed to overcome this difficulty. Schematically, they may be classifies as analytical (or purely asymptotic) and semi-analytical (i.e. combining numerical treatments with some asymptotic properties of the solution). The key ideas of asymptotic theories, well further references, can be found in recent works [1]-[6]. Only few works are devoted to semi-analytical approaches, and a good representation of respective ideas is given in [7]. For volumetric obstacles this is applied to the Fredholm boundary integral equation. The main goal of the present work is to propose a new numerical method in the anti-plane diffraction problem for an interface crack, which is efficient for high and moderate frequencies and based principally on the constructed explicit high-frequency asymptotics.

II. Anti-plane Diffraction Problem

Let us consider the SH- (anti-plane) problem on diffraction of a plane incident wave by a straight finite-length crack $x \in (-a,a)$, $y=0$ located on the boarder between two different linear elastic isotropic spaces. The plane incident transverse wave arrives from infinity in the upper (first) medium, forming angle θ with respect to vertical axis y: $w^{inc}(x,y)=\exp[-ik_{1s}(x\sin\theta+y\cos\theta)]$, where k_{1s} is the transverse wave number for the upper half-plane ($y\geq 0$) and the time-dependent factor $\exp(-i\omega t)$ is hidden. Note that in the anti-plane problem the displacement vector is $\bar{u_j}(x,y,z)=\{0,0,w_j(x,y)\}$, $j=1,2$ where functions w_j satisfy the Helmholtz equations for the upper ($j=1$) and lower ($j=2$) half-planes respectively:

$$\frac{\partial^2 w_j}{\partial x^2}+\frac{\partial^2 w_j}{\partial y^2}+k_j^2 w=0, \quad k_j=\omega\sqrt{\frac{\rho_j}{\mu_j}}, \tag{2.1}$$

where μ_j and ρ_j designate elastic shear modulus and mass density for respective medium.

The boundary conditions correspond to stress-free faces of the crack, and the continuity of the displacement and the stress on the interface outside the crack. This implies:

$$y=0: \ \frac{\partial w_1}{\partial y}=\frac{\partial w_2}{\partial y}=0\,, \quad |x|\leq a\,;$$

$$w_1=w_2\,, \quad \mu_1\frac{\partial w_1}{\partial y}=\mu_2\frac{\partial w_2}{\partial y}\,, \quad |x|>a\,. \tag{2.2}$$

Let us represent the wave field in the upper medium as the sum of the incident and the scattered ones: $w_1=e^{-ik_1(x\sin\theta+y\cos\theta)}+w_1^{sc}$. By applying the Fourier transform along x-axis: $w_1(x,y)\Longrightarrow W_1(s,y)$, $w_2(x,y)\Longrightarrow W_2(s,y)$, one easily obtains from (2.1):

$$W_1=A_1(s)e^{-\gamma_1 y}+2\pi\delta(s-k_1\sin\theta)e^{-ik_1 y\cos\theta},$$
$$W_2=A_2(s)e^{\gamma_2 y}, \quad \gamma_j=\sqrt{s^2-k_j^2}\,, \tag{2.3}$$

where the following obvious relation (δ is Dirac's delta-function):

$$\int_{-\infty}^{\infty} e^{-ik_1 x\sin\theta}\,e^{ixs}\,dx=2\pi\delta(s-k_1\sin\theta) \tag{2.4}$$

has been used, and A_1, A_2 are two arbitrary functions of Fourier parameter s. It should be noted that expressions (2.3) automatically satisfy the radiation condition at infinity.

It follows from (2.2) that $\mu_1\partial w_1/\partial y=\mu_2\partial w_2/\partial y,\ y=0$ for all $|x|<\infty$. This implies:

$$-\mu_1[\gamma_1 A_1+2\pi i k_1\cos\theta\ \delta(s-k_1\sin\theta)]=\gamma_2\mu_2 A_2. \tag{2.5}$$

In order to obtain a second relation between two quantities A_1 and A_2, let us introduce the new unknown function $q(x)$, as follows:

$$y=0: \quad w_1-w_2=q(x), \quad |x|<\infty\,;$$
$$q(x)=0, \quad |x|>a, \tag{2.6}$$

where the trivial value of $q(x)$ outside the crack follows from the continuity of the displacement over the interface, see (2.2). Therefore, if $q(x)\Longrightarrow Q(s)$, then

$$Q(s)=W_1(s,0)-W_2(s,0)=A_1+2\pi\delta(s-k_1\sin\theta)-A_2=$$

M. Yu. Remizov and M. A. Sumbatyan are with the Faculty of Mathematics, Mechanics and Computer Science, Southern Federal University, Rostov-on-Don, Russia e-mail: remizov72@mail.ru, sumbat@math.rsu.ru

Manuscript received May 31, 2014; revised January 11, 2014.

$$= \left(1 - \frac{ik_1}{\gamma_1}\cos\theta\right)\delta(s - k_1\sin\theta) - \left(1 + \frac{\mu_2\gamma_2}{\mu_1\gamma_1}\right)A_2, \quad (2.7)$$

where the value of A_1 in terms of A_2 has been used, see Eq. (2.5). Now, Eqs. (2.3) and (2.7) imply:

$$W_2(s,y) = \mu_1\left[2\pi\frac{\gamma_1 - ik_1\cos\theta}{\mu_1\gamma_1 + \mu_2\gamma_2}\ \delta(s - k_1\sin\theta) - \right.$$
$$\left. - \frac{\gamma_1 Q(s)}{\mu_1\gamma_1 + \mu_2\gamma_2}\right]e^{\gamma_2 y}, \quad (2.8)$$

and the remaining still unused boundary condition in (2.2), namely $\partial w_2(x,0)/\partial y = 0$, $|x| \le a$, by applying the inverse Fourier transform to Eq. (2.8), results in the basic integral equation for the unknown function $q(x)$:

$$\int_{-ak_1}^{ak_1} q(\xi)K(x-\xi)d\xi = f(x), \ |x| \le ak_1; \quad (2.9)$$

$$K(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} L(s)e^{-ixs}ds = \frac{1}{\pi}\int_0^{\infty} L(s)\cos(xs)ds,$$

$$L(s) = \frac{\sqrt{s^2-1}\ \sqrt{s^2-k^2}}{\mu\sqrt{s^2-1} + \sqrt{s^2-k^2}}, \ \mu = \frac{\mu_1}{\mu_2}, \ k^2 = \frac{k_2^2}{k_1^2} = \frac{\mu_1\rho_2}{\mu_2\rho_1},$$

$$f(x) = \left[\frac{(\sqrt{s^2-1} - i\cos\theta)\sqrt{s^2-k^2}}{\mu\sqrt{s^2-1} + \sqrt{s^2-k^2}}e^{-ixs}\right]_{s=\sin\theta} = Ae^{-ix\sin\theta},$$

$$A = \frac{-2i\cos\theta\sqrt{\sin^2\theta - k^2}}{\sqrt{\sin^2\theta - k^2} - i\mu\cos\theta}$$

written in a dimensionless form.

First of all, let us notice that the denominator of the fraction in function $L(s)$ does not vanish. This follows from the consideration of the three possible cases: (i) Both square roots $\sqrt{s^2-1}$ and $\sqrt{s^2-k^2}$ are real-valued. In this case they both are positive, hence the sum of two positive quantities cannot possess the zero value; (ii) One of them is real-valued and the other one is imaginary. In this case the sum of these two square roots may vanish if and only if they both are trivial that is impossible; (iii) Both these square roots are imaginary. This case can be reduced to case (i), since the denominator is the same quantity as in (i) multiplied by $-i$.

It is clear that kernel $K(x)$ is even: $K(x) = K(|x|)$. Besides, the kernel is smooth: $K(|x|) \in C_1(0, 2ak_1]$ outside the origin. In order to estimate its behavior as $x \to 0$, let us extract explicitly the leading asymptotic term of function $L(s)$ at infinity: $L(s) = |s|/(\mu+1) + O(1/|s|)$, $s \to \infty$. Therefore,

$$K(x) = \frac{1}{\pi(\mu+1)}\int_0^{\infty} s\cos(xs)\,ds \ + \ K_0(x) =$$
$$= -\frac{1}{\pi(\mu+1)\,x^2} \ + \ K_0(x), \quad (2.10)$$

$$K_0(x) = \frac{1}{\pi}\int_0^{\infty}\left[\frac{\sqrt{s^2-1}\ \sqrt{s^2-k^2}}{\mu\sqrt{s^2-1} + \sqrt{s^2-k^2}} - \frac{s}{\mu+1}\right]\cos(xs)\,ds =$$
$$= O(\ln|x|), \quad x \to 0.$$

It thus can be seen that the leading term of the kernel's expansion for small x is hyper-singular, and kernel $K_0(x)$ has the weak (integrable) logarithmic singularity only. A stable direct numerical algorithm to solve integral equations with such kernels is described in [8].

III. Asymptotic Analysis of the Basic Integral Equation

In the high-frequency regime the numerical treatment of equation (2.9) becomes inefficient, because it is necessary to keep a fixed number of nodes per wave length. As a result, this leads to a huge size of the discrete mesh. For this reason, let us construct an asymptotic solution of integral equation (2.9), as $ak_1 \to \infty$. The method we use is allied to the classical "Edge Waves" technique [4]. Let us represent the solution of equation (2.9) as a combination of three functions:

$$q(x) = q_1(ak_1 + x) + q_2(ak_1 - x) - q_0(x), \quad (3.1)$$

satisfying, respectively, the following three equations:

$$\int_{-ak_1}^{\infty} q_1(ak_1+\xi)K(x-\xi)d\xi = f(x) +$$
$$+ \int_{-\infty}^{-ak_1}[q_2(ak_1-\xi) - q_0(\xi)]K(x-\xi)d\xi, \ -ak_1 < x < \infty, \quad (3.2a)$$

$$\int_{-\infty}^{ak_1} q_2(ak_1-\xi)K(x-\xi)\,d\xi = f(x) +$$
$$+ \int_{ak_1}^{\infty}[q_1(ak_1+\xi) - q_0(\xi)]K(x-\xi)\,d\xi, \ -\infty < x < ak_1, \quad (3.2b)$$

$$\int_{-\infty}^{\infty} q_0(\xi)\,K(x-\xi)\,d\xi = f(x), \quad -\infty < x < \infty\,. \quad (3.2c)$$

The equivalence of equation (2.9) and the system of three equations (3.2) is easily proved if one applies the combination (3.2a)+(3.2b)–(3.2c).

The leading asymptotic term of the solution can be constructed by rejecting the residual integrals in the right-hand sides of (3.2a) and (3.2b). Under such a treatment, these two equations contain integral operators only in their left-hand sides, becoming the Wiener-Hopf equations on semi-infinite intervals. As soon as they are solved, the correctness of the hypothesis, that the rejected right-hand-side tails are asymptotically small, can be checked by substituting the found solutions into those tail integrals. Physically, this means that the reciprocal wave influence of the edges to each other is asymptotically small, in the first approximation.

It should be noted that the third equation (3.2c) is a simple convolution integral equation on the infinite axis, and its solution is easily obtained by the Fourier transform ($f(x) \Longrightarrow F(s)$):

$$q_0(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{F(s)}{L(s)}e^{-ixs}ds = a\int_{-\infty}^{\infty}\frac{\delta(s-\sin\theta)}{L(s)}e^{-ixs}ds =$$

$$= \frac{Ae^{-ix\sin\theta}}{L(\sin\theta)} = 2\,e^{-ix\sin\theta}. \qquad (3.3)$$

It is very interesting to notice that $q_0(x)$ is the same as it could be predicted by Kirchhoff's physical diffraction theory [4].

The Wiener-Hopf equations (3.2a), (3.2b) discussed above, after evident change of variables $x' = ak_1 \pm x$, $\xi' = ak_1 \pm \xi$, can be rewritten in a more standard form, holding over interval $(0, \infty)$:

$$\int_0^\infty q_{1,2}(\xi')K(x'-\xi')\,d\xi' = f_{1,2}(x'), \quad 0 \le x' < \infty;$$

$$f_{1,2}(x') = f[\pm(x'-ak_1)]. \qquad (3.4)$$

As usually, in solving the Wiener-Hopf equations let us designate for any function $\varphi^+(x') = \varphi(x')$, $x' \ge 0$; $\varphi^+(x') = 0$, $x' < 0$; $\varphi^-(x') = \varphi(x')$, $x' \le 0$; $\varphi^-(x') = 0$, $x' > 0$; and $\Phi^+(x')$ ($\Phi^-(x')$) are the values on the real axes of $\Phi(z)$, analytic in upper (lower) half-planes of complex variable z. Then (3.4) is rewritten as follows

$$Q_j^+(s)L(s) = F_j^+(s) + D^-(s), \quad (j = 1, 2);$$

$$F_{1,2}^+(s) = \frac{Ae^{\pm iak_1\sin\theta}}{i\,(\pm\sin\theta - s)}, \qquad (3.5)$$

where $D^-(s)$ is an unknown function.

The next step is a factorization of the symbolic function [9]: $L(s) = L^+(s)L^-(s)$. After that eq. (3.5) reads as follows:

$$Q_j^+(s)L^+(s) = \frac{F_j^+(s)}{L^-(s)} + E^-(s), \qquad (3.6)$$

where $E^-(s)$ is another unknown function. Now, after the obvious decomposition

$$\frac{F_{1,2}^+(s)}{L^-(s)} = \frac{Ae^{\pm iak_1\sin\theta}}{i\,(\pm\sin\theta - s)L^-(s)} =$$

$$= \left\{\frac{Ae^{\pm iak_1\sin\theta}}{i\,(\pm\sin\theta - s)}\left[\frac{1}{L^-(s)} - \frac{1}{L^-(\pm\sin\theta)}\right]\right\}_- +$$

$$+ \left[\frac{Ae^{\pm iak_1\sin\theta}}{iL^-(\pm\sin\theta)(\pm\sin\theta - s)}\right]_+ =$$

$$= H^-(s) + H^+(s), \qquad (3.7)$$

relation (3.6) can be rewritten as follows:

$$Q_j^+(s)L^+(s) - H^+(s) = H^-(s) + E^-(s). \qquad (3.8)$$

Since the left-hand side here contains only functions analytical in the upper half-plane and the right-hand side – only functions analytical in the lower one, these two functions are in fact the same unique entire function. The physical condition claims that the solution, which is an opening of the crack faces, should vanish when approaching the crack's edges: $q_{1,2} \to 0, x \to 0$. A simple analysis shows that this implies that the entire function above must identically be trivial. This defines the solution of the Wiener-Hopf equation (3.4) in the form

$$Q_{1,2}^+(s) = \frac{H^+(s)}{L^+(s)} =$$

$$= \frac{Ae^{\pm iak_1\sin\theta}}{iL^-(\pm\sin\theta)(\pm\sin\theta - s)L^+(s)}, \qquad (3.9)$$

which after application of the inverse Fourier transform gives the solution to equation (3.4) and, as a consequence, the leading term (3.1) of the high-frequency analytical solution to the whole problem.

IV. Efficient factorization of function L(s) and a closed-form solution

It is obvious from the previous section that the key point of the method proposed is the factorization of the symbolic function $L(s)$. The representation of this function in its exact form (2.9) unlikely admits any explicit-form factorization. There are known some complex analytical formulas for factorization of arbitrary function, expressed in quadratures [9], however in practice calculation of such integrals turns out very hard problem.

In the present work we give an efficient approximation of function $L(s)$ which admits an evident simple factorization. Let us notice that the structure of $L(s)$ (2.9) is a combination of four square roots $(\sqrt{s+1}\,)_+$, $(\sqrt{s+k}\,)_+$, $(\sqrt{s-1}\,)_-$, $(\sqrt{s-k}\,)_-$, two of them being analytical in the upper half-plane, and the other two – in the lower half-plane (for more detail, see [9]). obviously, the factorization of the numerator in (2.9) is attained in a simple way: $\sqrt{s^2-1}\sqrt{s^2-k^2} = \left(\sqrt{s+1}\sqrt{s+k}\right)_+ \left(\sqrt{s-1}\sqrt{s-k}\right)_-$.

Let us approximate the denominator of $L(s)$ as follows:

$$\mu\sqrt{s^2-1} + \sqrt{s^2-k^2} \approx \qquad (4.1)$$

$$\approx \frac{\mu+1}{(B+1)^2}\left(B\sqrt{s+1} + \sqrt{s+k}\right)_+ \cdot \left(B\sqrt{s-1} + \sqrt{s-k}\right)_-.$$

It is obvious that the approximating function has the same asymptotic behavior as $s \to \infty$. Besides, this keeps all qualitative properties of the initial function, having the same branching points $s = \pm 1, \pm k$. The introduced parameter $B = B(\mu, k) > 0$ may be chosen, for given values of parameters μ and k, to provide better approximation uniformly over all finite real-valued values of variable $s \in (-\infty, \infty)$. It is also obvious that in the case when $\mu_1/\mu_2 = \rho_1/\rho_2$ parameter $k = 1$, hence the approximation is absolutely precise with $B = 0$. By calculating the maximum relative error ε, between exact and approximating complex-valued functions, for $s \in (-\infty, \infty)$, a numerical investigation shows that the worst precision takes place for opposite values of relations μ_1/μ_2 and ρ_1/ρ_2, i.e. when the former is extremely large (or small) and at the same time the latter is extremely small (or large). For example, in the case $\mu_1/\mu_2 = 1/10$, $\rho_1/\rho_2 = 10$ the relative error can be attained $\varepsilon = 7\%$ only, with $B = 0.2$. in the case $\mu_1/\mu_2 = 1/5$, $\rho_1/\rho_2 = 5$ the value $\varepsilon = 4\%$ can be attained with $B = 0.31$.

However, it should be noted that the cases with rough approximation described above are not realistic from the physical point of view. Indeed, if one material is more rigid than the other one, then in practice its both elastic modulus and mass density are greater than respective parameters of the second material. If we leave such unrealistic cases aside, keeping only the cases when both the relations are simultaneously

ISBN: 978-1-61804-240-8

less or greater than the unit value, then the precision of the approximation becomes considerably better. By allowing the materials' parameters to differ maximum by one order only: $1/10 \le \mu_1/\mu_2,\ \rho_1/\rho_2 \le 1$, the worst case is $\mu_1/\mu_2 = 1$, $\rho_1/\rho_2 = 1/10$ with $\varepsilon = 2\%$ attained for $B = 0.77$. In practice, staying far from the extremal values of the physical parameters, the maximum relative error is always less than 1–2%.

With the introduced approximation (4.1) the efficient factorization of symbolic function $L(s)$ is taken in the form

$$L(s) = \left[\frac{(B+1)\sqrt{s+1}\,\sqrt{s+k}}{\sqrt{\mu+1}\,(B\sqrt{s+1}+\sqrt{s+k})}\right]_+ \times$$

$$\times \left[\frac{(B+1)\sqrt{s-1}\sqrt{s-k}}{\sqrt{\mu+1}\,(B\sqrt{s-1}+\sqrt{s-k})}\right]_- = L^+(s)L^-(s). \quad (4.2)$$

Omitting some transformations, one comes to the following expression of Fourier transforms $Q^+_{1,2}(s)$

$$Q^+_{1,2}(s) = \frac{A(\mu+1)e^{\pm\, ia k_1 \sin\theta}}{(B+1)^2(\pm\sin\theta - s)} \times \quad (4.3)$$

$$\times \left(\frac{1}{\sqrt{1\mp\sin\theta}} + \frac{B}{\sqrt{k\mp\sin\theta}}\right)\left(\frac{1}{\sqrt{s+1}} + \frac{B}{\sqrt{s+k}}\right).$$

The Fourier inversion of this function may be performed by passing to inverse Laplace transform, with the change $is = -p,\ s = ip$, where p is the Laplace parameter. Expression (4.3) contains elementary functions with tabulated Laplace inversions [10]:

$$\frac{1}{(\alpha - s)\sqrt{s+\beta}} = \frac{e^{\pi i/4}}{(p+i\alpha)\sqrt{p-i\beta}}$$

$$\Longleftarrow \ \frac{ie^{-i\alpha x'}}{\sqrt{\alpha+\beta}}\,\mathrm{Erf}\left[e^{-\pi i/4}\sqrt{(\alpha+\beta)x'}\right], \quad (4.4)$$

where $\mathrm{Erf}(z)$ is the probability integral. Since $x' = ak_1 \pm x$, the inversion of (4.3) gives

$$q^{(1,2)}(ak_1 \pm x) = \frac{A(\mu+1)i}{(B+1)^2}\times$$

$$\times \left(\frac{1}{\sqrt{1\mp\sin\theta}} + \frac{B}{\sqrt{k\mp\sin\theta}}\right)e^{\mp ix\sin\theta}\times$$

$$\times \left\{\frac{\mathrm{Erf}\left[e^{-\pi i/4}\sqrt{(1\pm\sin\theta)(ak_1\pm x)}\right]}{\sqrt{1\pm\sin\theta}} + \right.$$

$$\left. + \frac{B\ \mathrm{Erf}\left[e^{-\pi i/4}\sqrt{(k\pm\sin\theta)(ak_1\pm x)}\right]}{\sqrt{k\pm\sin\theta}}\right\}. \quad (4.5)$$

For large argument the probability integral tends to 1, then one can see that

$$q^+_{1,2}(ak_1 \pm x) \sim \frac{A(\mu+1)i}{(B+1)^2}\,e^{\mp ix\sin\theta}\Big(\frac{1}{\sqrt{1\mp\sin\theta}} +$$

$$+\frac{B}{\sqrt{k\mp\sin\theta}}\Big)\times\left(\frac{1}{\sqrt{1\pm\sin\theta}} + \frac{B}{\sqrt{k\pm\sin\theta}}\right) =$$

$$= \frac{Ae^{-ix\sin\theta}}{L(\sin\theta)} = 2\,e^{-ix\sin\theta} = q_0(x), \quad x\to\pm\infty\,. \quad (4.6)$$

The second term of this asymptotic estimate (not written here for the sake of brevity) shows that the difference $q^+_{1,2}(ak_1 \pm x) - q_0(x)$ not only tends to zero as $x \to \pm\infty$ but also is integrable at infinity. This guarantees the right-hand-side tails in Eq. (3.2) to be asymptotically small as $ak_1 \to \infty$, that justifies the basic hypothesis permitting rejection of the tails.

Finally, we note that similar problems have been studied in [11] where a polynomial approximation form is applied to the factorization problem, and in [12] where a numerical treatment of factorization is performed. The principal distinctive feature of the factorization proposed in the present work is that it catches all qualitative properties of the basic branching complex-valued symbolic function, in the way permitting the explicit-form solution of the posed problem. It also provides any desired level of precision, by obvious combination of the proposed factorization (4.1) with the ideas presented in [11].

V. THE ESSENCE OF THE PROPOSED NUMERICAL METHOD

The basic idea of the proposed method is to seek the basic unknown function $q(x)$ as a product of the high-frequency asymptotic representation (3.1) and a certain slowly varying function $G(x)$, namely

$$q(x) = \left[q_1\,(ak_1 + x) + q_2\,(ak_1 - x) - q_0(x)\right]G(x),$$

$$ak_1 \to \infty. \quad (5.1)$$

Physically, for extremely high frequencies function G must tend to an identically unit value. For moderately high frequencies the new unknown function $G(x)$ is a certain slowly varying one, playing the role of modulating amplitude for the rapidly oscillating component. One thus may take a small number of nodes, to find this function from the main integral equation, uniformly for all high frequencies. The substitution of (5.1) into Eqs. (2.9) rewrites it in the form

$$\frac{1}{2\pi}\int_{-ak_1}^{ak_1}\left[q_1\,(ak_1+\xi) + q_2\,(ak_1-\xi) - q_0(\xi)\right]G(\xi)\,d\xi\times$$

$$\times\int_{-\infty}^{\infty} L(s)e^{is(\xi-x)}\,ds = f(x), \quad |x| \le ak_1. \quad (5.2)$$

It should be noted that functions $q_{(1,2)}(x)$ in (4.5) can be rewritten in terms of Fresnel integrals $S_2(x)$, $C_2(x)$ of real-valued arguments:

$$q_{(1,2)}(x) = D_{1,2}e^{\mp ix\sin\theta}[C_2(a_{1,2}x) + iS_2(a_{1,2}x)] +$$

$$+E_{1,2}e^{\mp ix\sin\theta}[C_2(b_{1,2}x) + iS_2(b_{1,2}x)];$$

$$D_{1,2} = \frac{B_{1,2}e^{-i\frac{3\pi}{4}}\sqrt{2}}{\sqrt{a_{1,2}}}; \quad a_{1,2} = k_s \pm \sin\theta;$$

$$E_{1,2} = \frac{b_{1,2}e^{-\frac{3\pi}{4}}\sqrt{2}}{\sqrt{b_{1,2}}}; \quad b_{1,2} = 1 \pm \sin\theta. \quad (5.3)$$

In discretization of Eq. (5.2) let us choose N equal subintervals over full interval $(-a, a)$, where N is the same for all large values of parameter ak_1. Then, assuming that the

ISBN: 978-1-61804-240-8

unknown function $G(\xi)$ is almost constant over each small sub-interval, we deduce ($G_m = G(\xi_m)$):

$$I = \frac{1}{2\pi}\int_{-ak_1}^{ak_1} G(\xi)[q_1\left(ak_1+\xi\right)+ \tag{5.4}$$

$$+q_2\left(ak_1-\xi\right)-q_0(\xi)]d\xi\int_{-\infty}^{\infty} L(s)e^{is(\xi-x)} \approx$$

$$\approx \frac{1}{2\pi}\sum_{m=1}^{N_1} G_m \int_{\left(x_m-\frac{hk_1}{2}\right)}^{\left(x_m+\frac{hk_1}{2}\right)} [q_1\left(ak_1+\xi\right)+$$

$$+q_2\left(ak_1-\xi\right)-q_0(\xi)]d\xi\int_{-\infty}^{\infty} L(s)\, e^{is(\xi-x_n)}\, ds;$$

$$\sum_{m=1}^{N} G_m I_{nm} = f_n, \tag{5.5}$$

$$I_{nm} = \frac{1}{2\pi}\int_{\left(x_m-\frac{hk_1}{2}\right)}^{\left(x_m+\frac{hk_1}{2}\right)} [D_1 e^{-i(ak_1+\xi)\sin\theta}\times$$

$$\times\left\{C_2\left(a_1\left[ak_1+\xi\right]\right)+iS_2\left(a_1\left[ak_1+\xi\right]\right)\right\}+$$

$$+E_1 e^{-i(ak_1+\xi)\sin\theta}\times$$

$$\times\left\{C_2\left(b_1\left[ak_1+\xi\right]\right)+iS_2\left(b_1\left[ak_1+\xi\right]\right)\right\}+$$

$$+D_2 e^{i(ak_1-\xi)\sin\theta}(C_2\left(a_2\left[ak_1-\xi\right]\right)+$$

$$+iS_2\left(a_2\left[ak_1-\xi\right]\right))+E_2 e^{i(ak_1-\xi)\sin\theta}\times$$

$$\times\left\{C_2\left(b_2\left[ak_1-\xi\right]\right)+iS_2\left(b_2\left[ak_1+\xi\right]\right)\right\}-$$

$$-2e^{-i\xi\sin\theta}]d\xi\int_{-\infty}^{\infty} L(s)e^{is(\xi-x_n)}\, ds;$$

$$f_n = f(x_n) = Ae^{-ix_n\sin\theta},\ h = 2a/N,\ n = 1, N.$$

Let us estimate the efficiency of the proposed algorithm, say for $ak_1 = 200$. The standard numerical treatment means to solve a certain 1200×1200 LAS (linear algebraic system). The proposed method requires to find slowly varying function $G(x)$, hence it is quite sufficient to take $N = 120$ nodes, i.e a certain 120×120 LAS, whose dimension is smaller in 10 times. Since the number of arithmetic operations in the standard Gauss elimination algorithm is proportional to the third power of dimension, the gain is $10^3 = 1000$ times. Obviously, for larger N the gain becomes even more significant.

It should be noted that the algorithm proposed works well for all high, moderate, and low frequencies. Really, in the cases of high and moderate frequencies it is discussed above. In the case of low frequencies the full solution as well as extracted oscillating exponential function become slowly varying functions, hence a small quantity of nodes is again required in this case for the numerical treatment.

REFERENCES

[1] L.B. Felsen, N. Marcuvitz, *Radiation and Scattering of Waves*, Prentice-Hall: Englewood Cliffs, New Jersey, 1973.

[2] V.M. Babich, V.S. Buldyrev, *Asymptotic Methods in Short-Wavelength Diffraction Theory*, Springer-Verlag: Berlin / Heidelberg, 1989.

[3] D.A.M. McNamara, C.W.I. Pistorius, J.A.G. Malherbe, *Introduction to the Uniform Geometrical Theory of Diffraction*, Artech House: Norwood, 1990.

[4] P.Ya. Ufimtsev, *Fundamentals of the Physical Theory of Diffraction*, John Wiley: Hoboken, New Jersey, 2007.

[5] E. Scarpetta, M.A. Sumbatyan, Explicit analytical representations in the multiple high-frequency reflection of acoustic waves from curved surfaces: the leading asymptotic term, *Acta Acust. Acust.*, 2011, **97**, 115–127.

[6] E. Scarpetta, M.A. Sumbatyan, An asymptotic estimate of the edge effects in the high-frequency Kirchhoff diffraction theory for 3d problems, *Wave Motion*, 2011, **48**, 408–422.

[7] M.Yu. Remizov, M.A. Sumbatyan, A semi-analytical method of solving problems of the high-frequency diffraction of elastic waves by cracks, *J. Appl. Math. Mech.*, 2013, **77**, 452-456.

[8] G. Iovane, I.K. Lifanov, M.A. Sumbatyan, On direct numerical treatment of hypersingular integral equations arising in mechanics and acoustics, *Acta Mech.*, 2003, **162**, 99–110. H. Bateman, A. Erdelyi, *Tables of Integral Transforms. V.1*, McGraw-Hill: New York, 1954.

[9] R. Mittra, S.W. Lee, *Analytical Techniques in the Theory of Guided Waves*, Macmillan: New York, 1971.

[10] H. Bateman, A. Erdelyi, *Tables of Integral Transforms. V.1*, McGraw-Hill: New York, 1954.

[11] D. Abrahams, On the solution of Wiener-Hopf problems involving noncommutative matrix kernel decompositions, *SIAM J. Appl. Math.*, 1997, **57**, 541–567.

[12] S.C. Pal, M.L. Ghosh, High frequency scattering of anti-plane shear waves by an interface crack, *Indian J. Pure Appl. Math.*, 1990, **21**, 1107–1124.

Robust Adaptive Control with Disturbance Rejection for Symmetric Hyperbolic Systems of Partial Differential Equations

Mark J. Balas and Susan A. Frost

***Abstract*— Given a linear continuous-time infinite-dimensional plant on a Hilbert space and disturbances of known and unknown waveform, we show that there exists a stabilizing direct model reference adaptive control law with certain disturbance rejection and robustness properties. The closed loop system is shown to be exponentially convergent to a neighborhood with radius proportional to bounds on the size of the disturbance. The plant is described by a closed densely defined linear operator that generates a continuous semigroup of bounded operators on the Hilbert space of states.**

Symmetric Hyperbolic Systems of partial differential equations describe many physical phenomena such as wave behavior, electromagnetic fields, and quantum fields. To illustrate the utility of the adaptive control law, we apply the results to control of symmetric hyperbolic systems with coercive boundary conditions.

Keywords: infinite dimensional systems, partial differential equations, adaptive control.

I. Introduction

Many control systems are inherently infinite dimensional when they are described by partial differential equations. Currently there is renewed interest in the control of these kinds of systems especially in flexible aerospace structures and the quantum control field [1]-[2]. It is especially of interest to control these systems adaptively via finite-dimensional controllers. In our work [3]-[6] we have accomplished direct model reference adaptive control and disturbance rejection with very low order adaptive gain laws for MIMO finite dimensional systems. When systems are subjected to an unknown internal delay, these systems are also infinite dimensional in nature. The adaptive control theory can be modified to handle this situation [7]. However, this approach does not handle the situation when partial differential equations describe the open loop system.

This paper considers the effect of infinite dimensionality on the adaptive control approach of [4]-[6]. We will show that the adaptively controlled system is globally stable, but the adaptive error is no longer guaranteed to approach the origin. However, exponential convergence to a neighborhood can be achieved as a result of the control design. We will prove a robustness result for the adaptive control which extends the results of [4].

M.J. Balas is with the Aerospace Engineering Department, Embry-Riddle Aeronautical University , Daytona Beach, FL 32119 (balsam@erau.edu).

S.A. Frost is with the Intelligent Systems Division, NASA Ames Research Center, Moffett Field, CA 94035(susan.frost@nasa.gov).

Our focus will be on applying our results to Symmetric Hyperbolic Systems of partial differential equations. Such systems, originated by K.O. Friedrichs and P. D. Lax, describe many physical phenomena such as wave behavior, electromagnetic fields, and the theory of relativistic quantum fields; for example, see [15]-[18]. To illustrate the utility of the adaptive control law, we apply the results to control of symmetric hyperbolic systems with coercive boundary conditions. Other closely related work on compensators for infinite dimensions from a different viewpoint can be found in [21].

II. Robustness of the Error System

We begin by considering the definition of Strict Dissipativity for infinite-dimensional systems and the general form of the "adaptive error system" to later prove stability. The main theorem of this section will be utilized in the following section to assess the stability of the adaptive controller with disturbance rejection for linear diffusion systems.

Noting that there can be some ambiguity in the literature with the definition of strictly dissipative systems, we modify the suggestion of Wen in [8] for finite dimensional systems and expand it to include infinite dimensional systems.

Definition ***1*****:** *The triple* (A_c, *B, C*) *is said to be* ***Strictly Dissipative*** *if* A_c *is a densely defined ,closed operator on* $D(A_c) \subseteq X$ *a complex Hilbert space with inner product* (x, y) *and corresponding norm* $\|x\| \equiv \sqrt{(x,x)}$ *and generates a* C_0 *semigroup of bounded operators* $U(t)$, *and* (B,C) *are bounded finite rank input/output operators with rank M where* $B: R^m \to X$ and $C: X \to R^m$. *In addition there exist symmetric positive bounded operator P* on X *such that* $p_{\min}\|x\|^2 \le (Px, x) \le p_{\max}\|x\|^2$, *i.e. P is bounded and coercive, and*

$$\begin{cases} \mathrm{Re}(PA_c e, e) \equiv \dfrac{1}{2}[(PA_c e, e) + (e, PA_c e)] \\ \qquad\qquad \le -\alpha \|e\|^2 \\ PB = C^* \end{cases} \tag{1}$$

where $\alpha > 0$ *and* $e \in D(A_c)$.

We also say that (A,B,C) is *Almost Strictly Dissipative (ASD)* when there exists a $G_* \in \Re^{mxm}$ such that (A_c,B,C) is strictly dissipative with $A_c \equiv A+BG_*C$.

Note that if $P=I$ in (1),by the Lumer-Phillips Theorem [11], p405, we would have $\|U_c(t)\| \le e^{-\sigma t}; \quad t \ge 0$.

The following theorem shows that convergence to a neighborhood with radius determined by the supremum norm of ν is possible for a specific type of adaptive error system. In the following, we denote $\|M\|_2 \equiv \sqrt{\mathrm{tr}(M\gamma^{-1}M^T)}$ as the trace norm of a matrix *M* where $\gamma > 0$.

Theorem 2: Consider the coupled system of differential equations

$$\begin{cases} \dfrac{\partial e}{\partial t} = A_c e + B\underbrace{\left(G(t)-G^*\right)}_{\Delta G} z + \nu \\ e_y = Ce \\ \dot{G}(t) = -e_y z^{\mathrm{T}}\gamma - aG(t) \end{cases} \tag{2}$$

where $e,\nu \in D(A_C), z \in R^m$ and $\begin{bmatrix} e \\ G \end{bmatrix} \in \bar{X} \equiv XxR^{mxm}$

is a Hilbert space with

inner product $\left(\begin{bmatrix} e_1 \\ G_1 \end{bmatrix},\begin{bmatrix} e_2 \\ G_2 \end{bmatrix}\right) \equiv (e_1,e_2) + \mathrm{tr}\left(G_1\gamma^{-1}G_2\right)$,

norm $\left\|\begin{bmatrix} e \\ G \end{bmatrix}\right\| \equiv \left(\|e\|^2 + \mathrm{tr}(G\gamma^{-1}G)\right)^{\frac{1}{2}}$ and where $G(t)$ is the *mxm* adaptive gain matrix and γ is any positive definite constant matrix, each of appropriate dimension. *Assume the following:*

i.) (A,B,C) *is ASD with* $A_c \equiv A+BG_*C$

ii.)there exists $M_G > 0$ *such that*

$$\sqrt{\mathrm{tr}(G^*G^{*T})} \le M_G$$

iii.) there exists $M_\upsilon > 0$ *such that*

$$\sup_{t\ge 0}\|\nu(t)\| \le M_\nu < \infty$$

iv.) there exists $\alpha > 0$ *such that* $a \le \dfrac{\alpha}{p_{\max}}$, *where*

$p_{\max}$ *is defined in Definition 1*

v.) the positive definite matrix γ *satisfies*

$$\mathrm{tr}(\gamma^{-1}) \le \left(\frac{M_\nu}{aM_G}\right)^2,$$

then the gain matrix, G(t), is bounded, and the state, e(t) exponentially with rate e^{-at} approaches the ball of radius

$$R_* \equiv \frac{\left(1+\sqrt{p_{\max}}\right)}{a\sqrt{p_{\min}}} M_\nu$$

Proof of Theorem 2:

First we note that if see [10] Theo 8.8 p 151. Consider the positive definite function,

$$V = \frac{1}{2}(Pe,e) + \frac{1}{2}\mathrm{tr}\left[\Delta G\gamma^{-1}\Delta G^{\mathrm{T}}\right] \tag{3}$$

where $\Delta G(t) \equiv G(t) - G^*$ and P satisfies (1). Taking the time derivative of (3) (we assume this can be done in X) and substituting (2) into the result yields

$$\begin{aligned} \dot{V} &= \frac{1}{2}[(PA_c e,e) + (e,PA_c e)] + (PBw,e) \\ &\quad + \mathrm{tr}\left[\Delta\dot{G}\gamma^{-1}\Delta G^{\mathrm{T}}\right] + (Pe,\nu) \end{aligned}$$

where $w \equiv \Delta G z$. Invoking the equalities in the definition of Strict Dissipativity in (1), using $x^{\mathrm{T}}y$ = tr[yx^{T}], and substituting (2) into the last expression (with $(PBw,e) = (w,Ce) = \langle e_y, w\rangle \equiv e_y^* w$), we obtain

$$\begin{cases} \dot{V} = \mathrm{Re}(PA_c e,e) + \langle e_y, w\rangle \\ \quad -a\cdot\mathrm{tr}\left[G\gamma^{-1}\Delta G^{\mathrm{T}}\right] \\ \quad -\underbrace{tr(e_y z^T \Delta G^T)}_{\langle e_y,w\rangle} + (Pe,\nu) \\ \quad \le -\|e\|^2 - a\cdot\mathrm{tr}\left[(\Delta G + G^*)\gamma^{-1}\Delta G^{\mathrm{T}}\right] \\ \quad +(Pe,\nu) \\ \quad \le -\left(\alpha\|e\|^2 + a\cdot\mathrm{tr}\left[\Delta G\gamma^{-1}\Delta G^{\mathrm{T}}\right]\right) \\ \quad +a\cdot\left|\mathrm{tr}\left[G^*\gamma^{-1}\Delta G^{\mathrm{T}}\right]\right| + |(Pe,\nu)| \\ \quad \le -\begin{pmatrix} \dfrac{2\alpha}{p_{\min}}\bullet\dfrac{1}{2}(Pe,e) \\ +2a\bullet\dfrac{1}{2}\mathrm{tr}\left[\Delta G\gamma^{-1}\Delta G^{\mathrm{T}}\right] \end{pmatrix} \\ \quad +a\cdot\left|\mathrm{tr}\left[G^*\gamma^{-1}\Delta G^{\mathrm{T}}\right]\right| + |(Pe,\nu)| \\ \quad \le -2aV + a\cdot\left|\mathrm{tr}\left[G^*\gamma^{-1}\Delta G^{\mathrm{T}}\right]\right| + |(Pe,\nu)| \end{cases}$$

Now, using the Cauchy-Schwartz Inequality

$$\left|\mathrm{tr}\left[G^*\gamma^{-1}\Delta G^{\mathrm{T}}\right]\right| \le \|G^*\|_2\|\Delta G\|_2$$

And

$$\left|(Pe,v)\right| \le \left\| P^{\frac{1}{2}} v \right\| \left\| P^{\frac{1}{2}} e \right\| = \sqrt{(Pv,v)} \bullet \sqrt{(Pe,e)}$$

We have

$$\dot{V} + 2aV \le a \cdot \left\| G^* \right\|_2 \left\| \Delta G \right\|_2 + \sqrt{p_{\max}} \left\| v \right\| \sqrt{(Pe,e)}$$

$$\le a \cdot \left\| G^* \right\|_2 \left\| \Delta G \right\|_2 + (\sqrt{p_{\max}} M_v) \sqrt{(Pe,e)}$$

$$\le \left(a \left\| G^* \right\|_2 + \sqrt{p_{\max}} M_v \right) \sqrt{2} \underbrace{\left[\frac{1}{2}(Pe,e) + \frac{1}{2} \left\| \Delta G \right\|_2^2 \right]^{\frac{1}{2}}}_{V^{\frac{1}{2}}}$$

Therefore,

$$\frac{\dot{V} + 2aV}{V^{\frac{1}{2}}} \le a \left(\left\| G^* \right\|_2 + \sqrt{p_{\max}} M_v \right) \sqrt{2}$$

Now, using the identity $\mathrm{tr}[ABC] = \mathrm{tr}[CAB]$,

$$\left\| G^* \right\|_2 \equiv \left[\mathrm{tr}\left(G^* \gamma^{-1} (G^*)^T \right) \right]^{\frac{1}{2}} = \left[\mathrm{tr}\left((G^*)^T G^* \gamma^{-1} \right) \right]^{\frac{1}{2}}$$

$$\le \left[\left(\mathrm{tr}\left((G^*)^T G^* (G^*)^T G^* \right) \right)^{\frac{1}{2}} \left(\mathrm{tr}(\gamma^{-1} \gamma^{-1}) \right)^{\frac{1}{2}} \right]^{\frac{1}{2}}$$

$$= \left[\mathrm{tr}\left(G^* (G^*)^T \right) \right]^{\frac{1}{2}} \left[\mathrm{tr}(\gamma^{-1}) \right]^{\frac{1}{2}}$$

$$\le \frac{M_v}{a M_G} \bullet M_G = \frac{M_v}{a}$$

which implies

$$\frac{\dot{V} + 2aV}{V^{\frac{1}{2}}} \le \left(1 + \sqrt{p_{\max}} \right) M_v \sqrt{2} \tag{4}$$

From

$$\frac{d}{dt}(2e^{at} V^{\frac{1}{2}}) = e^{at} \frac{\dot{V} + 2aV}{V^{\frac{1}{2}}}$$

$$\le e^{at} \left(1 + \sqrt{p_{\max}} \right) M_v \sqrt{2}$$

Integrating this expression we have:

$$e^{at} V(t)^{1/2} - V(0)^{1/2} \le \frac{\left(1 + \sqrt{p_{\max}} \right) M_v}{a} \left(e^{at} - 1 \right)$$

Therefore,

$$V(t)^{1/2} \le V(0)^{1/2} e^{-at} + \frac{\left(1 + \sqrt{p_{\max}} \right) M_v}{a} \left(1 - e^{-at} \right) \tag{5}$$

The function $V(t)$ is a norm function of the state $e(t)$ and matrix $G(t)$. So, since $V(t)^{1/2}$ is bounded for all t, then $e(t)$ and $G(t)$ are bounded. We also obtain the following inequality:

$$\sqrt{p_{\min}} \left\| e(t) \right\| \le V(t)^{1/2}$$

Substitution of this into (5) gives us an exponential bound on state $e(\tau)$:

$$\left\| e(t) \right\| \le \frac{e^{-at}}{\sqrt{p_{\min}}} V(0)^{1/2} + \frac{\left(1 + \sqrt{p_{\max}} \right) M_v}{a \sqrt{p_{\min}}} \left(1 - e^{-at} \right) \tag{6}$$

Taking the limit superior of (6), we have

$$\overline{\lim_{\tau \to \infty}} \left\| e(t) \right\| \le \frac{\left(1 + \sqrt{p_{\max}} \right)}{a \sqrt{p_{\min}}} M_v \equiv R_* \tag{7}$$

End of Proof.

III. Robust Adaptive Regulation with Disturbance Rejection

In order to accomplish some degree of disturbance rejection in a MRAC system, we shall make use of a definition, given in [7], for the persistent disturbance:

Definition 2: *A disturbance vector* $u_D \in R^q$ *is said to be* ***persistent*** *if it satisfies the* ***disturbance generator equations****:*

$$\begin{cases} u_D(t) = \theta z_D(t) \\ \dot{z}_D(t) = F z_D(t) \end{cases} \text{or} \begin{cases} u_D(t) = \theta z_D(t) \\ z_D(t) = L \phi_D(t) \end{cases}$$

where F is a marginally stable matrix and $\phi_D(t)$ is a vector of known functions forming a basis for all the possible disturbances. This is known as "disturbances with known waveforms but unknown amplitudes".

Consider the Linear Infinite Dimensional Plant with Persistent Disturbances given by:

$$\frac{\partial x}{\partial t}(t) = Ax(t) + Bu(t) + \Gamma u_D(t) \tag{8a}$$

$$Bu \equiv \sum_{i=1}^{m} b_i u_i \tag{8b}$$

$$y(t) = Cx(t), y_i \equiv (c_i, x(t)), i = 1...m \tag{8c}$$

where $x(0) \equiv x_0 \in D(A)$, $x \in D(A)$ is the plant state, $b_i \in D(A)$ are actuator influence functions, $c_i \in D(A)$ are sensor influence functions, $u, y \in \Re^m$ are the control input and plant output m-vectors respectively, u_D is a

disturbance with known basis functions ϕ_D. *We assume the columns of Γ are linear combinations of the columns of B (denoted Span(Γ) $\subseteq$ Span(B)).* This can be relaxed a bit by using ideal trajectories, but we will leave that to another time.

The above system must have *output regulation to a neighborhood:*

$$y \xrightarrow[t\to\infty]{} N(0,R) \tag{9}$$

Since the plant is subjected to unknown bounded signals, we cannot expect better regulation than (9). The adaptive controller will have the form:

$$\begin{cases} u = G_e y + G_D \phi_D \\ \dot{G}_e = -y y^T \gamma_e - a G_e \\ \dot{G}_D = -y \phi_D^T \gamma_D - a G_D \end{cases} \tag{10}$$

Using Theorem 1, we have the following corollary about the corresponding direct adaptive control strategy:

Corollary 1: *Assume the following:*

i.) There exists a gain, G_e^* such that the triple $(A_C \equiv A + BG_e^* C, B, C)$ is SD, i.e. (A,B,C) is ASD,

ii.) A is a densely defined ,closed operator on $D(A) \subseteq X$ and generates a C_0 semigroup of bounded operators $U(t)$,

iii.) Span(Γ) $\subseteq$ Span(B)

Then the output $y(t)$ exponentially approaches a neighborhood with radius proportional to the magnitude of the disturbance, υ, for sufficiently small α and γ_i. Furthermore, each adaptive gain matrix is bounded.

Proof: Since Span(Γ)$\subseteq$Span(B), there exists a transformation G_D^* such that $\Gamma\theta L + BG_D^* = 0$

Let, $G \equiv \begin{bmatrix} G_e & G_D \end{bmatrix}$, $G^* \equiv \begin{bmatrix} G_e^* & G_D^* \end{bmatrix}$, and $\Delta G \equiv G - G^*$. Then $u = G\eta = G_e^* y + G_D^* \phi_D + \Delta G \eta$

where, $\eta \equiv \begin{bmatrix} y \\ \phi_D \end{bmatrix}$. The error differential equation becomes

$$\begin{cases} \dfrac{\partial x}{\partial t} = (\underbrace{A + BG_e^* C}_{A_c}) x + (\underbrace{BG_D^* + \Gamma\theta L}_{=0}) \varphi_D + \nu \\ \quad = A_C x + Bw + \nu \\ w \equiv \Delta G \eta \\ \nu \equiv \text{ bounded signal} \end{cases}$$

Since *B, C* are finite rank operators, so is $BG_e^* C$. Therefore, $A_c \equiv A + BG_e^* C$ with $D(A_c) = D(A)$ generates a C_0 semigroup $U_c(t)$ because A does; see [9] Theo. 2.1 p. 497. Using equations (10), we have

$$\begin{aligned} \Delta\dot{G} &= \dot{G} - \dot{G}^* \\ &= \dot{G} \\ &= -y\eta^T \gamma - aG \end{aligned}$$

where $\gamma \equiv \begin{bmatrix} \gamma_e & 0 \\ 0 & \gamma_D \end{bmatrix} > 0$. By Theorem 1, the corollary follows for α *and* γ_i sufficiently small.

End of Proof.

Corollary 1 provides a control law that is robust with respect to persistent disturbances and unknown bounded disturbances, and, exponentially with rate e^{-at}, produces:

$$\overline{\lim_{\tau\to\infty}} \|y(t)\| \le \frac{\left(1 + \sqrt{p_{\max}}\right)}{a\sqrt{p_{\min}}} \|B\| M_v .$$

IV. Symmetric Hyperbolic Systems

We will illustrate the above robust adaptive controller on the following *m* input, *m* output Symmetric Hyperbolic Problem*:*

$$\begin{cases} \dfrac{\partial x}{\partial t} = Ax + B(u + u_D) + v \\ y = Cx \equiv \begin{bmatrix} (c_1, x) \\ (c_2, x) \\ (c_3, x) \\ \cdots \\ (c_m, x) \end{bmatrix} \end{cases} \tag{11}$$

with inner product $(v,w) \equiv \int_\Omega (v^T w) dz$ and Ω is a bounded open set with smooth boundary, and where

$$B \equiv \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \cdots \\ b_m \end{bmatrix}^T : \Re^m \to X \text{ linear}; b_i \in D(A),$$

$x(0) \equiv x_0 \in D(A) \subseteq X \equiv L_N^2(\Omega)$, and

$C : X \to \Re^m$ linear; $c_i \in D(A)$.

For this application we will *assume the disturbances are step functions*. Note that the disturbance functions can be any

basis function as long as φ_D is bounded, in particular sinusoidal disturbances are often applicable. So we have $\varphi_D \equiv 1$ and $\begin{cases} u_D = (1) z_D \\ \dot{z}_D = (0) z_D \end{cases}$ which implies $F = 0$ and $\theta_D = 1$.

Let the adaptive control law be $u = G_e y + G_D$ with

$$\begin{cases} \dot{G}_e = -yy^T \gamma_e - \alpha G_e \\ \dot{G}_D = -y\gamma_D - \alpha G_D \end{cases}.$$

Now we define the closed linear operator A with domain $D(A)$ dense in the Hilbert space $X \equiv L^2(\Omega)$ with inner product $(v_1, v_2) \equiv \int_\Omega (v_1^T v_2) dz$ as:

$$Ax \equiv \sum_{i=1}^{N} A_i \frac{\partial x}{\partial z_i} + A_0 x$$

where A_i are *NxN* symmetric constant matrices, A_0 is a real *NxN* constant matrix, and x is an *Nx1* column vector of functions.

Thus (11) is a symmetric Hyperbolic System of first order partial differential equations with

$$A(\xi) \equiv \sum_{i=1}^{N} \xi_i A_i$$

which is an NxN symmetric matrix [15]. The Boundary Conditions which define the operator domain $D(A)$ will be *coercive*, i.e. $h^T n = 0$ where

$h(x) \equiv \frac{1}{2}\left[x^T A_1 x \quad x^T A_2 x \quad x^T A_3 x \quad \ldots \quad x^T A_N x \right]$ and $n(z)$ is the outward normal vector on boundary $\partial\Omega$ of the domain $\Omega \subseteq \Re^N$.

Now use

$u = G_e y + G_D \phi_D = G_e^* y + \underbrace{G_D^* \phi_D}_{u_D} + \underbrace{\Delta G_D \eta}_{w}$ where

$$\eta \equiv \begin{bmatrix} y \\ \phi_D \end{bmatrix}$$

which implies

$$x_t = [\underbrace{Ax + BG_e^* Cx}_{A_c x}] + Bw + v$$

which implies

$A_c = A + BG_e^* C$.

Since the boundary conditions are coercive, we use the Divergence Theorem to obtain

$$\begin{aligned} (A_c x, x) &= (Ax, x) + (BG_e^* Cx, x) \\ &= \int_\Omega (x^T \sum_{i=1}^{N} A_i \frac{\partial x}{\partial z_i}) dz + (A_0 x, x) + (BG_e^* Cx, x) \\ &= \frac{1}{2} \int_\Omega \underset{Div(h)}{(\nabla \circ h)} dz + (A_0 x, x) + (BG_e^* Cx, x) \\ &= \frac{1}{2} \int_\Omega \underset{=0}{(h^T n)} dz + (A_0 x, x) + (BG_e^* Cx, x) \\ &= (A_0 x, x) + (BG_e^* Cx, x) \end{aligned}$$

Assume $b_i = c_i$ or $B^* = C$ and $G_e^* \equiv -g_e^* < 0$.

Then we have

$$\begin{aligned} (A_c x, x) &= (A_0 x, x) + (BG_e^* Cx, x) \\ &= (A_0 x, x) - g_e^* (Cx, B^* x) \\ &= (A_0 x, x) - g_e^* \|Cx\|^2 \\ &\leq 0 \end{aligned}$$

which implies

$\mathrm{Re}(A_c x, x) = (A_0 x, x) - g_e^* \|Cx\|^2$ and $B^* = C$

which is **not** quite strictly dissipative.

But we have the following result:

Theorem 2: *A_c has compact resolvent; hence it has discrete spectrum, in the sense that it consists only of isolated eigenvalues with finite multiplicity.*

Proof*:* See Appendix I.

Consider that

$X = E_s \oplus E_u$ where E_s is the stable eigenspace and E_u is the unstable eigenspace with corresponding projections P_s, P_u. Assume that

$\dim E_u \equiv N_u$ and $E_u^\perp = E$. This implies that

P_s, P_u are bounded self adjoint operators.

Choose $C \equiv P_u$; this is possible when the unstable subspace is finite-dimensional.

Then we have the following result:

Theorem 3:

$\mathrm{Re}(A_0 x, x) \leq -\alpha \|P_s x\|^2$ for all $x \in D(A)$

implies that (A, B, C) is almost strictly dissipative (ASD).

Proof*:*

$$\begin{aligned}\mathrm{Re}(A_c x, x) &= \mathrm{Re}(Ax, x) - g_e^* \|Cx\|^2 \\ &= \mathrm{Re}(A_0 x, x) - g_e^* \|Cx\|^2 \\ &\le -\alpha \|P_s x\|^2 - g_e^* \|P_u x\|^2 \\ &\le -\alpha \left(\|P_s x\|^2 + \|P_u x\|^2 \right) \\ &= -\alpha \|x\|^2\end{aligned}$$

by choosing $g_e^* \ge \alpha$.

Therefore, $\mathrm{Re}(A_c x, x) \le -\alpha \|x\|^2$ and $C = B^*$ implies that there exists

$G_e^* \equiv -g_e^* < 0$ such that

$(A_c = A + BG_e^* C, B, C)$ is strictly dissipative with $P = I$.

Here is a simple first order symmetric hyperbolic system example to illustrate some of the above:

$$\begin{cases} x_t = \underbrace{\begin{bmatrix} \varepsilon & 1 \\ 1 & 0 \end{bmatrix}}_{A_1} x_z + \underbrace{\begin{bmatrix} -\varepsilon & 0 \\ 0 & 0 \end{bmatrix}}_{A_0} x + \underbrace{\begin{bmatrix} 0 \\ 1 \end{bmatrix}}_{B} \left(u + u_D \right) \\ y = \underbrace{\begin{bmatrix} 0 & 1 \end{bmatrix}}_{C} x \end{cases}$$

where $\varepsilon > 0$ is small.

If we use

$G_e^* = -g_* < 0$

this implies that

$$\begin{aligned}\mathrm{Re}\left(A_c x, x\right) &= (A_0 x, x) - g_e^* \|Cx\|^2 \\ &= -\varepsilon \|q_1\|^2 - g_* \|q_2\|^2 \\ &\le -\underbrace{\min(g_*, \varepsilon)}_{\alpha > 0} \left(\|q_1\|^2 + \|q_2\|^2 \right) \\ &\le -\alpha \|x\|^2\end{aligned}$$

where $x \equiv \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}$.

Then $(A_c = A + BG_e^* C, B, C)$ is strictly dissipative with $P = I$ and we can apply Theo. 1 and Cor. 1.

V. Conclusions

In Theorem 1 we proved a robustness result for adaptive control under the hypothesis of almost strict dissipativity for infinite dimensional systems. This idea is an extension of the concept of m-accretivity for infinite dimensional systems; see [9] pp278-280. In Cor 1, we showed that adaptive regulation to a neighborhood was possible with an adaptive controller modified with a leakage term. This controller could also mitigate persistent disturbances. The results in Theo. 1 can be easily extended to cause model tracking instead of regulation. Also we can relax the requirement that the disturbance enters through the same channels as the control.

We applied these results to general symmetric hyperbolic systems using m actuator and m sensors and adaptive output feedback.We showed that under some limitations on operator spectrum that we can accomplish robust adaptive control. This allows the possibility of rather simple direct adaptive control which also mitigates persistent disturbances for a large class of applications in wave behavior, electromagnetic fields, and some quantum fields.

Appendix

Proof of Theorem 2.

We will **assume** the operator A is closed. If not we can work with the closure of A. It is easy to see that A is skew self-adjoint, i.e. $(Af, g) = -(f, Ag)$ for all $f, g \in D(A)$. Therefore the spectrum of A must lie on the imaginery axis and any complex $\lambda \in \rho(A)$ if it has nonzero real part.

With coercive boundary conditions, it can be shown that for all $x \in D(A)$, we have

$$\sum_{i-1}^{N} \left\| \frac{\partial x}{\partial z_i} \right\|^2 \le K^2 \left(\|Ax\|^2 + \|x\|^2 \right)$$

and so $D(A) \subseteq H^1$.

Let $\lambda \in \rho(A)$, the resolvent set of A and consider the resolvent operator

$$R(\lambda) \equiv (\lambda I - A)^{-1} : X = L^2(\Omega) \to D(A)$$

.

We want to show that this resolvent operator is (sequentially) compact:

Take a bounded sequence

$\{x_k\}_{k=1}^{\infty} \subseteq X$ and define $h_k \equiv R(1) x_k \in D(A)$.

Then $\sum_{i-1}^{N} \left\| \frac{\partial h_k}{\partial z_i} \right\|^2 \le K^2 \left(\|Ah_k\|^2 + \|h_k\|^2 \right)$.

But $h_k - Ah_k = x_k$ or $Ah_k = h_k - x_k$

which implies

$$\|Ah_k\|^2 = \|h_k - x_k\|^2 \le \left(\|h_k\| + \|x_k\| \right)^2.$$

Therefore

$$\begin{aligned}\|h_k\|_1^2 &\equiv \|h_k\|^2 + \sum_{i=1}^{N}\left\|\frac{\partial h_k}{\partial z_i}\right\|^2 \\ &\le \|h_k\|^2 + K^2\left(\|Ah_k\|^2 + \|h_k\|^2\right) \\ &\le \|h_k\|^2 + K^2\left(\left(\|h_k\| + \|x_k\|\right)^2 + \|h_k\|^2\right) \\ &= (1+K^2)\|h_k\|^2 + K^2\left(\|h_k\| + \|x_k\|\right)^2\end{aligned}$$

Now $\{x_k\}$ is bounded by assumption and

$h_k \equiv R(1)x_k$ is bounded because $R(1)$ is a bounded operator.

Therefore, $\|h_k\|_1$ is a bounded and so $\{\mathrm{h_k}\}_{k=1}^{\infty}$ is a bounded sequence in H_1.

Consequently by the Rellich Compactness Theorem (see e.g [19]Theo. 8.38 p175 or [20] Theo. 2 p246),

H^1 is compactly embedded in $X \equiv L^2(\Omega)$ because Ω is a bounded open set with smooth boundary.

Therefore there exists a convergent subsequence of $\{h_k\}_{k=1}^{\infty}$ in $X \equiv L^2(\Omega)$ and the resolvent operator $R(1) = (I-A)^{-1}$ is a compact operator. Then, by Theo. 6.29 [9]p187, $R(\lambda)$ is compact for all $\lambda \in \rho(A)$ and the spectrum is discrete, in the sense that it consists only of isolated eigenvalues with finite multiplicities.

End of Proof.

REFERENCES

[1] A. Pazy,Semigroups of Linear Operators and Applications to partial Differential Equations,Springer 1983.

[2] D. D'Alessandro, Introduction to Quantum Control and Dynamics, Chapman &Hall, 2008.

[3] Balas, M., R. S. Erwin, and R. Fuentes, "Adaptive control of persistent disturbances for aerospace structures", AIAA GNC, Denver, 2000.

[4] R. Fuentes and M. Balas, "Direct Adaptive Rejection of Persistent Disturbances", Journal of Mathematical Analysis and Applications, Vol 251, pp 28-39, 2000

[5] Fuentes, R and M. Balas, "Disturbance accommodation for a class of tracking control systems", AIAA GNC, Denver, Colorado, 2000.

[6] Fuentes, R. and M. Balas, "Robust Model Reference Adaptive Control with Disturbance Rejection", Proc. ACC, 2002.

[7] M. Balas, S. Gajendar, and L. Robertson, "Adaptive Tracking Control of Linear Systems with Unknown Delays and Persistent Disturbances (or Who You Callin' Retarded?)", Proceedings of the AIAA Guidance, Navigation and Control Conference, Chicago, IL,Aug 2009.

[8] Wen, J., "Time domain and frequency domain conditions for strict positive realness", IEEE Trans Automat. Contr., vol. 33, no. 10, pp.988-992, 1988.

[9] T. Kato, Perturbation Theory for Linear Operators ,corrected 2nd edition, Springer, 1980.

[10] R. Curtain and A. Pritchard, Functional Analysis in Modern Applied Mathematics, Academic Press, 1977.

[11] M. Renardy and R. Rogers, An Introduction to Partial Differential Equations,Springer, 1993.

[12] M. Balas, "Trends in Large Space Structure Control Theory: Fondest Hopes, Wildest Dreams", IEEE Trans Automatic Control, AC-27, No. 3, 1982.

[13] M.Balas and R. Fuentes, "A Non-Orthogonal Projection Approach to Characterization of Almost Positive Real Systems with an Application to Adaptive Control", Proc of American Control Conference, 2004.

[14] P. Antsaklis and A. Michel, A Linear Systems Primer,Birkhauser, 2007.

[15] L. Bers, F. John, and M. Schecter, Partial Differential Equations, J. Wiley and Sons, New York, 1964.

[16] F. Treves, Basic Linear partial Differential Equations, Academic Press, New York, 1975.

[17] M. Taylor, Partial Differential Equations: Basic Theory, Springer, New York, 1996.

[18] S. Schweber, An Introduction to Relativistic Quantum Field Theory, Dover, Mineola, NY, 1989.

[19] A. Bressan, Lecture Notes on Functional Analysis with Applications to Linear Partial Differential Equations, American Mathematical Society, Vol 143, providence, RI, 2013.

[20] P. Lax, Functional Analysis, J. Wilet and Sons, NY, 2002.

[21] R. Curtain, M. Demetriou, and K. Ito, "Adaptive Compensators for perturbed positive Real Infinite Dimensional Systems",AFOSR Report, 03 June 1998.

Mathematical modeling of crown forest fires spread taking account firebreaks

Valeriy Perminov

National Research Tomsk Polytechnic University, e-mail: perminov@tpu.ru

Abstract

It is developed mathematical model of heat and mass transfer processes at crown forest fire spread which takes into account fire breaks. The paper gives a new mathematical setting and method of numerical solution of this problem. It is based on numerical solution of two dimensional Reynolds equations for turbulent flow taking into account diffusion equations for chemical components and equations of energy conservation for gaseous and condensed phases. To obtain discrete analogies a method of finite volume was used. Numerical solutions of crown forest fire propagation taking account breaks and glades were found. It possible to obtain a detailed picture of the change in the temperature and component concentration fields with time, and determine as well as the limiting condition of fire propagation in forest with these firebreaks.

Keywords— control volume, crown fire, fire spread, forest fire, mathematical model, numerical method.

1. INTRODUCTION

Many mathematical models have been developed to calculate forward rates of spread for various fuel complexes. A detailed list is given Grishin A.M.[1]. Crown fires are initiated by convective and radiative heat transfer from surface fires. However, convection is the main heat transfer mechanism [2]. The theory proposed by Van Wagner depends on three simple crown properties: crown base height, bulk density of forest combustible materials and moisture content of forest fuel. Also, crown fire initiation and hazard have been studied and modelled in details later by another authors [3-9]. The more complete discussion of the problem of crown forest fires is provided by co-workers at Tomsk University [1,10]. In particular, a mathematical model of forest fires was obtained by Grishin [1] based on an analysis of known and original experimental data Konev [11], and using concepts and methods from reactive media mechanics. The physical two-phase models used by Morvan and Dupuy [12, 13] may be considered as a continuation and extension of the formulation proposed by A.M. Grishin[1].

This study gives a two dimensional averaged mathematical setting and method of numerical solution of a problem of a forest fire spread. The boundary-value problem is solved numerically using the method of splitting according to physical processes. It was based on numerical solution of two dimensional Reynolds equations for the description of turbulent flow taking into account for diffusion equations chemical components and equations of energy conservation for gaseous and condensed phases, volume of fraction of condensed phase (dry organic substance, moisture, condensed pyrolysis products, mineral part of forest fuel). One aspect of this research is to study of the conditions when the forest fire spreads through firebreaks and glades. The purpose of this paper is to describe detailed picture of the change in the temperature and component concentration fields with time, and determine as well as the limiting condition of fire propagation in forest with fire breaks.

2. MATHEMATICAL MODEL

It is assumed that the forest during a forest fire can be modeled as 1) a multi-phase, multistoried, spatially heterogeneous medium; 2) in the fire zone the forest is a porous-dispersed, two-temperature, single-velocity, reactive medium; 3) the forest canopy is supposed to be non - deformed medium (trunks, large branches, small twigs and needles), which affects only the magnitude of the force of resistance in the equation of conservation of momentum in the gas phase, i.e., the medium is assumed to be quasi-solid (almost non-deformable during wind gusts); 4) let there be a so-called "ventilated" forest massif, in which the volume of fractions of condensed forest fuel phases, consisting of dry organic matter, water in liquid state, solid pyrolysis products, and ash, can be neglected compared to the volume fraction of gas phase (components of air and gaseous pyrolysis products); 5) the flow has a developed turbulent nature and molecular transfer is neglected; 6) gaseous phase density doesn't depend on the pressure because of the low velocities of the flow in comparison with the velocity of the sound. Let the point x_1, x_2 , x_3= 0 is situated at the centre of the surface forest fire source at the height of the roughness level, axis $0x_1$ directed parallel to the Earth's surface to the right in the direction of the unperturbed wind speed, axis $0x_2$ directed perpendicular to $0x_1$ and axis $0x_3$ directed upward (Fig. 1).

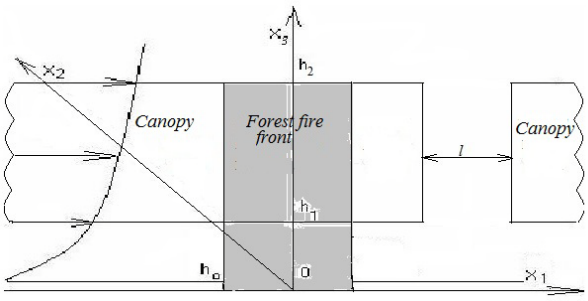


Fig.1. Scheme of the domain.

Because of the horizontal sizes of forest massif more than height of forest, system of equations of general mathematical model of forest fire was integrated between

the limits from height of the roughness level - 0 to top boundary of forest crown.

$$\frac{\partial \rho}{\partial t}+\frac{\partial}{\partial x_j}(\rho v_j)=Q-(\dot{m}^- -\dot{m}^+)/h,\ j=1,2,3; \quad (1)$$

$$\rho\frac{dv_i}{dt}=-\frac{\partial p}{\partial x_j}+\frac{\partial}{\partial x_j}(-\rho\overline{v_i'v_j'})-\rho s c_d v_i|\vec{v}|-\rho g_i- \quad (2)$$
$$-Qv_i+(\tau_i^- -\tau_i^+)/h,\ i=1,2,3;$$

$$\rho c_p\frac{dT}{dt}=\frac{\partial}{\partial x_j}(-\rho c_p\overline{v_j'T'})+q_5R_5-\alpha_v(T-T_s)+ \quad (3)$$
$$-(q_T^- -q_T^+)/h;$$

$$\rho\frac{dc_\alpha}{dt}=\frac{\partial}{\partial x_j}(-\rho\overline{v_j'c_\alpha'})+R_{5\alpha}-Qc_\alpha+ \quad (4)$$
$$+(J_\alpha^- -J_\alpha^+)/h,\ \alpha=1,5;$$

$$\frac{\partial}{\partial x_j}\left(\frac{c}{3k}\frac{\partial U_R}{\partial x_j}\right)-k(cU_R-4\sigma T_S^4)+(q_R^- -q_R^+)/h=0; \quad (5)$$

$$\sum_{i=1}^{4}\rho_i c_{pi}\varphi_i\frac{\partial T_S}{\partial t}=q_3R_3-q_2R_2+k(cU_R-4\sigma T_S^4) \quad (6)$$
$$+\alpha_V(T-T_S);$$

$$\sum_{\alpha=1}^{5}c_\alpha=1, P_e=\rho RT\sum_{\alpha=1}^{5}\frac{c_\alpha}{M_\alpha}, \vec{g}=(0,0,g), \sum_{i=1}^{5}\varphi_i=1.$$

The system of equations (1)–(6) must be solved taking into account the initial and boundary conditions

$$t=0: v_1=0, v_2=0, v_3=0, T=T_e, c_\alpha=c_{\alpha e}, T_s=T_e, \varphi_1=\varphi_{ie}; \quad (7)$$

$$x_1=-x_{1e}: v_1=V_e, v_2=0, \frac{\partial v_3}{\partial x_1}=0, T=T_e, c_\alpha=c_{\alpha e}, \quad (8)$$
$$-\frac{c}{3k}\frac{\partial U_R}{\partial x_1}+cU_R/2=0;$$

$$x_1=x_{1e}: \frac{\partial v_1}{\partial x_1}=0,\ \frac{\partial v_2}{\partial x_1}=0, \frac{\partial v_3}{\partial x_1}=0,\ \frac{\partial c_\alpha}{\partial x_1}=0, \quad (9)$$
$$\frac{\partial T}{\partial x_1}=0, \frac{c}{3k}\frac{\partial U_R}{\partial x_1}+\frac{c}{2}U_R=0;$$

$$x_2=x_{20}: \frac{\partial v_1}{\partial x_2}=0,\ \frac{\partial v_2}{\partial x_2}=0,\ \frac{\partial v_3}{\partial x_2}=0, \frac{\partial c_\alpha}{\partial x_2}=0, \quad (10)$$
$$\frac{\partial T}{\partial x_2}=0, -\frac{c}{3k}\frac{\partial U_R}{\partial x_2}+\frac{c}{2}U_R=0;$$

$$x_2=x_{2e}: \frac{\partial v_1}{\partial x_2}=0,\ \frac{\partial v_2}{\partial x_2}=0, \frac{\partial v_3}{\partial x_2}=0, \frac{\partial c_\alpha}{\partial x_2}=0, \quad (11)$$
$$\frac{\partial T}{\partial x_2}=0, \frac{c}{3k}\frac{\partial U_R}{\partial x_2}+\frac{c}{2}U_R=0.$$

$$x_3=0: v_1=0,\ v_2=0, \frac{\partial c_\alpha}{\partial x_3}=0, -\frac{c}{3k}\frac{\partial U_R}{\partial x_3}+\frac{c}{2}U_R=0, \quad (12)$$
$$v_3=v_{30}, T=T_g, |x_1|\le\Delta, |x_2|\le\Delta$$
$$v_3=0,\ \ T=T_e, |x_1|>\Delta, |x_2|>\Delta;$$

$$x_3=x_{3e}: \frac{\partial v_1}{\partial x_3}=0,\ \frac{\partial v_2}{\partial x_3}=0, \frac{\partial v_3}{\partial x_3}=0, \frac{\partial c_\alpha}{\partial x_3}=0, \quad (13)$$
$$\frac{\partial T}{\partial x_3}=0, \frac{c}{3k}\frac{\partial U_R}{\partial x_3}+\frac{c}{2}U_R=0.$$

Here and above $\frac{\mathrm{d}}{\mathrm{d}t}$ is the symbol of the total (substantial) derivative; α_V is the coefficient of phase exchange; ρ - density of gas – dispersed phase, t is time; v_i - the velocity components; T, T_S, - temperatures of gas and solid phases, U_R - density of radiation energy, k - coefficient of radiation attenuation, P - pressure; c_p – constant pressure specific heat of the gas phase, c_{pi}, ρ_i, φ_i – specific heat, density and volume of fraction of condensed phase (1 – dry organic substance, 2 – moisture, 3 – condensed pyrolysis products, 4 – mineral part of forest fuel, 5 – gas phase), R_i – the mass rates of chemical reactions, q_i – thermal effects of chemical reactions; k_g , k_S - radiation absorption coefficients for gas and condensed phases; T_e - the ambient temperature; c_α - mass concentrations of α - component of gas - dispersed medium, index α=1,2,3, where 1 corresponds to the density of oxygen, 2 - to carbon monoxide *CO*, 3 - to carbon dioxide and inert components of air, 4 - to particles of black, 5 - to particles of smoke; R – universal gas constant; M_α , M_C, and M molecular mass of α - components of the gas phase, carbon and air mixture; g is the gravity acceleration; c_d is an empirical coefficient of the resistance of the vegetation, s is the specific surface of the forest fuel in the given forest stratum. In system of equations (1)-(6) are introduced the next designations:

$$\dot{m}=\rho v_3, \tau_i=-\rho\overline{v_i'v_3'}, J_\alpha=-\rho\overline{v_3'c_\alpha'}, J_T=-\rho\overline{v_3'T'}$$

Upper indexes "+" and "-" designate values of functions at $x_3=h$ and $x_3=0$ correspondingly. It is assumed that heat and mass exchange of fire front and boundary layer of atmosphere are governed by Newton law and written using the formulas

$$(q_T^- -q_T^+)/h=-\alpha(T-T_e)/h,$$
$$(J_\alpha^- -J_\alpha^+)/h=-\alpha(c-c_{\alpha e})/hc_p.$$

To define source terms which characterize inflow (outflow of mass) in a volume unit of the gas-dispersed phase, the following formulae were used for the rate of formulation of the gas-dispersed mixture $\dot{m}$, outflow of oxygen R_{51}, changing carbon monoxide R_{52}

$$Q=(1-\alpha_c)R_1+R_2+\frac{M_c}{M_1}R_3, R_{51}=-R_3-\frac{M_1}{2M_2}R_5,$$
$$R_{52}=\nu_g(1-\alpha_c)R_1-R_5, R_{53}=0.$$

Here ν_g – mass fraction of gas combustible products of pyrolysis, α_4 and α_5 – empirical constants. Reaction rates

of these various contributions (pyrolysis, evaporation, combustion of coke and volatile combustible products of pyrolysis) are approximated by Arrhenius laws whose parameters (pre-exponential constant k_i and activation energy E_i) are evaluated using data for mathematical models [1].

$$R_1 = k_1\rho_1\varphi_1 \exp\left(-\frac{E_1}{RT_s}\right), R_2 = k_2\rho_2\varphi_2 T_s^{-0.5} \exp\left(-\frac{E_2}{RT_s}\right),$$

$$R_3 = k_3\rho\varphi_3 s_\sigma c_1 \exp\left(-\frac{E_3}{RT_s}\right), R_5 = k_5 M_2 \left(\frac{c_1 M}{M_1}\right)^{0.25} \frac{c_2 M}{M_2} T^{-2.25} \exp\left(-\frac{E_5}{RT}\right).$$

The initial values for volume of fractions of condensed phases are determined using the expressions:

$$\varphi_{1e} = \frac{d(1-\nu_z)}{\rho_1}, \varphi_{2e} = \frac{Wd}{\rho_2}, \varphi_{3e} = \frac{\alpha_c\varphi_{1e}\rho_1}{\rho_3}$$

where d -bulk density for surface layer, ν_z – coefficient of ashes of forest fuel, W – forest fuel moisture content. It is supposed that the optical properties of a medium are independent of radiation wavelength (the assumption that the medium is "grey"), and the so-called diffusion approximation for radiation flux density were used for a mathematical description of radiation transport during forest fires. To close the system (1)–(6), the components of the tensor of turbulent stresses, and the turbulent heat and mass fluxes are determined using the local-equilibrium model of turbulence [1]. The system of equations (1)–(6) contains terms associated with turbulent diffusion, thermal conduction, and convection, and needs to be closed. The components of the tensor of turbulent stresses $\rho\overline{v_i'v_j'}$, as well as the turbulent fluxes of heat and mass $\overline{\rho v_j' c_p T'}$, $\overline{\rho v_j' c_\alpha'}$ are written in terms of the gradients of the average flow properties using the formulas

$$-\rho\overline{v_i'v_j'} = \mu_t\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right) - \frac{2}{3}K\delta_{ij},$$

$$-\rho\overline{v_j c_p T'} = \lambda_t \frac{\partial T}{\partial x_j}, \; -\rho\overline{v_j c_\alpha'} = \rho D_t \frac{\partial c_\alpha}{\partial x_j},$$

$$\lambda_t = \mu_t c_p / \mathrm{Pr}_t, \rho D_t = \mu_t / Sc_t, \mu_t = c_\mu \rho K^2/\varepsilon,$$

where μ_t, λ_t, D_t are the coefficients of turbulent viscosity, thermal conductivity, and diffusion, respectively; Pr_t, Sc_t are the turbulent Prandtl and Schmidt numbers, which were assumed to be equal to 1. In dimensional form, the coefficient of dynamic turbulent viscosity is determined using local equilibrium model of turbulence [1]. The thermodynamic, thermophysical and structural characteristics correspond to the forest fuels in the canopy of a different (for example pine) type of forest. The system of equations (1)–(6) must be solved taking into account the initial and boundary conditions.

3. NUMERICAL METHOD AND RESULTS

The boundary-value problem (1)–(6) is solved numerically using the method of splitting according to physical processes. In the first stage, the hydrodynamic pattern of flow and distribution of scalar functions was calculated. The system of ordinary differential equations of chemical kinetics obtained as a result of splitting was then integrated. A discrete analog was obtained by means of the control volume method using the SIMPLE like algorithm [15]. The accuracy of the program was checked by the method of inserted analytical solutions. Analytical expressions for the unknown functions were substituted in (1)–(6) and the closure of the equations were calculated. This was then treated as the source in each equation. Next, with the aid of the algorithm described above, the values of the functions used were inferred with an accuracy of not less than 1%. The effect of the dimensions of the control volumes on the solution was studied by diminishing them. The time step was selected automatically.

Fields of temperature, velocity, component mass fractions, and volume fractions of phases were obtained numerically. The distribution of basic functions shows that the process of crown forest fire initiation goes through the next stages. The first stage is related to increasing maximum temperature in the fire source. At this process stage the fire source a thermal wind is formed a zone of heated forest fire pyrolysis products which are mixed with air, float up and penetrate into the crowns of trees. As a result, forest fuels in the tree crowns are heated, moisture evaporates and gaseous and dispersed pyrolysis products are generated. Ignition of gaseous pyrolysis products of the ground cover occurs at the next stage, and that of gaseous pyrolysis products in the forest canopy occurs at the last stage. As a result of heating of forest fuel elements of crown, moisture evaporates, and pyrolysis occurs accompanied by the release of gaseous products, which then ignite and burn away in the forest canopy. At the moment of ignition the gas combustible products of pyrolysis burns away, and the concentration of oxygen is rapidly reduced. The temperatures of both phases reach a maximum value at the point of ignition. The ignition processes is of a gas - phase nature. Note also that the transfer of energy from the fire source takes place due to radiation; the value of radiation heat flux density is small compared to that of the convective heat flux. At $V_e \neq 0$, the wind field in the forest canopy interacts with the gas-jet obstacle that forms from the forest fire source and from the ignited forest canopy and burn away in the forest canopy. Figures 2 - 5 present the distribution of temperature $\overline{T}\,(\overline{T} = T/T_e, T_e = 300K)$ (1- 2., 2 – 2.6, 3 – 3, 4 – 3.5, 5 – 4.) for gas phase, concentrations of oxygen $\overline{c}_1$ (1 – 0.1, 2 – 0.5, 3 – 0.6, 4 – 0.7, 5 – 0.8, 6 – 0.9) and volatile combustible products of pyrolysis $\overline{c}_2$ (1 – 1., 2- 0.1, 3 – 0.05, 4 – 0.01) ($\overline{c}_\alpha = c_\alpha / c_{1e}, c_{1e} = 0.23$) and temperature of condensed phase $\overline{T}_S\,(\overline{T}_S = T_S/T_e, T_e = 300K)$ (1- 2., 2 – 2.6, 3 – 3, 4 – 3.5, 5 – 4.) for wind velocity V_e= 10 m/s at h=*10* m: 1) t=3 sec., 2) t=10 sec, 3) t=18 sec., 4) t= 24 sec. The boundary-value problem is solved numerically using the

method of splitting according to physical processes. Fields of temperature, velocity, component mass fractions, and volume fractions of phases were obtained numerically. As a result of heating of forest fuel elements of crown, moisture evaporates, and pyrolysis occurs accompanied by the release of gaseous products, which then ignite and burn away in the forest canopy. At the moment of ignition the gas combustible products of pyrolysis burns away, and the concentration of oxygen is rapidly reduced. The temperatures of both phases reach a maximum value at the point of ignition. The ignition processes is of a gas - phase nature. Note also that the transfer of energy from the fire source takes place due to radiation; the value of radiation heat flux density is small compared to that of the convective heat flux. At $V_e \neq 0$, the wind field in the forest canopy interacts with the gas-jet obstacle that forms from the forest fire source and from the ignited forest canopy and burn away in the forest canopy. The isotherms and lines of equal levels of gas phase components concentrations were deformed and moved in the forest canopy by the action of wind. It is concluded that the forest fire begins to spread. Mathematical model and the result of the calculation give an opportunity to consider forest fire spread for different wind velocity, canopy bulk densities and moisture forest fuel. The results obtained in this papers show the decrease of the wind induces a decrease of the rate of fire spread.

One of the objectives of this paper could be to develop modeling means to reduce forest fire hazard in forest or near towns. In this paper it presents numerical results to study forest fire propagation through fire breaks and around glades. The results of numerical calculation present the forest fire front movement using distributions of temperature at different instants of time for various sizes of firebreaks. The fire break is situated in the middle of domain. In the first case the fire could not spread through this fire break. If the fire break reduces to 4 meters the fire continue to spread but the isotherms of forest fire is decreased after overcoming of fire break. The dependence of critical fire break value for different wind velocities is presented in paper. Of course the size of safe distance depends not only of wind velocity, but type and quality of forest combustible materials, its moisture, height of trees and others conditions. This model allows studying an influence all these main factors. The isotherms and lines of equal levels are moved in the forest canopy and deformed by the action of wind. Similarly, the fields of component concentrations are deformed. It is concluded that the forest fire begins to spread.

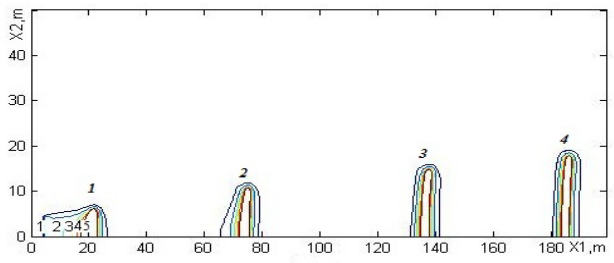


Fig.2. Field of isotherms of the forest fire spread (gas phase).

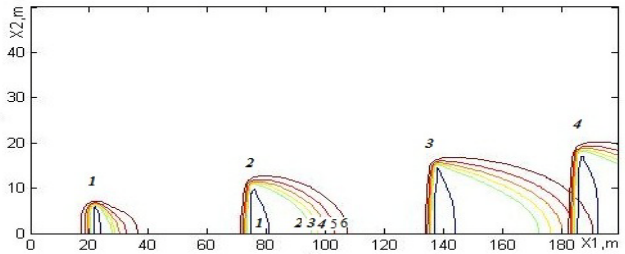


Fig.3. The distribution of oxygen $\overline{C}_1$

Mathematical model and the result of the calculation give an opportunity to consider forest fire spread for different wind velocity. Figures 6 (*a*, *b*, *c*, *d*) present the distribution of temperature for gas phase, concentration of oxygen and volatile combustible products of pyrolysis $\overline{C}_2$ concentrations and temperature of condensed phase for wind velocity V_e= 5 m/s at *h=10* m: 1) *t*=3 sec., 2) *t*=10 sec, 3) *t*=18 sec., 4) *t*= 20 sec., 5) *t*=31 sec., 6) *t*= 40 sec.

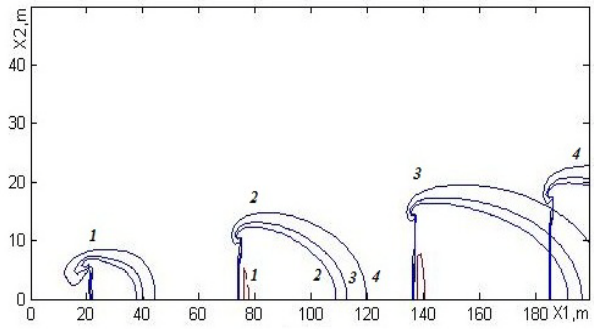


Fig.4 The distribution of $\overline{C}_2$

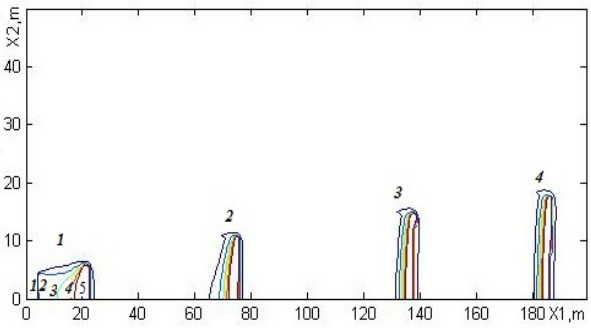


Fig.5 Field of isotherms of the forest fire spread (solid phase).

The results reported in Fig. 6 show the decrease of the wind induces a decrease of the rate of fire spread. One of the objectives of this paper could be to develop modeling means to reduce forest fire hazard in forest or near towns. In this paper it presents numerical results to study forest fire propagation through firebreak. This problem was considered by Zverev [16] in one dimensional mathematical model approach.

ISBN: 978-1-61804-240-8

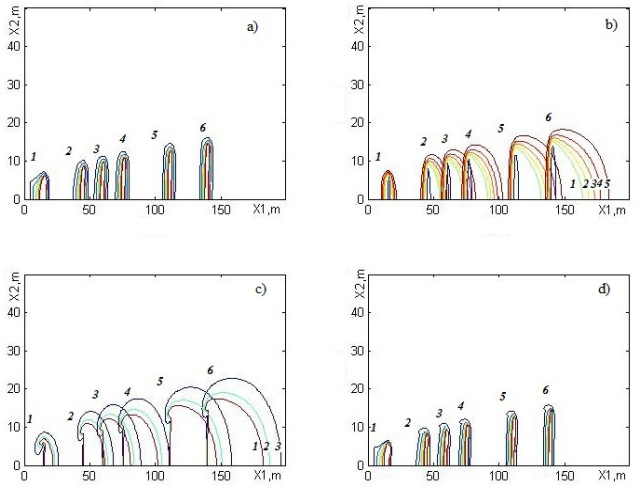


Fig.6 Fields of isotherms of gas (a) and solid phase (d), isolines of oxygen(b) and gas products of pyrolysis(c).

Figures 7 and 8 (Figure 8 b is a continuation of Figure 8 a) present the forest fire front movement using distributions of temperature at different instants of time for two sizes of firebreaks (4.5 and 4 meters). The fire break is situated in the middle of domain ($x_1 = 100\ m$). In the first case the fire could not spread through this fire break.

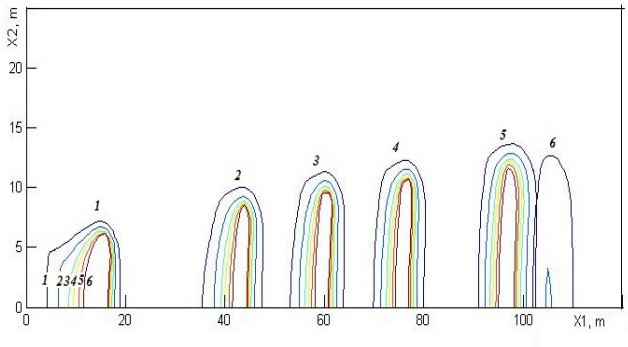


Fig.7 Fields of isotherms for firebreak equals 4.5 m.

If the fire break reduces to 4 meters (Figure 8) the fire continue to spread but the isotherm (isotherm 5) of forest fire is decreased after overcoming of fire break. In the Figure 9. the dependence of critical fire break value for different wind velocities is presented. Of course the size of safe distance depends not only of wind velocity, but type and quality of forest combustible materials, its moisture, height of trees and others conditions. This model allows to study an influence all these main factors.

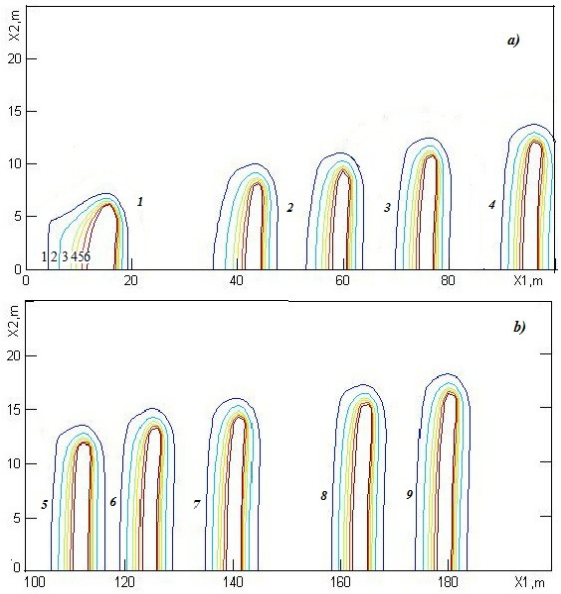


Fig.8 Field of isotherms. Firebreak equals 4 m.

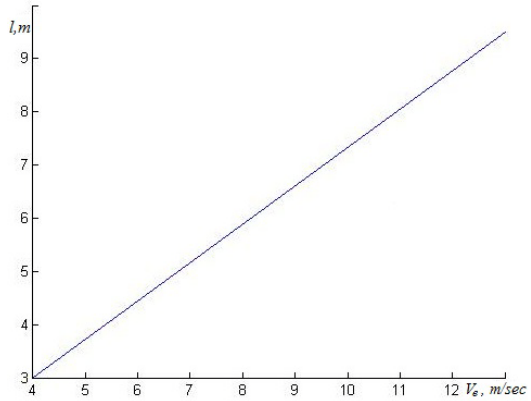


Fig.9 The influence of wind velocity at the size of firebreak.

Figure 10(a, b, c) show the results of numerical simulation of a forest fire spreading around the glade under the action of wind blowing through it at a speed 5 m/s in the direction of the Ox_1-axis. Initially, the source of the fire has the shape of a rectangular. Then isotherms are deformed under the action of wind and the contour of forest fire is look as crescent (Fig. 10 a, curves I). When the fire (isotherms II in Fig.10 a) moves around the forest glade it is divided in two parts. But after that two fire fronts were joined in united fire (isotherms III in Fig.10 a). Figures 10 (b, c) present the distribution of concentration of oxygen and volatile combustible products of pyrolysis $\overline{C}_2$ for this case.

ISBN: 978-1-61804-240-8

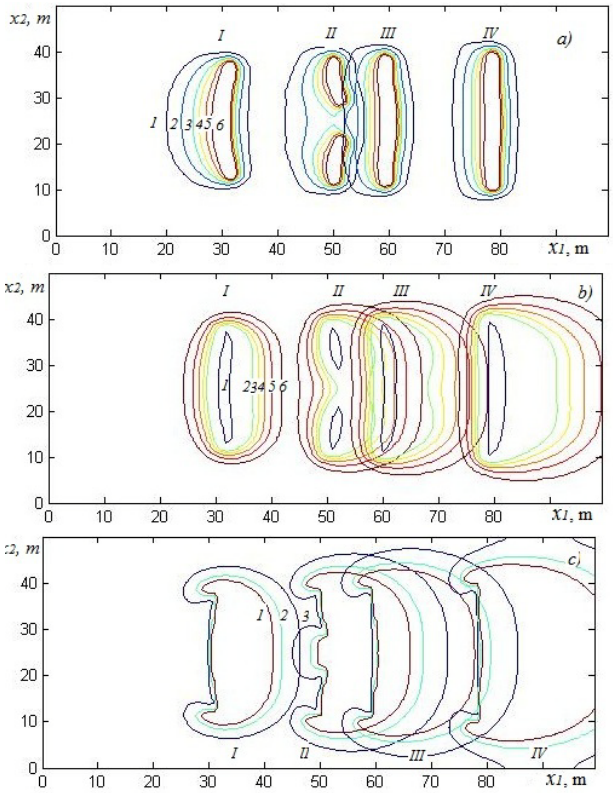


Fig.10 Fields of isotherms of gas phase (a), isolines of oxygen(b) and gas products of pyrolysis(c).

Using the model proposed in this paper it is possible to estimate the sizes of firebreaks which look likes as glades.

4. CONCLUSION

The results of calculation give an opportunity to evaluate critical condition of the forest fire spread, which allows applying the given model for preventing fires. The model proposed there gives a detailed picture of the change in the temperature and component concentration fields with time, and determine as well as the influence of different conditions on the crown forest fire initiation. It allows to investigate dynamics of forest fire initiation and spread under influence of various external conditions: a) meteorology conditions (air temperature, wind velocity etc.), b) type (various kinds of forest combustible materials) and their state(load, moisture etc.). The results obtained agree with the laws of physics and experimental data [1,11].

REFERENCES

[1] Grishin A.M., *Mathematical Modeling Forest Fire and New Methods Fighting Them*, Tomsk: Publishing House of Tomsk University (Russia), 1997.

[2] Van Wagner C.E., Conditions for the start and spread of crown fire, *Canadian Journal of Forest Research* 7, 1977, pp. 23-34.

[3] Alexander V.E., Crown fire thresholds in exotic pine plantations of Australasia, PhD diss., Australian National University, 1998.

[4] Van Wagner C.E., Prediction of crown fire behavior in conifer stands, *In Proceedings of 10th conference on fire and forest meteorology. Ottawa, Ontario. (Eds D. C. MacIver, H. Auld and R. Whitewood)*, 1989.

[5] Xanthopoulos, G., *Development of a wildland crown fire initiation model*,. Ph.D diss., University of Montana, 1990.

[6] Rothermel R.C., Crown fire analysis and interpretation, *In Proceedings of 11th International conference on fire and forest meteorology, Missoula, Montana (USA)*, 1991.

[7] Cruz M.G. et al., Predicting crown fire behavior to support forest fire management decision-making, *In Proceedings of IV International conference on forest fire research. Luso-Coimbra, Portugal. (Ed. D. X. Viegas)*, 11 [CD-ROM]. (Millpress), 2002.

[8] Albini F.A. et al,. Modeling ignition and burning rate of large woody natural fuels, *Int. Journal of Wildland fire*, 5, 1995, pp 81-91.

[9] Scott J.H. et al, Assessing crown fire potential by linking models of surface and crown fire behavior, *USDA Forest Service, Rocky Mountain Forest and Range Experiment Station. Fort Collins: RMRS-RP-29, (Colorado, USA)*, 2001.

[10] Grishin A.M., Perminov V.A., Mathematical modeling of the ignition of tree crowns, *Combustion, Explosion, and Shock Waves*, 34, 1998, pp. 378-386.

[11] Konev E.V., *The physical foundation of vegetative materials combustion*, Novosibirsk: Nauka, 1977.

[12] Morvan D., Dupuy J.L., Modeling of fire spread through a forest fuel bed using a multiphase formulation, *Combustion and Flame*. 127, 2001, pp. 1981-1994.

[13] Morvan D., Dupuy J.L., Modeling the propagation of wildfire through a Mediterranean shrub using a multiphase formulation. *Combustion and Flame*, 138, 2004, pp. 199-210.

[14] Perminov V.A., Mathematical modeling of crown forest fire initiation, *In Proceedings of III International conference on forest fire research and 14th conference on fire and forest meteorology. Luso, Portugal. (Ed. D.X.Viegas)*, 1998, pp. 419-431.

[15] Patankar S.V., *Numerical Heat Transfer and Fluid Flow. New York*, Hemisphere Publishing Corporation, 1981.

[16] Zverev V.G., *Mathematical modeling of aerodynamics and heat and mass transfer at crown forest fire spread*, Ph.D. diss, Tomsk State University, 1985.

Valeriy A. Perminov obtained his PhD at Tomsk University for a thesis on the mathematical modeling of forest fires. In 2011 he became a Doctor of Physical and Mathematical Sciences. Since 2011 he is an Associate Professor of Tomsk Polytechnic University, Department of Ecology and Basic Safety. His scientific interest is connected with mathematical and computational models for physical processes and their applications to ecology and forest fires initiation and development and environmental pollution. Dr. Perminov is a member in Council on combustion and explosion of Siberian Department of Russian Academy of Science and a member of American Chemical Society. He is an Editor-in-Chief of “Open Journal of Forestry”.

Analytical solution for some MHD problems on a flow of conducting liquid in the initial part of a channel in the case of rotational symmetry

Elena Ligere, Ilona Dzenite

***Abstract*—**This paper presents the analytical solution of magnetohydrodynamical (MHD) problems on a developing flow of conducting liquid in the initial part of a channel for the case, when conducting fluid flows into the channel through its wall in the presence of the rotational symmetry in the geometry of the flow. The problems are solved in Stokes and inductionless approximation, and on using integral transforms. The velocity field of the flow is analyzed numerically by means of the obtained solutions.

***Keywords*—** Magnetohydrodynamics, Navier-Stokes equations, analytical solution, channel flow.

I. Introduction

MAGNETOHYDRODYNAMICS (MHD) is a separate discipline combining the classical fluid mechanics and electrodynamics. The flow of a conducting fluid in the external magnetic field produces new effects, absent in the ordinary hydrodynamics, and which arise due to the electromagnetic Lorenz force generated by the interaction of the moving fluid with electromagnetic field. MHD analyzes these phenomena and it also studies a flow of a conducting fluid caused by the current passing through the fluid.

Nowadays MHD effects are widely exploited both in technical devices (e.g., in pumps, flow meters, generators) and industrial processes in metallurgy, material processing, chemical industry, industrial power engineering and nuclear engineering. Channels, in particular narrow and circular channels, are common parts of many MHD devices. Therefore, investigation of MHD phenomena in channels with conducting fluids is quite important.

The motion of conducting fluid in external magnetic field is described by the system of MHD equations, containing Navier-Stokes equation for the motion of incompressible viscous fluid with the additional term corresponding to the Lorentz force (see [3], [4]).

In magnetohydrodinamics the number of exact solutions, obtained analytically, is limited due to the nonlinearity of the Navier-Stokes equation. The exact solutions have been obtained only for very specific problems. Therefore, numerical methods are widely used for solving these problems.

Analytical solutions are mostly obtained for the simplified flow models and on using some approximations. In the present paper the following two approximations are used. These are the Stokes approximation, when the nonlinear term is neglected in the Navier-Stokes equations, and the inductionless approximation, for which the induced currents are taken into account, but the magnetic field created by these currents is neglected.

In this paper two problems on a flow of conducting liquid in the initial part of a channel are considered for the case, when conducting fluid flows into the channel through its wall in the presence of rotational symmetry in the geometry of the flow. These problems are solved analytically by using integral transforms.

The first problem is the problem on an inflow of a conducting fluid in the plane channel through a round hole of finite radius in its lateral side. In the authors' work [5], this problem was considered for the longitudinal magnetic field, but the case of the strong transverse magnetic field was just briefly mentioned. In the present paper, the case of the strong transverse magnetic field is considered in detail and some new numerical results are also presented.

Additionally, the problem on an inflow of a conducting liquid into a circular channel through a split of finite length in its lateral side is briefly considered. This problem was considered in the author's work [2] and its solution was obtained in the form of convergent improper integrals on using Stokes and inductionless approximation. In [2] on obtaining the solution, the Fourier transform was used together with the assumption that the velocity and pressure gradient are equal to zero in channel at the sufficient distance from the entrance region. But this assumption is not correct, since in longitudinal magnetic field in a round channel the Poiseuille flow appears far away from the entrance region. In the present paper the correct way of obtaining the analytical solution of the problem is considered, although it is shown that the final results obtained in [2] are the same and correct.

II. Problem Formulation

A plane channel with conducting fluid is located in region $D=\{0\le\tilde{r}\le+\infty,\ 0\le\tilde{\varphi}\le 2\pi,\ -h\le\tilde{z}\le h\}$, where $\tilde{r}$, $\tilde{\varphi}$, $\tilde{z}$ are cylindrical coordinates. There is a round hole of finite radius

Elena Ligere is with the Department of Engineering Mathematics, Riga Technical University, Riga, Latvia (e-mail: jelena.ligere@rtu.lv).

Ilona Dzenite with the Department of Engineering Mathematics, Riga Technical University, Riga, Latvia (e-mail: ilona.dzenite@rtu.lv).

$\tilde{R}$ in the channel later side, through which a conducting fluid flows into the channel with the constant velocity $V_0\vec{e}_z$ (see Fig.1).

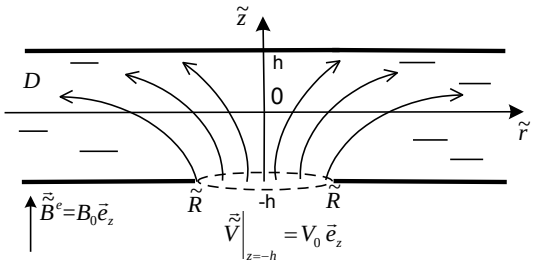


Fig. 1. The geometry of the flow.

The case of transverse magnetic field is considered, i.e., when the external magnetic field $\vec{B}^e = \mathrm{B}_0\vec{e}_z$ is parallel to the $\tilde{z}$ axis. It is also assumed that the channel walls $\tilde{z} = \pm h$ are non-conducting and induced streams do not flow through the hole $\{\ \tilde{z} = -h\ ,\ 0 < \tilde{r} < \tilde{R}\ \}$.

On introducing the dimensionless variables, when the half-width of channel h is used as a length scale, the magnitude of the velocity of fluid in the entrance region V_0 - as a velocity scale, and B_0, V_0B_0 , $\rho\nu V_0/h$ - as scales of magnetic field, electrical field and pressure, respectively, where σ is the conductivity, ρ is the density and ν is the viscosity of the fluid, and on using Stokes and inductionless approximations, the dimensionless MHD equations in cylindrical coordinates take the form

$$\Delta\vec{V} + Ha^2(\vec{E} + \vec{V}\times\vec{e}_B)\times\vec{e}_B = \nabla P\,, \tag{1}$$

$$div\vec{V} = 0\,, \tag{2}$$

where $\vec{e}_B$ is the unit vector of external magnetic field,

$\vec{V} = V_r(r,z)\vec{e}_r + V_z(r,z)\vec{e}_z$ is the velocity of the fluid,

$$\Delta\vec{V} = \vec{e}_r(L_0V_r - \frac{V_r}{r^2}) + \vec{e}_z(L_0V_z)\,,\quad L_0 = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{\partial^2}{\partial z^2}\,.$$

P is the pressure, $Ha = B_0h\sqrt{\sigma/\rho\gamma}$ is the Hartmann number, characterizing the ratio of electromagnetic force to viscous one.

Projecting (1) and (2) onto the r and z axes, and taking into account that $\vec{e}_B = \vec{e}_z$ and the intensity of electrical field $\vec{E} = 0$ for this problem (see [3],[4]), the problem takes the form

$$-\frac{\partial P}{\partial r} + L_1V_r - Ha^2V_r = 0\,, \tag{3}$$

$$-\frac{\partial P}{\partial z} + L_0V_z = 0\,, \tag{4}$$

$$\frac{\partial V_z}{\partial z} + \frac{1}{r}\frac{\partial}{\partial r}(r\cdot V_r) = 0\,. \tag{5}$$

with the following boundary conditions

$$z=-1:\qquad V_r = 0\,,\ V_z = \begin{cases}1, & 0\le r\le R\\ 0, & r>R\end{cases} \tag{6}$$

$$z=1:\qquad V_r = 0\,,\quad V_z = 0\,, \tag{7}$$

$$r\to\pm\infty:\ V_r\to 0\,,\ \partial P/\partial r\to 0\,, \tag{8}$$

where $L_1f = L_0f - f/r^2\,,\quad R = \tilde{R}/h\,.$

III. Problem Solving

Due to the rotational symmetry of the problem with respect to r, the Hankel transform (see [1]) is used for the problem solving. The Hankel transform of order 1 with respect to r is applied to the functions V_r and $\partial P/\partial r$, but the Hankel transform of order 0 is applied to the functions V_z and $\partial P/\partial z$:

$$\hat{V}_r(\lambda,z) = \int_0^\infty V_rJ_1(\lambda r)rdr\,,\qquad \hat{V}_z(\lambda,z) = \int_0^\infty V_zJ_0(\lambda r)rdr\,,$$

$$\hat{P}(\lambda,z) = \int_0^\infty PJ_0(\lambda r)rdr\,, \tag{9}$$

where $J_\nu(\lambda r)$ is the Bessel functions of order ν.

On applying the Hankel transform to the system (3)-(5), one gets the system of ordinary differential equations for the Hankel transforms $\hat{V}_r(\lambda,z)$, $\hat{V}_z(\lambda,z)$, $\hat{P}(\lambda,z)$:

$$\lambda\hat{P} - \lambda^2\hat{V}_r + \frac{d^2\hat{V}_r}{dz^2} - Ha^2\hat{V}_r = 0\,, \tag{10}$$

$$\frac{d\hat{P}}{dz} + \lambda^2\hat{V}_z - \frac{d^2\hat{V}_z}{dz^2} = 0\,, \tag{11}$$

$$\frac{d\hat{V}_z}{dz} + \lambda\hat{V}_r = 0 \tag{12}$$

with the boundary conditions:

$$z=-1:\quad \hat{V}_r = 0\,,\quad \hat{V}_z = RJ_1(\lambda R)/\lambda \tag{13}$$

$$z=1:\quad \hat{V}_r = 0\,,\ \hat{V}_z = 0\,. \tag{14}$$

On eliminating $\hat{V}_r$ and $\hat{P}$ from (10)-(12), the following differential equation is obtained for $\hat{V}_z$.

$$\hat{V}_z^{(4)} - \left(2\lambda^2 + Ha^2\right)\cdot\hat{V}_z'' + \lambda^4\cdot\hat{V}_z = 0\,. \tag{15}$$

The general solution of (15) has the form

$$\hat{V}_z = C_1\sinh k_1z + C_2\sinh k_2z + C_3\cosh k_1z + C_4\cosh k_2z\,, \tag{16}$$

where $k_1 = \mu + \sqrt{\mu^2+\lambda^2}$, $k_2 = \mu - \sqrt{\mu^2+\lambda^2}$, (17)

$\mu = Ha/2$, C_1 - C_4 are arbitrary constants.

In order to reduce the number of the constants C_1-C_4 in (16) and simplify the problem solving, the problem is divided into two sub-problems: an odd and even problem with respect to z, on considering a plane channel with two holes in its lateral sides $\tilde{z} = \pm h$ in the region $0 < \tilde{r} < \tilde{R}$.

In the **odd problem** with respect to z the fluid with velocities $\mp(V_0\vec{e}_z)/2$ flows into the channel through the both holes at $\tilde{z} = \pm h$. The geometry of the flow for this problem is presented in Fig. 2.

In the **even problem** with respect to z the fluid with velocity $(V_0\vec{e}_z)/2$ flows into the channel through the hole at $\tilde{z} = -h$ and flows out with the same velocity through the hole at $\tilde{z} = h$. The geometry of the flow is presented in Fig. 3.

Then the solution of the **general problem** is equal to the sum of solutions of the odd end even problems.

A. Solution of the Odd Problem

The odd problem with respect to z is the problem on an inflow of fluid into the channel through both holes at $\tilde{z} = \pm h$.

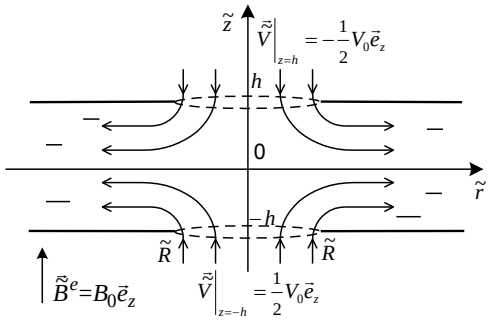


Fig. 2. The odd problem with respect to z.

The dimensionless boundary conditions for the problem are

$$z = \pm 1: \quad V_r = 0, \quad V_z = \begin{cases} \mp 1/2, & 0 \le r \le R \\ 0, & r > R \end{cases} \tag{18}$$

$$r \to \pm\infty: \quad V_x \to 0, \quad \partial P/\partial x \to 0. \tag{19}$$

On applying the Hankel transforms (9) to the boundary conditions (18)-(19), one gets

$$z = \pm 1: \quad \hat{V}_r = 0, \quad \hat{V}_z = \mp R \cdot J_1(\lambda R)/(2\lambda) \tag{20}$$

For the odd problem, $\hat{V}_z$ is the odd function with respect to z and, therefore, $C_3 = C_4 = 0$ in (16), i.e.

$$\hat{V}_z(\lambda, z) = C_1 \sinh k_1 z + C_2 \sinh k_2 z \tag{21}$$

In order to determine C_1 and C_2, the boundary condition (20) are used together with the additional boundary condition obtained from (12), i.e.,

$$z = 1: \quad \hat{V}_z = -R \cdot J_1(\lambda R)/(2\lambda) \quad \text{and} \quad d\hat{V}_z/dz = 0. \tag{22}$$

As a result, one obtains

$$\hat{V}_z = \frac{k_1 \cosh k_1 \sinh k_2 z - k_2 \cosh k_2 \sinh k_1 z}{\Delta_1} \cdot \frac{\psi}{\lambda}, \tag{23}$$

where k_1, k_2 are given by (17),

$$\Delta_1 = k_2 \cosh k_2 \cdot \sinh k_1 - k_1 \cosh k_1 \cdot \sinh k_2. \tag{24}$$

$$\psi = \psi(\lambda) = RJ_1(\lambda R)/2 \tag{25}$$

The function $\hat{V}_r$ is determined from (12):

$$\hat{V}_r = \frac{\cosh k_1 \cosh k_2 z - \cosh k_2 \cosh k_1 z}{\Delta_1} \cdot \psi \tag{26}$$

Functions $\lambda\hat{P}$ and $d\hat{P}/dz$ are determined from (10)-(11):

$$\lambda\hat{P} = Ha^2 \cdot \frac{k_2 \cosh k_2 \cosh k_1 z - k_1 \cosh k_1 \cosh k_2 z}{\Delta_1} \cdot \psi \tag{27}$$

$$\frac{d\hat{P}}{dz} = Ha \cdot \frac{\cosh k_2 \sinh k_1 z - \cosh k_1 \sinh k_2 z}{\Delta_1} \cdot \lambda\psi \tag{28}$$

Then on using the inverse complex Hankel transform, the solution of the problem (3)–(5) with boundary conditions (18)-(19) is obtained in the form of the convergent improper integrals:

$$V_r = \int_0^\infty \frac{\cosh k_1 \cosh k_2 z - \cosh k_2 \cosh k_1 z}{\Delta_1} \psi \lambda J_1(\lambda r) d\lambda, \tag{29}$$

$$V_z = \int_0^\infty \frac{k_1 \cosh k_1 \sinh k_2 z - k_2 \cosh k_2 \sinh k_1 z}{\Delta_1} \psi J_0(\lambda r) d\lambda \tag{30}$$

$$\frac{\partial P}{\partial r} = Ha^2 \int_0^\infty \frac{k_2 \cosh k_2 \cosh k_1 z - k_1 \cosh k_1 \cosh k_2 z}{\Delta_1} \psi \lambda J_1(\lambda r) d\lambda \tag{31}$$

$$\frac{\partial P}{\partial z} = Ha \int_0^\infty \frac{\cosh k_2 \sinh k_1 z - \cosh k_1 \sinh k_2 z}{\Delta_1} \psi \lambda^2 J_0(\lambda r) d\lambda \tag{32}$$

B. Solution of the Even Problem

The geometry of the flow for the even problem with respect to z is shown in Fig. 3.

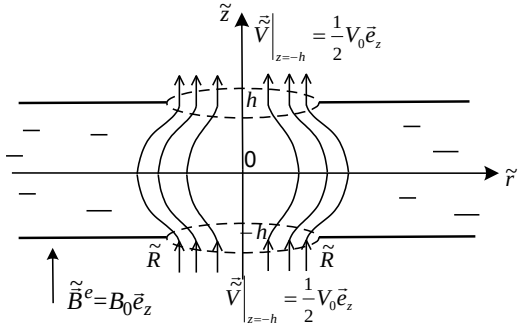


Fig. 3. The even problem with respect to z.

The dimensionless boundary conditions for this problem are

$$z=\pm 1:\ V_r=0,\quad V_z=\begin{cases}1/2, & 0\le r\le R\\ 0, & r>R\end{cases} \tag{33}$$

$$r\to\pm\infty:\quad V_x\to 0,\quad \partial P/\partial x\to 0. \tag{34}$$

On applying the Hankel transform (9) to boundary conditions (32)-(33), one gets

$$z=\pm 1:\quad \hat{V}_r=0,\quad \hat{V}_z=RJ_1(\lambda R)/(2\lambda) \tag{35}$$

For the even problem, the function $\hat{V}_z$ is even with respect to z, therefore, $C_1=C_2=0$ in (16), i.e.,

$$\hat{V}_z(\lambda,z)=C_3\cosh k_1 z+C_4\cosh k_2 z. \tag{36}$$

In order to determine C_3 and C_4, the boundary condition (35) and (12) are used, as a result, one gets

$$\hat{V}_z=\frac{k_2\sinh k_2\cosh k_1 z-k_1\sinh k_1\cosh k_2 z}{\Delta_2}\cdot\psi \tag{37}$$

where k_1, k_2 are given by (17), ψ is given by (25) and $\Delta_2=k_2\cosh k_1\cdot\sinh k_2-k_1\cosh k_2\cdot\sinh k_1$.

The function $\hat{V}_r$ is determined from (12):

$$\hat{V}_r=\frac{\sinh k_1\sinh k_2 z-\sinh k_2\sinh k_1 z}{\Delta_2}\cdot\psi \tag{38}$$

The functions $\lambda\hat{P}$ and $d\hat{P}/dz$ are determined from (10)-(11):

$$\lambda\hat{P}=Ha\cdot\frac{k_2\sinh k_2\sinh k_1 z-k_1\sinh k_1\sinh k_2 z}{\Delta_2}\cdot\psi \tag{39}$$

$$\frac{d\hat{P}}{dz}=Ha\cdot\frac{\sinh k_1\cosh k_2 z-\sinh k_2\cosh k_1 z}{\Delta_2}\cdot\lambda\psi \tag{40}$$

On applying the inverse complex Hankel transform to (37)-(40), the solution to the problem (3)–(5) with boundary conditions (33), (34) is obtained in the form of convergent improper integrals:

$$V_r=\int_0^\infty\frac{\sinh k_2\sinh k_1 z-\sinh k_1\sinh k_2 z}{\Delta_2}\psi\lambda_1(\lambda r)d\lambda, \tag{41}$$

$$V_z=\int_0^\infty\frac{k_2\sinh k_2\cosh k_1 z-k_1\sinh k_1\cosh k_2 z}{\Delta_2}\psi J_0(\lambda r)d\lambda \tag{42}$$

$$\frac{\partial P}{\partial r}=Ha\int_0^\infty\frac{k_1\sinh k_1\cdot\sinh k_2 z-k_2\sinh k_2\cdot\sinh k_1 z}{\Delta_2}\psi\lambda J_1(\lambda r)d\lambda \tag{43}$$

$$\frac{\partial P}{\partial z}=Ha\int_0^\infty\frac{\sinh k_1\cdot\cosh k_2 z-\sinh k_2\cdot\cosh k_1 z}{\Delta_2}\psi\lambda^2 J_0(\lambda r)d\lambda \tag{44}$$

where k_1, k_2 are given by (17) and ψ by (25).

IV. Numerical Results

A. Numerical Results for the Odd Problem

On the base of obtained solution, the velocity field is studied numerically. Results of calculations of the velocity radial component V_r for the odd problem at the Hartmann numbers Ha=10 and Ha=50 are presented graphically in Fig. 4. The component V_r is an odd function with respect to z. It can be seen from Fig.4 that V_r has the M-shaped profiles only near the entrance hole ($1\le r\le 1.1$ at Ha=10 and $1\le r<1.1$ Ha=50). Even at a small distance from the entrance, the profiles of V_r take the shape peculiar to the Hartmann flow in a plane channel in transverse magnetic field. The magnitude of the velocity is inversely proportional to the distance from the hole.

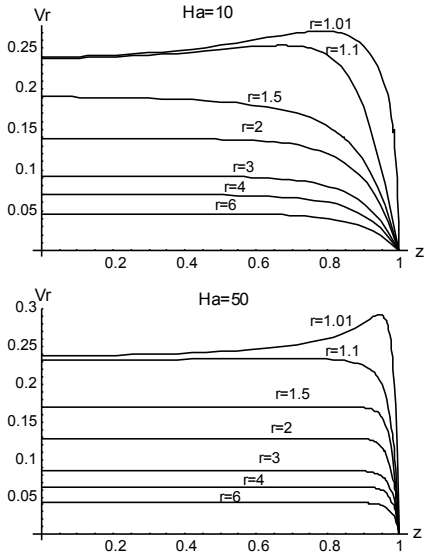


Fig. 4. Profiles of the velocity radial component V_r for the odd problem at R=1.

B. Numerical Results for the Even Problem

Fig. 5 presents the results of calculation of V_r by means of formula (41). One can see that V_r differs from zero only near the entrance region. Additionally, in Fig. 5 V_r is positive for some values of r, e.g., for $0<r\le 1$ and Ha=10. However, since the fluid flows out through the hole at $z=1$, the r-component of the velocity must be negative for $0<r<1$ at *Ha=0*. It means that in the transverse magnetic field in the region $0<z<1$ there exists an opposite flow, which occurs due to vortices generated in the channel (see Fig. 6). The vector

field of the velocity for Ha=10 is shown in Fig. 6.

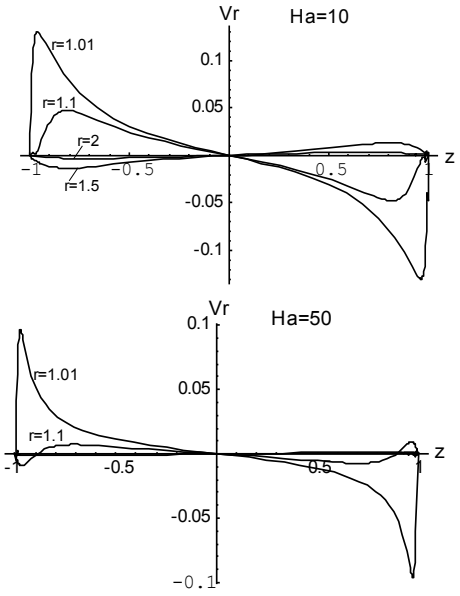


Fig. 5. Profiles of V_r for the even problem at R=1.

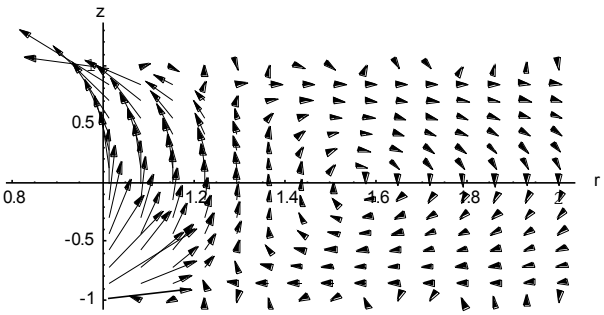


Fig. 6. Velocity field for the even problem at R=1 and Ha=10.

C. Numerical Results for the General Problem

The solution of the general problem is equal to the sum of solutions to the odd and even problems with respect to z. Results of calculation of V_r for the general problem at the Hartmann numbers Ha=10 and Ha=50 are presented in Fig. 7 and Fig. 8.

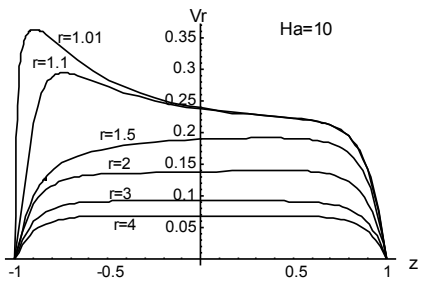


Fig. 7. Profiles of V_r for the general problem at R=1 and Ha=10.

One can see that as in the previous case, the profiles of the velocity component V_r differ from the Hartmann flow profiles only near the entrance region. The magnitude of the velocity is inversely proportional to the distance from the hole.

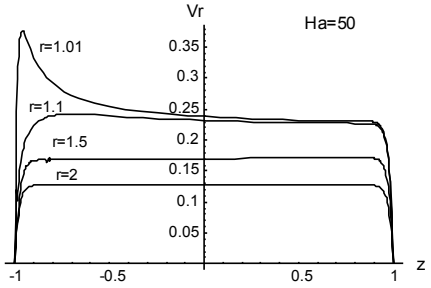


Fig. 8. Profiles of V_r for the general problem at R=1 and Ha=50.

V. Remarks to the Solution of MHD Problem on an Inflow of Conducting Fluid into a Circular Channel through the Channel's Lateral Side

A. Formulation of the Problem

A circular channel is located in the region $D=\{\,0\le\tilde r<R,\ 0\le\tilde\varphi<2\pi,\ -\infty<\tilde z<+\infty\,\}$. There is a split in the channel lateral surface in the region $\{\,\tilde r=R,\ -\tilde d\le\tilde z\le\tilde d\,\}$, through which a conducting fluid flows into the channel with the constant velocity $\vec{\tilde V}=-V_0\vec e_r$ (see Fig.9). The case of longitudinal magnetic field is considered, i.e., when the external magnetic field $\vec B^e=B_0\vec e_z$ is parallel to the $\tilde z$ axis.

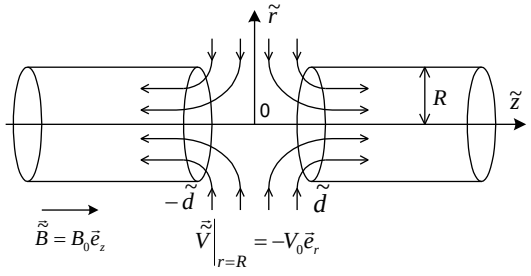


Fig. 9. The geometry of the flow in the circular channel.

It is supposed that walls $\tilde r=R$ are nonconducting and induced streams do not flow through the split $\tilde r=R$, $-\tilde d<\tilde z<\tilde d$ in the region $R<\tilde r<+\infty$. In this problem $\vec E=0$ (see [3]). The dimensionless variables are introduced similarly as it was done for the first problem, but the chanel radius R is used as the length scale.

The problem is described by the system of equations (3)-(5) with the boundary conditions:

$$r=1:\ V_z=0,\quad V_r=\begin{cases}-1, & z\in(-d,d)\\ 0, & z\notin(-d,d)\end{cases} \tag{45}$$

$z \to \pm\infty: V_z \to V_\infty(r)\cdot \mathrm{sign}(z)$,

$$\frac{\partial P}{\partial z} \to \frac{\partial P_\infty}{\partial z}\cdot sign(z) \equiv A\cdot sign(z)\ , \tag{46}$$

where $d = \tilde{d}/R$, $A = const$. The functions $V_\infty(r)$ and dP_∞/dz are the velocity of the flow and the pressure gradient in the channel at the sufficient distance from the entrance region, and which satisfy the following equation (see [3], [4]):

$$-\frac{dP_\infty}{dz} + \frac{1}{r}\frac{d}{dr}\left(r\cdot\frac{dV_\infty}{dr}\right) = 0 \tag{47}$$

with the boundary condition: $r=1$: $V_\infty(r) = 0$.

B. Solution of the Problem.

On solving the problem, the symmetry of the problem with respect to z is used, i.e., the velocity component $V_r(r,z)$ and pressure $P(r,z)$ are even functions with respect to z, but the component $V_z(r,z)$ is the odd function with respect to z. It means that the functions $V_r(r,z)$ and $V_z(r,z)$ satisfy additional boundary conditions:

$$z = 0,\ 0 < r < 1: \quad V_z = 0,\ \partial V_r/\partial z = 0. \tag{48}$$

The problem can be solved by the Fourier cosine and Fourier sine transforms, but since V_z and $\partial P/\partial z$ do not tends to zero at $z \to \pm\infty$, the new functions for the velocity and pressure gradient are to be introduced before using these transforms:

$$\vec{V}^{new} = \vec{V} - \frac{2}{\pi}\arctan(z)\cdot V_\infty(r)\cdot\vec{e}_z \tag{49}$$

$$\frac{\partial P^{new}}{\partial z} = \frac{\partial P}{\partial z} - \frac{2}{\pi}\arctan(z)\cdot A \tag{50}$$

As a result, the problem has the form:

$$-\frac{\partial P^{new}}{\partial r} + (L_1 - Ha^2)V_r = 0\ , \tag{51}$$

$$-\frac{\partial P^{new}}{\partial z} + L_0 V_z^{new} - \frac{2}{\pi}V_\infty(r)\cdot\frac{2z}{\left(1+z^2\right)^2} = 0\ , \tag{52}$$

$$\frac{\partial V_z^{new}}{\partial z} + \frac{1}{r}\frac{\partial}{\partial r}(rV_r) + \frac{2}{\pi}V_\infty(r)\cdot\frac{1}{1+z^2} = 0\ . \tag{53}$$

Boundary conditions are

$$r=1:\ V_z^{new} = 0\ , \quad V_r = \begin{cases} -1, & z\in(-d,d) \\ 0, & z\notin(-d,d) \end{cases} \tag{54}$$

$$z \to \pm\infty: \quad V_z^{new} \to 0\ , \quad \partial P^{new}/\partial z \to 0\ . \tag{55}$$

The Fourier cosine transform with respect to z is applied to (51), (53) and to V_r in boundary conditions (54), but the Fourier sine transform is applied to (52) and to V_z in boundary conditions (54):

$$V_r^c(r,\lambda) = F^c[V_r(r,z)] = \sqrt{\frac{2}{\pi}}\int_0^\infty V_r(r,z)\cos\lambda z\,dz\ ,$$

$$V_z^s(r,\lambda) = F^s\left[V_z^{new}(r,z)\right] = \sqrt{\frac{2}{\pi}}\int_0^\infty V_z^{new}(r,z)\sin\lambda z\,dz\ ,$$

$$P^c(r,\lambda) = F^c\left[P^{new}(r,z)\right] = \sqrt{\frac{2}{\pi}}\int_0^\infty P^{new}(r,z)\cos\lambda z\,dz\ .$$

It is also used that

$$F^s\left[\frac{2z}{\left(1+z^2\right)^2}\right] = \lambda F^c\left[\frac{1}{1+z^2}\right]. \tag{56}$$

As a result, the following system of ordinary differential equations for the unknown functions V_r^c, V_z^s, P^c is obtained:

$$-\frac{dP^c}{dr} + (L_{1r} - Ha^2)V_r^c = 0\ , \tag{57}$$

$$\lambda P^c + L_{0r}V_z^s - \frac{2}{\pi}V_\infty(r)\cdot\lambda\cdot F^c(\lambda) = 0\ , \tag{58}$$

$$\lambda V_z^s + \frac{1}{r}\frac{d}{dr}\left(rV_r^c\right) + \frac{2}{\pi}V_\infty(r)\cdot F^c(\lambda) = 0\ , \tag{59}$$

were $L_{0r} = \dfrac{d^2}{dr^2} + \dfrac{1}{r}\dfrac{d}{dr} - \lambda^2$, $L_{1r} = L_{0r} - \dfrac{1}{r^2}$,

un $F^c(\lambda) = F^c\left[\dfrac{1}{1+z^2}\right]$.

The boundary conditions are

$$\text{r=1:}\quad V_r^c(r,\lambda) = -\sqrt{\frac{2}{\pi}}\frac{\sin\lambda L}{\lambda}\ , \qquad V_z^s = 0 \tag{60}$$

On eliminating the functions V_z^s, P^c from system (57)-(59), the equation for V_r^c is obtained in the form

$$\frac{d}{dr}(L_r\tilde{L}_{0r})V_r^c - 2\lambda^2\tilde{L}_{1r}V_r^c + \lambda^2(\lambda^2 + Ha^2)V_r^c = 0. \tag{61}$$

where $\tilde{L}_{0r} = \dfrac{1}{r} + \dfrac{d}{dr}$, $L_r = \dfrac{d^2}{dr^2} + \dfrac{1}{r}\dfrac{d}{dr}$, $\tilde{L}_{1r} = L_r - \dfrac{1}{r^2}$.

Differential equation (61) completely coincides with differential equation for V_r^c obtained in [2], therefore, the solution of this equation is the same as in [2], i.e.,

$$V_r^c(r\lambda) = c_1(\lambda)I_1(k_1 r) + c_2(\lambda)I_1(k_2 r) \tag{62}$$

where

$$c_1(\lambda) = A(\lambda)k_2 I_0(k_2)/\Delta, \quad c_2(\lambda) = -A(\lambda)k_1 I_0(k_1)/\Delta\ , \tag{63}$$

$$k_1 = \sqrt{\lambda^2 + iHa\lambda}, \quad k_2 = \sqrt{\lambda^2 - iHa\lambda}, \tag{64}$$

$$A(\lambda) = \sqrt{\frac{2}{\pi}}\frac{\sin\lambda L}{\lambda}, \qquad \Delta = k_1 I_0(k_1) I_1(k_2) - k_2 I_0(k_2) I_1(k_1).$$

The function $\hat{V}_z^{\,s}$ is determined from (59):

$$V_z^s(\mathrm{r},\lambda) = -\frac{1}{\lambda}\left(C_1 k_1 I_0(k_1 r) + C_2 k_2 I_0(k_2 r)\right) - \frac{2}{\pi}\cdot\frac{F^c(\lambda)V_\infty(r)}{\lambda} \tag{65}$$

The functions λP^c and dP^c/dr are determined from (57) and (58) on taking into account the following formulas

$$\tilde{L}_{1_r} I_1(k\cdot r) = k^2 I_1(k\cdot r) \quad \text{un} \quad L_r I_0(k\cdot r) = k^2 I_0(k\cdot r).$$

Then

$$\frac{dP^c}{dr} = \tilde{c}_1(\lambda) I_1(k_1 r) - \tilde{c}_2(\lambda) I_1(k_2 r), \tag{66}$$

$$-\lambda P^c = iHa\cdot\left(c_2 k_2 I_0(k_2 r) - c_1 k_1 I_0(k_1 r)\right) - \frac{2}{\pi}\cdot\frac{F^c(\lambda)A}{\lambda}. \tag{67}$$

Note, that $V_z^{\,s}$ and $-\lambda P^c$ differ from result obtained in [2] by only last terms. On applying the inverse cosine and sine Fourier transforms to the functions $V_r^{\,c}$, $V_z^{\,s}$, dP^c/dr and $-\lambda P^c$, the solution of the problem is obtained and it has the form of convergent improper integrals, which coincide with the solution of the problem obtained in [2]:

$$V_r = \sqrt{\frac{2}{\pi}}\int_0^\infty \left[c_1(\lambda) I_1(k_1 r) + c_2(\lambda) I_1(k_2 r)\right]\cos\lambda z\, d\lambda, \tag{68}$$

$$V_z = -\sqrt{\frac{2}{\pi}}\int_0^\infty \left[c_1(\lambda) k_1 I_0(k_1 r) + c_2(\lambda) k_2 I_0(k_2 r)\right]\frac{\sin\lambda z}{\lambda}\, d\lambda, \tag{69}$$

$$\frac{\partial P}{\partial r} = \sqrt{\frac{2}{\pi}}\int_0^\infty \left[\tilde{c}_1(\lambda) I_1(k_1 r) - \tilde{c}_2(\lambda) I_1(k_2 r)\right]\cos\lambda z\, d\lambda, \tag{70}$$

$$\frac{\partial P}{\partial z} = iHa\sqrt{\frac{2}{\pi}}\int_0^\infty \left[c_2(\lambda) k_2 I_0(k_2 r) - c_1(\lambda) k_1 I_0(k_1 r)\right]\sin\lambda z\, d\lambda, \tag{71}$$

where $\tilde{c}_1(\lambda) = Ha\cdot(i\lambda - Ha)\cdot c_1(\lambda)$,

$\tilde{c}_2(\lambda) = Ha\cdot(i\lambda + Ha)\cdot c_2(\lambda)$

and $c_1(\lambda)$, $c_2(\lambda)$ are given by (63).

REFERENCES

[1] Antimirov M.Ya., Kolyshkin A.A., Vaillancourt R. *Applied Integral Transforms*.- Rhole Island USA: American Mathematical Society, 1993.

[2] Antimirov M., Ligere E. "Analytical solution for magnetohydrodynamical problems at flow of conducting fluid in the initial part of round and plane channels", *Magnetohydrodynamics*. vol.36, no.3.,pp. 241-250, 2000.

[3] Bojarevich V.B, Freiberg Ja.G., Shilova E.I., Shcherbinin E.V. *Electrically indjused vortical flows*. Dordreht; Boston; London: KLUWER Acad. Publ. 1989

[4] Davidson P.A. *An Introduction to Magnetohydrodynamics*, New York: Cambridge university press, 2001.

[5] Ligere E., Dzenite I. "Application of Integral Transforms for Solving Some MHD Problems" in *Proc 14th WSEAS Int. Conf. on Mathematical and Computational Methods in Science and Engineering.- Advances in Mathematical and Computational Methods*, Sliema (Malta), September 7-9, 2012, pp. 286.-291.

Some Properties of Operators with Non-Analytic Functional Calculus

Cristina Şerbănescu and Ioan Bacalu
Faculty of Applied Sciences,
University Politehnica of Bucharest,
Romania

Abstract— This paper is dedicated to the study of some properties of the operators which admit residually non-analytic functional calculus initiated in [14]. The concepts of $\mathcal{A}_S$-spectral function, respectively $\mathcal{A}_S$-decomposable and $\mathcal{A}_S$-spectral operators are introduced and characterized here and several elementary properties concerning them are studied. These operators are natural generalizations of the notions of $\mathcal{A}$-scalar, $\mathcal{A}$-decomposable and $\mathcal{A}$-spectral operators studied in [8] and appear, in generally, as restrictions or quotients of the last one.

Keywords— $\mathcal{A}$-spectral ($\mathcal{A}_S$-spectral) function; $\mathcal{A}$-scalar ($\mathcal{A}_S$-scalar); $\mathcal{A}$-decomposable ($\mathcal{A}_S$-decomposable); $\mathcal{A}$-spectral ($\mathcal{A}_S$-spectral); restrictions and quotients of operators.

I. Introduction

Let X be a Banach space, let $\mathbf{B}(X)$ be the algebra of all linear bounded operators on X and let $\mathbb{C}$ be the complex plane. If $T\in\mathbf{B}(X)$ and $Y\subset X$ is a (closed) invariant subspace to T, let us denote by $T|Y$ the restriction of T to Y, respectively by $\dot{T}$ the operator induced by T in the quotient space $\dot{X}=X/Y$. In what follows, by subspace of X we understand a closed linear manifold of X. Recall that Y is a *spectral maximal space* of T if it is an invariant subspace such that for any other subspace $Z\subset X$ also invariant to T, the inclusion $\sigma(T|Z)\subset\sigma(T|Y)$ implies $Z\subset Y$ ([8]). A family of open sets $G_S\bigcup\{G_i\}_{i=1}^n$ is an *S-covering* of the closed set $\sigma\subset\mathbb{C}$ if $G_S\bigcup\left(\bigcup_{i=1}^n G_i\right)\supset\sigma\bigcup S$ and $\overline{G}_i\cap S=\varnothing\ \ (i=1,2,...,n)$ (where $S\subset\mathbb{C}$ is also closed) ([12]).

The operator $T\in\mathbf{B}(X)$ is *S-decomposable* (where $S\subset\sigma(T)$ is compact) if for any finite open S-covering $G_S\bigcup\{G_i\}_{i=1}^n$ of $\sigma(T)$, there is a system $Y_S\bigcup\{Y_i\}_{i=1}^n$ of spectral maximal spaces of T such that $\sigma(T|Y_S)\subset G_S$, $\sigma(T|Y_i)\subset G_i\ \ (i=1,2,...,n)$ and $X=Y_S+\sum_{i=1}^n Y_i$ ([4]). If $\dim S=0$, then $S=\varnothing$ and T is decomposable ([8]). An open set $\Omega\subset\mathbb{C}$ is said to be a *set of analytic uniqueness* for $T\in\mathbf{B}(X)$ if for any open set $\omega\subset\Omega$ and any analytic function $f_0:\omega\to X$ satisfying the equation $(\lambda I-T)f_0(\lambda)\equiv 0$ it follows that $f_0(\lambda)\equiv 0$ in ω ([12]). For $T\in\mathbf{B}(X)$ there is a unique maximal open set Ω_T of analytic uniqueness ([12]). We shall denote by $S_T=\complement\Omega_T=\mathbb{C}\setminus\Omega_T$ and call it *the analytic spectral residuum of* T. For $x\in X$, a point λ is in $\delta_T(x)$ if in a neighborhood V_λ of λ, there is at least an analytic X-valued function f_x (called *T-associated to x*) such that $(\mu I-T)f_x(\mu)\equiv x$, for $\mu\in V_\lambda$. We shall put $\gamma_T(x)=\complement\delta_T(x)=\mathbb{C}\setminus\delta_T(x)$, $\rho_T(x)=\delta_T(x)\cap\Omega_T$ $\sigma_T(x)=\complement\rho_T(x)=\mathbb{C}\setminus\rho_T(x)=\gamma_T(x)\bigcup S_T$ and $X_T(F)=\{x\in X;\ \sigma_T(x)\subset F\}$

where $S_T\subset F\subset\mathbb{C}$ ([12], [13]).

An operator $T\in\mathbf{B}(X)$ is said to have *the single-valued extension property* if for any analytic function $f:\omega\to X$ (where $\omega\subset\mathbb{C}$ is an open set), with $(\lambda I-T)f(\lambda)=0$, it follows that $f(\lambda)\equiv 0$ ([10]). T

has the single-valued extension property if and only if $S_T = \varnothing$; then we have $\sigma_T(x) = \gamma_T(x)$ and there is in $\rho_T(x) = \delta_T(x)$ a unique analytic function $x(\lambda)$, T-associated to x, for any $x \in X$. We shall recall that if $T \in \mathbf{B}(X)$, $S_T \neq \varnothing$, $S_T \subset F$ and $X_T(F)$ is closed, for $F \subset \mathbb{C}$ closed, then $X_T(F)$ is a spectral maximal space of T ([12]).

We say that two operators $T_1, T_2 \in \mathbf{B}(X)$ are *quasinilpotent equivalent* if

$$\lim_{n\to\infty} \left\| (T_1 - T_2)^{[n]} \right\|^{\frac{1}{n}} = \lim_{n\to\infty} \left\| (T_2 - T_1)^{[n]} \right\|^{\frac{1}{n}} = 0$$

where

$$(T_1 - T_2)^{[n]} = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} T_1^k \, T_2^{n-k}$$ ([8]).

Definition 1.1. ([14]) Let Ω be a set of the complex plane $\mathbb{C}$ and let $S \subset \overline{\Omega}$ be a compact subset. An algebra $\mathcal{A}_S$ of $\mathbb{C}$-valued functions defined on Ω is called S-*normal* if for any finite open S-covering $G_S \bigcup \{G_i\}_{i=1}^n$ of $\overline{\Omega}$, there are the functions, f_S, $f_i \in \mathcal{A}_S$ $(1 \le i \le n)$ such that:

1) $f_S(\Omega) \subset [0, 1]$, $f_i(\Omega) \subset [0, 1]$ $(1 \le i \le n)$;

2) $\operatorname{supp}(f_S) \subset G_S$, $\operatorname{supp}(f_i) \subset G_i$ $(1 \le i \le n)$;

3) $f_S + \sum_{i=1}^{n} f_i = 1$ on Ω

where *the support* of $f \in \mathcal{A}_S$ is defined as: $\operatorname{supp}(f) = \overline{\{\mu \in \Omega;\ f(\mu) \neq 0\}}$.

Definition 1.2. ([14]) An algebra $\mathcal{A}_S$ of $\mathbb{C}$-valued functions defined on Ω is called S-*admissible* if:

1) $\lambda \in \mathcal{A}_S$, $1 \in \mathcal{A}_S$ (where λ and 1 denote the functions $f(\lambda) = \lambda$ and $f(\lambda) = 1$);

2) $\mathcal{A}_S$ is S-normal;

3) for any $f \in \mathcal{A}_S$ and any $\xi \notin \operatorname{supp}(f)$, the function

$$f_\xi(\lambda) = \begin{cases} \dfrac{f(\lambda)}{\xi - \lambda}, & \text{for } \lambda \in \Omega \setminus \{\xi\} \\ 0, & \text{for } \lambda \in \Omega \bigcap \{\xi\} \end{cases}$$

belongs to $\mathcal{A}_S$.

Definition 1.3. ([14]) An operator $T \in \mathbf{B}(X)$ is said to be $\mathcal{A}_S$-*scalar* if there are an S-admissible algebra $\mathcal{A}_S$ and an algebraic homomorphism $U : \mathcal{A}_S \to \mathbf{B}(X)$ such that $U_1 = I$ and $U_\lambda = T$ (where 1 is the function $f(\lambda) = 1$ and λ is the function $f(\lambda) = \lambda$). The mapping U is called $\mathcal{A}_S$-*spectral homomorphism* ($\mathcal{A}_S$-*spectral function* or $\mathcal{A}_S$-*functional calculus*) for T.

If $S = \varnothing$, then we put $\mathcal{A} = \mathcal{A}_\varnothing$ and we obtain an $\mathcal{A}$-spectral function and an $\mathcal{A}$-scalar operator ([8]).

The support of an $\mathcal{A}_S$-spectral function U is denoted by $\operatorname{supp}(U)$ and it is defined as the smallest closed set $F \subset \overline{\Omega}$ such that $U_f = 0$ for $f \in \mathcal{A}_S$ with $\operatorname{supp}(f) \bigcap F = \varnothing$.

A subspace Y of X is said to be *invariant with respect to* an $\mathcal{A}_S$-spectral function $U : \mathcal{A}_S \to \mathbf{B}(X)$ if $U_f Y \subseteq Y$, for any $f \in \mathcal{A}_S$.

We recall several important properties of an $\mathcal{A}$-spectral function U ([8]), because we want to obtain similar properties for an $\mathcal{A}_S$-spectral function:

1. U_λ has the single-valued extension property, where λ is the identical function $f(\lambda) \equiv \lambda$;

2. $\sigma_{U_\lambda}(U_f x) \subset \operatorname{supp}(f)$, for any $f \in \mathcal{A}$ and $x \in X$;

3. If $\sigma_{U_\lambda}(x) \bigcap \operatorname{supp}(f) = \varnothing$, then $U_f(x) = 0$, for any $f \in \mathcal{A}$ and $x \in X$;

4. $x \in X_{U_\lambda}(F) = \{x \in X;\ \sigma_{U_\lambda}(x) \subset F\} \Leftrightarrow U_f(x) = 0$, for any $f \in \mathcal{A}$ with property $\operatorname{supp}(f) \bigcap F = \varnothing$, $F \subset \Omega$ closed;

5. $\operatorname{supp}(U) = \sigma(U_\lambda)$;

6. U_λ is decomposable.

Theorem 1.4. *Let $T \in \mathbf{B}(X)$ be an $\mathcal{A}_S$-scalar operator and let U be an $\mathcal{A}_S$-spectral function for T. Then we have*:

$\operatorname{supp}(U) \subset \sigma(T) \bigcup S$ *and* $\sigma(T) \subset \operatorname{supp}(U) \bigcup S$.

Proof. Let us consider $f \in \mathcal{A}_S$ such that $\operatorname{supp}(f) \cap (\sigma(T) \cup S) = \varnothing$. If $\xi \notin \operatorname{supp}(f)$ and λ is the identical function $f(\lambda) = \lambda$, then we have

$$(\xi I - U_\lambda) U_{f_\xi} = U_{(\xi - \lambda) f_\xi} = U_f$$

hence

$U_{f_\xi} = \Re(\xi, U_\lambda)\, U_f$, for $\xi \in \rho(U_\lambda) \cap \complement \operatorname{supp}(f)$.

The function

$$F(\xi) = \begin{cases} \Re(\xi, T) U_f, & \text{for } \xi \in \rho(U_\lambda) \\ U_{f_\xi}, & \text{for } \xi \in \complement \operatorname{supp}(f) \end{cases}$$

is entire and $\lim_{|\xi| \to \infty} \|F(\xi)\| = 0$, therefore $F \equiv 0$. It follows that $U_{f_\xi} = 0$ on $\complement \operatorname{supp}(f)$ and $U_f = 0$, hence

$$\operatorname{supp}(U) \subset \sigma(T) \cup S.$$

Let now $\xi_0 \notin \operatorname{supp}(U) \cup S$, let V_{ξ_0} be an open neighborhood of ξ_0 and let W be an open neighborhood of $\operatorname{supp}(U) \cup S$ such that $V_{\xi_0} \cap W = \varnothing$. Because the algebra $\mathcal{A}_S$ is S-normal, then there is a function $f \in \mathcal{A}_S$ with $f(\mu) = 1$ on W and $f(\mu) = 0$ for $\mu \in V_{\xi_0}$. Consequently

$$\operatorname{supp}(1 - f) \cap (\operatorname{supp}(U) \cup S) = \varnothing$$

hence

$$U_{1-f} = 0, \text{ i.e. } U_f = I.$$

Whence

$$U_{f_{\xi_0}} (\xi_0 I - U_\lambda) = (\xi_0 I - U_\lambda) U_{f_{\xi_0}} = U_f = I$$

therefore we finally have $\xi_0 \notin \sigma(U_\lambda) = \sigma(T)$ and hence $\sigma(T) \subset \operatorname{supp}(U) \cup S$.

Theorem 1.5. (Properties of $\mathcal{A}_S$-spectral functions)

Let U be an $\mathcal{A}_S$-spectral function (particularly, U is an $\mathcal{A}_S$-spectral function for an $\mathcal{A}_S$-scalar operator $T \in \mathbf{B}(X)$, $T = U_\lambda$). Then we have the following properties:

(1) *The spectral analytic residuum S_T has the property*: $S_T \subset S$; *when* $S_T = \varnothing$ (*particularly*, $S = \varnothing$), *then T has the single-valued extension property*;

(2) *If* $(\lambda_0 I - U_\lambda) x_0 = 0$, *with* $x_0 \neq 0$ *and* $f \in \mathcal{A}_S$ *with* $f(\lambda) = c$, *for* $\lambda \in G \cap \Omega$, *where G is a neighborhood of* λ_0, *then* $U_f x_0 = c x_0$;

(3) *If* $f \in \mathcal{A}_S$ *and* $x \in X$, *then* $\gamma_T(U_f x) \subset \operatorname{supp}(f)$; *moreover, if* $\operatorname{supp}(f) \supset S$, *then* $\sigma_T(U_f x) \subset \operatorname{supp}(f)$;

(4) *If* $f \in \mathcal{A}_S$ *such that* $\sigma_{U_\lambda}(x) \cap \operatorname{supp}(f) = \varnothing$ *and* $S_T = \varnothing$, *then* $U_f x = 0$;

(5) *If* $F \subset \Omega$ *closed, with* $F \supset S$, $x \in X$ *and* $S_T = \varnothing$, *then* $x \in X_{U_\lambda}(F)$ *if and only if* $U_f x = 0$, *for any* $f \in \mathcal{A}_S$ *with the property* $\operatorname{supp}(f) \cap F = \varnothing$;

(6) U_λ *is S-decomposable.*

Proof. The assertions (1) and (2) are proved in [14], Theorem 3.2, respectively Lemma 3.1.

(3) We observe that for any $\xi \notin \operatorname{supp}(f)$ we have $f_\xi \in \mathcal{A}_S$ and the X-valued function $\xi \to U_{f_\xi} x$ is analytic. Consequently,

$$(\xi I - T) U_{f_\xi} x = (\xi I - U_\lambda) U_{f_\xi} x = U_f x,$$

therefore $\xi \in \delta_T(U_f x)$, hence $\gamma_T(U_f x) \subset \operatorname{supp}(f)$.

Furthermore, for $f \in \mathcal{A}_S$ with $\operatorname{supp}(f) \supset S$, we deduce

$$\sigma_T(U_f x) = S_T \cup \gamma_T(U_f x) \subset S \cup \gamma_T(U_f x) \subset \operatorname{supp}(f)$$

(4) Let $x(\xi)$ be the unique analytic X-valued function defined on $\rho_{U_\lambda}(x)$ which satisfies the equality

$$(\xi I - U_\lambda) x(\xi) = x \text{ on } \rho_{U_\lambda}(x).$$

It results that

$$(\xi I - U_\lambda) U_f x(\xi) = U_f (\xi I - U_\lambda) x(\xi) = U_f x \text{ on } \rho_{U_\lambda}(x)$$

hence the following inclusions are obtained

$$\rho_{U_\lambda}(x) \subset \rho_{U_\lambda}(U_f x) \text{ and}$$

$$\sigma_{U_\lambda}(U_f x) \subset \sigma_{U_\lambda}(x).$$

From assertion (3),

$$\sigma_{U_\lambda}(U_f x) = \sigma_T(U_f x) = S_T \cup \gamma_T(U_f x) = \gamma_T(U_f x) \subset \operatorname{supp}(f),$$

hence

$$\sigma_{U_\lambda}\left(U_f x\right)\subset \operatorname{supp}(f)\bigcap\sigma_{U_\lambda}(x)=\varnothing,$$

therefore according to Proposition 1.1.2, [8], it follows that $U_f x=0$.

The property (5) can be obtained by using (4), as in the proof of Proposition 3.1.17, [8] and will be omitted.

The proof of (6) is presented in [14], Theorem 3.3.

Lemma 1.6. *Let U be an $\mathcal{A}_S$-spectral function. If G_1 is an open neighborhood of $\operatorname{supp}(U)$, $G_1\supset\operatorname{supp}(U)$ and G_2 is an open set such that $G_1\bigcup G_2\supset\overline{\Omega}$, $G_2\bigcap\operatorname{supp}(U)=\varnothing$ (i.e. $\{G_1,G_2\}$ is an open covering of $\overline{\Omega}$), then by S-normality of the algebra $\mathcal{A}_S$ it results that there are tow functions $f_1,f_2\in\mathcal{A}_S$ such that*:

$$0\le f_1(\lambda)\le 1,\ 0\le f_2(\lambda)\le 1,\ \lambda\in\Omega,$$

$$\operatorname{supp}(f_1)\subset G_1,\ \operatorname{supp}(f_2)\subset G_2 \text{ and}$$

$$f_1+f_2=1 \text{ on } \Omega.$$

With these conditions we have:

a) $U_{f_1}=I$, $U_{f_2}=0$

b) *For $f\in\mathcal{A}_S$ having the property that $f=1$ on a neighborhood of $\operatorname{supp}(U)$, it results that $U_f=I$.*

Proof. We have

$$\operatorname{supp}(1-f_1)=\operatorname{supp}(f_2)\subset G_2$$

$$\operatorname{supp}(1-f_1)\bigcap\operatorname{supp}(U)=\varnothing$$

hence

$$0=U_{1-f_1}=U_1-U_{f_1}$$

therefore

$$U_{f_1}=U_1=I \text{ and } U_{f_2}=0.$$

Moreover, for $f\in\mathcal{A}_S$ with the property that $f=1$ on a neighborhood of $\operatorname{supp}(U)$ we have

$$U_f=I$$

because it can be chosen in this case: $f_1=f$, $f_2=g$, with $\operatorname{supp}(g)\bigcap\operatorname{supp}(U)=\varnothing$, hence $U_g=0$ and accordingly

$U_{f+g}=U_1=I=U_f+U_g$, whence $U_f=I$.

Remark 1.7. From Lemma 1.6, it results that if $f\in\mathcal{A}_S$ and $f=1$ in a neighborhood of $\operatorname{supp}(U)$, then $U_f=I$. If we denote by $\bigvee_{f\in\mathcal{A}_0}U_f Y$ the linear subspace of X generated by $U_f Y$, where $Y\subset X$ and $\mathcal{A}_0$ is the set of all functions in $\mathcal{A}_S$ with compact support, then we have:

$$\bigvee_{f\in\mathcal{A}_0}U_f X=X.$$

Definition 1.8. Let U be an $\mathcal{A}_S$-spectral function. For any open set $G\in\mathcal{G}_S$ we denote

$$X_{[U]}(G)=\bigvee_{\operatorname{supp}(f)\subset G}U_f X$$

and for any closed set $F\in\mathcal{F}_S$ we put

$$X_{[U]}(F)=\bigcap_{G\supset F}X_{[U]}(G).$$

where $\mathcal{F}_S$ (respectively, $\mathcal{G}_S$) is the family of all closed (respectively, open) subsets $F\subset\mathbb{C}$ (respectively, $G\subset\mathbb{C}$) having the property: either $F\bigcap S=\varnothing$ or $F\supset S$ (respectively, $G\bigcap S=\varnothing$ or $G\supset S$).

Theorem 1.9. *Let U be an $\mathcal{A}_S$-spectral function. Then*

$$X_{[U]}(F)=X_{U_\lambda}(F)=\left\{x\in X;\ \sigma_{U_\lambda}(x)\subset F\right\}, \text{ for}$$

$$F\in\mathcal{F}_S,\ F\supset S.$$

Proof. If $\sigma_{U_\lambda}(x)\subset F$, for $F\in\mathcal{F}_S$, with $F\supset S$, let us consider $G\in\mathcal{G}_S$ an open set with $G\supset F\supset S$. Then by S-normality of $\mathcal{A}_S$ there is a function $f\in\mathcal{A}_S$ such that

$$f(\xi)=\begin{cases}1, \text{ for } \xi \text{ in a neighborhood of } \Omega\bigcap F\\ 0, \text{ for } \xi \text{ in a neighborhood of } \Omega\setminus(G\bigcap\Omega)\end{cases}$$

and therefore $\operatorname{supp}(f)\subset G$, whence

$$\operatorname{supp}(1-f)\bigcap\sigma_{U_\lambda}(x)\subset\operatorname{supp}(1-f)\bigcap F=\varnothing.$$

According to Theorem 1.5, $U_{1-f}x=0$, hence

$$x=U_f x\in X_{[U]}(G).$$

$G\in\mathcal{G}_S$ being an arbitrary open set with $G\supset F$, $F\in\mathcal{F}_S$, we have

$$x\in X_{[U]}(F), \text{ i.e. } X_{U_\lambda}(F)\subseteq X_{[U]}(F).$$

Conversely, let us show that $X_{[U]}(F)\subset X_{U_\lambda}(F)$, for any $F\in\mathcal{F}_S$, $F\supset S$.

Let $x\in X_{[U]}(F)\subset X_{[U]}(G)$, for any open set $G\in\mathcal{G}_S$, $G\supset F\supset S$, and let $G_1\in\mathcal{G}_S$ be an arbitrary

open set containing $\overline{G}$. By S-normality of $\mathcal{A}_S$, there is a function $f_1 \in \mathcal{A}_S$ such that

$$f_1(\xi) = \begin{cases} 1, & \text{for } \xi \in \overline{G} \cap \Omega \\ 0, & \text{for } \xi \in \Omega \setminus (G_1 \cap \Omega) \end{cases}$$

hence $\operatorname{supp}(f_1) \subset \overline{G}_1$. Therefore for any $f \in \mathcal{A}_S$ with $\operatorname{supp}(f) \subset G$ we have $f_1 f = f$, so that

$U_{f_1} U_f = U_f$, i.e. $U_{f_1} \mid X_{[U]}(G) = I \mid X_{[U]}(G)$

whence

$$U_{f_1} x = x.$$

According to Theorem 1.5 it follows that

$$\sigma_{U_\lambda}(x) = \sigma_{U_\lambda}(U_{f_1} x) = \gamma_{U_\lambda}(U_{f_1} x) \cup S_{U_\lambda} = \gamma_{U_\lambda}(U_{f_1} x) \subset \overline{G}_1$$

and hence

$$\sigma_{U_\lambda}(x) = \bigcap_{\substack{G_1 \in \mathcal{G}_S \\ G_1 \supset \overline{G}}} \overline{G}_1 = \overline{G}.$$

$G \in \mathcal{G}_S$ being an arbitrary open set, $G \supset F \supset S$, we obtain

$$\sigma_{U_\lambda}(x) \subset \bigcap_{\substack{G \in \mathcal{G}_S \\ G \supset F \supset S}} \overline{G} = F, \text{ hence } x \in X_{U_\lambda}(F).$$

Corollary 1.10. *If U is an $\mathcal{A}_S$-spectral function, then for any $F \in \mathcal{F}_S$ with $F \supset S$, $X_{[U]}(F)$ is a maximal spectral space for U_λ.*

Proof. It results easily from the previous theorem.

Theorem 1.11. *Let $T_1, T_2 \in \mathbf{B}(X)$. If T_1 is S-decomposable (in particular, decomposable) and T_1, T_2 are spectral equivalent, then T_2 is also S-decomposable (in particular, decomposable) and*

$$X_{T_1}(F) = X_{T_2}(F),$$

for any $F \subset \mathbb{C}$ closed, $F \supset S$ (when $S = \varnothing$, for any $F \subset \mathbb{C}$ closed).

If T_1 and T_2 are decomposable, then T_1 is spectral equivalent to T_2 if and only if their spectral spaces $X_{T_1}(F)$ and $X_{T_2}(F)$ are equal, i.e. $X_{T_1}(F) = X_{T_2}(F)$, *for any $F \subset \mathbb{C}$ closed* ([8], 2.2.1, 2.2.2).

If T_1 and T_2 are S-decomposable and spectral equivalent, then their spectral spaces are equal, i.e. $X_{T_1}(F) = X_{T_2}(F)$, for any $F \subset \mathbb{C}$ closed, $F \supset S$, but conversely is not true.

II. Spectral equivalence of $\mathcal{A}_S$-scalar operators. $\mathcal{A}_S$-decomposable and $\mathcal{A}_S$-spectral operators

For decomposable (respectively, spectral, S-decomposable, S-spectral) operators, we have several important results with respect to spectral equivalence property. Thus if $T_1, T_2 \in \mathbf{B}(X)$, T_1 is decomposable (respectively, spectral, S-decomposable, S-spectral) and T_1, T_2 are spectral equivalent, then T_2 is also decomposable (respectively, spectral, S-decomposable, S-spectral). Furthermore, if T_1 and T_2 are decomposable (respectively, spectral), then T_1, T_2 are spectral equivalent if and only if the spectral maximal spaces $X_{T_1}(F)$, $X_{T_2}(F)$ of T_1 and T_2, corresponding to any closed set $F \subset \mathbb{C}$, are equal (respectively, the spectral measures E_1, E_2 of T_1 and T_2 are equal) ([8], 2.2.1, 2.2.2, 2.2.4). For S-decomposable (respectively, S-spectral) operators, the equality of the spectral spaces (respectively, the equality of S-spectral measures) does not induce the spectral equivalence of the operators, but only their S-spectral equivalence.

The behaviour of $\mathcal{A}$-scalar and $\mathcal{A}_S$-scalar operators with respect to spectral equivalence is completely different. If $T_1 \in \mathbf{B}(X)$ is $\mathcal{A}$-scalar (respectively, $\mathcal{A}_S$-scalar) and $T_2 \in \mathbf{B}(X)$ is spectral equivalent to T_1, then T_2 is not $\mathcal{A}$-scalar (respectively, $\mathcal{A}_S$-scalar), in general; in this situation, we still know that T_2 is decomposable (respectively, S-decomposable) and then T_2 is said to be $\mathcal{A}$-decomposable (respectively, $\mathcal{A}_S$-decomposable). If in addition T commutes with one of its $\mathcal{A}$-spectral (respectively, $\mathcal{A}_S$-spectral) functions U, i.e. $T U_f = U_f T$, for any $f \in \mathcal{A}$ (respectively, for any $f \in \mathcal{A}_S$), then T is said to be $\mathcal{A}$-spectral (respectively, $\mathcal{A}_S$-spectral).

Definition 2.1. An operator $T \in \mathbf{B}(X)$ is called $\mathcal{A}_S$-*decomposable* if there is an $\mathcal{A}_S$-spectral function U such that T is spectral equivalent to U_λ.

In case that $S=\varnothing$, we have $\mathcal{A}_{\varnothing}=\mathcal{A}$, $\mathcal{A}_{\varnothing}$-spectral function is $\mathcal{A}$-spectral function, $\mathcal{A}_{\varnothing}$-decomposable operator is $\mathcal{A}$-decomposable operator ([8]).

Theorem 2.2. *Let* $T\in\mathbf{B}(X)$ *such that we consider the following two assertions*:

(I) *There is an* $\mathcal{A}_S$*-spectral function* U *such that* T *is spectral equivalent to* U_λ (i.e. T *is* $\mathcal{A}_S$*-decomposable*);

(II) *There is an* $\mathcal{A}_S$*-spectral function* U *such that for any closed set* $F\subset\mathbb{C}$, $F\supset S$, *we have*:

(a) $TX_{U_\lambda}(F)\subset X_{U_\lambda}(F)$

(b) $\sigma\left(T\middle|X_{U_\lambda}(F)\right)\subset F$.

Then the assertion (I) *implies the assertion* (II), *and for case* $S=\varnothing$, *the assertions* (I) *and* (II) *are equivalent*.

Proof. Let us suppose that there is an $\mathcal{A}_S$-spectral function U such that T and U_λ are spectral equivalent. Since U_λ is S-decomposable (Theorem 1.5), then, according to Theorem 1.11, it results that T is S-decomposable and we have

$$X_T(F)=X_{U_\lambda}(F) \qquad \textbf{(1)}$$

for any $F\subset\mathbb{C}$ closed, $F\supset S$. But $X_T(F)$ is invariant to T and $\sigma\left(T\middle|X_T(F)\right)\subset F$ (Theorem 2.1.3, [6]), whence it follows (by (1)) that

$$TX_{U_\lambda}(F)\subset X_{U_\lambda}(F)$$

and

$$\sigma\left(T\middle|X_{U_\lambda}(F)\right)\subset F.$$

In case $S=\varnothing$, if the assertion (II) is fulfilled, according to Theorem 2.2.6, [8], we deduce that T is decomposable and that the equality (1) holds for any closed set $F\subset\mathbb{C}$. Then T is spectral equivalent to U_λ (Theorem 2.2.2, [8]) and therefore (I) is verified.

Remark 2.3. If $T\in\mathbf{B}(X)$ is $\mathcal{A}_S$-decomposable and U is one of its $\mathcal{A}_S$-spectral functions, then:

1) T is S-decomposable;

2) $X_T(F)=X_{U_\lambda}(F)$, for any $F\subset\mathbb{C}$ closed, $F\supset S$;

3) If V is another $\mathcal{A}_S$-spectral function of T, then U_λ and V_λ are spectral equivalent (in particular, V_λ is spectral equivalent to T);

4) For $S=\varnothing$, if $\mathcal{A}$ is an inverse closed algebra of continuous functions defined on a closed subset of $\mathbb{C}$ and V is another $\mathcal{A}$-spectral function of T, then U_f and V_f are spectral equivalent, for any $f\in\mathcal{A}$ (see [8]).

Definition 2.4. An operator $T\in\mathbf{B}(X)$ is called $\mathcal{A}_S$-*spectral* if it is $\mathcal{A}_S$-decomposable and commutes with one of its $\mathcal{A}_S$-spectral functions, hence T is $\mathcal{A}_S$-spectral if there is an $\mathcal{A}_S$-spectral function U commuting with T such that T is spectral equivalent to U_λ.

For $S=\varnothing$, we have that an $\mathcal{A}_{\varnothing}$-spectral operator is an $\mathcal{A}$-spectral operator ([8]).

Theorem 2.5. *For an operator* $T\in\mathbf{B}(X)$ *we consider the following four assertions*:

(I) T *is* $\mathcal{A}_S$*-decomposable and commutes with one of its* $\mathcal{A}_S$*-spectral functions* (i.e. T *is* $\mathcal{A}_S$*-spectral*);

(II) $(\mathrm{II}1)$ T *is* S*-decomposable*;

$(\mathrm{II}2)$ *There is an* $\mathcal{A}_S$*-spectral function* U *commuting with* T, i.e. $U_fT=TU_f$, *for any* $f\in\mathcal{A}_S$;

$(\mathrm{II}3)$ $X_T(F)=X_{U_\lambda}(F)$, *for any* $F\subset\mathbb{C}$ *closed*, $F\supset S$;

(III) $(\mathrm{III}1)$ *There is an* $\mathcal{A}_S$*-spectral function* U *commuting with* T;

$(\mathrm{III}2)$ $\sigma\left(T\mid X_{U_\lambda}(F)\right)\subset F$, *for any* $F\subset\mathbb{C}$ *closed*, $F\supset S$;

(IV) $T=S+Q$, *where* S *is an* $\mathcal{A}_S$*-scalar operator and* Q *is a quasinilpotent operator commuting with an* $\mathcal{A}_S$*-spectral function of* S (*not to be confused the compact subset* S *with the operator* S *from the equality* $T=S+Q$, S *being the scalar part of* T *and* Q *the radical part of* T).

Then the assertions (I) *and* (IV), *respectively* (II) *and* (III) *are equivalent*, (I) *implies* (II), *respectively* (III), *and finally* (IV) *implies* (II).

Proof. $(\mathrm{I})\Rightarrow(\mathrm{II}),(\mathrm{III})$. Assuming (I) fulfilled, we prove that the assertions (II) and (III) are verified. If T is $\mathcal{A}_S$-decomposable and commutes with one of its $\mathcal{A}_S$-spectral functions U, then U_λ is spectral-equivalent to T.

Furthermore, U_λ being S-decomposable (Theorem 1.5), then T is S-decomposable (Theorem 1.12) and we have the equality:

$$X_T(F) = X_{U_\lambda}(F)$$

for any $F \subset \mathbb{C}$ closed, $F \supset S$, hence (II) is fulfilled. From Theorem 2.2, it follows that

$$\sigma\left(T \middle| X_{U_\lambda}(F)\right) = \sigma\left(T \middle| X_T(F)\right) \subset F$$

for any $F \subset \mathbb{C}$ closed, $F \supset S$, hence (III) is also verified.

$(\mathrm{I}) \Rightarrow (\mathrm{IV})$ T being $\mathcal{A}_S$-spectral, there is an $\mathcal{A}_S$-spectral function U commuting with T, i.e. $TU_f = U_f T$, for any $f \in \mathcal{A}_S$ (in particular, $TU_\lambda = U_\lambda T$) such that T is spectral equivalent to U_λ. But the operator U_λ is S-decomposable (Theorem 1.5), hence by Theorem 1.12, T is also S-decomposable and the following equality is verified

$X_T(F) = X_{U_\lambda}(F)$, for any $F \subset \mathbb{C}$ closed, $F \supset S$.

Using the fact that T and U_λ commute, it follows that $T - U_\lambda$ is a quasinilpotent operator commuting with U, because

$$(T - U_\lambda)^{[n]} = \sum_{k=0}^{n} (-1)^{n-k} T^k U_\lambda^{n-k} = (T - U_\lambda)^n$$

and the quasinilpotent equivalence of T and U_λ is given by

$$\lim_{n\to\infty} \left\| (T - U_\lambda)^{[n]} \right\|^{\frac{1}{n}} = \lim_{n\to\infty} \left\| (U_\lambda - T)^{[n]} \right\|^{\frac{1}{n}} = 0$$

(we remember that an operator T is quasinilpotent if $\lim_{n\to\infty} \left\| T^n \right\|^{\frac{1}{n}} = 0$ or, equivalently, $\sigma(T) = 0$). We remark that if U is an $\mathcal{A}_S$-spectral function, then U_λ is an $\mathcal{A}_S$-scalar operator. Putting $S = U_\lambda$ and $Q = T - U_\lambda$, we have

$$T = S + Q$$

where S is $\mathcal{A}_S$-scalar and Q is quasinilpotent (S is the scalar part of T and Q is the radical part of T).

$(\mathrm{IV}) \Rightarrow (\mathrm{I})$ By the hypothesis of assertion (IV), since S is an $\mathcal{A}_S$-scalar operator, we deduce that there is at least one $\mathcal{A}_S$-spectral function U of S such that: $S = U_\lambda$, the quasinilpotent operator Q commutes with U and S is S-decomposable (Theorem 1.5). It also results that $T = S + Q$ commutes with U (since we obviously have $U_\lambda U_f = U_f U_\lambda = = U_{\lambda f}$) and since $Q = T - S$ is quasinilpotent, then T is spectral equivalent to S, consequently T is $\mathcal{A}_S$-spectral.

$(\mathrm{III}) \Rightarrow (\mathrm{II})$ Assume that there is an $\mathcal{A}_S$-spectral function U commuting with T such that $\sigma\left(T \mid X_{U_\lambda}(F)\right) \subset F$, for $F \subset \mathbb{C}$ closed, $F \supset S$. On account of the definition and the properties of an $\mathcal{A}_S$-spectral function and of an $\mathcal{A}_S$-scalar operator, we remark that U_λ is an $\mathcal{A}_S$-scalar operator, hence U_λ is S-decomposable (Theorem 1.5) and we have $X_{U_\lambda}(F) = X_{[U]}(F)$, $F \subset \mathbb{C}$ closed, $F \supset S$ (Theorem 1.9). But $X_{U_\lambda}(F)$ is a spectral maximal space of U_λ (Theorem 2.1.3, [6]), hence it is ultrainvariant to U_λ (Proposition 1.3.2, [8]); therefore $X_{U_\lambda}(F)$ is invariant to T and then the restriction $T \mid X_{U_\lambda}(F)$ makes sense and $\sigma\left(T \mid X_{U_\lambda}(F)\right) \subset F$.

$(\mathrm{II}) \Rightarrow (\mathrm{III})$ The operator T being S-decomposable, according to Theorem 2.1.3, [6], we have that $X_T(F)$ is a spectral maximal space of T, for any $F \subset \mathbb{C}$ closed, $F \supset S$ and

$$\sigma\left(T \mid X_T(F)\right) \subset F \cap \sigma(T)$$

hence (by $(\mathrm{II3})$)

$$\sigma\left(T \middle| X_{U_\lambda}(F)\right) = \sigma\left(T \middle| X_T(F)\right) \subset F.$$

$(\mathrm{IV}) \Rightarrow (\mathrm{II})$ S being $\mathcal{A}_S$-scalar, there is an $\mathcal{A}_S$-spectral function U such that $S = U_\lambda$. But from Theorem 1.5, S is S-decomposable and applying Theorem 1.11 to T and S, we get that T is S-decomposable and

$$X_T(F) = X_S(F) = X_{U_\lambda}(F)$$

for any $F \subset \mathbb{C}$ closed, $F \supset S$.

The function U commutes with the quasinilpotent operator Q, i.e. $QU_f = U_f Q$, for $f \in \mathcal{A}_S$, hence $T = S + Q$ commutes with U.

Remark 2.6. With the same conditions as in Theorem 2.4, if $S = \varnothing$, then the four assertions above are equivalent (see [8]).

Remark 2.7. Let $T_1, T_2 \in \mathbf{B}(X)$ be two spectral equivalent operators. Then we have:

1) If $T_1 \in \mathbf{B}(X)$ is $\mathcal{A}_S$-scalar (respectively, $\mathcal{A}$-scalar), then T_2 is not $\mathcal{A}_S$-scalar (respectively, $\mathcal{A}$-scalar).

2) If $T_1 \in \mathbf{B}(X)$ is $\mathcal{A}_S$-decomposable (respectively, $\mathcal{A}$-decomposable), then T_2 is $\mathcal{A}_S$-decomposable (respectively, $\mathcal{A}$-decomposable).

3) If $T_1 \in \mathbf{B}(X)$ is $\mathcal{A}_S$-spectral (respectively, $\mathcal{A}$-spectral), then T_2 is not $\mathcal{A}_S$-spectral (respectively, $\mathcal{A}$-spectral).

III. Several properties of $\mathcal{A}_S$-scalar, $\mathcal{A}_S$-decomposable and $\mathcal{A}_S$-spectral operators.

In this section we study the behaviour of these three classes of operators with respect to direct sums, to restrictions and quotients with regard to an invariant subspace and to continuous algebraic homomorphisms, respectively.

Lemma 3.1.

1° *Let* $A_i, B_i \in \mathbf{B}(X_i)$ *such that* A_i *is quasinilpotent equivalent to* B_i, $i = 1,2$. *Then* $A_1 \oplus A_2$ *is quasinilpotent equivalent to* $B_1 \oplus B_2$;

2° *Let* $A, B \in \mathbf{B}(X)$ *such that* A *is quasinilpotent equivalent to* B *and let* Y *be a closed linear subspace of* X *invariant to both* A *and* B. *Then the restrictions* $A|Y$ *and* $B|Y$ *are quasinilpotent equivalent*;

3° *If* $A, B \in \mathbf{B}(X)$ *are two quasinilpotent equivalent operators and* $h : \mathbf{B}(X) \to \mathbf{B}(Y)$ *is a continuous homomorphism, then* $h(A)$ *is quasinilpotent equivalent to* $h(B)$. *Similarly, if* h *is an antihomomorphism.*

Proof. We remind that the multiplication between two operators A and B means here the composition of A with B. In general, A and B are not permutable, and the definition of quasinilpotent equivalence is reminded in Preliminaries; if A and B commute, the spectral equivalence is equivalent to the fact that the operator $A-B$ is quasinilpotent (i.e. $\lim_{n\to\infty} \left\| (A-B)^n \right\|^{\frac{1}{n}} = 0$).

We also remember that

$$(A_1 \oplus A_2)(B_1 \oplus B_2) = A_1 B_1 \oplus A_2 B_2$$

$$\alpha (A_1 \oplus A_2) = \alpha A_1 \oplus \alpha A_2$$

$$(A_1 \oplus A_2) + (B_1 \oplus B_2) = (A_1 + B_1) \oplus (A_2 + B_2)$$

hence $(A_1 \oplus A_2)^n = A_1^n \oplus A_2^n$.

If we use the notation $(A-B)^{[n]} = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} A^k B^{n-k}$, for any $A, B \in \mathbf{B}(X)$ then we have

$$\left((A_1 \oplus A_2) - (B_1 \oplus B_2)\right)^{[n]} =$$

$$\sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} (A_1 \oplus A_2)^k (B_1 \oplus B_2)^{n-k} =$$

$$= \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} \left(A_1^k \oplus A_2^k\right)\left(B_1^{n-k} \oplus B_2^{n-k}\right) =$$

$$= \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} \left(A_1^k B_1^{n-k} \oplus A_2^k B_2^{n-k}\right) =$$

$$= \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} A_1^k B_1^{n-k} \oplus \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} A_2^k B_2^{n-k} =$$

$$= (A_1 - B_1)^{[n]} \oplus (A_2 - B_2)^{[n]}$$

therefore

$$\left((A_1 \oplus A_2) - (B_1 \oplus B_2)\right)^{[n]} = (A_1 - B_1)^{[n]} \oplus (A_2 - B_2)^{[n]}$$

By the last equality and from the definition of the norm in the space $X_1 \oplus X_2$

$$\|x_1 \oplus x_2\|^2 = \|x_1\|^2 + \|x_2\|^2$$

we deduce that

$$\left(\left\| \left((A_1 \oplus A_2) - (B_1 \oplus B_2) \right)^{[n]} \right\|^2 \right)^{\frac{1}{n}} =$$

$$\left(\left\| (A_1 - B_1)^{[n]} \right\|^2 \oplus \left\| (A_2 - B_2)^{[n]} \right\|^2 \right)^{\frac{1}{n}}$$

and therefore assertion 1° is established.

2° For $A, B \in \mathbf{B}(X)$ and $Y \subset X$ a closed subspace invariant to both A and B, we obviously

have

$$\left(A-B\right)^{[n]}\left|Y=\left(\sum_{k=0}^{n}\left(-1\right)^{n-k}\binom{n}{k}A^{k}B^{n-k}\right)\right|Y=$$

$$\sum_{k=0}^{n}\left(-1\right)^{n-k}\binom{n}{k}\left(A^{k}B^{n-k}\left|Y\right.\right)=$$

$$=\sum_{k=0}^{n}\left(-1\right)^{n-k}\binom{n}{k}\left(A^{k}\left|Y\right.\right)\left(B^{n-k}\left|Y\right.\right)=$$

$$\sum_{k=0}^{n}\left(-1\right)^{n-k}\binom{n}{k}\left(A\left|Y\right.\right)^{k}\left(B\left|Y\right.\right)^{n-k}=$$

$$=\left(\left(A\left|Y\right.\right)-\left(B\left|Y\right.\right)\right)^{[n]}$$

and applying the inequality $\left\|T\left|Y\right.\right\|\le\left\|T\right\|$ and the fact that A and B are quasinilpotent equivalent, it follows that $A\left|Y\right.$ is spectral equivalent to $B\left|Y\right.$.

3° If $A, B\in\mathbf{B}\left(X\right)$ and $h:\mathbf{B}\left(X\right)\to\mathbf{B}\left(Y\right)$ is a continuous homomorphism (respectively, antihomomorphism), then we obtain

$$h\left(\left(A-B\right)^{[n]}\right)=h\left(\sum_{k=0}^{n}\left(-1\right)^{n-k}\binom{n}{k}A^{k}B^{n-k}\right)=$$

$$=\sum_{k=0}^{n}\left(-1\right)^{n-k}\binom{n}{k}h\left(A^{k}\right)h\left(B^{n-k}\right)=$$

$$\sum_{k=0}^{n}\left(-1\right)^{n-k}\binom{n}{k}\left(h\left(A\right)\right)^{k}\left(h\left(B\right)\right)^{n-k}=$$

$$=\left(h\left(A\right)-h\left(B\right)\right)^{[n]}$$

(respectively,

$h\left(\left(A-B\right)^{[n]}\right)=\left(-1\right)^{n}\left(h\left(A\right)-h\left(B\right)\right)^{[n]}$) and

$$\left\|\left(h\left(A\right)-h\left(B\right)\right)^{[n]}\right\|\le\left\|h\right\|\left\|\left(A-B\right)^{[n]}\right\|$$

hence if A and B are quasinilpotent equivalent, it results that $h\left(A\right)$ is spectral equivalent to $h\left(B\right)$.

Proposition 3.2. *Let $\mathcal{A}_S$ be an S-admissible algebra and let X_1 and X_2 be two Banach spaces. If $T_1\in\mathbf{B}\left(X_1\right)$ and $T_2\in\mathbf{B}\left(X_2\right)$ are $\mathcal{A}_S$-scalar (respectively, $\mathcal{A}_S$-decomposable or $\mathcal{A}_S$-spectral) operators, then $T_1\oplus T_2\in\mathbf{B}\left(X_1\oplus X_2\right)$ is also $\mathcal{A}_S$-scalar (respectively, $\mathcal{A}_S$-decomposable or $\mathcal{A}_S$-spectral).*

Proof. If T_1 and T_2 are $\mathcal{A}_S$-scalar, then there are two $\mathcal{A}_S$-spectral functions U^1 and U^2 such that

$$U^1_\lambda=T_1,\ U^1_1=I_{X_1}\text{ and }U^2_\lambda=T_2,\ U^2_1=I_{X_2}$$

(where λ and 1 are the functions $f\left(\lambda\right)=\lambda$ and $f\left(\lambda\right)=1$).

The mapping

$$U^1\oplus U^2:\mathcal{A}_S\to\mathbf{B}\left(X_1\oplus X_2\right),\ f\to U^1_f\oplus U^2_f,$$

is evidently an $\mathcal{A}_S$-spectral function for $T_1\oplus T_2$, because we have:

$$\left(U^1\oplus U^2\right)_\lambda=U^1_\lambda\oplus U^2_\lambda=T_1\oplus T_2$$

$$\left(U^1\oplus U^2\right)_1=U^1_1\oplus U^2_1=I_{X_1}\oplus I_{X_2}=I_{X_1\oplus X_2}$$

$\xi\to U^1_{f_\xi}\oplus U^2_{f_\xi}$ is analytic on $\mathbb{C}\,\mathrm{supp}\left(f\right)$,

since $\xi\to U^1_{f_\xi}$ $\xi\to U^2_{f_\xi}$ are analytic on $\mathbb{C}\,\mathrm{supp}\left(f\right)$,

hence $T_1\oplus T_2$ is $\mathcal{A}_S$-scalar.

If T_1 and T_2 are $\mathcal{A}_S$-decomposable, then there are two $\mathcal{A}_S$-spectral functions U^1 and U^2 such that T_1 is spectral equivalent to U^1_λ, respectively T_2 is spectral equivalent to U^2_λ. According to Lemma 3.1, we have that $T_1\oplus T_2$ is spectral equivalent to $U^1_\lambda\oplus U^2_\lambda$, hence $T_1\oplus T_2$ is $\mathcal{A}_S$-decomposable.

If T_1 and T_2 are $\mathcal{A}_S$-spectral, then there are two $\mathcal{A}_S$-spectral functions U^1 and U^2 such that U^1 commutes with T_1 and T_1 is spectral equivalent to U^1_λ, respectively U^2 commutes with T_2 and T_2 is spectral equivalent to U^2_λ.

It is obvious that $U^1\oplus U^2$ commutes with $T_1\oplus T_2$:

$\left(U^1 \oplus U^2\right)_f (T_1 \oplus T_2) = \left(U^1_f \oplus U^2_f\right)(T_1 \oplus T_2) =$
$U^1_f T_1 \oplus U^2_f T_2 = T_1 U^1_f \oplus T_2 U^2_f =$
$(T_1 \oplus T_2)\left(U^1_f \oplus U^2_f\right) = (T_1 \oplus T_2)\left(U^1 \oplus U^2\right)_f$,
$f \in \mathcal{A}_S$.

From Lemma 3.1, $T_1 \oplus T_2$ is spectral equivalent to $U^1_\lambda \oplus U^2_\lambda$, hence $T_1 \oplus T_2$ is $\mathcal{A}_S$-spectral.

Proposition 3.3. *Let $T \in \mathbf{B}(X)$ be an $\mathcal{A}_S$-scalar (respectively, $\mathcal{A}_S$-decomposable or $\mathcal{A}_S$-spectral) operators and let Y be a closed linear subspace of X which is invariant to T and to one of its $\mathcal{A}_S$-spectral functions. Then the restriction $T|Y$ is $\mathcal{A}_{S_1}$-scalar (respectively, $\mathcal{A}_{S_1}$-decomposable or $\mathcal{A}_{S_1}$-spectral), where $S_1 = S \cap \sigma(T|Y)$.*

Proof. Let us suppose that T is $\mathcal{A}_S$-scalar and let U be an $\mathcal{A}_S$-spectral function for T such that Y is invariant to both T and U. Then the restrictions $T|Y$, $U|Y$, $U_f|Y$, $f \in \mathcal{A}_S$ make sense. Putting $V_f = U_f|Y$ we obtain a $\mathbf{B}(Y)$-valued function V which is an $\mathcal{A}_{S_1}$-spectral function for $T|Y$, since:

$$V_\lambda = U_\lambda|Y = T|Y$$
$$V_1 = U_1|Y = I_Y$$

hence the operator $T|Y$ is $\mathcal{A}_{S_1}$-scalar, where $S_1 = S \cap \sigma(T|Y)$.

If T is $\mathcal{A}_S$-decomposable, then T is spectral equivalent to U_λ and according to Lemma 3.1, it results that $T|Y$ is spectral equivalent to $U_\lambda|Y = V_\lambda$, therefore $T|Y$ is $\mathcal{A}_{S_1}$-decomposable.

If T is $\mathcal{A}_S$-spectral, then T commutes with U and T is spectral equivalent to U_λ. It is clear that $T|Y$ commutes with V and according to Lemma 3.1, $T|Y$ is spectral equivalent to V_λ, hence $T|Y$ is $\mathcal{A}_{S_1}$-spectral.

Proposition 3.4. *Let X and Y be two Banach spaces, let $T \in \mathbf{B}(X)$ be an $\mathcal{A}_S$-scalar (respectively, $\mathcal{A}_S$-decomposable or $\mathcal{A}_S$-spectral) and let $h: \mathbf{B}(X) \to \mathbf{B}(Y)$ be a continuous homomorphism or antihomomorphism. Then $h(T)$ is also $\mathcal{A}_S$-scalar (respectively, $\mathcal{A}_S$-decomposable or $\mathcal{A}_S$-spectral).*

Proof. We remark that the $\mathcal{A}_S$-spectral functions U and $h(U)$ are defined using the same S-admissible algebra $\mathcal{A}_S$. The S-admissible algebra $\mathcal{A}_S$ with which it is defined the $\mathcal{A}_S$-spectral function U is also the same for the $\mathcal{A}_S$-spectral function $h(U)$ with the same S.

If $U: \mathcal{A}_S \to \mathbf{B}(X)$ is an $\mathcal{A}_S$-spectral function, then the mapping $h(U): \mathcal{A}_S \to \mathbf{B}(Y)$ defined by $h(U)_f = h(U_f)$, $f \in \mathcal{A}_S$ is also an $\mathcal{A}_S$-spectral function.

Let us suppose that T is an $\mathcal{A}_S$-scalar operator and let U be an $\mathcal{A}_S$-spectral function for T, i.e. $U_\lambda = T$, $U_1 = I_X$. Then $h(U)$ is an $\mathcal{A}_S$-spectral function for $h(T)$:

$$h(U)_\lambda = h(U_\lambda) = h(T)$$
$$h(U)_1 = h(U_1) = h(I_X) = I_Y$$

hence $h(T)$ is $\mathcal{A}_S$-scalar.

If T is $\mathcal{A}_S$-decomposable, then T is spectral equivalent to U_λ and according to Lemma 3.1, it results that $h(T)$ is spectral equivalent to $h(U_\lambda) = h(U)_\lambda$, therefore $h(T)$ is $\mathcal{A}_S$-decomposable.

If T is $\mathcal{A}_S$-spectral, then T commutes with U and T is spectral equivalent to U_λ. It is obvious that $h(T)$ commutes with $h(U)$ and according to Lemma 3.1, $h(T)$ is spectral equivalent to $h(U)_\lambda$, hence $T|Y$ is $\mathcal{A}_S$-spectral.

Corollary 3.5. *If $T \in \mathbf{B}(X)$ is an $\mathcal{A}_S$-scalar (respectively, $\mathcal{A}_S$-decomposable or $\mathcal{A}_S$-spectral), U is one of its $\mathcal{A}_S$-spectral functions and Y is a closed linear subspace of X invariant to both T and U, then $\dot{T} \in \mathbf{B}(\dot{X})$, defined by $\dot{T}\dot{x} = \dot{\overline{Tx}}$, $\dot{x} \in \dot{X}$, the quotient operator induced by T in*

the quotient space $\dot{X} = X / Y$, *is also an* $\mathcal{A}_S$*-scalar* (*respectively*, $\mathcal{A}_S$*-decomposable or* $\mathcal{A}_S$*-spectral*) *operator*.

Proof. This is an immediate consequence of the previous proposition.

IV. CONCLUSIONS

We will underline the relevance, importance and necessity of studying the $\mathcal{A}_S$-scalar (respectively, $\mathcal{A}_S$-decomposable or $\mathcal{A}_S$-spectral) operators, showing the consistence of this class, in the sense of how many and how substantial its subfamilies are.

These operators are natural generalizations of the notions of $\mathcal{A}$-scalar, $\mathcal{A}$-decomposable and $\mathcal{A}$-spectral operators studied in [8] and appear, in general, as restrictions or quotients of the last one.

We demonstrated some of their properties, leaving the challenge to proof and generalize many others.

REFERENCES

[1] ALBRECHT, E.J., ESCHMEIER, J., *Analytic functional models and local spectral theory*, Proc. London Math. Soc., **75**, 323-348, 1997.

[2] APOSTOL, C., *Spectral decompositions and functional calculus*, Rev. Roum. Math. Putes et Appl., **13**, 1481-1528 1968.

[3] BACALU, I., *On restrictions and quotients of decomposable operators*, Rev. Roum. Math. Pures et Appl., **18**, 809-813, 1973.

[4] BACALU, I., $S-$ *decomposable operators in Banach spaces*, Rev. Roum. Math. Pures et Appl., **20**, 1101-1107, 1975.

[5] BACALU, I., *Some properties of decomposable operators*, Rev. Roum. Math. Pures et Appl., **21**, 177-194, 1976.

[6] BACALU, I., *Descompuneri spectrale reziduale* (Residually spectral decompositions), St. Cerc. Mat. I (1980), II (1980), III (1981).

[7] BACALU, I., $S-$ Spectral Decompositions, Ed. Politehnica Press, Bucharest, 2008.

[8] COLOJOARĂ, I., FOIAŞ, C., *Theory of generalized spectral operators*, Gordon Breach, Science Publ., New York-London-Paris, 1968.

[9] DOWSON, H.R., *Restrictions of spectral operators*, Proc. London Math. Soc., **15**, 437-457, 1965.

[10] DUNFORD, N., SCHWARTZ, J.T., *Linear operators*, Interscience Publishers, New York, I (1958), II (1963), III (1971).

[11] LAURSEN, K.B., NEUMANN, M.M., *An Introduction to Local Spectral Theory*, London Math. Soc. Monographs New Series, Oxford Univ. Press., New-York, 2000.

[12] VASILESCU, F.H., *Residually decomposable operators in Banach spaces*, Tôhoku Math. Journ., **21**, 509-522, 1969.

[13] VASILESCU F.H., *Analytic Functional Calculus and Spectral Decompositions*, D. Reidel Publishing Co., Dordrecht, Editura Academiei, Bucharest, 1982.

[14] ZAMFIR, M., BACALU, I., $\mathcal{A}_S-$-scalar operators, U.P.B. Sci. Bull., Series A, 74, 89-98, 2012.

Pulsatile Non-Newtonian Flows in a Dilated Vessel

Iqbal Husain, Christian R Langdon and Justin Schwark

Abstract— The aim of this study is to investigate several mathematical models describing pulsatile blood flow through the cardiovascular system. Specifically, this study considers the numerical simulation of blood flow through a three-dimensional model of an aneurysm in the common carotid artery in order to better understand the hemodynamic that may contribute to the growth of this aneurysm. Four non-Newtonian blood models, namely the Power Law, Casson, Carreau and the Generalized Power Law, as well as the Newtonian model of blood viscosity, are used to investigate the flow effects induced by these different blood constitutive equations. Results show significant differences between modeling blood as a Newtonian and non-Newtonian fluid at low shear rates. The dependence of the flow on the degree of abnormality is examined and differences from the Newtonian case are discussed.

.

***Keywords*—**blood, finite element, non-Newtonian, pulsatile

INTRODUCTION

This paper examines the flow dynamics in a representative model of an aneurysm in the common carotid artery under physiologically realistic pulsatile conditions and compares it with a healthy carotid artery for various degree of dilation using five blood rheological models. The results of transient simulations are presented in this paper while a companion paper investigates steady state flow.

An aneurysm is an area of localized dilation of a blood vessel. An aneurysm in the carotid artery involves the two carotid arteries, the left and right common carotid arteries (CCAs) that supply blood to the brain. They supply the large, front part of the brain, which is responsible for our personality and our ability to think, speak and move. Aneurysms are frequently located at branching points of the major arteries. Most aneurysms are fusiforms. They are shaped like a spindle with widening all around the circumference of an artery. The inside walls are often lined with a laminated blood clot. Aneurysms are most common after 60 years of age. Men are more likely than women to be affected. The foremost health danger of this aneurysm is rupture which leads to death in up to 90% of the victims.

Iqbal Husain is with Luther College – University of Regina, Regina, SK, Canada S4S 0A2 (corresponding author phone: 306-585-45751; fax: 306-585-5267; e-mail: Iqbal.Husain@ uregina.ca).

Chris Langdon is with Luther College – University of Regina, Regina, SK, Canada.

Justin Schwark is with the Mathematics Department, University of Regina, Regina, SK, Canada.

The most common cause of an aneurysm is hardening of the arteries, called arteriosclerosis [1]. The arteriosclerosis can weaken the arterial wall and the pressure of the blood being pumped through the aorta causes expansion at the site of weakness. Consequently, hemodynamic factors such as blood velocity, wall shear stress and pressure play important roles in the pathogenesis of aneurysms and thrombosis. The geometry of the aneurysm, its volume and aspect ratio (depth/neck width) and its relation to the parent vessel are also important factors affecting its hemodynamic.

Although the rupture of an aneurysm is thought to be associated with a significant change in its size, there is still some debate over the size at which rupture occurs. The relationship between geometric features and rupture is closely associated with low flow conditions. The stagnation of blood flow in large aneurysms is commonly observed. Clearly, a better understanding of aneurysm growth and rupture is needed.

Recently, Valencia and Solis [2], examined blood flow dynamics in a saccular aneurysm model with elastic walls of the basilar artery. They found the shear stress on the aneurysm wall and its deformation dependent on the wall thickness and the elastic or hyperelastic wall model. Oshima *et al.* [3] employed the finite-element method to study the flow in a cerebral aneurysm. Their geometrical model was derived from computed tomography data. The finite-element method was also used by Kumar and Naidu [4], to perform 2D axisymmetric simulations in aneurysms models with 0 – 75% dilation. Their results examined the sensitivity of various flow parameters to dilation height. Neofytou and Tsangaris [5], used a finite volume scheme to numerical simulate the effects of various blood rheological models in flows through a stenosis and an abdominal aortic aneurysm. Their results indicated significant differences between modeling blood as Newtonian and non-Newtonian fluids.

There are three objectives of this study: first, to investigate the variation in wall shear stress in an aneurysm of the carotid artery at different flow rates and degrees of dilation; second, to compare the various blood models and hence quantify the differences between the models and judge their significance and lastly, to determine whether the use of the Newtonian blood model is appropriate over a wide range of shear rates.

ISBN: 978-1-61804-240-8

2 MATHEMATICAL MODELLING

2.1 Governing equations

We assumed the blood flow to be laminar and incompressible and therefore the governing Navier-Stokes equations for such flows are given by

$$\nabla \cdot V = 0 \tag{1}$$

$$\rho\left(\frac{\partial V}{\partial t} + V \,.\, \nabla V\right) = -\nabla . \tau - \nabla p \tag{2}$$

where V is the three-dimensional velocity vector, p pressure, ρ density and τ the shear stress term.

We considered four different non-Newtonian blood flow models and compared the results obtained with that from the simple Newtonian model in this study. The effects of these models on the flow field and the wall shear stress in the vicinity of the aneurysm are examined. These models are given below [6].

Blood Models

1. Newtonian model

$$\mu = 0.00345 \quad Pa \cdot s \tag{3}$$

2. Power Law Model

$$\mu = \mu_0 (\dot{\gamma})^{n-1} \tag{4}$$

where μ_0=0.01467 and $n = 0.7755$.

3. Casson Model

$$\mu = \frac{[\sqrt{\tau_y} + \sqrt{\eta|\dot{\gamma}|}]^2}{|\dot{\gamma}|} \tag{5}$$

where $\eta = \eta_0(1-H)^{-2.5}$ and $\tau_y = 0.1(0.625H)^3$ with η_0= 0.0012 Pa·s and $H = 0.37$.

4. Carreau Model

$$\mu = \mu_\infty + (\mu_0 - \mu_\infty)\left[1 + (\lambda\dot{\gamma})^2\right]^{(n-1)/2} \tag{6}$$

where $\mu_0 = 0.056$ Pa·s, $\mu_\infty = 0.00345$ Pa·s, $\lambda = 3.313$ s and $n = 0.3568$.

5. Generalized Power Law Model

$$\mu = \lambda|\dot{\gamma}|^{n-1} \tag{7}$$

where

$$\lambda(\dot{\gamma}) = \mu_\infty + \Delta\mu \exp\left[-\left(1+\frac{|\dot{\gamma}|}{a}\right)\exp\left(\frac{-b}{|\dot{\gamma}|}\right)\right],$$

$$n(\dot{\gamma}) = n_\infty - \Delta n \exp\left[-\left(1+\frac{|\dot{\gamma}|}{c}\right)\exp\left(\frac{-d}{|\dot{\gamma}|}\right)\right]$$

$\mu_\infty = 0.00345$, $n_\infty = 1.0$, $\Delta\mu = 0.25$, $\Delta n = 0.45$, $a = 50$, $b = 3$, $c = 50$ and $d = 4$.

These models have been created by various researchers Walburn and Schneck [7], Cho and Kensey [8], Fung [9], Ballyk et al. [10]and Johnston et al. [11], by fitting the input parameters in these models to experimental data for blood viscosity measured at certain shear rates.

2.2 Geometry

The diameters of the left and right common carotid artery show a significant amount of variation. A study of 17 healthy subjects [12] produced diameters in the range of 0.52 to 0.75 cm with an average diameter of 0.64 cm. In this study, two different model of the abdominal aortic aneurysm are used to investigate blood flow in the initial stages of its development. The carotid artery before and after the dilation is idealized as a straight rigid tube without any branching arteries.

The flow geometry then consists of straight rigid tube of diameter *d* and is divided into three regions, the inlet, the dilation and the outlet region. The lengths of these regions are 4*d*, 4*d* and 18*d*, respectively. The radius of the undeformed inlet and outlet is $r_0 = d/2$. Two different values of the diameter were used to model the carotid artery, namely, $d = 0.64$ cm and $d = 2.0$ cm.

The radius of the diseased region [12] is given by

$$r = r_0 + \left[a - \left(\frac{a^2 + (b/2)^2}{2a}\right) + \sqrt{\left(\frac{a^2 + (b/2)^2}{2a}\right) - (b/2 - x)^2}\right] \qquad 0 \le x \le b \tag{8}$$

where x is the distance from the start of the aneurysm, *a* is the degree of dilation, and *b is* the overall length of the diseased area as shown in Figure 1. Each model had an aspect ratio $\frac{b}{d} = 4$ which is typical of fusiform aneurysms.

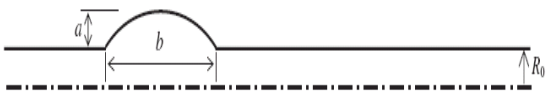


Fig. 1. Aneurysm Geometry

Three different degrees of dilation, 25%, 40% and 55% were used in this study.

2.3 Assumptions and boundary conditions

We assume the arterial walls to be rigid and apply the no-slip condition at the walls. At the outlet, stress-free conditions are applied and the pressure is set to zero. Symmetry is assumed at the centerline. Finally, the velocity profile at the inlet is regarded to be that of fully developed flow in a straight tube and is given by

$$u = \bar{u}\left[1 - \left(\frac{r}{r_0}\right)^2\right] \qquad 0 \le r \le r_0 \qquad (9)$$

where u is the velocity component in the $x-$ direction and $\bar{u}$ is the centerline velocity specified at the inlet.

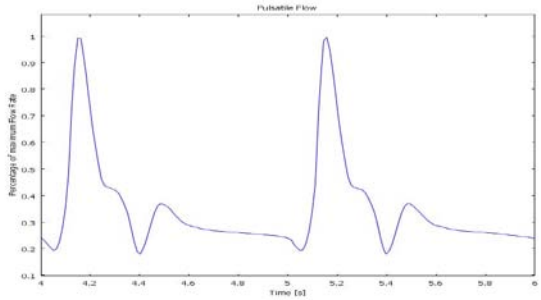


Fig. 2. Physiological flow waveform in the carotid artery used to drive the inlet velocity boundary condition as a function of time.

In transient flow, the pulsatile flow prescribed at the inlet of the artery is given by a time varying forcing function given in Figure 2. This physiological profile was obtained by averaging the pulsed Doppler spectra data collected from the left and right common carotid arteries of 17 normal volunteers [12]. The data was acquired over approximately 50 cardiac cycles and analyzed in both the time and frequency domains to determine the average properties and variability of human carotid waveform. In this study, this forcing function was scaled to yield a maximum inflow velocity of $\bar{u}$ with a heart rate of approximately 60 beats per minute.

2.4 Solution methodology

The governing equations are highly nonlinear and are solved numerically using techniques of computational fluid dynamics. In this study, these equations are solved using the finite element method as implemented by COMSOL (COMSOL Inc., Los Angeles, CA). The flow geometry for the aneurysm was first created using Matlab. Then a finite element mesh was placed on this geometry. Briefly, an inlet plane of the artery is meshed in 2D using triangles and this mesh is extruded along the centerline of the artery to create a 3D mesh consisting of hexadrel elements. The mesh used for all computations consisted of 17,696 elements and 27,132 nodes for the aneurysm.

Grid independence was determined by performing additional simulations using a greater number of nodes. A mesh size of 107, 350 nodes for the aneurysm was used and the results obtained differed from those on the original mesh by less than 1%.

3. RESULTS AND DISCUSSION

Pulsatile inflow simulations were performed using all five models given above. As stated, three different degrees of dilation of the aneurysm were examined namely 25%, 40% and 55%. Several flow rates $(\bar{u})$ were used in these simulations, from 0.04 m/sec to 0.22 m/sec, corresponding to the lower range of the average maximum systolic velocity of 1.08 m/sec in the common carotid artery as reported in [12].

A comparison of the streamlines patterns from various models in pulsatile flow simulation shows that the flow follows the contour of the wall (attached flow) throughout the aneurysm during the early systolic phase. During late systole, a vortex begins to form at the proximal end of the aneurysm. By the late diastolic stage, the flow becomes vortex dominated with the vortex filling the entire aneurysm. This flow pattern is similar in larger aneurysm except that the vortex strength and the translational speed increases. There exists minor differences in the recirculation regions shown by each model and these differences become more prominent at 55% dilation and higher flow rates, specifically the growth of the recirculation region and the vortex length. The Newtonian model shows the largest recirculation region and the Power law model, the smallest.

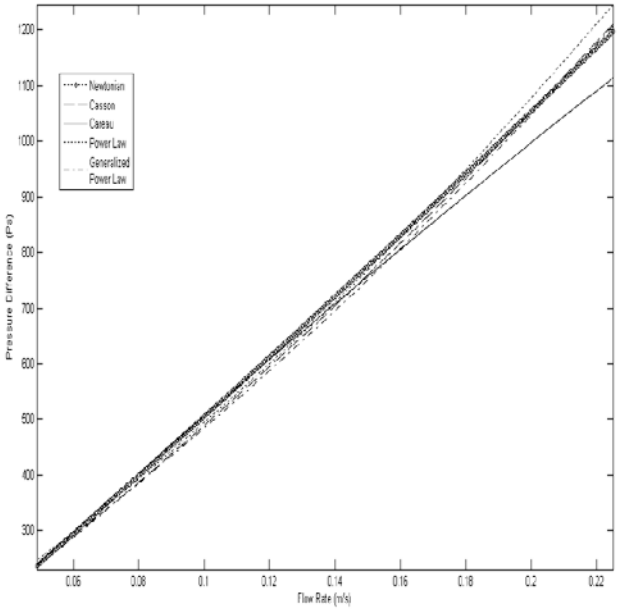


Fig. 3. Pressure difference distribution as a function of flow rate.

The distribution of maximum pressure with the flow rate is shown in Figure 3. This figure shows that all of the non-Newtonian models considered here produce a lower pressure difference than the Newtonian model at low flow rates.

Specifically, the lowest pressure drop is induced by the Generalized Power Law model. At higher flow rates, greater than 0.16 m/s, the Generalized Power Law model comes in close agreement with the Newtonian model whilst the Careau model begins to deviate, giving pressure differences that are significantly less than the Newtonian fluid. It is not clear why this is the case and further study is required to explain this behaviour. Also, at flow rates greater than 0.18 m/s, the Power Law model begins to breakdown, producing pressure differences significantly higher than the Newtonian model. The other non-Newtonian models show good agreement in pressure differences with the Newtonian case for various degrees of severity and as the flow rate increases. This result agrees well with that of [5] in the transient non-pulsatile case.

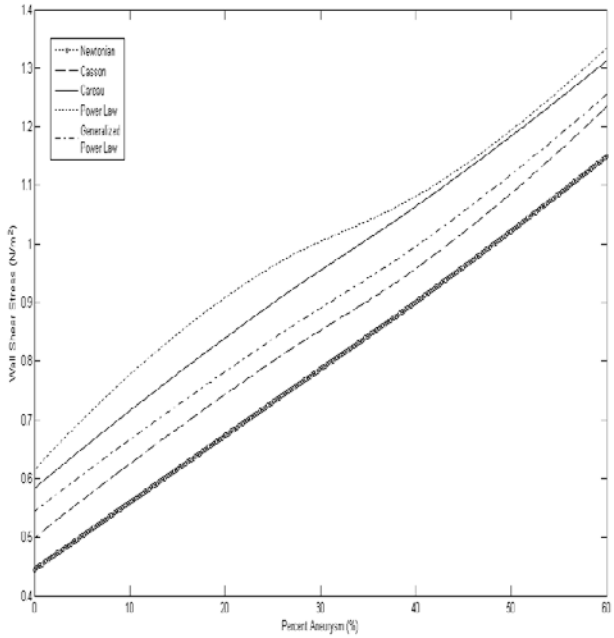


Fig. 4. Wall shear stress distribution for various degrees of dilation.

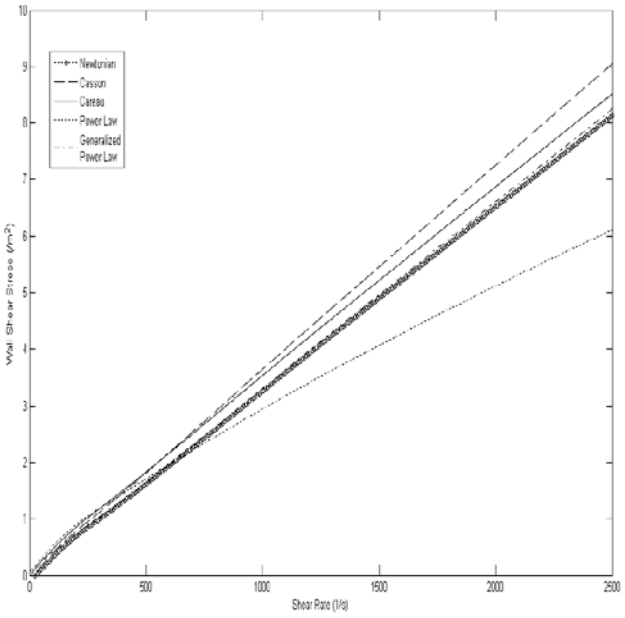


Fig. 5. Wall shear stress vs shear rate.

The distribution of the wall shear stress (WSS) is one of the most important hemodynamic parameter due to its direct relevance in artherosclerosis formation. The distribution of wall shear stress with the size of the aneurysm for the 0.64 cm diameter artery is shown in Figure 4. It is evident that WSS increases with increasing dilation. All of the non-Newtonian models give values that are higher than the Newtonian case for various flow rates, especially, the Power Law model. At all degrees of dilations, the WSS values predicted by this model are significantly higher than that of the Newtonian model. Figure 5 displays the distribution of maximum WSS with shear rate. Again, WSS increases with increasing shear rate with the Power Law model deviating significantly from the rest at higher shear rates. The Casson and the Carreau models produce higher WSS values compared to the Newtonian model at high shear rates but are in good agreement with the Newtonian values at low shear rates. The Generalized Power Law model compares very well with the Newtonian model at both high and low shear rates.

ISBN: 978-1-61804-240-8

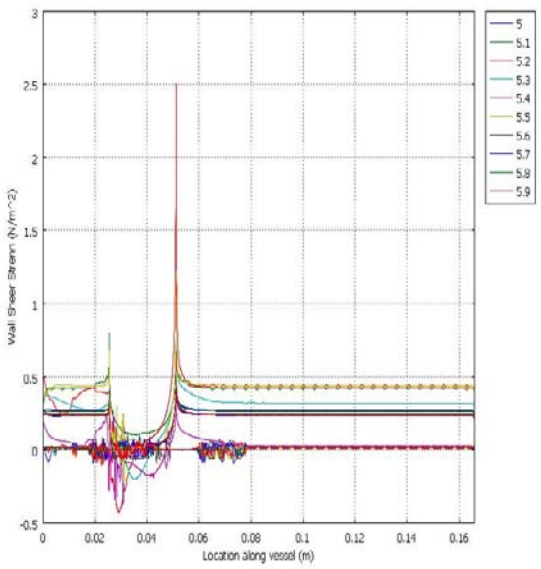


Fig. 6. Wall shear stress distribution for an artery with a 55% aneurysm using the generalized power law.

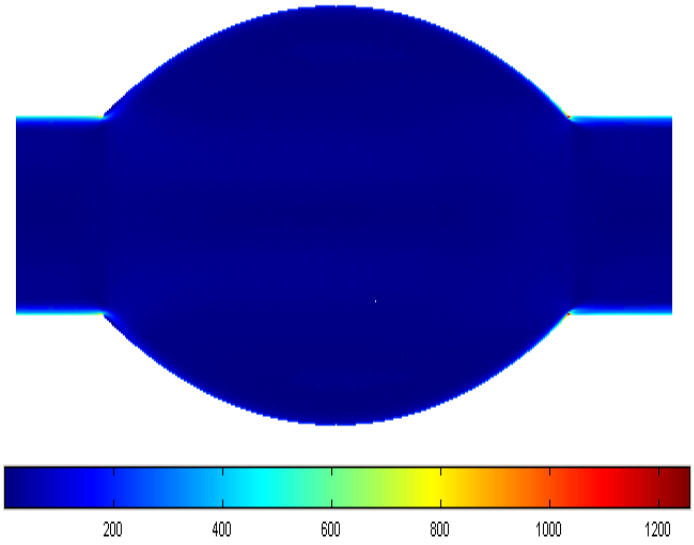


Fig. 7. Shear rate distribution in an artery containing a 55% aneurysm.

Figure 6 show the distributions of shear stress for a 55% aneurysm obtained from the Generalized Power Law at various times in a cardiac cycle. As can be seen, the shear stress drops abruptly as the flow enters the aneurysm. The magnitude of this drop increases with higher flow rates. This is followed by a sharp rise at the end of the aneurysm. Further downstream, the WSS rapidly regains its undisturbed value. The maximum wall shear stress occurs in the middle of the cycle corresponding to the maximum inflow velocity. The distribution of shear rates in a 55% dilated artery is shown in Figure 7. The high shear rates are confined to the small areas at the entrance and exit to the aneurysm and immediately downstream.

The maximum and minimum WSS values are in good agreement for the Generalized Power Law and the Newtonian models. The Power Law model gives a much lower value because it exhibits a lower viscosity at the entrance and exit of the aneurysm where the shear stress is high. As the flow rate increases, these WSS differences from the first two models become less prominent indicating insignificant differences in model behaviour at high shear rates.

Similar result are obtained when the diameter of the common carotid artery is assumed to be as large as 2.0 cm. The maximum wall shear stress and shear rates values are lower when compared to the 0.64 cm diameter artery but the differences in model behaviour are analogous.

It is evident from these results that the Power Law model tends to break down at higher shear rates in that it reduces the viscosity of the blood to levels below the Newtonian level which theoretically is not possible. This is clearly evident in the 55 % dilated model. The pressure difference predicted by this model is less than the Newtonian model, indicating a lower than Newtonian viscosity. This model also produces very low wall shear stress levels, dropping below Newtonian levels at fairly low shear rates, for example at the medium flow rate at 55% dilation, the WSS levels are less than Newtonian levels. This Power Law model is relatively easy to use but predict decreasing viscosity at higher strain, contrary to the generally accepted observation that blood behaves as a Newtonian fluid for strains above $100\ s^{-1}$.

At low shear rates the Casson model shows near Newtonian behavior. As the shear rate increases, this model begins deviating from the Newtonian case by producing higher WSS. This model takes the haematocrit factor *H* (the volume fraction of red blood cells in whole blood) into account, with the parameters given (obtained from data fitting) suggesting a value of *H* of 37%. However, it is reported that this yields a limiting viscosity at high shear slightly above the usual Newtonian value. The results obtained here suggest the same with WSS values above the Newtonian values at very high shear rates. This model appears to breakdown in the aneurysm at high shear rates, but is accurate at low shear rates.

The Carreau model generally produces values that are in close agreement with that of the Newtonian model at shear rates well above $100s^{-1}$. Our results do not indicate this to be the case. Both the WSS and the pressure difference deviate significantly from the Newtonian values at shear rates in excess of $1000s^{-1}$. This model by design reverts to Newtonian numbers as shear rates approach infinity. The basis for this model is the constant Newtonian viscosity, modified to non-Newtonian such that the modification tends to zero as the limit of the shear rate goes to infinity.

Finally, the Generalized Power Law model gave results that are in closest agreement with the Newtonian values at mid-range and high shear rates. At low shear rates, this model gives values that are close to that of the Power Law and the Carreau models. While the Power Law model breaks down at high shear, our results show a close agreement between the Generalized Power Law and the Newtonian models even at

high shear rates as shown in Figure 5. The Generalized Power Law model is widely accepted as a general model for non-Newtonian blood viscosity. It includes the Power Law model at low shear rate and the Newtonian model at mid-range and high shear rates. There is also good agreement between the Generalized Power Law and the Carreau model for low shear rates.

4. CONCLUSIONS

A study of the effects of modeling blood flow through an aneurysm using five different blood rheological models is presented. The flow field and wall shear stress distributions produced by each model are investigated for various flow rates and degrees of abnormality. In a specific dilated artery with a particular inlet velocity, the pattern of the WSS was found to be the same for all models studied. The only difference was in the magnitude of the WSS predicted by each model. These differences were significant at low shear rates and, in the case of the Power Law, the Carreau and the Casson model, at high shear rates. At mid-range to high shear rates, the Generalized Power Law and the Newtonian models produced almost identical results. The differences in magnitude of the WSS can be explained by the differences in the models themselves as discussed above.

The results show that there are significant differences between simulating blood as a Newtonian or non- Newtonian fluid at low shear rates. In particular, the Power Law model overestimates WSS at low shear rates and underestimates it at high shear. The Newtonian model under estimates WSS at low shear while the Generalized Power Law models provide a better approximation of WSS at low shear rates.

The distribution of shear rates in the dilated artery in Figure 7 shows a small region in the vicinity of the aneurysm where the shear rate is high. The shear rate in the rest of the artery is relatively low, indicating that non-Newtonian behavior is important. However, these simulations correspond to a heart rate of only 60 beats per minute and low inlet velocities (maximum of 0.225 m/s for the dilated artery). If the heart rate were increased to 100 beats per minute and/or the inlet velocities increased, the region (and periods in a cardiac cycle) over which the non-Newtonian behavior was important would decrease.

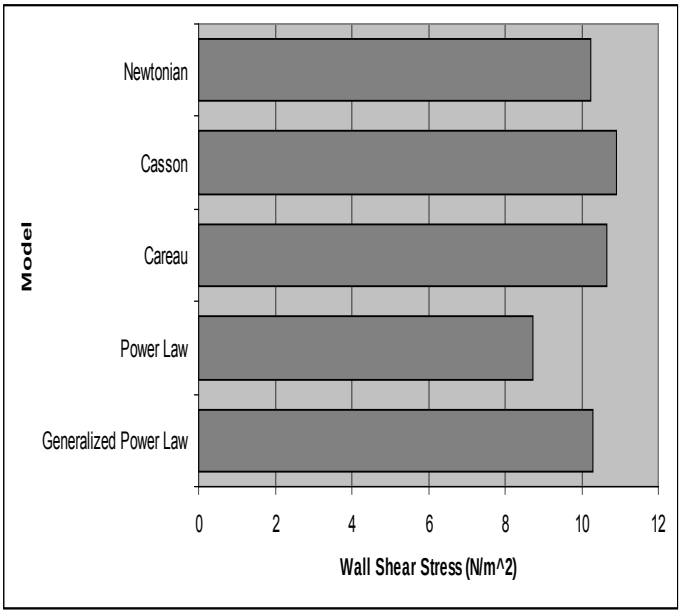


Fig. 8. Wall shear stress chart.

In a healthy left or right common carotid artery, the average maximum systolic velocity is 1.08 m/s [13]. In modeling the aneurysm, our simulation could only reach velocities as high as 0.375 m/s. We were successful in simulating a healthy artery at the maximum velocity and the results obtained from the various models are shown in Figure 8. The WSS values from all models are in good agreement except for the Power Law model. As in the dilated artery, the Generalized Power Law model gave the closest value to that of the Newtonian model.

In conclusion, in terms of the wall shear stress distribution, we found that the Newtonian model is a good approximation in regions of mid-range to high shear but the Generalized Power Law model provides a better approximation of wall shear stress at low shear. Whether the fact that the Newtonian model underestimates the WSS in regions of low shear is biologically significant is open to debate. A prudent approach would be to use the Generalized Power Law model since it predicts WSS better than the Newtonian model for low inlet velocities and regions of low shear and is effectively Newtonian at midrange to high shear.

These conclusions are presented under the assumption that the arterial walls are rigid and zero pressure is assumed at the outlet. A more realistic simulation would include elastic walls and incorporate the effects of upstream and downstream parts of the circulatory system into the boundary conditions. Simulations incorporating elastic walls are currently in progress.

References

[1] D.N. Ku, Blood flow in arteries, Annual Review of Fluid Mechanics (1997), Vol. 29 pp. 399-434.

[2] A. Valencia, F. Solis (2006), Blood flow dynamics and arterial wall interaction in a saccular aneurysm model of the basilar artery, Comp. Struc., Vol. 84 , pp.1326-1337.

[3] Oshima *et al.* (2001), Finite element simulation of blood flow in the cerebral artery, Comp. Meth. App. Mech. Eng., Vol. 191, pp. 661 – 671.

[4] B.V. Rathish Kumar and K.B. Naidu (1996), Hemodynamics in aneurysm, Comp. Biomed. Res., Vol. 2 , pp. 119 – 139.

[5] P. Neofytou, S. Tsangaris (2005), Flow effects of blood constitutive equations in 3D models of vascular anomalies, Int. J. Numer. Meth. Fluids, Vol. 51, pp. 489-510.

[6] T.J. Pedley (1980), The fluid mechanics of large blood vessels, Cambridge University Press Cambridge.

[7] F.J. Walburn, D.J. Schneck (1976), A constitutive equation for whole human blood, Biorheology Vol.13, pp. 201- 210.

[8] Y.I. Cho, K.R. Kensey (1991), Effects of the non-Newtonian viscosity of blood on flows in a diseased arterial vessel, Part I: steady flows, Biorheology, Vol. 28, pp. 241-262.

[9] Y.C. Fung (1993), Biomechanics: Mechanical properties of living tissue, 2nd Edition, Springer Berlin..

[10] P.D. Ballyk, D.A. Steinman, C.R. Ethier (1994), Simulations of non-Newtonian blood flow in an end-to-end anastomosis, Biorheology, Vol. 31 (5), pp. 565-586.

[11] B.M. Johnston et al (2004)., Non-Newtonian blood flow in human right coronary arteries: steady state simulations, J. Biomechanics, Vol. 37, pp. 709-720.

[12] D.W. Holdsworth et al. (1998), Characterization of common carotid artery blood-flow waveforms in normal human subjects, Physiol. Meas., Vol. 20, pp. 219-240.

Permutation codes: a branch and bound approach

Roberto Montemanni, János Barta and Derek H. Smith

***Abstract*—This paper presents a new approach for retrieving largest possible permutation codes. These have applications in error correction for telecommunication purposes. The method presented is based on combinatorial optimization concepts such as branch and bound techniques and incorporates new ad-hoc theoretical results. It is shown how the method can be applied to obtain new results for subproblems. These results for subproblems can be combined with other theoretical results to obtain new results for complete instances. It is shown how the new improved upper bound $M(7,5) \leq 124$ can be obtained with such techniques.**

***Index Terms*—Permutation codes; Branch and bound algorithms; Upper bounds.**

I. Introduction

THIS paper considers the application of branch and bound techniques to the construction of permutation codes. Permutation codes (sometimes referred to as permutation array) have received considerable attention in the literature [1], [2], [3], [4], [5], [6], [7]. This has been motivated by an application to powerline communications when M-ary Frequency-Shift Keying (FSK) modulation is used [3], [8], [9], [10], [11]. In this application permutations are used to ensure that power output remains as constant as possible while combatting impulsive noise permanent, narrow band noise from electrical equipment or magnetic fields, as well as the more common white Gaussian noise.

A permutation code is a set of permutations in the symmetric group $\mathscr{S}_n$ of all permutations on n elements. The codewords are the permutations and the code length is n. The error-correcting capability of a permutation code is related to the *minimum Hamming distance* of the code. The Hamming distance δ between two codewords is the number of elements that differ in the two permutations. Alternatively, two permutations σ_1 and σ_2 are at distance δ if $\sigma_1\sigma_2^{-1}$ has exactly $n-\delta$ fixed points. The minimum distance d is then the minimum δ taken over all pairs of distinct permutations. Such a code is then denoted an (n,d) permutation code.

Redundancy in an encoding is minimized if the number of codewords is as large as possible. Thus if $M(n,d)$ denotes the maximum number of codewords in an (n,d) permutation code it is important to determine $M(n,d)$, or if this is not possible to find good lower and upper bounds. The most complete contributions to lower bounds can be found is in [3], [12]. Recently, some improvements based on similar search techniques have been presented in [13], while in [14] a study on the structure of optimal codes has been presented. Results on a similar problem using a metric different from the one treated in this paper can be found in [15].

R. Montemanni and J. Barta are with the Dalle Molle Institute for Artificial Intelligence, University of Applied Sciences of Southern Switzerland, Galleria 2, 6928 Manno, Switzerland. Emails: {*roberto.montemanni, janos.barta*}*@supsi.ch*.

D.H. Smith is with the Division of Mathematics and Statistics, University of South Wales, Pontypridd, CF37 1DL, Wales, UK. Email: *derek.smith@southwales.ac.uk*.

$$\begin{aligned}\Gamma = \{&[012345],[021453],[034512],\\ &[045231],[102534],[130425],\\ &[153240],[205143],[243015],\\ &[251304],[310254],[324105],\\ &[341520],[425310],[432051],\\ &[450132],[504321],[513402]\}.\end{aligned}$$

Fig. 1. An optimal (6, 5) code.

II. Problem description

A permutation of the n-tuple $x_0 = [0,1,...,n-1] \in \mathbf{N}^n$ is a *codeword* of length n and the set of all codewords of length n is denoted by Ω_n. From an algebraic point of view the set Ω_n is the single orbit of the symmetric group of permutations $\mathscr{S}_n$, i.e.

$$\Omega_n = \{x \in \mathbf{N}^n | x = gx_0,\ g \in \mathscr{S}_n\}$$

Any subset Γ of Ω_n is a *permutation code* of length n. The main problem can now be stated as:

Definition 1. *Given a code length n and a Hamming distance d, the maximum permutation code problem (MPCP) consists of the determination of a code $\Gamma \subseteq \Omega_n$ with minimum distance d and the maximum possible number of codewords.*

Example 1. *The problem $(6,5)$ is to determine a maximal code of length $n=6$ with minimum distance $d=5$. As reported in [14], [3], [12] the optimal solution of this problem is $M(6,5)=18$. One of the many possible optimal $(6,5)$ codes is shown in Figure 1.*

III. A branch and bound algorithm

Branch and bound approaches work by building a search-tree that covers all possible assignments of permutations to solutions, but with most of the branches of the tree pruned by inferring information from lower and upper bounds. The main elements of the algorithm proposed here are described in this section.

A. Structure of the search-tree node

From now on, the set of nodes of the search tree will be denoted S, and the subtree of the search-tree rooted at node t will be denoted as $SubT(t)$. Each node t of the search-tree is then identified by the following elements:

ISBN: 978-1-61804-240-8

- $in(t)$: a list of permutations that are forced in the solutions associated with the search-tree nodes of $SubT(t)$;
- $feas(t)$: a list of permutations that are feasible according to the list of forced permutations $in(t)$, and to reduction rules (see Section III-D2);
- $lb(t)$: a lower bound for the number of permutations in the optimal solutions associated with the search-tree nodes in $SubT(t)$. The calculation of the lower bound is described in Section III-F.
- $ub(t)$: an upper bound for the number of permutations in the optimal solutions associated with the search-tree nodes in $SubT(t)$. The calculation of the upper bound is described in Section III-E.

From these four items all the information required by dominance rules and pruning can be derived.

B. Initialization and branching strategy

Initial lower and upper bounds $BestLB$ and $BestUB$ are provided as input to the algorithm (they can be 0 and $+\infty$, respectively). These initial values will be updated during the execution of the algorithm in case improved values are obtained. A permutation p is selected and the root r of the search-tree is the node initialized with $in(r) = \{p\}$, $lb(r) = BestLB$, $ub(r) = BestUB$ and $feas(r) = \{i \in \Omega_n : \delta(i,p) \geq d\}$. Initially, r will be the only node contained in S ($S := \{r\}$), the set of the nodes to be examined, referred to as *open nodes* in the remainder of the paper. Due to the symmetry of the problem, the first permutation included in $in(r)$ can be chosen arbitrarily. The set of *closed nodes* C is initialized as empty ($C := \emptyset$). This set will contain nodes already examined by the algorithm, and will be used by pruning and reduction techniques described in Section III-D.

At each iteration of the branch and bound algorithm, an open node t from the set S is expanded (see Section III-C for more details about the strategy used to select node t, and the rationale behind it), which means that node t is expanded by decomposing it into the associated subproblems. One new search-tree node t_p is created for each permutation p of $feas(t)$ in such a way that $in(t_p) = in(t) \cup \{p\}$ and the new set $feas(t_p)$ is determined, also taking into account the reduction and pruning rules described in Section III-D. Sets S and C are finally updated: $S = S \backslash \{t\}$, $C = C \cup \{t\}$. For each new node t_p the values of $lb(t_p)$ and $ub(t_p)$ are calculated, as described in Sections III-F and III-E respectively. In the case that the pruning test is positive (see Section III-D) the new node t_p is pruned and $C := C \cup \{t_p\}$, otherwise the set S of open node is incremented: $S := S \cup \{t_p\}$.

In case $\min_{t \in S} ub(t) \leq BestUB$, the global upper bound of the residual open problems has been improved, and the updating $BestUB := \min_{t \in S} ub(t)$ can take place. Also the value of $BestLB$ can be updated in case a new incumbent heuristic solution is found (in general $BestLB := \max_{t \in S} lb(t)$). All the open nodes u of the search tree are examined and pruned in case $ub(u) \leq BestLB$, since no improving solution can exist in the search-tree node rooted in u. In such a case $S := S \backslash \{u\}$ and $S := S \cup \{u\}$.

The branch and bound algorithm stops when the set S is empty (all the search-tree nodes have been processed or labelled as dominated).

C. Selection of the node to expand

Nodes are expanded in the same order they have previously been created. This strategy allows the best exploitation of the reduction and pruning rules described in Section III-D.

D. Reduction and pruning rules

This section discusses some rules useful to prune dominated search-tree nodes and to reduce the size of $feas(t)$ while generating a new search-tree node t. These results are based on the concept of *isomorphism* for graphs [16].

1) Pruning rule: Two graphs G and H are said to be *isomorphic* if a bijection between the vertex sets of G and H $f : V(G) \to V(H)$ exists, such that any two vertices u and v of G are adjacent in G if and only if $f(u)$ and $f(v)$ are adjacent in H.

Definition 2. *The* graph induced *by search-tree node t is defined as $G_t^I = \{V_t^I, E_t^I\}$, with $V_t^I = V \backslash in(t)$ and $E_t^I = \big\{\{i,j\} | i,j \in V_t^I, \delta(i,j) \geq d\big)\big\}$.*

Remark 1. *The graph induced by search-tree node t is considered instead of the subgraph of G with vertices set $feas(t)$ because $feas(t)$ might have already have benefited from reduction rules in previous iterations, and therefore isomorphisms could be more difficult to identify.*

Definition 3. *If the graph G_t^I induced by search-tree node t is isomorphic to the graph G_u^I induced by another search-tree node u, it will be said (for short) that node t is isomorphic to node u and written $t \cong u$.*

The following result allows one of two isomorphic nodes to be pruned from the branch and bound tree.

Theorem 1. *If a new search-tree node t is such that $t \cong u$ with $u \in S \cup C$, $|in(t)| = |in(u)|$ then the node t can be classified as dominated and moved to set C ($S = S \backslash \{t\}$ and $C = C \cup \{t\}$).*

Proof: The search-tree subtree associated with node t will provide an optimal solution with the same number of permutations of that of node u, that has been already expanded ($u \in C$), or is scheduled to be expanded ($u \in S$). ■

2) Reduction rule: During the branching of a search-tree node t, all potential new search-tree nodes obtained by expanding the set $feas(t)$ with each possible permutation will be considered.

Proposition 1. *While creating a new search-tree node u obtained by adding $p_u \in feas(t)$ into $in(u)$, permutation $p_k \in feas(u)$, with $k \cong v$, $v \in S \cup C$, $|in(k)| = |in(v)|$ (nodes already expanded at the same level) can be taken out of $feas(u)$: $feas(u) = feas(u) \backslash \{p_k\}$*

Proof: The best possible solution including permutations p_k and the permutations of $in(u)$ is equivalent to that of the problem v, that has been already expanded ($u \in C$), or is scheduled to be expanded ($u \in S$). Therefore all solutions

including permutation p_k are not of interest while solving the problem associated with $feas(u)$. ■

Remark 2. *When applying the reduction rule described in Proposition 1 it is necessary to expand search-tree nodes according to the basic strategy described in Section III-C, in order to avoid situations where the permutations isomorphic to that associated with node u_i are taken out from $feas(v_i)$ as a result of Proposition 1 at level i of the search-tree, but then the permutations associated with a descendant v_j of v_i are taken out of a descendant u_j of u_i at a level j, with $j > i$, of the tree. Such a situation would clearly lead to infeasible solutions since some regions of the search space are left unvisited by the search-tree.*

In the implementation described here the routines of *Nauty 2.5* described in [17] are used to identify graph isomorphisms.

E. Upper bound

The set of codewords Ω_n can be split into n subsets $W_0,...,W_{n-1}$, in such a way that for a fixed value $k \in \{0,...,n-1\}$ the subset W_i is defined as $W_i = \{x \in \Omega_n | x(k) = i\}$. In other words, the subset W_i contains all codewords with the k-th component having the value i. Since the partition is obtained by fixing the value of one component of the codewords, it is clear that the sets W_i are isomorphic to Ω_{n-1}. Furthermore, as the sets W_i form a partition of Ω_n it is well-known that an upper bound of $M(n,d)$ can be obtained by adding the upper bounds on the subsets W_i:

Theorem 2 (Deza and Vanstone [18])**.**

$$M(n,d) \leq n \cdot M(n-1,d) \tag{1}$$

The partitioning procedure described in Theorem 2 can be carried out on any subset of Ω_n. At each search-tree node t the algorithm generates a partition $T_0,...,T_{n-1}$ of the set $feas(t)$, such that $T_i = \{x \in feas(t) | x(k) = i\}$ and a partition $Q_0,...,Q_{n-1}$ of the set $in(t)$, such that $Q_i = \{x \in in(t) | x(k) = i\}$. For each subset T_i an upper bound $UB(T_i)$ is calculated using the Maximum Clique Problem (MCP) solver proposed in [19] (see also [20]), which is run for 10 seconds on every subproblem. The choice of this solver is motivated by its ability of proving optimality extremely fast on small/medium problems. From preliminary experiments the solver described in [19] is significantly faster than the one presented in [21] (see Section III-F) on the instances treated.

The new upper bound for the search tree node t can be expressed as follows.

Proposition 2.

$$\sum_{i=0}^{n-1} |Q_i| + \min\{UB(T_i); M(n-1,d) - |Q_i|\} \tag{2}$$

is a valid upper bound for the search-subtree rooted at the search-tree node t.

Proof: In case it can be shown that in a partition $\{T_i\}$ of $feas(t)$ the maximum clique has size smaller than $M(n-1,d)-|Q_i|$ a tighter upper bound for that partition is available. Combining the n partitions together, gives the global upper bound provided by (2). ■

Remark 3. *The result of Proposition 2 can be seen as a refinement of a previous result originally presented in [14].*

Remark 4. *As the index k of the fixed component in the codewords can be varied, there are n different partitions that can be generated. The algorithm computes for each partition an upper bound and finally chooses the lowest one.*

F. Lower bounds

The methods used to provide lower bounds for the optimal solution cost of the problem associated with the search-tree node t are based on MCP algorithms [21]. In detail, when examining node t, it is possible to associate a graph $G_t = \{V_t, E_t\}$ such that the set of vertices $V_t = feas(t)$ (where each vertex is associated with a permutation) and the set of edges $E_t = \{\{i,j\} | i,j \in V_t, \delta(i,j) \geq d)\}$. Notice that this graph is a subgraph of that induced by node t (see Section III-D). The problem is then equivalent to solving a MCP on graph G_t. This transformation usually has the side-effect that possible structure and information coming from group theory (and the possibility to exploit these) is lost. On the other hand, the very efficient methods developed for the MCP over the years can be used directly. In the remainder of this section it will be shown how to modify a general-purpose method to the current context, in order to insert permutation codes related concepts into its logic.

The original MCP method works as follows, according to the description provided in [21]. The algorithm considers the vertices of V_t in a given order $\{v_1, v_2, ..., v_{|V_t|}\}$. Initially, the method finds the largest clique C_1 that contains the vertex v_1. Then it finds C_2, the largest clique in $G_t \backslash \{v_1\}$ that contains v_2 and so on. Applying heuristics and pruning techniques, the search space can, however, be reduced dramatically. The notion of the *depth* is crucial for the algorithm. Suppose vertex v_1 is under investigation. At depth 2, all vertices adjacent to v_1 are considered. At depth 3, all vertices (that are already in depth 2) adjacent to the first vertex in depth 2 are considered and so on. When there are no more vertices left at a certain depth, a clique has been identified, and backtracking is activated. Pruning rules based on the number of vertices left at each level are triggered to have early backtracking. These rules are very quick to check but very effective. When the entire search-tree created by the method has been visited (or declared dominated) the computation is stopped and the largest clique has been retrieved.

Some context-dependent modifications of the general algorithm previously described were implemented. The modifications to the original method are introduced to anticipate the pruning in the branch and bound framework running internally in the MCP algorithm. In detail, during the execution a 2-dimension array *PosVal* is kept in memory, defined as follows.

Definition 4. *PosVal[i][j] contains at any time during the execution of the MCP algorithm [21] the current number of permutations with value j in position i that are still feasible*

according to the partial assignment under investigation, or are part of such a partial assignment.

Notice that the structure is dynamically updated during the execution of the algorithm, depending on the partial solution currently under investigation. The pruning is based on the following theoretical results.

Proposition 3. *When*

$$\sum_{j=1}^{n} \min\{M(n-1,d), PosVal[i][j]\} \leq BestLB \tag{3}$$

for some $i, 1 \leq i \leq n$ the partial solution under investigation will not lead to any improving result, it can be pruned and it is possible to backtrack.

Proof: It is known ([18]) that if a position i of the codewords of a code (n,d) is fixed to a value j then there cannot be more that $M(n-1,d)$ permutations with this property. This is valid for each value of i and j. During the execution of the MCP algorithm $PosVal[i][j]$ can become smaller than $M(n-1,d)$ for some i or j. In such a case the best upper bound for the corresponding subproblem is no longer $M(n-1,d)$, and the global upper bound is updated according to (3) by summing over all possible values j for position i. When the global upper bound is not greater than the cost of the best solution currently available ($BestLB$) there is no hope of finding an incumbent improved solution, and backtracking can be activated. ■

Proposition 4. *When*

$$\sum_{i=1}^{n} \min\{M(n-1,d), PosVal[i][j]\} \leq BestLB$$

for some $j, 1 \leq j \leq n$ the partial solution under investigation will not lead to any improving result, it can be pruned and it is possible to backtrack.

Proof: The proof is based on the same principles of that of Proposition 3 but now the sum is over all possible positions i for value j (by columns of $PosVal$ instead of by rows). ■

Remark 5. *The results of Propositions 3 and 4 can be seen as the adaptation of the upper bound of Proposition 2 to the context of the MCP algorithm, where it is preferable to have a less precise but much faster upper bound.*

In the context of the current permutation codes algorithm, the main target is to have the MCP algorithm complete the computation in order to prove optimality for the problem under investigation in the given time available (see Section IV). For this reason, to have pruning as early as possible, at each iteration one of the permutations with value j in position i is expanded such that

$$PosVal[i][j] = \min_{1 \leq k \leq n, 1 \leq l \leq n} PosVal[k][l] \tag{4}$$

The strategy described in (4) allows the anticipation of the application of Propositions 3 and 4 as much as possible. Different strategies might however be used in order to make it more likely to have heuristic solutions instead of early backtracking, if the algorithm is used in a different perspective.

The modified version of the algorithm originally proposed in [21] described previously is executed each time a new non-dominated search-tree node (of the external branch and bound method) is generated, and it is run for a maximum of 720 seconds.

IV. Experimental results

The approach discussed in Section III can be applied to subproblems to prevent ths size of the search tree increasing too much for large instances. Results on subproblems can then be propagated to full problems in order to obtain new theoretical results, as will be shown later in this section. All the experiments presented have been carried out on a computer equipped with a 2.3GHz AMD Quad Opteron processor (only one core has been used at a time), while the algorithm has been coded in ANSI C.

In the study presented here two subproblems of $(7,5)$ are considered, where a position of the permutations is restricted to values from a given set $F \subset \{0,1,...,n-1\}$. The largest possible code fulfilling the requirements with all permutations with the given position strictly from F is sought. Notice that either the position fixed or the values of F are not important due to the symmetry of the problem. The cardinality of F is the only important factor. Therefore subproblems can be defined as follows.

Definition 5. *Refer to the problem $(n,d)|_{|F|}$ as the subproblem of (n,d) where a position of the code is restricted to values from set F.*

Such a problem with $|F| = 1$ is equivalent to the problem $(6,5)$ since the largest possible set with a common symbol in the first position is sought. According to [18] this amounts to looking for the largest possible code of length $n-1$. More interesting are the cases when $2 \leq |F| \leq n-1$. Notice that for these problems a trivial upper bound, coming from the generalization of that described in (1), is the following one:

$$M(n,d)|_{|F|} \leq |F| \cdot M(n-1,d) \tag{5}$$

For $|F| = 2$ equation (5) returns an upper bound of 36. The branch and bound described in Section III has been able to retrieve a solution with 36 permutations, matching the upper bound. This result is not as obvious as it could appear, since traditional methods based on maximum clique solvers (e.g. [19], [21]) are not able to retrieve such a solution in a week of computation, while the method described here was able to close the problem in 15 382 seconds. The optimal solution retrieved is presented in Figure 2. It might turn out to be useful for future studies by other researcher, since inspecting such a solution might bring new insights about the general characteristics of solutions.

For $|F| = 3$ the branch and bound described in Section III has been used to obtain a new upper bound, which improves the one given by (5). The algorithm itself is not able to find significantly large heuristic solutions (and consequently lower bounds), but if it is run with a hypothetical initial lower bound of $BestLB = 53$, the algorithm is able to prove in 922 450 seconds that no solution with 54 permutations exists, leading to the following new result.

$$\Gamma = \{[0123456],[0134562],[0162345],$$
$$[0214635],[0236154],[0265413],$$
$$[0342516],[0351642],[0364251],$$
$$[0415326],[0426531],[0452163],$$
$$[0516243],[0521364],[0543621],$$
$$[0631425],[0645132],[0653214],$$
$$[1025634],[1043265],[1056423],$$
$$[1203546],[1240653],[1254360],$$
$$[1305462],[1326045],[1360524],$$
$$[1432605],[1450236],[1463052],$$
$$[1534026],[1546302],[1562430],$$
$$[1602354],[1624503],[1635240]\}.$$

Fig. 2. An optimal $(7, 5)|_2$ code.

Proposition 5.

$$M(7,5)|_3 \leq 53$$

The result of Proposition 5 has a remarkable implication: it allows us to improve the lower bound for the general problem $(7,5)$.

Proposition 6.

$$M(7,5) \leq 124$$

Proof: Partition the code in three parts, each part covering some possible values for a given position. The first and the second parts cover 3 possible distinct values each, while the third part covers the remaining value. By combining the results on subproblems the new global upper bound is obtained: $M(7,5) \leq M(7,5)|_3 \cdot 2 + M(7,5)|_1 \leq 53 \cdot 2 + 18 = 124$ ■

The result of Proposition 6 improves the previously best known upper bound of 126 that can be obtained with equation (5). It is interesting to mention that the best result reported in the literature so far was 140 instead of 126 (see, for example, [12]).

The results previously presented show that a novel approach like the one proposed has potential. Further refinement to the current branch and bound method could lead to new improvements.

V. Conclusions

A novel approach based on combinatorial optimization and branch and bound has been proposed to attack permutation codes. It is based on some ad-hoc new theoretical results that make the technique practical for real problems. Experimental results have been presented, aiming at clarifying the current potential of the method. New results on subproblems of permutation code instances have been described, and it has been shown how such results can be used to derive new theoretical results (upper bounds in this case) for full permutation code instances.

The technique appears to be capable of further improvements to handle larger problem instances. New theoretical results could be used to speed up computation and to make pruning even more effective. There is also a clear potential for such techniques to be adapted to other types of codes (e.g. binary codes). Such adaptions represent an important area for future work.

References

[1] I.F. Blake, *Permutation codes for discrete channels*, IEEE Transactions on Informormation Theory 20(1) 138–140, 1974.

[2] M. Bogaerts, *New upper bounds for the size of permutation codes via linear programming*, The Electronic Journal of Combinatorics 17(#R135), 2010.

[3] W. Chu, C.J. Colbourn and P. Dukes, *Constructions for permutation codes in powerline communications*, Designs, Codes and Cryptography 32, 51–64, 2004.

[4] P. Dukes and N. Sawchuck, *Bounds on permutation codes of distance four*, Journal of Algebraic Combinatorics 31 143–158, 2010.

[5] P. Frankl and M. Deza, *On maximal numbers of permutations with given maximal or minimal distance*, Journal of Combinatorial Theory Series A 22, 352–260, 1977.

[6] I. Janiszczak and R. Staszewski, *An improved bound for permutation arrays of length 10*, http://www.iem.uni-due.de/preprints/IJRS.pdf (downloaded 17^{th} March 2011).

[7] H. Tarnanen, *Upper bounds on permutation codes via linear programming*, European Journal of Combinatorics 20 101–114, 1999.

[8] C.J. Colbourn, T. Kløve and A.C.H. Ling, *Permutation arrays for powerline communication and mutually orthogonal latin squares*, IEEE Transactions on Information Theory 50, 1289–1291, 2004.

[9] A.J. Han Vinck, *Coded modulation for power line communications*, A.E.Ü. International Journal Electronics and Communications 54(1), 45–49, 2000.

[10] S. Huczynska, *Powerline communications and the 36 officers problem*, Philosophical Transactions of the Royal Socicety A. 364, 3199–3214, 2006.

[11] N. Pavlidou, A.J. Han Vinck, J. Yazdani and B. Honary, *Power line communications: state of the art and future trends*, IEEE Communications Magazine 41(4), 34–40, 2003.

[12] D.H. Smith and R. Montemanni, *A new table of permutation codes*, Design, Codes and Cryptography 63(2), 241–253, 2012.

[13] I. Janiszczak, W. Lempken, P.R.J. Östergård and R. Staszewski, *Permutation codes invariant under isometries*, Designs, Codes and Cryptography, to appear.

[14] J. Barta, R. Montemanni and D.H. Smith, *A branch and bound approach to permutation codes*, Proceedings of the IEEE 2^{nd} International Conference of Information and Communication Technology - ICOICT, 187–192, 2014.

[15] D.H. Smith and R. Montemanni, *Permutation codes with specified packing radius*, Design, Codes and Cryptography 69, 95–106, 2013.

[16] B.D. McKay, *Practical graph isomorphism*, Congressum Numerantium. 30 45–87, 1981.

[17] B.D. McKay and A. Piperno, *Practical graph isomorphism, II*, Journal of Symbolic Computation 60 94–112, 2014.

[18] M. Deza and S.A. Vanstone, *Bounds for permutation arrays*, Journal of Statistical Planning and Inference 2 197–209, 1978.

[19] P.R.J. Östergård, *A fast algorithm for the maximum clique problem*, Discrete Applied Mathematics 120, 197–207, 2002.

[20] P.R.J. Östergård, *A new algorithm for the maximum-weight clique problem*, Nordic Journal of Computing 8(4), 424–436, 2001.

[21] R. Carraghan and P.M. Pardalos, *An exact algorithm for the maximum clique problem*, Operations Research Letters, 9 375–382, 1990.

Unary operators

József Dombi

(Invited Paper)

***Abstract*—Modal operators play an important role in fuzzy theory, and in recent years researchers have devoted more effort on this topic. In our study, we will construct modal operators. We give a common form of all types of unary operators. We will characterise them using differential equations, Bayesian inference, and a functional equation. We will show that a special class of the kappa function is related to the sigmoid function, it and can be characterised by odds.**

***Keywords*—negation, modalities, hedges, sharpness operator, Pliant system**

I. Introduction

In logic theory, modal operators have a variety of applications and even from a theoretical perspective they are interesting to study. Here, we will present different approaches for obtaining the form of the necessity and possibility operators. These have a simple parametrical form. By changing the parameter value, we get different modalities. Cintula et al [1] made a study on fuzzy logic with an additive involutive negation operator. In Hájek's paper [2], the basic logic (BL) was defined. In a recent paper [3], we can find a survey paper that discusses the state-of-art of BL.

In standard modal systems, the basic modalities are called necessity (denoted by $\Box$) and possibility (denoted by $\diamondsuit$). They satisfy basic properties such as their interdefinability via negation $\Box p = \neg\diamondsuit\neg p$, and distributivity of $\Box$ over conjunction $\Box(p \wedge q) = \Box p \wedge \Box q$ and distributivity of $\diamondsuit$ over disjunction $\diamondsuit(p \wedge q) = \diamondsuit p \wedge \diamondsuit q$. Now consider a De Morgan operator system augmented with unary operators that represent the distributive modalities.

II. Negation

Definition 1: We say that $\eta(x)$ is a negation if $\eta\colon [0,1] \to [0,1]$ satisfies the following conditions:

C1:	$\eta : [0,1] \to [0,1]$ is continuous	(Continuity)
C2:	$\eta(0) = 1$, $\eta(1) = 0$	(Boundary conditions)
C3:	$\eta(x) < \eta(y)$ for $x > y$	(Monotonicity)
C4:	$\eta(\eta(x)) = x$	(Involution)

From C1, C2 and C3, it follows that there exists a fix point $\nu_* \in [0,1]$ of the negation where

$$\eta(\nu_*) = \nu_* \tag{1}$$

So another possible characterisation of negation is when we assign a so-called decision value ν for a given ν_0, i.e. a point (ν, ν_0) can be specified such that the curve must intersect. This tells us something about how strong the negation operator is.

$$\eta(\nu) = \nu_0 \tag{2}$$

If $\eta(x)$ has a fix point ν_*, we use the notation $\eta_{\nu_*}(x)$ and if the decision value is ν, then we use the notation $\eta_\nu(x)$. If $\eta(x)$ is employed without a suffix, then the parameter has no importance in the proofs. Later on we will characterise the negation operator in terms of the ν_*, ν_0 and ν parameters.

For the strong negation operator case, two representation theorems are known. Trillas [4] has shown that every involutive negation operator has the following form

$$\eta(x) = f^{-1}(1 - f(x)), \tag{3}$$

where $f : [0,1] \to [0,1]$ is a continuous strictly increasing (or decreasing) function. This generator function corresponds to the nilpotent operators (nilpontent t-norms). For the strictly monotonously increasing t-norms, another form of the negation operator is given in [5]:

$$\eta(x) = f^{-1}\left(\frac{1}{f(x)}\right),$$

where $f : [0,1] \to [0,\infty]$ is a continuous, increasing (or decreasing) function and f is the generator function of the strict monotone t-norm or t-conorm.

We can express these negation operators in terms of their neutral values and we get a new form of the negation operator.

For the strict monotone operators,

$$\eta_{\nu_*}(x) = f^{-1}\left(\frac{f^2(\nu_*)}{f(x)}\right) \tag{4}$$

The other form of the negation operator in terms of ν_0, and ν, and corresponding to (II), is

$$\eta_\nu(x) = f^{-1}\left(f(\nu_0)\frac{f(\nu)}{f(x)}\right) \tag{5}$$

In the following we will use (4) and (5) to represent the negation operator because here we are just considering strict monotone operators and we sketch the shape of the negation function.

Definition 2: If $\nu_1 < \nu_2$, then $\eta_{\nu_1}(x)$ is a stricter negation than $\eta_{\nu_2}(x)$.

Definition 3 (Drastic negation): We call $\eta_1(x)$ and $\eta_2(x)$ a drastic negation when

$$\eta_1(x) = \begin{cases} 1 & \text{if} \quad x \neq 1 \\ 0 & \text{if} \quad x = 1 \end{cases} \qquad \eta_0(x) = \begin{cases} 1 & \text{if} \quad x = 0 \\ 0 & \text{if} \quad x \neq 0 \end{cases}$$

J. Dombi is with the Institute of Informatics, University of Szeged, Hungary e-mail: dombi@inf.u-szeged.hu

Manuscript received June 19, 2014; revised July 04, 2014.

Theorem 4: The negation operators

$$\eta_\nu(x) = f^{-1}\left(f(\nu_0)\frac{f(\nu)}{f(x)}\right),$$
$$\eta_{\nu_*}(x) = f^{-1}\left(\frac{f^2(\nu_*)}{f(x)}\right)$$

have the following properties:

a. They are continuous.
b. They are strictly monotonous decreasing.
c. The correspondence principle is valid:

$$\eta_\nu(0)=1, \qquad \eta_\nu(1)=0, \qquad \eta_{\nu_*}(0)=1, \quad \eta_{\nu_*}(1)=0$$

d) The involutive property is valid:

$$\eta_\nu(\eta_\nu(x))=x, \quad \eta_{\nu_*}(\eta_{\nu_*}(x))=x.$$

e) The neutral value property holds:

$$\eta_\nu(\nu)=\nu_0, \quad \eta_{\nu_*}(\nu_*)=\nu_*.$$

Proof: Using the representation form of the negation operator, this is trivial.

■

In fuzzy theory, we utilise two types of negation operator. These are

$$\textbf{Yager:} \quad \eta_m(x) = \sqrt[m]{1-x^m} \tag{6}$$

and

$$\textbf{Hamacher, Sugeno:} \quad \eta_a(x) = \frac{1-x}{1+ax} \tag{7}$$

We can express the parameters of the negation operator in terms of its neutral values $n(\nu_*)=\nu_*$. So we have

$$\nu_* = \eta(\nu_*) = \sqrt[m]{1-\nu_*^m} \text{ and } m = -\frac{\ln(2)}{\ln(\nu_*)}$$

Then the Yager negation operator has the form

$$\eta_{\nu_*}(x) = \left(1-x^{-\frac{\ln 2}{ln\nu_*}}\right)^{-\frac{\ln\nu_*}{ln2}} \tag{8}$$

In a similar way, for the Hamacher negation operator,

$$\eta_\nu(x) = \frac{1}{1+\frac{1-\nu_0}{\nu_0}\frac{1-\nu}{\nu}\frac{x}{1-x}}, \quad \eta_{\nu_*}(x) = \frac{1}{1+(\frac{1-\nu_*}{\nu_*})^2\frac{x}{1-x}} \tag{9}$$

This form of the negation operator can be found in [6].

Definition 5: A negation $\eta_{\nu_1}(x)$ is stricter than $\eta_{\nu_2}(x)$, if $\nu_1 < \nu_2$.

III. Modalities induced by two negation operators

The linguistic hedge "very" always expresses a tight interval, whereas "more or less" expresses a looser interval (less tight). In this sense, "very" corresponds to the necessity operator and "'more or less"' the possibility operator.
With this starting point, the necessity and possibility operators used in fuzzy logic are based on an extension of modal logic to the continuous case. We begin with the negation operator and we make use of two types of this operator; one that is strict, and one that is less strict. We will show that with these two negation operators we can define the modal hedges.

Modal logic, which is an area of mathematical logic, can be viewed as a logical system obtained by adding logical symbols and inference rules.

We will construct linguistic modal hedges called necessity and possibility hedges. The construction is based on the fact that modal operators can be realised by combining two kinds of negation operators.

In intuitionistic logic, another kind of negation operator also has to be taken into account. Here $\sim_x$ means the negated value of x. $\sim_1 x$ and $\sim_2 x$ are two negation operators.

In modal logic, $\sim_1 x$ means "x" is impossible. In other words, $\sim_1$ is a stronger negation than not "x", i.e. $\sim_2 x$. Because $\sim_1 x$ in modal logic, it means "x is impossible".
We can write

$$impossible\ x = necessity(not\ x)$$

$$\sim_1 x := impossible\ x$$
$$\sim_2 x := not\ x$$

$$\sim_1 x = \Box \sim_2 x$$

We will show that both operators belong to the same class of unary operators, and also show that because they have a common form in the Pliant system, we will denote both of them by $\tau_\nu(x)$. Depending on the ν value, we get the necessity hedge or the possibility hedge.

As we mentioned above, in modal logic we have two more operators than in the classical logic case, namely necessity and possibility; and in modal logic there are two basic identities. These are

$$\sim_1 x = impossible(x) = necessity(not(x)) = \Box \sim_2 x \tag{10}$$
$$\Diamond x = possible(x) = not(impossible(x)) = \sim_2 (\sim_1 x) \tag{11}$$

In our context, we model $impossible(x)$ with a stricter negation operator than $not(x)$. Eq.(11) serves as a definition of the possibility operator.

If in Eq.(10) we replace x by $\sim_2 x$ and using the fact that $\sim_2 x$ is involutive, we get

$$\Box x = \sim_1 (\sim_2 x),$$

and with Eq.(11), we have

$$\Diamond x = \sim_2 (\sim_1 x).$$

Definition 6: The general form of the modal operators is

$$\tau_{\nu_1,\nu_2}(x) = \eta_{\nu_1}\left(\eta_{\nu_2}(x)\right),$$

where ν_1 and ν_2 are neutral values. If $\nu_1 < \nu_2$, then $\tau_{\nu_1,\nu_2}(x)$ is a necessity operator and if $\nu_2 < \nu_1$, then $\tau_{\nu_1,\nu_2}(x)$ is a possibility operator.
From the above definition, we get

$$\tau_{\nu_1,\nu_2}(x) = f^{-1}\left(f(\nu_1)\frac{f(x)}{f(\nu_2)}\right),$$

This can be rewritten as

$$\tau_{\nu,\nu_0}(x) = f^{-1}\left(f(\nu_0)\frac{f(x)}{f(\nu)}\right).$$

Dombi operator case:

$$f_c(x) = \left(\frac{1-x}{x}\right)^{\alpha} \qquad \text{and} \qquad f_d(x) = \left(\frac{x}{1-x}\right)^{\alpha}$$

$$\tau_{\nu_c}(x) = \frac{1}{1+\frac{1-\nu_0}{\nu_0}\left(\frac{\nu_c}{1-\nu_c}\frac{1-x}{x}\right)}$$

$$\tau_{\nu_d}(x) = \frac{1}{1+\frac{1-\nu_0}{\nu_0}\left(\frac{\nu_d}{1-\nu_d}\frac{1-x}{x}\right)}$$

IV. Bayesian inference and modalities

Theorem 7: Here, we show that Bayes' theorem can be reformulated so we have the modal operator of the Pliant system.

Proof: Bayes' theorem modifies a probability value, given new evidence, in the following way

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

$P(E|H)$ is called the conditional probability. It is also known as a likelihood function when it is regarded as a function of H for fixed E ($P(E|H_x)$). $P(H)$ is called an a priori probability, $P(E)$ is called a marginal probability and $P(H|E)$ is called an a posterior probability. Because $P(E)$ can be rewritten as

$$P(E) = P(E|H)P(H) + P(E|\bar{H})P(\bar{H}),$$

where $\bar{H}$ is the complementer event ($H \cup \bar{H} = X$ ($H, \bar{H}$ span over all possibilities) and $H \cap \bar{H} = \varnothing$). We can employ the identity $P(\bar{H}) = 1 - P(H)$. Hence, we can rewrite Bayes' formula as

$$\begin{aligned} P(H|E) &= \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\bar{H})P(\bar{H})} \\ &= \frac{1}{1+\frac{P(E|\bar{H})}{P(E|H)}\frac{1-P(H)}{P(H)}} \\ &= \frac{1}{1+\frac{\nu}{1-\nu}\frac{1-x}{x}} = \tau_\nu(x) \end{aligned} \tag{12}$$

where $P(H) = x$, $\nu = \frac{1}{1+\frac{P(E|H)}{P(E|\bar{H})}}$ and $\tau_\nu(x)$ is modal operator in the Pliant system. The ratio of two likelihood functions is called the likelihood ratio. So

$$\wedge_E = \frac{P(E|H)}{P(E|\bar{H})}$$

and we get

$$\nu = \frac{1}{1+\wedge_E} \quad \text{or} \quad P(H|E) = \frac{1}{1+\frac{1}{\wedge_E}\frac{1-P(H)}{P(H)}}$$

$P(H|E)$ can be expressed in terms of $\wedge_E$. From Eq. 12, we get

$$\frac{P(H|E)}{1-P(H|E)} = \wedge_E \frac{P(H)}{1-P(H)} \quad \text{or} \quad odd(H|E) = \wedge_E odd(H), \tag{13}$$

where $odd(H)$ and $odd(H|E)$ are called the odds of the a priori probability and a posteriori probability. That is, the posteriori probability is a linear function of the odds of prior probability.

With two pieces of evidence E_1 and E_2 that are marginally and conditionally independent of each other, successively applying Bayes' theorem yields

$$P(H|E_1 \wedge E_2) = \frac{P(E_1|H)P(E_2|H)P(H)}{P(E_1|H)P(E_2|H)P(H)+P(E_1|\bar{H})P(E_2|\bar{H})P(\bar{H})}$$

Thus, we get

$$\begin{aligned} P(H|E_1 \wedge E_2) &= \frac{1}{1+\frac{1}{\wedge_1\wedge_2}\frac{1-P(H)}{P(H)}} \\ &= \frac{1}{1+\frac{\nu}{1-\nu}\frac{1-P(H)}{P(H)}} = \tau_\nu(P(H)) \end{aligned}$$

So we get again the modal operator of the Pliant system, where

$$\wedge_1 = \frac{P(E_1|H)}{P(E_1|\bar{H})}, \quad \wedge_2 = \frac{P(E_2|H)}{P(E_2|\bar{H})}, \nu = \frac{1}{1+\frac{1}{\wedge_1\wedge_2}}$$

More generally, we can infer that

$$odd\left(H|\bigcap_{i=1}^{n} E_i\right) = odd(H)\prod_{i=1}^{n}\wedge_i, \quad \text{where} \quad \wedge_i = \frac{P(E_i|H)}{P(E_i|\bar{H})}$$

and $\prod_{i=1}^{n}\wedge_i = \wedge$ is called Bayes' factor

$$odd\left(H|\bigcap_{i=1}^{n} E_i\right) = \tau_\nu(P(H))$$

and

$$\nu = \frac{1}{1+\prod\wedge_i}.$$

■

V. Introduction: hedges in the Zadeh's sense

In the early 1970s, Zadeh [7] introduced a class of powering modifiers that defined the concept of linguistic variables and hedges. He proposed computing with words as an extension of fuzzy sets and logic theory (Zadeh [8], [9]). The linguistic hedges (LHs) change the meaning of primary term values. Many theoretical studies have contributed to the computation with words and to the LH concepts (see De Cock and Kerre [10]; Huynh, Ho, and Nakamori [11]; Rubin [12]; Türksen [13]).

As pointed out by Zadeh [14], [15], [16], linguistic variables and terms are closer to human thinking (which emphasise importance more than certainty) and are used in everyday life. For this reason, words and linguistic terms can be used to model human thinking systems (Liu et al. [17]; Zadeh [18]).

1) General form of modifiers: Three types of modifiers were introduced earlier. These are the

1) Negation operator:

$$\eta_{\nu,\nu_0}(x) = f^{-1}\left(f(\nu_0)\frac{f(\nu)}{f(x)}\right)$$

2) Hedge operators, necessity and possibility operators:

$$\tau_{\nu,\nu_0}(x) = f^{-1}\left(f(\nu_0)\frac{f(x)}{f(\nu)}\right) \tag{14}$$

3) Sharpness operator:

$$\chi^{(\lambda)}(x) = f^{-1}\left(f^{\lambda}(x)\right) \tag{15}$$

These three types of operators can be represented in a common form.

Definition 8: The general form of the modifier operators is

$$\kappa^{(\lambda)}_{\nu,\nu_0}(x) = f^{-1}\left(f(\nu_0)\left(\frac{f(x)}{f(\nu)}\right)^{\lambda}\right) \tag{16}$$

Theorem 9: Negation (1), hedge (14) and sharpness (15) are special cases of this modifier.

Proof:

$$\begin{array}{rcll} \lambda & = & -1 & \text{is the negation operator} \\ \lambda & = & 1 & \text{is the hedge operator} \\ f(\nu_0) = f(\nu) & = & 1 & \text{is the sharpness operator} \end{array}$$

■

VI. Characterization of the modifiers by a differential equation

We saw previously that $\kappa(x)$ is closely related to the generator function. Namely, it is an isomorphic mapping of the abstract space of objects onto the real line. If we have different types of mapping, then we will have different operators. If we change the isomorphic function f we get conjunctive, disjunctive or aggregation operators. See Figure 1.
We will characterize $\kappa(x)$ by this f function.

Let us introduce the effectiveness notion for $f(x)$ by defining the following

$$r(x) = \frac{f'(x)}{f(x)}$$

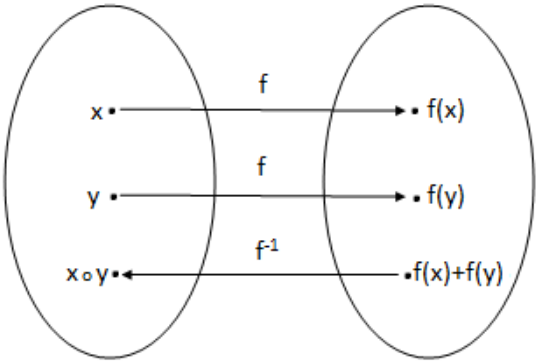


Fig. 1. Interpretation of $x \circ y = f^{-1}(f(x) + f(y))$

Definition 10: We say that the effectiveness of $\kappa(x)$ is normal if

$$r(\kappa(x)) = \lambda r(x)$$

and the boundary condition holds

$$\nu_0 = \kappa(\nu)$$

Theorem 11: Let $\kappa(x)$ be an arbitrary, continuous and differentiable modifier on $[0, 1]$ with the property

$$\kappa(\nu) = \nu_0.$$

$\kappa(x)$ is normal iff

$$\kappa(x) = f^{-1}\left(f(\nu_0)\left(\frac{f(x)}{f(\nu)}\right)^{\lambda}\right),$$

where $\lambda \neq 0$. This is the general form of the modifier operator.

Proof: Let f be a strictly monotonous transformation of $\kappa(x)$. Then

$$r(\kappa(x)) = \frac{f(\kappa(x))'}{f(\kappa(x))} = \lambda\frac{f'(x)}{f(x)} = r(x) \tag{17}$$

or, rewriting this equation, we have

$$\frac{f(\kappa(x))'}{f'(x)} = \lambda\frac{f(\kappa(x))}{f(x)},$$

That is, the ratio of the speed of the transformed value $\kappa(x)$ and the speed of x are the same as the ratio of the transformed value and the value x multiplied by a constant λ.

Recall that

$$(\ln(f(x)))' = \frac{f'(x)}{f(x)},$$

so Eq.(17) can be written in the following form:

$$(\ln(f(\kappa(x))))' = (\lambda\ln(f(x)))'$$

Integrating both sides, we get

$$\ln(f(\kappa(x))) = \lambda \ln(f(x)) + C$$

and

$$f(\kappa(x)) = Cf^{\lambda}(x)$$

Expressing this in terms of $\kappa(x)$, we find that

$$\kappa(x) = f^{-1}\left(Cf^{\lambda}(x)\right)$$

Using the boundary condition $\kappa(\nu) = \nu_0$,

$$f(\nu_0) = Cf^{\lambda}(\nu) \quad \Rightarrow \quad C = \frac{f(\nu_0)}{f^{\lambda}(\nu)},$$

which is the desired result. ■

Corollary 12: From (17), we see that

$$\frac{f'(\kappa(x))\kappa'(x)}{f(\kappa(x))} = \lambda \frac{f'(x)}{f(x)}$$

and so

$$\kappa'(x) = \lambda \frac{f(\kappa(x))}{f'(\kappa(x))} \frac{f'(x)}{f(x)}$$

is related to negation operators

VII. Sigmoid and kappa function

The sigmoid function plays an role in economics, biology, chemistry and other sciences. A key feature of the sigmoid function is that the solution of the differential equation $\sigma^{'}(x) = \lambda\sigma(x)(1-\sigma(x))$ is the sigmoid function.

Let us replace $\sigma(x)$ function by $\frac{\kappa(x)}{x}$ and $1-\sigma(x)$ by $\frac{1-\kappa(x)}{1-x}$, then with this substitution, we can characterize the kappa function. Here, $\sigma(x)$ is the value and $\frac{\kappa(x)}{x}$ is the relative value; and similarly $1-\sigma(x)$ is the value and $\frac{1-\kappa(x)}{1-x}$ is the relative value.

Theorem 13: Let $\kappa(x)$ be an arbitrary continuous and differentiable function on $[0,1]$ with the properties

$$\kappa(\nu) = \nu_0 \qquad \text{and} \qquad \kappa'(x) = \lambda \frac{\kappa(x)(1-\kappa(x))}{x(1-x)}.$$

If the above conditions hold, then the $\kappa(x)$ function is

$$\kappa(x) = \frac{1}{1 + \frac{1-\nu_0}{\nu_0}\left(\frac{1-\nu}{\nu}\frac{x}{1-x}\right)^{\lambda}} \tag{18}$$

Proof: Making use of (18), we can write

$$\begin{aligned} y' = \kappa'(x) &= \lambda \frac{1}{x(1-x)} \frac{1}{\frac{1}{\kappa(x)(1-\kappa(x))}} \\ &= \lambda \frac{1}{x(1-x)} \frac{1}{\frac{1}{y(1-y)}} \\ &= \frac{f_1(x)}{f_2(y)} \end{aligned}$$

where

$$f_1(x) = \frac{\lambda}{x(1-x)} \qquad f_2(x) = \frac{1}{y(1-y)}$$

The differential equation

$$y' = \frac{f_1(x)}{f_2(y)}$$

is separable, so the solution is:

$$\int f_2(y)dy = \int f_1(x)dx + c$$

In our case

$$\int \frac{1}{y(1-y)}dy = \lambda \int \frac{1}{x(1-x)}dx + c$$

so

$$\ln(y) - \ln(y-1) = \lambda\ ln(x) - ln(x-1) + ln(a)$$

We get

$$\frac{y}{1-y} = a\left(\frac{x}{1-x}\right)^{\lambda},$$

so the unary operator is

$$y = \frac{1}{1 + a\left(\frac{1-x}{x}\right)^{\lambda}}$$

■

VIII. Characterization of the kappa function by odds

Let us denote the odds of the input variable by X, i.e.

$$X = \frac{x}{1-x} \tag{19}$$

and the odds of the output (transformed) value by Y, i.e.

$$Y = \frac{\kappa(x)}{1-\kappa(x)}. \tag{20}$$

Let $F(X)$ be the corresponding function between the input and output odds

$$Y = F(X). \tag{21}$$

It is natural to assume that

$$F(X_1X_2) = C_1F(X_1)F(X_2) \tag{22}$$

or

$$F\left(\frac{X_1}{X_2}\right) = C_2\frac{F(X_1)}{F(X_2)}, \tag{23}$$

where $C_i > 0, \quad i = 1, 2.$

Theorem 14: The general solution of (21) and (23) is

$$F(X_1, X_2) = \frac{1}{C_1}X^{\lambda} \tag{24}$$

$$F(X_1, X_2) = C_2X^{\lambda} \tag{25}$$

Then (24) and (25) have the same form, i.e. $C_1 = \frac{1}{C_2}$.

Proof: See [19]. ■

Theorem 15: If (22) (or (23)) is true, then $\kappa(x)$ has the form

$$\kappa(x) = \frac{1}{1 + C\left(\frac{1-x}{x}\right)}$$

Proof: Using (19), (20) and (24), we get

$$\frac{\kappa(x)}{1-\kappa(x)} = \frac{1}{C}\left(\frac{x}{1-x}\right)^{\lambda}. \tag{26}$$

■

A. Extension of $\kappa_{\nu}^{\lambda}(x)$ to $[a,b]$ interval

We can extend K on the $[a,b]$ interval. $K : [a,b] \to [0,1]$

Here, $\kappa(x)$ is defined on $[0,1]$. We can extend it to the $[a,b]$ interval by applying a linear transformation. That is,

$$x := \frac{x-a}{b-a}$$

and

$$\nu := \frac{x_\nu - a}{b-a}$$

Then we get

$$K_{a,b}^{(\lambda)}(x) = \frac{1}{1 + \frac{1-\nu_0}{\nu_0}\left(\frac{x_\nu - a}{b - x_\nu}\frac{b-x}{x-a}\right)^{\lambda}}$$

$K_{a,b}^{(\lambda)}(x)$ has the following properties:

$$K_{a,b}^{(\lambda)}(x_\nu) = \nu_0,$$
$$K_{a,b}^{(\lambda)}(a) = 0 \quad \text{and} \quad K_{a,b}(b) = 1$$

$K_{a,b}$ has the same form as that in [20], where we showed that this is the general form for a certain class of membership function.

We can extend K such that $K^* : [a,b] \to [A,B]$. This $K^*(x)$ function can be written in the following implicit form:

$$\frac{K^*(x) - K^*(a)}{K^*(b) - K^*(x)}\frac{K^*(b) - K^*(x_\nu)}{K^*(x_\nu) - K^*(a)} = \left(\frac{x-a}{b-x}\right)^{\lambda}\left(\frac{b-x_\nu}{x_\nu - a}\right)^{\lambda} \tag{27}$$

Now let

$$X := \left(\frac{x-a}{b-x}\right)^{\lambda}\left(\frac{b-x_\nu}{x_\nu - a}\right)^{\lambda}\frac{K^*(x_\nu) - K^*(a)}{K^*(b) - K^*(x_\nu)} \tag{28}$$

Then we get the explicit form of $K(x)$

$$K^*(x) = \frac{XK^*(b) + K^*(a)}{1+X} \tag{29}$$

IX. Conclusions

In this study, we provided a general and theoretical basis for modalities. Here, we defined necessity and possibility operators, then we defined them via using a generator function of the Pliant operators. We gave a theoretical basis for hedges and the main result is a unified formula for unary operators. The subroutine can be downloaded from the following website: http://www.inf.u-szeged.hu/~dombi/.

Acknowledgements

This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013).

References

[1] P. Cintula, E. Klement, R. Mesiar, and M. Navara, "Fuzzy logics with an additional involutive negation," *Fuzzy Sets and Systems*, vol. 161, pp. 390–411, 2010.

[2] P. Hájek, *Metamathematics of fuzzy logic.* Kluwer Academic Publishers, Dordrecht, 1998.

[3] S. Gottwald and P. Hájek, "Triangular norm-based mathematical fuzzy logics," *Logical Algebraic, Analytic and Probabilistic Aspects of Triangular Norms*, pp. 275–299, 2005.

[4] E. Trillas, "Sobre functiones de negacion en la teoria de conjuntas difusos," *Stochastica*, vol. 3, pp. 47–60, 1979.

[5] J. Dombi., "De Morgan systems with an infinitely many negations in the strict monotone operator case," *Information Sciences*, vol. 181, pp. 1440–1453, 2011.

[6] ——, "Towards a general class of operators for fuzzy systems," *IEEE Transactions on Fuzzy Systems*, vol. 16, pp. 477–484, 2008.

[7] L. A. Zadeh, "A fuzzy-set - theoretic interpretation of linguistic hedges." *Journal of Cybernetics*, vol. 2(3), pp. 4–34, 1972.

[8] ——, "Fuzzy logic = computing with words," *IEEE Transactions on Fuzzy Systems*, vol. 4, pp. 103–111, 1996.

[9] ——, "From computing with numbers to computing with words-From manipulation of measurements to manipulation of perceptions." *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications*, vol. 45(1), pp. 105–119, 1999.

[10] M. D. Cock and E. E. Kerre., "Fuzzy modifiers based on fuzzy relations." *Information Sciences*, vol. 160, pp. 173–199, 2004.

[11] V. N. Huynh, T. B. Ho, and Y. Nakamori., "A parametric representation of linguistic hedges in Zadeh's fuzzy logic." *International Journal of Approximate Reasoning*, vol. 30, pp. 203–223, 2002.

[12] S. H. Rubin, "Computing with words." *IEEE Systems, Man and Cybernetics, Part B*, vol. 29(4), pp. 518–524, 1999.

[13] I. B. Turksen., "A foundation for CWW: Meta-linguistic axioms." *IEEE Fuzzy Information*, vol. Processing NAFIPS'04, pp. 395–400, 2004.

[14] L. A. Zadeh., "The concept of a linguistic variable and its application to approximate reasoning, part 1." *Information Sciences*, vol. 8, pp. 199–249, 1975.

[15] ——, "The concept of a linguistic variable and its application to approximate reasoning, part 2." *Information Sciences*, vol. 8, pp. 301–357, 1975.

[16] ——, "The concept of a linguistic variable and its application to approximate reasoning, part 3." *Information Sciences*, vol. 9, pp. 43–80, 1975.

[17] B. D. Liu, C. Y. Chen, and J. Y. Tsao., "Design of adaptive fuzzy logic controller based on linguistic-hedge concepts and genetic algorithms." *IEEE Systems, Man and Cybernetics, Part B*, vol. 31(1), pp. 32–53, 2001.

[18] L. A. Zadeh., "Quantitative fuzzy semantics." *Information Sciences*, vol. 3, pp. 159–176, 1971.

[19] J. Aczél, "Lectures on functional equations and their applications, acad," *Press, New York*, vol. 1966, 1966.

[20] J. Dombi, "Membership function as an evaluation," *Fuzzy Sets and Systems*, vol. 35, pp. 1–21, 1990.

József Dombi degrees earned at University of Szeged. Academic degrees: University doctor's degree (1977, Summa cum laude), Candidate mathematical sciences (1994, CSc, Title of dissertation: Fuzzy sets' structure from the aspect of multicriteria decision aid.). Visiting positions: Leningrad (1971, 6 months), Leipzig (1974, 6 months), Bukarest (1978, 1 month), DAAD Scholarship, Aachen (1978, 12 months), Alexander von Humboldt Scholarship, Aachen (1986, 12 months), European Scholarship, Bristol (1987, 3 months), Academic research exchange (Paris, Helsinki, Turku, Tampere), CEEPUS guest professor, Linz (2000 and 2009, 1 month), and Klagenfurt (2008, 1 month). Awards: 1991 Minister of Education award, 1997 Pro Sciencia award for student, 1998 László Kalmár award, in 1997 DataScope won Information Technology award in Brussels, won the Best Software of the Year award in 2000 at COMDEX, Las Vegas. Editorial membership: editorial boardâs member of the Foundations of Computing and Decision Sciences magazine until 1997, editorial boardâs member of the International Journal of Systems Sciences and editorial boardâs member of the International Journal of Advanced Intelligence Paradigms. Membership in international organizations: IFSA (International Fuzzy System Association), Member of European COST Action on Multicriteria Decision Making, ESIGMA (European Special Interest Group on Multicriteria Analysis), European Working Group on Multiple Criteria Decision Aid, MTA Public Body (Operational Research), MOT (Hungarian Association of Operational Research), Hungarian Humboldt Association. Business: 1993 founder and head of Cygron Research Ltd., 2000 scientific consultant of Mindmaker Ltd., 2002 founder and head of Dopti research and development Ltd. Research interest: computational intelligence, theory of fuzzy sets, multicriteria decision making, genetic and evolutional algorithms, operation research.

ISBN: 978-1-61804-240-8

MHD mixed convection flow of a second-grade fluid on a vertical surface

Fotini Labropulu, Daiming Li and Ioan Pop

***Abstract*—**An analysis of the steady magnetohydrodynamis (MHD) mixed convection flow of a viscoelastic fluid stagnating orthogonally on a heated or cooled vertical flat plate has been studied. Using similarity variables, the governing equations are transformed into a system of two coupled non-linear ordinary differential equations which then are solved numerically using the spectral method.

***Keywords*—**non -Newtonian; mixed convection; incompressible; stagnation-point; spectral method.

I. Introduction

INTEREST in the study of non-Newtonian fluids has become more evident over the past decades due to the occurrence of these fluids in many engineering and industrial applications. Non-Newtonian fluids is a broad class of fluids in which the relation between the shear stress and the shear rate is nonlinear and hence there exist many constitutive relations. One such relation describes the so-called second-grade fluid. The equations of motion of second-grade fluid are highly non-linear and one order higher than the Navier-Stokes equations.

The mixed convection flow is encountered in many industrial and technological applications which include electronic devices cooled by fans, solar central receivers exposed to wind currents, etc. (Seshadri et al. [1]). The mixed convection in stagnation flow is important when the buoyancy forces due to the temperature difference between the surface and the free stream become large. Consequently, both the flow and thermal fields are significantly affected by the buoyancy forces. Hiemenz [2] derived an exact solution of the Navier-Stokes equations which describes the steady flow of a viscous and incompressible fluid directed perpendicularly (orthogonally) to an infinite flat plate. An analytic technique, namely the homotopy analysis method (HAM), has been recently used by Hayat et al. [3] to study the steady mixed convection in two-dimensional stagnation flows of a viscoelastic fluid around a heated vertical surface for the case when the temperature of the wall varies linearly with the distance from the stagnation point.

The aim of this paper is to analyze the steady two-dimensional mixed convection flow of a second grade fluid stagnating orthogonally on a heated or cooled vertical flat plate. The governing equations are transformed into a system of two coupled non-linear ordinary differential equations using similarity variables. Employing the spectral method (Canuto et al. [4]), the resulting equations are then solved numerically.

F. Labropulu is with Luther College, University of Regina, Regina, SK, Canada S4S 0A2 (phone: 306-585-5040; fax: 306-585-5267; e-mail: fotini.labropulu@uregina.ca).

D. Li , was with University of Regina, Regina, SK Canada. He is now with the Department of Petroleum Engineering, University of Calgary, Calgary, AB, Canada (e-mail: lidaiming@hotmail.com).

I. Pop is with the Faculty of Mathematics, University of Cluj, R-3400 Cluj, CP 253, Romania (e-mail: popm.ioan@yahoo.co.uk).

II. Basic equations

Consider the steady orthogonal mixed convection flow of a second grade fluid close to the stagnation point on a vertical surface. The oblique velocity of the inviscid (potential) fluid is $\bar{v}_e(\bar{u}_e, \bar{v}_e)$ and it is assumed that the temperature of the plate is $T_w(\bar{x})$, while the uniform temperature of the ambient fluid is T_∞, where $T_w(\bar{x}) > T_\infty$ corresponds to a heated plate (assisting flow) and $T_w(\bar{x}) < T_\infty$ corresponds to a cooled plate (opposing flow), respectively. The potential velocity $\bar{v}_e$ has the components $\bar{u}_e = a\bar{x} + b\bar{y}$ and $\bar{v}_e = -a\bar{y}$, where a and b are positive constants. It is also assumed that the wall temperature $T_w(\bar{x})$ varies linearly with $\bar{x}$ being of the form $T_w(\bar{x}) = T_\infty + c\bar{x}$, where c is positive or negative constant. The bar on a variable denotes its dimensional form.

Under these assumptions along with the Boussinesq approximation, the steady two-dimensional mixed convection flow of a second grade fluid can be written in dimensionless form as follows:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{1}$$

$$u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}=-\frac{\partial p}{\partial x}+\nabla^2 u+W_e\left\{\frac{\partial}{\partial x}\left[2u\frac{\partial^2 u}{\partial x^2}+2v\frac{\partial^2 u}{\partial x\partial y}+4\left(\frac{\partial u}{\partial x}\right)^2+2\frac{\partial v}{\partial x}\left(\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\right)\right]+\frac{\partial}{\partial y}\left[\left(u\frac{\partial}{\partial x}+v\frac{\partial}{\partial y}\right)\left(\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\right)+2\frac{\partial u}{\partial x}\frac{\partial u}{\partial y}+2\frac{\partial v}{\partial x}\frac{\partial v}{\partial y}\right]\right\}+\lambda_1\frac{\partial}{\partial x}\left[\left(\frac{\partial u}{\partial x}\right)^2+\left(\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\right)^2\right]-Mu\pm\lambda T \tag{2}$$

$$u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}=-\frac{\partial p}{\partial y}+\nabla^2 v+W_e\left\{\frac{\partial}{\partial x}\left[2\frac{\partial u}{\partial x}\frac{\partial u}{\partial y}\left(u\frac{\partial}{\partial x}+v\frac{\partial}{\partial y}\right)\left(\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\right)+2\frac{\partial v}{\partial x}\frac{\partial v}{\partial y}\right]+\frac{\partial}{\partial y}\left[2u\frac{\partial^2 v}{\partial x\partial y}+4\left(\frac{\partial v}{\partial y}\right)^2+2\frac{\partial u}{\partial y}\left(\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\right)+2v\frac{\partial^2 v}{\partial y^2}\right]\right\}+\lambda_1\frac{\partial}{\partial y}\left[\left(\frac{\partial v}{\partial y}\right)^2+\left(\frac{\partial v}{\partial x}+\frac{\partial u}{\partial y}\right)^2\right] \tag{3}$$

$$u\frac{\partial T}{\partial x}+v\frac{\partial T}{\partial y}=\frac{1}{\Pr}\nabla^2 T \tag{4}$$

and the boundary conditions in dimensionless form are

$$v=0,\quad u=0,\quad T=T_w(x)=x\quad \text{at}\quad y=0$$
$$u_e=x,\quad v_e=-y,\quad T=0,\quad p=p_e\quad \text{as}\quad y\to\infty \tag{5}$$

where $\Pr$ is the Prandtl number, W_e is the Weissenberg number, M is the Hartmann number, λ_1 is a viscoelastic parameter and λ is the constant mixed convection parameter.

It should be mentioned that $\lambda>0$ (heated plate) corresponds to assisting flow and $\lambda<0$ (cooled plate) corresponds to opposing flow, while $\lambda=0$ corresponds to forced convection flow, respectively.

Using Eqs. (2) and (3), and boundary conditions (5), the non-dimensional pressure $p=p_e$ of the inviscid or far flow can be expressed as

$$p=p_e=-\frac{1}{2}(x^2+y^2)+Const. \tag{6}$$

The physical quantity of interest are the shear stress or skin friction at the wall and the local heat flux from the flat plate, which can be easily shown that in dimensionless form are given by

$$\tau_w=\left\{\frac{\partial u}{\partial y}+\frac{\partial v}{\partial x}+W_e\left[u\left(\frac{\partial^2 u}{\partial x\partial y}+\frac{\partial^2 v}{\partial x^2}\right)+v\left(\frac{\partial^2 u}{\partial y^2}+\frac{\partial^2 v}{\partial x\partial y}\right)-2\frac{\partial v}{\partial y}\frac{\partial u}{\partial y}+2\frac{\partial v}{\partial x}\frac{\partial v}{\partial y}\right]\right\}_{y=0} \tag{7}$$

$$q_w=-\left(\frac{\partial T}{\partial y}\right)_{y=0} \tag{8}$$

The boundary conditions (5) suggest that Eqs. (1) to (4) have a solution of the form

$$u=xF'(y),\quad v=-F(y),\quad T=x\theta(y) \tag{9}$$

where prime denotes differentiation with respect to y. Substituting (9) into Eqs. (2) to (4) and eliminating the pressure p by cross differentiation of Eqs. (2) and (3) it results in, after one integration of the resulting equation, the following ordinary differential equations

$$F'''+FF''-F'^2-We\left(FF^{iv}-2F'F'''+F''^2\right)-MF'\pm\lambda\theta+C_1=0 \tag{10}$$

$$\frac{1}{\Pr}\theta''+F\theta'-F'\theta=0 \tag{11}$$

where C_1 is a constant of integration. The boundary conditions (5) become

$$F(0)=0,\ F'(0)=0,\ F'(\infty)=1$$
$$\theta(0)=1,\quad \theta(\infty)=0 \tag{12}$$

Taking the limit $y\to\infty$ in Eq. (10) and using the boundary conditions $F'(\infty)=1$ and $\theta(\infty)=0$, we get $C_1=1$. Further, from an analysis of the boundary layer equation (10) it results in that $F(y)$ behaves as

$$F(y)=y+A\quad \text{as}\quad y\to\infty \tag{13}$$

where $A=A(We)$ is a constant accounts for the boundary layer displacement

Employing (9), the dimensionless skin friction and the local heat transfer given by equations (7) and (8) can now be written as

$$\tau_w=xF''(0) \tag{14}$$

$$q_w=-x\theta'(0) \tag{15}$$

where the values of $F''(0)$ and $\theta'(0)$ can be calculated from equations (10) to (11) with the boundary conditions (12) for various values of the parameters $We,\ \lambda,\ M$ and $\Pr$.

ISBN: 978-1-61804-240-8

III. Method of solution, results and discussion

The coupled equations (10) and (11) subject to boundary conditions (12) has been solved numerically for various values of $We,\ \lambda,\ M$ and $\Pr$ using the highly accurate and at the same time stable spectral method (Canuto et al., [4]).

Numerical values of $F''(0)$ and $-\theta'(0)$ for assisting flows are shown in Tables 1 to 2 for various values of $\Pr$ when $We = 0.0$, $\lambda = 1$ and $M = 0.$ These values are in good agreement with previously reported values of Ramachandran et a. [5], Lok et al. [6] and Ishak et al. [7]. Numerical values of $F''(0)$ and $-\theta'(0)$ for assisting and opposing flows are shown in Tables 3 to 4 for various values of $\Pr$ and We when $\lambda = 0.2$ and $M = 0.$ It is observed from Tables 3 and 4 that the skin friction coefficient and the local heat transfer are decreasing when the Weissenberg number We is increasing in both the assisting and opposing flows. On the other hand, the skin friction is decreasing and the local heat transfer is increasing when Pr is increasing in the case of assisting flow. In the case of opposing flow, the skin friction and the local heat transfer are increasing as Pr is increasing.

The variation of $F'(y)$ for various values of λ when $We = 0.3$ and $\Pr = 1$ for assisted flow is shown in Figure 1. Figure 2 depicts the variation of $F'(y)$ for various values of λ when $We = 0.3$ and $\Pr = 1$ for opposed flow. Figure 3 illustrates the variation of $\theta(y)$ for various values of λ when $We = 0.3$ and $\Pr = 1$ for assisted flow. The variation of $\theta(y)$ for various values of λ when $We = 0.3$ and $\Pr = 1$ for opposed flow are shown in Figure 4.

References

[1] R. Seshadri, N. Sreeshylan and G. Nath, "Unsteady mixed convection flow in the stagnation region of a heated vertical plate due to impulsive motion", Int. J. Heat Mass Transfer, 45 (2002), pp. 1345-1352.

[2] K. Hiemenz, "Die Grenzschicht an einem in den gleichförmigen Flüssigkeitsstrom eingetauchten geraden Kreiszylinder", Dingler Polytech, J. 326 (1911), pp. 321-324.

[3] T. Hayat, Z. Abbas and I. Pop, "Mixed convection in the stagnation point flow adjacent to a vertical surface in a viscoelastic fluid", Int. J. Heat Mass Transfer 51 (2008), pp. 3200-3206.

[4] C. Canuto, M.Y. Hossaini, A.Quartcroni and T.A. Zang, "Spectral Methods in Fluid Dynamics", Springer, Berlin, 1987.

[5] N. Ramachandran, T.S. Chen and B.F. Armaly, "Mixed convection in stagnation flows adjacent to vertical surfaces", J. Heat Transfer 110 (1988), pp. 373-377.

[6] Y.Y. Lok, . Amin, D. Campean, I. Pop, "Steady mixed convection flow of a micropolar fluid near the stagnation point on a vertical surface", Int. J. Num. Meth. Heat Fluid Flow 15 (2005), pp. 654-670.

[7] A. Ishak, R. Nazar and I. Pop, "Mixed convection boundary layer flow adjacent to a vertical surface embedded in a stable stratified medium", Int. J. Heat Mass Transfer 51 (2008), pp. 3693-3695.

Table 1. Numerical values of $F''(0)$ for various values of $\Pr$ for assisting flow when $\lambda = 1, We = 0$ and $M = 0.$

Pr	Ramachandran et al. [12]	Lok et al. [13]	Ishak et al. [14]	Present Results
0.7	1.7063	1.706376	1.7063	1.7063
1	-	-	1.6754	1.6754
7	1.1579	1.517952	1.5179	1.5179
10	-	-	1.4928	1.4928
20	1.4485	1.448520	1.4485	1.4485
40	1.4101	1.410094	1.4101	1.4101
60	1.3903	1.390311	1.3903	1.3903
80	1.3774	1.377429	1.3774	1.3774

Table 2. Numerical values of $-\theta'(0)$ for various values of $\Pr$ for assisting flow when $\lambda = 1, We = 0$ and $M = 0.$

Pr	Ramachandran et al. [12]	Lok et al. [13]	Ishak et al. [14]	Present Results
0.7	0.7641	0.764087	0.7641	0.7641
1	-	-	0.8708	0.8708
7	1.7224	1.722775	1.7224	1.7224
10	-	-	1.9446	1.9446
20	2.4576	2.458836	2.4576	2.4576
40	3.1011	3.103703	3.1011	3.1011
60	3.5514	3.555404	3.5514	3.5514
80	3.9095	3.194882	3.9095	3.9095

Table 3. Numerical values of $F''(0)$ for various values of Pr and We when $\lambda = 0.2$ and $M = 0$.

We	Pr	Assisting Flow	Opposing flow
0.0	0.2	1.35426	1.10711
0.2	0.2	1.15591	0.95607
0.5	0.2	0.98230	0.81854
0.7	0.2	0.90441	0.75554
1.0	0.2	0.81738	0.68434
1.5	0.2	0.71694	0.60129
2.0	0.2	0.64713	0.54310
0.2	0.0	1.19920	0.92091
	0.5	1.14411	0.96893
	0.7	1.13961	0.97378
	1.0	1.13482	0.97892

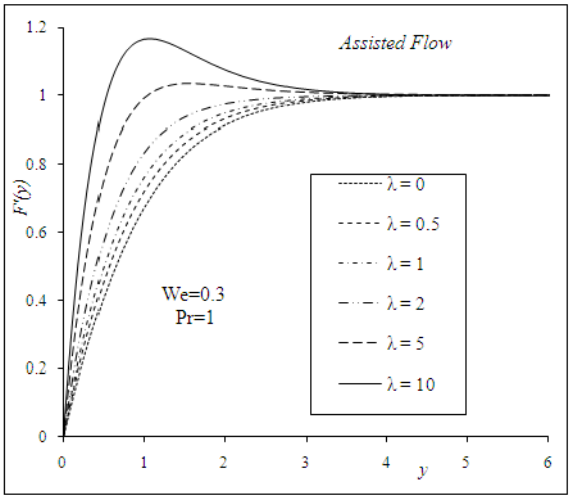


Figure 1: Variation of $F'(y)$ for various values of λ when $We = 0.3$ and $\Pr = 1$ for assisted flow.

Table 4. Numerical values of $-\theta'(0)$ for various values of Pr and We when $\lambda = 0.2$ and $M = 0$.

We	Pr	Assisting Flow	Opposing flow
0.0	0.2	0.44198	0.42351
0.2	0.2	0.42606	0.40958
0.5	0.2	0.40990	0.39499
0.7	0.2	0.40177	0.38753
1.0	0.2	0.39189	0.37837
1.5	0.2	0.37922	0.36652
2.0	0.2	0.36944	0.35729
0.2	0.0	0.03221	0.03221
	0.5	0.60841	0.58753
	0.7	0.69073	0.66820
	1.0	0.78862	0.76435

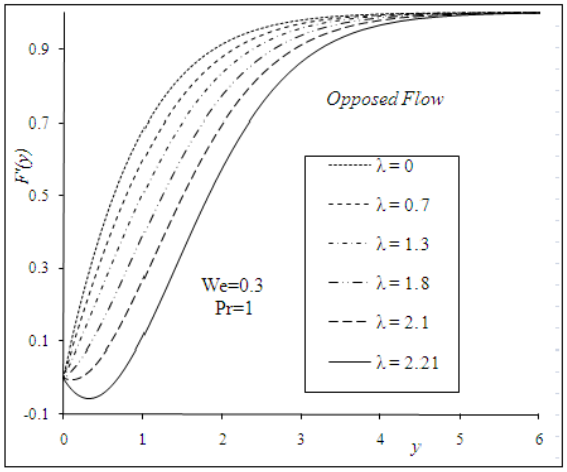


Figure 2: Variation of $F'(y)$ for various values of λ when $We = 0.3$ and $\Pr = 1$ for opposed flow.

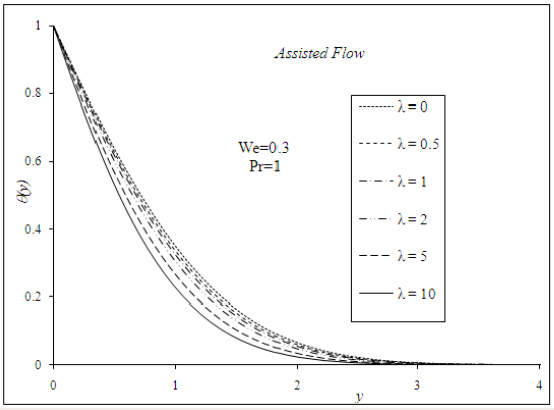


Figure 3: Variation of $\theta(y)$ for various values of λ when $We = 0.3$ and $\Pr = 1$ for assisted flow.

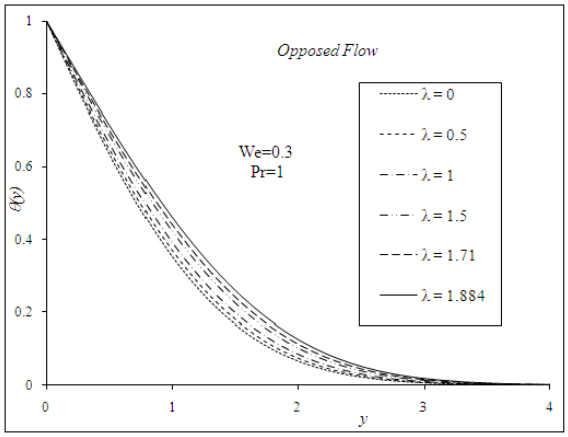


Figure 4: Variation of $\theta(y)$ for various values of λ when $We = 0.3$ and $\Pr = 1$ for opposed flow.

Workflow Analysis - A Task Model Approach

Glória Cravo, *Member of Center for Linear Structures and Combinatorics, University of Lisbon*

***Abstract*—In this paper we describe the structure of a workflow as a graph whose vertices represent tasks and the arcs are associated to workflow transitions. To each task an input/output logic operator is associated and this logic operator can be the logical AND $(\bullet)$, the OR $(\otimes)$, or the XOR -exclusive-or - $(\oplus)$. Furthermore, we associate a Boolean term to each transition present in the workflow.**

The main contribution of this paper is the analysis of a workflow through its tasks which allows us to describe the dynamism of the workflow in a very simple way.

Finally, we describe the logical termination of workflows and we present conditions under which this property is valid.

***Index Terms*—Graphs, Propositional Logic, Workflows, Business Processes.**

I. Introduction

A Workflow is an abstraction of a business process that consists on the execution of a set of tasks to complete a process (for example, hiring process, loan application, sales order processing, etc.). Tasks represent unities of work to be executed that can be processed by a combination of resources, such as a computer program, an external system, or human activity. In the literature several papers are devoted to the study of workflows, see for example [1], [2], [3], [4], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] and the references therein. In general, these approaches contemplate the use of Petri Nets, State and Activity Charts, Event-Condition-Action rules, Temporal Logic and Markov Chains.

In this paper we present a formalism to describe and analyse the structure of workflows based on the concept of graph and Propositional Logic. An important highlight of this paper is the emphasis on the tasks present in the workflow, which allows us to identify easily the dynamism present in the workflow. Finally, we describe the logical termination in a very intuitive form and we present conditions under which this property is valid.

II. Workflow Structure

In this Section we provide a complete description of the structure of workflows. We start by introducing the formal concept of a workflow. This workflow structure can be also found in [5], [6], [7]. Notice this type of graphs has and input/output logic operator associated with each vertex.

Definition 2.1: [5], [6], [7] *A workflow is a tri-logic acylic directed graph* $WG = (T, A, A', M)$, *where* $T = \{t_1, t_2, \ldots, t_n\}$ *is a finite nonempty set of vertices representing workflow tasks. Each task* t_i (*i.e., a vertex*) *has attributed an input logic operator* (*represented by* $\succ t_i$) *and an output logic operator* (*represented by* $t_i \prec$). *An input/output logic operator can be the logical AND* $(\bullet)$, *the OR* $(\otimes)$, *or the XOR -exclusive-or -* $(\oplus)$. *The set* $A = \{a_\sqcup, a_\sqcap, a_1, a_2, \ldots, a_m\}$ *is a finite nonempty set of arcs representing workflow transitions. The transition* $a_\sqcup$ is the tuple $(\sqcup, t_1)$ and transition $a_\sqcap$ is the *tuple* $(t_n, \sqcap)$, where the *symbols* $\sqcup$ *and* $\sqcap$ *represent abstract tasks which indicate the entry and ending point of the workflow*, *respectively. Every transition* $a_i, i \in \{1, \ldots, n\}$ corresponds to a tuple of the form (t_k, t_l), where $t_k, t_l \in T$.

We use the symbol $'$ *to reference the label of a transition, i.e.,* a'_i *references transition* a_i, $a_i \in A$. *The elements* a'_i *are called Boolean terms and form the set* A'.

Given $t_i \in T$, the *incoming transitions for task* t_i are the tuples of the form $(t_l, t_i), t_l \in T$, and the *outgoing transitions* are the tuples of the form $(t_i, t_l), t_l \in T$.

The *incoming/outgoing condition of task* t_i is the Boolean expression $a'_{k_1}\varphi \ldots \varphi a'_{k_l}$, where $\varphi \in \{\bullet, \otimes, \oplus\}$, $a'_{k_1}, \ldots, a'_{k_l} \in A'$ and $a_{k_1}, \ldots, a_{k_l}$ are the *incoming/outgoing* transitions of task t_i. The terms $a'_{k_1}, \ldots, a'_{k_l}$ are connected with the logic operator $\succ t_i$, $t_i \prec$, respectively. If task t_i has only one *incoming/outgoing* transition we assume that the condition does not have logic operator.

An Event-Action (EA) *model for task* t_i *is an implication of the form* $t_i : f_E \rightsquigarrow f_C$, *where* f_E *and* f_C *are the incoming and outgoing conditions of task* t_i, *respectively. An* EA *model has the behavior with two distinct modes: when* f_E *is evaluated to* $true$, f_C *is also evaluated to* $true$; *when* f_E *is evaluated to* $false$, f_C *is always* $false$. *And the* EA *model* $t_i : f_E \rightsquigarrow f_C$ *is true if both* f_E, f_C are true, otherwise it is false. We say that the EA *model* $t_i : f_E \rightsquigarrow f_C$ *is positive if its Boolean value is true, otherwise it is said to be negative.*

We denote by M the set of all EA models present in WG.

Task t_i is said to be executed if the EA *model* $t_i : f_E \rightsquigarrow f_C$ *is positive. In this case,* task t_i has *attributed the Boolean value true.*

Remark 1: Given an expression whose Boolean value is true (respectively, false), we simply can represent this fact by 1, (respectively, 0).

Remark 2: Given an EA model $t_i : f_E \rightsquigarrow f_C$, if f_E is false, then task t_i disables all its outgoing transitions. Consequently f_C is also false.

Notice the workflow starts its execution by enabling transition $a_\sqcup$, i.e., by asserting $a'_\sqcup$ to be *true*. In other words, the workflow starts its execution by executing task t_1.

Notice that a'_i is *true* if transition a_i is enabled, otherwise a_i is *false*. Transitions can be enabled by a user or by an external event. If the EA model $t_i : f_E \rightsquigarrow f_C$ is negative, then both f_E, f_C are *false*. In this case, all the transitions of f_C are disabled.

Glória Cravo is with the Center of Exact Sciences and Engineering, University of Madeira, 9020-105 Funchal, Madeira, Portugal, e-mail: gcravo@uma.pt (see http://www.uma.pt).

ISBN: 978-1-61804-240-8

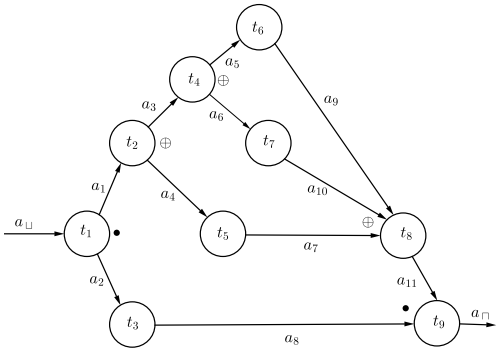


Fig. 1. Example of a workflow.

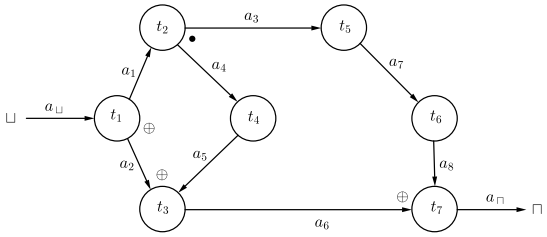


Fig. 2. Example of a workflow.

Example 1: In Figure 1 we present a workflow $WG = (T, A, A', M)$, where $T = \{t_1, t_2, \ldots, t_9\}$, $A = \{a_{\sqcup}, a_{\sqcap}, a_1, a_2, \ldots, a_{11}\}$, $A' = \{a'_{\sqcup}, a'_{\sqcap}, a'_1, a'_2, \ldots, a'_{11}\}$, $M = \{t_1 : a'_{\sqcup} \rightsquigarrow a'_1 \bullet a'_2,\ t_2 : a'_1 \rightsquigarrow a'_3 \oplus a'_4,\ t_3 : a'_2 \rightsquigarrow a'_8,\ t_4 : a'_3 \rightsquigarrow a'_5 \oplus a'_6,\ t_5 : a'_4 \rightsquigarrow a'_7,\ t_6 : a'_5 \rightsquigarrow a'_9,\ t_7 : a'_6 \rightsquigarrow a'_{10},\ t_8 : a'_7 \oplus a'_9 \oplus a'_{10} \rightsquigarrow a'_{11},\ t_9 : a'_8 \bullet a'_{11} \rightsquigarrow a'_{\sqcap}\}$.

The output logic operator of task t_2 ($t_2 \prec$) is a *XOR* ($\oplus$), while the input logic operator of task t_9 ($\succ t_9$) is an *AND* ($\bullet$).

The incoming transition for task t_2 is $a_1 = (t_1, t_2)$ and its outgoing transitions are $a_3 = (t_2, t_4)$ and $a_4 = (t_2, t_5)$. Hence the incoming condition for task t_2 is a'_1, while its outgoing condition is $a'_3 \oplus a'_4$.

Task t_2 is executed if the EA model $t_2 : a'_1 \rightsquigarrow a'_3 \oplus a'_4$ is positive, i.e., if a'_1 is true and only one of the Boolean terms a'_3, a'_4 is true.

Proposition 2.2: Let $WG = (T, A, A', M)$ be a workflow. Let $a_l = (t_i, t_j) \in A$, $t_i, t_j \in T$. If a'_l is true, then t_i is necessarily executed.

Proof 1: Let us assume that a'_l is true. Let $t_i : f_{E_i} \rightsquigarrow f_{C_i}$ be the EA model associated to task t_i. If task t_i is not executed, then the EA model $t_i : f_{E_i} \rightsquigarrow f_{C_i}$ is negative. Since the EA model is negative, all outgoing transitions of task t_i are disabled, in particular a_l is disabled, i.e., a'_l is false, wich is a contradiction. Hence task t_i is executed. ■

Remark 3: The condition of Proposition 2.2 is not sufficient. For example in the workflow from Figure 1, if task t_2 is executed, then the EA model $t_2 : a'_1 \rightsquigarrow a'_3 \oplus a'_4$ is positive. For $a'_1 = true$, $a'_3 = true$, $a'_4 = false$, $a_4 = (t_2, t_5)$, t_2 is executed, but a'_4 is false.

Remark 4: Let us consider the Boolean term a'_l where $a_l = (t_i, t_j) \in A$, $t_i, t_j \in T$. If a'_l is true, task t_j is not necessarily executed. For example, in the workflow from Figure 2, let us assume that $a'_{\sqcup} = true$, $a'_1 = true$, $a'_2 = false$, $a'_3 = true$, $a'_4 = true$, $a'_5 = true$, $a'_6 = true$, $a'_7 = true$, $a'_8 = true$, $a'_{\sqcap} = false$. Hence, for this assignment the EA model $t_7 : a'_6 \oplus a'_8 \rightsquigarrow a'_{\sqcap}$ is negative, which means that task t_7 is not executed. Nevertheless, $a_8 = (t_6, t_7)$ and a'_8 is true.

Next we introduce the concept of logical termination. This is a very important structural property, since its analysis will allow to verify if a workflow will eventually finish, according to the initial specifications.

Definition 2.3: Let $WG = (T, A, A', M)$ be a workflow. We say that WG logically terminates if task t_n is executed whenever task t_1 is executed.

In the following result we establish a necessary and sufficient condition for the logical termination.

Theorem 2.4: Let $WG = (T, A, A', M)$ be a workflow. Then WG logically terminates if and only if $a'_{\sqcap}$ is true whenever $a'_{\sqcup}$ is true.

Proof 2: Let us assume that WG logically terminates, i.e., task t_n is executed whenever task t_1 is executed. This means that the EA model $t_n : f_{E_n} \rightsquigarrow a'_{\sqcap}$ is positive whenever the EA model $t_1 : a'_{\sqcup} \rightsquigarrow f_{C_1}$ is positive. Bearing in mind that WG starts its execution by executing task t_1, then the EA model $t_1 : a'_{\sqcup} \rightsquigarrow f_{C_1}$ is positive. Hence the EA model $t_n : f_{E_n} \rightsquigarrow a'_{\sqcap}$ is also positive. Consequently, $a'_{\sqcup}$, f_{C_1}, f_{E_n}, $a'_{\sqcap}$ are true. Thus, $a'_{\sqcap}$ is true whenever $a'_{\sqcup}$ is true.

Conversely, let us assume that $a'_{\sqcap}$ is true whenever $a'_{\sqcup}$ is true. Let us assume that task t_1 is executed. This means that the EA model $t_1 : a'_{\sqcup} \rightsquigarrow f_{C_1}$ is positive. Bearing in mind that $a'_{\sqcap}$ is true, according to the behavior of the EA models, necessarily f_{E_n} is true. Hence the EA model $t_n : f_{E_n} \rightsquigarrow a'_{\sqcap}$ is positive, which means that task t_n is executed. So we can conclude that task t_n is executed whenever task t_1 is executed, which means that WG logically terminates. ■

Example 2: It is not hard to check that in the workflow from Figure 1, $a'_{\sqcap}$ is true whenever $a'_{\sqcup}$ is true. Thus, the workflow logically terminates.

Next we address our study on the dynamism present in a workflow. Obviously the dynamism is associated to the sequencial execution of its tasks. In the workflow from Figure 1 the execution of task t_1 implies the execution of both tasks t_2, t_3; the execution of task t_2 implies the execution of only one of the tasks t_4, t_5; the execution of task t_4 implies the execution of only one of the tasks t_6, t_7; the execution of only one of the tasks t_5, t_6, t_7 implies the execution of task t_8. Finally, the execution of both tasks t_3, t_8 implies the execution of taks t_9. Hence, we can state the execution of task t_1 implies the execution of $t_2 \bullet t_3$; the execution of task t_2 implies the execution of $t_4 \oplus t_5$; the execution of task t_4 implies the execution of $t_6 \oplus t_7$; the execution of $t_5 \oplus t_6 \oplus t_7$ implies the execution of task t_8; the execution of $t_3 \bullet t_8$ implies the

execution of taks t_9. Notice that when we consider $t_2 \bullet t_3$, the operator $\bullet$ is the output logic operator of task t_1,while when we consider $t_5 \oplus t_6 \oplus t_7$, $\oplus$ is the input logic operator of task t_8.

These remarks led us to introduce the following concept.

Definition 2.5: Let $WG = (T, A, A', M)$ be a workflow. The compound tasks of WG are the elements of the following form: $t_{i_1}\varphi t_{i_2}\varphi \dots \varphi t_{i_k}, t_{i_1}, t_{i_2}, \dots t_{i_k} \in T, \varphi \in \{\bullet, \otimes, \oplus\}$. The set of all compound tasks of WG is denoted by T', i.e.:

$$T' = \{t_{i_1}\varphi t_{i_2}\varphi \dots \varphi t_{i_k} : t_{i_1}, t_{i_2}, \dots t_{i_k} \in T, \varphi \in \{\bullet, \otimes, \oplus\}\}.$$

Example 3: In the workflow from Figure 1, $T' = \{t_2 \bullet t_3, t_4 \oplus t_5, t_6 \oplus t_7, t_5 \oplus t_6 \oplus t_7, t_3 \bullet t_8\}$.

Remark 5: Since every task t_i has associated a Boolean value, according to its execution, it is also natural to attribute a Boolean value to the compound tasks of WG. The natural attribution is the following. Given any compound task of WG, $t_{i_1}\varphi t_{i_2}\varphi \dots \varphi t_{i_k}, \varphi \in \{\bullet, \otimes, \oplus\}$:

If $\varphi = \bullet$, then the Boolean value of $t_{i_1}\varphi t_{i_2}\varphi \dots \varphi t_{i_k}$ is 1 if and only if the Boolean value of all tasks $t_{i_1}, t_{i_2}, \dots, t_{i_k}$ is equal to 1;

If $\varphi = \otimes$, then the Boolean value of $t_{i_1}\varphi t_{i_2}\varphi \dots \varphi t_{i_k}$ is 1 if and only if there exists at least one of the tasks $t_{i_1}, t_{i_2}, \dots, t_{i_k}$ whose Boolean value is equal to 1;

If $\varphi = \oplus$, then the Boolean value of $t_{i_1}\varphi t_{i_2}\varphi \dots \varphi t_{i_k}$ is 1 if and only if there exists only one of the tasks $t_{i_1}, t_{i_2}, \dots, t_{i_k}$ with Boolean value equal to 1.

Naturally, we can state that a compound task $t_{i_1}\varphi t_{i_2}\varphi \dots \varphi t_{i_k}$ is executed if and only if its Boolean value is equal to 1, which means that the compound task $t_{i_1}\varphi t_{i_2}\varphi \dots \varphi t_{i_k}$ is positive. In other words, $t_{i_1}\varphi t_{i_2}\varphi \dots \varphi t_{i_k}$ is executed if and only if:

If $\varphi = \bullet$, all tasks $t_{i_1}, t_{i_2}, \dots, t_{i_k}$ are executed;

If $\varphi = \otimes$, at least one of the tasks $t_{i_1}, t_{i_2}, \dots, t_{i_k}$ is executed;

If $\varphi = \oplus$, only one of the tasks $t_{i_1}, t_{i_2}, \dots, t_{i_k}$ is executed.

Definition 2.6: Let $WG = (T, A, A', M)$ be a workflow. Let $t_i,\ t_j,\ t_{i_1}, t_{i_2}, \dots, t_{i_k},\ t_{j_1}, t_{j_2}, \dots, t_{j_l} \in T,\ \varphi,\ \psi\{\bullet, \otimes, \oplus\}$. A compound task model is an implication with one of the following forms:

(1) $t_i \hookrightarrow t_{j_1}\psi t_{j_2}\psi \dots \psi t_{j_l}$;

(2) $t_{i_1}\varphi t_{i_2}\varphi \dots \varphi t_{i_k} \hookrightarrow t_j$;

(3) $t_{i_1}\varphi t_{i_2}\varphi \dots \varphi t_{i_k} \hookrightarrow t_{j_1}\psi t_{j_2}\psi \dots \psi t_{j_l}$.

Usually we represent a compound task model by $t_{I_i} \hookrightarrow t_{O_i}$, where t_{I_i} is called the incoming task and t_{O_i} is called the outgoing task. We say that a compound task model $t_{I_i} \hookrightarrow t_{O_i}$ is positive if both incoming and outgoing tasks are positive, i.e., if both tasks t_{I_i}, t_{O_i} are executed.

In particular, the implication of the form $t_i \hookrightarrow t_j$ is called a simple task model. Clearly, it is positive if both tasks t_i, t_j are executed.

The set of all simple and compound task models present in WG is called the set of task models of WG and is denoted by TM.

The task models have the behavior with two distinct modes: if its incoming task is true, necessarily its outgoing task is true; if the incoming task is false, the outgoing task is false. In other words, if $t_{I_i} \hookrightarrow t_{O_i}$ is a compound task model, then t_{I_i} is executed if and only if t_{O_i} is executed.

Notice that in a compound task model $t_{I_i} \hookrightarrow t_{O_i}$, at least one of the tasks t_{I_i}, t_{O_i} is compound.

Example 4: In the workflow from Figure 1, the set of its task models is: $TM = \{t_1 \hookrightarrow t_2 \bullet t_3,\ t_2 \hookrightarrow t_4 \oplus t_5,\ t_4 \hookrightarrow t_6 \oplus t_7,\ t_5 \oplus t_6 \oplus t_7 \hookrightarrow t_8,\ t_3 \bullet t_8 \hookrightarrow t_9\}$.

From now on, we use the symbol $\longleftrightarrow$ with the following meaning: $X \longleftrightarrow Y$ means that the compound statements X and Y are logically equivalent.

According to simple rules of Logic and taking into account the behavior of the task models, we can infer the following rules that allow us to identify new task models present in the workflow.

Remark 6: Let $WG = (T, A, A', M)$ be a workflow.

(a) If both tasks models $t_{I_i} \hookrightarrow t_{O_i}$ and $t_{I_j} \hookrightarrow t_{O_j}$ belong to TM and $t_{O_i} \longleftrightarrow t_{I_j}$ then the model $t_{I_i} \hookrightarrow t_{O_j}$ still holds in WG.

(b) If both task models $t_{I_i} \hookrightarrow t_{O_i}$ and $t_{I_j} \hookrightarrow t_{O_j}$ belong to TM, where $t_{O_i} \longleftrightarrow t_L\varphi t_{I_j}$, $\varphi \in \{\bullet, \otimes, \oplus\}$ then the task model $t_{I_i} \hookrightarrow t_L\varphi t_{O_j}$ still holds in WG.

(c) If both task models $t_{I_i} \hookrightarrow t_{O_i}$ and $t_{O_j} \hookrightarrow t_{I_j}$ belong to TM, where $t_{O_i} \longleftrightarrow t_L\varphi t_{I_j}$, $\varphi \in \{\bullet, \otimes, \oplus\}$ then the task model $t_{I_i} \hookrightarrow t_L\varphi t_{O_j}$ still holds in WG.

(d) If both task models $t_{I_i} \hookrightarrow t_{O_i}$ and $t_{I_j} \hookrightarrow t_{O_j}$ belong to TM, where $t_{I_i} \longleftrightarrow t_L\varphi t_{I_j}$, $\varphi \in \{\bullet, \otimes, \oplus\}$ then the task model $t_L\varphi t_{O_j} \hookrightarrow t_{O_i}$ still holds in WG.

(e) If both task models $t_{I_i} \hookrightarrow t_{O_i}$ and $t_{O_j} \hookrightarrow t_{I_j}$ belong to TM, where $t_{I_i} \longleftrightarrow t_L\varphi t_{I_j}$, $\varphi \in \{\bullet, \otimes, \oplus\}$ then the task model $t_L\varphi t_{O_j} \hookrightarrow t_{O_i}$ still holds in WG.

The previous remark allow us to identify new task models, as it is described below.

Definition 2.7: Let $WG = (T, A, A', M)$ be a workflow. An extended task model is a model obtained by applying a finite sequence of some of the rules presented in Remark 6. We denote by TM' the set of all extended task models of WG.

Example 5: In the workflow From Figure 1, bearing in mind that $t_1 \hookrightarrow t_2 \bullet t_3$, $t_2 \hookrightarrow t_4 \oplus t_5 \in TM$, according to Remark 6 we can conclude that the model $t_1 \hookrightarrow (t_4 \oplus t_5) \bullet t_3$ still holds in WG. Therefore, we can state that $t_1 \hookrightarrow (t_4 \oplus t_5) \bullet t_3$ is an extended task model of WG.

Notice we adopt the same notation of the task models to represent the extended task models. Furthermore, the extended task models verify the same properties of the task models. In particular, given an extended task model $B \hookrightarrow C$, necessarily both B, C have the same Boolean value.

Definition 2.8: Let $WG = (T, A, A', M)$ be a workflow. We define the closure of TM as the set of all task models and extended task models in WG. This set is denoted by TM^*. In other words, $TM^* = TM \cup TM'$.

Example 6: As we saw in Example 4 in the workflow from Figure 1, $TM = \{t_1 \hookrightarrow t_2 \bullet t_3,\ t_2 \hookrightarrow t_4 \oplus t_5,\ t_4 \hookrightarrow t_6 \oplus t_7,\ t_5 \oplus t_6 \oplus t_7 \hookrightarrow t_8,\ t_3 \bullet t_8 \hookrightarrow t_9\}$. Since $t_1 \hookrightarrow t_2 \bullet t_3, t_2 \hookrightarrow t_4 \oplus t_5 \in TM$, according to Remark 6 we can deduce that $t_1 \hookrightarrow (t_4 \oplus t_5) \bullet t_3 \in TM^*$. Now bearing in mind that $t_4 \hookrightarrow t_6 \oplus t_7 \in TM$, applying again Remark 6 we can conclude that $t_1 \hookrightarrow ((t_6 \oplus t_7) \oplus t_5) \bullet t_3 \in TM^*$. As $(t_6 \oplus t_7) \oplus t_5 \longleftrightarrow t_5 \oplus t_6 \oplus t_7$ we can state that $t_1 \hookrightarrow (t_5 \oplus t_6 \oplus t_7) \bullet t_3 \in TM^*$. Bearing in mind that $t_5 \oplus t_6 \oplus t_7 \hookrightarrow t_8$, applying once more Remark 6 we infer that $t_1 \hookrightarrow t_8 \bullet t_3 \in TM^*$. As $t_8 \bullet t_3 \longleftrightarrow t_3 \bullet t_8$, applying again Remark 6 we conclude that $t_1 \hookrightarrow t_9 \in TM^*$.

Notice the workflow from Figure 1 logically terminates and $t_1 \hookrightarrow t_9 \in TM^*$. Furthermore, we studied many other examples of workflows that logically terminates and simultaneously $t_1 \hookrightarrow t_n \in TM^*$. The analysis of these different cases led us to formulate the following conjecture.

Conjecture 1: Given a workflow $WG = (T, A, A', M)$, then WG logically terminates if and only if $t_1 \hookrightarrow t_n \in TM^*$.

III. Conclusion

In this paper we develop a formalism to describe and analyse the structure of workflows. Furthermore, our analysis allows us to study the logical termination of workflows. In particular we present conditions under which this property is valid.

It is important to point out that our main emphasis is the analysis of a workflow through the study of its tasks. Another relevant aspect of our approach is the introduction of the concept of compound tasks. This concept allows us to identify new task models based on the existing ones. Through these new task models we are able to describe the dynamism present in a workflow. Clearly, the study of the dynamism of a workflow is equivalent to analyse the sequential execution of its tasks.

Finally, we conjecture that a workflow (WG) logically terminates if and only if $t_1 \hookrightarrow t_n$ is in the closure of WG.

References

[1] W. V. M. P. v. d. Aalst. The application of petri nets to workflow management. *Journal of Circuits, Systems and Computers*, 8(1):21–66, 1998.

[2] W. V. M. P. v. d. Aalst. Workflow verification: Finding control-flow errors using petri-net-based techniques. In W. V. M. P. v. d. Aalst, J. Desel, and A. Oberweis, editors, *Business Process Management: Models, Techniques, and Empirical Studies*, pages 161–183. Springer-Verlag, Berlin, 2000.

[3] P. Attie, M. Singh, A. Sheth, and M. Rusinkiewicz. Specifying and enforcing intertask dependencies. In *Proceedings of 19th International Conference on Very Large Data Bases*, pages 134–145, Dublin, Ireland, 1993. Morgan Kaufman.

[4] J. Cao, C. Chan, and K. Chan. Workflow analysis for web publishing using a state-activity process model. *Journal of Systems and Software*, 76(3):221–235, 2005.

[5] G. Cravo. Applications of propositional logic to workflow analysis. *Applied Mathematics Letters*, 23:272–276, 2010.

[6] G. Cravo. Logical termination of workflows: An interdisciplinary approach. *Journal of Numerical Analysis, Industrial and Applied Mathematics*, 5(3-4):153–161, 2011.

[7] G. Cravo and J. Cardoso. Termination of workflows: A snapshot-based approach. *Mathematica Balkanica*, 21(3-4):233–243, 2007.

[8] U. Dayal, M. Hsu, and R. Ladin. Organizing long-running activities with triggers and transactions. In *ACM SIGMOD International Conference on Management of Data Table of Contents*, pages 204–214, Atlantic City, New Jersey, 1990. ACM Press, New York, NY, USA.

[9] J. Eder, H. Groiss, and H. Nekvasil. A workflow system based on active databases. In G. Chroust and A. Benczur, editors, *Proceedings of CON' 94, Workflow Management: Challenges, Paradigms and Products*, pages 249–265, Linz, Austria, 1994.

[10] F. Gottschalk, W. V. M. P. v. d. Aalst, M. H. Jansen-Vullers, and M. La Rosa. Configurable workflow models. *International Journal of Cooperative Information Systems*, 17(2):223–255, 2008.

[11] F. Gottschalk, W. V. M. P. v. d. Aalst, M. H. Jasen-Vullers, and H. M. W. Verbeek. Protor2CPN: Using colored petri nets for configuring and testing business processes. *International Journal on Software Tools for Technology Transfer*, 10(1):95–111, 2008.

[12] A. H. M. t. Hofstede and E. R. Nieuwland. Task structure semantics through process algebra. *Software Engineering Journal*, 8(1):14–20, 1993.

[13] J. Klingemann, J. Wäsch, and K. Aberer. Deriving service models in cross-organizational workflows. In *Proceedings of RIDE-Information Technology for Virtual Enterprises (RIDE-VE' 99)*, pages 100–107, Sydney, Australia, 1999.

[14] P. Muth, D. Wodtke, J. Weissenfels, G. Weikum, and K. Dittrich. Enterprise-wide workflow management based on state and activity charts. In A. Dogac, L. Kalinichenko, T. Ozsu, and A. Sheth, editors, *Proceedings NATO Advanced Study Institute on Workflow Management Systems and Interoperability*. Springer-Verlag, 1998.

[15] A. Rozinat, R. S. Mans, M. Song, and W. V. M. P. v. d. Aalst. Discovering colored petri nets from event logs. *International Journal on Software Tools for Technology Transfer*, 10(1):57–74, 2008.

[16] M. P. Singh and K. van Hee. Semantical considerations on workflows: An algebra for intertask dependencies. In *Fifth International Workshop on Database Programming Languages*, Umbria, Italy, 1995. Springer.

[17] H. M. W. Verbeek, W. V. M. P. v. d. Aalst, and A. H. M. Hofstede. Verifying workflows with cancellation regions and OR-joins: An approach based on relaxed soundness and invariants. *Computer Journal*, 50(3):294–314, 2007.

Glória Cravo holds a first degree in Mathematics from the University of Lisbon, Portugal (1992), and a MSc. Degree in Mathematics from the University of Lisbon, Portugal (1997). In 2003 she obtained her PhD in Mathematics, from University of Lisbon, Portugal. The main contribution of her PhD thesis is in the area of the so-called Matrix Completion Problems. Between 1992 and 1997, she worked at the Department of Mathematics of the University of Madeira as auxiliary teacher. From 1997 to 2003 she worked at the Department of Mathematics of the University of Madeira as teaching assistant. Since 2003 she works at the Department of Mathematics and Engineering of the University of Madeira as assistant professor. Her main areas of research are Matrix Completion Problems, Graph Theory, Control Theory, Logic, Computer Science and Wireless Sensor Networks. She is an active research member of the Center for Linear Structures and Combinatorics.

Degrees of Freedom and Advantages of Different Rule-Based Fuzzy Systems

Marco Pota, and Massimo Esposito

***Abstract*—**Rule-based fuzzy systems are gaining increasing importance for classification in many fields of application. Various degrees of freedom for the construction of rule-based fuzzy models are analyzed here, comprising fuzzy sets shape, different types of norms, axes rotation, and weights for antecedents and consequents of each rule and for different rules. Results of application on an example dataset are discussed in terms of classification performances, taking into account interpretability at the same time.

***Keywords*—**classification, norms, rule-based fuzzy systems, fuzzy set shapes, weights.

I. Introduction

RULE-BASED fuzzy systems are gaining increasing importance in a widening variety of fields of application. In particular, classification problems can be tackled by using fuzzy systems, basically constituted by a fuzzy partition of influencing variables, and a rule base connecting them to different classes to which each data sample should be associated. However, some structural decisions to be made in designing this type of systems have not been extensively considered yet. For example, at the best of our knowledge, the choices of the shape of fuzzy sets, often triangular or Gaussian, and of the type of T-norm and S-norm used for inference, have been studied only few times with rationale motivations [1-4]. Moreover, while a simple fuzzy system could be promptly interpretable by the user, many complications can be added to the system, making it gradually more similar to a neural network, aiming at improving performances, but returning the drawback of hindering the overall transparency and interpretability.

In this work, the classification performances of different rule-based fuzzy systems are studied, starting from a simple system, and going towards a more complicated one. A recently developed method [5] is used to extract knowledge from data, and to obtain a reference fuzzy system.

Firstly, on the simplified version of the reference system, the choice of the shape of MFs is evaluated, by considering the differences among a crisp system, a fuzzy system with linear membership functions (MFs), and the substitution with smoothed MFs. Then, using a family of T-norms and S-norms, and their soft couterparts, the types of norm which are more appropriate for different uses in the inference process are individuated. Moreover, possible complications added to the simplified system, such as the use of linear combinations of the original variables, the use of weights for different antecedents of each rule, weights for different rules, and the definition of rule consequents as a set of classes with respective probabilities instead of a single class, are evaluated in terms of their power of improving performances, and the results are discussed in order to decide whether such improvements could be great enough to justify the associated loss of system interpretability.

A well-known dataset, i.e. the Wisconsin Breast Cancer Dataset [6], is used as a proof of concepts. In particular, only a couple of independent variables is considered, in order to allow the 3D visualization of surfaces representing the posterior probabilities of different classes, and thus discuss the changes in the results while structural changes are disposed.

This paper is organized as follows. In Section II, some possible improvements to a simple fuzzy system for classification are listed, and followed by an example of a recently developed system. In Section III, the results of a comparison among different systems are given and discussed. Finally, Section IV concludes the work.

II. Variants of a Fuzzy System

A. A Simple Fuzzy System for Classification

A fuzzy system is basically made of the fuzzy partitions of the variables of interest (once these have been selected), and of a rule base.

Each fuzzy partition is made of a collection of fuzzy sets, representing the terms of the associated linguistic variable. Suppose that n variables are considered. Then, suppose that the range of the j-th variable is partitioned into T_j fuzzy sets $F_{t_j}^{(j)}$, with $t_j = 1,...,T_j$. Each of the N data samples $\mathbf{x}_i = \left\{x_i^{(1)},...,x_i^{(n)}\right\}$, $i = 1,...,N$, belongs to the fuzzy set $F_{t_j}^{(j)}$ with the membership grade $\mu_{t_j}^{(j)}\left[x_i^{(j)}\right]$.

A complete rule base [7] is made of a set of $T_1 \cdot ... \cdot T_n$ rules

M. Pota is with the Institute for High Performance Computing and Networking (ICAR) of National Research Council of Italy (CNR), Napoli, 80131 Italy (corresponding author to provide phone: +39 3384174424; fax: +39 0816139531; e-mail: marco.pota@na.icar.cnr.it).

M. Esposito is with the Institute for High Performance Computing and Networking (ICAR) of National Research Council of Italy (CNR), Napoli, 80131 Italy (e-mail:massimo.esposito@na.icar.cnr.it).

$r_{\{t_1,...,t_n\}}$, with $\{t_1,...,t_n\} \in \{1,...,T_1\} \times ... \times \{1,...,T_n\}$, of the type:

$$\begin{cases} if\ x^{(1)}\ is\ F_1^{(1)}\ and\ ...\ and\ x^{(n)}\ is\ F_1^{(n)}\ then\ C_{\{1,...,1\}} \\ ... \\ if\ x^{(1)}\ is\ F_{T_1}^{(1)}\ and\ ...\ and\ x^{(n)}\ is\ F_{T_n}^{(n)}\ then\ C_{\{T_1,...,T_n\}} \end{cases}, \quad (1)$$

where antecedents are all possible combinations of fuzzy sets of the partitions, and consequents are fuzzy sets [8,9].

Each data sample $\mathbf{x}_i$ fires the rule $r_{\{t_1,...,t_n\}}$ with a strength

$$FS_{\{t_1,...t_n\}} = \underset{j=1,...,n}{T}\left[\mu_{t_j}^{(j)}\left(x_i^{(j)}\right)\right], \quad (2)$$

while the implication of the consequence is usually modeled as:

$$IMP_{\{t_1,...,t_n\}} = T\left[FS_{\{t_1,...,t_n\}}, C_{\{t_1,...,t_n\}}\right], \quad (3)$$

and different implications are aggregated as:

$$AGG = \underset{\{t_1,...,t_n\} \in \{1,...,T_1\} \times .. \times \{1,...,T_n\}}{S}\left[IMP_{\{t_1,...,t_n\}}\right], \quad (4)$$

where T is a T-norm and S is an S-norm.

Note that FS in (2) is a number, while IMP in (3) and AGG in (4) are fuzzy sets. Suppose that there are K possible classes $C_1,...,C_K$, then the fuzzy set AGG assume a set of values $AGG_1,...,AGG_K$ in correspondence of different classes. Therefore, a defuzzification step is required, which can be accomplished in different ways: usually, the class k which takes the greatest AGG_k is chosen (winner-takes-all strategy); however, since a result constituted by different classes with respective confidence grades should be preferred [10], then the defuzzification can be implemented as a normalization [4]:

$$DEF_k = AGG_k / \sum_{\kappa=1}^{K} AGG_\kappa \ . \quad (5)$$

B. Shape of Membership Functions

Only few works have discussed on the opportunity of using a certain shape for MFs [1-3], however, results seem to be application-dependent. In general, the following considerations could be made.

First of all, the difference between crisp intervals (binary MFs) and fuzzy intervals (trapezoidal MFs) can be studied, even if the advantages of fuzzy logic [11] are widely recognized [12,13].

Fuzzy numbers (triangular MFs) can be viewed as particular cases of fuzzy intervals. Even if they are used very often, their advantage with respect to fuzzy intervals can be ascribed only to a lower number of parameters (3 instead of 4), which reflects the difference between a number and an interval. However, intervals are generally more appropriate to model the terms of a linguistic variable. On the other hand, intervals of the type $x < x_A$ or $x > x_B$ cannot be fuzzified with increasing and then decreasing MFs, like triangular and Gaussian ones, but "shoulders" (only decreasing or only increasing MFs) are required to model extreme linguistic terms instead. Therefore, triangular, Gaussian and bell-shaped MFs will not be considered here, even if Gaussian MFs require a lower number of parameters.

Trapezoidal MFs can be provided with some improvements, in order to reach: *i)* derivability, i.e. to obtain MFs smoothly going from the plateaus to the increasing/decreasing segments; *ii)* non-linearity, i.e. to obtain MFs overcoming the limits of linear models between input and output variables. These issues can be settled by considering sigmoidal MFs.

A trapezoidal or a sigmoidal MF can be represented by 4 parameters a, b, c, d, while a "shoulder" by 2 parameters. A whole partition with T MFs can be described by $2(T-1)$ parameters. Binary MFs are represented here with the same number of parameters. Therefore, the MFs are modeled here as follows: right "shoulders" $\mu_R[x;c,d]$ are binary

$$\mu_{RB}[x;c,d] = \begin{cases} 0, & x \le (c+d)/2 \\ 1, & (c+d)/2 \le x \end{cases}, \quad (6)$$

trapezoidal

$$\mu_{RT}[x;c,d] = \begin{cases} 0, & x \le c \\ \dfrac{x-c}{d-c}, & c < x < d \\ 1, & d \le x \end{cases}, \quad (7)$$

or sigmoidal

$$\mu_{RS}[x;c,d] = \frac{1}{1+e^{\left(t\cdot\frac{d+c}{d-c} - 2\cdot t\cdot\frac{x}{d-c}\right)}}, \quad (8)$$

respective left "shoulders" are

$$\mu_L[x;a,b] = 1 - \mu_R[x;a,b], \quad (9)$$

while internal MFs are

$$\mu_I[x;a,b,c,d] = 1 - \mu_L[x;a,b] - \mu_R[x;c,d], \quad (10)$$

with $a \le b \le c \le d$ and $t = Log[1/\varepsilon - 1]$, where $\varepsilon \ll 1$ (in the following $\varepsilon = 0.01$) is a positive constant fixed to approximate sigmoidal MFs to normal, i.e. $\varepsilon \le \mu[x] \le 1-\varepsilon$. Using these settings, binary sets divide the variable range in intervals, trapezoids fuzzify these intervals, and trapezoidal and sigmoidal MFs with the same parameters result very similar, the latter being smoother.

In the following, a comparison of performances will be made among binary, trapezoidal and sigmoidal MFs. In Fig. 1, the three types of partition are shown, all with a number of terms $T = 3$ and the same parameters.

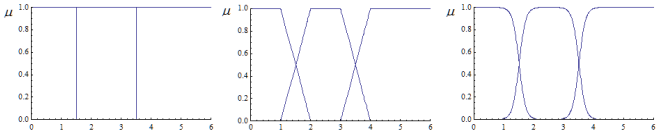


Fig. 1 Binary, trapezoidal, and sigmoidal fuzzy partitions.

C.A Family of (Soft) T-norms and S-norms

Many types of T-norms and S-norms were developed. In particular, Mamdani, Product, Łukasiewicz and Drastic T-norms are the most used. However, very few attempts [4] were done on deciding which type is the most suitable for different applications. At the same time, some parameterized families of norms were developed, which smoothly changes one norm into the others by varying just one parameter *p*. Here, the Frank T-norm family [14] is considered:

$$T\left[a_1,...,a_d;p\right]=\begin{cases}\min\left[a_1,...,a_d\right], & p=0\\ \prod_{\delta=1}^{d}a_\delta, & p=1\\ \max\left[\sum_{\delta=1}^{d}a_\delta-d+1,0\right], & p=+\infty\\ \log_p\left[1+\frac{\prod_{\delta=1}^{d}\left(p^{a_\delta}-1\right)}{p-1}\right], & otherwise\end{cases}, \quad (11)$$

with $p\geq 0$. The corresponding S-norm is calculated as:

$$S\left[a_1,...,a_d;p\right]=1-T\left[1-a_1,...,1-a_d;p\right]. \quad (12)$$

Moreover, following [15], a soft version of both norms, which is a middle way between each norm and the arithmetic mean, can be defined as follows:

$$\tilde{T}\left[a_1,...,a_d;p,\alpha\right]=(1-\alpha)\frac{1}{d}\sum_{\delta=1}^{d}a_\delta+\alpha T\left[a_1,...,a_d;p\right], \quad (13)$$

$$\tilde{S}\left[a_1,...,a_d;p,\alpha\right]=(1-\alpha)\frac{1}{d}\sum_{\delta=1}^{d}a_\delta+\alpha S\left[a_1,...,a_d;p\right]. \quad (14)$$

In the following, the dependence of system performances on both p and α parameters will be evaluated. In particular, their use for different T-norms and S-norms, i.e. p_F and α_F in (2), p_I and α_I in (3), p_A and α_A in (4), is investigated, and different optimal values of them can be found.

D.Preferred Directions

Some of the best regression [16,17] or classification [18,19] systems are based on the observation that there exist a set of directions in the variables space, which is better than others in explaining the behavior of the system. E.g., in [16] these preferential directions are those which maximize the variance and result more useful to study the output variable, while in [18,19] they define a hyperplane which divides the space where the samples of one class are situated from the other where there is the other class. Generally, these directions do not correspond to the original variables of the dataset, but are linear combinations of them. They can be viewed as a possible reduction of dimensionality and axes rotation.

On the one hand, preferred directions could be surely useful to improve the system performances. On the other hand, they are scarcely used in fuzzy systems where interpretability is a main objective, since their use undoubtedly complicates the comprehensibility of the system.

In the following, the advantages of axes rotation will be studied by substituting a couple of original variables x_1 and x_2 with another orthogonal couple obtained by linearly combining them through a parameter *D*, i.e.

$$\left\{x_1,x_2\right\}\Rightarrow\left\{x_1+Dx_2,x_2-Dx_1\right\}. \quad (15)$$

E. Weights

Different importance can be assigned to antecedents of each rule, and to different rules of a rule base.

In the case of rule antecedents, assigning them different weights $w_1,...,w_n$ corresponds to giving different importance to the variables, and can be modeled [4] by substituting in (2) the T-norm with a weighted T-norm:

$$WT\left[a_1,...,a_n;w_1,...,w_n\right]=T\left[1-w_1(1-a_1),...,1-w_n(1-a_n)\right]. \quad (16)$$

In case of *R* different rules, their different impact on the result can be modeled [4] by substituting in (4) the S-norm with a weighted S-norm:

$$WS\left[a_1,...,a_R;W_1,...,W_R\right]=S\left[W_1a_1,...,W_Ra_R\right]. \quad (17)$$

All $w_1,...,w_n$ and $W_1,...,W_R$ must be in the interval $[0,1]$ [4].

The dependence on the system performances on the use of weights will be studied in the following by comparing weighted and non-weighted firing strength and aggregation.

F. Consequent definition

In case of regression, each consequent of (1) can be a singleton (i.e. a fuzzy set whose support is a single value with a membership grade of 1) or a proper fuzzy set [8], or a function [9]. In case of a classifier, which is examined in this work, each consequent can be a singleton whose support is a class [8], or a fuzzy set defined on different classes [8].

Suppose that there are *K* possible classes $C_1,...,C_K$. In the simplified model, each consequent is a singleton whose support is a class. Deeper knowledge can be modeled by a more complicated system. For example, if one knows that, in the space restriction modeled by antecedents of a rule $r_{\{t_1,...,t_n\}}$, different classes have respective probabilities $p_{\{t_1,...,t_n\}-1},...,p_{\{t_1,...,t_n\}-K}$, then these can be assigned to the rule consequent, which can be regarded as a fuzzy sets with different non-zero membership grades in correspondence of different classes. In this case, the result of defuzzification (5) can be viewed as the posterior probability [5]:

$$P\left[C_k\mid\mathbf{x}\right]=AGG_k/\sum_{\kappa=1}^{K}AGG_\kappa. \quad (18)$$

G.A Reference Method for Knowledge Extraction

The method [5], which was recently developed by authors, is considered as a reference one. It can be summarized as follows.

Each range of the original variables is partitioned into a collection of T_j fuzzy sets described by sigmoidal MFs, $\mu_{t_j}^{(j)}\left[x^{(j)}\right]$, $t_j = 1,...,T_j$, such that:

$$P\left[C_k \mid x^{(j)}\right] \simeq \sum_{t_j=1}^{T_j}\left(\lambda_{t_j-k} \cdot \mu_{t_j}^{(j)}\left[x^{(j)}\right]\right), \quad (19)$$

with $k = 1,...,K$ and $j = 1,...,n$.

Note that parameters of MFs and the number of fuzzy sets constituting each partition are optimized at the same time.

Then, a rule base as (1) is constructed, by considering a complete set of rules $r_{\{t_1,...,t_n\}}$, each one having a weight $W_{\{t_1,...,t_n\}}$ and different probabilities of classes $p_{\{t_1,...,t_n\}-k}$. Weights and probabilities are calculated as follows:

$$W_{\{t_1,...,t_n\}} = \frac{\sum_{k=1}^{K} w_{\{t_1,...,t_n\}-k}}{\sum_{\tau_1=1}^{T_1}\cdots\sum_{\tau_n=1}^{T_n}\sum_{k=1}^{K} w_{\{\tau_1,...,\tau_n\}-k}} \quad (20)$$

$$p_{\{t_1,...,t_n\}-k} = \frac{w_{\{t_1,...,t_n\}-k}}{\sum_{k=1}^{K} w_{\{t_1,...,t_n\}-k}}, \quad (21)$$

where

$$w_{\{t_1,...,t_n\}-k} = \frac{\lambda_{t_1-k} \cdot ... \cdot \lambda_{t_n-k}}{P\left[C_k\right]^{n-1}}, \quad (22)$$

and $P\left[C_k\right]$ are prior probabilities of classes.

The types of norms used are product T-norm in (2), product T-norm in (3), and weighted sum for S-norm in (4). After normalization (5), it is demonstrated that the results approximate posterior probabilities as in (18).

In the following, this method will be used to calculate parameters of membership functions, rule weights and rule consequents. Its results will be compared with those of the systems obtained by simplifying it and then: *i)* substituting different shapes of MFs; *ii)* modifying types of T-norms and S-norms; *iii)* using soft T-norms and S-norms; *iv)* rotating axes; *v)* adding weights to antecedents; *vi)* adding weights to rules; *vii)* substituting singletons with fuzzy consequents.

III. Results and Discussion

In this section, the Wisconsin Breast Cancer Dataset [6] is used as a proof of concepts. 699 samples are classified into benign (C_B) and malignant (C_M). Two input variables of the dataset are used here, which are Uniformity of Cell Shape (UCS) and Bland Chromatin (BC).

The choice of only two variables derives from the need of showing the behavior of 3D surfaces representing the functions of the posterior probabilities of classes (CPPFs). In the following (see Figs 3-7, 9), the surface representing the probability of C_B, given different values of UCS and BC, i.e. $P\left[C_B \mid \mathbf{x}\right]$, in red, and that representing $P\left[C_M \mid \mathbf{x}\right]$ in green, are showed together. The lateral view of the images is given when the height of the surfaces is to be shown, while the top view is preferred to emphasize the separation of the variables space into regions associated to different classes (e.g., if $P\left[C_B \mid \mathbf{x}\right] > P\left[C_M \mid \mathbf{x}\right]$, then the point $\mathbf{x}$ of the space is assigned to the class C_B and in the top view it is red).

Results of different systems are evaluated in terms of classification error (*CE*) and squared classification error (*SCE*):

$$CE = \frac{1}{N} \cdot \sum_{i=1}^{N}\left[1 \ \ if \ \ \hat{C}_i = C_i, 0 \ \ otherwise\right], \quad (23)$$

$$SCE = \frac{1}{N}\frac{1}{K}\sum_{i=1}^{N}\sum_{k=1}^{K}\left(\alpha_i^k - \delta_i^k\right)^2, \quad (24)$$

where $\hat{C}_i$ is the predicted class, C_i the real one, α_i^k the activation of *k*-th class for the *i*-th data sample, and δ_i^k is 1 if the correct class for the *i*-th data sample is the *k*-th one, 0 otherwise. In particular, while *CE* is simply the fraction of wrongly classified data items, *SCE* takes into account that high (low) confidence should be assigned to right (wrong) solutions. Both should be minimized.

The method described in Section II.G is used to determine parameters of membership functions and the number of terms of each partition. In particular, only two fuzzy sets constitute each of the two partitions, and the parameters of the MFs (8) and (9) results: 1.09 and 5.37 for UCS, 1.34 and 6.01 for BC. The obtained partitions are shown in Fig.2.

Moreover, the reference method is used to determine optimal rule weights (20) and rule consequents (21), of a rule base as (1), which are listed in the first column of Table I.A, together with the resulting performances.

In order to evaluate the usefulness of system complications, let us begin by considering a simple system, as described in Section II.A, where parameters p and α of norms are set as follows: $p_F = p_I = p_A = 0$ and $\alpha_F = \alpha_I = \alpha_A = 1$. Original variables are used, no weights are assigned to rule antecedents nor to different rules, and consequents are defined as singletons, obtained by substituting the highest probability found by (21) with 1, the others with 0.

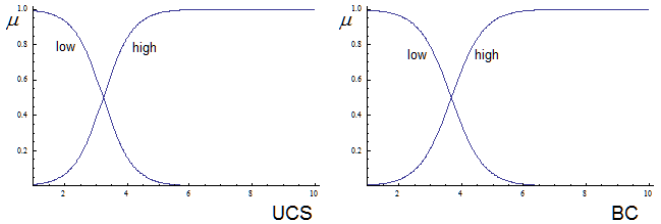


Fig. 2 Fuzzy partitions of two variables calculated by the reference method.

The first comparison can be made among different shapes of MFs, going from binary, through trapezoidal, to sigmoidal MFs. CPPFs given by the simple system in these three cases are shown in Fig.3, where CPPFs reflect the shape chosen for MFs, and the observations given in Section II.B can be repeated. Measures of performance are listed in Table I.A, where it can be seen that in all three cases the same *CE* is obtained, while an improvement of *SCE* is obtained with fuzzy systems with respect to crisp one. Using trapezoidal or sigmoidal MFs gives very similar results, therefore for the following runs the sigmoidal shape is chosen, in order to ensure derivability.

If the type of norms (11) and (12) is changed by varying p_F and p_I, as explained in Section II.C, then the CPPFs result very similar. If p_A is varied (see Fig. 4), then the first difference can be detected by comparing the case of $p_A = 0$ with others: in the null case, the separation between regions in which different classes are associated can be drawn by orthogonal segments connected by an angle, while in the other cases, the segments are connected by a curve, as can be seen by comparing Fig. 4 (a) and (c) (top views). Another significant variation happens for values of $p_A < 2$, different from 0 and 1, i.e. CPPFs flatten around the probability of 0.5, as can be seen in Fig. 4 (b). The evaluation of performances reveals that only varying p_A has significant influence on results, and in particular, if it assumes values of 0, 1 and $+\infty$ the best results are obtained, as can be seen by Table I.A. Therefore, since p_F and p_I can be chosen equal to 0, 1 or $+\infty$, in order to have simple norms, the choices of 1 or $+\infty$ can be done for p_A in order to improve performances and obtain curve class separation. For the following runs, the same settings of the reference method are chosen.

If soft norms are used as in (13) and (14), by varying parameters α_F, α_I and α_A as explained in Section II.C, then the class posterior probabilities are mainly influenced by the variation of α_F and α_I, as can be seen in Fig. 5. In particular, surfaces flatten when these two parameters decrease. On the other hand, in Table I.A it can be seen that the best performances are obtained when soft norms are not used. If a soft norm is used for aggregation, no significant effect is gained, but a complication is added. Therefore, for the following runs, $\alpha_F = \alpha_I = \alpha_A = 1$.

In Fig. 6 (top views), the axes rotation due to the variation of D in (15) from -0.2 to 0.2 is shown. From the analysis of results, a value of $D = -0.12$ was found to minimize *CE*, while a value of $D = -0.08$ minimizes *SCE*, as reported in Table I.B. This is in accordance with the observation that original variables are not necessarily the best, and the use of their linear combination can often improve system performances. However, in this case the improvements are not great enough to justify the great loss of comprehensibility due to the use of linear combination of variables in the rule base, especially for applications like medicine, where interpretability has fundamental importance. In the following, original variables are kept.

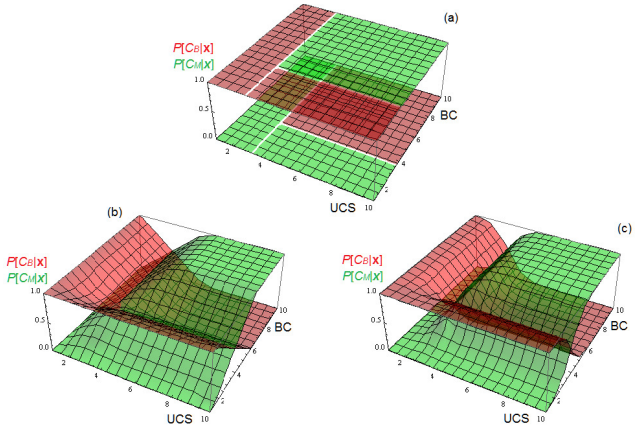


Fig. 3 Functions of class posterior probabilities by using (a) binary, (b) trapezoidal and (c) sigmoidal MFs.

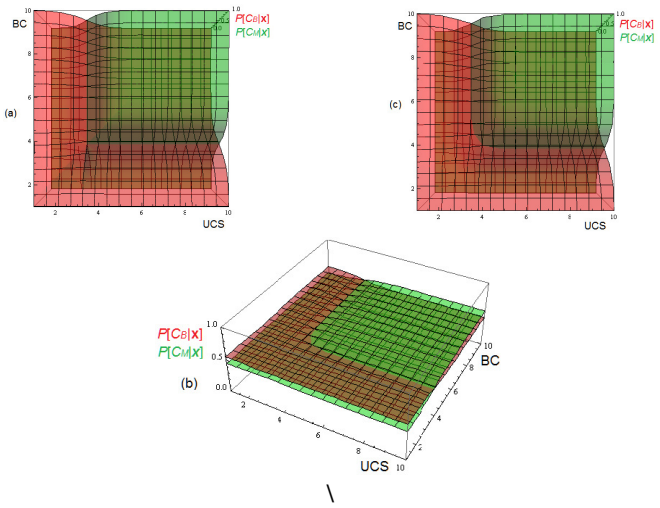


Fig. 4 Functions of class posterior probabilities by using (a) $p_A = 0$, (b) $p_A = 0.5$ and (c) $p_A = 1$.

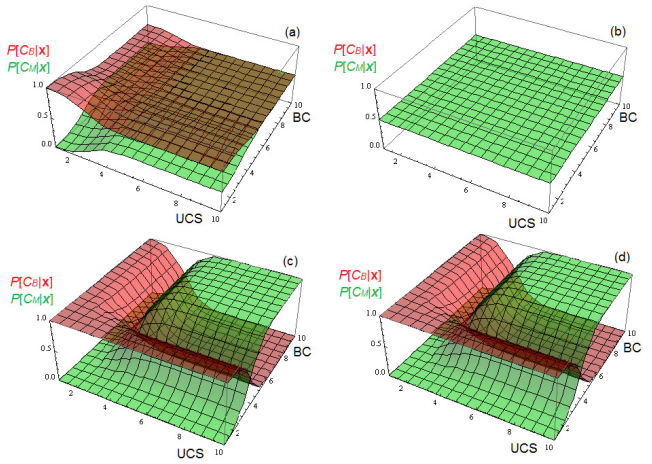


Fig. 5 Functions of class posterior probabilities by using soft norms: (a) $\alpha_F = 0$, $\alpha_I = 1$, $\alpha_A = 1$, (b) $\alpha_F = 1$, $\alpha_I = 0$, $\alpha_A = 1$, (c) $\alpha_F = 1$, $\alpha_I = 1$, $\alpha_A = 0$, (d) $\alpha_F = 1$, $\alpha_I = 1$, $\alpha_A = 1$.

ISBN: 978-1-61804-240-8

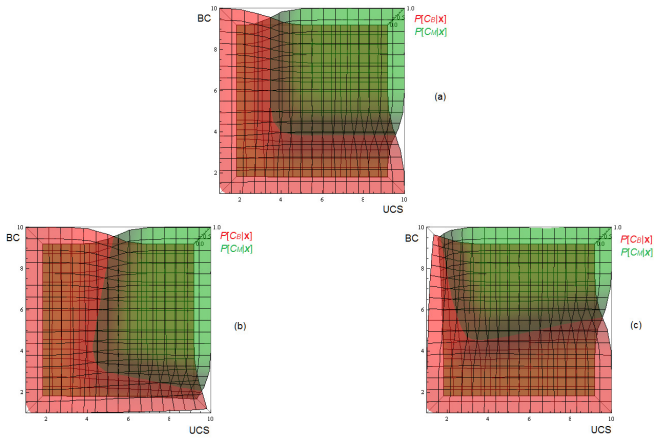


Fig. 6 Axes rotation: (a) $D=0$, (b) $D=-0.2$, (c) $D=0.2$.

If different weights are assigned to the antecedents of the rules, the CPPFs change the amount of dependence from different variables, as can be seen in Fig. 7: at the limit of some weight equal to 0, the dependence from the corresponding variable disappears; in this case, the surfaces approach the value of 0.5 in the region where there are two different rules in which the antecedent with weight 0 changes, the antecedent with weight 1 is the same, and the conclusions are opposite. Even if the use of these weights adds a meaningful complication to the interpretability of the system, it could be an advantage for performances in some applications. However, in the examined case, the optimal settings correspond to not applying them, as can be seen in Fig. 8. Therefore, antecedent weights are not applied in the following.

TABLE I.A
PERFORMANCES OF DIFFERENT SYSTEMS

	Settings										
	Reference	Shapes			Norm types				Soft norms		
Shape: Binary=B Trapezoidal=T Sigmoidal=S	S	B	T	S	S	S	S	S	S	S	S
p_F	1	0	0	0	$[0,2]$	$+\infty$	1	1	1	1	1
p_I	1	0	0	0	$[0,+\infty]$	0	1	1	1	1	1
p_A	$+\infty$	0	0	0	$\{0,1\}$	0	$]0,2]$	$+\infty$	$+\infty$	$+\infty$	$+\infty$
α_F	1	1	1	1	1	1	1	1	0	1	1
α_I	1	1	1	1	1	1	1	1	1	0	1
α_A	1	1	1	1	1	1	1	1	1	1	0
D	0	0	0	0	0	0	0	0	0	0	0
w_{UCS}	1	1	1	1	1	1	1	1	1	1	1
w_{BC}	1	1	1	1	1	1	1	1	1	1	1
$W_{\{low,low\}}$	0.35	1	1	1	1	1	1	1	1	1	1
$W_{\{low,high\}}$	0.02	1	1	1	1	1	1	1	1	1	1
$W_{\{high,low\}}$	0.04	1	1	1	1	1	1	1	1	1	1
$W_{\{high,high\}}$	0.59	1	1	1	1	1	1	1	1	1	1
$p_{\{low,low\}-B}$	1	1	1	1	1	1	1	1	1	1	1
$p_{\{low,low\}-M}$	0	0	0	0	0	0	0	0	0	0	0
$p_{\{low,high\}-B}$	1	1	1	1	1	1	1	1	1	1	1
$p_{\{low,high\}-M}$	0	0	0	0	0	0	0	0	0	0	0
$p_{\{high,low\}-B}$	0.66	1	1	1	1	1	1	1	1	1	1
$p_{\{high,low\}-M}$	0.34	0	0	0	0	0	0	0	0	0	0
$p_{\{high,high\}-B}$	0	0	0	0	0	0	0	0	0	0	0
$p_{\{high,high\}-M}$	1	1	1	1	1	1	1	1	1	1	1
	Performances										
CE	0.047	0.112	0.112	0.112	0.112	0.112	0.112	0.112	0.345	-	0.112
SCE	0.039	0.112	0.069	0.077	$[0.077,0.079]$	0.084	$[0.079,0.248]$	0.080	0.117	0.250	0.080

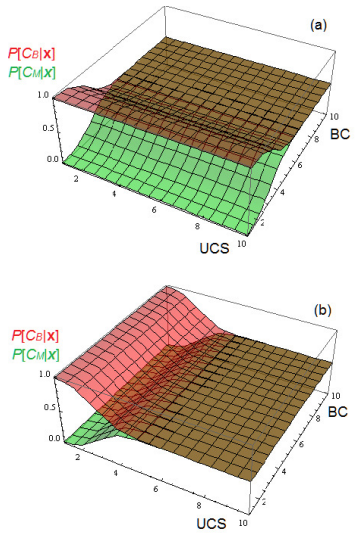


Fig. 7 Antecedent weights: (a) $w_{UCS}=0$ and $w_{BC}=1$, (b) $w_{UCS}=1$ and $w_{BC}=0$.

Rule weights and conclusions have been already optimized by means of the reference method. The influence on CPPFs of each one and both of them is shown in Fig. 9. Starting with parameters described above, it can be noticed that the application of optimized rule conclusions scales the CPPFs, therefore the interclass separation only slightly changes, but the height of CPPFs changes instead; on the other hand, if rule weights are applied, then the regions ascribed to different classes are significantly changed. As a consequence, as can be seen in Table I.B, the use of optimized rule conclusions does not improve much *CE*, but improves *SCE* significantly if rule weights are not applied. The application of rule weights significantly improves both *CE* and *SCE*, and in this case using optimized conclusions gives a further improvement. Therefore, the use of rule weights and conclusions in the form of different classes with respective probabilities, in particular of rule weights, can be very advantageous; however, the choice depends on the relative importance given to performances and interpretability.

TABLE I.B
PERFORMANCES OF DIFFERENT SYSTEMS

	Settings							
	Axes rotation			Antecedent weights		Rule weights	Rule conclusions	Reference
Shape: Binary=B Trapezoidal=T Sigmoidal=S	S	S	S	S	S	S	S	S
p_F	1	1	1	1	1	1	1	1
p_I	1	1	1	1	1	1	1	1
p_A	+∞	+∞	+∞	+∞	+∞	+∞	+∞	+∞
α_F	1	1	1	1	1	1	1	1
α_I	1	1	1	1	1	1	1	1
α_A	1	1	1	1	1	1	1	1
D	0	-0.12	-0.08	0	0	0	0	0
w_{UCS}	1	1	1	0	1	1	1	1
w_{BC}	1	1	1	1	0	1	1	1
$W_{\{low,low\}}$	1	1	1	1	1	0.35	1	0.35
$W_{\{low,high\}}$	1	1	1	1	1	0.02	1	0.02
$W_{\{high,low\}}$	1	1	1	1	1	0.04	1	0.04
$W_{\{high,high\}}$	1	1	1	1	1	0.59	1	0.59
$p_{\{low,low\}-B}$	1	1	1	1	1	1	1	1
$p_{\{low,low\}-M}$	0	0	0	0	0	0	0	0
$p_{\{low,high\}-B}$	1	1	1	1	1	1	1	1
$p_{\{low,high\}-M}$	0	0	0	0	0	0	0	0
$p_{\{high,low\}-B}$	1	1	1	1	1	1	0.66	0.66
$p_{\{high,low\}-M}$	0	0	0	0	0	0	0.34	0.34
$p_{\{high,high\}-B}$	0	0	0	0	0	0	0	0
$p_{\{high,high\}-M}$	1	1	1	1	1	1	1	1
	Performances							
CE	0.112	0.102	0.103	0.345	0.345	0.052	0.112	0.047
SCE	0.080	0.077	0.076	0.135	0.117	0.041	0.060	0.039

ISBN: 978-1-61804-240-8

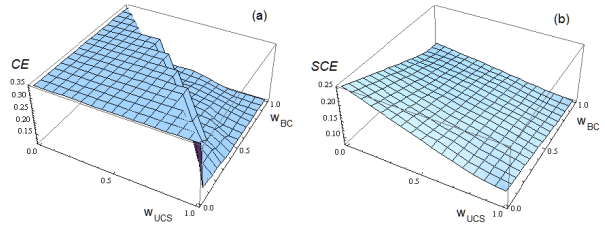


Fig. 8 (a) *CE* and (b) *SCE* as antecedent weights are varied.

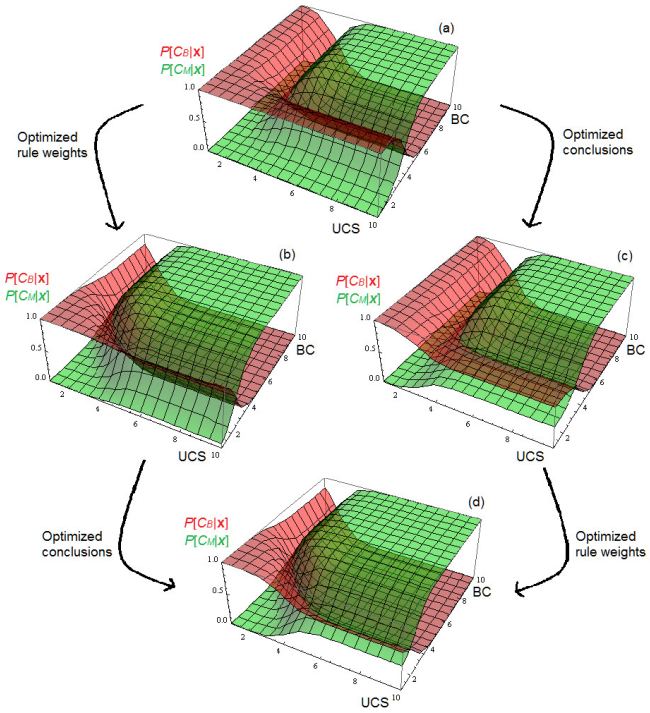


Fig. 9 (a) No rule weights and single class conclusions; (b) optimized rule weights; (c) optimized rule conclusions; (d) reference method.

IV. Conclusion

Various degrees of freedom which characterize the modeling of a rule-based fuzzy system for classification were analyzed, with the aim of individuating the possible settings and complications of a simple system which can improve the classification performances, taking into account the loss of interpretability connected with them. The following conclusions can be drawn: *i)* the trapezoidal shape of fuzzy sets is preferable with respect to others frequently used, such as triangular and Gaussian, and the sigmoidal shape adds derivability with the same number of parameters; *ii)* simple norms, like Mamdani (not for aggregation), Product, and Łukasiewicz ones can be adopted, while the use of soft norms should be avoided; *iii)* axes rotation can be advantageous, but the performance improvement can be small with respect to the loss of comprehensibility; *iv)* using weights for rule antecedents can be advantageous, but not in the examined case, however the associated loss of interpretability is significant; *v)* the use of rule weights strongly improves the system performances, therefore it is recommended, even if interpretability somehow decreases at the same time; *vi)* the definition of rule conclusions in the form of different classes with respective probabilities furnishes an improvement with respect to single class conclusions, with an associated slight increase of complexity. A recently developed method [5] results in accordance with these conclusions, and allows to obtain the best performances with respect to all the other cases.

References

[1] N. Gupta, S. K. Jain, "Comparative analysis of fuzzy power system stabilizer using different membership functions," *International Journal of Computer and Electrical Engineering*, vol. 2 (2), pp. 262-267, 2010.

[2] J. G. Monicka, N. Sekhar, K. R. Kumar, "Performance evaluation of membership functions on fuzzy logic controlled AC voltage controller for speed control of induction motor drive," *International Journal of Computer Applications*, vol. 13 (5), 2011.

[3] J. Zhao, B. Bose, "Evaluation of membership functions for fuzzy logic controlled induction motor drive," in *Proc. 28th IEEE Annual Conference of the Industrial Electronics Society*, Svilla, Spain, November 2002, pp. 229-234.

[4] L. Rutkowski, K. Cpalka, Flexible Neuro-fuzzy systems, *IEEE Transactions on Neural Networks*, vol. 14 (3), pp. 554-574, 2003.

[5] M. Pota, M. Esposito, G. De Pietro, "Combination of Interpretable Fuzzy Models and Probabilistic Inference in Medical DSSs," in *Proc. 8th International Conference on Knowledge, In-formation and Creativity Support Systems*, Krakow, Poland, 2013, pp. 541-552.

[6] K. Bache, M. Lichman, UCI Machine Learning Repository, 2013. <http://archive.ics.uci.edu/ml>. Irvine, CA: University of California, School of Information and Computer Science.

[7] P. Y. Glorennec, *Algorithmes d'apprentissage pour systèmes d'inférence floue*, Editions Hermès, Paris, 1999.

[8] E. H. Mamdani, S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies*, vol. 7 (1), pp. 1-13, 1975.

[9] M. Sugeno, *Industrial applications of fuzzy control*, Elsevier Science Pub. Co., 1985.

[10] M. Pota, M. Esposito, G. De Pietro, "Fuzzy partitioning for clinical DSSs using statistical information transformed into possibility-based knowledge," *Knowledge-Based Systems*, 2014, http://dx.doi.org/10.1016/j.knosys.2014.06.021

[11] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338-353, 1965.

[12] T. E. Rothenfluh, K. Bögl, K.-P. Adlassnig, "Representation and Acquisition of Knowledge for a Fuzzy Medical Consultation System," in *Studies in Fuzziness and Soft Computing*, P. S. Szczepaniak, P. J. G. Lisboa, J. Kacprzyk, Eds., vol. 41, Fuzzy systems in Medicine, Physica-Verlag HD, 2000, pp. 636-651.

[13] J .M. Alonso, C. Castiello, M. Lucarelli, C. Mencar, "Modelling interpretable fuzzy rule-based classifiers for Medical Decision Support," in *Medical Applications of Intelligent Data Analysis: Research advancements*, R. Magdalena, E. Soria, J. Guerrero, J. Gómez-Sanchis, A.J. Serrano, Eds., IGI Global, 2012, pp. 255-272.

[14] M. Frank, "On the simultaneous associativity of $F(x,y)$ and $x+y-F(x,y)$," *Aequationes Math.*, vol. 19, pp. 141-160, 1979.

[15] R. R. Yager, D. P. Filev, *Essentials of Fuzzy Modeling and Control*, New York: Wiley, 1994.

[16] M. G. Kendall, *A course in Multivariate Analysis*, London: Griffin, 1957.

[17] A. Hōskuldsson, "PLS regression methods," *Journal of Chemometrics*, vol. 2 (3), pp. 211-228, 1988.

[18] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7 (2), pp. 179-188, 1936.

[19] R. C. Rao, "The utilization of multiple measurements in problems of biological classification," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 10 (2), pp. 159-203, 1948.

Research method of energy-optimal spacecraft control during interorbital maneuvers

Sokolov N.L.
Mission Control Centre
Central Research Institute of Machine Building (FSUE/TSNIIMASH)
4, Pionerskaya St., Korolev, Moscow region, 141070
RUSSIAN FEDERATION
sokolov@mcc.rsa.ru

Abstract **-** This work is focused on the research of energy optimal scheme of SC thrust vector control during the interorbital maneuvers.

The methodological novelty of the proposed decision is development of transformation algorithms for initial systems of differential equations and connection formulae between the unknown parameters of SC motion and conjugate variables. It helps to avoid the above-mentioned difficulties. Thus, the algorithm is proposed for analytical decision of differential equations for the conjugate variables which provides the possibility to determine the structure of optimal control avoiding the complicated calculation procedures. Besides the proof is given of the Hamiltonian identical equality to zero and the conjugate variable corresponding to a subsatellite-point longitude along all the flightpath. It makes possible to elaborate additional dependencies, connecting the unknown parameters in the initial point of trajectory. Along with the use of transversality conditions in the boundary points of trajectories these transformations also allow reducing the multi-parameter boundary problem of the SC optimal trajectory search to the two-parameter one, which provides high level of efficiency during calculation.

The structure of thrust vector optimal control is determined. The fact is proved that for a wide range of boundary conditions, weight and power characteristics of SC the maximum number of SC engine burns equals two. At the first burn the SC is transferred from the initial orbit to the interim one, which has a cross point with the final orbit, where velocity and radius-vector of SC are respectively equal to the specified values. At the second burn, in the cross-point of orbits, the trajectory angle is adjusted.

On the whole the proposed methodological approach can be used as the basis of wide range of tasks aimed at optimization of interorbital flights and corrections and it can be introduced for the planning of perspective missions of near and outer space.

Key words-Spacecraft, interorbital maneuvers, minimization of energy consumption, transform algorithms, accelerated calculation algorithm.

Sokolov Nikolay Leonidovich, Candidate of Science (Engineering), senior researcher, Deputy Head of the Mission Control Centre of Federal State Unitary Enterprise "The Central Research Institute of Machine Building". E-mail: sokolov@mcc.rsa.ru. Research interests: optimal control, flight dynamics, ballistics, probability theory, mathematical modelling.

I. INTRODUCTION

The control process of spacecraft quite often requires the conducting of dynamic operations. Basing on the SC operation experience one can see the need for the space debris avoidance maneuvers from time to time and correction for maintaining of orbital parameters. Besides, the powered maneuvers are conducted during SC deorbiting which is the final stage for the most SC.

At the same time, all the mentioned tasks connected with conducting of powered maneuvers are focused on the research for energy optimal schemes of SC control.

The task solution of SC optimal control entails considerable estimate time and difficulties of computational architecture. That's why it would be reasonable to use quasioptimal algorithms of variational task solution which meet the requirements due to their effective operation speed: the limit of errors of the algorithm output must not exceed several per cent with the total qualitative coincidence with the results of optimal control tasks.

II. STATEMENT OF THE PROBLEM

The motion of SC is described by differential equations, which is a special case of the system, described in the works [1,2], without regard to aerodynamic forces, enabling the SC to maneuver in the atmosphere, as well as centrifugal and Coriolis forces:

$$\begin{aligned}
\frac{dV}{dt} &= -\frac{\rho V^2}{2P_x} - \frac{\mu}{r^2}\sin\theta + \frac{P}{m}\cos\alpha\cos\beta, \\
\frac{d\theta}{dt} &= -\frac{\mu}{r^2 V}\cos\theta + \frac{V}{r}\cos\theta + \frac{P}{mV}\sin\alpha\cos\beta, \\
\frac{d\varepsilon}{dt} &= -\frac{V}{r}\cos\theta\cos\varepsilon\, tg\varphi + \frac{P}{mV\cos\theta}\sin\beta, \\
\frac{dr}{dt} &= V\sin\theta, \qquad \frac{d\lambda}{dt} = \frac{V}{r}\frac{\cos\theta\cos\varepsilon}{\cos\varphi}, \\
\frac{d\varphi}{dt} &= \frac{V}{r}\cos\theta\sin\varepsilon, \qquad \frac{dm}{dt} = -\frac{P}{P_{spec}\mathrm{g}_E},
\end{aligned} \tag{1}$$

where V– spacecraft velocity, θ– burnout angle, ε– angle between projection of velocity onto the local horizon and local parallel, r– radius vector connecting Earth's centre and spacecraft centre of mass, λ and φ – geographic longitude and latitude respectively, m – spacecraft mass, ρ – atmospheric density, μ – product of gravitational constant by Earth's mass, P_x – front surface reduced load, C_x – aerodynamic drag coefficient, P – engine thrust, P_{spec} –

ISBN: 978-1-61804-240-8

specific thrust, g_E – gravitational acceleration on earth's surface, α – angle between thrust vector projection on a motion plane and spacecraft velocity vector, β – angle between thrust vector and SC motion plane.

Besides, in case of SC unperturbed motion with the switched off engine the correlation is true between the course angle ε and latitude φ

$$\cos\varepsilon\cos\varphi = C. \qquad (2)$$

The SC was controlled by change of propulsive efforts, characterized by the thrust value P and its orientation relative to SC velocity vector α and β:

$$0 \le P \le P_{mix}, \quad -\pi \le \alpha \le \pi, \quad -\pi \le \beta \le \pi. \quad (3)$$

The SC initial state was determined by the orbital parameters of earth satellite vehicle and it's mass in the fixed moment of time t_o:

$$V_0 = V(t_0), \quad \theta_0 = 0, \quad \varepsilon_0 = \varepsilon(t_0),$$
$$r_0 = r(t_0), \quad \lambda_0 = \lambda(t_0),$$
$$\varphi_0 = \varphi(t_0), \quad m_0 = m(t_0). \qquad (4)$$

End of trajectory is the point on the earth surface (h_R=0) with the specified geographic coordinates

$$\lambda_R = \lambda(h_R), \quad \varphi_R = \varphi(h_R). \qquad (5)$$

The intermediate conditions were also taken into account: velocity and burnout angle at the moment when SC reaches the atmospheric conditional boundary were set (h_{end}=100 km):

$$V_{ent} = V(h_{ent}), \quad \theta_e = \theta(h_{ent}). \qquad (6)$$

The research for the development of approximately optimal control algorithms for SC were based on the optimal control theory: for a SC which motion is described by the equations system (1) and the relation (2), it is necessary to find the control laws P(t), α(t), β(t), providing extremum of functional $J = \Delta m_F = m_0 - m_f - \min$ with the constraints (3), edge (4), (5) and intermediate (6) conditions.

The peculiarity of the given problem solution as compared to the SC maneuver tasks in the atmosphere is that the control process finishes before SC re-entry, i.e. the landing point deviation from the specified one is fully determined by the SC state vector at the moment of engines' closedown. It allows performing the optimal control task only in the extra-atmospheric phase, considering that the propulsion system is switched on at the initial moment of time: $P_o=P_{max}$, and the end point is determined by the conditions (6) and lateral displacement of the re-entry point L_{ent} in relation to the plane of initial orbit. The value L_{ent} is calculated depending on the orbital parameters (4) and geographic coordinates λ_R and φ_R.

III. OPTIMALITY CONDITIONS

We used maximum principle of Pontryagin in order to solve the variational task [3].

We introduce the Hamiltonian

$$H = PF_1 + F_2, \qquad (7)$$

where

$$F_1 = \frac{\psi_1}{m}\cos\alpha\cos\beta + \frac{\psi_2}{mV}\sin\alpha\cos\beta + \frac{\psi_3}{mV\cos\theta}\sin\beta - \frac{\psi_7}{P_{spec}g_E},$$

$$F_2 = -\frac{\mu}{r^2}\sin\theta\psi_1 - \frac{\mu}{r^2V}\cos\theta\psi_2 + \frac{V}{r}\cos\theta\psi_2 - \frac{V}{r}\cos\theta\cos\varepsilon\, tg\varphi\psi_3 + V\sin\theta\psi_4 + \frac{V}{r}\frac{\cos\theta\cos\varepsilon}{\cos\varphi}\psi_5 + \frac{V}{r}\cos\theta\sin\varepsilon\psi_6.$$

The conjugate variables φ_i (i=1,2,…,7) are determined by the following relations:

$$\frac{d\psi_1}{dt} = -\frac{\partial H}{\partial V} = \frac{\rho V\psi_1}{P_x} - \frac{\mu}{r^2V^2}\cos\theta\psi_2 - \frac{1}{r}\cos\theta\psi_2 + \frac{P}{mV^2}\sin\alpha\cos\beta\psi_2 + \frac{1}{r}\cos\theta\cos\varepsilon\, tg\varphi\psi_3 + \frac{P}{mV^2\cos\theta}\sin\beta\psi_3 - \sin\theta\psi_4 - \frac{1}{r}\frac{\cos\theta\cos\varepsilon}{\cos\varphi}\psi_5 - \frac{1}{r}\cos\theta\sin\varepsilon\psi_6,$$

$$\frac{d\psi_2}{dt} = -\frac{\partial H}{\partial\theta} = \frac{\mu}{r^2}\cos\theta\psi_1 - \frac{\mu}{r^2V}\sin\theta\psi_2 + \frac{V}{r}\sin\theta\psi_2 - \frac{V}{r}\sin\theta\cos\varepsilon\, tg\varphi\psi_3 - \frac{P}{mV\cos^2\theta}\sin\theta\sin\beta\psi_3 - V\cos\theta\psi_4 + \frac{V}{r}\frac{\sin\theta\cos\varepsilon}{\cos\varphi}\psi_5 + \frac{V}{r}\sin\theta\sin\varepsilon\psi_6, \qquad (8)$$

$$\frac{d\psi_3}{dt} = -\frac{\partial H}{\partial\varepsilon} = -\frac{V}{r}\cos\theta\sin\varepsilon\, tg\varphi\psi_3 + \frac{V}{r}\frac{\cos\theta\sin\varepsilon}{\cos\varphi}\psi_5 - \frac{V}{r}\cos\theta\cos\varepsilon\psi_6,$$

$$\frac{d\psi_4}{dt} = -\frac{\partial H}{\partial r} = -\frac{2\mu}{r^3}\sin\theta\psi_1 - \frac{2\mu}{r^3V}\cos\theta\psi_2 + \frac{V}{r^2}\cos\theta\psi_2 - \frac{V}{r^2}\cos\theta\cos\varepsilon tg\varphi\psi_3 + \frac{V}{r^2}\frac{\cos\theta\cos\varepsilon}{\cos\varphi}\psi_5 + \frac{V}{r^2}\cos\theta\sin\varepsilon\psi_6,$$

$$\frac{d\psi_5}{dt} = -\frac{\partial H}{\partial\lambda} = 0,$$

$$\frac{d\psi_6}{dt} = -\frac{\partial H}{\partial\varphi} = \frac{V}{r}\frac{\cos\theta\cos\varepsilon}{\cos^2\varphi}\psi_3 - \frac{V}{r}\frac{\cos\theta\cos\varepsilon\sin\varphi}{\cos^2\varphi}\psi_5,$$

$$\frac{d\psi_7}{dt} = -\frac{\partial H}{\partial m} = \frac{P}{m^2}\cos\alpha\cos\beta\psi_1 +$$

$$+\frac{P}{m^2V}\sin\alpha\cos\beta\psi_2 + \frac{P}{m^2V\cos\theta}\sin\beta\psi_3 .$$

The laws of parameter variation α, β and P with optimal control are determined from the Hamiltonian maximization condition. The relations for calculation of optimal values α and β are obtained from the conditions $\partial H / \partial\alpha =$ and $\partial H / \partial\beta = 0$:

$$tg\alpha = \frac{\psi_2}{V\psi_1}, \qquad (9)$$

$$tg\beta = \frac{\psi_3}{\left(V\psi_1\cos\alpha + \psi_2\sin\alpha\right)\cos\theta} = \frac{\psi_3\cos\alpha}{V\psi_1 cos\theta}. \qquad (10)$$

With a help of inequation $\partial^2 H / \partial\alpha^2 < 0,\ \partial^2 H / \partial\beta^2 < 0$ we will establish the membership of angles α and β to one of the two quadrants:

$$\cos\alpha = sign\psi_1 , \qquad (11)$$

$$\cos\beta = sign\left(\psi_1\cos\alpha + \frac{\psi_2}{V}\sin\alpha\right) =$$

$$= sign\left(\frac{\psi_1}{\cos\alpha}\right) . \qquad (12)$$

The engine thrust possesses the boundary values:

$$P = P_{max} \quad \text{with } F_1 > 0 \text{ and}$$

$$P = 0 \qquad \text{with } F_1 < 0 \qquad (13)$$

Let's prove that there are no more than two powered flight phases.

The expression F_1 has a switchover function in the thrust optimal control. In order to determine the number of burns it is necessary to examine the function F_1, which, according to [4], is calculated from the equation V_{ent} and L_{ent},

$$F_1 = \frac{1}{m}(\psi_1\cos\alpha\cos\beta + \frac{\psi_2}{V}\sin\alpha cos\beta +$$

$$+\frac{\psi_3}{V\cos\theta}\sin\beta)$$

Considering the formulae (9) and (10) we get

$$F_1 = \frac{1}{m}\left(\frac{\psi_1}{\cos\alpha\cos\beta}\right). \qquad (14)$$

IV. ANALYTIC ALGORITHM OF OPTIMAL MANEUVER CALCULATION

In order to investigate the behavior of curve F_1 we transform the mathematical model of SC motion, as the equation (14) cannot be solved analytically. Let's consider that SC flight with the running engine is determined only by active forces and during the coast flight – only by gravitational forces. Suppose that angles α and β, as well as trajectory angle θ change insignificantly during the powered flight. Then the differential equations for conjugate variable ψ_1, influencing the function F_1, take the form:

$$\frac{d\psi_1}{dt} = \frac{P}{mV^2}\sin\alpha\cos\beta\psi_2 +$$

$$+\frac{P}{mV^2}\sin\beta\psi_3 \qquad \text{with } P \neq 0,$$

$$\frac{d\psi_1}{dt} = -\frac{\mu}{r^2V^2}\psi_2 - \frac{1}{r}\psi_2 \qquad \text{with } P = 0.$$

In contexts of the made assumptions in both cases

$$\psi_2 = 0, \text{ that is } \psi_2 = C^*.$$

We show that $\psi_1(t) \geq 0$ with $t_0 \leq t \leq t_R$. The equation for ψ_1 with $P \neq 0$ considering (9) and (10) is transformed in the following way:

$$\frac{d\psi_1}{dt} = \frac{P\psi_1}{mV\cos\alpha\cos\beta}\left(sin^2\alpha cos^2\beta + sin^2\beta\right).$$

As the expression $\psi_1 / \cos\alpha$ has the same sign as $\cos\beta$ (q. v. (12)), the inequality is true $\psi_1 \geq 0$ with $P \neq 0$.

The sign of the variable ψ_1 with P=0 depends on the sign of the constant $\psi_2 = C^*$, which is determined from the following considerations. In order to transfer SC from earth satellite orbit to a descending trajectory, angle α in the process of powered flight should be in the range of $-\pi \leq \alpha < -\pi/2$. Then from the equation (11) we obtain that $\psi_1 \leq 0$, and with the help of equation (9) determine that $\psi_2 = C^* \leq 0$. Hence, $\psi_1 \geq 0$ during the coast flight. At that, the analysis of dependence ψ_1 showed that the conjugate variable ψ_1 jumps at the moment of engine thrust P switching. Hence we can conclude that the variable ψ_1 during all the considered flight phase of SC changes its sign no more than two times.

Thus, considering the condition (14) we may conclude that the maximum number of zeros of function F_1, as well as the number of SC engine thrust switchings, equals two. Besides in this case the switchings are made from $P=P_{max}$ to P=0, and then to $P=P_{max}$ again.

Thus, the laws of thrust vector optimal control are characterized by the dependencies (9), (10), (13). In order to obtain the numerical evaluation of optimal trajectory it's necessary to solve the boundary value problem which consists in iteration of 10 N(=10) parameters at the beginning of a trajectory (seven-parameters vector of conjugate variables ψ_{i0} and values of control functions (P_0, α_0, β_0), with which the final edge conditions are fulfilled.

Some boundary values of conjugate variables can be obtained on account of transversability condition [4]:

$$[\left(\psi_1 - 1\right)\delta m - H\delta t + \psi_3\delta\varepsilon +$$

$$+\psi_5\delta\lambda + \psi_6\delta\varphi]_{t_0}^{t_k} = 0. \qquad (15)$$

As the variations δε and δφ with $t=t_0$ are interdependent ones, the transversality condition is fulfilled in case if the following equality occurs

$$\psi_{30}\delta\varepsilon + \psi_{60}\delta\varphi = 0. \qquad (16)$$

On the other hand, the variations δε and δφ are related by

$$\frac{\partial q}{\partial\varepsilon}\delta\varepsilon + \frac{\partial q}{\partial\varphi}\delta\varphi = 0$$

where $q = \cos\varepsilon\cos\varphi - C.$

Therefore the condition (16) will be achieved if the initial values ψ_{30} and ψ_{60} are chosen to realize the relation

$$\left(\frac{\partial q}{\partial \varphi}\right)_0 \psi_{30} = \left(\frac{\partial q}{\partial \varepsilon}\right)_0 \psi_{60} \qquad \text{or}$$

$$\psi_{60} = tg\varphi_0 ctg\varepsilon_0 \psi_{30}. \qquad (17)$$

With $t=t_R$ the equation (15) due to arbitrariness of variations δm, δt, $\delta\lambda$ is possible only if

$$\psi_{7R} = 1, \quad H = 0, \quad \psi_{5R} = 0.$$

As the variables t and λ do not constitute in an explicit form the right parts of the system (1), we can conclude that, the Hamiltonian H and the conjugate variable ψ_5 are identically equal to zero:

$$H \equiv 0, \quad \psi_5 \equiv 0. \qquad (18)$$

The equations (9), (10), (17), (18) and the equality defined in the problem statement $P(t_0) = P_{max}$ provide six constraint equations between ten initial parameters. The deficient four equations necessary for initial estimate of boundary problem solution can be got introducing the assumptions about vanishing of switching function at the initial instant and impulse character of engines' operation. From the first assumption follows:

$$\frac{\psi_{10}}{m_0}\cos\alpha_0 cos\beta_0 + \frac{\psi_{20}}{m_0 V_0}\sin\alpha_0 cos\beta_0 +$$

$$+\frac{\psi_{30}}{m_0 V_0 \cos\theta_0} sin\beta_0 - \frac{\psi_{70}}{P_{spec} g_E} = 0. \qquad (19)$$

In view of equation (18) and (19) we get:

$$-\frac{\mu}{r_0^2}\sin\theta_0\psi_{10} - \frac{\mu}{r_0^2 V_0}\cos\theta_0\psi_{20} + \frac{V_0}{r_0}\cos\theta_0\psi_{20} -$$

$$-\frac{V_0}{r_0} cos\theta_0 cos\varepsilon_0 tg\varphi_0\psi_{30} + V_0\sin\theta_0\psi_{40} +$$

$$+\frac{V_0}{r_0}\frac{\cos\theta_0 cos\varepsilon_0}{\cos\varphi_0}\psi_{50} + \frac{V_0}{r_0}\cos\theta_0 sin\varepsilon_0\psi_{60} = 0. \quad (20)$$

Using the second equation and solving the Keplerian motion equations [5], we may determine initial orientation angles of thrust vector α_0 and β_0 at the moment of SC reentry:

$$\alpha_0 = -\pi + \arcsin\left(V_1 \sin\frac{\theta_1}{\Delta V}\right). \qquad (21)$$

where

$$V_0 = \sqrt{\frac{\mu}{r_0}}, \qquad V_1 = \sqrt{V_{ent}^2 + \frac{2\mu}{r_0} - \frac{2\mu}{r_{ent}}},$$

$$r_0 = r_1 = R_E + h_{0,}$$

$$\cos\theta_1 = \frac{r_{ent} V_{ent}\cos\theta_{ent}}{r_1 V_1},$$

$$\Delta V = \sqrt{V_0^2 + V_1^2 - 2V_0 V_1\cos\theta_1},$$

$$r_{ent} = R_E + h_{ent,}$$

$$\beta_0 = arctg\left\{\frac{L_{ent}}{R_E\left[\arcsin A(r_{ent}) - \arcsin A(r_0)\right]}\right\}, \quad (22)$$

where

$$A(r) = \frac{\mu - C_2/r}{\sqrt{\mu^2 - C_1 C_2}},$$

$$C_1 = \frac{2\mu}{r_{ent}} - V_{ent}^2, \qquad C_2 = r_{ent}^2 V_{ent}^2 cos^2\theta_{ent}.$$

If the single-burn SC transfer to the finite point with the coordinates $r=r_{ent}$, $V=V_{ent}$, $\theta=\theta_{ent}$ and $L_L=L_{ent}$ the two-burn transfer is considered. The first burn of the value ΔV_1 with $\alpha_0=\pi$ provides SC reentry with the specified values V_{ent} and L_{ent}, and with the help of the second burn ΔV_2, given with $r = r_{ent}$ the final value θ is adjusted. At that the initial value of angle β_0 is always determined by the formula (22).

Thus, the mentioned analytical dependencies enable us to find the first approximation for the boundary task solution.

V. PERFORMANCE ANALYSIS OF MANEUVER SCHEMES

The conducted analysis of numerical results reveals the structure of thrust vector optimal control with minimum of active mass. The calculations were made with variation of altitude h_0 and inclinations i_0 of earth satellite circular orbits, initial SC mass m_0, reduced frontal surface load P_x, thrust P and specific thrust P_{spec} of the propulsion device, spacecraft reentry conditions V_{ent}, θ_{ent}, L_{ent} in the range:

$$300 \le h_0 \le 700\,\text{km}, \qquad 40^0 \le i_0 \le 80^0,$$

$$500 \le m_0 \le 2500\,\text{kg}, \qquad 500 \le P_x \le 200\,\frac{\text{kg}}{\text{m}^2},$$

$$1000 \le P \le 5000\,\text{kg}, \quad 250 \le P_{spec} \le 450\,\text{s},$$

$$6{,}5 \le V_{ent} \le 8\,\text{km/s}, \quad -15^0 \le \theta_{ent} \le -5^0,$$

$$0 \le L_{ent} \le 800\,\text{km}. \qquad (23)$$

The following values were used as nominal values of varied parameters:

$$h_0 = 500\,\text{km}, \qquad i_0 = 55^0, \quad m_0 = 2000\,\text{kg},$$

$$P_x = 1000\,\text{kg/m}^2, \quad P = 2000\,\text{kg}, \quad P_{spec} = 320\,\text{s},$$

$$V_{ent} = 7{,}4 \quad km/s, \quad \theta_{ent} = -11^0. \qquad (24)$$

It is showed that for all the variation range of parameters the optimal control consists in two- burn ignition: at the 1st burn the SC is transferred from the satellite orbit to the descending trajectory, at the 2nd burn the reentry parameters are corrected. The thrust vector orientation at the 1st burn consists in the following: angle α is constant and equals ~ 180^0, angle β slightly changes from the initial value β_0, lying in the range from 120^0 to 180^0, depending on the value of lateral displacement L_{ent}, by $\pm 1 \div 2^0$.

The analysis of results provides opportunity to establish general principles of optimal control and derive a non-iterative algorithm on their basis. It involves the use of control programs with the constant thrust vector orientation at 1st burn: $\beta=\beta_0$, where β_0 is calculated by the formula (23).

The finish of 1st burn corresponds to SC velocity to the value providing during the further coasting flight the reentry with the specified velocity V_{ent}. At the 2nd burn, angle $\beta \approx 0$, and angle α either equals 90^0 if there's a need to reduce

angle with the increase of $|\theta_{ent}|$. The start of 2nd burn is chosen so that its finish will correspond to the moment of SC reentry ($h=h_{ent}$=100 km). As one would expect, the duration of the 1st burn is considerably greater than of the 2nd burn.

Application of such algorithm will not lead to the considerable increase in fuel consumption Δm_F in comparison with Δm_{Fmin}: the differences do not exceed 1÷2%. The supposed control will be called suboptimal control.

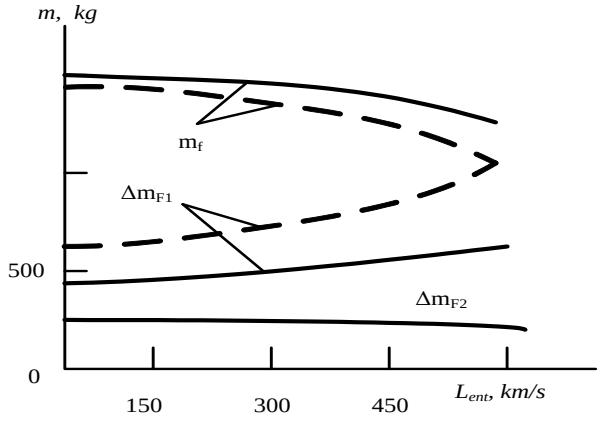


Fig. 1. Dependencies of SC final mass and mass Δm_{F1} and Δm_{F2} on the lateral displacement L_{ent} (the variant of nominal initial data)
Solid lines – suboptimal control;
dashed lines – the one-burn scheme

Figure 1 presents the dependencies of SC finite mass m_f, spent at the 1st and 2nd burns m_{F1} and Δm_{F2} , on the lateral displacement L_{ent}. For comparison there are the results of fuel mass calculation Δm_{F1}, necessary with the use of spacecraft one-burn deorbiting scheme and corresponding to the spacecraft finite mass m_f. Analyzing this data, we should note a considerable efficiency of the proposed deorbiting scheme. Thus, the increase of SC finite mass δm_f is ~ 180 kg.

As we should expect, with approximately optimal control the fuel mass Δm_{F1} monotonically increases with the increase of value L_{ent}: the change of L_{ent} from 0 to 600 km leads to increase of mass Δm_{F1} from 430 to 660 kg. The fuel mass Δm_{F2} changes less: in the same variety range L_{ent} mass Δm_{F1} reduces from 190 to 160 kg. On the whole for nominal values of variable parameters (24) the SC finite mass is ~ 1200-1400 kg.

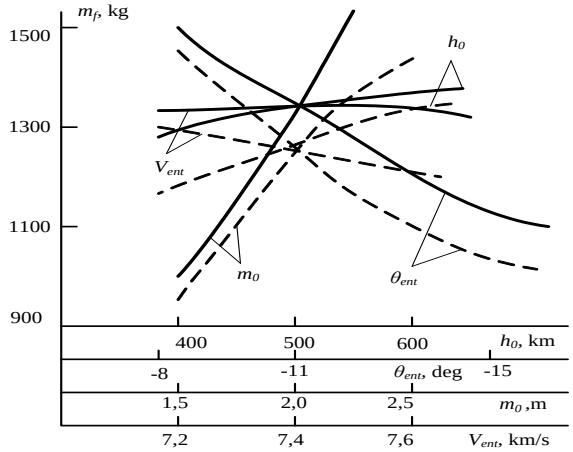


Fig. 2. Dependencies of SC final mass on altitude h_0, initial mass m_0, velocity V_{ent} , trajectory angle θ_{ent}
Solid lines – suboptimal control;
dashed lines – the one-burn scheme

The results presented on Fig. 2, demonstrate specific influence of variable parameters (h_0, m_0, V_{ent}, θ_{ent}) on the finite mass m_f and on the efficiency of mass increase δm_f using two-burn SC deorbiting pattern. Considering them one can see that the finite mass m_f is increasing with the altitude increase of earth satellite orbit h_0, initial mass m_0 and with reducing of SC reentry velocity V_{ent} and absolute value of trajectory angle θ_{ent}. Thus, the change of h_0 from 400 to 600 km leads to increase of m_f from 1265 to 1380 kg, mass change m_0 from 1,5 to 2,3 t leads to increase of m_f from 995 to 1660 kg, reducing of reentry velocity V_{ent} from 7,6 to 7,2 km/s leads to increase of m_f from 1310 to 1325 kg, change of trajectory angle from -15^0 to $-8{,}5^0$ – results in the increase of m_f from 1110 to 1490 kg (L_{ent}=300km).

The power characteristics of engine P and P_{spec}, reduced frontal surface load P_x and inclination of earth satellite orbit i_0 scarcely affect the finite mass of SC.

The efficiency of δ_{mf} increase due to application of the proposed scheme is provided in all the variation range of variable parameters, presented on fig. 2. Besides the high intensity of m_f increase is revealed for greater values m_0, absolute values θ_{ent} and smaller altitudes h_0: δ_{mf} achieves ~ 270 kg.

VI. CONCLUSION

Thus the represented materials show the possibility and big power gain of SC two-burn deorbiting pattern in the wide range of boundary conditions, design, mass and power characteristics of SC and engine. It should be noted that the proposed methodological approach can be applied also for the solution of tasks of thrust vector optimal control during interorbital maneuvers and correction for maintenance of SC orbital parameters in the specified limits.

VII. LIST OF REFERENCES

[1] Avduevsky V.S., Antonov B.M., Anfimov. N.A. and others. The theory of space flight, M.: Mashinostroyeniye, 1972, p.345.

[2] Ivanov N.M., Dmitrievsky A.A., Lysenko L.N. Ballistics and navigation of spacecrafts. M.: Mashinostroyenie. 1986, p.345.

[3] Pontryagin L.S., Boltyanskiy V.P., Gamkrelidze R.V., Mishchenko E.F. Mathematical theory of optimal processes. M.: Science. 1969, p.393.

[4] Letov A.M. Flight dynamics and control. M.: Science. 1969, p.360.

[5] Elyasberg, P. E. Introduction to the theory of flight of artificial earth satellites. M.: Science. 1965, p.537.

Keywords Extraction from Articles' Title for Ontological Purposes

Sylvia Poulimenou, Sofia Stamou, Sozon Papavlasopoulos, and Marios Poulos

***Abstract*—** in this paper, we introduce a novel keyword extraction algorithm, which identifies text descriptive terms in the titles of scientific articles. The identified terms can serve as search keywords for retrieving research articles form digital scientific databases and/or repositories. Our algorithm relies on the vector space model to represent the article titles and treats every term in a title as a vector with each vector containing three features namely, the number of characters a term has, the importance or else the strength of the term in the title and the relative order or else the position of the term in the title. Based on the weight our algorithm computes for each of the above features it assigns every term in the article's title with a score indicating the suitability of that term as search keyword that could retrieve its corresponding article in the top ranking position. The experimental evaluation of our proposed algorithm on real scientific data proves its effectiveness in detecting text descriptive keywords and verifies our assumption that in the case of scientific publications title terms are expressive of the articles'.

***Keywords*—** TFIDF, Vector Space Model, Data Mining, Text Analysis, Distance Measure.

I. Introduction

Keyword-based approach is user friendly and easy to apply with an acceptable retrieval precision, while semantically rich ontology addresses the need for complete descriptions of text retrieval and improves the precision of retrieval [1]. Keyword extraction is a significant technique for text retrieval, Web page retrieval, text clustering, summarization, text mining, and so on. By extracting appropriate keywords, we can easily choose which document to read to learn the relationship among documents [2]. A trivial algorithm for indexing is the unsupervised Term Frequency Inverse Document Frequency (TFIDF) measure [3], which extracts keywords that perform frequently in a text, but that don't perform frequently in the rest of the corpus [4,5]. It is computationally efficient and performs reasonably well [6]. The term "keyword extraction" is used in the context of text mining [3]. Keyword extraction has also been treated as a supervised learning problem [4, 5, 7], where a classifier is used to classify candidate words into positive or negative instances using a set of features. Other research for keyword extraction has also taken advantage of semantic resources [8], Web-based metric, such as PMI score (point-wise mutual information) [7], or graph-based algorithms (e.g., [9] that attempted to use a reinforcement approach to do keyword extraction and summarization simultaneously) [6].

Relative scientific work has been carried out, where Frequent Terms (FT) are considered crucial [10], thus are extracted first, in order to extract keywords from a single document. Another interesting research has been carried out, where an alternative approach occurs. A proposed algorithm [11], where not only term frequency is measured or other statistics, but in addition the usage of professional indexers is implemented.

Furthermore, TFIDF weighting measure is combined with the vector space method for this purpose [12]. The TFIDF very relevant for the document is commonly used in IR to compare a query vector with a document vector using a similarity or distance function such as the cosine similarity function. However, the problem focuses on the disharmony with the Shannon-like theory [13]. More details are given in section *II.A*. The algorithm proposed in this paper can be considered novel, since it is not based not on the frequency where a term occurs (TF/IDF) which considers every word in a document equally weighted, but it inserts 3 variables that constitute each word uniquely identified. In a correlation between the two algorithms, in TF/IDF algorithm the angle of a word vector and the component vector represents the term frequency when in the proposed angle it represents 3 unique variables that provide a unique vector identity for each word.

It is known, that the Ontologies are related to a model of knowledge, and knowledge in turn to information. Thus, it makes sense to introduce the concept of entropy and mutual information, as defined by Shannon for information theory in [14], on ontologies. Entropy and mutual information in turn enables us to define a distance measure formally. With this distance a sound foundation is given for the capturing of the inherent structure of ontology. Consequently, in our work, we attempt to use a new algorithm which is creating in the philosophy of TFIDF algorithm giving simultaneously a solution to Shannon-like model disharmony.

This paper is organized as follows:

- Section II provides the theoretical and practical implications of this study which includes the related TF/IDF algorithm, the basis of the algorithm and the ranking criterion.
- Section III describes the experimental design analyzed the data collection part, the implementation of the algorithm as well as the retrieval and evaluation results
- Section IV presents the discussion and the future plan.

II. Methodology

A. Vector Space Model

The TF/IDF algorithm is the best known weighting scheme for terms in information retrieval until know. This is based on three components

- Document Frequency (DF)
- Inverse Document Frequency (IDF)
- Term Frequency (TF)

And is implemented by the formula

TF / IDF = Term _ Frequency _ X _ Inverse _ Document _ Frequency

Or via the equation 1.

$$w_{t,d} = \left(1 + \log tf_{t,d}\right) * \log_{10}\left(\frac{N}{df_t}\right) \quad (1)$$

A characterization of algorithm is that the TF/IDF score (or, weight) for a term increases with the number of occurrences in a document (TF component). Moreover it increases when considering how rare the term is across the entire collection (IDF component). However, this is based on a continuously dynamic training procedure. Furthermore, this technique demands multi relationships between the terms and texts. The most significant problems are referred in the searching procedure using weighted terms [12, 13].

In particular, the core of the problem is that it is difficult to identify a single event space and probability measure within which all the related random variables can be determined. Without such a unifying event space, any procedure that contains matching and/or mixing different measures local and global that may well be unacceptable in terms of a Shannon-like theory [13]. A more serious problem occurs, when we search using weighted terms, we typically take the query terms ignoring all the other terms in the vocabulary. It is difficult to realize how such a practice could make sense in a Shannon-like model: every term in the document must be supposed to transmit information as well as any other. That is, the existence of a term k_i would carry. Then an amount of information log P (k_i) is produced irrespective of whether or not it is in the query. There is not any connect for the amount of information to the specific query. So we would have no explanation for leaving it out of the estimation [13]. However, the similarity of the typical IDF formulation to a component of entropy has inspired other researchers [14] to make connections, sometimes somewhat differently from the connection suggested above.

B. The Basis of the Algorithm

The feature extraction is based on vector space model theory. In more details, we considered that the title of a published document depicts a determined vector ($\mathbf{v}_s$) which is unique. This technique creates an auto-correlation mechanism which is unique from each document and this practice gives a solution to Shannon-like model because each document is downgraded in term locally. Thus, a stable-single event space is created for each document with computing probability measure. Taking into account this consideration, we assume that each word of a specific title represents of a vector ($\mathbf{v}_w$) which consists of three (3) features namely, the number of characters (w), the strength (s), and the order in the title (n) (see equation 2).

$$\vec{V}_w^{(i)}\Big|_{i=0}^{n} = \left[\, i \quad s_i \quad w_i \,\right] \quad (2)$$

Where, $s_0 = 0 \wedge w_0 = 0$

For this reason, we normalized the vector ($\mathbf{v}_w$) and we obtained the equivalent vector

$$\vec{V}_{we}^{\,i}\Big|_{i=0}^{n} = \frac{\vec{V}_w^{\,i}\Big|_{i=0}^{n}}{\left\|\vec{V}_w^{\,i}\Big|_{i=0}^{n}\right\|} \quad (3)$$

As strength (s) we use an encoding measure and for simplification reason we defined the ASCII encoding [15], more details of this procedure is given in the experimental part.

Furthermore, the vector (vs) depicted as the resultant vector (vs), see equation 3.

$$\vec{V}_s = \sum_{i=0}^{n} \vec{V}_{we}^{(i)} \quad (4)$$

C. The Ranking criterion

The proximity rankings of documents in a keyword search can be estimated, by using the assumptions of document similarities theory [16], by comparing the deviation of angles between each document vector and the original query vector is the resultant vector (vs) which represents the dominant orientation vector of the title. The calculation of the cosine of the angle between the vectors, is defined by equation (5).

$$\cos\theta^{(i)} = \frac{\vec{V}_{we}^{(i)} \vec{V}_s}{\left\|\vec{V}_{we}^{(i)}\right\| \left\|\vec{V}_s\right\|} \quad (5)$$

The new ranking of $r_{j+1}\Big|_{j=0}^{n} = \min\left[\cos\theta^{(i)} - r_0\right]$ (6)

Where $r_0 = \min\left[\cos\theta^{(i)}\right]$

D. The novelty of the Algorithm

Taking into account the sections *II.A* and *II.B* the calculation of documents vectors can be calculated into two different techniques.

In the classical vector space model theory using the TF/IDF algorithm the document vector is extracted using multivariate components i.

$$U_s = \left[\, w_1, w_2, ..., w_i \,\right] \quad (7)$$

In the proposed method the document vector v_s is extracted using the three aforementioned components $\left[\, i \quad s_i \quad w_i \,\right]$. In other words the proposed method uses stable size of document vector (dimension 3).

In the same way, the query vector of the proposed vector has the same dimensionality while the classical vector has i dimensionality.

A vector space model $\vec{V}_i$ is adopted instead of vector $\vec{U}_s$ where the i variable is replaced by variable r (see equation 6).

The reason of this replacement took place because variable r is considered as a very important factor in the semantic theory [17]. The replacement was made deliberately since vector $\vec{V}_i$ represents the influence's degree through variable r because this will lead in a new ranking procedure according to its influence.

$$\vec{U}_i = [r_i \quad s_i \quad w_i] \tag{8}$$

Finally, in both cases the relevance rankings of documents in a keyword search can be calculated, using the assumptions of document similarities theory, by comparing the deviation of angles between each document vector and the original query vector where the query is represented as the same kind of vector as the documents (see fig. 1).

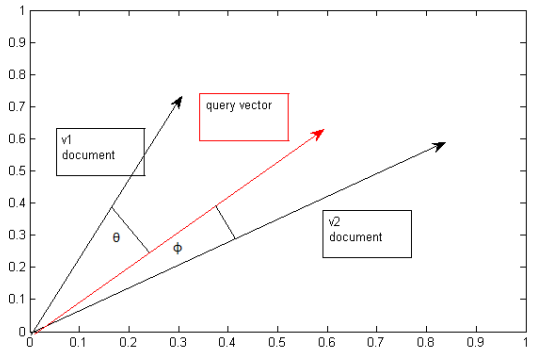


Fig. 1. The query procedure between query vector and document vector

The matching calculation is implemented via the cosine of the angles (θ,φ example see fig. 1) between the vectors, instead of the angle itself, see equations (5,6).

III. Experimental Part

A. The Data Collection

The next step of the experiment lies in using an internet browser to visit Google Scholar in order to gather titles from 3 different Classes from Dewey Classification. In detail, from each Class we choose 10 titles from 10 different subclasses, that way we gather a sample of 181 titles totally from 3 different Dewey Classes. These Titles are: knowledge, systems, bibliographies, catalogs, libraries, biographies, topology, publishing, manuscripts and Algebra. Therefore our total title sample is approximately 181 titles. Continuing with the methodology steps followed, after already explaining the algorithm that extracts the 3 most dominant keywords and before inserting the title to the algorithm for processing, the step that needs to be done is a preprocessing of the title. The title is filtered in an automatic way in order to remove words considered as "non keywords", such as stop words. In this paper we consider as a valid keyword, words that belong to the part of speech noun or compound noun. After the title is filtered, the next step is to apply the algorithm to each title. All the results for each separate title are being registered. The output that is generated every time returns the three most dominant words that are considered as keywords for every title. In the next step we test the output in order to measure our experiment success rate. We search in Google Scholar (in title) using the exact output of the algorithm and carry out a statistical research of the total results. In order to be more precise we will present in detail an example/sample taken from our research. The whole process begins once we have selected a title of an article, which was obtained by using Google Scholar. Our search term was "Knowledge" and the title obtained "Advances in knowledge discovery and data mining". Now, the obtained title must be preprocessed, before applying the algorithm. By using POS software, we keep words that based on the part of speech they belong they can be considered as valid keywords. All stop words must be cleared and also part of speech that for the purposes of this research must be excluded. Therefore, from the original title what is left is "Advances knowledge discovery data mining". At this point we can apply the algorithm to the title that is left in order to export the three most dominant keywords of the sentence. After the algorithm is applied, we continue the experiment by testing whether our results are successful. To do so, we must visit Google Scholar again and by searching only in title, with the three most dominant keywords we check the results. The first result we find out of 181 results is the original title.

B. Implementation of the Algorithm

According to sections (*II.A, II.B, III.A*) we implement the proposed procedure using the aforementioned example according to the following steps:

Step 1. In the pre-processing procedure the title "Advances in knowledge discovery and data mining" is filtered in the eliminated one "Advances knowledge discovery data mining" using the rule of section (*III.A*)

*Step 2.*_We applied the algorithm by using the equations 1& 2 for the extraction of the v_w features. This implemented by the following code of Matlab and the results are presented in Table I:

```
num1=double('Advances knowledge discovery data mining');
 k1=find (num1==32);
 k2=length(k1);
 sol1=[];
 for i=2:1:k2;
    j=i-1;
    t=num1(k1(j)+1:k1(j+1)-1);
    sol=[i,length(t),sum(t)];
   sol1=[sol1;sol];
 end
        t=num1(1:k1(1)-1);
       solbeg=[1,length(t),sum(t)];
                    t=num1(k1(k2)+1:length(num1));
       sollast=[k2+1,length(t),sum(t)];
       sol1=[solbeg;sol1;sollast];
 % sol1=Vw word vector
 % word normalization vector Vwe
    d1=sol1/norm(sol1);
```

Table I. The Features of V_w

Words	*Order*	*No. Characters*	*Strength*
Advances	1	8	805
knowledge	2	9	960
discovery	3	9	984
data	4	4	410
mining	5	6	642

Step 3. We applied the algorithm by using the equation 3 for the extraction of the resultant vector (vs) features. This implemented by the following code of Matlab " u=sum(d1)" and the results are presented in equation 7:

$$\vec{V}_s = \sum_{i=0}^{5} \vec{V}_{we}^{(i)} \vec{V}_s^{(i)} \Big|_{i=0}^{5} = [0.0085 \quad 0.0204 \quad 2.1524] \tag{7}$$

Step 4. This implementation according to equations 4 and 5 give the following results in Table II.

Table II. The angle between the resultant ($\mathbf{v}_s$) and the word ($\mathbf{v}_w$)

Words	*Cosθ*	*Ranking*
Advances	0.1572	3
knowledge	0.1069	2
discovery	0.0547	1
data	0.3332	5
mining	0.2202	4

As result of this procedure the words "*discovery, knowledge and Advances"* are selected as the optimum triple keywords.

C. The Retrieval Results

According to previous stage, 181 triple keywords are extracted from corresponding Scholar Google titles. In the retrieval procedure each of the 181 triple keywords are given as queries in Scholar Google machine in order to obtain the corresponding answers. Furthermore, we investigate if the primary keyword for collection of the articles contains in the extracted triple keywords. More details, about the order of this answers regarding the score of the successful trial of retrieval the correct title is given in Table III.

Table III. Retrieval Results regarding to 181 triple words

Order of success	*Queries*	*Present of primary Keyword*
1	156	155
2	14	12
3	3	2
4	0	0
5	1	1
6	0	0
7	0	0
8	0	0
9	0	0
10	1	0
11	2	0
12	1	0
13	1	0
14	0	0
15	0	0
16	0	
17	1	

Furthermore, if we considered that x is the order of the successful retrieval using the triple words and y is number of queries. Then there is a polynomial n degree

$$p(x) = p_1 x^n + p_2 x^{n-1} + \ldots + p_n x + p_{n+1} \tag{8}$$

Where

$$\min\left(\sum_i \left(p(x_i) - y_i \right)^2 \right) \tag{9}$$

Then using the following code of Matlab

```
c = polyfit(x,y,5);
xfit = linspace(min(x),max(x),length(x));
yfit = polyval(c,xfit);
plot(x,y,'o',xfit,yfit,'--')
```

we calculate that the robust fit polyonym is n=5 degree (see fig. 2)

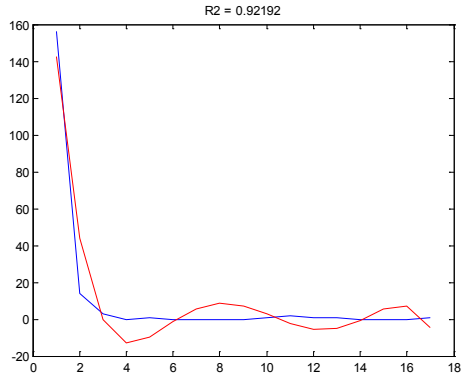


Fig. 2. The polynomial fit regression with r2=0.9219

Then the polynomial of this experiment is determined by equation (8)

$$p(x) = -0.0071x^5 + 0.3566x^4 - 6.7615x^3 + 59.1588x^2 - 233.5057x + 323.0271 \tag{8}$$

D. Evaluation of the Results

For the evaluation of the collected data the probability density function (pdf) is used. The aim of this adoption focuses on the determination of the maximum width of the interval (window)

x=[1,…,17] values. These values correspond to the order of successful article retrieval in Scholar Google search. In this way, we ask to found the robust interval window of values with the highest probability of searching scores. We typically proceed by investigating at a diagram (see fig. 2) of our data, and demanding to match it to the form of a smooth probability Gaussian density function, and rather one that is easy to work with. The Gaussian distribution is normally useful when the values a random variable takes are grouped near its mean, with the probability that the value falls below the mean corresponding to the probability that it falls above the mean [18]-[19].

In this stage we depict the density probability of integer values of polynomial with x=[1,…,17] which are presented in Table IV:

Table IV. Polynomial Values Calculation

x values	*P values*
1	142.2682
2	44.0373
3	-0.4620
4	-13.1717
5	-10.0314
6	-1.8303
7	4.9408
8	7.2375
9	4.6098
10	-1.6499
11	-9.3588
12	-16.2957
13	-21.0530
14	-23.8887
15	-27.5784
16	-38.2673
17	-66.3222

If $f(x)$ is the normal distribution $\chi \in \{p_1, p_{2,\ldots,} p_{17}\}$ then we calculated the normal density probability function (see equation 10)

$$\Pr[a \le X \le b] = \int_a^b f_X(x)dx \qquad (10)$$

a=1 and b=17 (see fig. 3), then we ascertain that the most probable value of x is 3.

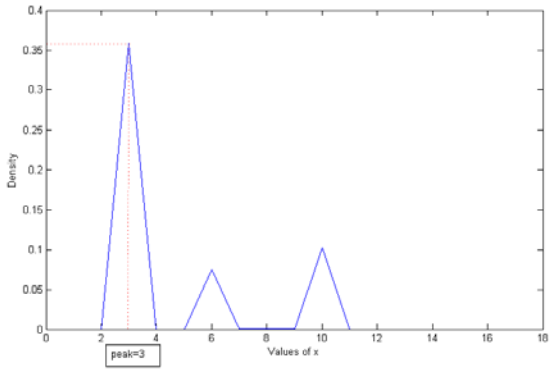


Fig. 3. The probability density function of p values

As an interpretation of this result is that the proposed method for a triple keyword query gives an accurate retrieval ranking into the first 3 positions.

Furthermore, in order to evaluate the sensitivity and specificity of our method [20] we consider that the success scores are ranged between 1 and 3 interval space, as obtained by the probability density function (see fig. 2). In our case, we considered as true positive scores, all the articles which retrieved in the first second and third order via the triple query vector. As false positives scores are considered the articles which the primary keyword does not present in the triple keywords. Finally, true negative scores consider the rest scores (beyond of the three first orders). Then, the Table III is transformed into Table IV.

Table V. Retrieval Success Results regarding to 181 triple words

True Positive	*False Positive*	*True Negative*
151	11	8
sensitivity	0.90	
specificity	0.58	

The Sensitivity relates to the test's ability to identify positive results.

This can also be written as:

$$sensitivity = \frac{number\_of\_true\_positives}{number\_of\_true\_positives + number\_of\_false\_positives} = \frac{151}{168} = 0.90$$

Specificity relates to the test's ability to identify negative results. This can also be written as:

$$specificity = \frac{number\_of\_true\_negatives}{number\_of\_true\_negatives + number\_of\_false\_positives} = \frac{11}{19} = 0.58$$

IV. Conclusions

We have presented a new keyword extraction algorithm that detects the most important terms in the titles of scientific papers, which are then employed as search keywords for retrieving articles form digital scientific online resources. In particular, the core of the problem lies in the fact that it is

rather difficult to identify a single event space and probability measure within which all the related random variables can be determined. Our algorithm is based on the vector space model to represent the article titles and treats every term in a title as a vector, with each vector containing three features namely, the number of characters a term has, the importance of the term in the title and the relative order of the term in the title. In the classical vector space model theory, using the TF/IDF algorithm, the document vector is extracted by using multivariate components *j*, while the classical vector has j dimensionality.

Based on the weight our algorithm computes for each of the above features it assigns every term in the article's title with a score indicating the suitability of that term as a search keyword that could retrieve its corresponding article in the top ranking position. The experimental evaluation of our proposed algorithm regarding real data proves its effectiveness in detecting the most suitable keywords in the articles' title and indicates that title terms may be sufficient for representing the article semantics, when it comes to scientific publications, which is to be investigated in the future. Currently, we are in the process of enriching our algorithm with additional features from the articles' contents such as the abstract and the keywords authors indicated for their articles. Moreover, to address the issue of synonyms and polysemy, that are not included in this experiment. In addition, we plan to employ the identified keywords for building an ontology that could be employed as a repository of search term for conducting text retrieval tasks in scientific online repositories.

Furthermore, in order to evaluate the sensitivity of our method we consider that the success scores are ranged between 1 and 3 interval space, such as obtained by the probability density function.

Finally, we intend to release the ontology that will be developed so that it could be employed by other researchers for performing semantic retrieval tasks. Entropy and mutual information in turn enables us to define a distance measure formally. With this distance a sound foundation is given for the capturing of the inherent structure of ontology. Consequently, in our work, we attempt to use a new algorithm which will be created in the philosophy of TFIDF algorithm giving simultaneously a solution to Shannon-like model disharmony.

REFERENCES

[1] H. Wang, S. Liu, L.T. Chia. "Does ontology help in image retrieval? A comparison between keyword, text ontology and multi-modality ontology approaches". In *Proceedings of the 14th Annual ACM International Conference on Multimedia*. 2006. pp. 109–112.

[2] Y. Matsuo, M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information". *International Journal on Artificial Intelligence Tools*. 2004. Vol. 13, pp. 157–169.

[3] M. Rajman, R. Besancon "Text mining – knowledge extraction from unstructured textual data". In *Proceedings of the 6th Conference of International Federation of Classification Societies*. 1998.

[4] E. Frank, G.W. Paynter, I.H. Witten, C. Gutwin, C.G. Nevill-Manning. "Domain-specific keyphrase extraction". In *Proceedings of IJCAI*, 1999, pp. 688–673.

[5] Y.H. Kerner, Z. Gross, A. Masa, "Automatic extraction and learning of keyphrases from scientific articles". *Computational Linguistics and Intelligent Text Processing*, 2005, pp. 657–669.

[6] Liu F. Liu, Y Liu. "Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion". *In Proceedings of IEEE SLT*.

[7] P. Turney, "Coherent keyphrase extraction via web mining". In *Proceedings of IJCAI*, 2003, pp. 434–439.

[8] G. Carenini, R.T. Ng, X. Zhou, "Summarizing emails with conversational cohesion and subjectivity," in *Proceedings of ACL/HLT*, 2008.

[9] A. Janin, D. Baron, J. Edwards, D. Ellis, G. Gelbart, N. Norgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The icsi meeting corpus," in *Proceedings of ICASSP*, 2003.

[10] Matsuo, Y., Ishizuka, M. "Keyword extraction from a single document using word co-occurrence statistical information". *International Journal on Artificial Intelligence Tools*. 2004. Vol 13, pp. 157–169.

[11] Hulth, A., "Improved automatic keyword extraction given more linguistic knowledge". In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. 2003. Association for Computational Linguistics, pp. 216–223

[12] P. Soucy και G. W. Mineau, 'Beyond TFIDF weighting for text categorization in the vector space model'. *IJCAI*, 2005, Vol. 5, pp 1130–1135.

[13] Robertson, S. "Understanding Inverse Document Frequency: On theoretical arguments for IDF". *Journal of Documentation*, 2004, Vol. 60 (5), pp. 503–520.

[14] C. Blake, 'A comparison of document, sentence, and term event spaces', In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 2006, pp. 601–608.

[15] M. Poulos, S. Papavlasopoulos, V. Chrissikopoulos, "A text categorization technique based on a numerical conversion of a symbolic expression and an onion layers algorithm". *Journal of Digital Information*, 2006, Vol. 1 (6).

[16] Salton, G., Wong, A., Yang, C. S. "A vector space model for automatic indexing". *Communications of the ACM*. 1975. Vol. 18, pp. 613–620.

[17] Harispe S. et al. "Semantic measures for the comparison of units of language, concepts or entities from text and knowledge base analysis". *Arxiv*. 2013. Vol. 1310, 1285. pp. 1-159.

[18] Parzen, E. "On estimation of a probability density function and mode". *The annals of mathematical statics.* 1962. Vol. 33, pp. 1065-1076

[19] Abramowitz, M., Stegun, I. A. (Eds.). "Probability Functions." Ch. 26 in Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing. New York: Dover, 1972, pp. 925-964,

[20] M. Poulos, G. Bokos, N. Kanellopoulos, S. Papavlasopoulos and M. Avlonitis. "Specific selection of FFT amplitudes from audio sports and news broadcasting for classification purposes". *Journal of Graph Algorithms and Applications*. 2007. Vol. 11(1). pp. 277–307

Architecture of an Agents-Based Model for pulmonary tuberculosis

M.A. Gabriel Moreno Sandoval[1]

William David Peña Peña[2]

Abstract- This text generally describes the architecture of an agent-based model for pulmonary tuberculosis in the zone of Usme (Bogotá, Colombia), product of a master's thesis in Information Sciences and Computing of District University "Francisco José de Caldas" [1]. First comes the introduction, then the tools and concepts that allow the simulation, then the model is exposed from a conceptual approach to technology-society relationship, and finally it is disclosed the author's findings.

Key words: Agents-Based Models, Geographic Automata System (GAS), Epidemiologic simulation, complex Systems.

I. Introduction

The technological edge thinks forward, looking for ways to understand it, study and intervene in the world over time towards a particular purpose. So it makes models that attempt to reproduce actual facts and allows a better understanding of the relationship of humans with them, appearing technologies with that purpose as Agent-Based Models and Geographic Automata System, among others [2].

But the future is thought from the uncertainty, because the world is complex and multiple, set of systems with the general characteristics of not to be reversible, not to be accurately predicted, to develop nonlinear interdependent relationships, and to have appearance of order (emergency), being necessary to go beyond the deterministic mathematical formal schemes [3, 4].

The bio-social systems, being complex, cannot be predicted with accuracy nor are reversible, so they have a constitution order from the simple to the complex, allowing the passage from chaos to order at the time called edge of chaos, where self-organization emerges, which is the coordination of the parts' behaviors of a system without any central power or external coercion that lead them. Jhon Stwart Kauffman formulated a hypothesis of a global connection between all parts of a physical system, that after a certain time, due to the energy accumulated between these, by inertia was that these associated themselves synergistically and generate patterns (self- adaptive systems) [3, 4].

[1]*Department of Academic Research, "Manuela Beltran" University.* gabrielmoreno10@gmail.com

[2] *Department of Academic Research, UNIMINUTO.* wpenapena1@hotmail.com

Current knowledge have the task of describing the complexity of the world we inhabit, so in this case it is intended to represent the nonlinear and interdependent relationships of a community in the middle of a tuberculosis epidemic, in order to predict the scenarios of this complex system [4, 5], explaining the construction of a Agent-Based Model for Tuberculosis in Usme's zone, and thinking about the origin and social horizon of technology.

II. Tool and Basic Concepts for the Model

A. Geographic Automata System (GAS)

Between 1999 and 2001, Paul M. Torrens and Itzhak Benenson created Geographic Automata System (GAS) for modeling phenomena on real spaces, from individual computer entities within the system: agents. On the one hand, they modified the classical Cellular Automata (CA), which is a system in one, two or three dimensions, consisting of partitions called cells, which acquire ways to present: states, from a default set of them, ranging from relations with its neighboring cells, their neighborhood (Figure 1), through pre-established transition- state rules, in a sequence of moments called evolution system [6, 7].

And on the other hand, they placed those entities that change over time (A. Figure 2) on a Geographic Information System (GIS), which is a set of layers that describe the geographic characteristics of a place, from the general to the specific (Figure 2 B). Thus, agents are related to real spaces, preserving the neighborhood concept in his role for the change of state (Figure 1), but not only on generic objects such as cells, but sometimes on entities with own characteristics and mobility: specific sites , people, vehicles, etc. (C. Figure 2) [2, 6].

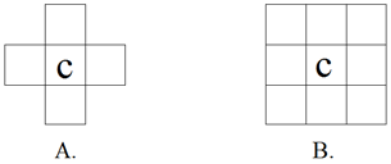


Fig 1. Types of neighborhood in cellular automata which allows the evolution of the system. Neighborhood A: Von Newman neighborhood, with octagonal neighbor cells. Neighborhood B: Moore neighborhood, with octagonal and diagonal neighbor cells [5].

A Geographic Automata System (G), is defined as the set of automatas (K), which vary from state (S) over time, through transition rules (Ts), developed at specific locations (L), from

ISBN: 978-1-61804-240-8

mobility rules (if any) of each one (ML), considering its neighborhood (N) at every time, called Tick, and the criteria of relationship they have with this (RN) (Figure 3) [7, 8]. Where, thanks to this set, It is able to see the interdependent relationships between the conditions of each area (population density, healthiness, etc.), and each agent (location, nutrition, origin, economic status, etc.) [9].

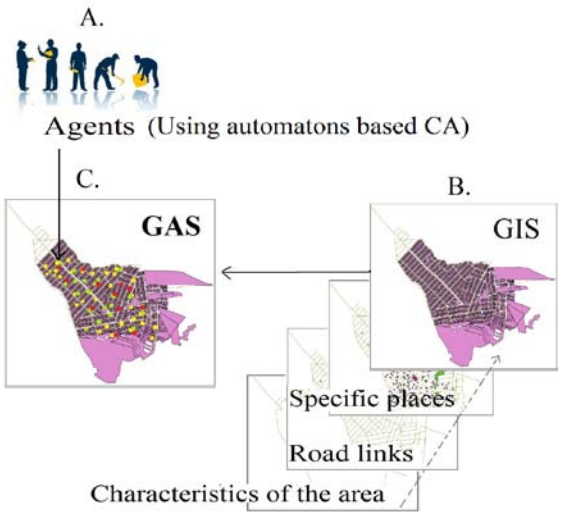


Fig 2. Integration of CA and GIS in the constitution of GAS.

$$G \sim (\mathbf{K};\ \mathbf{S}, \mathbf{T_s};\ \mathbf{L}, \mathbf{M_L};\ \mathbf{N}, \mathbf{R_N})$$

Fig 3. Representation of a Geographic Automata System (GAS) [8].

First, the geographic information of Usme Central zone, was registered with ArgGis, application that allows to enter raster data type of a land, generating a GIS shapefiles (.shp) [10]. And then again, as Framework it was used Repast, which is a set of Java classes and methods linking agents with a GIS, allowing focusing on the modeling of the relevant attributes of the phenomenon, because this program already contains the multi-threaded programming, where in pseudo-parallel way for each Tick, transition rules are executed for all agents run, and also it provides a graphical interface of the simulation [11].

B. Usme zone

Usme is a local and administrative subdivision of Bogotá, which was incorporated into the city in 1990, because before that it was a town. So today some rural practices persist (agriculture, small farms, artisanal slaughtering), where according to figures consulted (data from 1999 to 2007 that varied in the study of late 2011, the same year it was supported the thesis on the model), 84.9% of its land (18,306.52 hectares) was rural and 15% urban, preserving some natural resources (now damaged by mining, urbanization and the tanning of leather) as water sources (21 in urban sector, 23 in rural sector, 11 rivers, 2 dams and 3 ponds.) [12].

Its population consisted of 51% female and 49% male, 34.8% of the population was under 15 years and 2.5% over 60, there being a high economic dependence, where on average 100 people were dependent of every 59, workers mostly with low levels of schooling and informal jobs. In addition, from 1995 to 2005, amidst the paramilitary phenomenon nationwide, this zone received 8.2% of the displaced population that came to Bogotá by the violence [12].

Thus, according to the Unsatisfied Basic Needs Indicator (NBI, by the Spanish acronym), which considers the shortages at home: a) housing with physical or structural defects, b) lack of basic services or deficiencies in drinking water and feces disposal, c) overcrowding (2 persons/5m2), d) high economic dependence (1 productive person / 3 or more dependents), e) truancy (at least one child between 7 and 11 do not regularly attend a school), and f) Misery, when the home has two or more of the above conditions, it was found that Usme was 9, 1% of homes in NBI, with 1% in misery. Fact also reflected in the 51% of the population below the "poverty line" indicator that arises from considering the minimum subsistence income for a person [13].

C. Tuberculosis Pulmonar (TB)

Illness from the bacteria Mycobacterium tuberculosis, native bovine and adapted to the human (zoonoses), which attacks places rich in mucous membranes (such as the lungs) and develops according to: the strength of the micro-organism to survive and be transmitted (virulence), the opposition that the body makes to the micro-organism (resistance), ease that people have to acquire and develop the disease (hypersensitivity) and the morphology of the affected tissues (genesis of the pattern infectious) being vulnerable to ultra-violet rays. It manifests in cough with coughing up phlegm or blood, evening sweating, fever, fatigue and unintentional weight loss.. Phases of development are: the attack or the arrival of bacillus in the body, then its logarithmic growth and the progressive activation of cells infected into other tissues, then there is an immunity, delayed as the body's inadequate response to disease, thus leading to the destruction of tissue and the transmission of new one [14].

Diagnosis is made by biochemical reaction of cultivated samples. To eradicate it is used treatment shortened supervised (DOTS), the antibiotic rifampicin, isoniazid, Pyrazinamide, and Ethambutol, over a period of 48 weeks, extending in the case of a relapse to 63, and whose success or failure (even fatal), depend on conditions of life of the infected as food, hygiene, wholesomeness in the habitat and permanence in the treatment, since deaths from TB are commonly associated with poverty and undernutrition [15].

III. Agent-Based Model for Tuberculosis in Usme

A. Simulation's time and space

In architecture model time and space simulation, as the first condition for the interdependence between agents is carried out in the middle of tuberculosis outbreak as a complex system were established. On the one hand *TimeProject* whose *TimeProject* constructor class is created generated internal time model, allowing the passage of the method *step()*. And moreover the *Global* class, entering shape files and generates graphical output interface was created, also it determines the

amount of Tick equivalent to every hour (1 hour = 1 Tick) , the number of initial agents in each state, that the epidemiological model of pulmonary tuberculosis (SIR model) are: Susceptible (which could be infected), infected (disease carrier) and Recovered (who overcomes the infection), and the amount of each type of agent , which under Usme socioeconomic conditions are: *Housewife*, *Worker*, *Student*, *Deplaced* and *Homeless*, which enter as parameters via *setValuesFromParameters()* method of that class [1] (A. Figure 4).

Therefore, States and agents that is different in the middle of a space and a time shared, determine that the Organization of the information in a complex system should be a dynamic structure that contains the particularity of the elements and their no-lineal relationships, and don't simply a collection of attributes.

B. Contexts

Contexts are joint that grouped the agents of a system and the relationships between them. Thus, *MainContext* class is the generic context that through the *build()* method, load the global timeline, originates and controls the contexts of spaces and people, and is the time and the particular space of these agents. Subsequently determined the context of each place: Home (*HomeContext*), workstation (*WorkPlaceContext*), study location (*StudyPlaceContext*) and entertainment venue (*EntertainmentPlaceContext*), dynamically created by reflection (in the execution of the program), through the method *createSubContext()* in the class *CityContext*, referencing their positions from the tract (class *RoadContext*), the intersections of these and the boundaries between places (class *JunctionContext*). In the same way, with the class *PersonContext* was created the possibility of dynamically generating the contexts of agents -sets of sites and other agents-: *HousewifeContext*, *WorkerContext*, *StudentContext*, *DeplacedContext*, *HomellessContext* (B. Figure 4).

The location of the agents was determined by the rules of movement (ML) of each, using for this parametric class *AgentContext* and *getGeography()* method. And the location of the sites was performed using the same method name *PlaceContext* parametric class (C. Figure 4).

Due to the needed for information on the TB outbreak (major focus of infection, more infected population trend of spread of the disease, etc.), which would act on this, the dynamic generation of contexts allowed the development of the non-linear relationships between the different agents in a period of time, loading and processing the increased volume of information in the model thanks to the architecture of the program, through reflection, extended the possibilities of development of the system.

C. Places

Parametric classes system can evolve on the geographical conditions of Usme, which is very important for the relationship between the demographic conditions of a region and the spread of an epidemic (TB in this case), where the different types of places and their characteristics were determined and verified with the abstract class *Place*, with the attribute *listPerson* created an arrangement for certain amount of agents inside, regardless of their type, verify with the method *getOvercrowdingCondition()* the overpopulation (boolean data), on the basis of the capacity of the place (method *getPeople Capacity()*), and the number of people in it (method *getAmountPerson()*) [1].

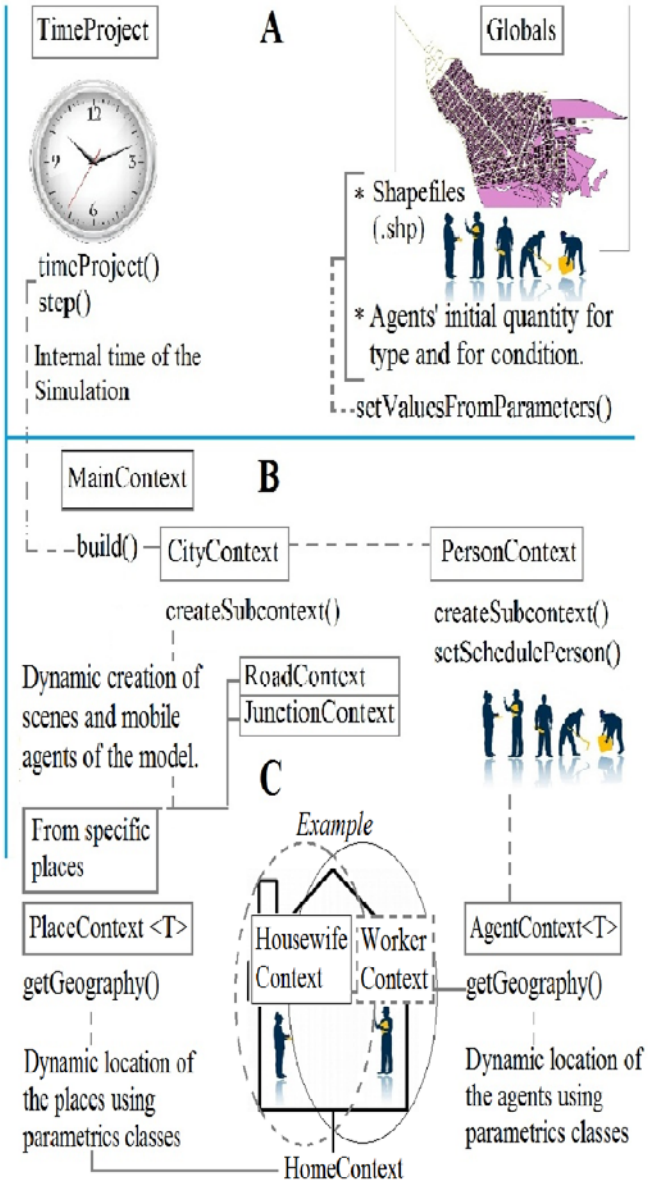


Fig 4. Time and space model, and training contexts through parametric classes.

The specific locations and attributes were defined with *HomePlace* classes (households), *WorkPlace* (work sites), *StudyPlace* (study sites), *EntertainmentPlace* (entertainment venues) and *Route* (tracks), which inherit of the class *Place* attributes as the maximum area place (*MaxArea...*), high capacity (*MaxAmount...*), many people (*amountPersons*), capacity (*peopleCapacity*), use, management, geometry, number of people per state (*amountSusceptible*, *amountInfected*, *amountRecovered*), among others.

Calculation of amounts by State was performed on the method *step()* -different class *TimeProject step()* method, and later exposed the *Person* class *step()* method-, and the distribution of agents by States was random from classes of each place: *Work Place*, *Study Place*, *Home Place*,

Entertainment Place, trough the method *getRandomStudyPlace*, *getRandomHomePlace*, *getRandomWorkPlace*, and *getRandom EntertainmentPlace*, respectfully (A. Figure 5).

The description of Usme geography through Shape files and their inclusion in the system, allowed to generate scenarios appropriate for studying the TB epidemic in the actual conditions of this place, as a requirement of a model oriented to care for the life of the population, through the generation and the study of patterns in the spread of this disease.

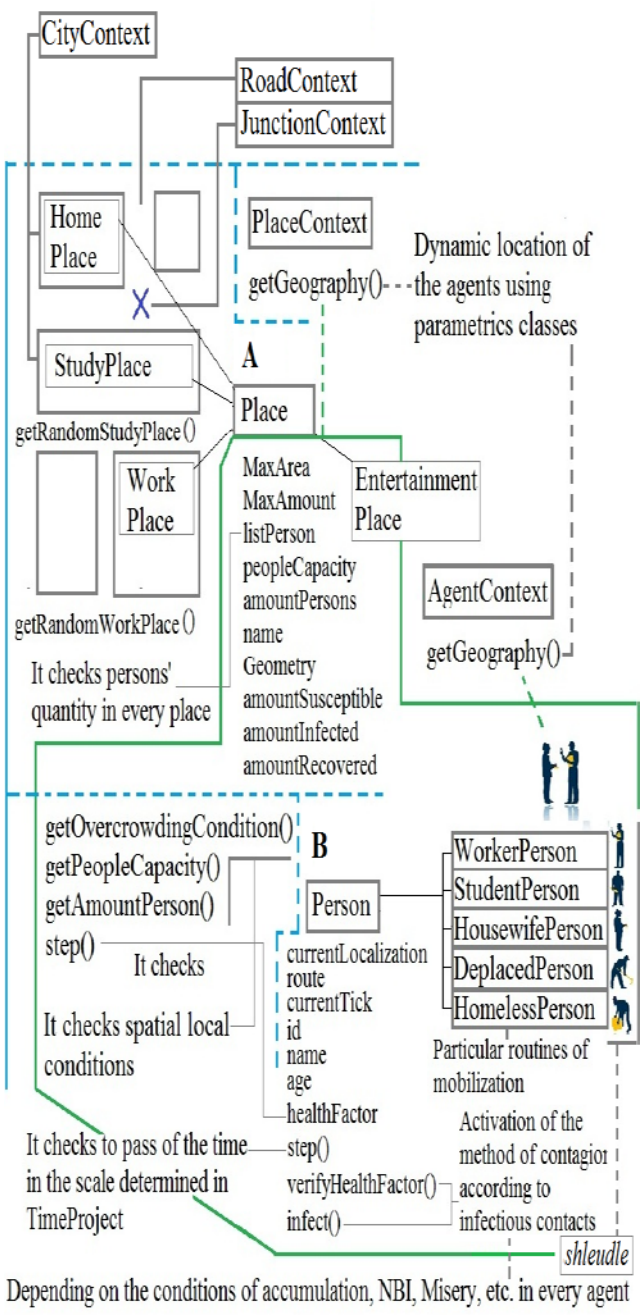


Fig. 5. Architecture of sites Agents model.

D. Agents

It determined each agent-specific property so that they can evolve relationship tailored to the reality of the TB outbreak: daily places of movement, origin in the case of displaced persons, etc. Why the particular agents were created 'Worker', 'Housewife', 'Student', 'Deplaced for violence' and 'Street people', through classes, *Worker*, *Housewife*, *Student*, *Deplace* and *Homeless*, respectively.

Previous classes inherit from the Person class attributes of localization in each place (*currentLocalization*), routes of movement (*route*), location at every moment of the system (*currentTick*), identification (*id*), employment status, medical service, stratum, previous infection by HIV and TB, type of person, State of health (*healthFactor*) and number of health (*numberHealthFactor*), which is verified by the *verifyHealthFactor()* method According to the infections activate method *infect()*, which passes the person State Susceptible to infected (first passive, then contagious).

Although it must be clarified that home and the places of development of each agent are attributes in each particular class (*Worker*, *Student*, *Housewife*, *Homeless*, *Deplaced* etc.), and is thus allowing you to determine routines on different types of agent through parameters that are entered as *Schedule* (specific cases) in the classes of every kind of person (B. Figure 5).

E. Rules of transition

Rules of transition between the States of SIR epidemiological model are identical for all agents (TABLE I.): susceptible to passive infected, and hence a contagious agent, then be recovered, however the times varied considering multiple living conditions of Usme (TABLE II.), where the method *infect()* of class *Person* activated particular rules contained in the classes of each type of agent making it easier in some cases to the acquisition of disease, and likewise, hampering its recovery (average increase of vulnerability, TABLE II) [1].

TABLE I: TRANSITION RULES.

State change	Susceptible to Infected liability	Infected liability. From Infected Contagious	Infected to Recovered
Time (1 Tick = actual 1 hour).	12 Tick	96 Tick	Tick 4032 (24 weeks): Assuming completion of treatment of 48 weeks.

According to the conditions of life in each agent, for example, the time 12 Tick in the vicinity of a contagious, necessary to move from Susceptible to passive infected, agent reduced by the average decrease in 1/8, 1/6, or 1/4 of the number healht Factor attribute, as there was an average delay of 1/8, 1/6, or 1/4 in raising that attribute according to the conditions of table 2, to move from infected to retrieved, for a population chosen randomly according to the percentage of population affected, with the following conditions: 1 in every 4 people in overcrowded housing was inadequate, 1 in 3 people in destitution was in misery by NBI, 1 in 4 people with poor nutrition in the NBI1 in 3 people with chronic malnutrition was on the same indicator, as gaps in living conditions have problems shared, setting such conditions through schedule in each type of agent [9].

TABLE II.: IMPACT ON VULNERABILITY FACTORS ATTRIBUTE PERCENTAGE WITH POPULATION AFFECTED *NUMBERHEALHTFACTOR.*

Vulnerability Factor	**Average increase in vulnerability. Attribute numberHealhtFactor () <= 0.2 = infected. Scale of 0-1.**	**% Affected Population [3]**
Poor nutrition	1/8	30%
Chronic malnutrition	1/6	15%
Overcrowding	1/4	5%
Housing inadequate	1/6	2%
Households with NBI (1 Home = Average of 4 People)	1/8	9.1% households = 36.4% population
Misery as NBI	¼	1% of households = 4% population
Homelessness	¼	6%

F. A CASE

With a total amount of 500 agents, 20% love House, workers 25%, 25%, 15% and 15% displaced students of street, 5% initial of infected, recovered 20% and 75% susceptible, were obtained data from Table III.

TABLE III: AGENT TYPES OF INFECTION BY TICK.

Agents	Infected% (of the total) for Tick. T= 1	T= 200	T= 528	T= 4032
Housewife	1%	1.2%	3.8%	0.2%
Worker	1%	1.8%	4.2%	0.4%
Student	1%	1.4%	3.4%	0.2%
Displaced	1%	2%	3.6%	0.8%
Homeless	1%	1.6%	4.2%	0.6%
TOTAL INFECTION	5%	8%	19.2%	2.2%

The resulting data is inferred that infection levels rise first in patients with immune deficiencies related to bad nutrition, poverty, and inadequate conditions of habitat (housing overcrowding or free services: NBI), nothing that although the highest proportion of agents corresponded to a type (workers and students), harder those infected in that eradicating the disease, they are those who described under very poor conditions of existence (Street and displaced inhabitants), which, if they had not taken into account would have caused a uniform tendency in the simulation behavior, without allowing to see the auto-organising patterns of the disease from the actual conditions to be able to act on this.

[3] Estimated figures weighting of survey data quality of life and social diagnosis of the town, considering older vulnerabilities for IDPs and resident street [9, 10].

IV. CONCLUSIONS

1) The dynamic construction of contexts from parametric classes (T:Microsoft.VisualStudio.Test Tools. execution. agent context and Place Context) in the architecture of the model allowed a greater volume of information loaded and processed, reducing the lines of code and optimizing the performance of the machine, but projecting a greater precision of the simulation to the real conditions of the phenomenon in Usme (special agents(: street people, displaced persons, and specific conditions: NBI, nutrition, overcrowding) at all times, making this program as a support tool that could be used by local health authorities.

2) The creation of Time Project and Global classes with their respective methods, and the union coherent types of agent to these, incorporating in a single set all elements of the simulation, is the architecture of the system as a Framework for the creation of models of epidemics-targeting GAS, regardless specific geography in which develops and simulated disease now that the creation of common space and time, and the possibility of agents to act is through these by means of dynamic contexts, create the basic conditions for any model can develop.

3) The model is created for the study of infection and the spread of tuberculosis taking into account the social dynamics, adopt to different fields and learn behaviors and theoretical but reliable of a possible epidemic statistics without ethical implications, and allowing the application of preventive methods and control large scale.

REFERENCES

[1] Moreno Sandoval, L. G. Epidemiological model of the tuberculosis based on agents in Usme, Bogota, Colombia. Thesis of mastery in Sciences of the Information and the Communication. University Distrital " Francisco Jose of Caldas ". June, 2007 - March, 2012.
[2] Benenson I., Torrens P. M. Geographic Automata Systems: A New Paradigm for Integrating GIS and Geographic Simulation. In Proceedings of the AGILE 2003 Conference on Geographic Information Science, Pages 367-384 (Lyon, France). April 24th- 26th, 2003.
[3] Lucas C. Self-Organization. In Magazine Thinking the complexity, 3(6): 5-15, 2009.
[4] S. M. Manson., S. Sun., D. Bonsal, "Agent-Based Modeling and Complexity" in *Agent-Based Models of Geographical Systems* A.J. Heppenstall, A.T. Crooks, L.M. See, M. Batty, Eds. Dordrecht, Heidelberg, London, New York, Springer, 2012, ch, 4, pp. 125-140.
[5] Perez Martínez A. Stuart Kauffman's work. The problem of the complex order and his philosophical implications. In Magazine Thinking the complexity, 3(6): 21-38, 2009.
[6] Benenson I., Torrens P. M. Geographic Automata Systems. In International Journal of Geographical Information Science. 19 (4): 385–412, April 2005.
[7] Kari J. Cellular Automata. Technical Report. University of Turku, Finland, 2013.
[8] I. Benenson, V. Kharbash. Geographic Automata Systems: From The Paradigm to the Urban Modeling Software. In: Proceedings of the 8th International Conference on Geocomputation University of Michigan. United States of America. July 31 to August 3, 2005.
[9] M. Parvin, "Agent-based Modeling of Urban Phenomena in GIS" Capstone Project, thesis submitted in fulfillment of Masters in Urban Spatial Analytics, University of Pennsylvania, CA, 2007
[10] Autonomous university of Madrid. Manual (basic level) for the production of maps with ArcGis. December, 2011.
[11] Malleson N. Repast Simphony Tutorial. Technical Report. May 7, 2008.

[12] Secretariat of estate of the District. Crossing Usme 2004. In: physical Physical and socioeconomic Diagnosis of the localities of Bogota, D.C. Major mayoralty of Bogota. Bogota DC, Colombia. 2004.
[13] Secretariat of Health of the District. The health and the quality of life. Locality 5 - Usme. Major mayoralty of Bogota DC. 2009.
[14] Dr. S. Invertrz. Pathogeny of the tuberculosis. University of Buenos Aires. Buenos Aires, Argentina. 2008.
[15] Dr. Ospina S. The tuberculosis, a historical - epidemiological perspective. Magazine Infectio. Vol.5 no. 4 pp. 241-250, Bogota October - December, 2001.

Authors

Luis Gabriel Moreno Sandoval: He was born on March 8, 1986 in Bogotá, Colombia. System Engineer, Master of Science in Information and Communication of the University District "Francisco Jose de Caldas" (Bogota). Researcher tecnolgías vanguard as agent-based models and natural language processing, manager and director of leading academics and business groups on innovation and technology. Currently he is finishing his master's degree in administration from the University "Shown of Colombia".

William David Peña Peña: He was born on February 12, 1987, Philosopher (University "Free" Colombia) and transdisciplinary research. Manager various academic areas in the humanities and speaker at international events (Scientific Convention CUJAE, Cuba, 2012. Innoed 2013, Cuba. CLEI2014 And soon, Uruguay. Among others) on the application of technology to solve social problems related to education and health. Currently he is doing based on developments in the particular cognitive processes educational technologies.

Flanged wide reinforced concrete beam subjected to fire - numerical investigations

A. Puskás and A. Chira

***Abstract*—**When using wide pre-stressed reinforced concrete beams for realization of prefabricated concrete slabs beside their mechanical disadvantages with respect to the regular ones, their increased risk of brittle failure and other uncertainty in their behavior have to be considered. Since in structural design structural height might be imposed, wide beams with all their disadvantages become the right solution. But how the mechanical behavior of the prefabricated floor system can be influenced in case of fire? The paper presents numerical modeling for prefabricated floor system using wide reinforced concrete beams considered together with the corresponding double floor with precast slab and concrete topping, subjected to fire in different fire scenarios, establishing the scenario with the highest risk on the structural stability.

***Keywords*—**non-linear analysis, transient coupled temperature-displacement, wide pre-stressed reinforced concrete beams

I. Introduction

In practice of precast concrete structures several structural solutions are widely spread. Double floor systems are assuring the necessary speed in structure realization as well as the structural flexibility, which, in combination with use of wide pre-stressed reinforced concrete beams, presents countless benefits, but at the same time raises questions with respect to their mechanical behavior when subjected to loadings. Provisions of existing design codes are not clearly covering all the possible load situations according to [1], neither use of wide pre-stressed reinforced concrete since either code provisions are not covering or the structural system used [1] has no or limited references in codes and practice. From fire resistance point of view the degree of fire resistance is established according to [3], which is establishing the necessary fire resistance of each structural element. According to provisions of [2] fire resistance is assured by foreseeing specific concrete cover of reinforcement. Use of pre-stressed slab elements can avoid excessive deformations and postpone the appearance of cracks [4], and as consequence the excessive exposure of reinforcement to fire. Due to the decreased structural height of wide beam floor systems their sensibility to fire is crucial. Studies on wide reinforced concrete beams subjected to fire shows that in numerical analysis the beam supposed to different fire scenarios has an adequate behavior for the imposed fire resistance degree [5], but their behavior as part of the floor system needs further investigations.

This work was realized as consequence of the research concerns of the Technical University of Cluj-Napoca, Faculty of Civil Engineering.

A. Puskás is now with the Department of Structures, Faculty of Civil Engineering, Technical University of Cluj-Napoca, G. Baritiu street no. 25, 400027, Cluj-Napoca, Romania (phone:+40-264-401545, e-mail: attila.puskas@dst.utcluj.ro).

A. Chira is working at the Department of Structural Mechanics, Faculty of Civil Engineering, Technical University of Cluj-Napoca, G. Baritiu street no. 25, 400027, Cluj-Napoca, Romania. He is now postdoctoral researcher of the Department of Building Structures, Faculty of Civil Engineering, Department of Building Structures, Thákurova 7, 166 29 Praha 6, Czech Republic (e-mail: alexandru.chira@ fsv.cvut.cz).

II. Problem formulation

For studying the behavior of the wide pre-stressed reinforced concrete beam subjected to different fire scenarios a floor system using this type of beam is considered [6], as presented in Fig. 1.

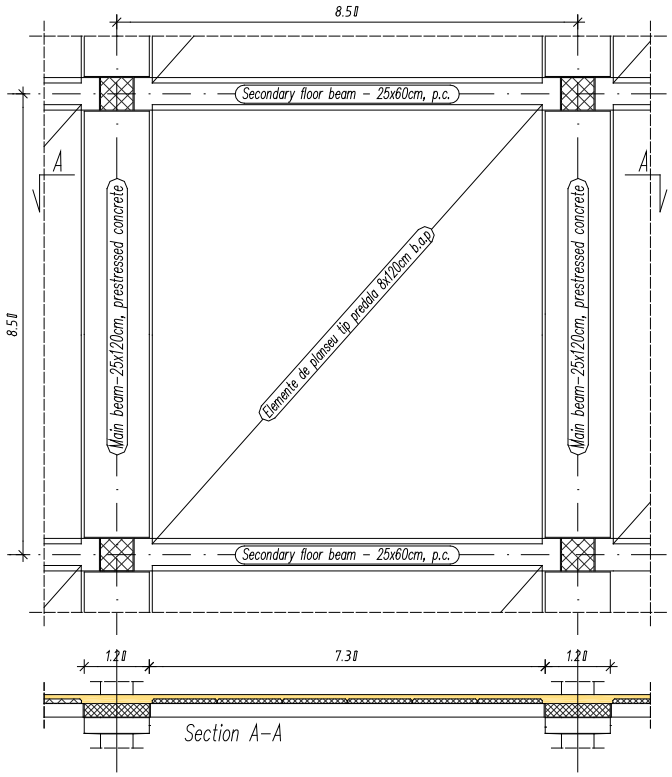


Fig. 1 floor system layout

For the modeling the floor system effective width of the beam has been considered (Fig. 2) according to the clause 5.3.2.1 of [1], taking into consideration the used concrete and reinforcement quantity, quality and disposal.

ISBN: 978-1-61804-240-8

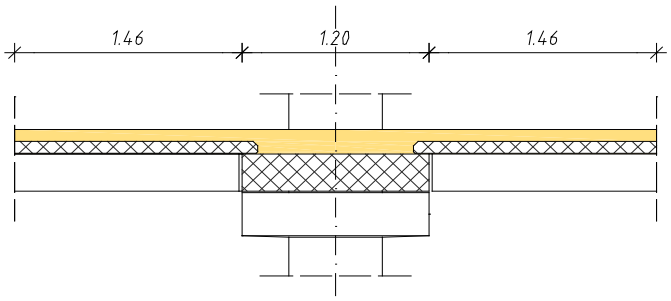


Fig. 2 effective width of the beam

The total double floor thickness is 17 cm, build up by a 8 cm precast floor of C40/50 class concrete and a 8 cm topping of C25/30. The main wide beam reinforcement is presented on Fig. 3. Its total length is 7.86 a, while the cross-section is 25x120 cm. The concrete quality used is C30/37, with longitudinal and transversal reinforcements PC52 type and active reinforcements of St1660 type having 12.9 mm diameter. For the analysis Abaqus FEM code [7] has been used, with different material law for both concrete and steel, for every change of temperature. For concrete it has been used the C3D8T solid elements: an 8-node thermally coupled brick, tri-linear displacement and temperature and for the reinforcements the T3D2T elements: a 2-node 3-D thermally coupled truss. The analysis was done in three steps: in the first one the pre-tensioning was done and in the second one the gravity loads were applied for both a static general analysis being used. In the third step a transient coupled temperature-displacement analysis have been used [8][9][10][11][12].

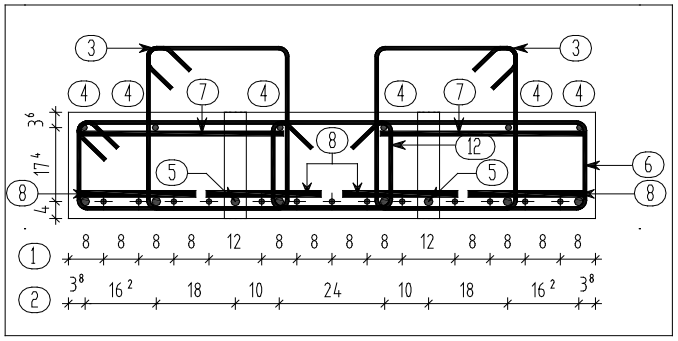


Fig. 3: main wide beam reinforcement

For the analysis of the beam taking into account the effective width of the beam three different fire scenarios have been used [5],considering two hours from the curve presented in figure 4.

Fig. 4: ISO 834 standard fire curve [2]

III. MODELING OF THE BEAM ON FIRE

In order to investigate the flanged wide reinforced concrete beam subjected to fire analyses have been performed using Abaqus finite element analysis. Results for the three scenarios are presented in the followings:

A. *Scenario I*

The first scenario of fire takes into consideration acting of the fire along the whole length and aside the whole width of the flanged beam (Fig. 5).

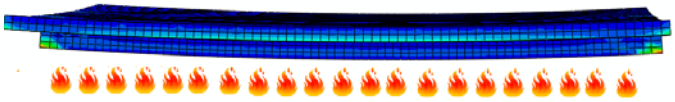


Fig. 5: Scenario I of fire action

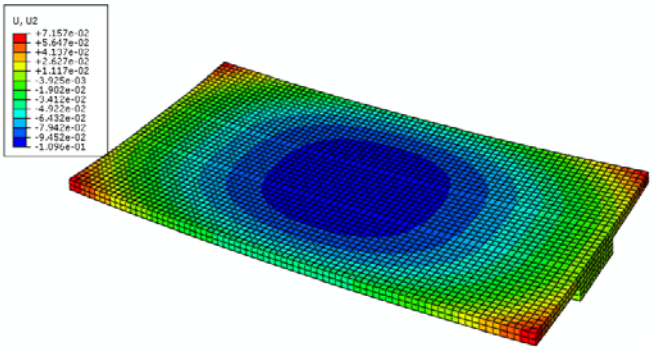


Fig. 6: Displacement after two hours of fire, d=10.96 cm

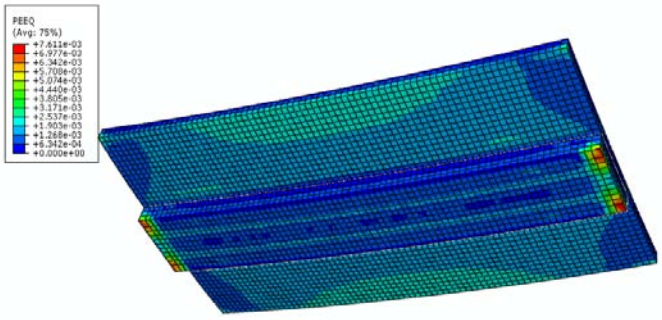


Fig. 7: Concrete plastic equivalent strain from compression, PEEQ=7.61E-3

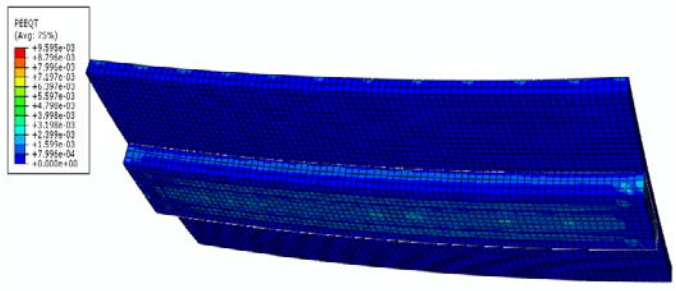


Fig. 8: Concrete plastic equivalent strain from tension, PEEQT=9.59 E-3

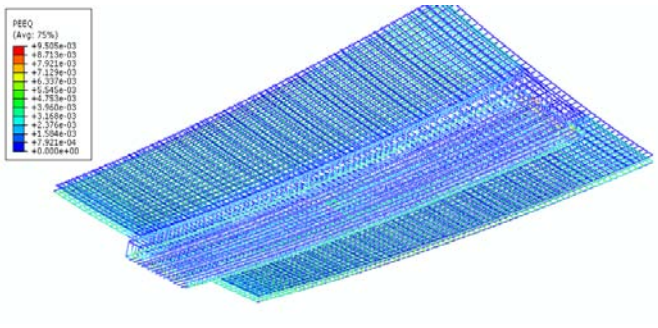


Fig. 9: Steel plastic equivalent strain, PEEQ=9.50E-3

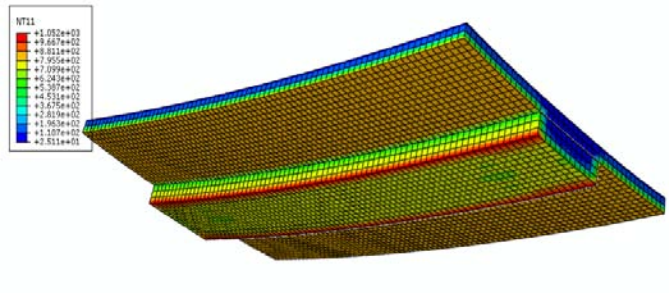


Fig. 10: Temperature distribution on concrete after two hours of fire, T_{max}=1052ºC

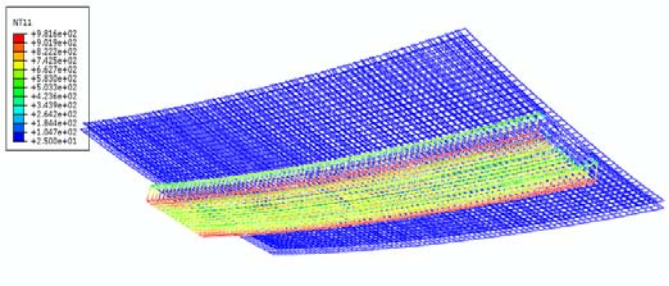


Fig. 11: Temperature distribution on steel after two hours of fire, T_{max}=981.6 ºC

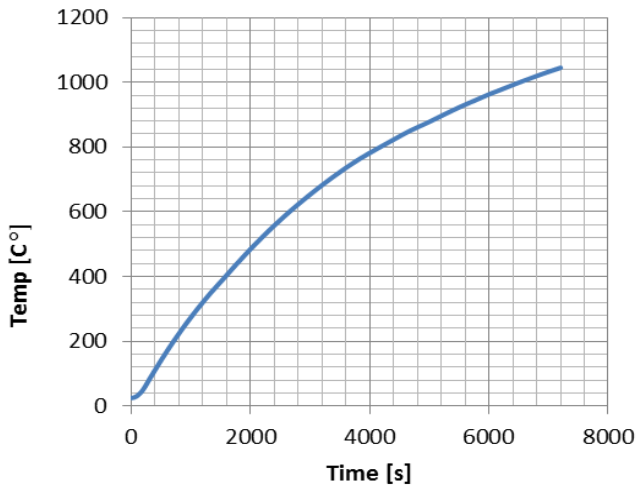


Fig. 12: Time versus maximum temperature curve on concrete

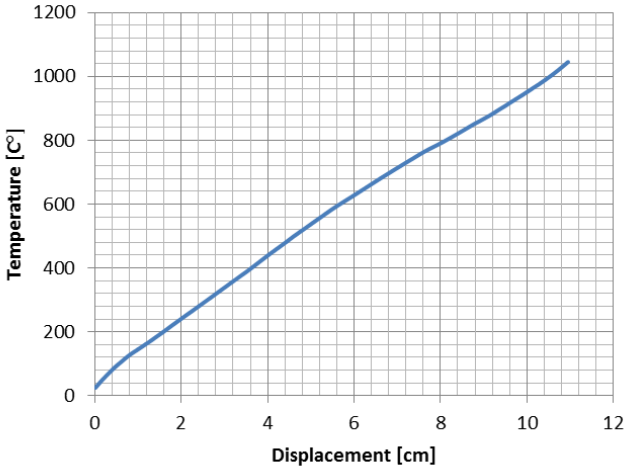


Fig. 13: Displacement – maximum temperature diagram

B. Scenario II

The second scenario of fire considers the fire acting in the middle of the opening on a strip of 0.50 m wide, aside the whole width of the flanged beam (Fig. 14).

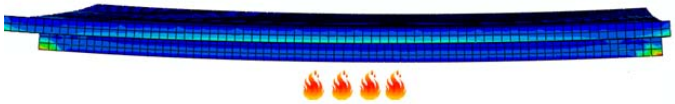


Fig. 14: Scenario II of fire action

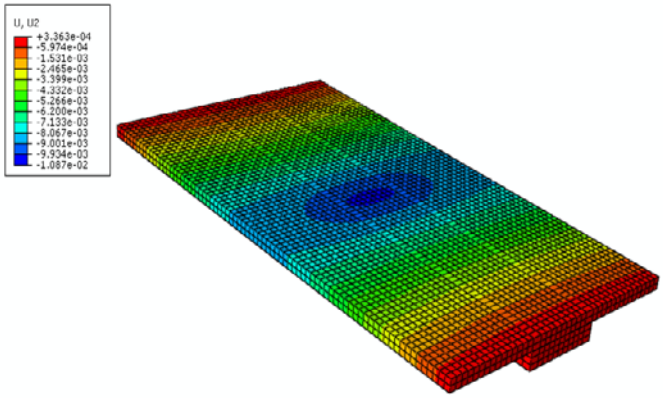


Fig. 15: Displacement after two hours of fire, d=1.087 cm

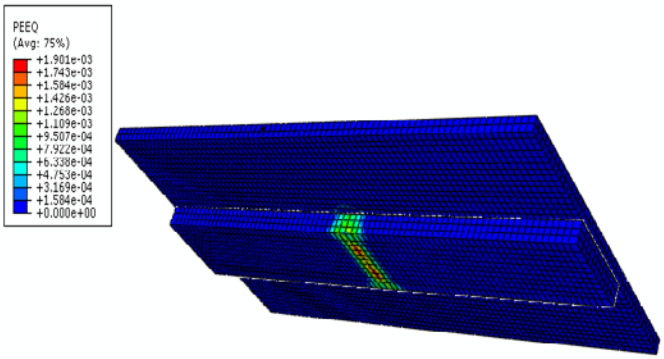


Fig. 16: Concrete plastic equivalent strain from compression, PEEQ=1.90E-3

ISBN: 978-1-61804-240-8

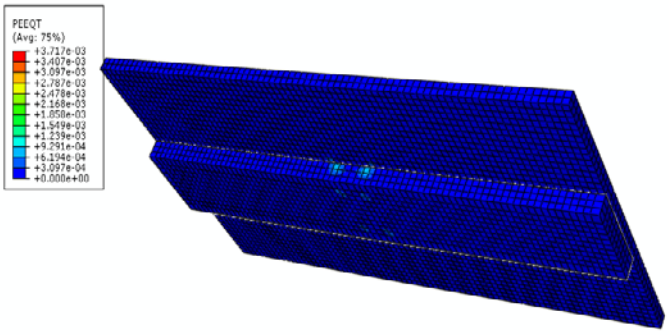


Fig. 17: Concrete plastic equivalent strain from tension, PEEQT=3.71E-3

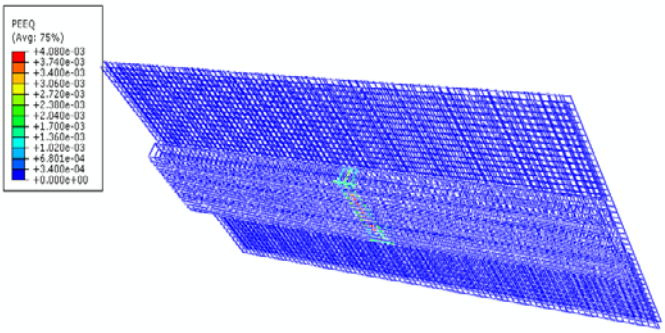


Fig. 18: Steel plastic equivalent strain, PEEQ=4.08E-3

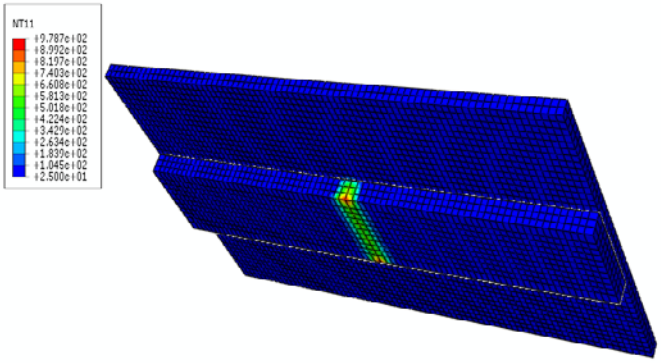


Fig. 19: Temperature distribution on concrete after two hours of fire, T_{max}=978.7ºC

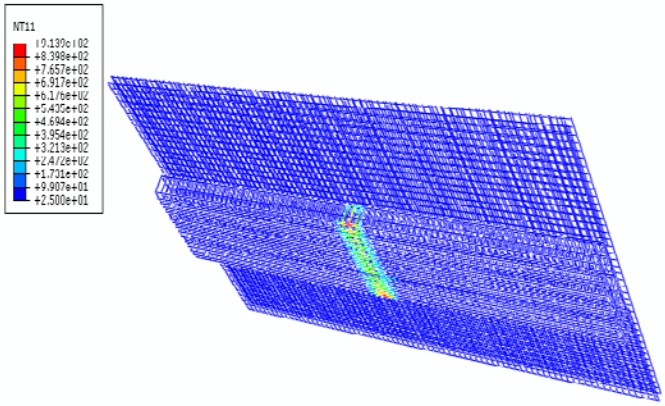


Fig. 20: Temperature distribution on steel after two hours of fire, T_{max}=913.9 ºC

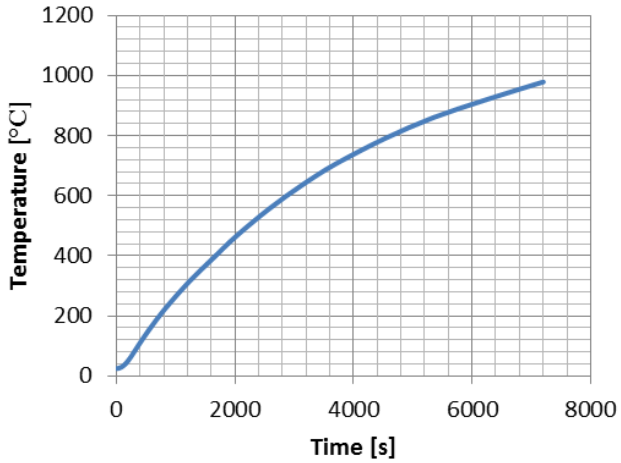


Fig. 21: Time versus maximum temperature curve on concrete

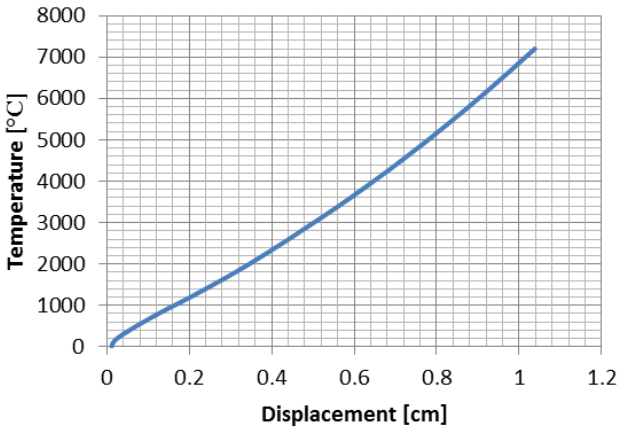


Fig. 22: Displacement – maximum temperature diagram

C. Scenario III

In the third scenario of fire it have been considered acting on a strip of 0.50 m wide near the support, aside the whole width of the flanged beam (Fig. 23).

Fig. 23: Scenario III of fire action

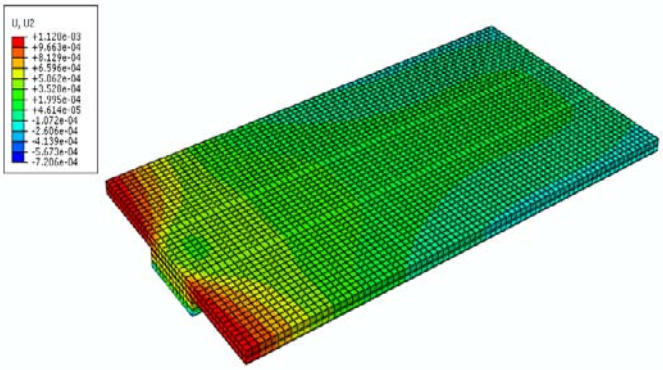


Fig. 24: Displacement after two hours of fire, d=0.072 cm

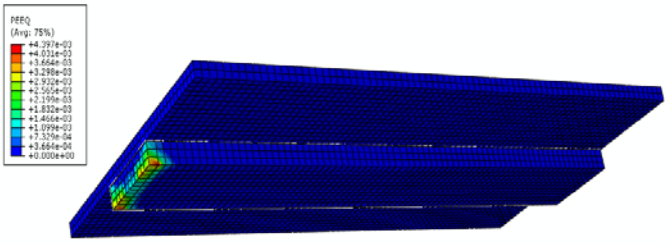


Fig. 25: Concrete plastic equivalent strain from compression, PEEQ=4.39E-3

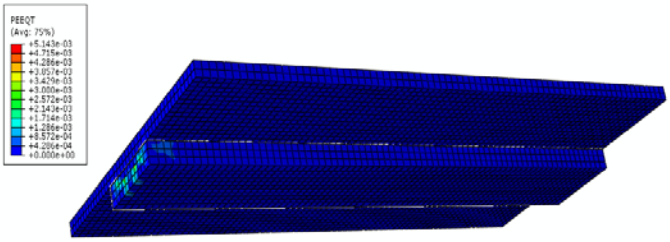


Fig. 26: Concrete plastic equivalent strain from tension, PEEQT=5.14E-3

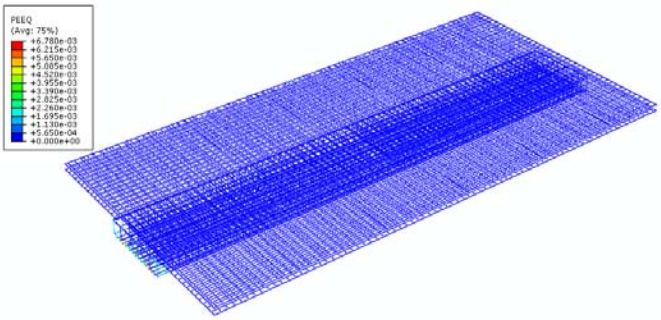


Fig. 27: Steel plastic equivalent strain, PEEQ=6.78E-3

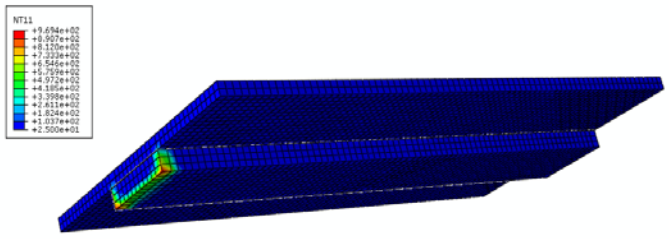


Fig. 28: Temperature distribution on concrete after two hours of fire, T_{max}=969.4ºC

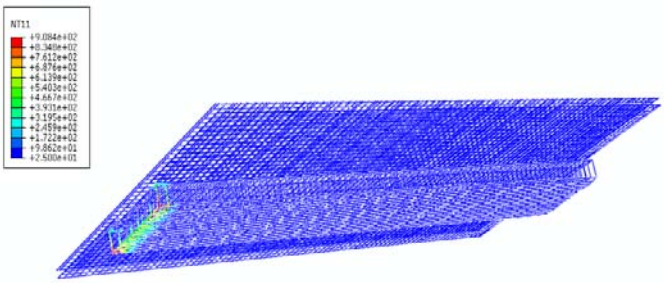


Fig. 29: Temperature distribution on steel after two hours of fire, T_{max}=908.4 ºC

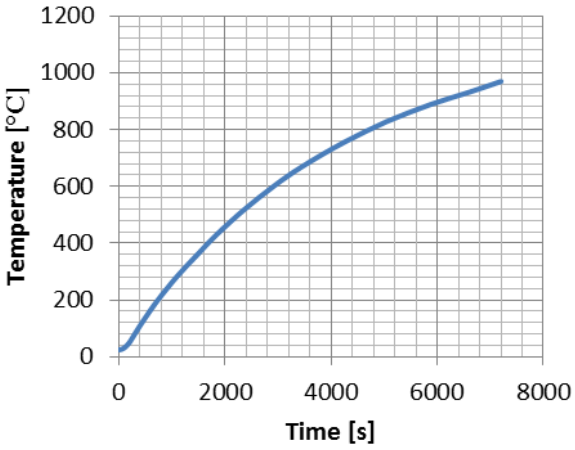


Fig. 30: Time versus maximum temperature curve on concrete

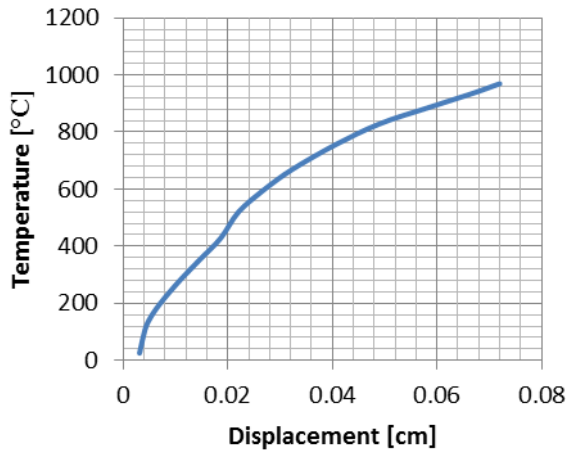


Fig. 31: Displacement – maximum temperature diagram

IV. Discussion

The three scenarios taken into consideration in the investigation presents possible situation of fire acting on the beam. When comparing results obtained for flanged wide reinforced concrete beam with respect to the independent wide reinforced concrete beam [5] one can remark similar behavior of the beams under the same external load, but deflection of the beam, internal stresses in concrete and reinforcements as well as internal temperature are of more reduced values.

Comparison of the results for the three scenarios is unnecessary since the load given by the fire in the first scenario is incomparable with the other two scenarios. Even so we can remark the increased risk for the stability of the element for scenario I since after one hour it reaches excessive deformation (beyond *l/100*), where the concrete cover of the tensioned reinforcements is already inexistent. For scenarios II and III the deformation occurred is almost negligible under the external and fire loads.

ISBN: 978-1-61804-240-8

V. Conclusion

The behavior of the flanged wide reinforced concrete beam subjected to fire according to the presented numerical investigation can be considered highly satisfying, taking in consideration the designed concrete cover and the imposed fire resistance of the element of 15 minutes. Fire resistance of the beam is improved when joint with the flanges is considered and pre-stressing for the beam is applied, avoiding excessive deflection even after one hour of fire load. Temperature in the reinforcements reaches dangerous values after one hour of fire load.

The displacements after two hours of fire are less for the wide beam interacting with the precast slab then the results only on the wide beam alone [5], 3 times less for the first and second scenario and almost 8 times less for the third scenario.

Modeling the interaction of the precast slab with the prestressed wide beam gives a more accurate representation on the behavior of the beam subjected to gravity and fire loads. For a better understanding of the mechanical behavior the authors will have to do some experimental investigations in order to see if the numerical model is close to reality.

References

[1] *** SR EN 1992-1-1-2004, Eurocode 2, Design of concrete structures. Part 1-1: General rules and rules for buildings, 2004.

[2] *** SR EN 1992-1-2-2004, *Eurocode 2, Design of concrete structures*. Part 1-2: General rules – structural fire design, 2004.

[3] ***. (1999). P118-99: Normativ de siguran□ă la foc a construc□iilor

[4] A. Puskas, Z. Kiss, "Testing of a wide reinforced concrete beam", *The 7th Central European Congress on Concrete Engineering*, Balatonfüred, Hungary, 22-23 September 2011, pp. 315-318

[5] A. Puskas, A. Chira, "Numerical Investigations on a Wide Reinforced Concrete Beam Subjected to Fire", *Proceedings of the 4th International Conference on Mathematical Models for Engineering Science - MMES '13*, Brasov, Romania, June 1-3, 2013, pp. 169-174, ISBN: 978-1-61804-194-4

[6] Z. Kiss, K. Bálint, A. Puskás, "Steel or concrete structure – prefabricated or cast in situ? The design of a multistory building in Bucharest for Kika", *III, Medunarodna Savetovanke*, Subotica, 8-9. Octobar 2009, p. 79-93.

[7] Abaqus. *Abaqus Analysis User's Manual*.

[8] I. Moga, L. Moga, "Heat flow simulation through the window together with the wall in which is fitted in", *IASTED conference "Applied Simulation and Modeling"* Palma de Mallorca, Spain, 29 – 31 August, 2007, ISBN: 978-0-88986-687-4

[9] A. Faris, A. Nadjai, S. Choi, "Numerical and experimental investigations of the behavior of high strength concrete columns in fire", *Elsevier, Engineering structures*, 2010.

[10] K. Venkatesh, R. Nikhil, "A simplified approach for predicting fire resistance of reinforced concrete columns under biaxial bending", *Elsevier, Engineering structures*, 2012.

[11] Qing-Hua Tan,Lin-HaiHan n, Hong-XiaYu, Fire performance of concrete filled steel tubular (CFST) column to RC beam joint, Fire Safety Journal, 2012

[12] Anil Agarwal , Lisa Choe , Amit H. Varma ,,Fire design of steel columns: Effects of thermal gradients'' *Elsevier, Journal of Constructional Steel Research*, 2014.

A. Puskás is Assistant Professor at Faculty of Civil Engineering of Technical University of Cluj-Napoca, Romania, since 2007. He received the B.S. and Ph.D. degrees in civil engineering in 1995 and 2012, respectively, from Technical University of Cluj-Napoca, Romania, and M.S. degree in Business Administration from Faculty of Business of Babes-Bolyai University, Cluj-Napoca, Romania, in 2005. In 2013 he graduated a Postgraduate Course in Sustainable Urbanization at Technical University of Cluj-Napoca, Romania and participated in a short course in Sustainability: Principles and Practice at Massachusetts Institute of Technology, Cambridge, United States.

He joined Technical University of Cluj-Napoca, Romania in 2003 as Teaching Assistant. From 2000 he have also worked as Structural Designer, leading or participating in design of several steel, concrete, masonry or wooden structured industrial and public buildings. Since 2005 he is also Technical Director of a privately owned construction company, with extensive activity in industrial and public building design and realization. He has authored more than 30 Journal and Conference papers. His current interests include pre-stressed concrete design, sustainable structural solutions, sustainability of structures and their environmental impact as well as waste recycling in construction industry.

Dr. Puskás is member of The International Federation for Structural Concrete, The American Concrete Institute, Association of Environmental Engineering and Science Professors, Romanian Green Building Council and Association of Structural Designer Civil Engineers.

A. Chira is Assistant Professor at Faculty of Civil Engineering of Technical University of Cluj-Napoca, Romania, since 2014. He received the B.S. and Ph.D. degrees in civil engineering in 2008 and 2011, respectively, from Technical University of Cluj-Napoca, Romania,

He joined Technical University of Cluj-Napoca, Romania in 2010 as Teaching Assistant. In 2007 he started to work as Structural Designer, being involved in the design of concrete, steel or masonry both industrial and public buildings.

Dr. Chira is member of research team ,,Computational Modeling and Advanced Simulation in Structural and Geotechnical Engineering''.

A new open source project for modeling and simulation of complex dynamical systems

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*—** OpenMVL project is a research project devoted to mathematical problems of equation-based modelling and simulation of complex dynamical systems. Open source tool *OpenMVLSHELL* developed by A.A. Isakov is presented. *OpenMVLSHELL* transforms a model written in Model Vision Language to a system of algebraic-differential equations. Hereby *OpenMVLSHELL* automatically builds, analyzes, reduces, and solves systems of algebraic-differential equations. The numerical software of *OpenMVLSHEL* is available for augmenting by users. User can test and compere effectiveness of his own methods with *OpenMVLSHELL's* methods using built-in set of test problems.

***Keywords*—** complex dynamical systems, modeling languages, equation-based models, hybrid systems.

I. Introduction

Visual tools for modeling and simulation of complex dynamical systems are used in research, industry, and education. Simulink (StateFlow, Simmechanics, SimPowerSystems, etc. - MathWorks), Dymola, Ptolemy are only a few well-known names of such tools [1]. It is possible marking out universal and unified tools among them. Universality assumes ability of building a model of any needed type using universal tool only. Commonality requires using unified modeling language for model specification and its well-defined interpretation (See Unified Modeling Language (UML) for example). Modeling and simulation of complex dynamical systems using universal and unified tools may be disjoint on stages:

- Specification of an equation-based model.
- Building a system of equations using component equations and connection equations.
- Reducing of a system.
- Numerical solution of current equations.
- Visualizing of solution (behavior of model).

Modules of tools for modeling and simulation of complex dynamical systems answerable for Building, Reducing, and Numerical Solution may be considered as special kind of numerical software. Numerical software and numerical libraries are generally accessible and have open source usually (Netlib for example).

Open source tool *OpenMVLSHELL* developed by A.A. Isakov deals with mathematical and numerical problems of modeling and simulation of complex dynamical systems. OpenMVL Project (http://dcn.ftk.spbstu.ru/) opens for Users numerical software of universal and unified visual tools MvStudium and Rand Model Designer (www.mvstudim.com; www.rand-service.com), developed by MvStudium Group.

II. Open source tools for modeling and simulation of complex dynamical systems

Modern tools automatize modeling and simulation of physical or technical, real or projectible complex objects. Complex objects demand complex models: classical dynamical and hybrid systems are among them. They keep on key role in equation-based modeling. Traditional problems of numerical solution of large scale and sparse hybrid algebraic-differential equations are supplemented by problems of their building, analyzing, and reducing. This is specific character and main distinction of numerical software for modeling and simulation of complex dynamical systems from traditional numerical libraries for solving algebraic, differential, and algebraic-differential equations. OpenMVL project is a research project devoted to mathematical problems of equation-based modelling and simulation of complex dynamical systems.
The OpenModelica Project [5, 6] has been choosing as prototype of OpenMVL [7].
OpenModelica:

- is open source and free of charge tool,
- provides users object-oriented technology of modeling (without graphical interface),
- allows playing with the model and carry out complex computational experiments.

If OpenModelica Project draws user's attention and demonstrates possibilities of object-oriented modeling, then OpenMVL deals with problems of numerical software needed for object-oriented modeling of complex dynamical systems only.
A hybrid system with local continuous behavior in the form of sparse and large scale algebraic-differential equations is the main mathematical model for complex dynamical systems. Blocks with «contacts-flows» connection equations («acausal» blocks [5]) and hybrid systems as their component model may

ISBN: 978-1-61804-240-8

cause difficulties if we want building executable code for all permissible current systems of equations for whole model beforehand (Fig.1).

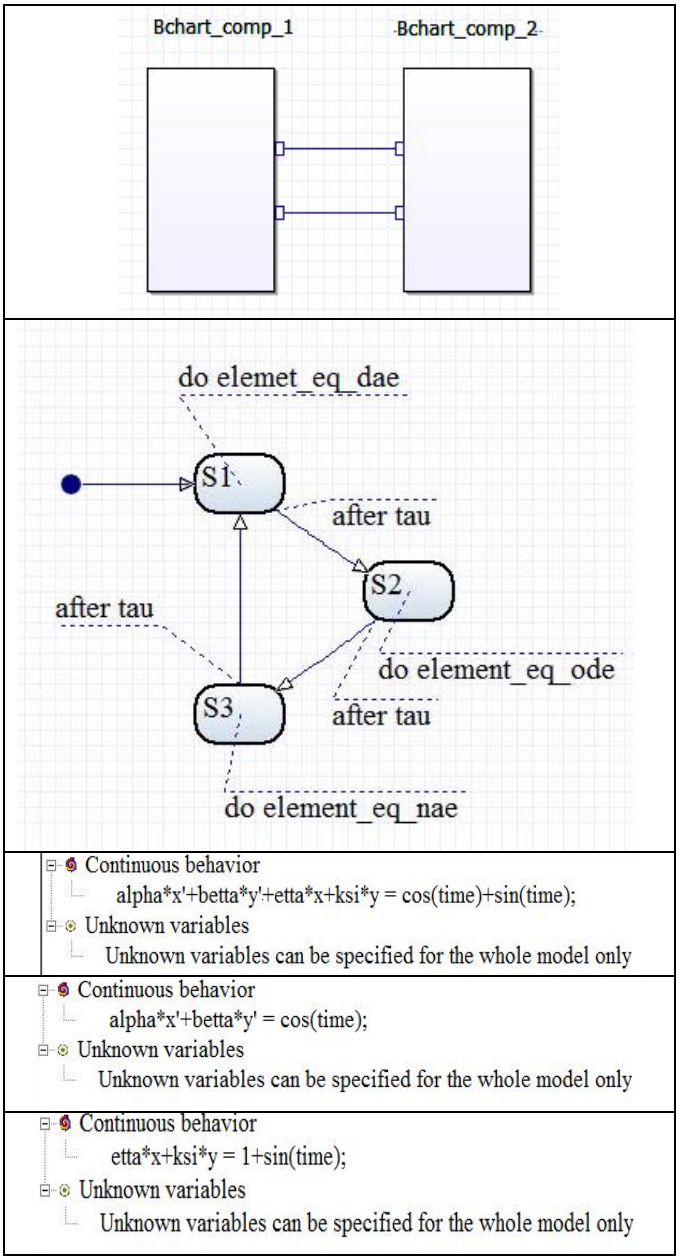


Fig. 1. «Acasual» blocks with «contacts-flows» and hybrid component equations.

Tools used Modelica overcome this problem restricting local behavior equations in hybrid system. Model Vision Language (MVL) has no such limitations. This is only one distinction between Modelica, MVL, and other Modeling Languages [1]. There are many others, and it is important starting discussion about Standard for tools of modeling and simulation of complex dynamical systems. Open source tools such as *OpenModelica* or *OpenMVLSHELL* may be considered as prototypes for workable Standard. Any Standard fixes progress and issues the challenges. Standard implies reproducibility modeling and simulation results for different tools if they follow Standard.

Standard should touch on classification of permissible:

- model types (systems of algebraic equations, system of ordinarily differential equations, systems of algebraic-differential equation),
- methods of decomposition and aggregation of component models (causal, «acausal» blocks, «agents» and so on) ,
- methods of analyzing, reducing and approximation of models,

from the point of view of numerical software for modeling and simulation of complex dynamical systems.

Standard is able to prescribe compulsory list of

- numerical methods,
- and instruments for computational experiments

for comparison results of simulation.

We are going argues only classification of models and list on numerical methods used in MvStudium Group's tools in this paper.

Component «Analyzer» recognizes a model types and component «Solver» suggests and calls acceptable numerical method in *OpenMVLSHELL*.

III. *OpenMVL* Project

OpenMVL Project with OpenMVLShell (Fig. 2) tool:

- is open source project,
- based on Model Vision object-oriented modeling Language with UML-based diagrams for hybrid systems (Behavior-Charts or for short B-Charts) for specification of a model under consideration (Fig. 3.).

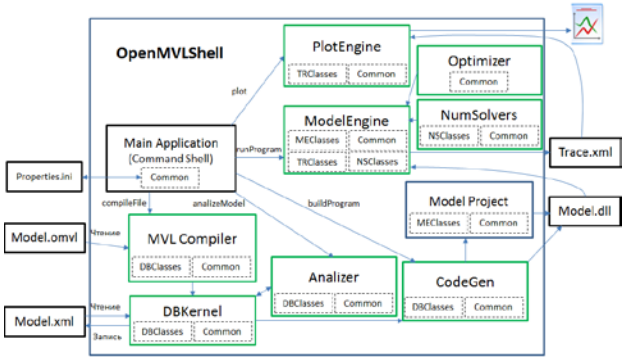


Fig 2. OpenMVLShell structure diagram

OpenMVLShell:

- has Editor for model specification;

- build executable code interacting with User;
- has Test-bench for plying with model and plotting results (Fig.4).

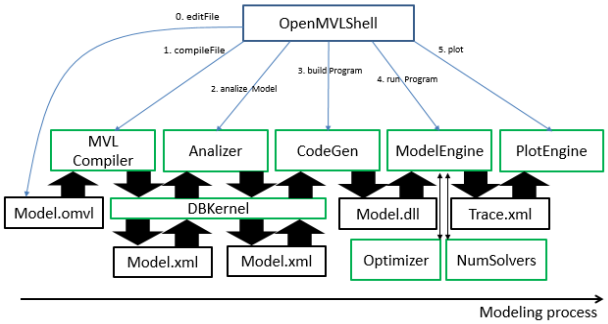


Fig 3. Modelling in OpenMVLShell.

For building and playing with Model:

- download a Model written in MVL (any Model Vision tool can save graphical specification of a model as MVL-text);
- build executable Model with help of OpenMVLShell;
- call executable Model.

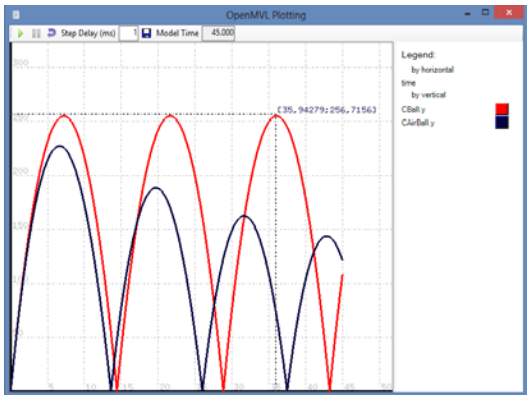


Fig 4. A numerical experiment in OpenMVLShell.

Component «Solver» calls numerical methods in compliance with results of syntactic analysis of equations written by User. In the rest of the paper:

$t \in \Re^1(time); \mathbf{x} \in \Re^n(unknowns);$

$Sub(\mathbf{x}) : \Re^n \to \Re^n(substitutions);$

$\mathbf{F}, \mathbf{G} : \Re^m \to \Re^k;$

$\mathbf{A}, \mathbf{B}, \mathbf{b}$ – *real matrices and vectors.*

«Analyzer» distinguishes between:

1. Systems of nonlinear algebraic equations (NAE) begin by substitutions $Sub()$.

$$\text{NAE: } \begin{cases} \mathbf{x}_1 = \mathbf{Sub}(\mathbf{x}_1) \\ \mathbf{F}(\mathbf{x}_1, \mathbf{x}_2, t) = 0 \end{cases}.$$

2. Systems of linear algebraic equations (LAE):

$$\text{LAE: } \begin{cases} \mathbf{x}_1 = \mathbf{Sub}(\mathbf{x}_1) \\ \mathbf{A} \cdot \mathbf{x}_2 = \mathbf{b}(t) + \mathbf{G}_1(\mathbf{x}_1) \end{cases}.$$

3. Systems of ordinary differential equations:

$$\text{ODE\_general: } \begin{cases} \mathbf{x}_1 = \mathbf{Sub}(\mathbf{x}_1) \\ \mathbf{F}(\mathbf{x}_1, \frac{d\mathbf{x}_2}{dt}, \mathbf{x}_2, t) = 0 \end{cases}$$

$$\text{ODE\_normal: } \begin{cases} \mathbf{x}_1 = \mathbf{Sub}(\mathbf{x}_1) \\ \frac{d\mathbf{x}_2}{dt} = \mathbf{F}(\mathbf{x}_1, \mathbf{x}_2, t) \end{cases}$$

$$\text{ODE\_linear: } \begin{cases} \mathbf{x}_1 = \mathbf{Sub}(\mathbf{x}_1) \\ \frac{d\mathbf{x}_2}{dt} = \mathbf{A} \cdot \mathbf{x}_2 + \mathbf{b} + \mathbf{G}(\mathbf{x}_1) \end{cases}$$

4. Systems of algebraic-differential equations:

$$\text{DAE\_semi-explicit: } \begin{cases} \mathbf{x}_1 = \mathbf{Sub}(\mathbf{x}_1) \\ \frac{d\mathbf{x}_2}{dt} = \mathbf{F}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 t) \\ 0 = \mathbf{G}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 t) \end{cases}$$

$$\text{DAE\_semi-explicit\_1: } \begin{cases} \mathbf{x}_1 = \mathbf{Sub}(\mathbf{x}_1) \\ \frac{d\mathbf{x}_2}{dt} = \mathbf{F}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 t) \\ \mathbf{A} \cdot \mathbf{x}_3 = \mathbf{b} + \mathbf{G}_1(\mathbf{x}_1, \mathbf{x}_2) \end{cases}$$

$$\text{DAE\_semi-explicit\_2: } \begin{cases} \mathbf{x}_1 = \mathbf{Sub}(\mathbf{x}_1) \\ \mathbf{B} \cdot \frac{d\mathbf{x}_2}{dt} = \mathbf{A} \cdot \mathbf{x}_2 + \mathbf{F}(\mathbf{x}_1, t) + \mathbf{b} \end{cases}$$

ISBN: 978-1-61804-240-8

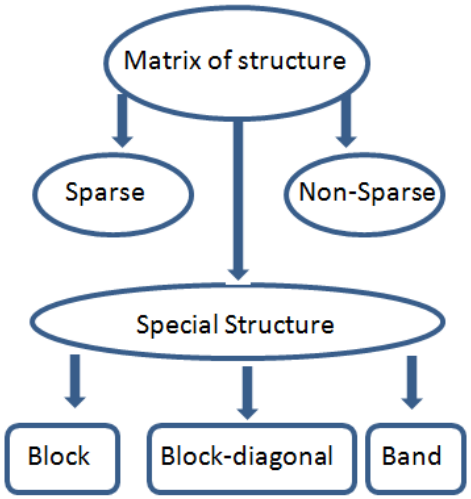


Fig 5. Possible structures of solved systems.

Effectiveness depends on structure of solved system. OpenMVLShell distinguishes between sparse and non-sparse systems, and systems with the special structure (Fig. 5). Solvers of systems of linear algebraic equations take into account their sparseness.

«Solver» for NAE, ODE, and DAE can call «automaton», standard (built-in), users', and debugging methods (Fig. 6). «Automaton» is used on default. The goal of «Automaton» is to find solution at any cost. It may be a set of sequentially executable software implementation of numerical methods.

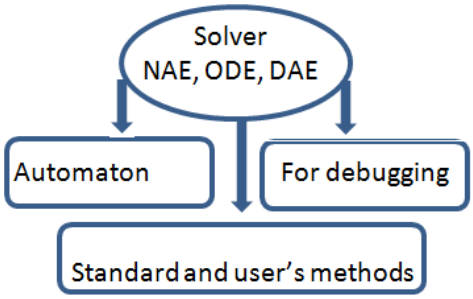


Fig. 6. Structure of a «Solver».

OpenMVLShell and all MvStudium Group tools have two-level numerical library. Lower level contains software implementation of numerical methods for NAE, LAE, ODE, DAE written in Fortran.

NAE: Newton and Powell methods.

LAE: LINPACK solvers for non-sparse systems and systems with special structure (band and so on); Sparse Solvers.

ODE: Solvers from ODEPACK [9]; Hairer and Wanner Solvers from [8].

DAE: ODEPACK, DASSL [10], DASPK [11].

IV. Conclusion

Number and complexity of tools for modeling and simulation of complex dynamical systems increases permanently. They should be standardized. OpenMVL is open project used now for research and education [2, 3, 4] (http://dcn.ftk.spbstu.ru/). We suggest using it as start point for future Standard. Joint us!

References

[1] Popper N., Breitenecker F., Extended and Structural Features of Simulators. Simulation News Europe, 2008 - Dec., pp.27-38.

[2] Isakov A.A., Senichenkov Yu. B. OpenMVL is a tool for modelling and simulation. //Computing, measuring, and control systems, St. Petersburg: SPBSPU. - 2010 - pp. 84-89.

[3] Isakov A.A., Senichenkov Yu. B. Program Complex OpenMVL for Modeling Complex Dynamical Systems // «Differential equations and control processes» Journal, Mathematics and Mechanics Faculty of Saint-Petersburg State University, St. Petersburg - 2011.

[4] Isakov A.A., OpenMVLShell – research tool for modeling and simulation. // All-Russian competition for students and post-graduate students "Innovation technologies in Education", Belgorod state National University - 2011 - pp. 11-15.

[5] Fritzson P. Principles of Object-Oriented Modeling and Simulation with Modelica 2.1, Wiley-IEEE Press, 939 pages.

[6] Fritzson. P. Introduction to Modeling and Simulation of Technical and Physical Systems with Modelica 2011. Wiley-IEEE Press , 232 pages.

[7] https://www.modelica.org/publications

[8] Hairer E., Wanner G. (1996), Solving oridinary differential equations II. Stiff and differential-algebraic problems, Springer-Verlag, Berlin.

[9] A. C. Hindmarsh, "ODEPACK, A Systematized Collection of ODE Solvers," in Scientific Computing, R. S. Stepleman et al. (eds.), North-Holland, Amsterdam, 1983 (vol. 1 of IMACS Transactions on Scientific Computation), pp. 55-64.

[10] http://www.netlib.org/ode/ddassl.

[11] http:/www-users.cs.umn.edu/~petzold/DASPK

Isakov Alexander 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.

Timed ignition of separated charge

Michal Kovarik

***Abstract*—**The typical design of the weapon system using separated propellant charge is not beneficial enough to the muzzle velocity because of the ignition delay of additional charge. The new method of the precisely timed electric ignition has been developed. Contributions of the timed ignition to the ballistic output of weapon have been modelled and model has been supported by experimental shootings results. The muzzle energy of the projectile fired with timed ignition has been increased by more than 25 percent compared to the projectile fired from typical weapon.

***Keywords*—**Interior ballistics, gun, separated charge, muzzle velocity, timed ignition, electric ignition.

I. Introduction

THE concept of gun using separated propellant charge is described in [1,4,6,7], it is a weapon system designed to produce high muzzle velocity alongside preserving the maximum pressure limit. The design of the weapon is distinguished with the use of at least one additional chamber placed alongside the barrel bore. The additional propellant charge is placed inside the additional chamber and ignited during the shot. The only method of ignition of the additional propellant had been the way of utilization of hot powder gases from barrel bore produced during the shot by burning of basic charge. This method does inflict ignition delay. The ballistic efficiency does drop as a result of any delay of ignition. If the new method of electronically driven timed ignition is used, the ballistic performance of the system is increased significantly, whereas the main advantage of not breaking the maximal ballistic pressure limit is maintained.

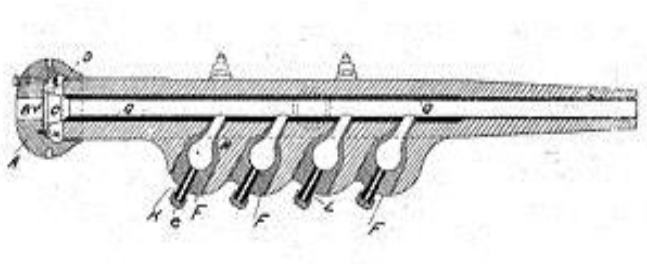


Fig. 1 – Historical concept of weapon system using separated propellant charge with ignition by hot gases (by Lyman and Haskell in 1883).

M. Kovarik is with the Weapons and Ammunition Department, Faculty of Military Technology, University of Defence, Brno, Czech Republic (e-mail: 4401@unob.cz).

II. Problem formulation

The ballistic performance of the weapon could be measured by the kinetic energy of the projectile. If the energy conservation equation (1) is observed and the work done by propellant gases is expressed, the weapon performance could be enhanced by increasing the ballistic pressure of propellant gases accelerating the projectile:

$$s\int_0^{l_{ú}} p \, \mathrm{d}l = \frac{1}{2} m_q v_{ú}^{\,2}, \qquad (1)$$

where cross-section surface is represented by s, ballistic pressure by p, the l does stand for projectile travel, $l_{ú}$ for barrel length, m_q is mass of projectile and $v_{ú}$ is projectile's muzzle velocity. However the high pressures inside the barrel are limited to the highest value of pressure p_{max}, because of the barrel strength limitation. This fact does lead to pressure development optimization resulting in the ideal rectangular pressure course illustrated in the following figure.

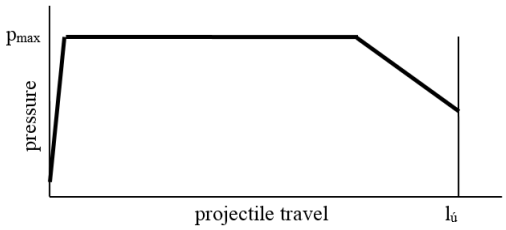


Fig. 2 – Ideal pressure course.

Typical pressure courses are not closed enough to the ideal one, it could be refined by the way of placing additional charges alongside the barrel (see Fig.1) to help reduce the effect of the pressure curve decreasing branch.

The process of shot evolution does begin ordinary, starting with initiation of basic propellant charge placed inside the main cartridge chamber. After the certain amount of time projectile is moving inside the barrel bore, being accelerated by the propellant gases of the basic charge. While the projectile does travel, it releases volume in which the basic propellant charge does burn and does enable propellant gases to expand and ballistic pressure to drop. At this time projectile should pass the flash hole and should connect the volume of burning gases to the volume of additional chamber with additional propellant. Hot gases intrude the additional chamber through flash hole and ignite additional propellant in order to produce more gases and reverse the pressure drop behind the projectile.

ISBN: 978-1-61804-240-8

The described process of ignition of additional propellant is significantly time-consuming and does impede full utilization of system with separated propellant charge (see Fig.4). If the ignition of additional propellant charge could be timed precisely to the moment when projectile does pass the flash hole, the obstacle of ignition delay would be removed. Therefore the pressure course should be improved and higher muzzle velocity should be produced.

III. Model

Following the objective of physical action description the mathematical model was assessed. The core of the model of is based on the thermodynamical description of geometrical theory of propellant combustion consisting of following equations:

$$\psi = \kappa z + \kappa\lambda z^2 + \kappa\mu z^3, \tag{2}$$

$$p = \frac{f\omega\psi - \frac{1}{2}\Theta\varphi m_q v^2}{s\left(l_\psi + l\right)} + p_z, \tag{3}$$

$$\varphi m_q \frac{\mathrm{d}v}{\mathrm{d}t} = s\left(p + p_z\right), \tag{4}$$

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \frac{u_1\left[m + \left(p + p_z\right)^\nu\right]}{I_k}, \tag{5}$$

$$l_\psi = l_0\left[1 - \frac{\Delta}{\delta} - \Delta\psi\left(\alpha - \frac{1}{\delta}\right)\right], \tag{6}$$

$$\frac{\mathrm{d}l}{\mathrm{d}t} = v, \tag{7}$$

$$T = T_V\left[1 - \frac{1}{\psi}\left(\frac{v}{v_{\lim}}\right)^2\right], \tag{8}$$

where ψ stands for the relative quantity of burnt-out powder, κ, λ, μ are geometric characteristics of the powder grain, z is the relative burnt thickness of the powder grain, p is the ballistic pressure and p_z is the primer pressure, f is the specific energy of propellant, ω is the mass of the propellant charge, mq is the mass of the projectile, φ is the fictivity coefficient of the projectile, s is the cross-section area of bore, m, u_1, ν are rate of burning coefficients, I_k is the pressure impulse of ballistic pressure, l is the projectile travel as l_ψ is the relative length of initial combustion volume, Θ is the heat parameter of powder expansion, Δ stands for the loading density, δ for the powder mass density and α is the covolume of powder gases.

This model of the interior ballistics has to be modified in order to description of the weapon utilizing additional serial chambers. In the case of system with additional propellant charge, every combustion space has to be described separately. The main issue is mathematical description of the propellant gases flow between the chamber and barrel bore, the problem is mainly represented by the assuming the flow rate of the propellant gases through the ignition channel and the generation of the propellant gases mixtures.

The flow rate of the propellant gases m_{pr} is dependent on the difference in the internal pressure in barrel bore p and internal pressure in the additional chamber p_1. If the condition

$$p_1 \le p\left(\frac{2}{\Theta+2}\right)^{\frac{\Theta+1}{\Theta}} \tag{9}$$

is met, the propellant gases will flow by the critical velocity and flow rate will be

$$m_{pr} = \xi^* \kappa_k \frac{p}{\sqrt{f\frac{T}{T_v}}}, \tag{10}$$

where the outflow coefficient ξ^* is given by

$$\xi^* = \sqrt{\Theta+1}\left(\frac{2}{\Theta+2}\right)^{\frac{\Theta+2}{2\Theta}}. \tag{11}$$

If the condition (9) is not met, the coefficient ξ^* is substituted by the outflow coefficient ξ given by

$$\xi = \sqrt{\frac{2\left(\Theta+1\right)}{\Theta}\left[\left(\frac{p_1}{p}\right)^{\frac{2}{\Theta+1}} - \left(\frac{p_1}{p}\right)^{\frac{\Theta+2}{\Theta+1}}\right]}. \tag{12}$$

The coefficient κ_k defines local flow conditions of the particular channel design, it has been assessed according to the Cibulevky theory.

The equations for mathematical description of the propellant gases mixture are based on the Dalton's law $p = \sum p_i$, the total pressure exerted by the mixture of non-reactive gases is equal to the sum of the partial pressures of individual gases. When β_i represents partial mass of i-th portion

$$\beta_i = \frac{\omega_i}{\omega}, \tag{13}$$

then the mixture could be characterized by the quantities:

$$f = \frac{1}{\psi}\sum \beta_i \psi_i f_i, \tag{14}$$

$$\Theta = \frac{\psi f}{\sum \beta_i \psi_i \frac{f_i}{\Theta_i}}, \tag{15}$$

$$\alpha = \frac{1}{\psi}\sum \beta_i \psi_i \alpha_i, \tag{16}$$

and

$$T = \frac{1}{\psi r}\sum \beta_i \psi_i r_i T_i, \tag{17}$$

where gas constant r is equal to $r = \psi^{-1}\sum\beta_i \cdot \psi_i \cdot r_i$.

Described mathematical model conforms the preconditions and simplifications of no heat transfer between barrel and propellant gases, uniform dispersion of the propellant burning products inside the barrel bore, neglect of wave process in the gases, merely simplification of the influence of the primer,

quantification of energetic losses are by means of the fictiveness coefficient, minimization of the gases leakage, constant the heat capacity ratio during the shot generation and others which are perceptible from the equations notation itself, for example combustion rate law $u = u_1(m+p^v)$.

The solution of the presented system could be found by the numerical method only. The Runge-Kutta 4th order method has been used and the appropriate procedures were built under the environment of Matrix Laboratory software.

Parameters and constant values were given by characteristics and properties of the gun and ammunition used during following experimental shootings. The calculated pressure course of chosen gun fired as an ordinary weapon (with basic propellant charge only) is illustrated in following figure.

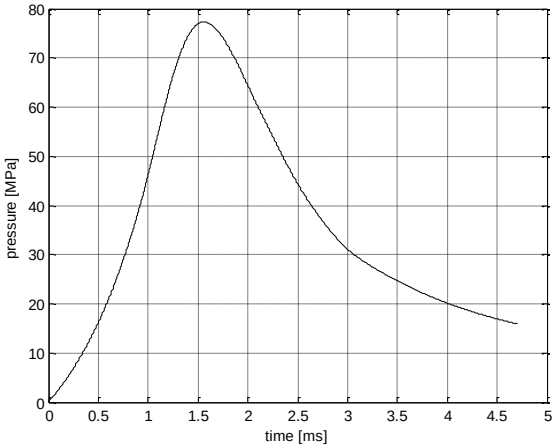


Fig.3 – Pressure course of chosen gun.

The comparison of the numerical model results for the chosen case of weapon system designed as a weapon system with one additional propellant charge utilizing the 'gas ignition' and the same weapon system with the precisely timed ignition are depicted in the next figures (the safe reduced experimental additional propellant charge was modelled for correspondence between the computation and the experiment).

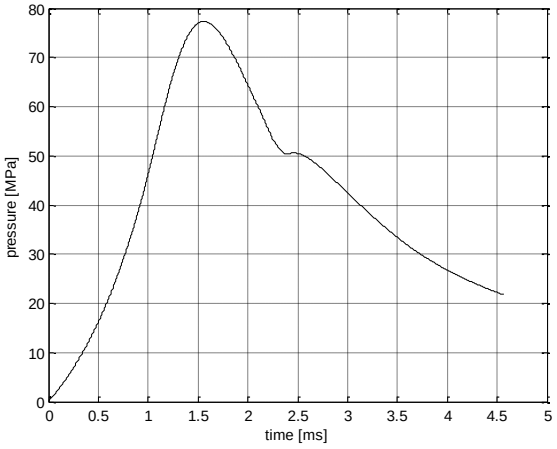


Fig.4 – Pressure course in the case of 'gas ignition'.

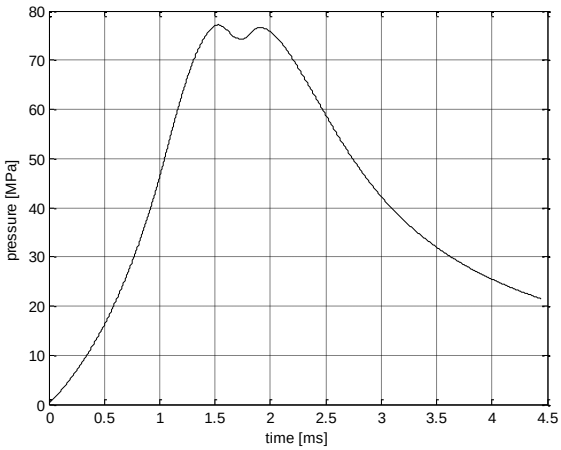


Fig.5 – Pressure course in the case of timed ignition.

The numerical analysis of the mass of additional propellant charge and time of ignition for the chosen weapon system was made and the resulting pressure course for the most profitable feasible solution is shown in the Fig. 6.

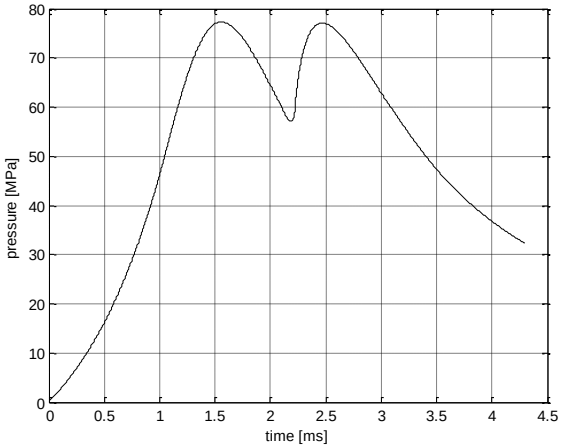


Fig.6 – The best feasible pressure course.

Numerical values of the calculated maximum pressures and relevant muzzle velocities are summarized in the following table.

Table I. – Calculated values of velocities and maximal pressures.

Charge arrangement	Velocity [m·s^{-1}]	Maximum pressure [MPa]
Typical gun	543	77,3
'Gas ignition'	613	77,3
'Timed ignition'	632	77,3
'Best solution'	722	77,2

ISBN: 978-1-61804-240-8

IV. Experiment

A. Weapon

The experimental weapon was the modified smooth-bored barrel of calibre 12.7×107 and length of 100 bores. The three additional chamber were placed alongside the bore, just first was utilized.

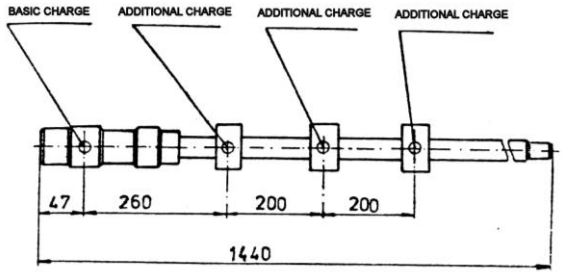


Fig.7 – The best feasible pressure course.

The additional chamber was designed as a hollow volume inside the screw and placed into the first position. Flash holes were drilled to the barrel and inclined to the barrel bore axis. The barrel was supported at the placement of additional chamber by the steel rings and the additional chamber was mechanically sealed. The electric ignition wire passage opening in the screw's head was left.

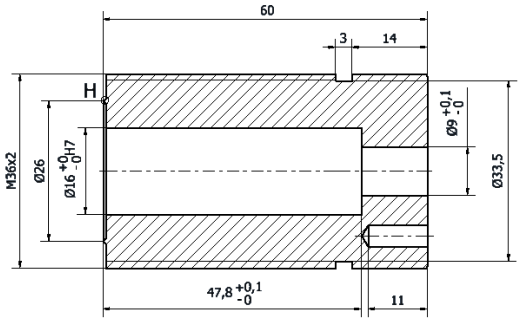


Fig.8 – The additional chamber.

The barrel was fit to the universal breech UZ-2000 and gun carriage STZA providing remote initiation.

B. Ammunition

The experimental ammunition consist of projectile, basic charge and additional charge. Monolithic brass projectiles were used instead of the common line production because of the mass variation.

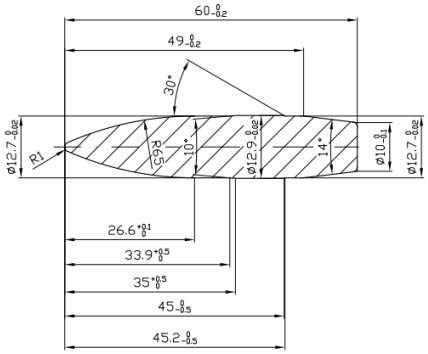


Fig.9 – The experimental projectile.

The basic charge was placed inside the brass case adjusted to the pressure measurement according to C.I.P. standards. The propellant was 7-hole single based deterred powder. The mass of the basic charge was set to constant 9 g. Every charge was reweighted.

The additional propellant charge was integrated into the screw of additional chamber. It consist of electric initiator F3 (EMS PATVAG) in the hollow volume of the support casing, composite contact casing of crezol-formaldehyde resin, and the assembly of contact screw placed inside the composite pin. In the case of 'gas ignition' the assembly without electric equipment was used.

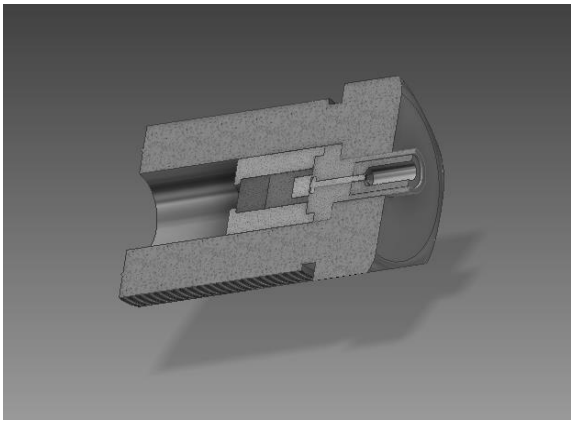


Fig.10 – The additional charge assembly.

The double-based small-grained spherical powder with the high vivacity was used as the additional propellant. Because of the manipulation reasons it was placed inside the micro-foil pre-perforated bag.

C. Arrangement

The arrangement of the measuring chain is depicted in the Fig. 12. The action started by the time delay setup on the timing unit from personal computer by the means of the developed digital interface. Then the sufficient electric charge was stored in the capacitive high-voltage unit. The mechanical initiation of the basic propellant charge caused the rise of the ballistic pressure. The value of the pressure was measured by the piezoelectric sensor. The charge signal was transformed to voltage signal and observed by the timing unit in real-time. After the pressure threshold had been reached, the timing unit counted delay and sent the control signal to the switch unit. The high-voltage circuit was closed and initiator ignited the additional propellant. The laser gates were used to measure projectile velocity and ballistic analyzer recorded signals.

ISBN: 978-1-61804-240-8

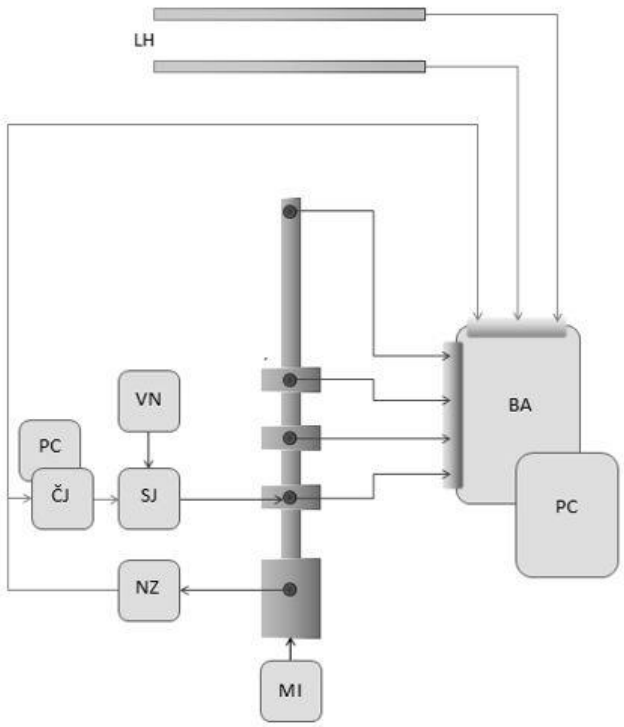


Fig.12 – The measuring chain (PC-personal computers, BA-ballistic analyzer, NZ-charge-voltage transducer, ČJ-timing unit, SJ-switch unit, VN-high voltage unit, LH-laser gates, arrows-signal lines).

D. Results

The experimental comparison between systems of typical gun, 'gas ignition' and 'timed ignition' has been made. The same types and amounts of propellants were used. The pressure courses are compared in the Fig. 13 (refers to Fig.3-5). Analogously the 'best solution' pressure course was measured (Fig.14 does relate to Fig.6).

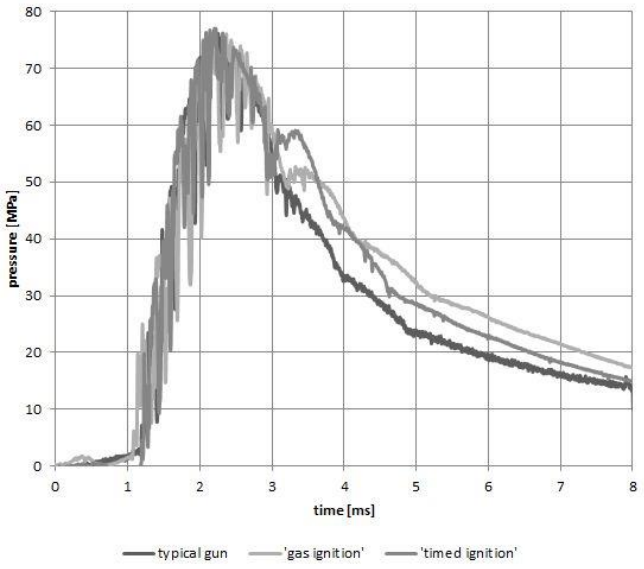


Fig.13 – The comparison of measured pressure courses.

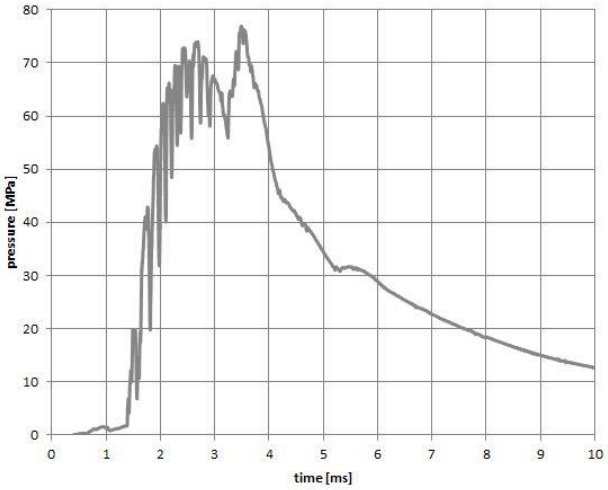


Fig.14 – The measured best feasible solution pressure course.

Numerical values of the measured maximum pressures and relevant muzzle velocities are summarized in the following table.

Table II. – Measured values of velocities and maximal pressures.

Charge arrangement	Velocity [m·s^{-1}]	Maximum pressure [MPa]
Typical gun	544	76,7
'Gas ignition'	547	76,1
'Timed ignition'	557	77,1
'Best solution'	613	76,9

V. Conclusion

The timed ignition of separated propellant charge does result in the muzzle velocity increase compared both to the typical gun and to the gun utilizing the gas ignition of the separated propellant charge. Although the reliability of the model is limited, the effect to pressure curves is in accordance. If the more detailed description of the weapon system with separated propellant charge was focused, it should incorporate the hydrodynamic theory, though the solution would become complicated due to its demands for values acquaintance of considerable amount of unknown quantities. The utilization of one additional chamber with timed ignition does produce equivalent increase in muzzle velocity compared to use of three additional chambers ignited by hot gases [13].

REFERENCES

[1] M. Kovarik, "Serial chamber gun (Published Conference Proceedings style)", in *Proceedings of the 17th International Conference New Trends and Research of Energetic Materials*, Pardubice, 2014.

[2] M. Kovarik, "Projectile velocity increase by the use of separated propellant charge (Published Conference Proceedings style)", in *Proceedings of the 4th International Conference on Fluid Mechanics and Heat & Mass Transfer)*. World Scientific and Engineering Academy and Society, Dubrovnik, 2013.

[3] E. Davis, "Advanced propulsion study", Warp Drive Metrics, Las Vegas, 2004.

[4] L. Stiefel, "Gun propulsion technology (Book style), New Jersey, AIAA,1997.

[5] A. Horst, "Brief Journey Through The History of Gun Propulsion (Book style)", Army Research Lab, Aberdeen proving ground MD, 2005.

[6] A. Siegel, Theory of high-muzzle-velocity guns, *AGARDogograph Vol 91*, US Naval Ordnance Laboratory, Silver Spring, 1965.

[7] А. Златин, "Баллистические установки и их примениние в экспериментальных исследованниях (Book style)," Наука, 1974.

[8] P. Oosthuisen, "Compressible fluid flow (Book style)", New York, McGraw-Hill, 1997.

[9] E. Hairer, The numerical solution of differential-algebraic systems by Runge-Kutta methods. Berlin, Springer, 1989.

[10] J. Kusák, Thermodynamical devices of high velocity of ejected body (Study style)", Prototypa ZM, Brno, 2007.

[11] L. Jedlička, "Modelling of pressure gradient in the space behind the projectile", In: *Proceedings of the 7th WSEAS international conference on System science and simulation in engineering*. World Scientific and Engineering Academy and Society, 2008.

[12] M. Hajn, "Possible advantages of the weapon system using separated propellant charge (Published Conference Proceedings style)", in *Proc. Armament and Equipment of Land Forces*, Lipt. Mikuláš, 2012.

[13] M. Hajn, "Projectile velocity increase by the use of separated propellant charge (Published Conference Proceedings style)", in *Proceedings of the 4th International Conference on Fluid Mechanics and Heat & Mass Transfer)*. World Scientific and Engineering Academy and Society, Dubrovnik, 2013.

[14] S. Ying, "The mechanism analysis of interior ballistics of serial chamber gun (Report style)", Ballistic Laboratory of China, 2008.

Analysis of physical health index and Children obesity or overweight in Western China

Jingya Bai, Ye He*, Xiangjun Hai, Yutang Wang, Jinquan He, Shen Li
Faulty of Medicine, Northwest University for Nationalities, China
(*Corresponding Author: jingyabai@gmail.com)

Abstract—To investigate and analyze the obesity or overweight situation among 7 to 12 year old children in 2013 and its relation to blood pressure, pulse, vital capacity index. To provide scientific evidence for different level governments to make corresponding preventative strategies and intervention measures. This paper take 7 to 12 year old children in 11 primary schools as investigation objects; measure their height, weight, blood pressure, pulse and vital capacity index and screen the overweight and obesity according to the standard made by WGOC. The overweight rate of 7 to 12 city boys, city girls, country boys, country girls in 2013 is respectively 14.54%,7.06%,7.35%,1.97%. The obesity or overweight rate of them is respectively 23.85%,13.02%,11.38%,3.82%. The overweight rate and the obesity rate have different trends as the age grows. The overweight rate and the obesity rate at different ages are different in genders and town-country difference. The overweight children's and obesity children's blood pressure, pulse of are higher than the normal children's, but their vital capacity index is lower than the normal children's. BMI is very positively related to the diastolic blood pressure and systolic blood pressure (except girls), while BMI is obviously negatively related to vital capacity index. The school and the children's parents cooperate to strengthen the food management of obesity or overweight children, train them good dietary habits, do exercise to build their physique, improve their health level.

Keywords: Children obesity or overweight, physical health index, Western China

I. INTRODUCTION

In recent years, as the dietary pattern and life style changes and the living standards improves, Childhood obesity is developing in Children, which seriously influences the children's health, and it has been an important public health matter. According to the survey, childhood obesity cannot only cause social psychological problems to the children [1], but also increase the risk of obesity to them in their adulthood and blood pressure, diabetes, dyslipidemia and other diseases related to obesity which often happen in adulthood occur in their childhood [2-5].

In [11], in comparisons among age−sex−BMI percentile groups, systolic and diastolic blood pressure values were higher in obese and overweight groups than in normal weight groups for both sexes. Although BMI among girls was higher than among boys in all three percentile groups, there were no significant differences between sexes with respect to blood pressure values. The present findings emphasize the importance of the prevention of obesity in order to prevent future related problems such as hypertension in children and adolescents. In [12], this paper aimed to investigate the ability of BMI and waist circumference, single and combined, in identifying children who are at risk of hypertension and in influencing absolute blood pressure values. High blood pressure is strongly associated with excess weight. Waist circumference improves the ability of BMI to identify hypertension in obese children. Waist circumference is related to absolute blood pressure values in all weight classes. In [13], this paper aimed to determine the prevalence of overweight in US children using the most recent national data with measured weights and heights and to examine trends in overweight prevalence. The prevalence of overweight among children in the United States is continuing to increase, especially among Mexican-American and non-Hispanic black adolescents. In [14], this paper aimed to examine the extent of blood lipid abnormalities in overweight children and to determine whether the prevalence of dyslipidemia is different in overweight children with elevated blood pressure (BP) compared with overweight children with normal BP (NBP). The high prevalence of dyslipidemia found in this overweight sample supports recent recommendations to collect plasma lipid levels in not only overweight children with BP≥90th percentile but also in all overweight children. In [15], this paper aimed to examine tracking and predictiveness of childhood lipid levels, blood pressure, and body mass index for risk profile in adulthood and the best age to measure the childhood risk factor levels. Childhood blood pressure, serum lipid levels, and body mass index correlate strongly with values measured in middle age. These associations seemed to be stronger with increased age at measurements. In [16], although the influence of obesity on ventilatory function has long been recognized, the nature of the relationship and the mechanisms are not yet clear. The purpose of this report was to examine the effects of overall obesity and fat distribution on ventilatory function. Body fat distribution has independent effects on ventilatory function after adjustment for overall obesity in men. The finding that age modifies this association has implications for future research.

The paper analyzed the 7-12 year old children's physical health monitoring data in Lanzhou, gained the children's overweight, obesity popularity situation among different economic social groups in Lanzhou cities and countries, analyzes the relation between childhood obesity and blood pressure, pulse, vital capacity index and then provide scientific evidence for different level government to make corresponding preventative strategies and intervention measures.

II. 2. Research objects and methods

A. Research objects

Through taking whole group at random as sample, classify Lanzhou according to the town and the country. According to the social economic and cultural development level, divide them into up, middle and low level. According to the requirements of National Student Physique and Health Research Group, it choose 4188 students at 7 to 12 years old in 11 schools as researching objects, among which are 1279 city boys, 991 city girls, 1116 country boys and 812 country girls.

B. Investigation Methods

The age of children is full age according to the birthdates and the measure dates. It strictly obeys the National Student Physics and Health Monitoring Implementing Rules to measure their heights, weight, blood pressure, pulse and vital capacity. The survey screws are professional medical personnel who have taken uniform training, and they use the same mode measuring apparatus. Before being used, the apparatuses has been checked and corrected in uniform, and the site quality control measures all meet the demand. They applies vertical height and weight complex measurement apparatus, the unit of height being m (meter) and being precise to 0.01m and the unit of weight being kg (kilogram) and being precise to 0.1 kg. When measuring the blood pressure, the children tested should sit down and be quiet with the up arm at the same level as the heart, measure the brachial artery on right up arm for twice, and take the mean value of the two continuous tests as the tested children's blood pressure (kpa).It applies cylinder vital capacity measurement apparatus to measure vital capacity. Before being used, the cylinder vital capacity measurement apparatus should be checked and corrected, measure it for three times and take the largest vital capacity value as the vital capacity value, finally account the vital capacity index[vital capacity (mL)/(kg)]. When measuring the pulse, the checkers put their fore-fingers and middle fingers onto the cross stripes in wrist of the children and near the thumbs. Normal pulse is in regular rhythm and uniform force, the fingers are flexible.

C. Diagnostic Criteria

Count the index (BMI) by using height and weight. BMI= Weight/height2 (kg/m^2). Screen the overweight and obesity according to the standard made by Working Group of Obesity in China ,WGOC. In order to get rid of the interference caused by malnutrition, they set up a comparison between the malnutrition children and the normal children according to the *Students' Standard Height and Weight Table*.

D. Statistical Method

Input the data by using Excel software and analyze the data by using SPSS19.0 software. The measurement data shall be shown in mean data±s. The comparison of rate$\chi 2$ to check and exact probabilities in fourhold table. T check will be used to analyze when comparison between two groups is made, but one-way ANOVA T Test is will be used to analyze between more than two groups. BMI and blood pressure, pulse, vital capacity index will be used as Pearson to analyze.

III. 3. Results

A. Overweight and obesity situation

1) General situation

As shown in Table 1, the city boys' overweight rate, obesity rate and overweight+ obesity rate are highest, while the country girls' are lowest. By comparison, we learn that the city boys overweight rate, obesity rate, overweight + obesity rate is respectively higher than the city girls' ($P<0.01$) and the country boys' ($P<0.01$). The city girls' and the country boys' overweight rate, obesity rate, overweight+ obesity possible rate are higher than the country girls ($P<0.01$).

2) The overweight, obesity possible rate in different gender children

As shown in table 1, by comparing the city boys' and the city girls', the boys' overweight rate at different age groups are higher than the girls, among which the 7 year old boys' overweight rate is a bit higher than the 7 year old girls; the 8 to 12 year old boys' overweight rage is obviously higher than girls (more than doubled.) The boys' obesity rate at different ages are higher than the girls, among which 9 year old boys' obesity rate is close to the girls', boys' obesity rate at other ages are 2 to 7 percent higher than the girls'. The boys' overweight+ obesity possible rates at different ages are higher than the girls', among which the difference at 11-year-old reaches 15.77 percent. By comparing the city boys and city girls in the same ages, city boys overweight rate, overweight + obesity possible rate (except 7 year-old group) at the same ages are significantly different, and the 10-year-old group boys obesity rate is very different from the city girls($P<0.05$ or $P<0.01$).

Through comparison between the country boys' and country girls', the country girls' overweight rate, obesity rate, overweight + obesity rate is very low, in detail, the 8 to 10 year-old group country girls' possible rate is 0, 11 to 12 year old country girls' possible rate is higher than the country boys'. By comparing the country boys' and country girls' at the same age, 8 to 11 year-old group country boys' overweight rate, 7-year old and 9-year old groups obesity rate, 7 to 11 year old group overweight + obesity possible rate is significantly different from the country girls'.

3) The city children's and the country children's overweight, obesity possible rate

As shown in table 1, the city boys' overweight rate, obesity rate, overweight + obesity rate at different ages are higher than country boys', among which the differences in 10-year old group overweight rate, overweight + obesity rate and 7-year old group obesity rate are largest, they respectively are 11.68 percent, 18.16 percent, 8.26 percent. By comparing the boys at the same age group, 8 to 10-year old group city boys' overweight rate, 7 to 8-year old group and 10-year old group overweight + obesity rate are obviously different from the country boys ($P<0.05$ or $P<0.01$).

By comparing the city girls and country girls, the 11-year old group country girls' obesity rate and 12-year old group overweight rate, obesity rate, overweight + obesity rate are higher than the city girls'; the overweight rate, obesity rate,

ISBN: 978-1-61804-240-8

overweight + obesity rate at other ages group are higher in the city girls. By comparing the city girls and country girls at the same age group, the 7-year old group city girls' overweight rate, overweight + obesity rate , 7 to 9-year old group obesity rate is greatly different from the country girls'(P<0.05 or P<0.01).

TABLE I. 7 TO 12 YEAR OLD TOWN AND COUNTRYSIDE CHILDREN'S OVERWEIGHT AND OBESITY POSITIVE RATE IN LANZHOU (2013)

Area	Age	Boy				Girl			
		Testing Person	Overweight	obesity	Overweight and obesity	Testing Person	Overweight	obesity	Overweight and obesity
City	7	216	20(9.26)	26(12.04)△△	46(21.30)△△	189	16(8.47)△	15(7.94)△△	31(16.40)△△
	8	237	29(12.24)*△△	22(9.28)△	51(21.52)**△△	169	10(5.92)△	8(4.73)△	18(10.65)△△
	9	236	31(13.14)**△△	19(8.05)	50(21.19)*△△	197	11(5.58)△	15(7.61)△△	26(13.20)△△
	10	226	44(19.47)*△△	25(11.06)*△	69(30.53)**△△	161	18(11.18)△△	7(4.35)	25(15.53)△△
	11	188	34(18.09)**	16(8.51)	50(26.60)**△	120	6(5.00)	7(5.83)	13(10.83)
	12	176	28(15.91)**	11(6.25)	39(22.16)**	155	9(5.81)	7(4.52)	16(10.32)
	Total	1279	186(14.54)**△△	119(9.30)**△△	305(23.85)**△△	991	70(7.06)△△	59(5.95)△△	129(13.02)△△
Countryside	7	238	15(6.30)	9(3.78)*	24(10.08)**	190	5(2.63)	0(0)	5(2.63)
	8	211	9(4.27)*	8(3.79)	17(8.06)**	137	0(0)	0(0)	0(0)
	9	189	8(4.23)*	8(4.23)*	16(8.47)**	139	0(0)	0(0)	0(0)
	10	186	15(8.06) **	8(4.30)	23(12.37)**	112	0(0)	0(0)	0(0)
	11	145	17(11.72)**	6(4.14)	23(15.86)*	102	0(0)	7(6.86)	7(6.86)
	12	147	18(12.24)	6(4.08)	24(16.33)	132	11(8.33)	8(6.06)	19(14.39)
	Total	1116	82(7.35)**	45(4.03)**	127(11.38)**	812	16(1.97)	15(1.85)	31(3.82)

Note: the number in ( ) is the possible rate/%. Comparison between children in different genders: Comparison between children in town and in country: Δ means P,0.05; ΔΔ means P<0.01.

4) Children's overweight, obesity possible rate at different age groups

Boys' overweight rate grows as the age increases. City boys' obesity rate falls as the age increases. The country boys' obesity rate stays at about 4 percent. City boys' overweight + obesity possible rate first grows and then falls. The country boys overweight + obesity possible rate first falls and then grows. Through inspection, the city (F= 12.82, P=0.025), the country (F=15.42, P=0.009) boys' overweight rate at different ages are greatly different.

The city girls' overweight, obesity, overweight + obesity possible rate all falls as the age grows. The country girls' overweight, obesity, overweight + obesity possible rate grows as the age grows. Through inspection, the country girls' overweight rate (F=37.95, P=0.0000), obesity rate (F=37.95, P=0.0000), overweight + obesity rate (F=58.91, P= 0.0000) at different age groups are greatly different.

B. Comparison of Obesity-or-overweigh children's and normal children's blood pressure, vital capacity index

The boys' and girls' diastolic blood pressure, systolic blood pressure and pulse: the obesity group > the overweight group> the normal weight group. Vital capacity index: the normal group > overweight group > obesity group. Through comparison, the boys' and the girls' overweight group, obesity group and the normal weight group's diastolic blood pressure, systolic blood pressure, pulse (except girls), vital capacity index is statistically different (P<0.05 or P< 0.01).

TABLE II. TABLE 2 COMPARISON OF LANZHOU OVERWEIGHT GROUP, OBESITY GROUP AND NORMAL WEIGHT GROUP CHILDREN'S BLOOD PRESSURE, PULSE, VITAL CAPACITY INDEX (±S)

Type	Boy					Girl				
	Testing Person	diastolic blood pressure (kPa)	systolic blood pressure (kPa)	Pulse (time/min)	Vital capacity index (mL)/(kg)	Testing Person	diastolic blood pressure (kPa)	systolic blood pressure (kPa)	Pulse (time/min)	Vital capacity index (mL)/(kg)
comparison	1703	69.32±7.61**	107.92±9.68**	86.18±7.37	48.95±11.68**	1408	69.34±7.14	106.66±8.96**	86.59±7.19	43.45±11.32**
Overweight	268	72.17±9.04**	112.54±10.13**	87.63±7.62*	40.29±8.47**	86	72.38±7.79**	111.63±9.01**	87.23±7.81	35.69±7.93**
obesity	164	75.36±8.46**	117.44±10.09**	88.99±8.04**	34.48±7.98**	74	74.63±8.25**	115.27±9.53*	88.47±10.52	29.12±4.93**

Note: * means P<0.05, ** means P<0.01.

C. 3.3. Correlation analysis between BMI and blood pressure, pulse, vital capacity index

TABLE III. CORRELATION ANALYSIS BETWEEN CHILDREN'S BMI AND BLOOD PRESSURE, PULSE, VITAL CAPACITY INDEX IN LANZHOU

Gender	Index	R value	P value
Boy	**diastolic blood pressure**	0.194	0.000
	systolic blood pressure	0.190	0.000
	Pulse	-0.020	0.475
	Vital capacity index	-0.365	0.000
Girl	**diastolic blood pressure**	0.046	0.161
	systolic blood pressure	0.055	0.097
	Pulse	-0.002	0.939
	Vital capacity index	-0.270	0.000

It applies Pearson to analyze the correlation of BMI and diastolic blood pressure, systolic blood pressure, pulse and vital capacity. BMI is positively related to the diastolic blood pressure and systolic blood pressure (except girls); BMI is negatively related to vital capacity index.

IV. 4. DISCUSSION

A. Current situation of Children obesity or overweight in Lanzhou 2013 and the causes

The overweight rate announces the early popular trend, the obesity or overweight rate reflects the popular situation, the obesity rate reflects the popular level [6].The thesis shows that 7 to 12 –year old city boys' , city girls', country boys' an country girls' overweight rate is respectively 14.54 percent, 7.06 percent, 7.35 percent and 1.97 percent. The obesity rate is respectively 9.30 percent, 5.95 percent, 4.03 percent and 1.85 percent. The obesity or overweight possible rate is respectively 23.85 percent, 13.02 percent, 11.38 percent, 3.82 percent. The boys' is higher than the girls' and the city's is higher than the country's, which is in accordance with the reported results [7]. The city boys are the most easy to join the group of obesity-or-overweight, so measures should be taken. The overweight and obesity possible rate differs in genders and city-country at all the age groups, among which the city and country boys' overweight rate at 10 to 12-year old group is higher than those at other age groups; country girls' overweight rate, obesity rate at 11 to 12 –year old grows sharply, and they are the most dangerous group to get obesity or overweight. The latest American report shows that the childhood obesity rate in 2009 to 2010 is 16.9 percent [8]. In cities of Guangdong province, 7 to 14-year old students' overweight rate is 11.1 percent, the obesity rate is 7.2 percent [9]. In Zhangjiakou city, 7 to 12-year old children's overweight rate is 13.01 percent, obesity rate is 14.25 percent [10]. By comparison, Lanzhou- children overweight rate, obesity rate in 2013 are lower than that in developed countries and the large or medium cities in China. Compared with the Lanzhou (2013) city boys', city girls' and country boys' overweight rate, obesity rate grows, while the country girls' overweight rate and obesity rate falls.

B. The relation between childhood obesity and blood pressure, pulse, vital capacity

The human beings' cardio system and breath system is a symbol of a person's health and they also influence the human beings' lifespan, working duration and working efficiency. Blood pressure, pulse and vital capacity is a key physical index of the heart, vessels and lungs, thus they mean a lot to the body and the health.

Obesity is an important factor to cause high blood pressure. The gradually increasing blood pressure rate coexists with the sharply growing obesity or overweight. The research shows, the overweight children and the obesity children's diastolic blood pressure, systolic blood pressure are higher than normal weight children. BMI is positively related to the diastolic blood pressure, systolic blood pressure, which is in accordance with the research results in recent years[11-13]: prove that overweight and obesity is an important factor to cause the children blood pressure to be higher. Obesity children's growing blood pressure may be a compensatory mechanism. As the adipose tissue increases, the vascular bed increases, then the capacity increases, thus the output per pulse and the output from the heart increases. As the heart contains a larger capacity in a long time, the chambers of the heart grow, the myocardium grows thicker and the weight of myocardium increases. Meanwhile, as the adipose tissue in and out of the myocardium is easy to cause damage to the myocardium, the diastolic function of left chamber falls[14]. It's reported that the blood pressure has tracks. The original high blood pressure at adulthood can date back to the childhood. As the blood pressure in childhood grows, the blood pressure in adulthood and the risk of high blood pressure at adulthood rise greatly [15]. Therefore, to conduct children blood pressure and prevent the children from high blood pressure is very important.

ISBN: 978-1-61804-240-8

Vital capacity is one of the most important indexes to reflect the human beings' lung function. The vital capacity is not only related to the anatomy capacity of the lungs, but also related to the ventilator function and the contractility of the lungs. Obesity causes the ventilator function to fall, thus the blood circulation property becomes abnormal and the hypoxemia forms, which finally damages the ventilator function. The research founds that overweight, obesity children's vital capacity index is lower than the normal weight children's. BMI is negatively related to the vital capacity index. Note: the obesity children's ventilator functions grows far behind the normal children's, because the extra fat accumulating in the chest and the belly can take mechanism action to the chest and the belly, action of the chest and the diaphragm is limited, then the ventilator block grows causing the ventilator function to fall, finally the vital capacity and the ventilator adjustment changes [16]. The research has not found the relation between pulse and BMI, but the close relation between pulse and individual action may influence the relation between pulse and BMI.

Childhood obesity mainly is a kind of pure obesity. The parents and the school should strengths the health education onto their children or students; help them to believe in the conception that health is the most important. They should take some necessary long-term interfere actions to control the dietary habit of "high energy, high fat", increase their exercise strength, change their bad lifestyle to prevent or control obesity and other complications.

ACKNOWLEDGMENT

Chengguan of Lanzhou of Science and Technology Project: Application of bioelectrical impedance measuring body composition feasibility study in the diagnosis of childhood obesity (2012-1-11).

REFERENCE

[1] DANIELS SR.Complications ofobesity in children and adolescents.Int JObes(Lond), 2009,33(Suppl1):S60-S65.

[2] Christodoulos AD, Douda HT, Tokmakidis SP, Cardiorespiratory fitness, metabolic risk and inflammation in children.INT J Pediatr,2012,2012:270515

[3] Chen X, Wang Y, T racking of blood pressure from childhood to adulthood: a systematic review and met aregression analysis[J].Circulation,2008,117(25):3171-3180

[4] Weiss R, Dziura J, Burgert TS, et al. Obesity and the Metabolic Syndrome in Children and Adolescents.N Engl J Med, 2004, 350:2362-2374.

[5] DANIELSSR. The consequences of childhood overweight and obesity. Future Child, 2006, 16(1):47-67.

[6] Working Group on Obesity in China (WGOC). Guidelines for prevention and control of overweight and obesity in Chinese adults. Biom Environ Sci, 2004, 17(Suppl):1-36.

[7] OGDEN CL,CARROLL MD,CURTIN LR,et al. Prevalence of overweight and obesity in the United States,1999-2004[J].JAMA,2006,295(13):1549-1555

[8] Ogden CL, Carroll MD, Kit BK, et al. Prevalence of obesity and trends in body mass index Among US children and adolescents, 1999-2010[J].JAMA,2012,307(5):483-490.

[9] Liu W, Lin R, Lin U, et al. Prevalence and association between obesity and metabolic syndrome Among Chinese elementary school children: a school-based survey. BMC Public Health, 2010,10:780-785

[10] Christodoulos, A. D., Douda, H. T., & Tokmakidis, S. P. (2012). Cardiorespiratory fitness, metabolic risk, and inflammation in children.International journal of pediatrics, 2012.

[11] Gundogdu Z. Relationship between BMI and blood pressure in girls and boys.Public Health Nutr, 2008,11(10):1085-1088.

[12] Genovesi S, Antolini L, Giussani M, et al. Usefulness of waist circumference for the identification of childhood hypertension. J Hypertens, 2008,26(8):1563-1570.

[13] Ogden CL, Flegal KM, Carroll M, et al. Prevalence and trends in overweight among US children and adolescents 1999-2000[J].JAMA,2002,288(14):1728-1732.

[14] Boyd, G. S., Koenigsberg, J., Falkner, B., Gidding, S., & Hassink, S. (2005). Effect of obesity and high blood pressure on plasma lipid levels in children and adolescents. Pediatrics, 116(2), 442-446.

[15] Juhola J, Magnussen CG, Viikari JS, et al, Tracking of serum lipid levels, blood pressure, and body mass index from childhood to adulthood : the cardiovascular risk in young finns Study. J Pediatr,2011,159(4):584-590

[16] Lazarus, R., Sparrow, D., & Weiss, S. T. (1997). Effects of obesity and fat distribution on ventilatory function the normative aging study. CHEST Journal,111(4), 891-898.

New Tuning Method of the Wavelet Function for Inertial Sensor Signals Denoising

Ioana-Raluca Edu[1,*],
Felix-Constantin Adochiei[2],
Radu Obreja[2],
Constantin Rotaru[1],
Teodor Lucian Grigorie[2]

[1]*Military Technical Academy, Faculty of Mechatronics and Armament Integrated Systems, Bucharest, Romania*
[2]*University of Craiova, Faculty of Electrical Engineering, Department of Electric, Energetic and Aerospace Engineering*
*Corresponding author (E-mail: edu_ioana_raluca@yahoo.com)

Abstract — The current research is aimed at implementing and validating software procedures by proposing new algorithms that receive data from inertial navigation systems/sensors (data acquisition equipment) and provide accurate navigation information. In the first phase, our concern was to implement a real-time evaluation criterion with the intention of achieving real-time data from an accelerometer. It is well known that most errors in the detection of position, velocity and attitude in inertial navigation occur due to difficult numerical integration of noise. The main goal of this study was to propose a signal processing algorithm, based on the Wavelet filter, used along with a criterion for evaluating and updating the Wavelet's optimal level of decomposition, for reducing the noise component. We performed numerical simulations using signals received from an accelerometer and analyzed the numerical results to see whether the improved Wavelet (proposed method) can be used to achieve more precise information on a vehicle.

Keywords — signal processing; wavelet transform, partial directed coherence method.

I. Introduction

The aim of the our scientific research is to develop advanced algorithms able to determine the optimal level of decomposition for the Wavelet method and to implement these algorithms in a miniaturized inertial measurement units in order to obtain accurate data regarding the vehicle displacement.

Apart from the indisputable benefits of size reduction, reliability, manufacturing costs and power consumption, the miniaturization of sensors and by default of inertial measurement systems (*INS*) caused a number of problems related to their performance degradation. As a result of miniaturization, stochastic (noise of the system) and deterministic errors occurred [1-4].

The inertial sensors noise, major source of errors for inertial navigation systems, is characterized by a constant power throughout the frequency spectrum, that reflects the dynamics of mobile systems - which are intended to be monitored (generally in the range 0-100 Hz). Therefore, this type of noise filtering in the 0-100 Hz band is not recommended.

This noise component that overlaps over the output of the sensors, cannot be totally eliminated but it can be influenced by stochastic processes [5].

The development and also the optimization of advanced algorithms for improving the performances of miniaturized inertial sensor and inertial measurement units are extremely important topics in the field of aerospace navigation systems.

The practical challenge of the study under discussion was to develop and validate a complex algorithm that would process the signals received from accelerometers, and later from *INS*, remove the noise and offer precise information regarding the vehicle displacement.

Therefore, an improved wavelet filter was proposed, used to remove the noise detected during measurements, in order to obtain a better accuracy of the measurements. The optimal order of the wavelet filter (the optimal decomposition level) was calculated using a correlation analysis function applied to the signals achieved from the accelerometers and the real speed signals applied to the accelerometers (considered as reference signals).

II. Proposed Method

The Wavelet transform is a very powerful tool for the signal feature extraction and noise reduction and offers effective localization in time and frequency domains.

In order to analyze signals, the continuous Wavelet transform (*CWT*) can be considered as a tree decomposition of the signal (the Wavelet decomposition tree), a combination of a set of basic functions, obtained by means of dilation and translation of a single prototype Wavelet function $\psi(t)$ called the mother Wavelet, as illustrated in Fig. 1 [6].

In practice, the dual tree consists of two discrete Wavelet transforms working in parallel. The branches of the tree are interpreted as the real part, respectively, the imaginary one of a complex wavelet. Thus the complex wavelet transform is obtained.

The continuous Wavelet transform $(W_\psi f)(s,\tau)$, of the signal $f(t) \in L^2(R)$ can be determined as:

$$(W_\psi f)(s,\tau) = \int_{-\infty}^{+\infty} f(t) \cdot \frac{1}{\sqrt{s}} \cdot \overline{\psi\left(\frac{t-\tau}{s}\right)} \cdot dt \quad (1)$$

where, $\overline{\psi}$ denote the complex conjugate of ψ .

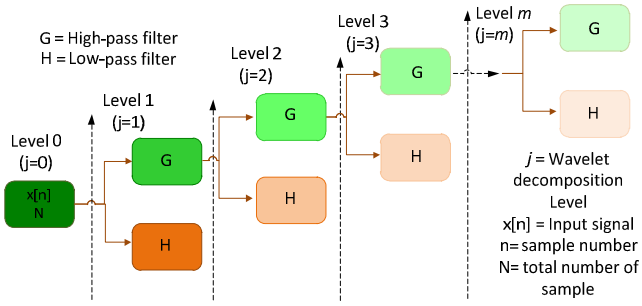


Fig.1- Wavelet decomposition tree

A wavelet filter acts as an averaging filter or a filter that detects details when a signal is decomposed with wavelets. A part of the consequent wavelet coefficients match with details in the data set. The detail significance is proportional with the amplitude of the waves - if they are small, they can be left out, without essentially influencing the data set main features [7]. The main idea of thresholding is to set all coefficients that are below a specific threshold at zero value. In order to rebuild the data set these coefficients are utilized in an inverse wavelet transform [8]. Deficiencies and design issues of such decomposition lead to the development of new processing methods.

We decided to propose a new method for estimating the optimal level of decomposing for the wavelet filter.

We are introducing a new time frequency approach, an extension of the Partial Directed Coherence (*PDC)* method, to assess coupling dynamics information in multivariate dynamic systems [9].

PDC approach is able to detect direct and indirect couplings between two time series. *PDC* is based on an *m*-dimensional multichannel autoregressive model (*MAR)* and uses an *MAR* process with order *p*:

$$\begin{bmatrix} x_1(n) \\ \vdots \\ x_N(n) \end{bmatrix} = \sum_{r=1}^{p} A_r(n) \begin{bmatrix} x_1(n-r) \\ \vdots \\ x_N(n-r) \end{bmatrix} + \begin{bmatrix} W_1(n) \\ \vdots \\ W_N(n) \end{bmatrix} \quad (2)$$

Where the *w* vector is the white noise and A_r matrices are expressed by means of the next formulation:

$$A_r(n) = \begin{bmatrix} a_{11}(r,n) & \cdots & a_{1N}(r,n) \\ \vdots & a_{ij}(r,n) & \vdots \\ a_{N1}(r,n) & \cdots & a_{NN}(r,n) \end{bmatrix} \quad (3)$$

with *r=1,.., p* model order. a_{ij} parameters represent the linear interaction effect of $x_j(n-r)$ onto $x_i(n)$. They are estimated by means of an adaptive autoregressive approach [10], the main advantage of which is the possibility of analyzing time-varying signals by updating the calculations for each time sample under investigation.

By calculating the Fourier transform of $A_r(n)$ matrix, more specific by calculating *A(n, t)* coefficient matrix in frequency domain:

$$A(n,f) = I - \sum_{r=1}^{p} A_r(n) z^{-r}\big|_{z=e^{i2\pi f}} \quad (4)$$

where I is the identity matrix, a number of time-varying measurements of connectivity can be established.

The *PDC* coupling estimation between two time series (*Xi* and *Xj*) was defined by Baccala et al. [11] as:

$$\pi_{ij}(n) \overset{\Delta}{=} \frac{A_{ij}(n,f)}{\sqrt{a_j^H(n,f) a_j(n,f)}} \quad (5)$$

where $\pi_{ij}(n)$ is the correlation parameter, $(.)^H$ the Hermitian transpose, $A_{ij}(n,f)$ the $A_r(n)$,the Fourier transform of the matrix in the frequency domain, $a_j(n,f)$ the *j'th* column number of the matrix *A(n,f)*, *n* the number of windows and *f*, the frequency.

The π_{ij} parameter normalization conditions in the frequency domain ($\pi_{ij}(f)$) were defined as:

$$0 \le \left|\pi_{ij}(f)\right| \le 1, \quad \sum_{i=1}^{m} \left|\pi_{ij}(f)\right| = 1 \quad (6)$$

for all $1 \le j \le m$ values.

These measures were considered to provide information on the presence and level of causal correlation between two time series (*Xi* and *Xj*) as follows:

a) high values reflecting a directionally linear influence from *Xj* to *Xi*, meaning that, for values equal to 1, all the causal influences originating from *Xj* are directed towards *Xi,*

b) low values (≈0) suggesting the absence of any causal correlation from *Xj* to *Xi*, meaning that *Xj* does not influence *Xi*.

In order to estimate the coupling level (*CL*) between two time series belonging to the same system and to estimate the optimal level of the wavelet filter, the calculation of a new parameter was proposed, by employing the following equations:

$$a = mean \;\; PDC(X_i \to X_j)$$
$$b = mean \;\; PDC(X_{i-1} \to X_{j-1})$$
$$CL = \begin{cases} WoptLvl = WactualLvl + 1, \; if \; a - b > 0 \\ WoptLvl = WactualLvl, \; if \; a - b = 0 \\ WoptLvl = WactualLvl - 1, \; if \; a - b < 0 \end{cases} \quad (7)$$

where, $WactualLvl$ is the Wavelet's actual level and $WoptLvl$ is the Wavelet's optimal level of decomposition. These measures provide information on wavelet coefficient as follows: a) if the previous value is lower than the current value then the order of the wavelet decomposition is equal to the previous value plus 1. b) if the previous value is higher than or equal to the current value, then the optimal order of decomposition is equal to the previous value.

The main idea of the optimization algorithm is illustrated in Fig. 2, where a signal received from an accelerometer is processed and analyzed by using the Wavelet transform until an optimal level of decomposition is established and the useful signal is achieved [10]. This elementary structure was proposed and studied in order to obtain a further general tuning method for the inertial sensor denoising, with the wavelet

method. In this new general structure, the reference signals will be provided by a GPS, while the disrupted input signals in PDC are the outputs of the inertial navigation system (INS) (Fig. 3).

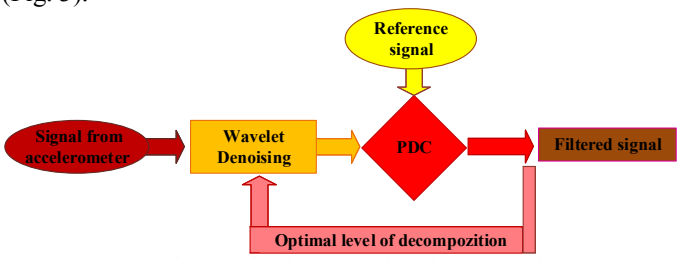


Fig.2 - The logical scheme/mathematical algorithm proposed for improving the signals acquired from an accelerometer

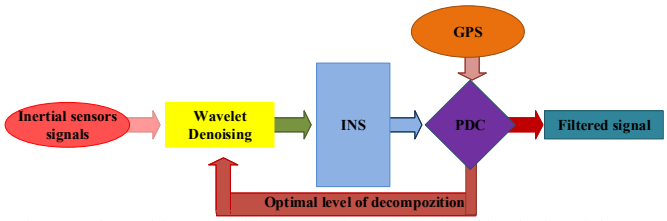


Fig. 3 – The architecture of the general tuning method for the inertial sensor denoising with wavelet method

For the here-presented study we simulated noisy and noiseless (clean) signals received from a miniaturized accelerometer in order to acknowledge an offline tuning of the wavelet function used in the denoising process of the accelerometer. The noiseless signals were used as reference signals (equivalent to the signals received from a GPS device) to correct the errors of the accelerometer. The noisy signals were correlated with the noiseless signals, by applying equation 7, in order to achieve the optimal level of decomposition of the wavelet filter leading to - after the accelerometer calibration - the achievement of more accurate acceleration data.

In simulations, a sinusoidal signal, generated by using the *wnoise Matlab* function for "Noisy wavelet test data", Fig. 4, was considered as a reference signal. This signal was corrupted by different types of noise (additive Gaussian white noise) as it may be seen in Fig. 5.

The method was implemented in Matlab for testing and validation after the mathematical problem was established.

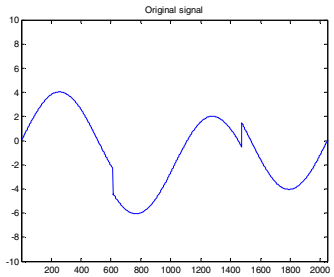


Fig 4 - Original signal

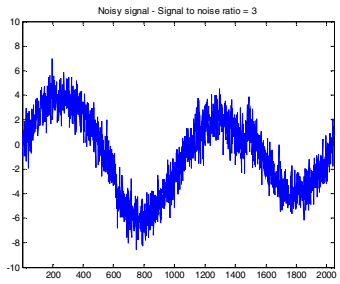


Fig. 5 - Signal corrupted with additive Gaussian white noise

III. Results and Discussions

By applying the proposed algorithm to the corrupted signal, using as reference the original signal (equivalent to the *GPS* signal), the $WoptLvl$ and CL achieved values were recorded in Table 1. According to equation 7, we can observe that the optimal level of decomposition is 10 for $CL = 0.851520$.

Table 1.

$WactualLvl$	CL
2	0.142700
3	0.289110
4	0.552780
5	0.795460
6	0.811420
7	0.811610
8	0.812420
9	0.835450
10	0.851520
11	0.811660
12	0.793500
13	0.778150
14	0.770220

For a visual comparison of the coupling level between the corrupted and the original signals, the coupling level diagrams for five different coupling levels were plotted by means of a short time implementation, with a window of 300 sample length, Figures 6 and 7. In both figures, *y*-axis represents the number of windows and the *x*-axis represents the normalized frequency between 0 and 1.

As it can be seen in all coupling diagrams, transitions from coupling to uncoupling, from strong level of couplings (represented in red color and shades of red) towards the absence of coupling (represented in blue color and shades of blue) are visible; an absence of coupling – a predominantly blue color is visible in Figure 6 suggesting that for $CL = 0.289110$, the investigated signals are uncorrelated or the level of correlation is very low and the achieved data corresponds in a limited propotion with the real data (the original signal). We are interested in achieving the highest level of coupling between the signals, CL values ≈ 1. A higher level of coupling can be seen in figure 8, for $CL = 0.851520$, where the predominantly red color suggests the presence of a strong level of coupling between the two signals. This level of coupling indicates that, for $CL = 0.851520$ $WactualLvl = 10$ is the $WoptLvl$.

ISBN: 978-1-61804-240-8

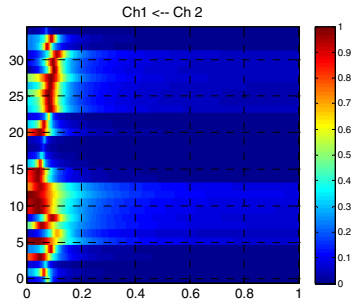


Fig 6 - Signal achieved for $WacualLvl$=3

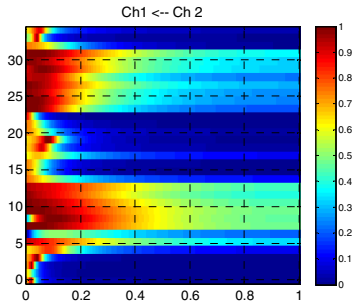


Fig. 7 – The coupling level diagram for $CL = 0.811420$, $WactualLvl = 6$

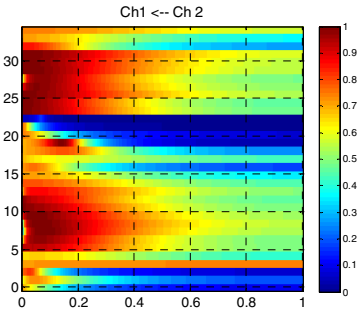


Fig. 8 – The coupling level diagram for $CL = 0.851520$, for which $WactualLvl = WoptLvl = 10$

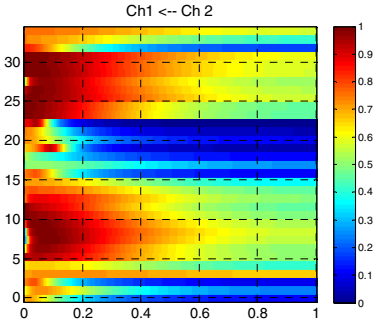


Fig. 9 - Signal achieved for $WactualLvl=14$

The sinusoidal signals achieved for different $WactualLvl$ were displayed in figures 10, 11, 12, 13 and 14.

After a careful visual inspection of figures 10 - 14 one can see that as the CL level increases, the signal achieved is more similar to the original signal but when/after CL = 14, the signal shape changes and turns into a sinusoidal signal, which loses the characteristics of the original signal. Also from the visual inspection we concluded that sinusoidal signal achieved for $WactualLvl$=10, which may be seen in Fig. 11 is/ was the optimum level of decomposition - $WoptLvl$.

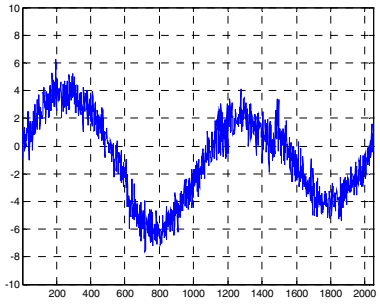


Fig. 10 - Signal achieved for $WactualLvl$=1

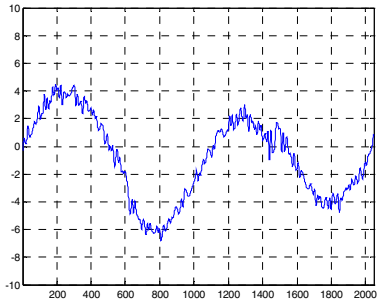


Fig. 11 - Signal achieved for $WactualLvl$=3

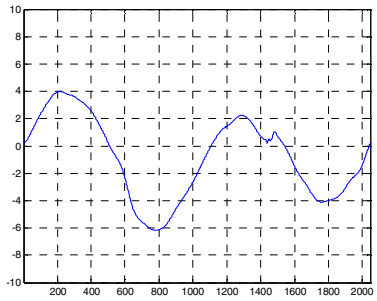


Fig. 12 - Signal achieved for $WactualLvl = 6$

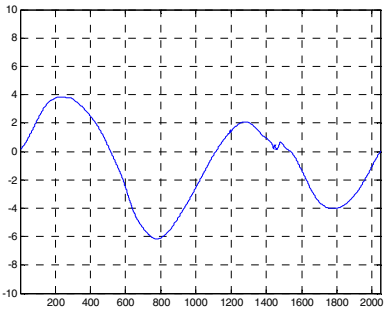


Fig. 13 - Signal achieved for $WactualLvl = WoptLvl$=10

ISBN: 978-1-61804-240-8

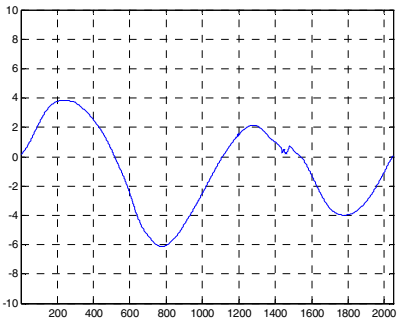


Fig. 14 - Signal achieved for $WactualLvl$=14

By using the proposed configuration from Figure 3, the correlation between useful signals, signals received from an accelerometer or an inertial measurement unit/*INS* can be tracked and corrected by using signals received from a GPS device (in a pre-calibration phase of the INS). This correlation becomes less clear when the signals achieved from the INS become correlated with the signals received from GPS, the correlation level reaches values close to 1, resulting in reduced errors of the navigation system (caused by the noise).

IV. CONCLUSIONS

This is a topical issue which brings significant improvement in the inertial navigation systems signal processing having a clear-cut role in positioning investigations.

The purpose of this research was to improve the performance of inertial navigation systems and their level of accuracy for situations when the GPS becomes unavailable. An improved version of the Wavelet filter was proposed for filtering/denoising the signals received from an accelerometer.

We intend to implement this algorithm for pre-calibrating a two-dimensional navigation system in the horizontal plan in order to improve its accuracy in positioning. By establishing the best coupling level of signals received from INS and GPS, using the GPS signal as reference, the optimal level of decomposition of the wavelet transform can be established and the proposed algorithm can be implemented in the inertial measurement unit as a real-time evaluation criterion with the purpose of achieving real-time data.

REFERENCES

[1] A. Lawrence, "Modern inertial technology: navigation, guidance and control.," *Springer Verlag, New York,* 1993.

[2] R. Ramalingam, G. Anitha, and J. Shanmugam, "Microelectromechnical Systems Inertial Measurement Unit Error Modelling and Error Analysis for Low-cost Strapdown Inertial Navigation System," *Defence Science Journal, Vol. 59, No. 6, Nov. , pp. 650-658,* 2009.

[3] T. D. Tan, L. M. Ha, N. T. Long, N. P. Thuy, and H. H. Tue, "Performance Improvement of MEMS-Based Sensor Applying in Inertial Navigation Systems," *Research - Development and Application on Electronics, Telecommunications and Information Technology, No. 2 , Posts, Telematics & Information Technology Journal, pp. 19-24,* 2007.

[4] D. G. Eqziabher, "Design and Performance Analysis of a Low-Cost Aided Dead Reckoning Navigator," *A Dissertation submitted to the Department of Aeronautics and Astronautics and the committee on graduate studies of Stanford University in partial fulfillment of the requirements for the degree of doctor of philosophy, Stanford University, February,* 2004.

[5] H. Haiying, " Modeling inertial sensors errors using Allan variance," *UCEGE reports number 20201, Master's thesis, University of Calgary, September,* 2004.

[6] T. Chan and C. J. Kuo, "Texture Analysis and Classification with Tree-Structured Wavelet Transform," *IEEE Transactions on Image Processing, Vol. 2, No. 4, October,* 1993.

[7] H. Khorrami and M. Moavenian, "A comparative study of DWT, CWT and DCT transforms in ECG arrhythmias classification " *Expert Systems with Applications,* 2010.

[8] M. Alfaouri and K. Daqrouq, "ECG Signal Denoising By Wavelet Transform Thresholding," *in American Journal of Applied Sciences 5 (3): pp: 276 – 281,* 2008.

[9] F.-C. Adochiei, S. Schulz, I.-R. Edu, H. Costin, and A. Voss, "A new normalised short time pdc for dynamic coupling analyses in hypertensive pregnant women.," *BMT 2013, GRAZ, 19-21 September,* 2013.

[10] T. Milde, L. Leistritz, L. Astolfi, W. H. R. Miltner, T. Weiss, F. Babiloni, and H. Witte, "A new Kalman filter approach for the estimation of high-dimensional time-variant multivariate AR models and its application in analysis of laser-evoked brain potentials " *Neuroimage 50 960-9,* 2010.

[11] L. A. Baccala and K. Sameshima, "Partial directed coherence: a new concept in neural structure determination.," *Biol. Cybern., 84, 463-474.*

An exploratory crossover operator for improving the performance of MOEAs

K. Metaxiotis, K. Liagkouras

***Abstract*—**In evolutionary algorithms the recombination operator considered to be one of the key operators that allows the population to progress towards higher fitness solutions. In this paper we re-examine the Simulated Binary Crossover (SBX) operator and propose an exploratory version of the SBX operator that allows the Multiobjective Evolutionary Algorithms (MOEAs) to search large landscapes efficiently. The proposed Exploratory Crossover Operator (ECO) is applied to the Non-dominated Sorting Genetic Algorithm II (NSGAII), under the Zitzler-Deb-Thiele's (ZDT) set of test functions. The relevant results are compared with the results derived by the same MOEA by using its typical configuration with the SBX operator. The experimental results show that the proposed Exploratory Crossover Operator outperforms the classical Simulated Binary Crossover operator, based on two performance metrics that evaluate the proximity of the solutions to the Pareto front.

***Keywords*—**Multiobjective optimization; evolutionary algorithms; crossover.

I. Introduction

RECOMBINATION or crossover operators are being used by the Evolutionary Algorithms along with the selection and mutation operators in order to evolve a set of solutions towards higher fitness regions of the search space. Each one of the aforementioned operators is designed in order to facilitate different needs, for instance the crossover operator is responsible for the search process while the mutation acts as a diversity preserving mechanism. Further, the various recombination techniques can be divided into distinct categories based on the selected representation. For instance, there is a substantial number of recombination operators that use binary encoding for the representation of chromosome, to name just a few recombination operators that belong to this category, one-point crossover, n-point crossover, uniform crossover and arithmetic crossover. Binary representation, however, can be problematic in tasks that require a high numerical precision [1]. Real coding is more suitable representation for continuous domain problems, where each gene represents a variable of the problem. One of the most popular real coded crossover operators that has been applied to a considerable number of multiobjective evolutionary algorithms (MOEAs), is the Simulated Binary Crossover (SBX) [2]. In this work we propose a new strategy to improve the performance of SBX operator by introducing an exploratory version of the SBX operator that allows the more efficient exploration of the search space.

The rest of the paper is structure as follows. In section II, a description of the Simulated Binary Crossover (SBX) is given and in section III the proposed Exploratory Crossover Operator (ECO) is presented. The experimental environment is presented in section IV. Section V presents the performance metrics. In section VI we test the performance of the proposed ECO by using the Zitzler-Deb-Thiele's (ZDT) set of test functions. Finally, section VII analyzes the results and concludes the paper.

II. Simulated Binary Crossover (SBX)

The simulated binary crossover (SBX) operator was introduced by Deb and Agrawal [2] in 1995. It uses a probability distribution around two parents to create two children solutions. Unlike other real-parameter crossover operators, SBX uses a probability distribution which is similar in principle to the probability of creating children solutions in crossover operators used in binary-coded algorithms. In SBX as introduced by [2] each decision variable x_i, can take values in the interval: $x_i^{(L)} \leq x_i \leq x_i^{(U)}, i = 1, 2, \dots, n$. Where $x_i^{(L)}$ and $x_i^{(U)}$ stand respectively for the lower and upper bounds for the decision variable *i*. In SBX, two parent solutions $y^{(1)}$ and $y^{(2)}$ generate two children solutions $c^{(1)}$ and $c^{(2)}$ as follows:

1. Calculate the spread factor β:

$$\beta = 1 + \frac{2}{y^{(2)} - y^{(1)}} \, min\left[(y^{(1)} - y^{(l)}), (y^{(u)} - y^{(2)})\right]$$

2. Calculate parameter α :

$$\alpha = 2 - \beta^{-(\eta_c + 1)}$$

3. Create a random number u between 0 and 1.

$$u \longrightarrow [0, 1];$$

K. Metaxiotis is an Associate Professor at the Decision Support Systems Laboratory, Department of Informatics, University of Piraeus, 80, Karaoli & Dimitriou Str., 18534 Piraeus, Greece (phone: +30210 4142378; fax: +30210 4142264; e-mail: kmetax@unipi.gr).

K. Liagkouras is a PhD Candidate at the Decision Support Systems Laboratory, Department of Informatics, University of Piraeus, 80, Karaoli & Dimitriou Str., 18534 Piraeus, Greece (e-mail: kliagk@unipi.gr).

4. Find a parameter β_q with the assistance of the following polynomial probability distribution:

$$\beta_q = \begin{cases} (au)^{1/(\eta_c+1)} & if\ u \le \frac{1}{a}, \\ \left(\frac{1}{2-au}\right)^{1/(\eta_c+1)} & otherwise \end{cases}$$

The above procedure allows a zero probability of creating any children solutions outside the prescribed range $[x^{(L)}, x^{(U)}]$. Where η_c is the distribution index for SBX and can take any nonnegative value. In particular, small values of η_c allow children solutions to be created far away from parents and large values of η_c allow children solutions to be created near the parent solutions.

5. The children solutions are then calculated as follows:

$$c^{(1)} = 0.5\left[\left(y^{(1)} + y^{(2)}\right) - \beta_q |y^{(2)} - y^{(1)}|\right]$$

$$c^{(2)} = 0.5\left[\left(y^{(1)} + y^{(2)}\right) + \beta_q |y^{(2)} - y^{(1)}|\right]$$

The probability distributions as shown in *step 4*, do not create any solution outside the given bounds $[x^{(L)}, x^{(U)}]$ instead they scale up the probability for solutions inside the bounds, as shown by the solid line in Fig. 1.

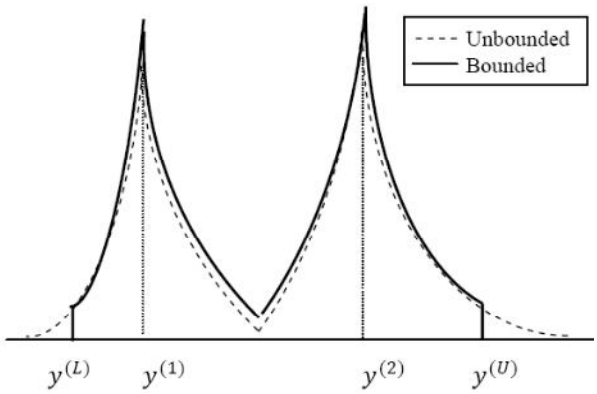


Fig. 1 Probabilities distributions for bounded and unbounded cases

III. Exploratory Crossover Operator (ECO)

The Evolutionary Algorithms (EAs) in order to be able to maintain a satisfactory level of progression towards higher fitness regions of the search space is necessary to exploit and explore efficiently the search space. The exploitation of the search space is done mainly through the selection operator and the exploration through the crossover operator. In this study we propose a new crossover operator named Exploratory Crossover Operator (ECO) due to its ability to explore efficiently the search space. We will start analyzing ECO mechanism by recalling the simulated binary crossover (SBX) as shown in section II, as the first two steps are common for both methods. Indeed as shown below first we calculate the spread factor β and then the parameter α in the same manner as the SBX.

However, in *step 3* we follow a different strategy. As shown in section II that illustrates the SBX operator, a random number $u \in [0,1]$ is generated. If $u \le 1/a$, it samples to the left hand side (region between $y^{(L)}$ and $y^{(i)}$, otherwise if $u > 1/a$ it samples to the right hand side (region between $y^{(i)}$ and $y^{(U)}$, where $y^{(i)}$ is the i^{th} parent solution.

In ECO at this particular point as shown below we follow a different methodology. Specifically, instead of generating a random number $u \in [0, 1]$, we generate two random numbers, $u_L \in [0, 1/a]$ to sample the left hand side and a random number $u_R \in (1/a, 1]$ to sample the right hand side of the probability distribution. From the aforementioned process emerge two values of β_q , the β_q^L that samples the left hand side of the polynomial probability distribution and the β_q^R that samples the right hand side of the polynomial probability distribution. Next, as shown below in *step 5* with the assistance of β_q^L and β_q^R are formulated two variants for each child solution. Specifically, $c_L^{(1)}$ and $c_R^{(1)}$ are the two variants that emerge by substituting the β_q^L and β_q^R to $c^{(1)}$. Respectively $c_L^{(2)}$ and $c_R^{(2)}$ are the two variants that emerge by substituting the β_q^L and β_q^R to $c^{(2)}$.

Then, by substituting to the parent solution vector at the position of the selected variable to be crossovered, respectively the $c_L^{(1)}$and $c_R^{(1)}$ we create two different child solution vectors (*csv*), the $csv_L^{(1)}$and $csv_R^{(1)}$. Thanks to the generated $csv_L^{(1)}$and $csv_R^{(1)}$ we are able to perform fitness evaluation for each one of the corresponding cases. As soon as we complete the fitness evaluation process, we select the best child solution between the two variants $c_L^{(1)}$and $c_R^{(1)}$ with the assistance of the Pareto optimality framework. The same procedure is followed for $c_L^{(2)}$and $c_R^{(2)}$. The proposed methodology allows us to explore more efficiently the search space and move progressively towards higher fitness solutions. Whenever, there is not a clear winner i.e. strong or weak dominance, between the $c_L^{(1)}$and $c_R^{(1)}$, or respectively between the $c_L^{(2)}$and $c_R^{(2)}$ the generation of a random number allows the random choice of one of the two alternative child solutions.

The procedure of computing children solutions $c^{(1)}$ and $c^{(2)}$ from two parent solutions $y^{(1)}$ and $y^{(2)}$ under the Exploratory Crossover Operator (ECO) is as follows:

1. Calculate the spread factor β:

$$\beta = 1 + \frac{2}{y^{(2)} - y^{(1)}}\, min\left[(y^{(1)} - y^{(l)}), (y^{(u)} - y^{(2)})\right]$$

2. Calculate parameter a :

$$\alpha = 2 - \beta^{-(\eta_c+1)}$$

3. Create 2 random numbers $u_L \in [0, 1/a]$ and $u_R \in (1/a, 1]$.

   $u_L \longrightarrow [0, 1/\alpha]$;

   $u_R \longrightarrow (1/\alpha, 1]$;

4. Find 2 parameters β_q^L and β_q^U with the assistance of the following polynomial probability distribution:

$$\beta_q^L = (au_L)^{1/(\eta_c+1)} \quad , \quad u_L \in [0, 1/a],$$

$$\beta_q^R = \left(\frac{1}{2-au_R}\right)^{1/(\eta_c+1)} \quad , \quad u_R \in (1/a, 1]$$

5. Thus, instead of a unique value for $c^{(1)}$ and $c^{(2)}$, we obtain two evaluations for each child solution that correspond to β_q^L and β_q^R respectivelly:

$$c_L^{(1)} = 0.5\left[\left(y^{(1)} + y^{(2)}\right) - \beta_q^L|y^{(2)} - y^{(1)}|\right]$$
$$c_R^{(1)} = 0.5\left[\left(y^{(1)} + y^{(2)}\right) - \beta_q^R|y^{(2)} - y^{(1)}|\right]$$

$$c_L^{(2)} = 0.5\left[\left(y^{(1)} + y^{(2)}\right) + \beta_q^L|y^{(2)} - y^{(1)}|\right]$$
$$c_R^{(2)} = 0.5\left[\left(y^{(1)} + y^{(2)}\right) + \beta_q^R|y^{(2)} - y^{(1)}|\right]$$

6. We perform fitness evaluation for each variant child solution, by substituting the candidate solutions into the parent solution vector.

7. We select the best variant between the $c_L^{(1)}$and $c_R^{(1)}$, based on the Pareto optimality framework. The same procedure is followed for $c_L^{(2)}$and $c_R^{(2)}$.

Whenever, there is not a clear winner i.e. strong or weak dominance, between the $c_L^{(1)}$and $c_R^{(1)}$, or respectively between the $c_L^{(2)}$and $c_R^{(2)}$ the generation of a random number allows the random choice of one of the two alternative child solutions.

IV. Experimental Environment

All algorithms have been implemented in Java and run on a personal computer Core 2 Duo at 1.83 GHz. The jMetal [3] framework has been used to compare the performance of the proposed, Exploratory Crossover Operator (ECO) against the Simulated Binary Crossover (SBX) operator with the assistance of NSGAII. In all tests we use, binary tournament and polynomial mutation (PLM) [2] as, selection and mutation operator, respectively. The crossover probability is $P_c = 0.9$ and mutation probability is $P_m = 1/n$, where n is the number of decision variables. The distribution indices for the crossover and mutation operators are $\eta_c = 20$ and $\eta_m = 20$, respectively. Population size is set to *100*, using *25,000* function evaluations with *100* independent runs.

V. Performance Metrics

A. Hypervolume

Hypervolume [4], is an indicator of both the convergence and diversity of an approximation set. Thus, given a set S containing m points in n objectives, the hypervolume of S is the size of the portion of objective space that is dominated by at least one point in S. The hypervolume of S is calculated relative to a reference point which is worse than (or equal to) every point in S in every objective. The greater the hypervolume of a solution the better considered the solution.

B. Epsilon Indicator I_ε

Zitzler et al. [5] introduced the epsilon indicator (I_ε). There are two versions of epsilon indicator the multiplicative and the additive. In this study we use the unary additive epsilon indicator. The basic usefulness of epsilon indicator of an approximation set A ($I_{\varepsilon+}$) is that it provides the minimum factor ε by which each point in the real front R can be added such that the resulting transformed approximation set is dominated by A. The additive epsilon indicator is a good measure of diversity, since it focuses on the worst case distance and reveals whether or not the approximation set has gaps in its trade-off solution set.

VI. Experimental Results

A number of computational experiments were performed to test the performance of the proposed Exploratory Crossover Operator (ECO) for the solution of the ZDT1-4, 6 set of test functions [6]. The performance of the proposed ECO operator is assessed in comparison with the Simulated Binary Crossover (SBX) operator with the assistance of a well-known MOEA, namely the Non-dominated Sorting Genetic Algorithm II (NSGAII). The evaluation of the performance is based on two performance metrics that assess both the proximity of the solutions to the Pareto front.

A. The Zitzler-Deb-Theile (ZDT) test suite

The Zitzler-Deb-Theile (ZDT) test suite [6] is widely used for evaluating algorithms solving MOPs. The following five bi-objective MOPs named ZDT1, ZDT2, ZDT3, ZDT4 and ZDT6 were used for comparing the proposed Exploratory Crossover Operator (ECO) against the Simulated Binary Crossover (SBX). They have been used extensively for testing MOEAs and their Pareto front shapes are convex, nonconvex, disconnected, multimodal and non-uniform. ZDT1, ZDT2 and

ZDT3 use 30 decision variables and ZDT4 and ZDT6 use 10 decision variables respectively.

Zitzler-Deb-Thiele's function N.1 (ZDT1) problem:

$$Min = \begin{cases} f_1(x) = x_1 \\ f_2(x) = g(x)h(f_1(x), g(x)) \\ g(x) = 1 + \frac{9}{29}\sum_{i=2}^{30} x_i \\ h(f_1(x), g(x)) = 1 - \sqrt{\frac{f_1(x)}{g(x)}} \end{cases}$$
$$for\ 0 \le x_i \le 1\ and\ 1 \le i \le 30$$

Zitzler-Deb-Thiele's function N.2 (ZDT2) problem:

$$Min = \begin{cases} f_1(x) = x_1 \\ f_2(x) = g(x)h(f_1(x), g(x)) \\ g(x) = 1 + \frac{9}{29}\sum_{i=2}^{30} x_i \\ h(f_1(x), g(x)) = 1 - \left(\frac{f_1(x)}{g(x)}\right)^2 \end{cases}$$
$$for\ 0 \le x_i \le 1\ and\ 1 \le i \le 30$$

Zitzler-Deb-Thiele's function N.3 (ZDT3) problem

$$Min = \begin{cases} f_1(x) = x_1 \\ f_2(x) = g(x)h(f_1(x), g(x)) \\ g(x) = 1 + \frac{9}{29}\sum_{i=2}^{30} x_i \\ h(f_1(x), g(x)) = 1 - \sqrt{\frac{f_1(x)}{g(x)}} - \left(\frac{f_1(x)}{g(x)}\right)\sin(10\pi f_1(x)) \end{cases}$$
$$for\ 0 \le x_i \le 1\ and\ 1 \le i \le 30$$

Zitzler-Deb-Thiele's function N.4 (ZDT4) problem

$$Min = \begin{cases} f_1(x) = \\ f_2(x) = g(x)h(f_1(x), g(x)) \\ g(x) = 91 + \sum_{i=2}^{10} (x_i^2 - 10\cos(4\pi x_i)) \\ h(f_1(x), g(x)) = 1 - \sqrt{\frac{f_1(x)}{g(x)}} \end{cases}$$
$$for\ \ 0 \le x_1 \le 1, \quad -5 \le x_i \le 5, \qquad 2 \le i \le 10$$

Zitzler-Deb-Thiele's function N.6 (ZDT6) problem

$$Min = \begin{cases} f_1(x) = 1 - \exp(-4x_1)sin^6(6\pi x_1) \\ f_2(x) = g(x)h(f_1(x), g(x)) \\ g(x) = 1 + 9\left[\frac{\sum_{i=2}^{10} x_i}{9}\right]^{0.25} \\ h(f_1(x), g(x)) = 1 - \left(\frac{f_1(x)}{g(x)}\right)^2 \end{cases}$$
$$for\ 0 \le x_i \le 1\ and\ 1 \le i \le 10$$

The results in the Tables I and II have been produced by using JMetal [3] framework. Table I presents the results of ZDT1-4, 6 test functions. Specifically, it presents the mean, standard deviation (STD), median and interquartile range (IQR) of all the independent runs carried out for Hypervolume (HV) and Epsilon indicator respectively.

Clearly, regarding the HV [4], [7] indicator the higher the value (i.e. the greater the hypervolume) the better the computed front. HV is able of capturing in a single number both the closeness of the solutions to the optimal set and to a certain degree, the spread of the solutions across the objective space [8]. The second indicator, the Epsilon [5] is a measure of the smaller distance that a solution set A, needs to be changed in such a way that it dominates the optimal Pareto front of this problem. Obviously the smaller the value of this indicator, the better the derived solution set.

Table II use boxplots to present graphically the performance of NSGAII under two different configurations, ECO and SBX respectively, for HV and Epsilon performance indicators. Boxplot is a convenient way of graphically depicting groups of numerical data through their quartiles.

TABLE I
MEAN, STD, MEDIAN AND IQR FOR HV AND EPSILON

	Problem: ZDT1 NSGAII	
	ECO	SBX
HV. Mean and Std	6.60e-01$_{2.9e-04}$	6.59e-01$_{3.2e-04}$
HV. Median and IQR	6.60e-01$_{3.5e-04}$	6.59e-01$_{4.8e-04}$
EPSILON. Mean and Std	1.26e-02$_{2.0e-03}$	1.35e-02$_{2.4e-03}$
EPSILON. Median and IQR	1.25e-02$_{3.0e-03}$	1.29e-02$_{2.5e-03}$

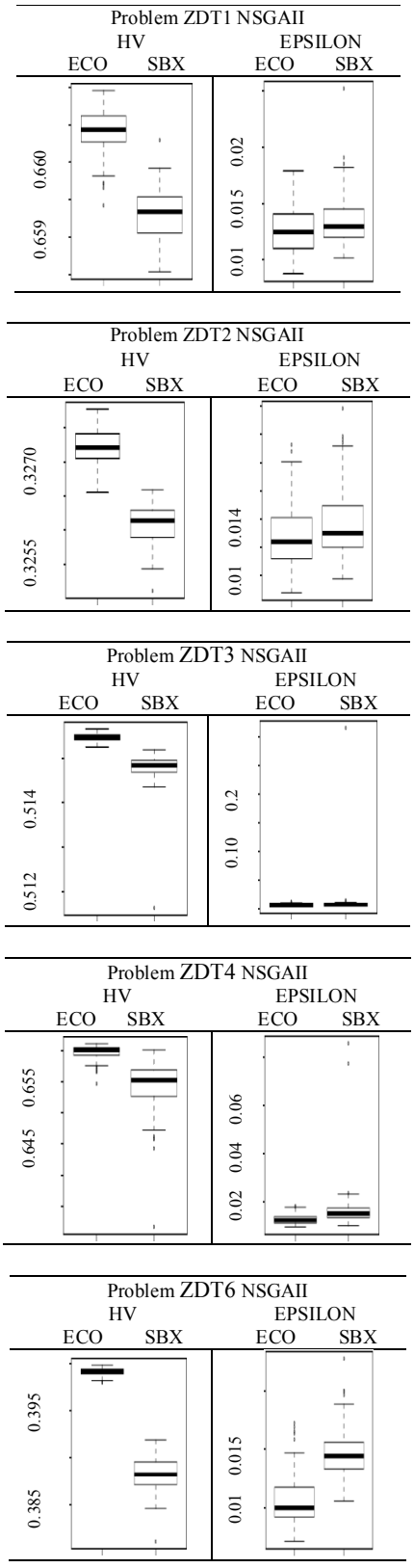
	Problem: ZDT2 NSGAII	
	ECO	SBX
HV. Mean and Std	3.27e-01$_{2.6e-04}$	3.26e-01$_{2.9e-04}$
HV. Median and IQR	3.27e-01$_{3.6e-04}$	3.26e-01$_{4.0e-04}$
EPSILON. Mean and Std	1.28e-02$_{2.1e-03}$	1.37e-02$_{2.5e-03}$
EPSILON. Median and IQR	1.24e-02$_{2.9e-03}$	1.30e-02$_{2.9e-03}$

	Problem: ZDT3 NSGAII	
	ECO	SBX
HV. Mean and Std	5.15e-01$_{8.5e-05}$	5.15e-01$_{3.6e-04}$
HV. Median and IQR	5.15e-01$_{1.1e-04}$	5.15e-01$_{2.7e-04}$
EPSILON. Mean and Std	7.87e-03$_{1.6e-03}$	1.14e-02$_{3.1e-02}$
EPSILON. Median and IQR	7.68e-03$_{1.8e-03}$	7.93e-03$_{1.9e-03}$

	Problem: ZDT4 NSGAII	
	ECO	SBX
HV. Mean and Std	6.60e-01$_{1.1e-03}$	6.54e-01$_{4.1e-03}$
HV. Median and IQR	6.60e-01$_{1.2e-03}$	6.55e-01$_{4.2e-03}$
EPSILON. Mean and Std	1.27e-02$_{2.0e-03}$	1.66e-02$_{9.7e-03}$
EPSILON. Median and IQR	1.23e-02$_{2.6e-03}$	1.50e-02$_{3.9e-03}$

	Problem: ZDT6 NSGAII	
	ECO	SBX
HV. Mean and Std	3.99e-01$_{4.3e-04}$	3.88e-01$_{1.7e-03}$
HV. Median and IQR	3.99e-01$_{5.1e-04}$	3.88e-01$_{2.4e-03}$
EPSILON. Mean and Std	1.06e-02$_{2.2e-03}$	1.46e-02$_{2.1e-03}$
EPSILON. Median and IQR	9.98e-03$_{2.6e-03}$	1.44e-02$_{2.3e-03}$

TABLE II
BOXPLOTS FOR HV AND EPSILON



VII. ANALYSIS OF THE RESULTS - CONCLUSIONS

In this section, we analyze the results obtained by applying the Exploratory Crossover Operator (ECO) and the Simulated Binary Crossover (SBX) operator respectively to the NSGAII for solving the ZDT1-4, 6 benchmark problems. The assessment of the performance of the proposed crossover operator is done with the assistance of two well known performance indicators, namely Hypervolume and Epsilon indicator.

Examining the results of both indicators we notice that the ECO performs better for all test functions examined, compared with results derived by the SBX operator. Moreover, by applying the Wilcoxon rank-sum test we validated that the observed difference in ECO and SBX performance is statistically significant with 95% confidence for all test functions examined. To conclude, the analysis of the results suggests that the proposed Exploratory Crossover Operator (ECO) demonstrates superior search ability than the Simulated Binary Crossover (SBX).

REFERENCES

[1] Koza, J.R. Genetic Programming. MIT Press, Cambridge; 1992.
[2] Deb, K. & Agrawal, R.B. (1995) Simulated binary crossover for continuous search space, Complex Systems 9 (2) 115–148.
[3] Durillo, J. J. & Nebro, A. J. (2011). jMetal: A Java framework for multi-objective optimization, Advances in Engineering Software 42, 760–771
[4] Zitzler, E., Brockhoff, D. & Thiele. L.(2007). The Hypervolume Indicator Revisited: On the Design of Pareto-compliant Indicators Via Weighted Integration. In Conference on Evolutionary Multi-Criterion Optimization (EMO 2007), pages 862–876. Springer.
[5] Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C.M. & Da Fonseca, V.G. (2003). Performance assessment of multiobjective optimizers: an analysis and review. IEEE Transactions on Evolutionary Computation, 7(2):117–132.
[6] Zitzler, E., Deb, K. & Thiele., L. (2000). Comparison of multiobjective evolutionary algorithms: Empirical results. Evolutionary Computation, 8(2):173–195, Summer.
[7] Fonseca C, Flemming P., Multiobjective optimization and multiple constraint handling with evolutionary algorithms - part ii: Application example. IEEE Transactions on System, Man, and Cybernetics, 28:38–47, 1998.
[8] Zeng, F., Decraene, J., Hean Low, M.Y., Hingston, P., Wentong, C., Suiping, Z. & Chandramohan, M. (2010). Autonomous Bee Colony Optimization for Multi-objective Function. Proceedings of the IEEE Congress on Evolutionary Computation.

K. Metaxiotis is an Associate Professor at the Department of Informatics, University of Piraeus. His research interests include: Decision Support Systems, ERPs, Artificial Intelligence, Expert Systems, Neural Networks, Genetic Algorithms and Knowledge Management. He is a widely recognised researcher, with more than 450 citations.

K. Liagkouras is a PhD Candidate at the Decision Support Systems Laboratory, Department of Informatics, University of Piraeus. He holds a BSc in Economics (Piraeus), a MPhil in International Finance (Glasgow) and a MSc in Computer Systems & Networking (London). His research interests include application of artificial intelligence techniques in finance and economics, operational research and decision sciences.

Modelling of high-temperature behaviour of cementitious composites

Jiří Vala, Anna Kučerová and Petra Rozehnalová

***Abstract*—The computational prediction of non-stationary behaviour of cementitious composites exposed to high temperature, supported by proper physical and mathematical formulations, is a rather complicated analysis, dealing with the coupled heat and (liquid and vapour) mass transfer and related mechanical effects. The macroscopic mass, momentum and energy conservation equations need to exploit available information from microstructural considerations. The derivation and implementation of a computational model is always a compromise between the model complexity and the reasonable possibility of identification of material characteristics of all components and their changes. The paper presents a possible approach to such analysis, supplied by an illustrative example, confronting computational results from a simplified model (taking only dominant physical processes into account), with experimental ones, performed in the specialized laboratory at the Brno University of Technology.**

***Index Terms*—Cementitious composites, high temperature behaviour, poro-mechanics of multi-phase media, computational simulation.**

I. Introduction

PROBLEMS concerning cementitions composites, especially concrete both in its regular and more advanced forms, as high-strength, stamped, (ultra-)high-performance, self-consolidating, pervious or vacuum concrete, limecrete, shotcrete, etc., exposed to elevated temperatures are significant and wide ranging. The increasing interest of designers of building structures in this field in several last decades has been driven by the exploitation of advanced materials, structures and technologies, reducing the external energy consumption, whose thermal, mechanical, etc. time-dependent behaviour cannot be predicted using the simplified semi-empirical formulae from most valid European and national technical standards. Such models as that of one-dimensional heat conduction with (nearly) constant thermal conductivity and capacity, give no practically relevant results for refractory or phase change materials, thermal storage equipments, etc. – cf. [11]. Fortunately, in such cases the correlation between the predicted and measured quantities, e. g. of the annual energy consumption of a building, can be available, to help to optimize the design of both new and reconstructed building structures. This is not true for high temperatures leading to partial or total collapse of a structure, as those caused by thermal radiation during a fire.

A lot of historical remarks and relevant references for the analysis of structures exposed to fire can be found in [5]. The physical background for the evaluation of simultaneous temperature and moisture redistributions under such conditions is based on the balance law of classical thermodynamics. However, it is not trivial to supply a resulting system of evolution by reasonable effective constitutive relations, valid at the microstructural level, despite of the complicated (typically porous) material structures. Most classical approaches to evaluate temperature and moisture redistributions rely on the (semi-)empirical relations from [1]. The attempts to improve them can be seen in various directions, with still open problems and difficulties everywhere. The physical and/or mathematical homogenization approaches applied to a reference volume element seem to be useful. The process of cement hydration, distinguishing between anhydrous cement scale, cement-paste scale, mortar scale and macro-scale, handled by specific physical and mathematical formulations, with a posteriori scale bridging applying least squares arguments, has been described in details, evaluating the hydration of particular clinker phases, in [14]. However, the thermal gains and losses from chemical reactions cannot be considered as deterministic ones, moreover corresponding computations just in the case of dehydration must suffer from the lack of relevant data at some scales. The formal mathematical two-scale or similar homogenization, introduced in [4], needs to remove assumptions of periodicity using stochastic or abstract deterministic homogenization structures; this leads to physically non-transparent and mathematically very complicated formulations, whose applicability to the construction of practical algorithms for the analysis of engineering problems is not clear. From such point of view, the most promising approach of the last years seems to be that of [9] and (in a substantially generalized version) of [8], replacing the proper homogenization by some arguments from the mixture theory; this will be the principal idea even in our following considerations in this short paper.

J. Vala, A. Kučerová and P. Rozehnalová are with the Brno University of Technology, Faculty of Civil Engineering, Czech Republic, 602 00 Brno, Veveří 95.

II. Physical and mathematical background

Following [8] (unlike [14]), for the quantification of the dehydration process we shall work with the hydration degree Γ, a number between 0 and 1, as the part of hydrated (chemically combined) water m^w in its standard mass, constituted usually during the early-age treatment of a cementitious composite. Although the evaluation of Γ from a simple algebraic formula is not available because it must take into account the chemical affinity and the fact that the accessibility of water for chemical reactions is controlled by the water content η^w inside the pores under certain temperature T, we shall apply Γ as a known function of such (or slightly transformed) variables.

The multiphase medium at the macroscopic level can be considered as the superposition of 4 phases: solid material, liquid water, water vapour and dry air, identified by their indices $\varepsilon \in \{s, w, v, a\}$. In the Lagrangian description of motion,

ISBN: 978-1-61804-240-8

following [16], the deformation tensor F^s can be derived using the derivatives of displacements of particular points with respect to Cartesian coordinate system $x = (x_1, x_2, x_3)$ in the 3-dimensional Euclidean space. If ω^ε is a source corresponding to certain scalar quantity ϕ^ε then the conservation of such quantity can be expressed by [3], p. 4, and [6], p. 9, as

$$\dot{\phi}^\varepsilon + (\phi^\varepsilon v_i^\varepsilon)_{,i} = \omega^\varepsilon \,; \tag{1}$$

this formula contains the dot notation for the derivative with respect to any positive time t, $(\dots)_{,i}$ means the derivative with respect to x_i where $i \in \{1,2,3\}$, $v_i^\varepsilon = \dot{u}_i^\varepsilon$, with u_i^ε referring to displacements related to the initial geometrical configuration $x_0 = (x_{01}, x_{02}, x_{03})$ (for $t = 0$) and later also $a_i = \ddot{u}_i$; the Einstein summation is applied to i and j from $\{1,2,3\}$ everywhere. For the brevity, ϕ_ε will be moreover used instead of $\phi^\varepsilon \eta^\varepsilon$ where $\eta^\varepsilon(n,S)$ is the volume fraction of the phase ε, a function of the material porosity n and the saturation S. Clearly $\det F^s = (1-n)/(1-n_0)$ with n_0 corresponding to x_0. The saturation S is an experimentally identified function of the absolute temperature T and of the capillary pressure p^c, needed later; the assumption of local thermal equilibrium yields the same values of T for all phases.

In addition to T, we have 4 a priori unknown material densities ρ^ε and 12 velocity components v_i^ε. Assuming that vapour and dry air are perfect gases, we are able to evaluate their pressures $p^v(\rho^v, T)$ and $p^a(\rho^a, T)$ from the Clapeyron law; the capillary pressure is then $p^c = p^v + p^a - p^w$, with the liquid water pressure p^w, or alternatively just with p^c, as an additional unknown variable. In the deterioration of a composite structure 2 crucial quantities occur: the mass m^w of liquid water lost from the skeleton and the vapour mass m^v caused by evaporation and desorption. The time evolution of mass m^w can be determined from the formally simple formula $m^w = \Gamma m_0^w$, with m_0 related to $t = 0$. The vapour fraction ζ remains to be calculated from the system of balance equations of the type (1), supplied by appropriate constitutive relations.

The mass balance works with

$$\phi^\varepsilon = \rho_\varepsilon \,,$$
$$\omega^s = -\dot{m}^w \,,\ \omega^w = \dot{m}^w - \dot{m}^v \,,\ \omega^v = \dot{m}^v \,,\ \omega^a = 0$$

in (1). No additional algebraic relations are necessary.

Since all phases are considered as microscopically non-polar, the angular momentum balance forces only the symmetry of the partial Cauchy stress tensor τ, i. e. $\tau_{ij} = \tau_{ji}$ for such stress components. The formulation of the linear momentum balance in 3 direction is more delicate. Introducing $w_i^\varepsilon = \rho_\varepsilon v_i^\varepsilon$ and choosing (for particular i)

$$\phi^\varepsilon = w_i^\varepsilon \,,$$

we can evaluate, using the Kronecker symbol δ, the total Cauchy stress σ in the form $\sigma_{ij} = \tau_{ij}\delta^{s\varepsilon}$ and finally set

$$\omega^\varepsilon = \sigma_{ij,j}^\varepsilon + \rho_\varepsilon(g_i - a_i^\varepsilon + t_i^\varepsilon)$$

in (1) where g_i denotes the gravity accelerations and t_i^ε the additional accelerations caused by interactions with other phases, whose evaluation is possible from the Darcy law, as explained lower. The constitutive relationships for the solid phase, e. g. those between τ and u^s, v^s, etc., need the multiplicative decomposition into a finite number m of matrix components $F^s = F^{s1} \dots F^{sm}$ (elasticity, creep, damage, etc. ones) to express $\tau(F^{s1} \dots F^{sm}, \dot{F}^{s1} \dots \dot{F}^{sm}, \dots)$. For $\varepsilon \neq s$,

Fig. 1. Fire simulation at the Brno University of Technology: laboratory setting (upper photo), detail of real experiment (lower photo).

introducing the dynamical viscosity μ^ε and the permeability matrix K_{ij}^ε, depending on ρ_ε again, we can formulate the Darcy law as

$$\mu^\varepsilon \rho_\varepsilon (v_i^\varepsilon - v_i^s) = K_{ij}^\varepsilon (\rho_\varepsilon (g_j - a_j^\varepsilon + t_j^\varepsilon) - p_{\varepsilon,j}) \,. \tag{2}$$

The energy balance inserts

$$\phi^\varepsilon = \frac{1}{2} w_i^\varepsilon v_i^\varepsilon + \rho_\varepsilon \kappa^\varepsilon$$

with κ^ε usually defined as $c^\varepsilon T$, using the thermal capacities c^ε, in general functions of T and p^c again, and some internal heat fluxes q_i^ε, and also

$$\omega^\varepsilon = (\sigma_{ij,j}^\varepsilon + q_i^\varepsilon)_{,j} + (g_i - a_i^\varepsilon + t_i^\varepsilon) + \varpi^\varepsilon \,,$$
$$\varpi^s = -\dot{m}^w h^w \,,\ \varpi^w = \dot{m}^w h^w - \dot{m}^v h^v \,,\ \varpi^v = \dot{m}^v h^v \,,$$
$$\varpi^a = 0$$

into (1); two new characteristics here are the specific enthalpies of cement dehydration h^w and evaporation h^v. The internal heat fluxes q_i^ε come from the constitutive relation

$$q_i^\varepsilon = -\lambda_{ij}^\varepsilon T_{,j} - \xi_{ij}^\varepsilon p_{,j}^c \,; \tag{3}$$

the first (typically dominant) additive term corresponds to the well-known Fourier law of thermal conduction, the second one to the Dufour effect, with some material characteristics λ_{ij}^ε and ξ_{ij}^ε dependent on T and p^c. Similarly to (3), it is possible derive the diffusive fluxes $r_i^\varepsilon = \rho_\varepsilon(v_i^\varepsilon - v_i^s)$ with $i \in \{v, w\}$ in the form

$$r_i^\varepsilon = -\varsigma_{ij}^\varepsilon T_{,j} - \gamma_{ij}^\varepsilon p_{,j}^c \,, \tag{4}$$

due to the Fick law (the second additive term), respecting the Soret effect (the first one), with some material characteristics $\varsigma_{ij}^\varepsilon$ and γ_{ij}^ε dependent on T and p^c.

Now we have 4 mass balance equations, 3×4=12 momentum ones and 4 energy ones, in total 20 partial differential equations of evolution for 3 groups of variables:

$$\mathcal{R} = (\rho^s, \rho^w, \rho^v, \rho^a) \,,$$
$$\mathcal{V} = (v_1^s, v_2^s, v_3^s, v_1^w, v_2^w, v_3^w, v_1^v, v_2^v, v_3^v, v_1^a, v_2^a, v_3^a) \,,$$
$$\mathcal{T} = (T, p^c, m^v, \zeta) \,,$$

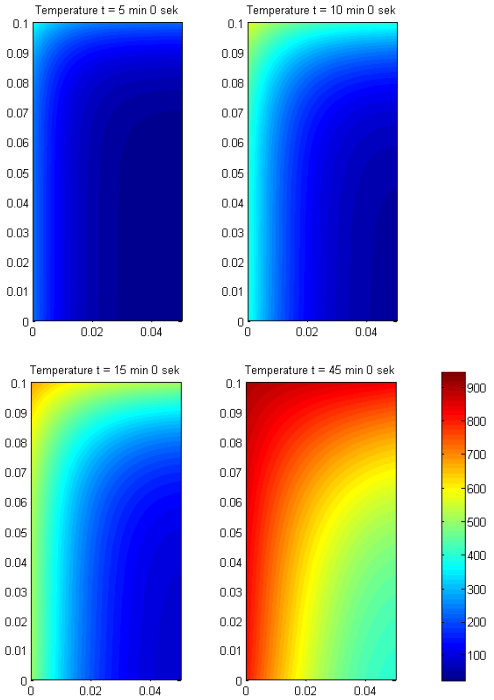


Fig. 2. Time development of temperature T.

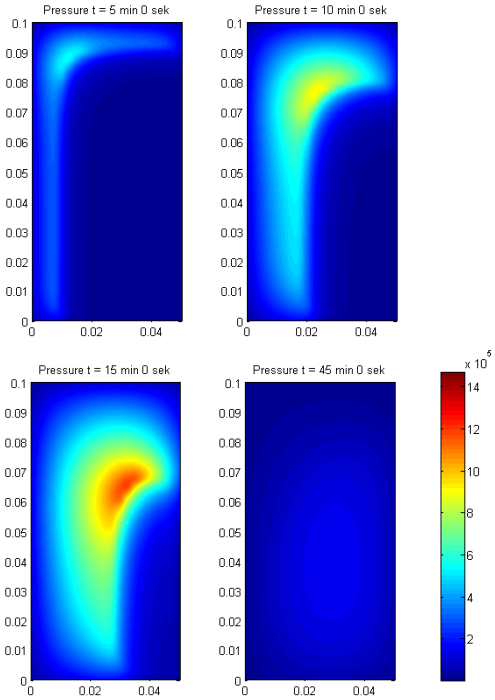


Fig. 3. Time development of pressure p^c.

supplied by appropriate initial and boundary conditions, e. g. for a priori known values of all variables for $t = 0$ of the Dirichlet, Cauchy, Neumann, Robin, etc. types, for local unit boundary outward normals $n(x) = (n_1(x), n_2(x), n_3(x))$ in particular

- $\sigma_{ij} n_j = \bar{t}_i$ with imposed tractions t_i,
- $(\rho_a(v_i^a - v_i^s) + r_i^a) n_i = \overline{r}^a$ with imposed air fluxes $\overline{r}^a$,
- $(\rho_w(v_i^w - v_i^s) + r_i^w + \rho_v(v_i^v - v_i^s) + r_i^v) n_i = \overline{r}^w + \overline{r}^v + \beta(\rho_v)$ with imposed liquid water and vapour fluxes $\overline{r}^w$, $\overline{r}^v$ and some mass exchange function β,
- $(\rho_w(v_i^w - v_i^s) h_v - \lambda_{ij} T_{,j} - \xi_{ij}^{\varepsilon} p_{,j}^c) n_i = \bar{q} + \alpha(T)$ with imposed heat fluxes and some heat exchange function α, e. g. by the Stefan - Boltzmann law, proportional to T^4.

This seems to be a correct and complete formulation for the analysis of time development of $\mathcal{R}$, $\mathcal{V}$ and $\mathcal{T}$.

III. Computational prediction

Unfortunately, a lot of serious difficulties is hidden in the above presented formulation, e. g. the still missing "Millenium Prize" existence result on the solvability of Navier - Stokes equations (cf. the "mysteriously difficult problem" of [15], p. 257), the physical and mathematical incompatibilities like [7], as well the absence of sufficiently robust, efficient and reliable numerical algorithms based on the intuitive time-discretized computational scheme:

1. set $\mathcal{R}$, $\mathcal{V}$ and $\mathcal{T}$ by the initial conditions for $t = 0$,
2. go to the next time step, preserving $\mathcal{R}$, $\mathcal{V}$ and $\mathcal{T}$,
3. solve some linearized version of (1) with the mass balance choice, evaluate and perform the correction $\epsilon_{\mathcal{R}}$,
4. solve some linearized version of (1) with the momentum balance choice, evaluate and perform the correction $\epsilon_{\mathcal{V}}$,
5. solve some linearized version of (1) with the energy balance choice, evaluate and perform the correction $\epsilon_{\mathcal{T}}$,
6. if $\epsilon_{\mathcal{R}}$, $\epsilon_{\mathcal{V}}$ and $\epsilon_{\mathcal{T}}$ are sufficiently small, return to 3,
7. stop the computation if the final time is reached, otherwise return to 2.

Consequently all practical computational tools make use of strong simplifications. The rather general approach of [8] introduces the additive linearized strain decomposition instead of the multiplicative finite strain one, ignores some less important terms, as the kinetic energy (the first additive term in ϕ^{ε}) in the energy balance interpretation of (1), as well as the Dufour and Soret effects in (3) and (4), and reduces the number of variables, comparing (2) with the differences $v_i^{\varepsilon} - v_i^s$ with $\varepsilon \in \{v, w, a\}$ from the momentum balance interpretation of (1), and presents a computational scheme of the above sketched type, applying the Galerkin formulation and the finite element technique. Nevertheless, most engineering approaches, as [2] or [12], endeavour to obtain a system of 2 equations of evolution for T and p^c (or some equivalent quantity) only, pre-eliminating or neglecting all remaining ones, using arguments of various levels: from micro-mechanically motivated physical considerations to formal simplification tricks.

The reliability of computational results depends on the quality and precision of input data, including the design and identification of all material characteristics. Probably all authors take some of them from the literature, not from their extensive original experimental work; the variability of forms of such characteristics and generated values, accenting those from (2), including the formulae from [8], has been discussed in [5]. Nevertheless, the proper analysis of uncertainty and significance of particular physical processes lead to even more complicated formulations, analyzed (for much easier direct model problems) in [17] using the spectral stochastic finite element technique, or in [10] the Sobol sensitivity indices, in both cases together with the Monte Carlo simulations.

ISBN: 978-1-61804-240-8

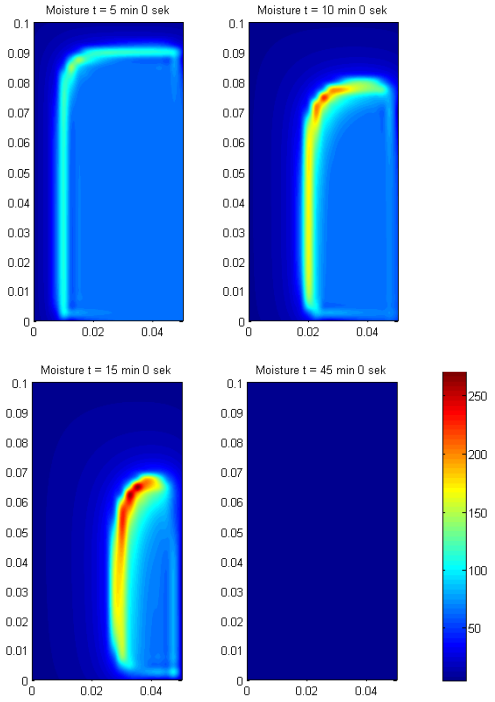


Fig. 4. Time development of moisture content $\rho_w + \rho_v$.

Unlike most experiments in civil and material engineering in laboratories and in situ, nearly no relevant results for advanced numerical simulations are available from real unexpected fires, as the most dangerous conditions for building structures, crucial for the reliable prediction of behaviour of concrete and similar structures. Since also laboratory experiments with cementitious composites under the conditions close to a real fire, as that documented on Figure 1, performed at the Brno University of Technology, are expensive, producing incomplete and uncertain quantitative data anyway, the reasonable goal of numerical simulation is to test simple numerical models with acceptable correlation with real data, as the first step for the development of more advanced multi-physical models.

Figures 2, 3 and 4 are the outputs from a rather simple two-dimensional isotropic model, neglecting mechanical loads, strains and stresses, compatible with the slightly modified and revisited approach of [2], combining the finite element and volume techniques together with the iterated time discretization scheme (the numerical construction of Rothe sequences), implemented in the MATLAB environment. The fire is considered as the boundary thermal radiation on the left and upper edges of the rectangle. Surprisingly some phenomena observed in situ can be explained from the deeper analysis of results of such seemingly simple calculations.

IV. Conclusion

Development of reasonable models and computational algorithms for the prediction of thermal, mechanical, etc. behaviour of cementitious composites and whole building structures, is strongly required by the applied research in civil engineering, which cannot be ignored, although the formal mathematical verification of such models is not available and the practical validation suffers from the lack of data from observations in situ, as well as of large databases from relevant (very expensive) experiments. The first steps, both in the sufficiently general formulation of the problem, containing most practical computational approaches as special ones, as well as the methodology of computational and experimental work, sketched in this paper, should be a motivation for further extensive research.

Acknowledgment

This research has been supported by the research project FAST-S-14-2490 at Brno University of Technology, in collaboration with the research project FAST-S-14-2346.

References

[1] Z. P. Bažant, W. Thonguthai, Pore pressure and drying of concrete at high temperature. *Journal of the Engineering Mechanics Division* 104, 1978, pp. 1059–1079.

[2] M. Beneš, R. Štefan and J. Zeman, Analysis of coupled transport phenomena in concrete at elevated temperatures. *Applied Mathematics and Computation* 219, 2013, pp. 7262–7274.

[3] A. Bermúdez de Castro, *Continuum Thermomechanics*. Birkhäuser, 2005.

[4] D. Cioranescu and P. Donato, *An Introduction to Homogenization*. Oxford University Press, 1999.

[5] C. T. Davie, C. J. Pearce and N. Bićanić, Aspects of permeability in modelling of concrete exposed to high temperatures. *Transport in Porous Media* 95, 2012, pp. 627–646.

[6] J. H. Ferziger and M. Perić, *Computational Methods for Fluid Dynamics*. Springer, 2002.

[7] T. Fürst, R. Vodák, M. Šír and M. Bíl, On the incompatibility of Richards' equation and finger-like infiltration in unsaturated homogeneous porous media. *Water Resource Research* 45, 2009, W0340, 12 pp.

[8] D. Gawin and F. Pesavento, An overview of modeling cement based materials at elevated temperatures with mechanics of multi-phase porous media. *Fire Technology* 48, 2012, pp. 753-793.

[9] D. Gawin, F. Pesavento and B. A. Schrefler, Simulation of damage-permeability coupling in hygro-thermo-mechanical analysis of concrete at high temperature. *Communications in Numerical Methods and Engineering* 18, 2002, pp. 113-119.

[10] J. Gottvald and Z. Kala, Sensitivity analysis of tangential digging forces of the bucket wheel excavator SchRs 1320 for different terraces. *Journal of Civil Engineering and Management* 18, 2012, pp. 609–620.

[11] P. Jarošová and S. Šťastník, Modelling of thermal transfer for energy saving buildings. *International Conference on Numerical Analysis and Applied Mathematics 2013 – Proceedings*, American Institute of Physics, 2013, pp. 1000–1003.

[12] V. Kočí, J. Maděra, T. Korecký, M. Jerman and R. Černý, Computational analysis of coupled heat and moisture transport in building envelopes: the effect of moisture content. *Thermophysics 2013 – Conference Proceedings* in Podkylava (Slovak Republic). Slovak Academy of Sciences in Bratislava, 2013, pp. 62–71.

[13] G. Nguetseng and N. Svanstedt, σ-convergence. *Banach Journal of Mathematical Analysis*, 5, 2011, pp. 101–135.

[14] Ch. Pichler, R. Lackner and H. A. Mang, Multiscale model of creep of shotcrete – from logarithmic type viscous behaviour of CSH at the μ-m scale to macroscopic tunnel analysis. *Journal of Advanced Concrete Technology* 6, 2008, pp. 91–110.

[15] T. Roubíček, *Nonlinear Partial Differential Equations with Applications*. Birkhäuser, 2005.

[16] L. Sanavia, B. A. Schrefler and P. Steinmann, A formulation for an unsaturated porous medium undergoing large inelastic strains. *Computational Mechanics* 1, 2002, pp. 137–151.

[17] N. Zabaras, Inverse problems in heat transfer. In: *Handbook on Numerical Heat Transfer* (W. J. Minkowycz, E. M. Sparrow and J. S. Murthy, eds.), Chap. 17. John Wiley & Sons, 2004.

Jiří Vala (*1956) is professor of the Institute of Mathematics and Descriptive Geometry at the Brno University of Technology, Faculty of Civil Engineering. **Anna Kučerová** (*1985) and **Petra Rozehnalová** (*1984) are Ph.D. students at the same university.

Gaussian Mixture Models Approach for Multiple Fault Detection - DAMADICS Benchmark

Erika Torres, Edwin Villareal

***Abstract*—In fault detection, data-driven methods are limited by the lack of knowledge about faults. This paper presents an approach based in statistical modeling, which uses normal operation information for training a Gaussian Mixture Model. Such model has been widely used in speaker verification, because is capable of mapping the observations into a multidimensional membership function, this feature makes this algorithm very convenient because works in a unsupervised way. This technique uses the normal behavior data to estimate the parameters and then classify future data according to the model. The proposed approach consists of two stages: an offline stage, where the Gaussian model is trained by using Expectation-Maximization algorithm; and the online stage, where the readings likelihood is evaluated according to the normal operation model. If the likelihood is found to be inferior to a certain threshold, a fault is detected. This approach may be extended to multiple processes in science and engineering; for instance, a case study is presented using the DAMADICS benchmark, where this approach was validated.**

***Keywords*—Fault Detection, Gaussian Mixture Models, Industrial Processes**

I. Introduction

As the industrial processes complexity keeps increasing, fault detection has become a critical issue in order to meet safety, enviromental and productivity requirements. In this regard, several approaches have been proposed, particularly, being artificial intelligence methods the current trend. However, one of the main drawbacks of these methods is the implementation complexity for a real industrial environment, which limits its appeal for most industrial applications. Hence, approaches that may incur in a lower complexity while keeping acceptable results are highly favored. The approach presented in this paper aims at this paradigm.

Computer-based solutions have been proposed to meet the aforementioned requirements, for instance, machine learning methods, see [1], like Bayesian Learning methods [2], Self-Organizing Maps [3], or PCA [4],etc. The main objective of these methods is to transform the monitoring data (usually sensors) into statistical models, in order to classify and detect anomalies. Since the analytical model is not taken into account, the method remains flexible for implementation in different environments.

However, a shortcoming of artificial intelligence methods is the need of data from different anomalies or failures to train the model. Some systems have low on-site failure rate, and as a consequence, the amount of failure data available is limited, see [5]. In this paper, a GMM approach for fault detection is proposed, to overcome the lack of failure data, as this method can be trained using only normal operation data.

Regarding industrial processes, failures usually occur in sensors, actuators or processes. The main purpose of fault detection is to evaluate symptoms, which indicate the difference between normal and faulty states. As such, and taking into account the performed observation, the methods of failure detection are divided into three categories: signal threshold, signal model and process model based approaches, see [6].

From the taxonomy of failure detection presented previously, the proposed approach in this paper falls within the signal model-based category, as it extracts the statistical features of the process from the sensor readings, where the training is taking place off-line. As such, this approach is suitable for real-time applications, because the operations required for classification can be calculated in small batches while new data is being read.

This paper is organized as follows. Firstly, A brief review of GMM is presented, fault detection by using GMM is described in Section 3. The proposed scheme is validated by simulation examples of Chemical Plant Benchmark (DAMADICS) in Section 4. Finally, we present the conclusions obtained from the implementation and discuss future work.

II. Gaussian Mixture Models

Gaussian mixture models are commonly used in statistical analysis. In most applications, its parameters are determined by using Maximum Likelihood and Expectation Maximization algorithm.

As industrial processes are influenced by multiple facts, like temperature, noise and environmental conditions, it is required to include adequate modelling to these variables. The GMM has the capability of representing a large class of sample distributions, which is a very useful tool for the characterization of multiple problems, like speaker recognition, see [7].

In the context of fault detection systems, the usage of a GMM for representing features distributions, is motivated by the notion that the individual component densities can model a set of hidden classes. In the same context, we assume that the features in the normal operation space correspond to different hidden classes. We assume that if we learn a Gaussian Mixture Model with sufficient variability of normal operation behavior, then the model will be capable of distinguish between any state different than the normal one. Because all the features used to train the GMM are unlabeled, the healthy state classes are hidden in a class of an observation that is unknown.

ISBN: 978-1-61804-240-8

The mixture model is a convex combination of unimodal gaussian functions:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i) \quad (1)$$

$$\sum_{i=1}^{M} w_i = 1 \quad (2)$$

where w_i are the *mixing coefficients* and the parameters of the of the component density functions $p(\mathbf{x}|\lambda)$ which usually vary with i.

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{(x-\mu_i)^T(x-\mu_i)}{2\Sigma_i}\right) \quad (3)$$

where $g(x|\mu_i, \Sigma_i)$ are the component Gaussian densities. The set of parameters is $\lambda_i = (w_i, \mu_i, \Sigma_i)$, where $i = 1, \ldots, M$, M is the GMM model size, μ_i and Σ_i are the mean and covariance matrix of the ith GMM component, respectively: Finally, D is the dimension of vector x.

To determine the "membership" of the data point x to the ith GMM component, the loglikelihood of the data point, of each individual model distribution is calculated using the Equation 4 . An implementation of this algorithm for Matlab can be found at [8].

$$\log p(x|\lambda) = \sum_{t=1}^{T} \log p(x_t|\lambda) \quad (4)$$

A. Expectation Maximization Algorithm for Gaussian Mixtures

Given a set of training vectors and a GMM configuration (model size, minimum variance, which are determined heuristically),the parameters of the GMM λ are left to be estimated. These parameters, should lead to the best match (in the maximum likelihood sense) for the distribution of the training features vectors. There are several techniques available for estimating the parameters of the GMM, being Maximum Likelihood (ML) the most used method. The aim of ML estimation is to find the model parameters which maximize the likelihood of the GMM given the training data. For a sequence of T training vectors $X = \{\mathrm{x}_1, \ldots, \mathrm{x}_T\}$, the GMM likelihood, assuming independence between the vectors, can be written as

$$p(X|\lambda) = \prod_{t=1}^{T} p(\mathrm{x}_i|\lambda) \quad (5)$$

Unfortunately, for this expression direct maximization is not possible. This equation, however, can be solved iteratively using a special case of the EM algorithm. The goal of this algorithm is to maximize the likelihood function (Equation 4) according to the parameters: means, covariances and the mixing coefficients for all the components. The process is divided in the following stages:

1) Initialize the means μ_i , covariances σ_i and mixing coefficients w_i, and evaluate the initial value of the likelihood.

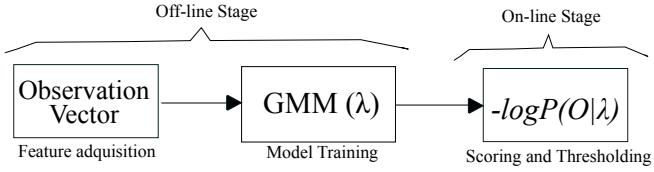


Fig. 1. Block diagram for fault detection by using GMM

2) Evaluate probabilistic distance using the current parameter values

$$\overline{w}_i = \frac{1}{T}\sum_{t=1}^{T} \Pr(i|\mathrm{x}_t, \lambda) \quad (6)$$

$$\overline{\mu}_i = \frac{\sum_{t=1}^{T} \Pr(i|\mathrm{x}_t, \lambda)\mathrm{x}_t}{\sum_{t=1}^{T} \Pr(i|\mathrm{x}_t, \lambda)} \quad (7)$$

$$\overline{\sigma}_i^2 = \frac{\sum_{t=1}^{T} \Pr(i|\mathrm{x}_t, \lambda)x_t^2}{\sum_{t=1}^{T} \Pr(i|\mathrm{x}_t, \lambda)} - \overline{\mu}_i \quad (8)$$

3) Calcule the aposteriori data probabilities.

$$\Pr(i|\mathrm{x}_t, \lambda) = \frac{w_i g(\mathrm{x}_t|\mu_i, \Sigma_i)}{\sum_{k=1}^{M} w_k g(\mathrm{x}_t|\mu_k, \Sigma_k)} \quad (9)$$

4) Calcule the error, which is the negative of the likelihood function.
5) Go back to step 2 until the error change is very small or the number of iterations is reached (usually 15-20).

where σ^2,x_t, and μ_i, refer to arbitrary elements of the vectors σ^2, $\mathbf{x}_t$, and μ_i respectively.

III. Fault detection using GMM

The main advantage of using Gaussian Mixture Models, is that the model can be trained with normal operation data, without further information about the multiple failure states, which can be unknown at the moment. This advantage, allows the detection of multiple failure states even if a particular failure is happening for the very first time.

The signals can be used directly or transformed to extract more interesting features for the detection task. In this article we use the raw signals to learn the model.

The principle of this method consists of two phases: a first phase during which an statistical model is learned using the Expectation Maximization algorithm; and a second phase where the learned model is used to score the current system condition, as it is shown in Figure 1. If the maximum likelihood score exceeds certain threshold, then the detection system triggers a failure alarm.

The proposed method described in the Algorithm 1, is a data-based technique which mainly uses the sensor data (consisting as well of normal operating data called $\mathbf{x}_{tr}$) gathered during a significant period of time to adjust the parameters of a Gaussian mixture model called λ of size M. The threshold is obtained using validation data (also normal operation data), called $\mathbf{x}_v$. The obtained model uses log Likelihood to distinguish between data that is "normal", and the data that is "anomalous", the data set for testing

is called $\mathbf{x}_t$. In the algorithm the testing data is fed to the model through a window of time, in order calculate the loglikelihood when sufficient data is gathered.

Algorithm 1 Fault detection by using Gaussian Mixture Models

1: **procedure** FDGMM(train data)
2: Organize train data in a matrix called $\mathbf{x}_{tr}$.
3: Set the parameters for the algorithm.
4: λ_0=*kmeans*($\mathbf{x}_{tr}$).
5: λ=*gmmEM*($\mathbf{x}_{tr}$,M,λ_0)
6: l=*loglikelihood*($\mathbf{x}_v$,λ).
7: Set the likelihood threshold ϵ according to l,to detect a failure.
8: **for** $\mathbf{x}_t$(1:$window$:end) **do**
9: l=*loglikelihood*($\mathbf{x}_t$,λ)
10: **if** l$<\epsilon$ **then**
11: Hey, something is really wrong.
12: **else**
13: Everything is cool
14: **end if**
15: **end for**
16: **end procedure**

IV. The testing benchmark DAMADICS

This benchmark was proposed in 2003 [9] to help the development of fault detection methods in a industry environment. The process has multiple sensors and actuators which can present different types of failure, abrupt and incipient. From the repository we use the normal behavior data to train the model, and the failure data to test the results. The faulty data is scarce, but we simulated 30 cases to test our approach with the simulink tool available in the repository.

V. Experiments

To validate the functionality of the algorithm we use the DAMADICS benchmark. This database was created to test fault detection algorithms with real data and simulations. The authors provided libraries to simulate 19 types of faults from three actuators with two different degrees of ocurrence (Abrupt, Incipient). In this experiment we monitor three signals: controller output, upstream pressure and downstream pressure.

In Table I, the analysed fault types are presented, and the fault scenarios that are derived from them. As the monitoring of root causes for a faulty event is a very difficult task for the operators, it follows to infer the state of the process, from key variables belonging to it.

A. Data generation

The actuator model is simulated using Matlab Simulink and the libraries provided in DAMADICS webpage. The available data is separated into three sets, one for training (1 st to 4th of November), a second for validation (5th of November) and other for testing (October 30th, November 9th and November 17th). The observation sequences are

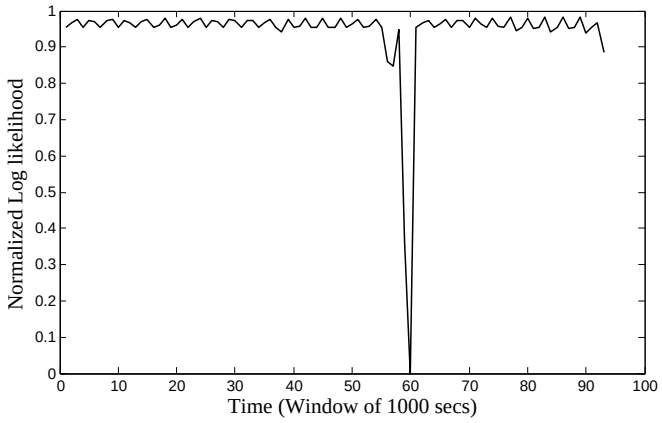


Fig. 2. Normalized Loglikelihood for fault 16, corresponding to November 30th. This picture indicates that the failure occurred near to the time window number 60.

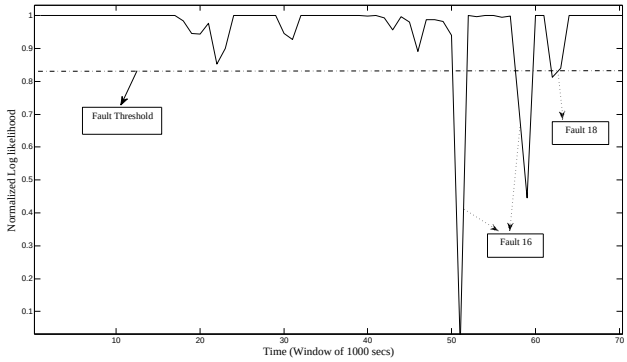


Fig. 3. Normalized Loglikelihood for data set, corresponding to November 17th, where multiple types of faults occurred. In this picture we can see that three faults occurred after the time window number 50.

sensors and valves readings, whose size depends on the failure mode, for this article we only consider the 16, 17 and 18 faults, which are general external failures. The model size was determined heuristically, as well the minimum variance and detection threshold.

B. Model identification

The goal of this stage is to obtain a GMM for normal operation data. The data set concerning the failures is used only in the testing stage. The first step uses the k-means algorithm to locate the centroids for the initial parameters. The second step relies in the EM algorithm to estimate

TABLE I. Faults under analysis

Date	Fault type	Description
October 30th	18	Partially opened bypass valve
November 9th	16	Positioner supply pressure drop
November 17th	17	Unexpected pressure drop across the valve

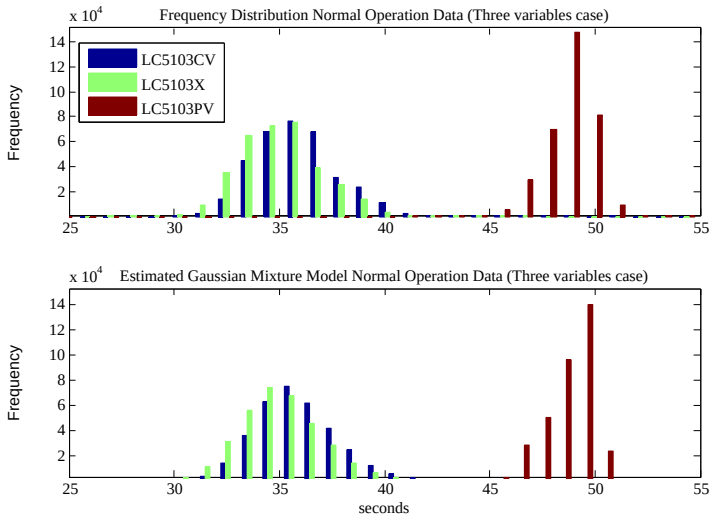


Fig. 4. Upper: Histogram of Normal operation data set corresponding to November 1st to 5th for training purposes. Below: Histogram of learned Gaussian Mixture model.

the GMM parameters, where the covariance matrices are set to be diagonal. The appropriate size of the model is determined, in order to obtain trustworthy results. Also, the best minimum covariance threshold is found, to ensure that the model converges and remains sensitive to variances in data. The model selection is based on the likelihood function evaluated over the validation set. The result is a model for the normal operation data which detects anomalies in the process by using loglikelihood function.

C. Testing the Model

A data set representing an operating condition as shown in Figure 4 is fed to the GMM model. This data set is composed by data of the three previously mentioned signals, between November 1st-4th. The whole data set has 345600 data points for each signal, which is equivalent to 4 days of operation.

From the data set, we calculate the loglikelihood score and normalize it by the number of samples in the window (set to 1000 secs). If the data set score is lower than the threshold ϵ, the algorithm detects an anomaly. The model is tested with different types of faults, being important to note that the user has to set the threshold according to the sensibility desired.

VI. Results and Discussion

A set of observation sequences belonging to a particular fault scenario are fed to the model. The results for both monitoring tasks and detection, are shown in the Figures 2 and 3. These figure show how the score decreases drastically when a fault is detected. The fault scenario corresponds to October 30th, and 17th of November. We set the threshold to 0.82 for normalized likelihood. Every one of the faults expected in every day, was detected by the algorithm. To prove the robustness of the algorithm, we use simulated data, for normal behavior data we use 100 tests and for faulty behavior we use 30 test. The results for false alarm was 6% and false rejection was 1%. This numbers change according to the threshold but these were the best found in the experiments.

VII. Conclusion

A fault detection system based on GMM was developed. Using the DAMADICS benchmark actuator system it was possible to verify its capabilites regarding monitoring tasks. With a model of normal operations was able to detect fault events immediately after its occurrence. This pattern recognition tool plays an important role when fault event data is scarce, and we have plenty of normal operation data. Besides, by using a temporal sequence of observations as model input, the model may be trained again if there are any changes in the behavior of the process, with the posibility of adding even more signals, which can be a daunting task with model based methods.

The results are very promising concerning to the application of GMM in chemical process monitoring. In the future, these results can be used along with decision trees to include a fault diagnosis stage, as a complement for the proposed system. It follows from the results obtained, that the proposed approach may have further applications in the monitoring of other industrial procesess as well, leading to prospective opportunities for research and development.

ISBN: 978-1-61804-240-8

In real world fault detection is important the aplicability of the solutions, in this article we aim to solve some issues that are common in data-based methods. Such as the lack of failure information, and the training of multiple models which can be very expensive and impractical. The results show that with only normal operation data is possible to learn a model that is capable of detect anomalies in multiple signals. This solution is very flexible, can be applied to many situations and only requires be trained once, this fact makes this solution ideal for real time applications. In the future this solution can be improved if is combined with Maximum Aposteriori adaptation and decision trees for diagnosis tasks.

REFERENCES

[1] Y. Guo, J. Wall, and J. L. andSam West, "A machine learning approach for fault detection in multi-variable systems," in *ATES in conjunction with Tenth Conference on Autonomous Agents and Multi-Agent Systems (AAMAS) AAMAS 2011*, 2011.

[2] J. P. Matsuura, A. Iees, P. Marechal, E. Gomes, T. Yoneyama, R. Kawakami, and H. Galvo, "Learning bayesian networks for fault detection," in *In Proceedings of the IEEE Signal Processing Society Workshop*, 2004, pp. 133–142.

[3] T. Chopra and J. Vajpai, "Classification of faults in damadics benchmark process control system using self organizing maps," *International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-1, Issue-3, July 2011*, vol. 1 (3), pp. 85–90, 2011.

[4] J. Sun, Y. Li, and C. Wen, "Fault diagnosis and detection based on combination with gaussian mixture models and variable reconstruction," in *Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on*, 2009, pp. 227–230.

[5] F. Nelwamondo and T. Marwala, "Faults detection using gaussian mixture models, mel-frequency cepstral coefficients and kurtosis," in *Systems, Man and Cybernetics, 2006. SMC '06. IEEE International Conference on*, vol. 1, 2006, pp. 290–295.

[6] A. Shui, W. Chen, P. Zhang, S. Hu, and X. Huang, "Review of fault diagnosis in control systems," in *Proceedings of the 21st annual international conference on Chinese control and decision conference*, ser. CCDC'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 5360–5369. [Online]. Available: http://dl.acm.org/citation.cfm?id=1714810.1715162

[7] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000, p. 2000.

[8] I. Nabney, *NETLAB: Algorithms for Pattern Recognition*, ser. Advances in Computer Vision and Pattern Recognition. Springer, 2002. [Online]. Available: http://books.google.com.co/books?id=LaAAJP1ZxBsC

[9] M. Syfert, R. Patton, M. Bartys, and J. Quevedo, "Development and application of methods for actuator diagnosis in industrial control systems (damadics): A benchmark study." *Proceedings of the IFAC Symposium Safe Process*, pp. 939–950, 2003.

Analytical and experimental modeling of the drivers spine

Veronica Argesanu, Raul Miklos Kulcsar, Ion Silviu Borozan, Mihaela Jula, Saša Ćuković, Eugen Bota

***Abstract*—**The aim of this study is to determine a analytical expression in the coronal plane of the drivers spine while driving along curved roads and also to determine ergonomic parameters for the car seat design. To determine the analytical expression and the ergonomic parameters, an experiment was developed to monitor the position variation in time of the vertebras in the coronal plane. The result lead to three sinusoidal equations. The amplitude values of the sinusoidal functions describing the variation in time of angles between the vertebras gives an image regarding the deformation degree of the intervertebral discs.

***Keywords*—**Spine, ergonomics, vehicle, musculoskeletal affections.

I. Introduction

THE possibility to drive in complete healthy and safety conditions not only for the professional drivers but also for the rest of the population which uses vehicles as frequent transportation means leads to efficiency by improving the quality of life.

In this context it is noted the following objectives and research directions: the development of modern mathematical models and principles to be included in a design or control algorithm.

The present study is based on the egronomical research regarding the spine's behavior while driving along curved roads.

II. Analytical Expression of the Spine in the Coronal Plane

The optimal ergonomic body posture of the driver sitting in the car seat is influenced by the structural characteristics of the seat. The body has to be constrained to the seat such way so that the spine's form is an ideal anatomical or ergonomic optimal shape. Therefore to design and construct the car seat, it is proposed to start from the ideal anatomical shape of the spine in the coronal plane (Fig. 1). [2, 3]

To determine the design parameters of the car seat is necessary to know the analytical form of the spine's shape in the coronal plane.

In the coronal plane, the shape of the spine can be expressed mathematically by the equation of a straight vertical line. Vertebrae centers are collinear. Considering a reference system as in figure 2, the vertical line's equation containing vertebras centers is considered to be x = 0.

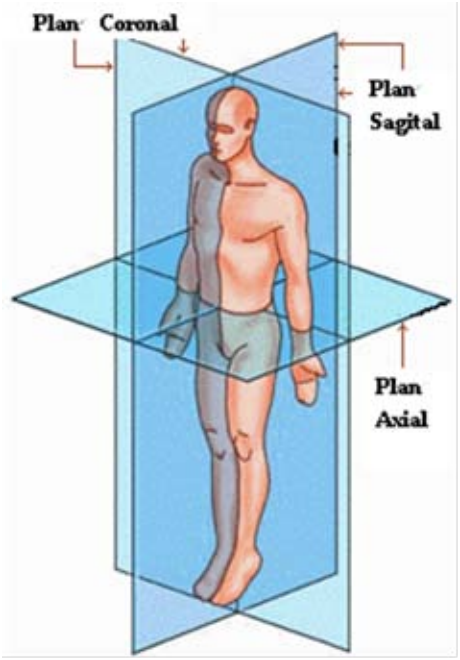


Fig. 1- Anatomical planes.

Point O, the origin of the coordinate system coincides with the lowest point of the coccyx.

The analytical expression x = 0 of the spine's shape in the coronal plane is only valid if the vehicle is at rest, or the vehicle travels on a rectilinear continuous road (unreal case).

Due to the centrifugal force acting on the human body while the vehicle is traveling along a, the human body changes its posture in the coronal plane in the opposite direction of the centrifugal force, to maintain the balance in the car seat. Thus the spine's shape changes depending on the vehicle's traveling speed and the curved path's radius, causing the spine shape mathematical expression in the coronal plane to be a motion law.

The spine shape is the line containing the centers of the vertebras. Anatomically, the shape and movement of the spinal column are shown by the relative rotational movement between the vertebras. According to anatomy and kinematic studies of the human spine, it is concluded that the center of rotation between two vertebras is the center of the intervertebral disc that connects the two vertebras. Thus

ISBN: 978-1-61804-240-8

intervertebral disc can be considered a ball joint with three degrees of freedom corresponding to rotation after three axes.

In figure 3 are shown as an example, L3 and L4 vertebras centers as CL3 and CL4 points, and the rotation centers of the L2, L3, L4 and L5 vertebras, as CrL2-L3-L4 and CrL4 CrL3-L5.

Fig. 2. – The spine in the coronal plane related to the coordinate system xOy.

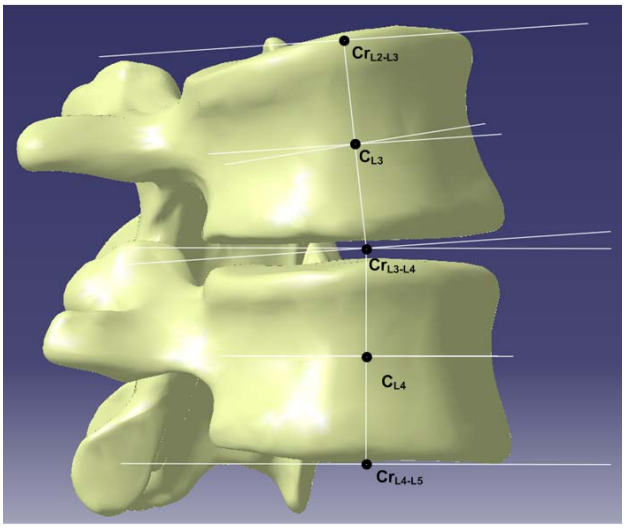


Fig. 3 - L3 and L4 vertebras centers (CL3 and CL4), and the rotation centers of the L2, L3, L4 and L5 vertebras (CrL2-L3-L4 and CrL4 CrL3-L5).

Considering the vertebras in the coronal plane as represented by segments connecting the rotation centers, the shape of the spine may be given by the angles α_i of these segments.

Figure 4 represents the lumbar segment in the coronal plane. The L1, L2 ... L5 vertebras are the CrT12-L1CrL2-L3-L3 CRL1-L2CrL2, CrL2-L3CrL3-L4-S1 ... CrL4-L5CrL5 segments. The relative rotation between two vertebras is given by the angle α_i between the segments representing the two vertebras.

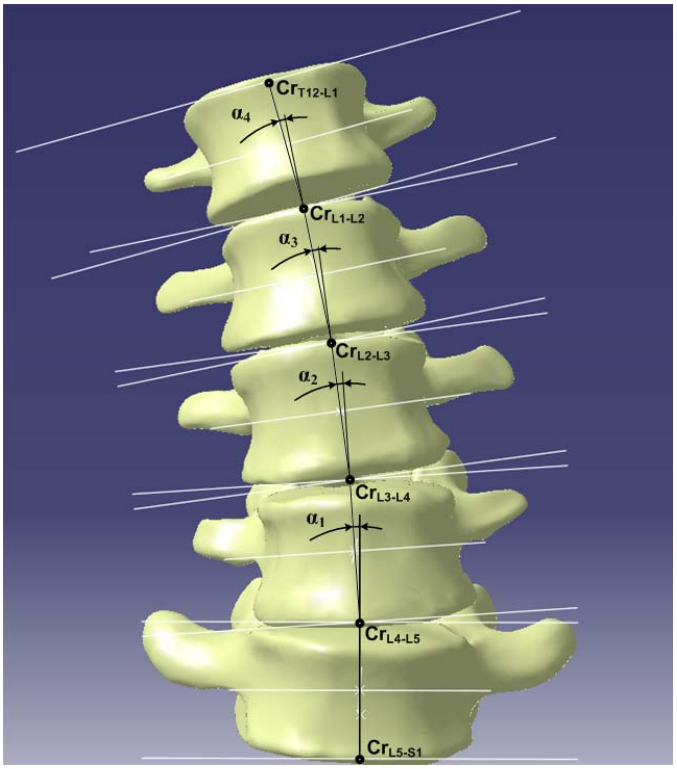


Fig. 4 – The lumbar spine with the segments representing L1, L2 ... L5 vertebras.

The motion law of the spine in the coronal plane can be expressed as a function of the vehicle speed (v_a), the curved trajectory radius (r_{tr}) and the upper body mass (m_{cs}), function that returns the values of the α_i angles.

$$f(v_a, r_{tr}, m_{cs}) \rightarrow \alpha_i \quad (1)$$

To determine the function given by relation (1), we created an experiment that for a given route and a constant driving speed, the upper body movements in the coronal plane were monitored.

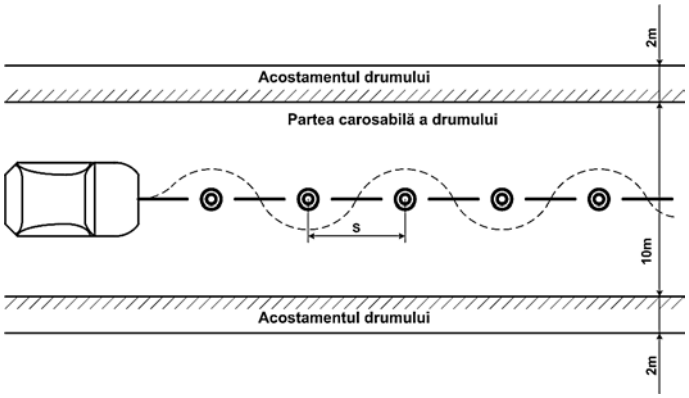


Fig. 5 - The route used in the experiment.

ISBN: 978-1-61804-240-8

The track used in the experiment is the same track used to determine the dynamic cornering ability of the vehicles (fig. 5).[29]

In the experiment we used the motion sensor manufactured by *PASCO scientific* and the PASCO CI-6400 Science Workshop 500 Interface (fig. 7).

The motion sensor MotionSensor II (fig. 6) operates on the sonar principle. The sensor can measure distances between 0.15m and 8m, between it and the object of interest (fig. 6).

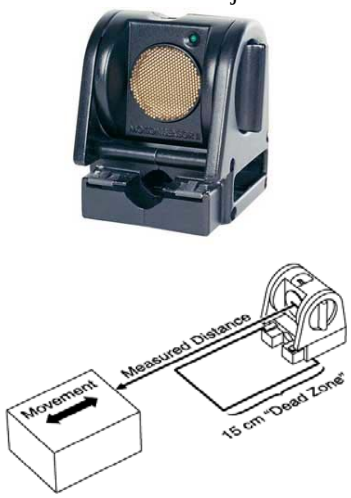


Fig. 6 – Motion Sensor II.

Fig. 7 - PASCO CI-6400 Science Workshop 500 Interface.

Before measurements, the motion sensor must be calibrated.

In this experiment the driver's upper body sideway movements in the coronal plane were monitored. To monitor the movements in the coronal plane the motion sensor was used to determine the positions in time of three points on the driver's body right side. In figure 8 is shown the positioning of the sensor. The first point is on the right side of the C1 vertebra, located at a distance of $d_C = 0.477$m from the sensor. The second point is placed on the right shoulder on the T4 vertebra's right side, located at a distance of $d_T = 0.39$m from the sensor. The third point is located next to the L1 vertebra located at a distance of $d_L = 0.419$m from the sensor.

In figure 9 is shown the sensor in the first position for determining the C1 vertebra movements. The r_C, r_T and r_L distances from the seat surface, were determined by anthropometric measurements of the driver's body in seated position. Thus $r_C = 0.8$m, $r_T = 0.585$m and $r_L = 0.4$m.

The experiment was carried out in three stages. In each stage the position in time of one of the three points is determined. In each stage the vehicle is traveling with a constant speed of 15km/h according to the vehicle dynamic steering ability tests. [5]

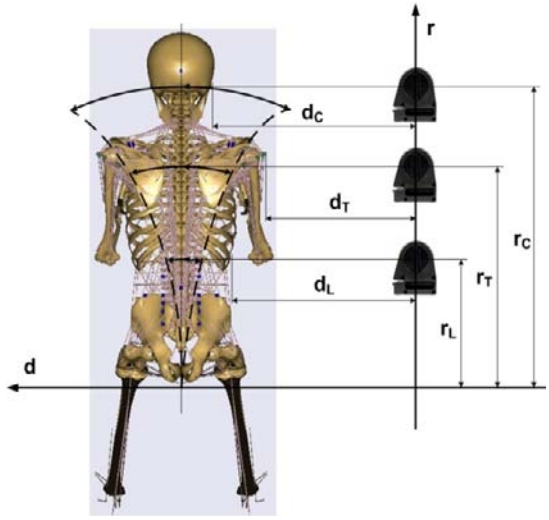


Fig. 8 - Points of interest for sensor positioning.

Fig. 9 - The sensor in the first position for determining the C1 vertebra movements.

III. The Experimental Results and Data Processing

The traveling time in one direction and performing a series of measurements, is about 15s.

Figures 10, 11 and 12 are presented graphically the results of series of measurements for the three points. At each step corresponding to a point were performed seven series of measurements.

For each point were averaged seven sets of measurements. Thus the results of processing experimental data are presented graphically in figure 13.

As a first analysis of the results obtained, it can be seen that the variation in time of the position of the three points can be expressed as a sinusoidal function with the same frequency but

different amplitudes.

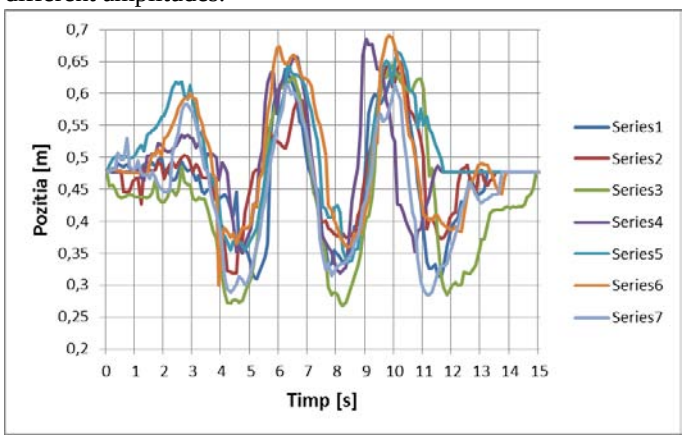


Fig. 10 - The series of measurements for the C1 vertebra.

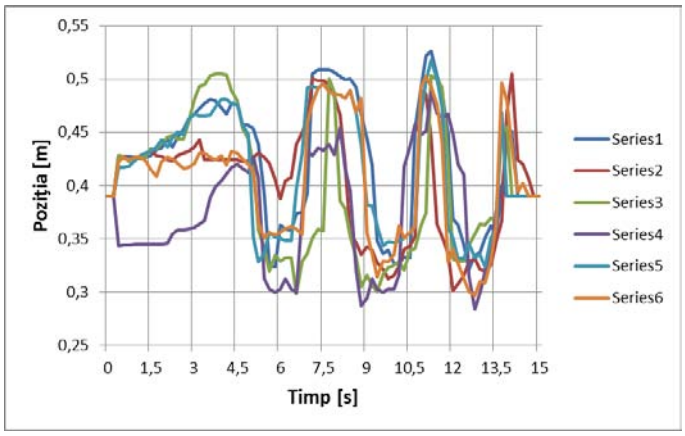


Fig. 11 - The series of measurements the T4 vertebra.

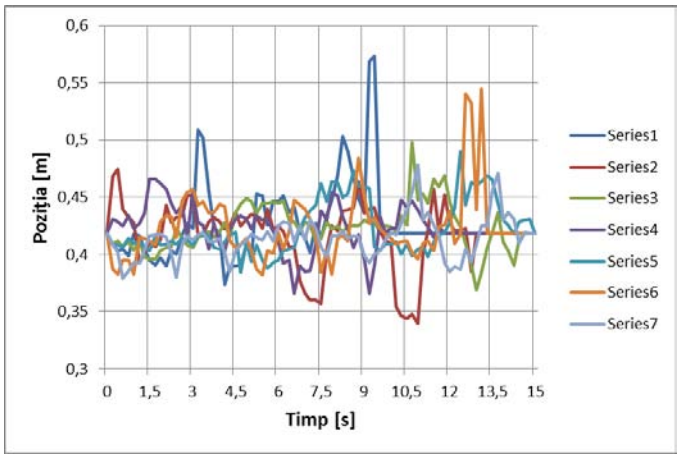


Fig. 12 - The series of measurements for the L1 vertebra.

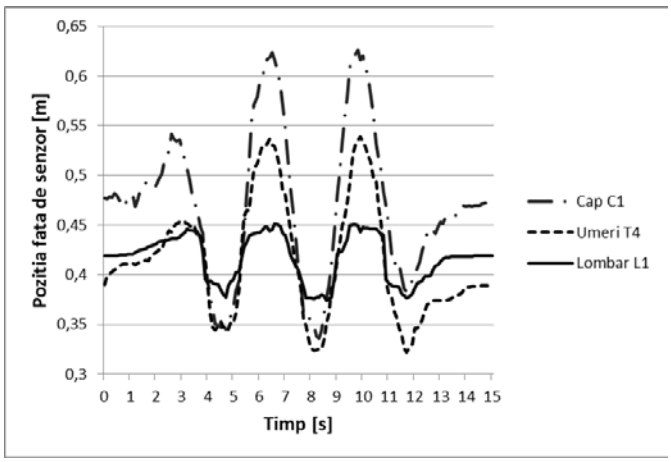


Fig. 13 – Graphical representation of the positions in time of the three points.

IV. Determination of the Sinusoidal Functions Describing the Variation in Time of the C1, T4 and L1 Vertebras Positions

Using the Mathcad software the position in time values for the three points were introduced as the following strings:

cp :=

	0
0	0.477
1	0.476
2	0.479
3	0.478
4	0.482
5	0.479
6	0.474
7	0.48
8	0.473
9	0.473
10	0.475
11	0.48
12	0.469
13	0.476
14	0.485
15	...

um :=

	0
0	0.39
1	0.399
2	0.401
3	0.403
4	0.405
5	0.407
6	0.409
7	0.411
8	0.411
9	0.411
10	0.411
11	0.411
12	0.41
13	0.412
14	0.413
15	...

lb :=

	0
0	0.418
1	0.418
2	0.418
3	0.419
4	0.419
5	0.419
6	0.419
7	0.42
8	0.42
9	0.42
10	0.42
11	0.421
12	0.422
13	0.424
14	0.425
15	...

The *cp* string corresponds to the C1 point, *um* string corresponds to the T4 point and *lb* string corresponds to the L1 point.

The next step is to determine the frequency of each string.

$$tcpmax_i := \begin{cases} i & \text{if } \max(cp) = cp_i \\ 0 & \text{otherwise} \end{cases}$$

$$\max(tcpmax) = 98$$

ISBN: 978-1-61804-240-8

$$tcpmin_i := \begin{cases} i & \text{if } min(cp) = cp_i \\ 0 & \text{otherwise} \end{cases}$$

$$max(tcpmin) = 83$$

$$tummax_i := \begin{cases} i & \text{if } max(um) = um_i \\ 0 & \text{otherwise} \end{cases}$$

$$max(tummax) = 99$$

$$tummin_i := \begin{cases} i & \text{if } min(um) = um_i \\ 0 & \text{otherwise} \end{cases}$$

$$max(tummin) = 117$$

$$tlbmax_i := \begin{cases} i & \text{if } max(lb) = lb_i \\ 0 & \text{otherwise} \end{cases}$$

$$max(tlbmax) = 70$$

$$tlbmin_i := \begin{cases} i & \text{if } min(lb) = lb_i \\ 0 & \text{otherwise} \end{cases}$$

$$max(tlbmin) = 86$$

The time interval between the maximum and minimum for each string is determined:

$$\Delta tcp = \frac{|max(tcpmax) - max(tcpmin)|}{10} = 1.5s \quad (2)$$

$$\Delta tum = \frac{|max(tummax) - max(tummin)|}{10} = 1.8s \quad (3)$$

$$\Delta tlb = \frac{|max(tlbmax) - max(tlbmin)|}{10} = 1.6s \quad (4)$$

In order to determine the single frequency in all three strings, the average of the three time periods is determined:

$$\Delta t = \frac{\Delta tcp + \Delta tum + \Delta tlb}{3} = 1.6333s \quad (5)$$

Thus, the frequency will be:

$$f_{\Delta t} = \frac{1}{\Delta t} \quad (6)$$

The amplitude of each string is determined as follows:

$$acp = \frac{max(cp) + min(cp)}{2} \quad (7)$$

$$aum = \frac{max(um) + min(um)}{2} \quad (8)$$

$$alb = \frac{max(lb) + min(lb)}{2} \quad (9)$$

The *cp* string amplitude is noted *acp*, the *um* string amplitude is noted with *aum*, and the amplitude of the *lb* string is noted *alb*.

The sinusoidal functions describing the position variation in time of the C1, T4 and L1 vertebraes points are the following:

$$ycp_i = cp_0 + acp \cdot \cos(f_{\Delta t} \cdot i \cdot \pi) \quad (10)$$

$$yum_i = um_0 + aum \cdot \cos(f_{\Delta t} \cdot i \cdot \pi) \quad (11)$$

$$ylb_i = lb_0 + alb \cdot \cos(f_{\Delta t} \cdot i \cdot \pi) \quad (12)$$

In the figures 14, 15 and 16 the sinusoidal functions are represented in comparison to the *cp*, *um* and *lb* strings graphic form. For each case can be seen that the sinusoidal functions allure is very close to the allure of the strings measured values.

In conclusion it can be considered that these sinusoidal functions can describe the position variation in time of the C1, T4 and L1 vertebra's points, while driving on a sinusoidal trajectory.

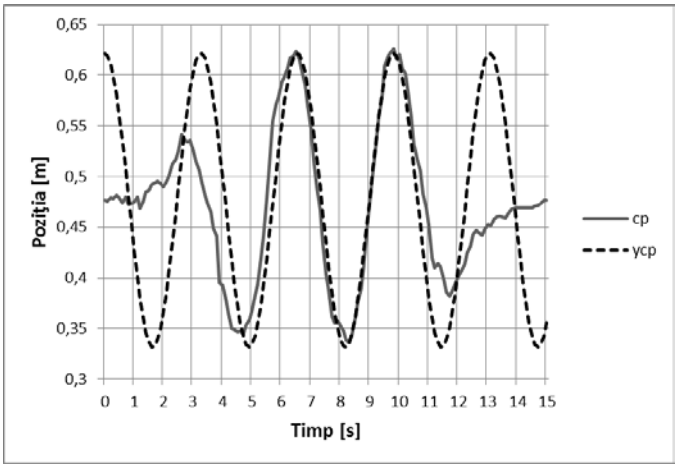


Fig. 14 - Graphical representation of the *ycp* sinusoidal function compared with the *cp* string.

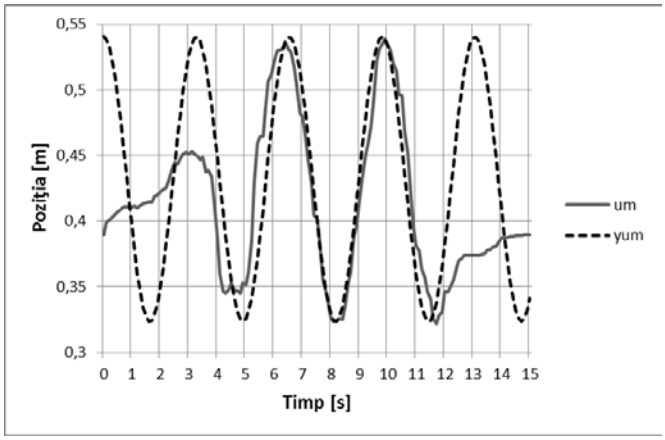


Fig. 15 - Graphical representation of the *yum* sinusoidal function compared with the *um* string.

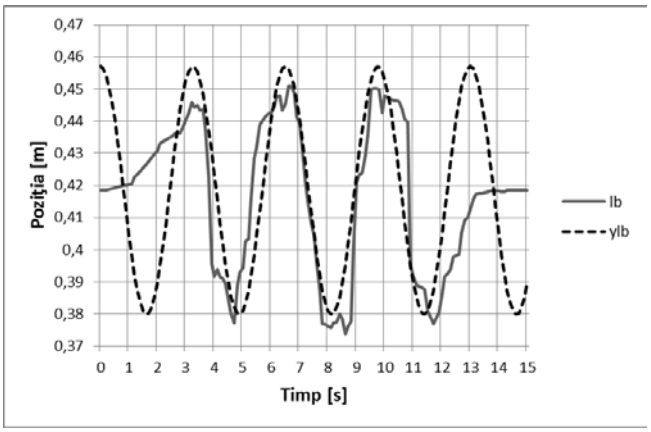


Fig. 16 - Graphical representation of the *ylb* sinusoidal function compared with the *lb* string.

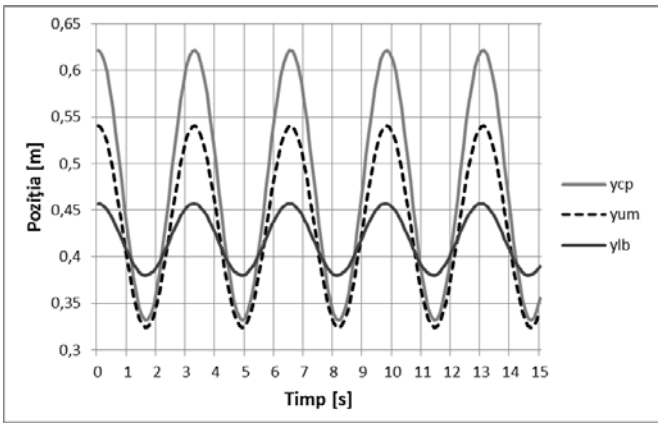


Fig. 17 – The sinusoidal functions describing the position variation in time of the C1, T4 and L1 vertebras points.

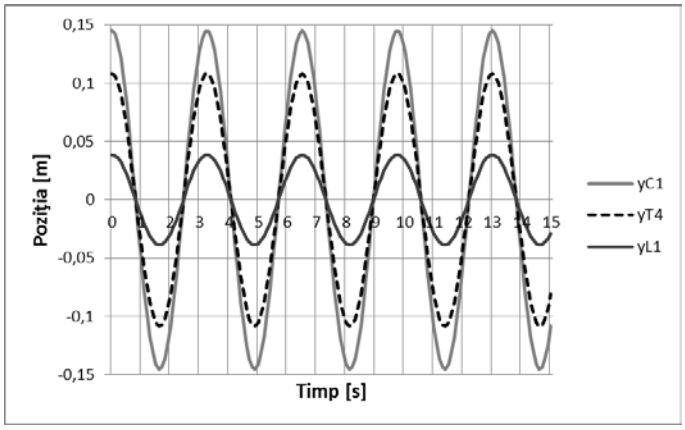


Fig. 18 – The sinusoidal functions describing the position variation in time of the C1, T4 and L1 vertebras.

The sinusoidal functions describing the position variation in time of the C1, T4 and L1 vertebras in the coronal plane are:

$$yC1_i = acp \cdot \cos(f_{\Delta t} \cdot i \cdot \pi) \quad (13)$$

$$yT4_i = aum \cdot \cos(f_{\Delta t} \cdot i \cdot \pi) \quad (14)$$

$$yL1_i = alb \cdot \cos(f_{\Delta t} \cdot i \cdot \pi) \quad (15)$$

V. Conclusions

The amplitude values of the sinusoidal functions describing the variation in time of angles between the vertebras gives an image regarding the deformation degree of the intervertebral discs.

A nonergonomic posture of the driver's body seated in the vehicle's seat implies the spine to be in a shape that subjects the intervertebral discs to uneven tensions causing deformations that in some cases can exceed the limits at which the musculoskeletal affections of the spine can be avoided or treated by physiotherapy.

References

[1] Adams M. ; Bogduk N. ; Burton K. ; Dolan P. (2006). The Biomechanics of Back Pain Second Edition, Churchill Livingstone Elsevier.

[2] Borozan I. S.; Maniu I.; Kulcsar R. M.; „Ergonomic analysis on driving an Automatic Gearbox equipped vehicle", SACI 2012 IEEE 7th International Symposium on Applied Computational Intelligence and Informatics, May 24-26, 2012, Timisoara, Romania.

[3] Borozan I. S.; Kulcsar R. M.**;** "Vertebral column bioengineering analysis at bending and torsion", International Conference on Human-Machine Systems, Cyborgs and Enhancing Devices HUMASCEND, Iasi, Romania, June 14-17, 2012.

[4] Goran Devedžić, Saša Ćuković, Vanja Luković, Danijela Milošević, K. Subburaj, Tanja Luković, "ScolioMedIS: web-oriented information system for idiopathic scoliosis visualization and monitoring", Journal of Computer Methods and Programs in Biomedicine, Vol.108, No.-, pp. 736-749, ISSN -, Doi 10.1016/j.cmpb.2012.04.008, 2012.

[5] Hilohi C., Untaru M., Soare I., Druța Gh., „Metode si mijloace de incercare a automobilelor", Editura Tehnică Bucure□ti, 1982.

[6] Hinza B., Seidel H., „The significance of using anthropometric parameters and postures of European drivers as a database for finite-

element models when calculating spinal forces during whole-body vibration exposure", International Journal of Industrial Ergonomics, Elsevier, 2008.

[7] Kolich M., „A conceptual framework proposed to formalize the scientific investigation of automobile seat comfort", Applied Ergonomics, Elsevier, 2008.

[8] Kulcsar R. M.; Madaras L.; "Ergonomical study regarding the effects of the inertia and centrifugal forces on the driver", MTM & Robotics 2012, The Joint International Conference of the XI International Conference on Mechanisms and Mechanical Transmissions (MTM) and the International Conference on Robotics (Robotics'12), Clermont-Ferrand, France, June 6-8, 2012 Applied Mechanics and Materials, Vol. 162, Mechanisms, Mechanical Transmissions and Robotics, ISBN-13:978-3-03785-395-5, pp. 84-91.

[9] Muksian R., Nash C.D.Jr. "A model for the response of seated humans to sinusoidal displacements of the seat", J. Biomechanics, vol.7, pp 209-215, Pergamon Press, 1974.

[10] Tae-Yun Koo1, Kee-Jun Park, „A Study on Driver's Workload of Telematics Using a Driving Simulator: A Comparison among Information Modalities", International journal of precision engineering and manufacturing vol. 10, no. 3, pp. 59-63, 2009.

Exponentially scaled point processes and data classification

Marcel Jiřina

***Abstract*—We use a measure for distances of neighbors' of a given point that is based on l_p metrics and a scaling exponent. We show that if the measure scales with scaling exponent mentioned, then distribution function of this measure converges to Erlang distribution. The scaling of distances is used for design of a classifier. Three variants of classifier are described. The local approach uses local value of scaling exponent. The global method uses the correlation dimension as the scaling exponent. In the IINC method indexes of neighbors of the query point are essential. Results of some experiments are shown and open problems of classification with scaling are dicussed.**

***Keywords*—Multivariate data, nearest neighbor, Erlang distribution, multifractal, scaling exponent, classification, IINC.**

I. Introduction

In this paper we use a model that there is some underlying process and in the process some events occur. We suppose that events occur randomly and independently one of another. The only information we have is d-dimensional data arising from events, i.e. by (often rather approximate) measurement or sampling.

Important notion is a scaling characterized by scaling exponent denoted also as fractal dimension. This dimension q is lesser than space dimension d, and usually is not an integer. The space dimension d is often called embedding dimension using concept that fractal is a q-dimensional formation plunged into larger d-dimensional space [16]. This concept can be applied to volume V of a ball of radius r. There is $V = c_q r^q$ for q-dimensional ball in d-dimensional space; c_q is a constant dependent on q and metrics used. Usually $q = d$ but the same holds for integer $q < d$, e.g. two dimensional circle in three dimensional Euclidean space. Keeping the concept consistent, q need not be an integer but there is no intuition how, say, 2.57-dimensional ball looks like.

The goal of this study is to analyze the distances of nearest neighbors from given point (location) in a multidimensional spatial point process in R^d with exponential scaling [5]. The result is that when using scaled measure for distance of the k-th neighbor, the distance can have the Erlang distribution of order k. We show here that scaling leads to simple polynomial transformation $z = r^q$. With the use of this transformation a classifier can be designed.

The work was supported by Ministry of Education of the Czech Republic under INGO project No. LG 12020.

M. Jiřina is with the Institute of Computer Science AS CR, Pod Vodarenskou vezi 2, 182 07 Praha 8, Czech Republic (e-mail: marcel@cs.cas.cz)

II. Multidimensional point processes and fractal behavior

A. Point processes

Let there be an „underlying process" U_P. This process is sampled randomly and independently so that random d-dimensional data

$$P = x_1, x_2, \ldots, x_i \in X \subset R^d \tag{1}$$

arose. This data (without respect to time or order in which individual samples x_i was taken) forms spatial point process in R^d and individual samples x_i are called points, in applications often events [6], samples, patterns or so.

We are interested in distances from one selected fixed point x to others; especially distance to the k-th nearest neighbor. From now we use numbering of points according to their order as neighbors of point x; x_k being the k-th nearest neighbor of point x. To distance l_k from x to its k-th nearest neighbor a probability is assigned. There is introduced

$$S_k(l) = Pr\{l < l_k\} = Pr\{N(l_k) < k\}$$

i.e. probability that a distance to the k-th nearest neighbor is larger than l that is equal to probability of finding k-1 points within distance l_k [4]. For k = 1 it is called avoidance probability and often denoted P_0. Function

$$F_k(l) = 1 - S_k(l)$$

is the distribution function of distance l to the k-th neighbor.

A scaling function is a real-valued function $c : R^d \rightarrow R_+$, that satisfies a self-similarity property with respect to a group of affine transformations [20]. There are several types of scaling functions, shifting, scaling, eventually reflections. General equation for scaling can have form

$$\mu(\vec{x} + \vec{a}) = c_\theta(\vec{x})$$

and in less general (fractal) case of exponential scaling

$$\mu(\vec{x} + \vec{a}) = a^{h(\vec{x})}$$

Here θ is l-dimensional parameter vector. When the scaling is location dependent, we speak about locally dependent point process.

ISBN: 978-1-61804-240-8

B. Fractal behavior

We admitt that an „underlying process“ U_P shows exponentially scaled characteristics. Let there be data in R^d, see (1).

One can introduce a distance between two points of P using l_p metrics, l_{ij} = $||x_i$ - $x_j||_p$, $x_i, x_j \in P$. In a bounded region $W \in R^d$ a cumulative distribution function of l_{ij}

$$C_I(l) = \lim_{N\to\infty} \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N-1} h(l - l_{ij}) \quad ,$$

is denoted as correlation integral; $h(.)$ is the Heaviside step function. Grassberger and Procaccia [10] introduced correlation dimension ν as limit

$$\nu = \lim_{l\to 0} \frac{C_I(l)}{l} \quad .$$

Having empirical data on P, distances between any two points of P is the only information yelded exactly with the use a relatively simple computation.

It is apparent that scaling of distances between any two points of P also holds for near neighbors' distances distribution. Let $F_k(l)$ be the distribution function of distance from some point x to the k-th neighbor. Let us define another function, the function $D(x,\ l)$ of neighbors' distances from one particular point x as follows [13], [14].

Definition

Probability distribution mapping function $D(x,\ l)$ of the neighborhood of the query point x is function $D(x,l) = \int_{B(x,l)} p(z)dz$, where l is the distance from the query point and $B(x,\ l)$ is the ball with center x and radius l.

In bounded region $W \subset P$ when using a proper rescaling, the DMF is, in fact, a cumulative distribution function of distances from given location $x \in W \subset P$ to all other points of P in W. We call it also near neighbors' distance distribution function. We use $D(x,\ r)$ mostly in this sense. It is easily seen that DMF can be written in form

$$D(x,l) = \lim_{N\to\infty} \frac{1}{N-1} \sum_{j=1}^{N-1} h(l - l_j).$$

The correlation integral can be decomposed into set of DMFs each corresponding to particular point $x_{0i} \in W \subset P$ as follows [14]

$$C_I(r) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{N-1} \sum_{j=1}^{N-1} h(r - l_{ij}) \right)$$

that means

$$C_I(l) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} D(x_{0i}, l) \quad .$$

Thus the correlation integral is a mean of probability distribution mapping functions for all points of $W \subset P$.

We introduce a local scaling exponent q according to the following definition.

Definition

Let there be a positive q such that $\frac{D(x,l)}{l^q} \to const$ for $l \to 0+$.

We call function

$$z(l) = l^q$$

a power approximation of the probability distribution mapping function and q is a distribution mapping exponent.

C. Common interesting behavior

It is common that measure $l(A)$ on R^d is usually a Lebesgue measure or based on it. Thus $l(A)$ depends on integer dimensionality d. Our intention is to deal with some q, $d \geq q > 0$ not necessary an integer.

Here we contract metric space $(X,\ \rho)$ to $(R^d,\ l_p)$, where l_p is Lebesgue p-norm. Let $q \in (0,\ d]$. We define measure $\mu(.)$ of neighbors distances so that for S = (a line between x_i and x_j) there is $\mu(S) = l_p^q(x_i - x) - l_p^q(x_j - x)$, $l_p(x_i - x) \geq l_p(x_j - x)$, $\mu(O) = 0$, $\mu(S_1 \cup S_2) = \mu(S_1) + \mu(S_2)$; $S_1 \cap S_2 = O$ $a.s.$

It is easily seen that $\mu(.)$ is a measure; it is nonnegative, it equals to zero for the empty set and for x_i = x_j , and is countable additive.

Then it holds a theorem that is a special but useful case of more general results about point processes [4], [5], [20].

Theorem 1: Let there be a point process P and bounded region $x \subset R^d$, where there is given point x and Nnearest neighbors of x. Let $D(x, l)$ scales with exponent q. Let process P in bounded region $W \subset R^d$ be mapped (by mapping M_{Ppx}) to process p in bounded interval $w \subset R+$. Then one-dimensional point process p in $w \subset R+$ is a homogenous Poisson process with intensity $\lambda = \lim_{N\to\infty} N/z_N$.

Proof: It is omitted here.

Theorem 1 can be applied to all points $x_0 \in P$. Supposing monofractal underlying process U_P and by point process P induced measure $\mu_{p\nu}(.)$ with correlation dimension ν as one if its parameters, the ν scales also the DMF of all points of P and then q = ν.

Corollary 1: Let there be a point process P and bounded region W, where there are given location x and N nearest neighbors of x. Let DMF $D(x,\ l)$ scales with exponent q. Then probability distribution of μ_k = $\mu_{pq}(x_k - x)$ of the k-th nearest neighbor x_k of given location x is Erlang distribution Erl($\mu_k,\ k,\ \lambda$), i.e.

$$F(\mu_k) = 1 - \exp(-\lambda\mu_k) \sum_{j=0}^{k-1} \frac{(\lambda\mu_k)^j}{j!}$$

$$f(\mu_k) = \frac{\lambda^k}{k!} (\mu_k)^{k-1} \exp(-\lambda\mu_k) \quad .$$

Proof: It is omitted here.

We found that when one can find a scaling of neighbors' distances measure, in form $z = r^q$, q is the distribution mapping exponent, then one can find a “Poisson process-like” behavior, i.e. Erlang distribution of neighbors' distances measure. Usually, a measure is considered that may depend

on the embedding space dimension d (integer), while we use more general distribution mapping exponent q that is a positive real number.

III. Classification using scaling

Here we show basic idea of multidimensional data classification using scaling and three variants of this approach.

A. Data

Let the learning set U of total N samples be given. Each sample $x_t = \{x_{t1}, x_{t2}, x_{td}\}; t = 1, 2, ...N, x_{tk} \in R; k = 1, 2, ..., d$ corresponds to a point in d-dimensional metric space M_d, where d is the sample space dimension. For each $x_t \in U$ a class function $T : R^d \rightarrow \{1, 2, ...C\} : T(x_t) = c$ is introduced. With the class function the learning set U is decomposed into disjoint classes $U_c = \{x_t \in U | T(x_t) = c\}; c \in \{1, 2, ..., C\}, \bigcup_{c=1}^{C} U_c, \cap U_b = \emptyset; c, b \in 1, 2, ..., C; c \neq b$. Cardinality of set U_c let be N_c. As we need to express which sample is closer or further from some given point x, we can rank points of the learning set according to distance r_i of point x_i from point x. Therefore, let points of U be indexed (ranked) so that for any two points $x_i, x_j \in U$ there is $i < j$ if $r_i < r_j; i, j = 1, 2, ...N$, and class $U_c = \{x_i \in U | T(x_i) = c\}$. Of course, the ranking depends on point x and eventually metrics of M_d. We use Euclidean (L_2) and absolute (Manhattan, L_1) metrics here. In the following indexing by i means ranking just introduced.

B. The DME method

This classifier uses the distribution mapping exponent already introduced.

1) Intuitive explanation: Let us consider the partial influences of the individual points to the probability that point x is of class c. Each point of class c in the neighborhood of point x adds a little to the probability that point x is of class c, where $c = 0, 1$ is the class mark. Suppose that this contribution is the larger the closer the point considered is to point x and vice versa. Let $p(c|x, i)$ be a partial contribution of the i-th nearest point to the probability that point x is of class c. Then:

For the first (nearest) point i = 1 and ther is $p(c|x, 1) \simeq \frac{1}{S_q r_1^q}$, where we use the distribution mapping exponent q instead of the data space dimensionality d; S_q is proportionality constant dependent on the dimensionality and metrics used. For the second point i = 2 there is $p(c|x, 2) \simeq \frac{1}{S_q r_2^q}$ ... And so on; generally for point No. i $p(c|x, i) \simeq \frac{1}{S_q r_i^q}$.
We add the partial contributions of individual points together by summing up

$$p(c|x) \simeq \sum_{x_i \in U_c} p(c|x, i) = \frac{1}{S_q} \sum_{x_i \in U_c} 1/r_i^q$$

(The sum goes over the indexes i for which the corresponding samples of the learning set are of class c). For both classes there is $p(0|x) + p(1|x) = 1$ and from it $S_q = \sum_{i=1}^{N} 1/r_i^q$ Thus we get the form suitable for practical computation

$$\hat{p}(c|x) = \frac{\sum_{x_i \in U_c} 1/r_i^q}{\sum_{i=1}^{N} 1/r_i^q} \tag{2}$$

(The upper sum goes over the indexes i for which the corresponding samples of the learning set are of class c). At the same time all N points of the learning set are used instead of some finite number as in the k-NN method. Moreover, we do not use the nearest point ($i = 1$) usually. It can be found that its influence is more negative than positive on the probability estimate here.

2) Theory: Here we come from an assumption that the best approximation of the probability distribution of the data is closely related to the uniformity of the data space around the query point x. In cases of uniform distribution - at least in the neighborhood of the query point - the best results are usually obtained. Therefore we approximate (polynomially expand) the true distribution so that at least in the neighborhood of the query point the distribution density function appears to be constant.

Now a question arises why influences of individual points of a given class to the final probability that point x is of the class are inversely proportional to the $z = r_i^q$. Let there be Z, the largest of all z for a given class. We have shown that variable $z = r^q$ has uniform distribution with some density p_z. It holds $Zp_z = 1$ because the integral of the distribution density function over its support $(0, Z)$ equals to one. If support would be $(0, Z_1), Z_1 < Z$, then the density must be larger proportionally to Z/Z_1. It means that shift of each point closer to point x will enlarge the density so that it will be inversely proportional to the distance of a point from point x.

Theorem 2: Let the task of classification into two classes be a given. Let the size of the learning set be N and let both classes have the same number of samples. Let q , $1 < q < d$ be the distribution mapping exponent, let i be the index of the i-th nearest neighbor of point x (without respect to class), and $r_i > 0$ its distance from point x. Then

$$p(c|x) = \lim_{N \rightarrow \infty} \frac{\sum_{x_i \in U_c} 1/r_i^q}{\sum_{i=1}^{N} 1/r_i^q} \tag{3}$$

(the upper sum goes for all points of class c only) is probability that point x belongs to class c.

Proof: can be found in [13]

3) Generalization: Up to now we suposed two classes only and the same number of samples of both classes in the learning set. For general number of C classes and of the different number of the samples $N_1, N_2, ...N_C$ of individual classes formula (3) must be completed. In fact, the last is only a recalculation of the relative representation of the different number of the samples in classes.

ISBN: 978-1-61804-240-8

$$p(c|x) = \frac{\lim\limits_{N\to\infty}(1/N_c \sum\limits_{x_i\in U_c} 1/r_i^q)}{\sum\limits_{k=1}^{C} \lim\limits_{N\to\infty}(1/N_k \sum\limits_{x_i\in U_k} 1/r_i^q)} \quad (4)$$

4) The DME classifier construction: This method represents a direct use of formula (2) eventually formula (4) in form

$$\hat{p}(c|x) = \frac{1/N_c \sum\limits_{x_i\in U_c} 1/r_i^q}{\sum\limits_{k=1}^{C}(1/N_k \sum\limits_{x_i\in U_k} 1/r_i^q)} \quad (5)$$

Note that the convergence of sums above is faster the larger DME q is. Usually, for multivariate real-life data the DME is also large (and the correlation dimension as well). Figs. 1 and 2 illustrate the convergence of the sum in the numerator above for one query point for the well-known "vote" data, see [1]. The task is to find whether a president elected will be republican or democrat. The data is 15-dimensional of two classes that have a different number of samples. In the learning set, there are 116 times republican and 184 times democrat. The distribution mapping exponent q varies between 4.52 and 14 with the mean value 10.22.

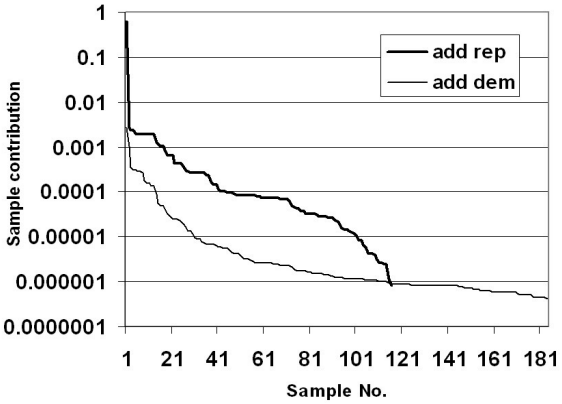


Fig. 1. Sample contribution to the sum in the numerator of (5) for the 15 dimensional data vote and one particular query point; $q = 7.22$. The upper line corresponds to the republican, the lower line to the democrat. Samples are sorted according to the distance r, i.e. also to the size of the sample contribution to the sum for one class. There are different numbers of samples of one and the other class in the learning set.

The classification procedure is rather straightforward. First, compute the distribution mapping exponent q for the query point x by standard linear regression, see the next section. Then, we simply sum up all the components excluding the nearest point.

In our approach, a true distribution is mapped to the uniform distribution. For uniform distribution, it holds that the i-th neighbor distance from a given point has an Erlang distribution of i-th order. For an Erlang distribution of i-th order, the relative statistical deviation, i.e. the statistical deviation divided by the mean, is equal to $1/\sqrt{i}$. Then the relative statistical

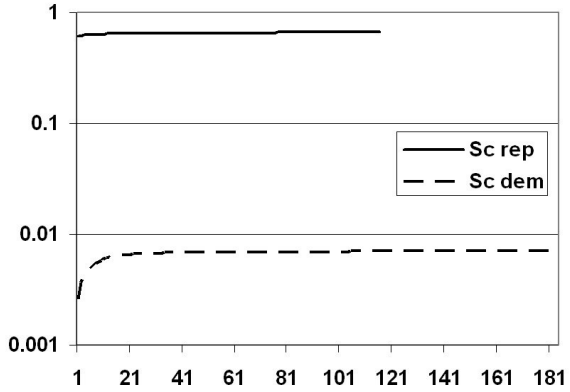


Fig. 2. The size of the total sum in the numerator of (5) for the 15-dimensional data "vote" and one particular query point; q = 7.22. The upper line corresponds to the republican, the lower line to the democrat. The samples are sorted according to the distance r, i.e. also to the size of the sample contribution to the sum for one class.

deviation diminishes with the index of the neighbor and for the nearest neighbor is equal to 1 which also follows from the fact that Erlang(1) distribution is exponential distribution. So, there is a large relative spread in the positions of the nearest neighbor and, at the same time, its influence is the largest. In practice, it appears better to eliminate the influence of the first nearest neighbor. Theorems for DME as well for CD method remains valid.

This is made for classes, simultaneously getting C sums for all classes. Then we can get the Bayes ratio or a probability estimate that point x belongs to class. The class that has largest probability estimate is taken as an estimated class of query point x. Eventually these probabilities can be weighted in the same way as in other classifiers.

5) Distribution mapping exponent estimation: Important issue of this method is a procedure how to determine the distribution mapping exponent.

To estimate the distribution mapping exponent q we use a similar approach, nearly identical, to the approach of Grassberger and Procaccia [10] for the correlation dimension estimation.

This is a task of estimating the slope of a straight line linearly approximating the graph of the dependence of the neighbor's index as a function of distance in log-log scale. Grassberger and Procaccia [10] proposed a solution by linear regression. Dvorak and Klaschka [7], Guerrero and Smith [11], Osborne and Provenzale [18] later proposed different modifications and heuristics. Many of these approaches and heuristics can be used for the distribution mapping exponent estimation, e.g. use of the square root of N_c nearest neighbors instead of N_c to eliminate the influence of a limited number of the points of the learning set. The accuracy of the distribution mapping exponent estimation is the same problem as the accuracy of the correlation dimension estimation. On the other hand, one can find that a small change of q does not essentially influence the classification results.

The approach described here has two other variants.

C. CD method - correlation dimension based approach

In this method it is supposed that distribution mapping exponents for individual query points differ only slightly and that one can use the value of correlation dimension ν instead. Computation has then two steps, in the first step the estimate of correlation dimension ν is computed using any known suitable method and then one uses formulas (2) or (5) where ν instead of q is used.

Again, as in Section III-B we exclude the first nearest neighbor of the query point. The convergence of sums is equally fast as in the DME method.

A relative advantage of this approach is that estimate of the correlation dimension is more exact than estimate of the distribution mapping exponent and that computation of the correlation dimension is done once only in difference of the DME that must be computed for each query point anew.

1) Correlation dimension estimation: For the approximation of probability of class at a given point and classification described above, a fast and reliable method for correlation dimension estimation is needed. Methods for the estimation of correlation dimension differ by approaches used and also by some kind of heuristics that usually optimize the size of radius r to get a realistic estimation of correlation dimension [17], [3], [25] as mentioned above.

Averaging method

The basic problem of correlation dimension estimation is the large number of pairs that arise even for a moderate learning set size. The idea of the correlation dimension estimation described below is based on the observation that distances between all pairs of points can be divided into groups, each group associated with one (fixed) point of the learning set.

Theorem 3: Let there be a learning set of N points (samples). Let the correlation integral be $C_I(r)$ and let $D(x,r)$ be the distribution mapping function corresponding to point x. Then, $C_I(r)$ is a mean of $D(x,r)$ for all points of U

Proof: For proof see [15].

We have found that for sufficiently good estimation of the correlation dimension one can use part of the data set only, for each point to estimate the distribution mapping exponent, and take the average. The part of the data set may be some number of points randomly selected from the data set. It suffice to use 100 points. The method of averaging need not be limited to the Grassberger-Procaccia algorithm. We use it analogically for Takens' algorithm [25] as well.

D. IINC method - the inverted indexes of neighbors classifier

1) Intuitive basis: Similar way as in Section III-B1 let us assume that the influence on the probability that point x is of class c of the nearest neighbor of class c is 1, the influence of the second nearest neighbor is 1/2, the influence of the third nearest neighbor is 1/3 etc. Again we add the partial influences of individual points together by summing up

$$\hat{p}(c|x) = \sum_{x_i \in U_c} p_1(c|x, r_i) = K \sum_{x_i \in U_c} 1/i.$$

The sum goes over indexes i for which the corresponding samples of the learning set are of class c. The estimation of the probability that the query point x belongs to class c is

$$\hat{p}(c|x) = \frac{\sum\limits_{x_i \in U_c} 1/i}{\sum\limits_{i=1}^{N} 1/i}.$$

In the denominator is the so-called harmonic number H_N, the sum of truncated harmonic series. The hypothesis above is equivalent to the assumption that the influence of individual points of the learning set is governed by Zipfian distribution (Zipf's law) [27], [23]. There is an interesting fact that the use of $1/i$ has a close connection to the correlation integral and correlation dimension and thus to the dynamics and true data dimensionality of processes that generate the data we wish to separate.

2) Theory:

Theorem 4: Let the task of classification into two classes be given. Let the size of the learning set be N and let both classes have the same number of samples. Let i be the index of the i-th nearest neighbor of point x (without considering the neighbor's class) and r_i be its distance from point x. Then

$$p(c|x) = \lim_{N \to \infty} \frac{\sum\limits_{x_i \in U_c} 1/i}{\sum\limits_{i=1}^{N} 1/i} \tag{6}$$

(the upper sum goes over indexes i for which the corresponding samples are of class c) is the probability that point x belongs to class c.

Proof: For proof see [15].

In the formula above it is seen that the approach is, in the end a kernel approach with rather strange kernel in difference to kernels usually used [12], [22].

It is easily seen that

$$\sum_{c=1}^{C} p(c|x) = \sum_{c=1}^{C} \lim_{N \to \infty} \frac{\sum\limits_{x_i \in U_c} 1/i}{H_N} = 1$$

and $p(c|x)$ is a "sum of relative frequencies of occurrence" of points of a given class c. A "relative frequencies of occurrence" of point i , i.e. of the i-th neighbor of query point x, is

$$f(i; 1, N) = \frac{1/i}{H_N}$$

In fact, $f(i; 1, N)$ is a probability mass function of Zipfian distribution (Zipf's law). In our case $p(c|x)$ is a sum of probability mass functions for all appearances of class c.

From these considerations Theorem 4 above was formulated.

3) The Classifier construction: Let samples of the learning set (i.e. all samples regardless of the class) be sorted according to their distances from the query point x. Let indexes be assigned to these points so that 1 is assigned to the nearest neighbor, 2 to the second nearest neighbor etc. This sorting is important difference to both methods described before

that need no sorting when distribution mapping exponent or cortrelation dimension are known. Let us compute sums $S_0(x) = \frac{1}{N_0} \sum_{x_i \in U_0} 1/i$ and $S_1(x) = \frac{1}{N_1} \sum_{x_i \in U_1} 1/i$, i.e. the sums of the reciprocals of the indexes of samples from class $c = 0$ and from class $c = 1$. N_0 and N_1 are the numbers of samples of class 0 and class 1, respectively, $N_0 + N_1 = N$. The probability that point x belongs to class 0 is

$$\hat{p}(c = 0|x) = \frac{S_0(x)}{S_0(x) + S_1(x)} \tag{7}$$

and similarly the probability that point x belongs to class 1 is

$$\hat{p}(c = 1|x) = \frac{S_1(x)}{S_0(x) + S_1(x)} \tag{8}$$

When some discriminant threshold θ is chosen then if $\hat{p}(c = 1|x) > \theta$ point x is of class 1 else it is of class 0. This is the same procedure as in other classification approaches where the output is an estimation of probability (naive Bayes) or any real valued variable (neural networks). The value of threshold can be optimized with respect to minimal classification error. The default value of the discriminant threshold here is $\theta = 0.5$.

4) Generalization: Formulas above hold for two class problem with equal number of samples of both classes in the learning set. For larger number of classes and a different number of samples of classes formula has the form similar to (5):

$$\hat{p}(c|x) = \frac{1/N_c \sum_{x_i \in U_c} 1/i}{\sum_{k=1}^{C} (1/N_k \sum_{x_i \in U_k} 1/i)} \tag{9}$$

It is only a recalculation of the relative representation of different numbers of samples of one and the other class. For classification into more than two classes we use this formula for all classes and we assign to the query point x a class c for which $\hat{p}(c|x)$ is the largest.

IV. Experiments

We demonstrate the features and the power of the classifier both on synthetic and real-life data.

A. Synthetic Data

Synthetic data according to Paredes and Vidal [19] is two-dimensional and consists of three two-dimensional normal distributions with identical a-priori probabilities. If μ denotes the vector of the means and C_m is the covariance matrix, there is

Class $A : \mu = (2, 0.5)^t, C_m = (1, 0; 0, 1)$ (identity matrix)
Class $B : \mu = (0, 2)^t, C_m = (1, 0.5; 0.5, 1)$
Class $C : \mu = (0, -1)^t, C_m = (1, -0.5; -0.5, 1)$.

Fig. 3 shows the results obtained by the different methods for the different learning sets sizes from 8 to 256 samples and a testing set of 5000 samples all from the same distributions and independent. Each point in the figure was obtained by averaging over 100 different runs. It is seen that in this synthetic experiment, the DME based method presented here reliably outperforms all other methods shown and for a large number of samples fast approaches to the Bayes limit.

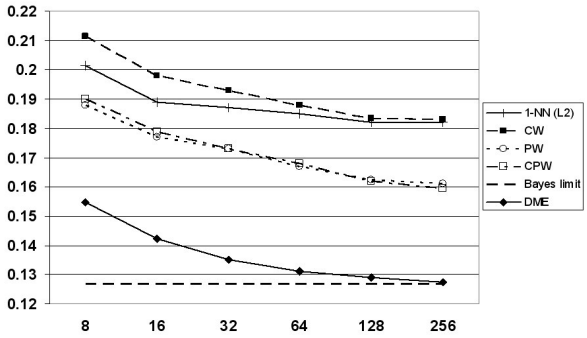


Fig. 3. Comparison of the classification errors of the synthetic data for the different approaches in dependence on the size of the learning set. In the legend, 1-NN(L2) means the 1-NN method with Euclidean metrics, CW, PW, and CPW are three variants of the method by Paredes and Vidal [19]; the points are estimated from this reference. Bayes means the Bayes limit, DME means the basic method presented here.

Note that in this test, the error of the DME estimation is combined with numerical errors, and with a negative influence of the low number of the samples giving the results presented in Fig. 3.

B. Data from Machine Learning Repository

The classification ability of the algorithm (DME) was tested using real-life tasks from the UCI Machine Learning Repository; see [1]. Seven databases have been used for the classification task, see Table 1.

TABLE 1. Classification mean square errors for four different methods including DME.

Data set	Attributes	Cross validation	$\sqrt{N}$ -NN	1-NN	Bayes	DME
Mushroom	22	1	0.0207	0	0.00764	0
Shuttle (Statlog)	9	1	0.00828	0.00259	0.01294	0.00207
Iris (see Friedman, 1994)	4	10	0.0488	0.0609	0.0854	0.0488
Congressional Voting ("Vote")	16	1	0.0602	0.1053	0.0977	0.0752
Spambase	57	1	0.113	0.0997	0.143	0.0886
Heart (Statlog)	13	9	0.158	0.245	0.182	0.166
Molecular Biology (Splice)	61	10	0.372	0.404	0.287	0.297

For the Shuttle data, the learning and testing sets are directly at hand and were used as they are. For smaller data sets a cross validation of 10 or 9 was used. The Iris data set was

modified into a two-class problem excluding the iris-setoza class according to Friedman [9]. The methods for comparison are

- 1-NN - standard nearest neighbor method
- Sqrt-NN - the k-NN method with k equal to the square root of the number of samples of the learning set
- Bayes - the naive Bayes method using ten bins histograms

For the k-NN, Bayes, and our method the discriminant thresholds were tuned accordingly. All procedures are deterministic (even Bayes algorithm) and then no repeated runs are needed. The testing shows the classification ability of the DME method for some tasks compared to the other published methods and results for the same data sets.

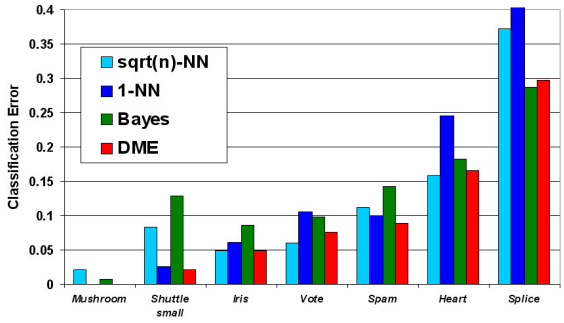


Fig. 4. Comparison of the classification errors for four different methods including the DME. Note that for the Mushroom data, both the 1-NN and DME algorithms give zero error. For the Shuttle data, the errors are ten times enlarged.

V. Open problems

We have shown the use of scaling for data classification. The three classifiers presented here were tested and can be used as they are similarly as one sometimes uses the 1-NN or k-NN methods. On the other hand, preprocessing like data editing or some kind of learning may essentially enhance classifier‘s behavior.

A. Editing

This is a way of learning data set modification that tries enhance especially borders between classes to make class recognition easier. Original idea of editing (or preclassification) [26] is to classify a sample of the learning set by the method for which edited learning data will be used. If classification result does not correspond to the sample class, remove this sample from the learning set else leave it there. After this is done, use the edited learning set for data classification by standard way. There are another ingenious methods that modify originally simple methods with help of learning, e.g. the learning weigting method [19] modifies the learning set by weighting classes and features and then uses simple 1-NN method similarly as [26].

B. Crossing phenomenon

The basic notion used here is the distribution mapping function. Depicted in the log-log coordinates it is approximately linearly growing function. When there are two classes we may have two such lines in one graph for a point x. If one line lies under the other, point x belongs to class of the lower line. But what if lines cross? And is the crossing point essential issue?

C. Scaled point processes but not exactly exponentially

The exponential scaling used here is a special case of more complex scaling functions. Transformation $z = r^q$ may have another form depending on the scaling function used. Main problem is scaling function identification [20], [21]. One can suppose that with the use of more realistic scaling function than exponential, may lead to modification of methods presented here and improving their behavior.

VI. Discussion

We found that when one can find a scaling of neighbors' distances measure, in form $z = r^q$, q is the distribution mapping exponent, then one can find a "Poisson process-like" behavior, i.e. Erlang distribution of neighbors' distances measure. Usually, a measure is considered that may depend on the embedding space dimension d (integer), while we use more general distribution mapping exponent q that is a positive real number.

Because the Erlang distribution converges to Gaussian distribution for index $k \to \infty$, the result according to Theorem 1 also relates to some results of e.g. [2], [8], [24] about convergence of near-neighbor distances.

The correlation dimension, eventually multifractal dimension, singularity (or Hölder) exponent or singularity strength, is often used for characterization of one dimensional or two-dimensional data, i.e. for signals and pictures. Our results are valid for multidimensional data that need not form a series because in this respect data is considered as individual points in a multidimensional space with proper metrics.

Our model of the polynomial expansion of the data space comes from the demand to have a uniform distribution of points, at least locally. There is an interesting relationship between the correlation dimension and the distribution mapping exponent. The former is a global feature of the fractal or data generating process; the latter is a local feature of the data set and is closely related to a particular query point. On the other hand, if linear regression were used, the computational procedure is almost the same in both cases. Moreover, it can be found that the values of the distribution mapping exponent lie sometimes in a narrow, sometimes in a rather wide interval around its mean value. Not surprisingly, the mean value of the distribution mapping exponent over all samples is not far from the correlation dimension. Introducing the notion of the distribution mapping exponent and the polynomial expansion of the distances may be a starting point for a more detailed description of the local behavior of the multivariate data and for the development of new approaches to the data analysis, including classification problems.

ISBN: 978-1-61804-240-8

Our experiments demonstrate that the simplest classifier based on the ideas introduced here can outperform other methods for some data sets. In all the tasks presented here, the distribution-mapping-exponent-based method outperforms or is comparable to the 1-NN algorithm and in six of the seven tasks, outperforms naive Bayes algorithm being only a little bit worse for the Splice data. All of these comparisons include an uncertainty in the computation of the distribution mapping exponent. By the use of the notion of distance, i.e. a simple transformation $E_n \rightarrow E_1$, the problems with dimensionality are easily eliminated at a loss of information on the true distribution of the points in the neighborhood of the query point which does not seem to be fundamental.

VII. Conclusion

This work was motivated by observation that near neighbors distances in homogenous Poisson processes in R^d have, in fact, the Erlang distribution modified so, that independent variable is substituted by term Kr^d, where K is a constant, r the distance of the neighbor and d the space dimension. This is the scaling function in exponential form. Here we answer a question, what if point process has arisen from underlying process with scaling exponent lesser than space dimension d.

This problem is solved by introduction of a distribution mapping function and its power approximation. It has been shown that the distribution mapping exponent of the power approximation is very close to the scaling exponent known from the theory of fractals and multifractals. It leads simplified, in the end, to strange scale measured by scaling exponent-power of neighbors' distances. It was then found that when using thus scaled measure for distance of the k-th neighbor one can construct simple and effective classifier; we have presented here three its variants and discussed some open problems.

References

[1] Bache, K., Lichman, M.(2013) UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, [online], 2013. Available: ¡http://archive.ics.uci.edu/ml¿.

[2] Bonetti M, Pagano M. (2005) The interpoint distance distribution as a descriptor of point patterns, with an application to spatial disease clustering. *Stat. Med.* Vol. 24, No. 5, pp. 753-773.

[3] Camastra, F. (2003) Data dimensionality estimation methods: a survey. Pattern Recognition, Vol. 6, pp. 2945-2954.

[4] Daley, D. J., Vere-Jones, D. (2005) *An Introduction to the Theory of Point Processes. Volume I, Elementary theory and methods. Second edition*, Springer.

[5] Daley, D.J., Verre-Jones, D. (2008) *An Introduction to the Theory of Point Processes. Volume II, General Theory and Structure. Second Edition*, Springer.

[6] Diggle, P.J. (2003) *A statistical Analysis of Spatial Point Processes*. Arnold, London.

[7] Dvorak,I., Klaschka,J. (1990) Modification of the Grassberger-Procaccia algorithm for estimating the correlation exponent of chaotic systems with high embedding dimension. *Physics Letters A,* Vol. 145, No. 5, pp. 225-231.

[8] Evans, D. (2008) A law of large numbers for nearest neighbour statistics. *Proc. R. Soc. A.* Vol. 464, pp. 3175–3192.

[9] Friedmann,J. H. (1994) Flexible Metric Nearest Neighbor Classification. *Technical Report, Dept. of Statistics, Stanford University*, 32 p.

[10] Grassberger, P., Procaccia, I. (1983) Measuring the strangeness of strange attractors, *Physica*, Vol. 9D, pp. 189-208.

[11] Guerrero, A., Smith, L.A. (2003) Towards coherent estimation of correlation dimension. *Physics letters A*, Vol. 318, pp. 373-379.

[12] Herbrich, R. (2002) *Learning Kernel Classifiers. Theory and Algorithms.* The MIT Press, Cambridge, Mass., London, England.

[13] Jiřina, M. (2013) Utilization of singularity exponent in nearest neighbor based classifier. *Journal of Classification (Springer)* Vol. 30, No. 1, pp. 3-29.

[14] Jiřina, M. (2014) Correlation Dimension-Based Classifier. *IEEE Transaction on Cybernetics* Vol. 44, in print.

[15] Jiřina, M., Jiřina, M., jr. (2014) Classification Using the Zipfian Kernel. *Journal of Classification (Springer).* Vol. 31, in print.

[16] Mandelbrot, B. B. (1982) *The Fractal Geometry of Nature*. W. H. Freeman & Co; ISBN 0-7167-1186-9.

[17] Mo, D., Huang, S.H. (2012) Fractal-Based Intrinsic Dimension Estimation and Its Application in Dimensionality Reduction. IEEE Trans on Knowledge and Data Engineering. Vol. 24 no. 1, pp. 59-71.

[18] Osborne,A. R., Provenzale,A.(1989) Finite correlation dimension for stochastic systems with power-law spectra. *Physica* D, Vol. 35, pp. 357-381, (1989).

[19] R. Paredes,R., Vidal,E.(2006) Learning Weighted Metrics to Minimize Nearest Neighbor Classification Error. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 7, pp. 1100-1110.

[20] Prokešovà, M., Hahn, U., and Jensen, E. B. V. (2006) Statistics for locally scaled point processes. *In Baddeley, A., Gregori, P., Mateu, J., Stoica, R., and Stoyan, D. (editors), Case Studies in Spatial Point Process Modelling, Lecture Notes in Statistics*, Springer, New York, Vol. 185, pp. 99-123.

[21] Prokešovà, M. (2010) Inhomogeneity in spatisal Cox point processes - location dependent thinning is not the only option. *Image Anal Stereol* Vol.29, pp. 133-141.

[22] Schlkopf, B., Smola, A.J. (2002) *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond.* The MIT Press, Cambridge, Mass., London, England.

[23] Schmuland, B.(2003) Random Harmonic Series. *American Mathematical Monthly* Vol 110, pp. 407-416, May 2003.

[24] Silverman, B. W. (1976) Limit theorems for dissociated random variables. *Advances in Applied Probability*, Vol. 8, pp.806-819.

[25] Takens, F. (1985) On the Numerical Determination of the Dimension of the Attractor. In: Dynamical Systems and Bifurcations, in: *Lecture Notes in Mathematics*, Vol. 1125, Springer, Berlin, p. 99-106.

[26] Wilson,D.L. (1972) Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans on System Man and Cybernetics*, Vol. SMC-2, No. 3, pp. 408-421. (July 1972)

[27] Zipf, G.K.(1968) The Psycho-Biology of Language. *An Introduction to Dynamic Philology* . The MIT Press, 1968.

A comparative study on principal component analysis and factor analysis for the formation of association rule in data mining domain

Dharmpal Singh[1], J.Pal Choudhary[2], Malika De[3]
[1]Department of Computer Sc. & Engineering, JIS College of Engineering
Block 'A' Phase III, Kalyani, Nadia-741235, West Bengal, INDIA
singh_dharmpal@yahoo.co.in
[2]Department of Information Technology, Kalyani Govt. Engineering College,
Kalyani, Nadia-741235, West Bengal, INDIA
jnpc193@yahoo.co.in
[3]Department of Engineering & Technological Studies, University of Kalyani
Kalyani, Nadia-741235, West Bengal, INDIA
demallika@yahoo.com

Abstract:
Association rule plays an important role in data mining. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data warehouse. Several authors have proposed different techniques to form the associations rule in data mining. But it has been observed that the said techniques have some problem. Therefore, in this paper an effort has been made to form the association using principal component analysis and factor analysis. A comparative study on the performance of principal component analysis and factor analysis has also been made to select the preferable model for the formation of association rule in data mining. A clustering technique with distance measure function has been used to compare the result of both the techniques. A new distance measure function named as Bit equal has been proposed for clustering and result has been compared with other exiting distance measure function.

Keywords: Data mining, Association rule, Factor analysis, Principal component analysis, Cluster, K-means and Euclidian distance, Hamming distance.

Introduction:
Data mining [1] is the process of extracting interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from large information repositories such as: relational database, data warehouses, XML repository, etc. Also data mining is known as one of the core processes of Knowledge Discovery in Database (KDD).
Association rule mining, one of the most important and well researched techniques of data mining, was first introduced in [2]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc.
Association rule mining has also been applied to e-learning systems for traditionally association analysis (finding correlations between items in a dataset), including, e.g., the following tasks: building recommender agents for on-line learning activities or 14 Enrique García, Cristóbal Romero, Sebastián Ventura and Toon Calders shortcuts [3], automatically guiding the learner's activities and intelligently generate and recommend learning materials [4], identifying attributes characterizing patterns of performance disparity between various groups of students [5], discovering interesting relationships from student's usage information in order to provide feedback to course author [6], finding out the relationships between each pattern of learner's behavior [7], finding students' mistakes that are often occurring together [8], guiding the search for best fitting transfer model of student learning [9], optimizing the content of an e-learning portal by determining the content of most interest to the user [10], extracting useful patterns to help educators and web masters evaluating and interpreting on-line course activities [3], and personalizing e-learning based on aggregate usage profiles and a domain ontology [11].

Association rule mining algorithms need to be configured before to be executed. So, the user has to give appropriate values for the parameters in advance (often leading to too many or too few rules) in order to obtain a good number of rules. A comparative study between the main algorithms that are currently used to discover association rules can be found in: Apriori [12], FP-Growth [13], MagnumOpus [14], and Closet [15].
Most of these algorithms require the user to set two thresholds, the minimal support and the minimal confidence, and find all the rules that exceed the thresholds specified by the user. Therefore, the user must possess a certain amount of expertise in order to find the right settings for support and confidence to obtain the best rules. Therefore an effort has been made to

form the association rule using the principal component analysis and factor analysis.

Markus Z¨oller[16] has discussed the ideas, assumptions and purposes of PCA (Principal component analysis) and FA. A comprehension and the ability to differ between PCA and FA have also been established.

Hee-Ju Kim [17] has examined the differences between common factor analysis (CFA) and principal component analysis (PCA). Further the author has opined that CFA (Common factor analysis) provided a more accurate result as compared to the PCA (Principal component analysis).

Diana D. Suhr [18] has discussed similarities and differences between PCA (Principal component analysis) and EFA. Examples of PCA (Principal component analysis) and EFA with PRINCOMP and FACTOR have also been illustrated and discussed.

Several authors have used data mining techniques [2-15] for association rule generation and the selection of best rules among the extracted rule. Certain authors have made a comparative study regarding the performance principal component analysis [16]-[18] and factor analysis [16]-[18].

During the extraction of association rule, it has been observed that some of rules have been ignored. In many cases, the algorithms generate an extremely large number of association rules, often in thousands or even in millions. Further, the component of association rule is sometimes very large. It is nearly impossible for the end users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Several strategies have been proposed to reduce the number of association rules, such as generating only “interesting” rules, generating only “nonredundant” rules, or generating only those rules satisfying certain other criteria such as coverage, leverage, lift or strength.

In order to eliminate the problems, the technique of factor analysis and principal component analysis has been applied on the available data to reduce the number of variables in the rules. Further it has been observed that if data are not in clean form poor result may be achieved. Therefore the data mining preprocessing techniques like data cleansing, data integration, data transformation and data reduction have to be applied on the available data to clean it in proper form to form the association rule in data mining. Thereafter, method of factor analysis and principal component analysis has been applied on the dataset. The comparative study of factor analysis and principal component analysis has been made by forming the different number of clusters on the total effect value as formed using factor analysis and principal component analysis. The distance measure function has been applied on the different number of cluster as formed using factor analysis and principal component analysis to select the preferable model. The bit equal distance measure function has been proposed to select the clustering elements based on the same number of bit of two elements. Here Iris Flower data set has been used as data set.

In first section, abstract and introduction have been furnished. In second section, brief methodologies of the models have been furnished. In third section, implementation has been furnished in detail. In fourth section, result and conclusion have been furnished.

2. Methodology:

2.1 Data Mining

A formal definition of data mining (DM) is also known as data fishing, data dredging (1960), knowledge discovery in databases (1990), or depending on the domain, as business intelligence, information discovery, information harvesting or data pattern processing.

Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

By data the definition refers to a set of facts (e.g. records in a database), whereas pattern represents an expression which describes a subset of the data, i.e. any structured representation or higher level description of a subset of the data. The term process designates a complex activity, comprised of several steps, while non-trivial implies that some search or inference is necessary, the straightforward derivation of the patterns is not possible. The resulting models or patterns should be valid on new data, with a certain level of confidence. The patterns have to be novel for the system and that have to be potentially useful, i.e. bring some kind of benefit to the analyst or the task. Ultimately, these should be interpretable, even if this requires some kind of result transformation.

An important concept is that of interestingness, which normally quantifies the added value of a pattern which combines validity, novelty, utility and simplicity. This can be expressed either explicitly, or implicitly, through the ranking performed by the DM (Data Mining) system on the returned patterns. Initially DM (Data Mining) has represented a component in the KDD (Knowledge Discovery in Databases) process which is responsible for finding the patterns in data.

2.2 Data Preprocessing

Data preprocessing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining. Data gathering methods handle resultant in out-of-range values (e.g., Income: 100), impossible data combinations (e.g., Gender: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is necessary to be reviewed before running an analysis.

If there exits irrelevant and redundant information or noisy and unreliable data, knowledge discovery during the training phase becomes difficult. Data preparation and filtering steps can take considerable amount of processing time. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. The data processing and data post processing depends on the user to form and represent the knowledge of data mining.

Data preprocessing includes the following techniques:
(1) Data cleaning: This technique includes fill in missing values, smooth noisy data, identify or remove outliers and resolve inconsistencies.
(2) Data integration: This technique includes integration of multiple databases, data cubes, or files.
(3) Data transformation: This technique includes normalization and aggregation.
(4) Data reduction: This technique is used to obtain reduced representation in volume but to produce the same or similar analytical results
(5) Data discretization: This is the part of data reduction but it is importance especially for numerical data.

2.3 Association Rule

The formal statement of association rule mining problem has been discussed by Agrawal et al. [18] in the year 1993. Let I= $\{I_1, I_2. \ldots I_m\}$ be a set of *m* distinct attributes, T be transaction that contains a set of items such that $T \subseteq I$, D be a database with different transaction records Ts. An association rule is an implication in the form of $X \Rightarrow Y$, where X, $Y \subset I$ are sets of items called item sets, and $X \cap Y = \phi$. X is called antecedent while Y is called consequent, X implies Y has formed as rule. There are two important basic measures for association rules, support(s) and confidence(c). Since the database is large and users concern only those frequently used items, usually thresholds of support and confidence are predefined by the users to drop those rules that are not so important or useful. The two thresholds are called minimal support and minimal confidence respectively, additional constraints of interesting rules also can be specified by the users. Support(s) of an association rule is defined as the percentage/fraction of records that contain $X \cup Y$ to the total number of records in the database. The count for each item is increased by one in every time when the item is encountered in different transaction T in database D during the scanning process. It means the support count does not take the quantity of the item into account. For example in a transaction a customer buys three bottles of beers but the support count number of {beer}is increased by one, in another word if a transaction contains an item then the support count of this item is increased by one. Support(*s*) is calculated by the following formula:

$$Support\ (XY) = \frac{Support\ count\ of\ XY}{Total\ number\ of\ transaction\ in\ D}$$

From the definition it can be mentioned that the support of an item is a statistical significance of an association rule. Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contains purchasing of this item. The retailer will not pay much attention to such kind of items that are not bought so frequently, obviously a high support is desired for more interesting association rules. Before the mining process, users can specify the minimum support as a threshold, which means the user are only interested in certain association rules that are generated from those item sets whose support value exceeds that threshold value. However, sometimes even the item sets are not as frequent as defined by the threshold, the association rules generated from them are still important. For example in the supermarket some items are very expensive, consequently these are not purchased so often as the threshold required, but association rules between those expensive items are as important as other frequently bought items to the retailer.

Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain $X \cup Y$ to the total number of records that contain X, where if the percentage exceeds the threshold of confidence an interesting association rule $X \Rightarrow Y$ can be generated.

$$Confidence\ (X/Y) = \frac{Support(XY)}{Support(X)}$$

Confidence is a measure of strength of the association rules, suppose the confidence of the association rule $X \Rightarrow Y$ is 80%, it means that 80% of the transactions that contain X and Y together, similarly to ensure the interestingness of the rules specified by minimum confidence is also pre-defined by users. Association rule mining is to find out association rules that satisfy the pre-defined minimum support and confidence from a given database. The problem is usually decomposed into two sub problems. One is to find those item sets whose occurrences exceed a predefined threshold in the database, those item sets are called frequent or large item sets. The second problem is to generate association rules from those large item sets with the constraints of minimal confidence. Suppose one of the large item sets is L_k. $L_k = \{ I_1, I_2,\ldots, I_{k-1}, I_k\}$ association rules with this item sets are generated in the following way, the first rule is $=\{ I_1, I_2,\ldots, I_{k-1}\} \Rightarrow \{I_k\}$, by checking the confidence this rule can be determined as important or not. Then other rules are generated by deleting the last items in the antecedent and inserting it into the consequent item, further the confidence of the new rules are checked to determine the importance of them. Those processes are iterated until the antecedent item becomes empty.

2.4 Factor Analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. In other words, it is possible, that variations in fewer observed variables mainly reflect the variations in total effect. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modeled as linear combinations of the potential factors, plus "error" terms. The information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset. Computationally this technique is equivalent to low rank approximation of the matrix of observed variables. Factor analysis originated in psychometrics, and is used in behavioral sciences, social sciences, marketing, product management, operations research, and other applied sciences that deal with large quantities of data.

2.5 Principal component analysis (PCA) is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called **principal components**. The number of principal components is less than or equal to the number of original variables. This

transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components. Principal components are guaranteed to be independent if the data set is jointly_normally distributed. PCA is sensitive to the relative scaling of the original variables.

2.6 Clustering

The Clustering method deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. The graphical example of cluster has shown in figure 1.9. In this case it is easily identified the 4 clusters into which the data can be divided, the similarity criterion is that two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called distance-based clustering.

2.6.1 K-Means Clustering

K-Means (MacQueen, 1967) clustering algorithm is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) with a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point it is needed to re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After these k new centroids have been obtained a new binding has to be done between the same data set points and the nearest new centroid. A loop has been formed. As a result of this loop it may be noticed that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

Finally, this algorithm aims at minimizing an objective function, generally a squared error function. The objective function as $J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j.

2.6.1.2 K-Means Algorithm

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group of centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers. The k-means algorithm can be run multiple times to reduce this effect. K-means algorithm is a simple algorithm that has been adapted to many problem domains.

2.6.3 Distance Measure

An important component of a clustering algorithm is the distance measure between data points. If the components of the data instance vectors are in the same physical units it is possible that the simple successfully group similar data instances. However, even in this case the Euclidean distance can sometimes be misleading.

2.6.3.1 Euclidean Distance

According to the Euclidean distance formula, the distance between two points in the plane with coordinates (x, y) and (a, b) is given by

$$\text{dist}((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

2.6.3.2 Hamming Distance

The Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different. In another way, it measures the minimum number of *substitutions* required to change one string into the other, or the minimum number of *errors* that could have transformed one string into the other.

The Hamming distance between:

- **1011101** and **1001001** is 2.
- **2173896** and **2233796** is 3.

2.6.3.3 Proposed Bit equal

It has been observed that hamming distance based on the minimum number of *errors* that could have transformed one string into the other. But clustering techniques based on grouping of object of similar types therefore here an effort has been made to group the object of similar type rather than based on the dissimilarity of strings (hamming distance). The bit equal distance measure function has been proposed to count the same number of bit of two elements. For an example, consider the strings 1011101 and 1001001, the hamming distance is 2, but in bit equal (Consider the strings have equal number bit) the similarity is 5. Here we can say

that five bits of the elements are same based the similarity properties of clustering.

3. Implementation:

The available data contains the information of Iris flowers with various items which have been furnished column wise in table 1. The data related to iris flower containing 40 data values which have been shown in table 2. It is to note that the quality of flower depends on the sepal length of the flower. If the sepal length of a particular flower is known, the quality of flower can be ascertained. Therefore the sepal length of flower (D) has been chosen as the objective item (consequent item) of the flowers. The other parameters i.e. sepal width (A), petal length (B) and petal width (C) have been chosen as the depending items (antecedent items).

The purpose of this work is to correlate the items A, B and C with D, so that based on any value of A, B and C, the value of D can be estimated. From the value of D the quality of that type of flower can be ascertained.

Table 1
Flower Characteristics

Item Name	Item description
A	Sepal width
B	Petal length
C	Petal width
D	Sepal length

Now it is necessary to check whether the data items are proper or not. If the data items are proper, extraction of information is possible otherwise the data items are not suitable for the extraction of knowledge. In that case preprocessing of data is necessary for getting the proper data. The set of 40 data elements has been taken which have been furnished in table 2.

Table 2
Available Data

Serial Number	A	B	C	D
1	3.5	1.4	0.2	5.1
2	3	1.4	0.2	4.9
3	3.2	1.3	0.2	4.7
4	3.1	1.5	0.2	4.6
5	3.6	1.4	0.2	5
6	3.9	1.7	0.4	5.4
7	3.4	1.4	0.3	4.6
8	3.4	1.5	0.2	5
9	2.9	1.4	0.2	4.4
10	3.1	1.5	0.1	4.9
11	3.7	1.5	0.2	5.4
12	3.4	1.6	0.2	4.8
13	3	1.4	0.1	4.8
14	3.2	4.7	1.4	7
15	3.2	4.5	1.5	6.4
16	3.1	4.9	1.5	6.9
17	2.3	4	1.3	5.5
18	2.8	4.6	1.5	6.5
19	2.8	4.5	1.3	5.7
20	3.3	4.7	1.6	6.3
21	2.4	3.3	1	4.9
22	2.9	4.6	1.3	6.6
23	2.7	3.9	1.4	5.2
24	2	3.5	1	5
25	3	4.2	1.5	5.9
26	2.2	4	1	6
27	3.3	6	2.5	6.3
28	2.7	5.1	1.9	5.8
29	3	5.9	2.1	7.1
30	2.9	5.6	1.8	6.3
31	3	5.8	2.2	6.5
32	3	6.6	2.1	7.6
33	2.5	4.5	1.7	4.9
34	2.9	6.3	1.8	7.3
35	2.5	5.8	1.8	6.7
36	3.6	6.1	2.5	7.2
37	3.2	5.1	2	6.5
38	2.7	5.3	1.9	6.4
39	3	5.5	2.1	6.8
40	2.5	5	2	5.7

The data mining preprocessing techniques like data cleansing, data integration, data transformation and data reduction have been applied on the available data as follows:

Data Cleansing

The data cleansing techniques include the filling in the missing value, correct or rectify the inconsistent data and identify the outlier of the data which have been applied on the available data.

It has been observed that each data set does not contain any missing value. The said data item does not contain any inconsistent data i.e. any abnormally low or any abnormally high value. All the data values are regularly distributed within the range of that data items. Therefore the data cleansing techniques are not applicable for the available data.

Data Integration

The data integration technique has to be applied if data has been collected from different sources. The available data have been taken from a single source therefore the said technique is not applicable here.

Data transformation

The data transformation techniques such as smoothing, aggregation, normalization, decimal scaling have to be applied to get the data in proper form. Smoothing technique has to be applied on the data to remove the noise from the data. Aggregation technique has to be applied to summarize the data. To make the data within specific range smoothing and normalization technique have to be applied. Decimal scaling technique has to be applied to move the decimal point to particular position of all data values. Out of these data transformation techniques, smoothing and decimal scaling techniques have been applied on the data as furnished in table 2 to get the revised data as furnished in table 3.

Table 3
Revised Data

Serial Number	A	B	C	D

ISBN: 978-1-61804-240-8

1	3500	1400	200	5100
2	3000	1400	200	4900
3	3200	1300	200	4700
4	3100	1500	200	4600
5	3600	1400	200	5000
6	3900	1700	400	5400
7	3400	1400	300	4600
8	3400	1500	200	5000
9	2900	1400	200	4400
10	3100	1500	100	4900
11	3700	1500	200	5400
12	3400	1600	200	4800
13	3000	1400	100	4800
14	3200	4700	1400	7000
15	3200	4500	1500	6400
16	3100	4900	1500	6900
17	2300	4000	1300	5500
18	2800	4600	1500	6500
19	2800	4500	1300	5700
20	3300	4700	1600	6300
21	2400	3300	1000	4900
22	2900	4600	1300	6600
23	2700	3900	1400	5200
24	2000	3500	1000	5000
25	3000	4200	1500	5900
26	2200	4000	1000	6000
27	3300	6000	2500	6300
28	2700	5100	1900	5800
29	3000	5900	2100	7100
30	2900	5600	1800	6300
31	3000	5800	2200	6500
32	3000	6600	2100	7600
33	2500	4500	1700	4900
34	2900	6300	1800	7300
35	2500	5800	1800	6700
36	3600	6100	2500	7200
37	3200	5100	2000	6500
38	2700	5300	1900	6400
39	3000	5500	2100	6800
40	2500	5000	2000	5700

Data Reduction

The data reduction method has to be applied on huge amount of data (tetra byte) to get the certain portion of data. Since here the amount of data is not large, the said technique has not been applied.

Association Rule Formation

The association rule has been formed using the data items as furnished in table 2 with attribute description as present in table 1 as follows:

Step 1

Rule 1 IF A = 3500 AND B = 1400 AND C = 200 THEN D = 5100

Rule 2 IF A = 3000 AND B = 1400 AND C = 200 THEN D = 4900

Rule 3 IF A = 3200 AND B = 1300 AND C = 200 THEN D = 4700

. . . . . .

. . . . . .

Rule 40 IF A = 2500 AND B = 5000 AND C = 2000 THEN D = 5700

It is to mention that each rule has been formed on the basis of the value of each data item. In table 3, the first row contains the value of A, B, C, D as 3500, 1400, 200, 5100 respectively. On the basis of values of data items, the rule1 has been formed. Accordingly all rules have been set up. Here the item D has been chosen as objective item (consequent item) and other items A, B, C have been taken as depending items (antecedent items). Since there are 40 data items, 40 rules have been formed based on different values of data items.

Step 2

The numerical values have been assigned to the attribute of the each rule which has been furnished in table 4. Here the range of A, B, C and D have been divided in three equal part. As for example, the numerical value as 100 has been assigned to attribute A if the value of A lies between 2000 to 2633.33. The value 100 has been termed as A_1. The numerical value as 150 has been assigned to attribute A if the value of A lies between 2633.33 to 3266.66. The value as 150 has been termed as A_2. Accordingly the numerical values have been assigned to all data items.

Table 4
Numerical Values for the Range

Item/ Range			
A	A_1(2000-2633.33) = 100	A_2(2633.33-3266.66) = 150	A_3(3266.66-3900) = 200
B	B_1(1300-3066.67) = 201	B_2(3066.67-4833.34) = 251	B_3(4833.34-6600) = 301
C	C_1(100-900) = 302	C_2(900-1700) = 352	C_3(1700-2500) = 402
D	D_1(4400-5466.67) = 600	D_2(5466.67-6533.34) = 650	D_3(6533.34-7600) = 700

Step 3

Numerical values have been assigned to each antecedent item (A, B, C) and consequent item (D). For rule number 1, the values of A, B, C and D are 3500, 1400, 200 and 5100 respectively. Using the values, numerical values have been assigned to the rule based on the range of values as specified in table 4.

For the rule number one, the numerical value of A, B and C are 200, 201 and 302 respectively (from table 3) therefore the sum of all antecedent numerical values are = 200 + 201 + 302 = 703 which has been termed as cumulative antecedent item. Using the range of the values as furnished in table 4, the numerical value for D has been assigned as 600. This procedure has been repeated for all set of rules. The cumulative antecedent item and consequent item for all rules have been furnished in table 5.

Table 5
Antecedent and Consequent Item

Serial Number	Antecedent Item	Consequent Item
1	703	600
2	653	600

3	653	600
4	653	600
5	703	600
6	703	600
7	703	600
8	703	600
9	653	600
10	653	600
11	703	600
12	703	600
13	653	600
14	753	700
15	753	650
16	803	700
17	703	650
18	753	650
19	753	650
20	803	650
21	753	600
22	753	700
23	753	600
24	703	600
25	753	650
26	703	650
27	903	650
28	853	650
29	853	700
30	853	650
31	853	650
32	853	700
33	753	600
34	853	700
35	803	700
36	903	700
37	853	650
38	853	650
39	853	700
40	803	650

Step 4

It is to note that numerical values have been assigned to each item by dividing the items values into certain range of values. Here the items A, B, C and D have been divided into three ranges of values. Further it is to note that it may happen that each item may have to divide into unequal range of values. In that case it may be difficult to handle range of values. Apart from that, it may happen that a particular consequent item may be related to a number of cumulative antecedent items with huge difference in data values.

The support and confidence has to be found on Iris data for the application of Apriori algorithm. In case of Iris flower data all the data in numeric form where it is not possible to calculate the confidence and support using Apriori algorithm method mentioned by Agrawal et al. [18] in the year 1993. Therefore an effort has to be made to calculate the confidence and support in following ways. To calculate the support and confidence all the data have to be converted into range which has been furnished in table 4. Base on the table 6 the all data have been assigned a range which has been furnished in table 6.

Table 6
Assigned Range of the Data Element

Serial Number	A	B	C	D	Range of Data			
1	3500	1400	200	5100	A3	B1	C1	D1
2	3000	1400	200	4900	A2	B1	C1	D1
3	3200	1300	200	4700	A2	B1	C1	D1
4	3100	1500	200	4600	A2	B1	C1	D1
5	3600	1400	200	5000	A3	B1	C1	D1
6	3900	1700	400	5400	A3	B1	C1	D1
7	3400	1400	300	4600	A3	B1	C1	D1
8	3400	1500	200	5000	A3	B1	C1	D1
9	2900	1400	200	4400	A2	B1	C1	D1
10	3100	1500	100	4900	A2	B1	C1	D1
11	3700	1500	200	5400	A3	B1	C1	D1
12	3400	1600	200	4800	A3	B1	C1	D1
13	3000	1400	100	4800	A2	B1	C1	D1
14	3200	4700	1400	7000	A2	B1	C2	D3
15	3200	4500	1500	6400	A2	B1	C2	D2
16	3100	4900	1500	6900	A2	B2	C2	D3
17	2300	4000	1300	5500	A1	B2	C2	D2
18	2800	4600	1500	6500	A2	B2	C2	D2
19	2800	4500	1300	5700	A2	B2	C2	D2
20	3300	4700	1600	6300	A3	B2	C2	D2
21	2400	3300	1000	4900	A1	B2	C2	D1
22	2900	4600	1300	6600	A3	B2	C2	D3
23	2700	3900	1400	5200	A2	B2	C2	D1
24	2000	3500	1000	5000	A1	B2	C2	D1
25	3000	4200	1500	5900	A2	B2	C2	D2
26	2200	4000	1000	6000	A1	B2	C2	D2
27	3300	6000	2500	6300	A3	B3	C3	D2
28	2700	5100	1900	5800	A2	B3	C3	D2
29	3000	5900	2100	7100	A2	B3	C3	D3
30	2900	5600	1800	6300	A2	B3	C3	D2
31	3000	5800	2200	6500	A2	B3	C3	D2
32	3000	6600	2100	7600	A2	B3	C3	D3
33	2500	4500	1700	4900	A1	B3	C3	D1
34	2900	6300	1800	7300	A2	B3	C3	D3
35	2500	5800	1800	6700	A1	B3	C3	D2
36	3600	6100	2500	7200	A3	B3	C3	D3
37	3200	5100	2000	6500	A2	B3	C3	D2
38	2700	5300	1900	6400	A2	B3	C3	D2
39	3000	5500	2100	6800	A2	B3	C3	D3
40	2500	5000	2000	5700	A1	B3	C3	D2

Here A1, A2 and A3 occurred 11, 22 and 7 times respectively. Similarly B1, B2, B3, C1, C2, C3, D1, D2 and D3 have been occurred 15, 12, 13, 13, 13, 14, 17, 15 and 8 respectively.

According to Apriori algorithm support of A1 will be 11/40 (27.5%), A2=22/40 (55%), A3=7/40 (17.5%), B1=15/40 (37.5%), B2=12/40, (30%) B3=13/40, (32.5%) C1=13/40 (32.5), D1= (42.5%), D2=15/40(37.5%) and D3=8/15(20%).

If support is less than 30% then ignore the rule. Here A3 will be ignored from the set of rule. Therefore it is necessary to ignore the rule where A3 has occurred. Therefore, seven rules have to be ignored from the rule set. This action further ignored the rules which are associated with A3. For an

example in step 1 A3 has occurred seven places out of 40 rules (as shown in table 7. If the A3 will be ignored therefore other values which have been associated with it has to be ignored to make the rule consistent. It has been seen that if rule has been ignored in this ways then almost all rule will be ignored.
In order to eliminate the problems, the techniques of factor analysis and principal component analysis have been applied on the available data which have been described as follows:

3.1 Multivariate Data Analysis

The methods of factor analysis and principal component analysis have been applied on the available input data to make a decision to select the optimal model for the formation of association rule.

3.1.1 Factor Analysis

Step 1

Using the values of A, B and C as furnished in table 2, the correlation of coefficient of the items A and B, A and C, B and C, A and A, B and B, C and C have been computed and these coefficients have been stored in table 7 termed as correlation matrix.

Table 7
Correlation matrix

1.0000	-0.3395	-0.2858
-0.3395	1.0000	0.9740
-0.2858	0.9740	1.0000

Step 2

The eigen value and eigen vectors of the elements A, B and C using the above correlation matrix have been calculated using Matlab Tools. The contributions of eigen value of each item among all other items has been calculated and it has been observed that percentage contribution of the item A (0.81%) is less as compared to other items (B (27.61%), C (71.49%)) therefore it has been ignored. The major factors have been calculated as per the formula $\sqrt{(}$eigen value × eigen vector) using the selected eigen value and eigen vector The major factors have been furnished in table 8.

Table 8
Contribution of Eigen Vector Corresponding Eigen Value

Data Attribute /Eigen Value	0.8308	2.1448
A	0.8479	-0.5263
B	0.2037	0.9696
C	0.2602	0.9563

Step 3

The cumulative effect value for all these data items have been calculated by adding the values row wise corresponding to each element of table 7. The cumulative values for all items have been calculating and furnished in table 9.

Table 9
Cumulative Effect Value of Items

Data Attribute	Cumulative Effect
A	0.32
B	1.17
C	1.22

Step 4

Now a relation has been formed by using the cumulative effect value of all the elements to produce total effect value.
Total effect value = (0.32) × A + (1.17) × B + (1.22) × C. Now using the relation, a resultant total effect value has been furnished in table 10.

Table 10
Total Effect Value

Serial Number	A	B	C	Total Effect Value
1	3500	1400	200	3002
2	3000	1400	200	2842
3	3200	1300	200	2789
4	3100	1500	200	2991
5	3600	1400	200	3034
6	3900	1700	400	3725
7	3400	1400	300	3092
8	3400	1500	200	3087
9	2900	1400	200	2810
10	3100	1500	100	2869
11	3700	1500	200	3183
12	3400	1600	200	3204
13	3000	1400	100	2720
14	3200	4700	1400	8231
15	3200	4500	1500	8119
16	3100	4900	1500	8555
17	2300	4000	1300	7002
18	2800	4600	1500	8108
19	2800	4500	1300	7747
20	3300	4700	1600	8507
21	2400	3300	1000	5849
22	2900	4600	1300	7896
23	2700	3900	1400	7135
24	2000	3500	1000	5955
25	3000	4200	1500	7704
26	2200	4000	1000	6604
27	3300	6000	2500	11126
28	2700	5100	1900	9149
29	3000	5900	2100	10425
30	2900	5600	1800	9676
31	3000	5800	2200	10430
32	3000	6600	2100	11244
33	2500	4500	1700	8139
34	2900	6300	1800	10495
35	2500	5800	1800	9782
36	3600	6100	2500	11339
37	3200	5100	2000	9431
38	2700	5300	1900	9383
39	3000	5500	2100	9957
40	2500	5000	2000	9090

Step 5

The total effect value has been formed using the value of the items A, B and C. This total effect value has been related to the value of D. The distribution of sorted total effect value and the corresponding value of D have been furnished in figure 1. The total effect value has been named as cumulative antecedent item and values of D has been named as consequent item.

ISBN: 978-1-61804-240-8

3.1.2 Principal Component Analysis

Step 1

Using the values of A, B and C as furnished in table 2, the covariance coefficient of the items A and B, A and C, B and C, A and A, B and B, C and C have been computed and these coefficients have been stored in table 11 termed as covariance matrix.

Table 11
Covariance Matrix

0.1731	-0.2551	-0.0933
-0.2551	3.2614	1.3803
-0.0933	1.3803	0.6158

Step 2

The eigen value and eigen vectors of the elements A, B and C using above covariance matrix have been calculated using Matlab Tools, it has been observed that percentage contribution of the item A (0.62%) is less as compared to other (B (3.82%), C (95.55%)) therefore it has been ignored. The major factors have been calculated as per the formula (√eigen value × eigen vector) using the selected eigen value eigen vector. The major factors have been furnished in table 12.

Table 12
Contribution of Eigen Vector Corresponding Eigen Value

Data Attribute/ Eigen Value	0.1548	3.8703
A	0.3866	-0.1442
B	0.0092	1.8073
C	0.0508	0.7707

Step 3

The cumulative effect value for all these data items have been calculated by adding the values row wise corresponding to each element of table 11. The cumulative value for all items has been furnished in table 13.

Table 13
Cumulative Effect Value of Items

Data Attribute	Cumulative Effect
A	0.24
B	1.82
C	0.82

Step4

Now a relation has been formed by using the cumulative effect value of all the elements to produce total effect value.

Total effect value = (0.24) × A + (1.82) × B + (0.82) × C. Now using the relation, a resultant total effect value has been computed which has been furnished in table 14.

Table 14
Total Effect Value

Serial Number	A	B	C	Total Effect Value
1	3500	1400	200	3552
2	3000	1400	200	3432
3	3200	1300	200	3298
4	3100	1500	200	3638
5	3600	1400	200	3576
6	3900	1700	400	4358
7	3400	1400	300	3610
8	3400	1500	200	3710
9	2900	1400	200	3408
10	3100	1500	100	3556
11	3700	1500	200	3782
12	3400	1600	200	3892
13	3000	1400	100	3350
14	3200	4700	1400	10470
15	3200	4500	1500	10188
16	3100	4900	1500	10892
17	2300	4000	1300	8898
18	2800	4600	1500	10274
19	2800	4500	1300	9928
20	3300	4700	1600	10658
21	2400	3300	1000	7402
22	2900	4600	1300	10134
23	2700	3900	1400	8894
24	2000	3500	1000	7670
25	3000	4200	1500	9594
26	2200	4000	1000	8628
27	3300	6000	2500	13762
28	2700	5100	1900	11488
29	3000	5900	2100	13180
30	2900	5600	1800	12364
31	3000	5800	2200	13080
32	3000	6600	2100	14454
33	2500	4500	1700	10184
34	2900	6300	1800	13638
35	2500	5800	1800	12632
36	3600	6100	2500	14016
37	3200	5100	2000	11690
38	2700	5300	1900	11852
39	3000	5500	2100	12452
40	2500	5000	2000	11340

Step 5

The total effect value has been formed using the value of the items A, B and C. This total effect value has been related to the value of D. The sorted total effect value and the corresponding value of D have been furnished in figure 1. The total effect value has been named as cumulative antecedent item and values of D has been named as consequent item.

ISBN: 978-1-61804-240-8

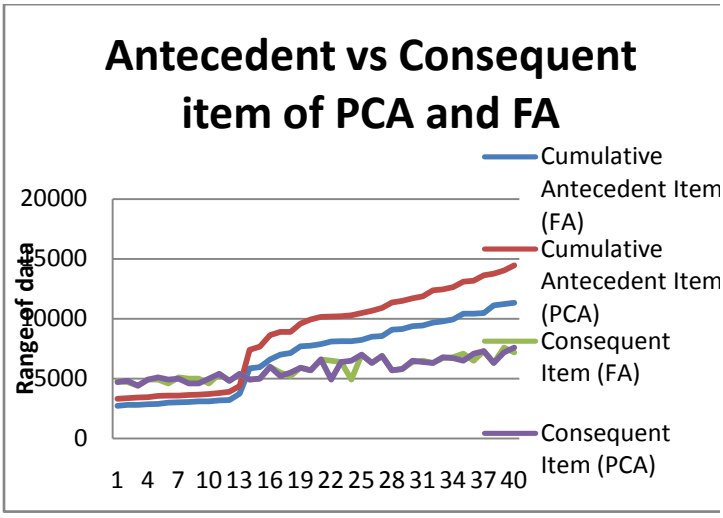


Figure 1:
Antecedent vs. Consequent item of PCA and FA

Step 6

K-means clustering has been applied on the total effect value using factor analysis and principal component analysis with a number of clusters (2, 3, 4, and 12).

Step 7

Several distance function have been used to calculate the cumulative distance value for the total cluster cumulative distance value has been calculated by summing the distance values of individual elements within the cluster from respective cluster.

Step 8

The cumulative distance value for different clusters have been furnished in table 15 using Iris Flower data set and Wine Data set and Boston city data set have been furnished in table 16 and table 17 respectively.

Table 15
Data used Iris data set
PCA vs. Factor Analysis Using Different Clusters

	Factor Analysis				Principal Component Analysis			
No. of cluster/ Distance Function	**(Two Cluster)**	**(Three Cluster)**	**(Four Cluster)**	**(Twelve Cluster)**	**(Two Cluster)**	**(Three Cluster)**	**(Four Cluster)**	**(Twelve Cluster)**
Euclidean	10577.91	6847.881	1858.28	555.77	13588.59	8419.52	6951.631	2779.303
Hamming	269	237	228	169	279	259	254	199
Bit Equal	192	181	176	129	219	233	238	169

Table 16
Data used Wine Data Set
PCA vs. Factor Analysis Using Different Clusters

	Factor Analysis				Principal Component Analysis			
No. of cluster / Distance Function	**(Two Cluster)**	**(Three Cluster)**	**(Four Cluster)**	**(Twelve Cluster)**	**(Two Cluster)**	**(Three Cluster)**	**(Four Cluster)**	**(Twelve Cluster)**
Euclidean	145889.2	118276	83136.7	29373.85	220367.5	178716.2	126225.7	53575.41
Hamming	369	347	314	269	403	399	366	292
Bit Equal	282	266	246	229	319	303	288	217

Table 17
Data used Boston city Data Set
PCA vs. Factor Analysis Using Different Clusters

	Factor Analysis				Principal Component Analysis			
No. of cluster / Distance Function	**(Two Cluster)**	**(Three Cluster)**	**(Four Cluster)**	**(Twelve Cluster)**	**(Two Cluster)**	**(Three Cluster)**	**(Four Cluster)**	**(Twelve Cluster)**
Euclidean	14637.05	11507.74	8696.66	2187.659	149111.2	11914 9.6	10382 3.9	30728.32
Hamming	267	247	214	198	303	289	256	196
Bit Equal	182	166	146	129	219	203	188	157

Step 9

The objective of the formation of cluster is to separate samples of distinct group by transforming them to space which minimizes their within class variability. From table 15, 16 and table 17 it has been observed that the factor analysis is more effective as compared to principal component analysis after the formation of clusters using data items i.e. Iris Flower, Wine Data set and Boston city data set.

Therefore factor analysis method is suitable multivariate analysis method using Iris data, Wine Data set and Boston city data set for the formation of association rule.

4. Result

The methods of multivariate analysis (factor analysis and principal component analysis) have been applied on the available input data to make a decision to select the optimal model for the formation of association rule. The cumulative antecedent item has been formed using these methods. . From table 15, 16 and table 17 it has been observed that the factor analysis is more effective as compared to principal component analysis after the formation of clusters using data items i.e. Iris Flower, Wine Data set and Boston city data set.

Conclusion:

It has been observed that AND methodology and Apriori algorithm were not useful to form the association rule for the Iris Flower data set, Wine Data set and Boston city data set. In order to eliminate the problems, the techniques of factor analysis and principal component analysis have been applied on the available data. From the table 15, 16 and 17 it has been observed that the cumulative distance value within the cluster is less in case of factor analysis as compared to principal component analysis using two, three, four and twelve clusters. Therefore cumulative antecedent value using factor analysis has to be used to from the association rule in data mining. In future, initially the same number of clusters has been formed using the objective item value as D. Thereafter, a relation has been formed using total effect value and objective item value of the data sets.

ISBN: 978-1-61804-240-8

Reference:

[1] Chen, M.-S., Han, J., and Yu, P. S, "Data mining: an overview from a database perspective", *IEEE Transaction on Knowledge and Data Engineering 8*, pp. 866-883, 1996.

[2] Agrawal, R., Imielinski, T., and Swami, A. N., "Mining association rules between sets of items in large databases", *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 207-216, 1993.

[3]. Zaiane, O., "Building a Recommender Agent for e-Learning Systems" Proc. of the Int.Conf. in Education, pp. 55-59, 2002.

[4] Lu, J." Personalized e-learning material recommender system", *In: Proc. of the Int. Conf. on Information Technology for Application*, pp. 374–379, 2004

[5] Minaei-Bidgoli, B., Tan, P., Punch, W., "Mining interesting contrast rules for a web-based educational system" *In: Proc. of the Int. Conf. on Machine Learning Applications*, pp. 1-8, 2004.

[6]. Romero, C., Ventura, S., Bra, P. D., "Knowledge discovery with genetic programming for providing feedback to courseware author: User Modeling and User-Adapted Interaction:" *The Journal of Personalization Research*, Vol. 14, No. 5, pp.425–464, 2004.

[7]. Yu, P., Own, C., Lin, L. "On learning behavior analysis of web based interactive Environment", *In: Proc. of the Int. Conf. on Implementing Curricular Change in Engineering Education*, pp. 1-10, 2001.

[8]. Merceron, A., & Yacef, K, "Mining student data captured from a web-based tutoring tool", *Journal of Interactive Learning Research*, Vol.15, No.4, pp.319–346, 2004.

[9]. Freyberger, J., Heffernan, N., Ruiz, C "Using association rules to guide a search for best fitting transfer models of student learning" *In: Workshop on Analyzing Student-Tutor Interactions Logs to Improve Educational Outcomes at ITS Conference* , pp. 1-10, 2004.

[10]. Ramli, A.A., "Web usage mining using apriori algorithm: UUM learning care portal case", *In: Proc. of the Int. Conf. on Knowledge Management* , pp.1-19, 2005.

[11] Markellou, P., Mousourouli, I., Spiros, S., Tsakalidis, A,"Using Semantic Web Mining
Technologies for Personalized e-Learning Experiences", *In: Proc. of the Int. Conference on Web-based Education* pp. 1-10, 2005

[12]Agrawal, R. and Srikant, R., "Fast algorithms for mining association rules", *In: Proc. of Int. Conf. on Very Large Data Bases*, pp. 487-499, 1996.

[13]. Han, J., Pei, J., and Yin, Y., "Mining frequent patterns without candidate generation" *In: Proc. of ACM-SIGMOD Int. Conf. on Management of Data* , pp.1-12, 1999.

[14]. Webb, G.I.: OPUS, "An efficient admissible algorithm for unordered search", Journal *of Artificial Intelligence Research* Vol. 3, pp. 431-465, 1995.

[15] Pei, J., Han, J., Mao, R., "CLOSET: An efficient algorithm for mining frequent closed itemsets", *In Proc. of ACM_SIGMOD Int. DMKD*, pp. 21-30, 2000.

[16]Markus Z¨oller "A Comparison between Principal Component Analysis and Factor Analysis", University Of Applied Sciences W¨Urzburg-Schweinfurt, pp 1-4, 2012.

[17] Hee-Ju Kim "Common Factor Analysis Versus Principal Component Analysis: Choice for Symptom Cluster Research", *Asian Nursing Research*, Vol. 2, No. 1, pp.17–24, 2000.

[18] Diana D. Suhr "Principal Component Analysis vs. Exploratory Factor Analysis", *University of Northern Colorado*, pp. 203-230.

[19] D. P. Singh and J. P. Choudhury, "Assessment of Exported Mango Quantity By Soft Computing Model," *in IJITKM-09 International Journal*, Kurukshetra University, pp. 393-395, June-July 2009.

[20] D. P. Singh, J. P. Choudhury and Mallika De, "Performance Measurement of Neural Network Model Considering Various Membership Functions under Fuzzy Logic," *International Journal of Computer and Engineering*, vol. 1, no. 2, pp.1-5, 2010.

[21] D. P. Singh, J. P. Choudhury and Mallika De, "Performance measurement of Soft Computing models based on Residual Analysis," *in International Journal for Applied Engineering and Research, Kurukshetra University*, vol.5, pp 823-832, Jan-July, 2011.

[22] D. P. Singh, J. P. Choudhury and Mallika De, "A comparative study on the performance of Fuzzy Logic, Bayesian Logic and neural network towards Decision Making," *in International Journal of Data Analysis Techniques and Strategies*, vol.4, no.2, pp. 205-216, March, 2012.

[23] D. P. Singh, J. P. Choudhury and Mallika De, "Optimization of Fruit Quantity by different types of cluster techniques," *in PCTE Journal Of Computer Sciences*., vol.8, no.2, pp .90-96 , June-July, 2011.

[24] D. P. Singh, J. P. Choudhury and Mallika De, " A Comparative Study on the performance of Soft Computing models in the domain of Data Mining," *International Journal of Advancements in Computer Science and Information Technology*, vol. 1, no. 1, pp. 35-49, September, 2011.

[25] D. P. Singh, J. P. Choudhury and Mallika De," A Comparative Study to Select a Soft Computing Model for Knowledge Discovery in Data Mining," *in International Journal of Artificial Intelligence and Knowledge Discovery*, Vol. 2, no. 2, pp. 6-19, April, 2012.

Bibliography:

Dharmpal Singh has received Bachelor of Computer Science & Engineering from West Bengal University of Technology and Master of Computer Science & Engineering also from West Bengal University of Technology. He has about six years of experience in teaching and research. At present he is with JIS College of Engineering, Kalyani, West Bengal, India as Assistant Professor. Now he is pursuing PhD in University of Kalyani. He has about 15 publications in National and International Journals and Conference Proceedings.

Complex Probability and Markov Stochastic Process

Bijan Bidabad, Behrouz Bidabad, Nikos Mastorakis

***Abstract*—**This paper discusses the existence of "complex probability" in the real world sensible problems. By defining a measure more general than conventional definition of probability, the transition probability matrix of discrete Markov chain is broken to the periods shorter than a complete step of transition. In this regard, the complex probability is implied.

***Keywords*—** Probability, Markov Chain, Stochastic Process.

I. INTRODUCTION

SOMTIMES analytic numbers coincide with the mathematical modeling of real world and make the real analysis of problems complex. All the measures in our everyday problems belong to R, and mostly to R^+. Probability of occurrence of an event always belongs to the range [0,1]. In this paper, it is discussed that to solve a special class of Markov chain which should have solution in real world, we are confronted with "analytic probabilities"!. Though the name probability applies to the values between zero and one, we define a special analogue measure of probability as complex probability where the conventional probability is a subclass of this newly defined measure.

Now define the well-known discrete time Markov chain $\{Y_n\}$ a Markov stochastic process whose state space is $s=\{1,2,...,N\}$ for which $T=\{0,1,2,...\}$. Refer to the value of Y_n as the outcome of the n^{th} trial. We say Y_n being in state i if $Y_n = i$. The probability of Y_{n+1} being in state j, given that Y_n is in state i (called a one–step transition probability) is denoted by $P_{ij}^{n,n+1}$, i.e.,

$$P_{ij}^{n,n+1}=P_{ij}=\Pr\{Y_{n+1}=j \mid Y_n=i\} \tag{1}$$

Therefore, the Markov or transition probability matrix of the process is defined by

$$\mathbf{P}=\|P_{ij}\|$$
$$P_{ij}\geq 0 \quad \forall i,j\in s$$
$$\sum_{j=1}^{N} P_{ij}=1 \quad \forall i\in s \tag{2}$$

The n-step transition probability matrix $\mathbf{P}^{(n)}=\|P_{ij}^{n}\|$, which P_{ij}^{n} denotes the probability that the process goes from state i to state j in n transitions. Formally,

$$P_{ij}^{n}=\Pr\{Y_{n+m}=j \mid Y_m=i\} \qquad \forall i,j\in S \tag{3}$$

According to Chapman – Kolmogorov relation for discrete Markov matrices (Karlin and Taylor (1975)), it can be proved that

$$\mathbf{P}^{(n)}=\mathbf{P}^{n} \quad n\in N\,(Natural\ numbers) \tag{4}$$

$\mathbf{P}^n$ that is $\mathbf{P}$ to the power n is a Markov matrix if $\mathbf{P}$ is Markov.

Now, suppose that we intend to derive the t-step transition probability matrix $\mathbf{P}^{(t)}$ where $t\geq 0$ from the above (3) and (4) definition of n-step transition probability matrix $\mathbf{P}$. That is, to find the transition probability matrix for incomplete steps. On the other hand, we are interested to find the transition matrix $\mathbf{P}^{(t)}$ when t is between two sequential integers. This case is not just a tatonnement example. To clarify the application of this phenomenon, consider the following example.

Example 1. Usually in population census of societies with N distinct regions, migration information is collected in an NxN migration matrix for a period of ten years. Denote this matrix by **M**. Any element of **M**, m_{ij} is the population who leaved region i and went to region j through the last ten years. By deviding each m_{ij} to sum of the i^{th} row of **M**, a value of P_{ij} is computed as an estimate of probability of transition from i^{th} to j^{th} regions. Thus, the stochastic matrix **P** gives the probabilities of going from region i to region j in ten years (which is one–step transition probability matrix). The question is: how we can compute the transition probability matrix for one year or one-tenth step and so on.

If we knew the generic function of probabilities in very small period of time we would be able to solve problems similar to example 1. But the generic function (5) is not obtainable. If it were, we would apply the continous time Markov procedure using the generic NxN matrix **A** as:

Bijan Bidabad is WSEAS Post Doctorate Researcher, (No. 2, 12th St., Mahestan Ave., Shahrak Gharb, Tehran, 14658, IRAN. (Tel.: +98-21-88360810, Mobile: +98-912-1090164, Fax: +98-21-88369336, email: bijan@bidabad.com , web: http://www.bidabad.com)

Behrouz Bidabad, Faculty of mathematics, Polytechnics University, Hafez Ave., Tehran, 15914, IRAN (email: bidabad@aut.ac.ir, web: http://www.aut.ac.ir/official/main.asp?uid=bidabad)

Nikos Mastorakis, Technical University of Sofia, Bulgaria, Department of Industrial Engineering, Sofia, 1000, BULGARIA (email: mastor@tu-sofia.bg, web: http://elfe.tu-sofia.bg/mastorakis)

$$\mathbf{A} = \lim_{h \to o^+} \frac{\mathbf{P}(h) - \mathbf{I}}{h} \tag{5}$$

Where **P**(h) denotes transition probability matrix at time h. Then the transition probability matrix at any time $t \geq 0$ might be computed as follows. (Karlin and Taylor (1975)).

$$\mathbf{P}(t) = e^{\mathbf{A}t} \tag{6}$$

Therefore a special procedure should be adopted to find the transition probability matrix $\mathbf{P}^{(t)}$ at any time t from discrete Markov chain information. As it will be show later the adopted procedure coincides with transition probability matrix with complex elements.

II. Breaking the Time in Discrete Markov Chain

Consider again matrix **P** defined in (2). Also, assume **P** is of full rank.

Assumption 1: **P** is of full rank.

This assumption assures that all eigenvalues of **P** are nonzero and **P** is diagnalizable, Searle (1982), Dhrymes (1978). This assumption is not very restrictive, since; actually, most of Markov matrices have dominant diagonals. That is probability of transition from state i to itself is more than the sum of probabilities from state i to all other states. The matrices having dominant diagonals are non-singular, Takayama (1974). Threfore, **P** can be decomposed as follows (Searle (1982), Klein (1973)).

$$\mathbf{P} = \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1} \tag{7}$$

Where **X** is an NxN matrix of eigenvectors $\mathbf{x}_i$, $i = 1,...,N$,

$$\mathbf{X} = [\mathbf{x}_1,....,\mathbf{x}_N] \tag{8}$$

and $\boldsymbol{\Lambda}$ the NxN diagonal matrix of corresponding eigenvalues,

$$\boldsymbol{\Lambda} = diag\{\lambda_1,...,\lambda_N\} \tag{9}$$

Using (7), (8) and (9) to break n-step transition probability matrix **P** to any smaller period of time $t \geq 0$, we do as follows. If $t_i \leq 0$ for all i∈{1,…,K}are fractions of n–step period and $\sum_{i=1}^{k} t_i = n$ for any n belonging to natural numbers then,

$$\mathbf{P}^n = \prod_{j=1}^{k} \mathbf{P}^{t\,i} = \mathbf{P}^{\sum_{i=1}^{k} t_i} \tag{10}$$

On the other hand, transition probability matrix of n-step can be broken to fractions of n, if sum of them is equal to n. Therefore, any $t \geq 0$ fraction of one-step transition probability matrix can be written as,

$$\mathbf{P}^t = \mathbf{X}\boldsymbol{\Lambda}^t\,\mathbf{X}^{-1} \tag{11}$$

where,

$$\boldsymbol{\Lambda}^t = diag\left\{\lambda_1^t\ ,....,\lambda_N^t\right\} \tag{12}$$

Before discussing on the nature of eigenvalues of **P** let us define the generalized Markov matrix.

Definition 1. Matrix **Q** is a generalized Markov matrix if the following conditions are fulfilled:

1) $q_{ij} \in C \qquad \forall\, i, j \in S$

2) $\text{Re}\,(q_{ij}) \in [0,1] \qquad \forall i, j \in S$

3) $\text{Im}(q_{ij}) \in [-1,1] \qquad \forall i, j \in S$

4) $\sum_{j=1}^{N} \text{Re}\,(q_{ij}) = 1 \qquad \forall i \in S$

5) $\sum_{j=1}^{N} \text{Im}(q_{ij}) = 0 \qquad \forall i \in S$

Remark 1. According to definition 1, matrix **Q** can be written as:

$$\mathbf{Q} = \mathbf{U} + i\mathbf{V} \tag{13}$$

Where **U** and **V** are NxN matrices of real and imaginary parts of **Q** with $i = \sqrt{-1}.$

Remark 2. Matrix **U** has all Properties of **P** defined by (2), thus, **P**⊂ **Q**.

Treorem 1. If **P** is a Markov matrix then $\mathbf{P}^t$ also satisfies Markovian properties.

Proof: According to Chapman–Kolmogorov relation for continuous Markov chain (Karlin and Taylor (1975)), we have

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s) \qquad t, s \geq 0 \tag{14}$$

That is, if **P**(t) and **P**(s), transition probability matrices at times t and s are Markovs, then the product of them **P**(t+s) is also Markov. Let t=1, then **P**(1) is a one-step transition probability matrix which is equivalent to (2). Hence, our discrete Markov matrix **P** is equivalent to its continuous analogue **P**(1). So

$$\mathbf{P} = \mathbf{P}(1) \tag{15}$$

If we show that

$$\mathbf{P}^t = \mathbf{P}(t) \tag{16}$$

Then according to (14)

$$\mathbf{P}^{t+s} = \mathbf{P}^t\mathbf{P}^s \tag{17}$$

We can conclude that if **P** is Markov then $\mathbf{P}^t$, $\mathbf{P}^s$ and $\mathbf{P}^{t+s}$ are also Markovs for $t, s \geq 0$ and the theorem is proved.

Rewrite **P**(t) in (6) as (18).

$$\mathbf{P}(t) = \mathbf{X}\boldsymbol{\Lambda}(t)\mathbf{X}^{-1} \tag{18}$$

Where λ_i , $i \in S$ are the eigenvalues of **A** defined by (5), and

$$\boldsymbol{\Lambda}(t) = diag\left\{\exp(\lambda_1^t),...,\exp(\lambda_N^t)\right\} \tag{19}$$

And **X** is the corresponding eigenmatrix of **A**. Take the natural logarithm of (18),

$$ln\,\mathbf{P}(t) = \mathbf{X}\boldsymbol{\Phi}(t)\mathbf{X}^{-1} \tag{20}$$

Where,

$$\Phi(t)=t\ diag\left\{\lambda_1,...,\lambda_N\right\} \quad (21)$$

So,

$$ln\,\mathbf{P}(t)=t\ \mathbf{X}\boldsymbol{\psi}\mathbf{X}^{-1} \quad (22)$$

where

$$\boldsymbol{\psi}=diag\left\{\lambda_1,...,\lambda_N\right\} \quad (23)$$

Write (22) for t=1 and multiply both side by t,

$$t\ ln\,\mathbf{P}(1)=t\ \mathbf{X}\boldsymbol{\psi}\mathbf{X}^{-1} \quad (24)$$

By comparison of (22) and (24) conclude that

$$ln\,\mathbf{P}(t)=t\ ln\,\mathbf{P}(1) \quad (25)$$

or,

$$\mathbf{P}(t)=\mathbf{P}^t(1) \quad (26)$$

Given (15), equation (26) is the same as (16) Q.E.D.

Result 1. Matrix $\mathbf{P}^t$ fulfils definition 1. Thus, $\mathbf{P}^t\subseteq\mathbf{Q}$. This comes from the following remarks.

Remark 3. Sum of each row of $\mathbf{P}^t$ is equal to one. Since $\mathbf{P}^t$ satisfies Markovian properties (theorem 1).

Remark 4. Sum of imaginary parts of each row is equal to zero. This immediately comes from remark 3.

Remark 5. If q_{ij} denotes the ijth element of $\mathbf{P}^t$ for $t\geq 0$, then $\left|q_{ij}\right|\leq 1$ for all i and j belonging to S. This remark can be concluded form theorem 1.

Remark 6. If $\mathbf{Q}=\mathbf{P}^t, t\geq 0$ equals to the complex matrix defined by (13), then $\left|V_{jk}\right|\leq 1 \quad \forall j,k\in S$. Since,

$$1\geq\left|q_{jk}\right|=\left|u_{jk}+iv_{jk}\right|\Rightarrow 1\geq\sqrt{u^2{}_{jk}+iv^2{}_{jk}}$$

$$1\geq u_{jk}^2+v_{jk}^2\Rightarrow 1\geq\left|v_{jk}\right|.$$

Remark 7. Given **Q** as in remark 6, then $u_{jk}\in[0,1]$. This also comes immediately from theorem 1.

III. Discussion on Broken Times

The broken time discrete Markov chain is not always a complex probability matrix defined by definition 1. Matrix $\mathbf{P}^t$ has different properties with respect to t and eigenvalues. λ_i may be real (positive or negative) or complex depending on the characteristic polynomial of **P**.

Since **P** is a non–negative matrix, Forbenius theorem (Takayama (1974), Nikaido (1970)) assures that **P** has a positive dominant eigenvalue

$$\lambda_1\succ 0 \quad \text{(Frobenius root)} \quad (27)$$

and

$$\left|\lambda_i\right|\leq\lambda_1 \qquad \forall i\in\left\{2,...,N\right\} \quad (28)$$

Furthermore, if **P** is also a Markov matrix then its Frobenius root is equal to one, (Bellman (1970), Takayama (1974)). Therefore,

$$\lambda_1=1 \quad (29)$$

$$\left|\lambda_i\right|\leq 1 \qquad \forall i\in S \quad (30)$$

With the above information, consider the following discussions.

$a)\ \lambda_i\in(0,1] \qquad \forall i\in S$

In this case all $\lambda_i^t\geq 0$ for $t\geq 0$ and no imaginary part occurs in matrix $\mathbf{P}^t$. λ_i are all positive for i belonging to S if we can decompose the matrix **P** to two positive semi-definite and positive definite matrices **B** and **C** of the same size (Mardia, Kent, Bibby (1982)) as

$$\mathbf{P}=\mathbf{C}^{-1}\mathbf{B}$$

$b)\ \lambda_i\in[-1,1], \lambda_i\neq 0, \forall i\in S$

λ_i^t , $t\geq 0$ belongs to sets of real and imaginary numbers based on the value of t. In this case $\mathbf{P}^t$ belongs to the class of generalized stochastic matrix **Q** of definition 1. For $\lambda_i\in R$, it is sufficient that **P** be positive definite.

$c)\ \lambda_i\in C, \left|\lambda_i\right|\in(0,1]\ \forall i\in S$

$\mathbf{P}^t$ in this case for $t\geq 0$ and $t\notin N$ belongs to the class of generalized Markov matrices of definition 1.

$d)\ t\in N$ (Natural numbers)

In all cases of a, b, and c we never coincide with complex probabilities. Since $\mathbf{P}^t$ can be drived by simply multiplying **P**, t times.

$e)\ t\in Z$ (Integer numbers)

In this case, $\mathbf{P}^t$ is a real matrix but does not always satisfy condition 2 of definition 1.

$f)\ t\in R^-$

$\mathbf{P}^t$ is a complex matrix but does always satisfy conditions 2 and 3 of definition 1.

IV. Complex Probability Justification

Interpretation of the "Complex probability" as defined by definition 1 is not very simple and needs more elaborations. The interesting problem is that, it exists in operational works of statistics as the example 1 discussed. Many similar examples like the cited may be gathered.

With this definition of probability, the moments of a real random variable are complex. Although the t–step distribution π_t of initial distribution π_0 with respect to $\mathbf{P}^t$ may be complex, they have the same total as π_0. That is, if

$$\boldsymbol{\pi}_0 = (\pi_{01}, \ldots, \pi_{0N}) \tag{32}$$

Then,

$$\boldsymbol{\pi}_t = \boldsymbol{\pi}_o \mathbf{P}^t = \boldsymbol{\pi}_o \mathbf{Q} = \boldsymbol{\pi}_o \mathbf{U} + i\, \boldsymbol{\pi}_o \mathbf{V} \tag{33}$$

And we have the following remark accordingly,

Remark 8. Sum of t-step distribution is equal to sum of initial distribution. That is,

$$\sum_{j=1}^{N} \pi_{oj} = \sum_{j=1}^{N} \pi_{tj} \tag{34}$$

This can be derived based on (32) and (33) as

$$(\pi_{t1}, \ldots, \pi_{tN}) = \left(\sum_{j=1}^{N} \pi_{oj} U_{j1}, \ldots, \sum_{j=1}^{N} \pi_{oj} U_{jn}\right) + i \left(\sum_{j=1}^{N} \pi_{oj} v_{j1}, \ldots, \sum_{j=1}^{N} \pi_{oj} v_{jN}\right) \tag{35}$$

And, sum of t–step distribution is

$$\sum_{j=1}^{N} \pi_{tj} = \sum_{j=1}^{N} \pi_{oj} (u_{j1} + \ldots + u_{jN}) + i \sum_{j=1}^{N} \pi_{oj} (v_{j1}, \ldots, v_{jn}) \tag{36}$$

The two parentheses in (36) are one and zero respectively based on conditions 4 and 5 of definition 1. Thus, (36) and (34) are the same.

The above remark 8 states that though there exists imaginary transition probabilities to move from state j to k, the total sum of "imaginary transitions" is equal to zero. On the other hand, after t^{th} step transition, the total distribution has no imaginary part.

V. Summary

By summarizing the discrete and continuous times Markov stochastic processes a class of real world problems was introduced which cannot be solved by each of the procedures. The solutions of these problems coincide with "Complex probabilities" of transitions that are inherent in mathematical formulation of the model. Complex probability is defined and some of its properties with respect to the cited class are examined. Justification of the idea of complex probability needs more work and is left for further research.

Acknowledgements

The authors are indebted to Dr. A. Monajemi who read the manuscript and gave valuable remarks.

References

[1] R. Bellman (1970), Introduction to matrix analysis, McGraw–Hill.

[2] P.J. Dhrymes (1978), Mathematics for econometrics. Springer-Verlag.

[3] W. Feller (1970, 1971), An Introduction to probability theory and its applications, vols. 1,2, Wiley, New York.

[4] P.G. Hoel, S.C. Port, C. J. Stone (1972), Introduction to stochastic processes. Houghton Mifflin, New York.

[5] S. Karlin, H.M.Taylor (1975), A first course in stochastic processes. Academic Press.

[6] E. Klein (1973), Mathematical methods in theoretical economics, topological and vector space foundations of equilibrium analysis, Academic Press.

[7] K.V. Mardia, J.T. Kent, J.M. Bibby (1982), Multivariate analysis, Academic Press.

[8] H. Nikaido (1970), Introduction to sets and mapping in modern economics. North Holland Publishing Co.

[9] S.S. Searle (1982), Matrix algebra useful for statistics. Wiley.

[10] A.Takayama (1974) Mathematical economics. Dyrden Press, Illinois.

[11] E. Wentzel, L. Ovcharov (1986) Applied problems in probability theory. Mir, Moscow.

Comparison of Homotopy Perturbation Sumudu Transform method and Homotopy Decomposition method for solving nonlinear Fractional Partial Differential Equations

[1]Rodrigue Batogna Gnitchogna, [2]Abdon Atangana,

***Abstract*—**We paid attention to the methodology of two integral transform methods for solving nonlinear fractional partial differential equations. On one hand, the Homotopy Perturbation Sumudu Transform Method (HPSTM) is the coupling of the Sumudu transform and the HPM using He's polynomials. On the other hand, the Homotopy Decomposition Method (HDM) is the coupling of Adomian Decomposition Method and Perturbation Method. Both methods are very powerful and efficient techniques for solving different kinds of linear and nonlinear fractional differential equations arising in different fields of science and engineering. However, the HDM has an advantage over the HPSTM which is that it solves the nonlinear problems using only the inverse operator which is basically the fractional integral. Additionally there is no need to use any other inverse transform to find the components of the series solutions as in the case of HPSTM. As a consequence the calculations involved in HDM are very simple and straightforward.

.

***Keywords*—**Homotopy decomposition method,Integral transforms,nonlinear fractional differential equation, Sumudu transform.

I. Introduction

Fractional Calculus has been used to model physical and engineering processes, which are found to be best described by fractional differential equations. It is worth nothing that the standard mathematical models of integer-order derivatives, including nonlinear models, do not work adequately in many cases. In the recent years, fractional calculus has played a very important role in various fields an excellent literature of this can be found in [1-10]. However, analytical solutions of these equations are quickly difficult to find.

One can find in the literature a wide class of methods dealing with approximate solutions to problems described by nonlinear fractional differential equations, asymptotic and perturbation methods for instance. Perturbation methods carry among others the inconvenient that approximate solutions engage series of small parameters which cause difficulties since most nonlinear problems have no small parameters at all. Even though a suitable choice of small parameters occasionally lead to ideal solution, in most cases unsuitable choices lead to serious effects in the solutions. Therefore, an analytical method which does not require a small parameter in the equation modelling the phenomenon is welcome. To deal with the pitfall presented by perturbation methods for solving nonlinear equations, we present a literature review in some new asymptotic methods aiming for the search of solitary solutions of nonlinear differential equations, nonlinear differential-difference equations, and nonlinear fractional differential equations; see in [11]. The homotopy perturbation method (HPM) was first initiated by He [12]. The HPM was also studied by many authors to present approximate and exact solution of linear and nonlinear equations arising in various scientific and technological fields [13–23]. The Adomian decomposition method (ADM) [24] and variational iteration method (VIM) [25] have also been applied to study the various physical problems. The Homotopy decomposition method (HDM) was recently proposed by [26-27] to solve the groundwater flow equation and the modified fractional KDV equation [26-27]. The Homotopy decomposition method is actually the combination of perturbation method and Adomian decomposition method. Singh et al. [28] studied solutions of linear and nonlinear partial differential equations by using the homotopy perturbation Sumudu transform method (HPSTM). The HPSTM is a combination of Sumudu transform, HPM, and He's polynomials.

II. Sumudu Transform

The Sumudu transform, is an integral transform similar to the Laplace transform, introduced in the early 1990s by Gamage K. Watugala [29] to solve differential equations and control engineering problems. It is equivalent to the Laplace-Carson transform with the substitution $p = 1/u$. Sumudu is a Sinhala word, meaning "smooth". The Sumudu transform of a function $f(t)$, defined for all real numbers $t \geq 0$, is the function $F_s(u)$, defined by: (2.1).

[1]Department of Pure and Applied Mathematics University of Namibia Bag 13301,340 MandumeNdemufayo Ave Windhoek, Namibia, E-mail: rbatogna@yahoo.fr

[2]Institute for Groundwater Studies, University of the Free State, P0 Box 399, Bloemfontein, South Africa, Email address: abdonatangana@yahoo.fr

$$S(f(t)) = F_s(u) = \int_0^\infty \frac{1}{u} exp\left[-\frac{t}{u}\right] f(t)dt$$

A. Properties of Sumudu Transform [30-33]

- The transform of a Heaviside unit step function is a Heaviside unit step function in the transformed domain.
- The transform of a Heaviside unit ramp function is a Heaviside unit ramp function in the transformed domain.
- The transform of a monomial t^n is the called monomial $S(t^n) = n!\,u^n$.
- If $f(t)$ is a monotonically increasing function, so is $F(u)$ and the converse is true for decreasing functions.
- The Sumudu transform can be defined for functions which are discontinuous at the origin. In that case the two branches of the function should be transformed separately.
- If $f(t)$ is C^ncontinuous at the origin, so is the transformation $F(u)$.
- The limit of $f(t)$ as t tends to zero is equal to the limit of $F(u)$ as u tends to zero provided both limits exist.
- The limit of $f(t)$ as t tends to infinity is equal to the limit of $F(u)$ as u tends to infinity provided both limits exist.
- Scaling of the function by a factor $c > 0$ to form the function $f(ct)$ gives a transform $F(cu)$ which is the result of scaling by the same factor.

.

III. Basic Definition of Fractional Calculus

Definition 1 A real function $f(x), x > 0$, is said to be in the spaceC_μ, μ ℝ if there exists a real number p > μ, such that f(x) = x^ph(x), where $h(x)\epsilon\, C\,[0,\infty)$, and it is said to be in space C_μ^m if $f^{(m)}\epsilon C_\mu$, m∈ℕ

Definition 2 The Riemann-Liouville fractional integral operator of order α≥0, of a function $f\epsilon C_\mu$, μ≥-1, is defined as

(3.1)

$$J^\alpha f(x) = \frac{1}{\Gamma(\alpha)}\int_0^x (x-t)^{\alpha-1} f(t)dt, \quad \alpha > 0, x > 0$$

$$J^0 f(x) = f(x).$$

Properties of the operator can be found in [1-4] we mention only the following:

For $f\epsilon C_\mu$, $\mu \geq -1$, α, $\beta \geq 0$ and $\gamma > -1$:

(3.2)

$$J^\alpha J^\beta f(x) = J^{\alpha+\beta} f(x),$$

$$J^\alpha J^\beta f(x) = J^\beta J^\alpha f(x) J^\alpha x^\gamma = \frac{\Gamma(\gamma+1)}{\Gamma(\alpha+\gamma+1)} x^{\alpha+\gamma}.$$

Lemma 1 If $m-1 < \alpha \leq m, m \in \mathbb{N} and f \in C_\mu^m, \mu \geq -1$, *then*

$$D^\alpha J^\alpha f(x) = f(x) and,$$

$$J^\alpha D_0^\alpha f(x) = f(x) - \sum_{k=0}^{m-1} f^{(k)}(0^+)\frac{x^k}{k!}\ ,\ x > 0 \quad (3.3)$$

Definition 3: Partial Derivatives of Fractional order

Assume now that $f(\boldsymbol{x})$ is a function of n variables $x_i i = 1, \ldots\ldots, n$also of class Con $D \in \mathbb{R}_n$. As an extension of definition 3 we define partial derivative of order α for $f(x)$ respect to x_i the function

(3.4)

$$a\partial_{\underline{x}}^\alpha f = \frac{1}{\Gamma(m-\alpha)}\int_a^{x_i} (x_i - t)^{m-\alpha-1}\, \partial_{x_i}^m f(x_j)|_{x_j=t} dt$$

If it exists, where $\partial_{x_i}^m$ is the usual partial derivative of integer order m.

Definition 4: The Sumudu transform of the Caputo fractional derivative is defined as follows [30-33]:

(3.5)

$$S[D_t^\alpha f(t)] = u^{-\alpha} S[f(t)] - \sum_{k=0}^{m-1} u^{-\alpha+k} f^{(k)}(0^+), (m-1 < \alpha \leq m)$$

IV. Solution by (HPSTM) and (HDM)

V.I. Basic Idea of HPSTM

We illustrate the basic idea of this method, by considering a general fractional nonlinear non-homogeneous partial differential equation with the initial condition of the form of general form:

$$D_t^\alpha U(x,t) = L(U(x,t)) + N(U(x,t)) + f(x,t), \quad \alpha > 0$$

(4.1)

subject to the initial condition

$$D_0^k U(x,0) = g_k, \qquad (k = 0, \dots \dots n-1), D_0^n U(x,0) = 0 \; and \; n = [\alpha]$$

where, D_t^α denotes without loss of generality the Caputo fraction derivative operator, f is a known function, N is the general nonlinear fractional differential operator and L represents a linear fractional differential operator.

Applying the Sumudu Transform on Both sides of equation (4.1), we obtain:

$$S[D_t^\alpha U(x,t]) = S[L\big(U(x,t)\big)] + S[N\big(U(x,t)\big)] + S[f(x,t)] \quad (4.2)$$

Using the property of the Sumudu transform, we have

$$S[U(x,t)] = u^\alpha S\big[L\big(U(x,t)\big)\big] + u^\alpha S\big[N\big(U(x,t)\big)\big] + u^\alpha S[f(x,t)] + g(x,t) \quad (4.3)$$

Now applying the Sumudu inverse on both sides of (4.3) we obtain:

$$U(x,t) = S^{-1}\Big[u^\alpha S\big[L\big(U(x,t)\big)\big] + u^\alpha S\big[N\big(U(x,t)\big)\big]\Big] + G(x,t) \quad (4.4)$$

$G(x,t)$represents the term arising from the known function $f(x,t)$ and the initial conditions.

Now we apply the HPM:

(4.5)

$$U(x,t) = \sum_{n=0}^{\infty} p^n U_n(x,t)$$

The nonlinear tern can be decomposed

(4.6)

$$NU(x,t) = \sum_{n=0}^{\infty} p^n \mathcal{H}_n(U)$$

using the He's polynomial$\mathcal{H}_n(U)$ [22] given as:

(4.7)

$$\mathcal{H}_n(U_0, \dots \dots \dots, U_n) = \frac{1}{n!}\frac{\partial^n}{\partial p^n}\left[N\left(\sum_{j=0}^{\infty} p^j U_j(x,t)\right)\right], n = 0,1,2 \dots \dots \dots$$

Substituting (4.5) and (4.6)$\sum_{n=0}^{\infty} p^n U_n(x,t) = G(x,t) + p\Big[S^{-1}\big[u^\alpha S[L(\sum_{n=0}^{\infty} p^n U_n(x,t))] + u^\alpha S[N(\sum_{n=0}^{\infty} p^n U_n(x,t))]\big]\Big]$(4.8)

which is the coupling of the Sumudu transform and the HPM using He's polynomials. Comparing the coefficients of like powers of$p$, the following approximations are obtained.

p^0: $U_0(x,t) = G(x,t)$,

p^1: $U_1(x,t) = S^{-1}\Big[u^\alpha S\big[L\big(U_0(x,t)\big) + H_0(U)\big]\Big]$,

p^2: $U_2(x,t) = S^{-1}\Big[u^\alpha S\big[L\big(U_1(x,t)\big) + H_1(U)\big]\Big]$,

(4.9)

p^3: $U_3(x,t) = S^{-1}\Big[u^\alpha S\big[L\big(U_2(x,t)\big) + H_2(U)\big]\Big]$,

p^n: $U_n(x,t) = S^{-1}\Big[u^\alpha S\big[L\big(U_{n-1}(x,t)\big) + H_{n-1}(U)\big]\Big]$,

Finally, we approximate the analytical solution $U(x,t)$by the truncated series:

(4.10)

$$U(x,t) = \lim_{N\to\infty} \sum_{n=0}^{N} U_n(x,t)$$

The above series solution generally converges very rapidly [33]

V.II. Basic Idea of HDM [26-27]

The method first step here is to transform the fractional partial differential equation to the fractional partial integral equation by applying the inverse operator $D_t^\alpha$of on both sides of equation (4.1) to obtain: (4.11)

$$U(x,t) = \sum_{j=1}^{n-1} \frac{g_j(x)}{\Gamma(\alpha - j + 1)} t^j + \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} [L(U(x,\tau)) + N(U(x,\tau)) + f(x,\tau)] d\tau$$

Or in general by putting

$$\sum_{j=1}^{n-1} \frac{f_j(x)}{\Gamma(\alpha - j + 1)} t^{\alpha-j} = f(x,t) \; or \; f(x,t) = \sum_{j=1}^{n-1} \frac{g_j(x)}{\Gamma(\alpha - j + 1)} t^j$$

We obtain:

(4.12)

$$U(x,t) = T(x,t) + \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} [L(U(x,\tau)) + N(U(x,\tau)) + f(x,\tau)] d\tau$$

In the homotopy decomposition method, the basic assumption is that the solutions can be written as a power series in p

$$U(x,t,p) = \sum_{n=0}^{\infty} p^n U_n(x,t) \quad (4.13)$$

$$U(x,t) = lim_{p\to 1}\, U(x,t,p) \qquad (4.14)$$

and the nonlinear term can be decomposed as

$$NU(x,t) = \sum_{n=0}^{\infty} p^n \mathcal{H}_n(U) (4.14)$$

where$p \in (0,\ 1]$ is an embedding parameter. $\mathcal{H}_n(U)$ [22] is the He's polynomials that can be generated by

(4.16)

$$\mathcal{H}_n(U_0, \cdots\cdots\cdots, U_n) = \frac{1}{n!}\frac{\partial^n}{\partial p^n}\left[N\left(\sum_{j=0}^{\infty} p^j U_j(x,t)\right)\right], n = 0,1,2\cdots\cdots\cdots$$

The homotopy decomposition method is obtained by the graceful coupling of homotopy technique with Abel integral and is given by (4.17)

$$\sum_{n=0}^{\infty} p^n U_n(x,t) - T(x,t) = \frac{p}{\Gamma(\alpha)}\int_0^t (t-\tau)^{\alpha-1}\left[f(x,\tau) + L\left(\sum_{n=0}^{\infty} p^n U_n(x,\tau)\right) + N\left(\sum_{n=0}^{\infty} p^n U_n(x,\tau)\right)\right] d\tau$$

Comparing the terms of same powers of $p$gives solutions of various orders with the first term:

$$U_0(x,t) = T(x,t) \qquad (4.18)$$

It is worth noting that, the term $T(x,t)$ is the Taylor series of the exact solution of equation (4.1) of order $n-1$.

V. Applications

In this section we solve some popular nonlinear partial differential equation with both methods.

Example 1:

Let consider the following one-dimensional fractional heat-like problem: (5.1)

$$D_t^{\alpha} u(x,t) = \frac{1}{2}x^2 u_{xx}(x,t), 0 < x < 1, 0 < \alpha \le 1, t > 0$$

Subject to the boundary condition:

$$u(0,t) = 0, \qquad u(1,t) = \exp[t]$$

and initial condition $u(x,0) = x^2$

Example 2

Consider the following time-fractional derivative in x,y −plane as (5.2)

$$D_t^{\alpha} u(x,y,t) = \frac{1}{2}\nabla^2 u(x,y,t), 1 < \alpha \le 2, x,y \in \mathbb{R}, t > 0$$

subject to the initial conditions (5.3)

$$u(x,y,0) = \sin(x+y), \qquad u_t(x,y,0) = -\cos(x+y)$$

Example 3

Consider the following nonlinear time-fractional gas dynamics equations [Kilicman]

$$D_t^{\alpha} U + \tfrac{1}{2}(U^2)_x - U(1-U) = 0, 0 < \alpha \le 1, (5.4)$$

with the initial conditions

$$U(x,0) = \exp[-x] \qquad (5.5)$$

Example 4: Consider the following three-dimensional fractional heat-like equation

$$\partial_t^{\alpha} u(x,y,z,t) = x^4y^4z^4 + \tfrac{1}{36}\left(x^2u_{xx} + y^2u_{yy} + z^2u_{zz}\right), 0 < x,y,z < 1, 0 < \alpha \le 1 \quad (5.6)$$

Subject to the initial condition:

$$u(x,y,z,0) = 0 (5.7)$$

V.I. Solution via HPSTM

Example1: Apply the steps involved in HPSTM as presented in section 4.1 to equation (5.1) we obtain the following:

p^0: $u_0(x,t) = x^2$,

(5.8)

p^1: $u_1(x,t) = S^{-1}\left[u^{\alpha}S\left(\frac{1}{2}x^2(u_0)_{xx}\right)\right] = \frac{x^2t^{\alpha}}{\Gamma(\alpha+1)}$,

p^2: $u_2(x,t) = S^{-1}\left[u^{\alpha}S\left(\frac{1}{2}x^2(u_1)_{xx}\right)\right] = \frac{x^2t^{2\alpha}}{\Gamma(2\alpha+1)}$,

(5.9)

p^3: $u_3(x,t) = S^{-1}\left[u^{\alpha}S\left(\frac{1}{2}x^2(u_2)_{xx}\right)\right] = \frac{x^2t^{3\alpha}}{\Gamma(3\alpha+1)}$,

(5.10)

p^n: $u_n(x,t) = S^{-1}\left[u^{\alpha}S\left(\frac{1}{2}x^2(u_n)_{xx}\right)\right] = \frac{x^2t^{n\alpha}}{\Gamma(n\alpha+1)}$,

Therefore the series solution is given as:

(5.11)

$$u(x,t) = x^2\left[1 + \frac{t^{\alpha}}{\Gamma(\alpha+1)} + \frac{t^{2\alpha}}{\Gamma(2\alpha+1)} + \frac{t^{3\alpha}}{\Gamma(3\alpha+1)} + \cdots + \frac{t^{n\alpha}}{\Gamma(3\alpha+1)} + \cdots\right]$$

This equivalent to the exact solution in closed form:

(5.12)

$$u(x,t) = x^2 E_{1,\alpha}(t^\alpha)$$

where $E_{1,\alpha}()$ is the Mittag-Leffler function.

Example 2: Applying the steps involved in HPSTM as presented in section 4.1 to equation (5.2) we obtain:

p^0: $u_0(x,y,t) = \sin(x+y) - \cos(x+y)t,$

$p^1: u_1(x,t) = S^{-1}\left[u^\alpha S\left(\frac{1}{2}x^2\left[(u_0)_{xx} + (u_0)_{yy}\right]\right)\right] =$

$-\sin(x+y)\frac{t^2}{2} + \cos(x+y)\frac{t^3}{3!},$

$$p^2: u_2(x,t) = S^{-1}\left[u^\alpha S\left(\frac{1}{2}x^2\left[(u_1)_{xx} + (u_1)_{yy}\right]\right)\right]$$
$$= \sin(x+y)\left[-\frac{t^2}{2!} + \frac{t^4}{4!} + \frac{t^{4-\alpha}}{\Gamma(5-\alpha)}\right]$$
$$+ \cos(x+y)\left[-\frac{t^3}{3!} + \frac{t^5}{5!} + \frac{t^{5-\alpha}}{\Gamma(6-\alpha)}\right]$$

$p^3: u_3(x,t) = S^{-1}\left[u^\alpha S\left(\frac{1}{2}x^2\left[(u_2)_{xx} + (u_2)_{yy}\right]\right)\right] =$

$\sin(x+y)\left[-\frac{t^2}{2!} + \frac{t^4}{4!} + \frac{t^6}{6!} + \frac{2t^{4-\alpha}}{\Gamma(5-\alpha)} - \frac{2t^{6-\alpha}}{\Gamma(7-\alpha)} - \frac{4^{\alpha-2}\sqrt{\pi}t^{6-2\alpha}}{(6-2\alpha)(5-2\alpha)\Gamma(3-\alpha)\Gamma(2.5-\alpha)}\right] + \cos(x+y)\left[\frac{t^3}{3!} - \frac{t^5}{5!} + \frac{t^7}{7!} + \frac{2t^{7-\alpha}}{\Gamma(7-\alpha)} + \frac{2t^{7-2\alpha}}{\Gamma(8-2\alpha)}\right],$

Therefore the series solution is given as:

(5.13)

$$u(x,y,t) = \sin(x+y)\left[1 - \frac{3t^2}{2!} + \frac{t^4}{8} + \frac{t^6}{6!} + \frac{3t^{4-\alpha}}{\Gamma(5-\alpha)} - \frac{2t^{6-\alpha}}{\Gamma(7-\alpha)} - \frac{4^{\alpha-2}\sqrt{\pi}t^{6-2\alpha}}{(6-2\alpha)(5-2\alpha)\Gamma(3-\alpha)\Gamma(2.5-\alpha)}\right] + \cos(x+y)\left[-t + \frac{t^3}{3!} - \frac{t^5}{5!} + \frac{t^7}{7!} + \frac{3t^{7-\alpha}}{\Gamma(7-\alpha)} + \frac{t^{7-2\alpha}}{\Gamma(8-2\alpha)}\right] + \cdots$$

It is important to point out that if $\alpha = 2$ the above solution takes the form: (5.14)

$$u_{N=4}(x,y,t) = \sin(x+y)\left[1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \frac{t^6}{6!}\right] - \cos(x+y)\left[t - \frac{t^3}{3!} + \frac{t^5}{5!} - \frac{t^7}{7!}\right]$$

which is the first four terms of the series expansion of the exact solution $u(x,y,t) = \sin(x+y-t)$

Example 3: Apply the steps involved in HPSTM as presented in section 4.1 to equation (5.4) Kilicman et al [33] obtained the following:

p^0: $u_0(x,t) = exp(-x),$

$p^1: u_1(x,t) = S^{-1}\left[u^\alpha S\left(\frac{1}{2}x^2(u_0)_{xx}\right)\right] = \frac{exp(-x)t^\alpha}{\Gamma(\alpha+1)},$

$p^2: u_2(x,t) = S^{-1}\left[u^\alpha S\left(\frac{1}{2}x^2(u_1)_{xx}\right)\right] = \frac{exp(-x)t^{2\alpha}}{\Gamma(2\alpha+1)},$

$p^3: u_3(x,t) = S^{-1}\left[u^\alpha S\left(\frac{1}{2}x^2(u_2)_{xx}\right)\right] = \frac{exp(-x)t^{3\alpha}}{\Gamma(3\alpha+1)},$

(5.15)

$p^n: u_n(x,t) = S^{-1}\left[u^\alpha S\left(\frac{1}{2}x^2(u_n)_{xx}\right)\right] = \frac{exp(-x)t^{n\alpha}}{\Gamma(n\alpha+1)},$

Therefore the series solution is given as:

(5.16)

$$u(x,t) = exp(-x)\left[1 + \frac{t^\alpha}{\Gamma(\alpha+1)} + \frac{t^{2\alpha}}{\Gamma(2\alpha+1)} + \frac{t^{3\alpha}}{\Gamma(3\alpha+1)} + \cdots + \frac{t^{n\alpha}}{\Gamma(3\alpha+1)} + \cdots\right]$$

Example 4: Applying the steps involved in HPSTM as presented in section 4.1 to equation (5.2) we obtain:

$$p^0: u_0(x,t) = x^4y^4z^4$$

$$p^1: u_1(x,t) = S^{-1}\left[u^\alpha S\left(\frac{1}{36}\left(x^2(u_0)_{xx} + y^2(u_0)_{yy} + z^2(u_0)_{zz}\right)\right)\right] = \frac{t^\alpha x^4y^4z^4}{\Gamma(\alpha+1)}$$

$$p^2: u_2(x,t) = S^{-1}\left[u^\alpha S\frac{1}{36}\left(x^2(u_1)_{xx} + y^2(u_1)_{yy} + z^2(u_1)_{zz}\right)\right] = \frac{t^{2\alpha}x^4y^4z^4}{\Gamma(2\alpha+1)}p^3: u_3(x,t) = S^{-1}\left[u^\alpha S\left(\frac{1}{36}\left(x^2(u_2)_{xx} + y^2(u_2)_{yy} + z^2(u_2)_{zz}\right)\right)\right] = \frac{t^{3\alpha}x^4y^4z^4}{\Gamma(3\alpha+1)}$$

(5.17)

$$p^n: u_n(x,t) = S^{-1}\left[u^\alpha S\left(\frac{1}{36}\left(x^2(u_{n-1})_{xx} + y^2(u_{n-1})_{yy} + z^2(u_{n-1})_{zz}\right)\right)\right] = \frac{t^{n\alpha}x^4y^4z^4}{\Gamma(n\alpha+1)}$$

Therefore the approximate solution of equation for the first n is given below as:

$$u_N(x,y,z,t) = \sum_{n=1}^{N}\frac{t^{n\alpha}x^4y^4z^4}{\Gamma(n\alpha+1)}. \quad (5.18)$$

V.II. Solution via HDM

Example 1: Apply the steps involved in HDM as presented in section 4.2 to equation (5.1) we obtain the following

(5.19)

$$\sum_{n=0}^{\infty} p^n \mathrm{u}_n(x,t) - x^2 = \frac{p}{\Gamma(\alpha)}\int_0^t (t-\tau)^{\alpha-1}\left[x^2\sum_{n=0}^{\infty} p^n \frac{\partial^2 u_n(x,\tau)}{\partial x^2}\right]d\tau$$

Comparing the terms of the same powers of p we obtain:

(5.20)

$$u_0(x,t) = x^2$$

$$u_1(x,t) = \frac{1}{\Gamma(\alpha)}\int_0^t (t-\tau)^{\alpha-1}\frac{x^2\partial^2 u_0(x,\tau)}{\partial x^2}d\tau = \frac{x^2 t^{\alpha}}{\Gamma(\alpha+1)}$$

$$u_2(x,t) = \frac{1}{\Gamma(\alpha)}\int_0^t (t-\tau)^{\alpha-1}\frac{x^2\partial^2 u_1(x,\tau)}{\partial x^2}d\tau = \frac{x^2 t^{2\alpha}}{\Gamma(2\alpha+1)}$$

$$u_3(x,t) = \frac{1}{\Gamma(\alpha)}\int_0^t (t-\tau)^{\alpha-1}\frac{x^2\partial^2 u_2(x,\tau)}{\partial x^2}d\tau = \frac{x^2 t^{3\alpha}}{\Gamma(2\alpha+1)}$$

$$u_n(x,t) = \frac{1}{\Gamma(\alpha)}\int_0^t (t-\tau)^{\alpha-1}\frac{x^2\partial^2 u_{n-1}(x,\tau)}{\partial x^2}d\tau = \frac{x^2 t^{n\alpha}}{\Gamma(n\alpha+1)}$$

The asymptotic solution is given by

$$u_N(x,t) = x^2\left[\frac{t^{\alpha}}{\Gamma(\alpha+1)} + \frac{t^{2\alpha}}{\Gamma(2\alpha+1)} + \frac{t^{3\alpha}}{\Gamma(3\alpha+1)} + \cdots + \frac{t^{N\alpha}}{\Gamma(N\alpha+1)} + \cdots\right]$$

$$\lim_{N\to\infty} u_N(x,t,\alpha) = u(x,t,\alpha) \quad (5.20)$$

$$\lim_{\substack{\alpha\to 1\\ N\to\infty}} u_N(x,t,\alpha) = x^2\exp(t)$$

This is the exact solution of (5.1) when $\alpha = 1$.

Example 2

Following the discussion presented earlier, applying the initial conditions and comparing the terms of the same power of p, integrating, we obtain the following solutions:

$$u_0(x,t) = \sin(x+y) - \cos(x+y)t$$

$$u_1(x,t) = -\sin(x+y)\frac{t^2}{2} + \cos(x+y)\frac{t^3}{3!}$$

$$u_2(x,t) = \sin(x+y)\left[-\frac{t^2}{2!} + \frac{t^4}{4!} + \frac{t^{4-\alpha}}{\Gamma(5-\alpha)}\right] + \cos(x+y)\left[-\frac{t^3}{3!} + \frac{t^5}{5!} + \frac{t^{5-\alpha}}{\Gamma(6-\alpha)}\right] \quad (5.21)$$

$$u_3(x,t) = \sin(x+y)\left[-\frac{t^2}{2!} + \frac{t^4}{4!} + \frac{t^6}{6!} + \frac{2t^{4-\alpha}}{\Gamma(5-\alpha)} - \frac{2t^{6-\alpha}}{\Gamma(7-\alpha)} - \frac{4^{\alpha-2}\sqrt{\pi}t^{6-2\alpha}}{(6-2\alpha)(5-2\alpha)\Gamma(3-\alpha)\Gamma(2.5-\alpha)}\right] + \cos(x+y)\left[\frac{t^3}{3!} - \frac{t^5}{5!} + \frac{t^7}{7!} + \frac{2t^{7-\alpha}}{\Gamma(7-\alpha)} + \frac{2t^{7-2\alpha}}{\Gamma(8-2\alpha)}\right]$$

Using the package Mathematica, in the same manner one can obtain the rest of the components. But for four terms were computed and the asymptotic solution is given by:

(5. 22)

$$u(x,y,t) = \sin(x+y)\left[1 - \frac{3t^2}{2!} + \frac{t^4}{8} + \frac{t^6}{6!} + \frac{3t^{4-\alpha}}{\Gamma(5-\alpha)} - \frac{2t^{6-\alpha}}{\Gamma(7-\alpha)} - \frac{4^{\alpha-2}\sqrt{\pi}t^{6-2\alpha}}{(6-2\alpha)(5-2\alpha)\Gamma(3-\alpha)\Gamma(2.5-\alpha)}\right] + \cos(x+y)\left[-t + \frac{t^3}{3!} - \frac{t^5}{5!} + \frac{t^7}{7!} + \frac{3t^{7-\alpha}}{\Gamma(7-\alpha)} + \frac{t^{7-2\alpha}}{\Gamma(8-2\alpha)}\right] + \cdots$$

It is important to point out that if $\alpha = 2$ the above solution takes the form:

$$u_{N=4}(x,y,t) = \sin(x+y)\left[1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \frac{t^6}{6!}\right] - \cos(x+y)\left[t - \frac{t^3}{3!} + \frac{t^5}{5!} - \frac{t^7}{7!}\right]$$

Which are the first four terms of the series expansion of the exact solution$u(x,y,t) = \sin(x+y-t)$

Example 3:

p^0: $u_0(x,t) = exp(-x)$,

p^1: $u_1(x,t) = \frac{exp(-x)}{\Gamma(\alpha+1)}t^{\alpha}$,

p^2: $u_2(x,t) = \frac{exp(-x)}{\Gamma(2\alpha+1)}t^{2\alpha}$,

p^3: $u_3(x,t) = \frac{exp(-x)}{\Gamma(3\alpha+1)}t^{3\alpha}$,

(5.23)

⋮.

p^n: $u_n(x,t) = \frac{exp(-x)}{\Gamma(n\alpha+1)} t^{n\alpha}$,

Therefore the series solution is given as:

(5.24)

$$u(x,t) = exp(-x)\left[1 + \frac{t^{\alpha}}{\Gamma(\alpha+1)} + \frac{t^{2\alpha}}{\Gamma(2\alpha+1)} + \frac{t^{3\alpha}}{\Gamma(3\alpha+1)} + \cdots + \frac{t^{n\alpha}}{\Gamma(3\alpha+1)} + \cdots\right]$$

Example 4: Following carefully the steps involved in the HDM, we arrive at the following equations

$$\sum_{n=0}^{\infty} p^n u_n(x,y,z,t)$$

$$= \frac{p}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} \left(x^4y^4z^4 + \frac{1}{36}\left(x^2 \left(\sum_{n=0}^{\infty} p^n u_n(x,y,z,t) \right)_{xx} + y^2 \left(\sum_{n=0}^{\infty} p^n u_n(x,y,z,t) \right)_{yy} + z^2 \left(\sum_{n=0}^{\infty} p^n u_n(x,y,z,t) \right)_{zz} \right) \right) d\tau$$

(5.25)

Now comparing the terms of the same power of p yields:

p^0: $u_0(x,y,z,t)$(5.26)

$$p^1\text{: } u_1(x,y,z,t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} x^4y^4z^4 d\tau$$

$$\vdots$$

$$p^n\text{: } u_n(x,y,z,t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} \left(\frac{1}{36}\left(x^2 (u_{n-1})_{xx} + y^2 (u_{n-1})_{yy} + z^2 (u_{n-1})_{zz} \right) \right) d\tau, u_n(x,y,z,0) = 0, n \geq 2$$

Thus the following components are obtained as results of the above integrals

$$u_0(x,y,z,t) = 0$$

$$u_1(x,y,z,t) = \frac{t^{\alpha} x^4y^4z^4}{\Gamma(\alpha+1)}$$

$$u_2(x,y,z,t) = \frac{t^{2\alpha} x^4y^4z^4}{\Gamma(2\alpha+1)}$$

$$u_3(x,y,z,t) = \frac{t^{3\alpha} x^4y^4z^4}{\Gamma(3\alpha+1)}$$

$$\vdots$$

$u_n(x,y,z,t) = \frac{t^{n\alpha} x^4y^4z^4}{\Gamma(n\alpha+1)}$(5.27)

Therefore the approximate solution of equation for the first n is given below as:

$$u_N(x,y,z,t) = \sum_{n=1}^{N} \frac{t^{n\alpha} x^4y^4z^4}{\Gamma(n\alpha+1)}$$

Now when $N \to \infty$ we obtained the follow solution

(5.28)

$$u(x,y,z,t) = \sum_{n=0}^{\infty} \frac{t^{n\alpha} x^4y^4z^4}{\Gamma(n\alpha+1)} - x^4y^4z^4 = x^4y^4z^4(E_{\alpha}(t^{\alpha}) - 1)$$

Where $E_{\alpha}(t^{\alpha})$ is the generalized Mittag-Leffler function. Note that in the case $\alpha = 1$

$$u(x,y,z,t) = x^4y^4z^4(\exp(t) - 1) \quad (5.29)$$

This is the exact solution for this case.

VI. Comparison of Methods

This section is devoted to the comparison between the two integral transform methods.

The two methods are very powerful and efficient techniques for solving different kinds of linear and nonlinear fractional differential equations arising in different fields of science and engineering. However, it can be noted that the HDM has an advantage over the HPSTM which is that it solves the nonlinear problems using only the inverse operator which is simply the fractional integral. There is no need to use any other inverse transform to find the components of the series solutions as in the case of HPSTM. In addition the calculations involved in HDM are very simple and straightforward. In conclusion, the HDM and the HPSTM may be considered as a nice refinement in existing numerical techniques and might find wide applications.

Table 1: Numerical results of equation (5.2) via mathematica

			HPSTM and HDM	HPSTM and HDM	HPSTM and HDM		
t	x	y	α $= 1.25$	α $= 1.75$	α $= 2.0$	Exact	Errors
0.2	0.	0.	0.6436	0.6660	0.6816	0.6816	0
5	5	5	24	50	39	39	0
	0.	1.	0.9015	0.9294	0.9489	0.9489	0
	5	0	11	40	85	85	0
	1.	0.	0.9015	0.9294	0.9489	0.9489	
	0	5	11	40	85	85	
	1.	1.	0.9386	0.9652	0.9839	0.9839	
	0	0	76	71	86	86	
0.5	0.	0.	0.3665	0.4402	0.4794	0.4794	0.0000
	5	5	63	70	25	26	06
	0.	1.	0.6913	0.7888	0.8414	0.8414	0
	5	0	70	06	71	71	0
	1.	0.	0.6913	0.7888	0.8414	0.8414	0
	0	5	70	06	71	71	
	1.	1.	0.8469	0.9442	0.9974	0.9974	
	0	0	06	15	95	95	
0.7	0.	0.	0.0670	0.1925	0.2474	0.2474	0.0000
5	5	5	24	05	02	04	02
	1.	1.	0.4215	0.6004	0.6816	0.6816	0.0000
	0	0	20	31	36	39	03
	0.	0.	0.4215	0.6004	0.6816	0.6816	0.0000
	5	5	20	31	36	39	03
	1.	1.	0.6728	0.8613	0.9489	0.9489	0.0000
	0	0	13	1	82	85	03
1.0	0.	0.	-	-	-	0.0000	0.0000
	5	5	0.2082	0.0543	0.0000	00	19
	1.	1.	46	57	019	0.4794	0.0000
	0	0	0.1417	0.3857	0.4794	26	25
	0.	0.	97	25	01	0.4794	0.0000
	5	5	0.1417	0.3857	0.4794	26	25
	1.	1.	97	25	01	0.8414	0
	0	0	0.4571	0.7313	0.8414	48	
			23	69	48		

The approximate solution of equation (5.2) obtained by the present methods is close at hand to the exact solution. It is to be noted that only the fourth-order term of the HDM and HPSTM were used to evaluate the approximate solutions for Figures 1. It is evident that the efficiency of the present method can be noticeably improved by computing additional terms of $u(x, t)$when the HDM is used.

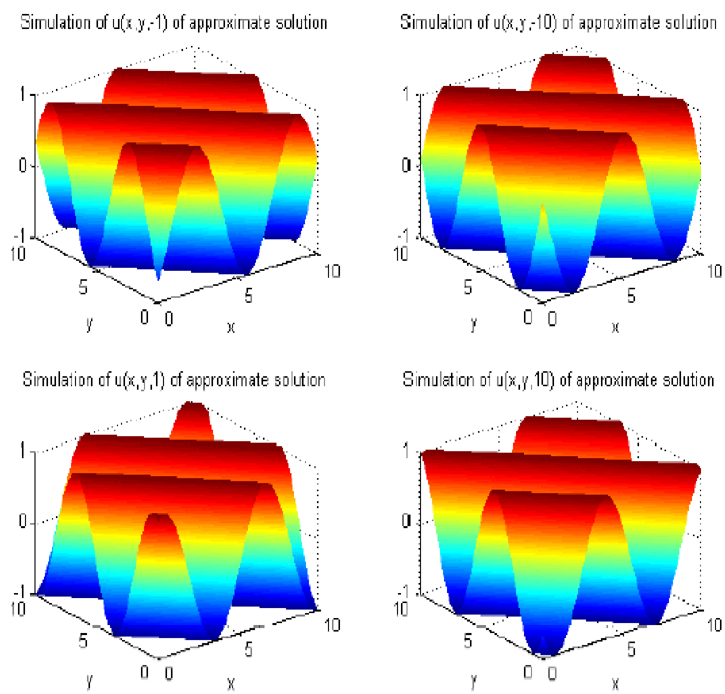


Figure 1: Numerical simulation of the approximated solution of equation (5.2)

VII. Conclusion

We studied two integral transform methods for solving fractional nonlinear partial differential equation. The first method namely homotopy perturbation Sumudu transform method is the coupling of the Sumudu transform and the HPM using He's polynomials. The second method namely Homotopy decomposition method is the combination of Adomian decomposition method and HPM using He's polynomials. These two methods are very powerful and efficient techniques for solving different kinds of linear and nonlinear fractional differential equations arising in different fields of science and engineering. However, the HDM has an advantage over the HPSTM which is that it solves the nonlinear problems using only the inverse operator which is simple the fractional integral. Also we do not need to use any order inverse transform to find the components of the series solutions as in the case of HPSTM. In addition the calculations involved in HDM are very simple and straightforward. In conclusion the HDM is a friendlier method.

ISBN: 978-1-61804-240-8

REFERENCES

[1] K. B. Oldham and J. Spanier, "The Fractional Calculus", Academic Press, New York, NY, USA, (1974).

[2] I. Podlubny, "Fractional Differential Equations", Academic Press, New York, NY, USA, (1999).

[3] A. A. Kilbas, H. M. Srivastava, and J. J. Trujillo, "Theory and Applications of Fractional Differential Equations", Elsevier, Amsterdam, The Netherlands, (2006).

[4] I. Podlubny, "Fractional Differential Equations", Academic Press, San Diego, Calif, USA, (1999).

[5] M. Caputo, "Linear models of dissipation whose Q is almost frequency independent, part II," Geophysical Journal International, vol. 13, no. 5, pp. 529–539, (1967).

[6] A. A. Kilbas, H. H. Srivastava, and J. J. Trujillo, "Theory and Applications of Fractional Differential Equations", Elsevier, Amsterdam, The Netherlands, (2006).

[7] K. S. Miller and B. Ross, "An Introduction to the Fractional Calculus and Fractional Differential Equations", Wiley, New York, NY, USA, (1993).

[8] S. G. Samko, A. A. Kilbas, and O. I. Marichev, "Fractional Integrals and Derivatives: Theory and Applications", Gordon and Breach, Yverdon, Switzerland, (1993).

[9] G. M. Zaslavsky, "Hamiltonian Chaos and Fractional Dynamics", Oxford University Press, (2005).

[10] A. Yildirim, "An algorithm for solving the fractional nonlinear Schrödinger equation by means of the homotopy perturbation method," International Journal of Nonlinear Sciences andNumerical Simulation, vol. 10, no. 4, (2009). pp. 445–450,

[11] J. H. He, "Asymptotic methods for solitary solutions and compactons," Abstract and Applied Analysis, vol. 2012, (2012). Article ID 916793, 130 pages,

[12] J.-H. He, "Homotopy perturbation technique," Computer Methods in Applied Mechanics and Engineering, vol. 178, no. 3-4, (1999), pp.257–262.

[13] J.-H. He, "Homotopy perturbation method: a new nonlinear analytical technique," Applied Mathematics and Computation, vol. 135, no. 1, (2003). pp. 73–79.

[14] J.-H. He, "New interpretation of homotopy perturbation method. Addendum: 'some asymptotic methods for strongly nonlinear equations'," International Journal of Modern PhysicsB, vol. 20, no. 18, (2006). pp. 2561–2568.

[15] D. D. Ganji, "The application of He's homotopy perturbation method to nonlinear equations arising in heat transfer," PhysicsLetters A, vol. 355, no. 4-5, (2006). pp. 337–341.

[16] A. Yildirim, "An algorithm for solving the fractional nonlinear Schrödinger equation by means of the homotopy perturbation method," International Journal of Nonlinear Sciences andNumerical Simulation, vol. 10, no. 4, (2009), pp. 445–450.

[17] D. D. Ganji and M. Rafei, "Solitary wave solutions for a generalized Hirota-Satsuma coupled KdV equation by homotopy perturbation method," Physics Letters A, vol. 356, no. 2, (2006)., pp. 131– 137.

[18] M. M. Rashidi, D. D. Ganji, and S. Dinarvand, "Explicit analytical solutions of the generalized Burger and Burger-Fisher equations by homotopy perturbation method," Numerical Methodsfor Partial Differential Equations, vol. 25, no. 2, (2009). pp. 409–417

[19] H. Aminikhah and M. Hemmatnezhad, "An efficient method for quadratic Riccati differential equation," Communications inNonlinear Science and Numerical Simulation, vol. 15, no. 4, (2010). pp. 835–839.

[20] S. H. Kachapi and D. D. Ganji, Nonlinear Equations: Analytical Methods and Applications, Springer, (2012).

[21] Atangana Abdon. "New Class of Boundary Value Problems", Inf. Sci. Lett. Vol 1 no. 2, (2012) pp 67-76

[22] Y. M. Qin and D.Q. Zeng, "Homotopy perturbation method for the q-diffusion equation with a source term," Communicationsin Fractional Calculus, vol. 3, no. 1, (2012). pp. 34–37,

[23] M. Javidi and M. A. Raji, "Combination of Laplace transform and homotopy perturbation method to solve the parabolic partial differential equations," Communications in FractionalCalculus, vol. 3, no. 1, , (2012).pp. 10–19.

[24] J. S. Duan, R. Rach, D. Buleanu, and A. M. Wazwaz, "A review of the Adomian decomposition method and its applications to fractional differential equations," Communications in FractionalCalculus, vol. 3, no. 2, (2012). pp. 73–99.

[25] D.D. Ganji, "Asemi-Analytical technique for non-linear settling particle equation of motion," Journal of Hydro-EnvironmentResearch, vol. 6, no. 4, (2012). pp. 323–327,

[26] A. Atangana and Aydin Secer. "Time-fractional Coupled-the Korteweg-de Vries Equations" Abstract Applied Analysis, In press (2013)

[27] Atangana A., Botha J.F. "Analytical solution of groundwater flow equation via Homotopy Decomposition Method, J Earth Sci Climate Change, vol 3, (2012) pp 115. doi:10.4172/2157-7617.1000115

[28] J. Singh, D. Kumar, and Sushila, "Homotopy perturbation Sumudu transform method for nonlinear equations,"

Advancesin Applied Mathematics and Mechanics, vol. 4, (2011), pp. pp 165–175.

[29] Watugala, G. K., “Sumudu transform: a new integral transform to solve differential equations and control engineering problems.” International Journal of Mathematical Education in Science and Technology vol 24, (1993), pp 35–43.

[30] Weerakoon, S., “Application of Sumudu transform to partial differential equations” International Journal of Mathematical Education in Science and Technology, vol 25 (1994), pp 277–283.

31-Asiru, M.A. “Classroom note: Application of the Sumudu transform to discrete dynamic systems” International Journal of Mathematical Education in Science and Technology, vol 34 no 6, (2003) , pages. 944-949

[32] Airu, M.A. “Further properties of the Sumudu transform and its applications” International Journal of Mathematical Education in Science and Technology, vol 33 no 3, (2002), pp. 441-449

[33] Jagdev Singh, Devendra Kumar, and A. Kılıçman “Homotopy Perturbation Method for Fractional Gas Dynamics Equation Using Sumudu Transform”. Abstract and Applied Analysis Volume 2013 (2013), pp 8

Noise Studies in Measurements and Estimates of Stepwise Changes in Genome DNA Chromosomal Structures

JORGE MUNOZ-MINJARES, YURIY S. SHMALIY, JESUS CABAL-ARAGON
Universidad de Guanajuato
Department of Electronics Engineering
Ctra. Salamanca-Valle, 3.5+1.8km, 36885, Salamanca
MEXICO
j.ulises_minjares@live.com, shmaliy@ugto.mx

Abstract: Measurements using the high resolution array-comparative genomic hybridization (HR-CGH) array are accompanied with large noise which strongly affects the estimates of the copy number variations (CNVs) and results in segmental errors as well as in jitter in the breakpoints. Based on the probabilistic analysis and algorithm designed, we show that jitter in the breakpoints can be well approximated with the discrete skew Laplace distribution if the local signal-to-noise ratios (SNRs) exceed unity. Using this distribution, we propose an algorithm for computing the estimate upper and lower bounds. Some measurements and estimates tested using these bounds show that the higher probe resolution is provided the more segmental accuracy can be achieved and that larger segmental SNRs cause smaller jitter in the breakpoints. Estimates of the CNVs combined with the bounds proposed may play a crucial role for medical experts to make decisions about true chromosomal changes and even their existence.

Key–Words: Genome copy number, estimate, jitter, breakpoint, error bound

1 Introduction

The deoxyribonucleic acid (DNA) of a genome essential for human life often demonstrates structural changes called copy-number variations (CNVs) associated with disease such as cancer [1]. The sell with the DNA typically has a number of copies of one or more sections of the DNA that results in the structural chromosomal rearrangements - deletions, duplications, inversions and translocations of certain parts [2]. Small such CNVs are present in many forms in the human genome, including single-nucleotide polymorphisms, small insertion-deletion polymorphisms, variable numbers of repetitive sequences, and genomic structural alterations [3]. If genomic aberrations involve large CNVs, the process was shown to be directly coupled with cancer and the relevant structural changes were called copy-number alterations (CNAs) [4]. A brief survey of types of chromosome alterations involving copy number changes is given in [5]. The copy number represents the number of DNA molecules in a cell and can be defined as the number of times a given segment of DNA is present in a cell. Because the DNA is usually double-stranded, the size of a gene or chromosome is often measured in base pairs. A commonly accepted unit of measurement in molecular biology is kilobase (kb) equal to 1000 base pairs of DNA [6]. The human genome with 23 chromosomes is estimated to be about 3.2 billion base pairs long and to contain $20000 - 25000$ distinct genes [1]. Each CNV may range from about one kb to several megabases (Mbs) in size [2].

One of the techniques employing chromosomal microarray analysis to detect the CNVs at a resolution level of 5–10 kbs is the array-comparative genomic hybridization (aCGH) [7]. It was reported in [8] that the high-resolution CGH (HR-CGH) arrays are accurate to detect structural variations (SV) at resolution of 200 bp. In microarray technique, the CNVs are often normalized and plotted as $\log_2 R/G = \log_2 \mathrm{Ratio}$, where R and G are the fluorescent Red and Green intensities, respectively [9]. An annoying feature of such measurements is that the Ratio is highly contaminated by noise which intensity does not always allow for correct visual identification of the breakpoints and copy numbers and makes most of the estimation techniques poor efficient if the number of segmental readings is small. It was shown in [10] that sufficient quality in the CNVs mapping can be achieved with tens of millions of paired reads of 29–36 bases at each. Deletions as small as 300 bp should also be detected in some cases. For instance, arrays with a 9-bp tiling path were used in [8] to map a 622-bp heterozygous deletion. So, further progress in the probe resolution

ISBN: 978-1-61804-240-8

of the CNVs measurements is desirable.

Typically, a chromosome section is observed with some average resolution $\bar{r}$, bp and M readings in the genomic location scale. The following distinct properties of the CNVs function were recognized [2, 5]:

1) It is piecewise constant (PWC) and sparse with a small number L of the *breakpoints* (edges) i_l, $l \in [1, L]$, on a long base-pair length. The breakpoints are places as $0 < i_1 < \cdots < i_L < \bar{r}M$ and can be united in a vector

$$\mathcal{I} = [i_1\, i_2 \ldots i_L]^T \in \mathcal{R}^L\,. \tag{1}$$

Sometimes, the genomic location scale is represented in the number of readings $n \in [1, M]$ with a unit step ignoring "bad" or empty measurements, where n represents the nth reading. In such a scale, the n_lth discrete point corresponds to the i_lth breakpoint in the genomic location scale and the points placed as $0 < n_1 < \cdots < n_L < M$ can be united in a vector

$$\mathcal{N} = [n_1\, n_2 \ldots n_L]^T \in \mathcal{R}^L\,. \tag{2}$$

An advantage of $\mathcal{N}$ against $\mathcal{I}$ is that it facilitates the algorithm design. However, the final estimates are commonly represented in the genomic location scale.

2) Its *segments* with constant copy numbers a_j, $j \in [1, L+1]$, are integer, although this property is not survived in the $\log_2 \mathrm{Ratio}$. The segmental constant changes can also be united in a vector

$$\mathbf{a} = [a_1\, a_2 \ldots a_{L+1}]^T \in \mathcal{R}^{L+1}\,, \tag{3}$$

in which a_j characterizes a segment between i_{j-1} and i_j on an interval $[i_{j-1}, i_j - 1]$.

3) The measurement noise in the $\log_2$ Ratio is highly intensive and can be modeled as additive white Gaussian.

The estimation theory offers several useful approaches for piecewise signals such as those generated by the chromosomal changes. One can employ the *wavelet*-based [11, 12] filters, *robust* estimators [12], adaptive *kernel smoothers* [13, 14], *maximum likelihood* (ML) based on Gauss's *ordinary least squares* (OLS), penalized *bridge* estimator [15] and *ridge* regression [16] (also known as Tikhonov regularization), fussed least-absolute shrinkage and selection operator (*Lasso*) [17], the *Schwarz information criterion*-based estimator [18, 19], and *forward-backward smoothers* [20–22].

We also find a number of solutions developed especially for needs of bioinformatics. Efficient algorithms for filtering, smoothing and detection were proposed in [11,12,19,23–28]. Methods for segmentation and modeling were developed in [10, 18, 24, 29–32].

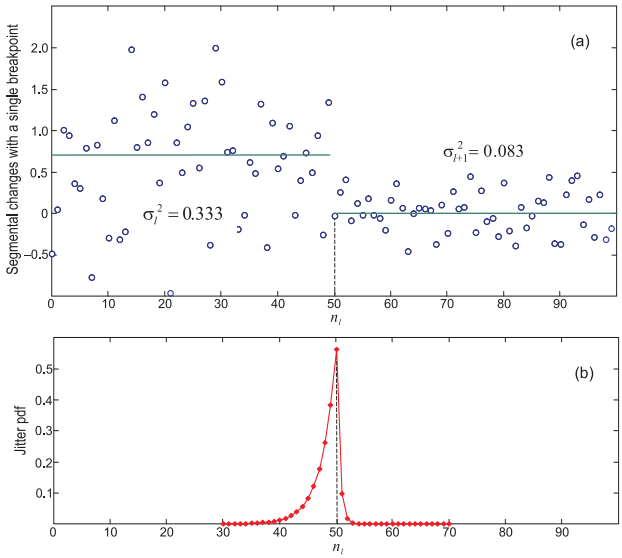


Figure 1: Simulated genome segmental changes with a single breakpoint at $n_l = 50$ and segmental variances $\sigma_l^2 = 0.333$ and $\sigma_{l+1}^2 = 0.083$ corresponding to segmental SNRs $\gamma_l = 1.47$ and $\gamma_{l+1} = 5.88$: (a) measurement and (b) jitter pdf. The jitter pdf was found by applying a ML estimator via a histogram over 10^4 runs.

Sparse representation based on penalized optimization and Bayesian learning were provided in [33–38]. These results show that a small number of readings N_j per a segment a_j in line with large measurement noise remain the main limiters of accuracy in the estimation of CNVs. Picard *et al.* have shown experimentally in [29] that each segmental estimate is accompanied with *errors* and each breakpoint has *jitter* which cannot be overcome by any estimator.

For clarity, we generalize an experiment conducted in [29] in Fig. 1. Here, a chromosomal part having two constant segments $a_l = 0.7$ and $a_{l+1} = 0$ and a breakpoint $n_l = 50$ is simulated in the presence of discrete white Gaussian noise having segmental variances $\sigma_l^2 = 0.333$ and $\sigma_{l+1}^2 = 0.083$ (Fig. 1a). For the local segmental signal-to-noise ratios (SNRs)

$$\gamma_l^- = \frac{\Delta_l^2}{\sigma_l^2}\,, \quad \gamma_l^+ = \frac{\Delta_l^2}{\sigma_{l+1}^2}\,, \tag{4}$$

where $\Delta_l = a_{l+1} - a_l$ is a local segmental change, it corresponds to $\gamma_l^- = 1.47$ and $\gamma_l^+ = 5.88$.

The breakpoint location n_l was detected in Fig. 1 using a ML estimator [22] (one can employ any other estimator). Measurements and estimations were repeated 10^4 times with different realization of noise. Then the histogram was plotted for the detected breakpoint locations and normalized to have a unit area. The jitter probability density function (pdf) obtained in such a way is sketched in Fig. 1b. Even a quick

ISBN: 978-1-61804-240-8

look at this figure assures that jitter at a level of 0.01 (jitter probability of 1%) has 10 points to the left (left jitter) and 2 points to the right (right jitter). In other words, with the probability of 99%, the breakpoint n_l can be found at any point between $n = 40$ and $n = 52$ that may be too rough for medical conclusions, especially if $\bar{r}$ is large. Let us add that simple averaging which is optimal for the estimation of PWC changes between the breakpoints is able to reduce the noise variance by the factor of N_l. Noise reduction may thus also be insufficient for medical applications if N_l is small. So, effect of noise needs more investigations and the CNVs estimate bounds are required.

2 Jitter in the Breakpoints

In follows from the experiment conducted in [29] and supported by Fig. 1 that jitter in the breakpoints plays a critical role in the estimation of the CNVs. Large jitter may cause wrong conclusions about the breakpoint locations. On the other hand, it may cause extra errors in the determination of segmental changes especially if N_l and segmental SNRs occur to be small.

2.1 Laplace-Based Approximation

The results published in [29] and our own investigations provided in [39] and generalized in Fig. 1b show that jitter in the breakpoints has approximately the skew Laplace distribution. The discrete skew Laplace distribution was recently derived in [40],

$$p(k|d_l, q_l) = \frac{(1-d_l)(1-q_l)}{1-d_l q_l} \begin{cases} d_l^k\,, & k \geqslant 0\,, \\ q_l^{|k|}\,, & k \leqslant 0\,, \end{cases} \tag{5}$$

where $d_l = e^{-\frac{\kappa_l}{\nu_l}} \in (0,1)$ and $q_l = e^{-\frac{1}{\kappa_l \nu_l}} \in (0,1)$ and in which $\kappa_l > 0$ and $\nu_l > 0$ are coefficients defined by the process. Below, we shall show that (5) can serve as a reasonably good approximation for jitter in the breakpoints of PWC signals such as that shown in Fig. 1a if the segmental SNRs exceed unity.

Let us consider N neighboring to n_l readings in each segment. We may assign an event A_{lj} meaning that all measurements at points $n_l - N \leqslant j < n_l$ belong to lth segment. Another event B_{lj} means that all measurements at $n_l \leqslant j < n_l + N - 1$ belong to $(l+1)$th segment. We think that a measured value belongs to one segment if the probability is larger than if it belongs to another segment. Because noise is Gaussian and the segmental variances are different, the Gaussian pdfs cross each other in two points, α_l and β_l. The events A_{lj} and B_{lj} can thus be specified as follows:

$$A_{lj} \quad \text{is} \quad \begin{cases} (\alpha_l < x_j) \wedge (x_j < \beta_l)\,, & \sigma_l^2 > \sigma_{l+1}^2\,, \\ x_j > \alpha_l\,, & \sigma_l^2 = \sigma_{l+1}^2\,, \\ \alpha_l < x_j < \beta_l\,, & \sigma_l^2 < \sigma_{l+1}^2\,, \end{cases} \tag{6}$$

$$B_{lj} \quad \text{is} \quad \begin{cases} \beta_l < x_j < \alpha_l\,, & \sigma_l^2 < \sigma_{l+1}^2\,, \\ x_j < \alpha_l\,, & \sigma_l^2 = \sigma_{l+1}^2\,, \\ (x_j < \alpha_l) \wedge (x_j > \beta_l)\,, & \sigma_l^2 > \sigma_{l+1}^2\,. \end{cases} \tag{7}$$

The inverse events meaning that at least one of the points do not belong to the relevant interval are $\bar{A}_{lj} = 1 - A_{lj}$ and $\bar{B}_{lj} = 1 - B_{lj}$.

Both A_{lj} and B_{lj} can be united into two blocks

$$\begin{aligned} \mathbf{A}_l &= \{A_{l(i_l-N)} A_{l(i_l-N+1)} \dots A_{l(i_l-1)}\}\,, \\ \mathbf{B}_l &= \{B_{l(i_l)} B_{l(i_l+1)} \dots B_{l(i_l+N-1)}\}\,. \end{aligned}$$

We think that if $\mathbf{A}_l$ and $\mathbf{B}_l$ occur simultaneously then the breakpoint n_l will be jitter-free. However, there may be found some other events which do not obligatorily lead to jitter. We ignore such events and define approximately the probability $P(\mathbf{A}_l\mathbf{B}_l)$ of the jitter-free breakpoint as

$$P(\mathbf{A}_l\mathbf{B}_l) = P(A_{i_l-N} \dots A_{i_l-1} B_{i_l} \dots B_{i_l+N-1})\,. \tag{8}$$

The inverse event $\bar{P}(\mathbf{A}_l\mathbf{B}_l) = 1 - P(\mathbf{A}_l\mathbf{B}_l)$ meaning that at least one point belongs to another event can be called the *jitter probability*.

In white Gaussian noise, all the events are independent and (8) thus can be rewritten as

$$P(\mathbf{A}_l\mathbf{B}_l) = P^N(A_l) P^N(B_l)\,, \tag{9}$$

where, following (6) and (7), the probabilities $P(A_l)$ and $P(B_l)$ can be specified as, respectively,

$$P(A_l) \quad = \quad \begin{cases} 1 - \int\limits_{\beta_l}^{\alpha_l} p_l(x)dx\,, & \sigma_l^2 > \sigma_{l+1}^2\,, \\ \int\limits_{\alpha_l}^{\infty} p_l(x)dx\,, & \sigma_l^2 = \sigma_{l+1}^2\,, \\ \int\limits_{\alpha_l}^{\beta_l} p_l(x)dx\,, & \sigma_l^2 < \sigma_{l+1}^2\,, \end{cases} \tag{10}$$

$$P(B_l) \quad = \quad \begin{cases} \int\limits_{\beta_l}^{\alpha_l} p_{l+1}(x)dx\,, & \sigma_l^2 > \sigma_{l+1}^2\,, \\ \int\limits_{-\infty}^{\alpha_l} p_{l+1}(x)dx\,, & \sigma_l^2 = \sigma_{l+1}^2\,, \\ 1 - \int\limits_{\alpha_l}^{\beta_l} p_{l+1}(x)dx\,, & \sigma_l^2 < \sigma_{l+1}^2\,, \end{cases} \tag{11}$$

where $p_l(x) = \frac{1}{\sqrt{2\pi\sigma_l^2}} e^{-\frac{(x-a_l)^2}{\sigma_l^2}}$ is Gaussian density.

Let us now think that jitter occurs at some point $n_l \pm k$, $0 \leqslant k \leqslant N$, and assign two additional blocks of events

$$\begin{aligned} \mathbf{A}_{lk} &= \{A_{i_l-N} \dots A_{i_l-1-k}\}, \\ \mathbf{B}_{lk} &= \{B_{i_l+k} \dots B_{i_l+N-1}\}. \end{aligned}$$

The probability $P_k^- \triangleq P_k^-(\mathbf{A}_{lk}\bar{A}_{l(i_l-k)} \dots \bar{A}_{i_l-1}\mathbf{B}_l)$ that jitter occurs at kth point to the left from n_l (left jitter) and the probability $P_k^+ \triangleq P_k^+(\mathbf{A}_l\bar{B}_{l(i_l+1)} \dots \bar{B}_{l(i_l+k-1)}\mathbf{B}_{lk})$ that jitter occurs at kth point to the right from n_l (right jitter) can thus be written as, respectively,

$$P_k^- = P^{N-k}(A_l)[1-P(A_l)]^k P^N(B_l), \quad (12)$$

$$P_k^+ = P^N(A_l)[1-P(B_l)]^k P^{N-k}(B_l). \quad (13)$$

By normalizing (12) and (13) with (9), we arrive at a function that turns out to be independent on N:

$$f_l(k) = \begin{cases} [P^{-1}(A_l)-1]^{|k|} & , \quad k<0, \quad \text{(left)} \\ 1 & , \quad k=0, \\ [P^{-1}(B_l)-1]^k & , \quad k>0. \quad \text{(right)} \end{cases} \quad (14)$$

Further normalization of $f_l(k)$ to have a unit area leads to the pdf $p_l(k) = \frac{1}{\phi_l} f_l(k)$, where ϕ_l is the sum of the values of $f_l(k)$ for all k,

$$\phi_l = 1 + \sum_{k=1}^{\infty} [\varphi_l^A(k) + \varphi_l^B(k)], \quad (15)$$

where $\varphi_l^A(k) = [P^{-1}(A_l)-1]^k$ and $\varphi_l^B(k) = [P^{-1}(B_l)-1]^k$. Now observe that, in the approximation accepted, $f_l(k)$ converges with k only if $0.5 < \tilde{P} = \{P(A), P(B)\} < 1$. Otherwise, if $\tilde{P} < 0.5$, the sum ϕ_l is infinite, $f_l(k)$ cannot be transformed to $p_l(k)$, and the lth breakpoint cannot be detected. Considering the case of $0.5 < \tilde{P} = \{P(A), P(B)\} < 1$, we conclude that $\ln\tilde{P} < 0$, $\ln(1-\tilde{P}) < 0$, and $\ln(1-\tilde{P}) < \ln\tilde{P}$. Next, using a standard relation $\sum\limits_{k=1}^{\infty} x^k = \frac{1}{x^{-1}-1}$, where $x < 1$, and after little transformations we bring (15) to

$$\phi_l = \frac{P(A_l) + P(B_l) - 1}{[1-2P(A_l)][1-2P(B_l)]}. \quad (16)$$

The *jitter pdf* $p_l(k)$ associated with the lth breakpoint can finally be found to be

$$p_l(k) = \frac{1}{\phi_l} \begin{cases} [P^{-1}(A_l)-1]^{|k|} & , \quad k<0, \\ 1 & , \quad k=0, \\ [P^{-1}(B_l)-1]^k & , \quad k>0, \end{cases} \quad (17)$$

where ϕ_l is specified by (16) and $0.5 < \{P(A_l), P(B_l)\} < 1$. By substituting $q_l = P^{-1}(A_l) - 1$ and $d_l = P^{-1}(B_l) - 1$, we find $P(A_l) = 1/(1+q_l)$ and $P(B_l) = 1/(1+d_l)$, provide the transformations, and finally go from (17) to the discrete skew Laplace distribution (5) in which κ_l and ν_l still need to be connected to (17). To find κ_l and ν_l, below we consider three points $k=-1$, $k=0$, and $k=1$. By equating (5) and (17), we first obtain $\frac{(1-d_l)(1-q_l)d_l}{1-d_lq_l} = \frac{1}{\phi_l}\frac{1-P(B_l)}{P(B_l)}$ for $k=1$ and $\frac{(1-d_l)(1-q_l)q_l}{1-d_lq_l} = \frac{1}{\phi_l}\frac{1-P(A_l)}{P(A_l)}$ for $k=-1$ that gives us $\nu_l = \frac{1-\kappa_l^2}{\kappa_l \ln\mu_l}$, where

$$\mu_l = \frac{P(A_l)[1-P(B_l)]}{P(B_l)[1-P(A_l)]}. \quad (18)$$

For $k=0$, we have $\frac{(1-d_l)(1-q_l)}{1-d_lq_l} = \frac{1}{\phi_l}$ and transform it to an equation $x_l^2 - \frac{\phi_l(1+\mu_l)}{1+\phi_l}x - \frac{1-\phi_l}{1+\phi_l}\mu_l = 0$, which proper solution is

$$x_l = \frac{\phi_l(1+\mu_l)}{2(1+\phi_l)}\left(1 - \sqrt{1 + \frac{4\mu_l(1-\phi_l^2)}{\phi_l^2(1+\mu_l)^2}}\right) \quad (19)$$

and which $x_l = \mu_l^{-\frac{\kappa_l^2}{1-\kappa_l^2}}$ gives us

$$\kappa_l = \sqrt{\frac{\ln x_l}{\ln(x_l/\mu_l)}}. \quad (20)$$

By combining ν_l with (19), we also provide a simpler form for ν_l, namely

$$\nu_l = -\frac{\kappa_l}{\ln x_l}. \quad (21)$$

The discrete skew Laplace distribution (5) can thus be used to represent jitter in the breakpoints statistically.

Now substitute the Gaussian pdf to (10) and (11), provide the transformations, and find

$$P(A_l) = \begin{cases} 1 + \frac{1}{2}[\mathrm{erf}(g_l^\beta) - \mathrm{erf}(g_l^\alpha)] & , \quad \gamma_l^- < \gamma_l^+, \\ \frac{1}{2}\mathrm{erfc}(g_1^\alpha) & , \quad \gamma_l^- = \gamma_l^+, \\ \frac{1}{2}[\mathrm{erf}(g_l^\beta) - \mathrm{erf}(g_l^\alpha)] & , \quad \gamma_l^- > \gamma_l^+, \end{cases} \quad (22)$$

$$P(B_l) = \begin{cases} \frac{1}{2}[\mathrm{erf}(h_l^\alpha) - \mathrm{erf}(h_l^\beta)] & , \quad \gamma_l^- < \gamma_l^+, \\ 1 - \frac{1}{2}\mathrm{erfc}(h_l^\alpha) & , \quad \gamma_l^- = \gamma_l^+, \\ 1 + \frac{1}{2}[\mathrm{erf}(h_l^\alpha) - \mathrm{erf}(h_l^\beta)] & , \quad \gamma_l^- > \gamma_l^+, \end{cases} \quad (23)$$

where $g_l^\beta = \frac{\beta_l - \Delta_l}{|\Delta_l|}\sqrt{\frac{\gamma_l^-}{2}}$, $g_l^\alpha = \frac{\alpha_l - \Delta_l}{|\Delta_l|}\sqrt{\frac{\gamma_l^-}{2}}$, $h_l^\beta = \frac{\beta_l}{|\Delta_l|}\sqrt{\frac{\gamma_l^+}{2}}$, $h_l^\alpha = \frac{\alpha_l}{|\Delta_l|}\sqrt{\frac{\gamma_l^+}{2}}$, $\mathrm{erf}(x)$ is the error function, $\mathrm{erfc}(x)$ is the complementary error function. If

$\gamma_l^- \neq \gamma_l^+$, the coefficients α_l and β_l are defined by

$$\alpha_l, \beta_l = \frac{a_l\gamma_l^- - a_{l+1}\gamma_l^+}{\Gamma_l} \mp \frac{|\Delta_l|}{\Gamma_l}\sqrt{\gamma_l^-\gamma_l^+ + 2\Gamma_l \ln\sqrt{\frac{\gamma_l^-}{\gamma_l^+}}} \tag{24}$$

where $\Gamma_l = \gamma_l^- - \gamma_l^+$. For $\gamma_l^- = \gamma_l^+$, set $\alpha_l = \Delta_l/2$ and $\beta_l = \pm\infty$. Using (22) and (23), below we investigate errors inherent to the Laplace-based approximation.

2.2 Errors in Laplace-Based Approximation

To realize how well the discrete skew Laplace distribution (5) fits real jitter distribution with different SNRs, we consider a measurement of length M with one breakpoint at $n = K$ and two neighboring segments with known changes a_l and a_{l-1}. The segmental variances σ_l^2 and σ_{l-1}^2 of white Gaussian noise are supposed to be known. In the ML estimator, the mean square error (MSE) is minimized between the measurement and the CNVs model in which the breakpoint location is handled around an actual value. Thereby, the breakpoint location is detected when the MSE reaches a minimum. In our experiments, measurements were conducted 10^4 times for different noise realizations and the histogram of the estimated breakpoint locations was plotted. Such a procedure was repeated several times and the estimates were averaged in order to avoid ripples. Normalized to have a unit area, the histogram was accepted as the jitter pdf. The relevant algorithm can easily be designed to have as inputs a_l, a_{l-1}, segmental SNRs γ_l^- and γ_l^+, M, K, and the number of point K_1 around K covering possible breakpoint locations. The algorithm output is the jitter histogram "Jitter". An analysis was provided for typical SNR values peculiar to the CNVs measurements using the HR-CGH arrays. As a result, we came up with the following conclusions:

1) The Laplace approximation is reasonably accurate in the lower bound sense if the SNRs exceed unity, $(\gamma_l^-, \gamma_l^+) > 1$. Figure 2 sketches the Laplace pdf and the experimentally found pdf (circled) for the case of $\gamma_l^- = 1.4$ and $\gamma_l^+ = 1.38$ taken from real measurements. Related to the unit change, the approximation error was computed as $\varepsilon, \% = (\text{ML estimate} - \text{Laplace approximation})\times 100$. As can be seen, $\varepsilon_{\max}$ reaches here about 10% at $n = K$ (Fig. 2b). That means that the Laplace distribution fits measurements well for the allowed probability of jitter-free detection of 90%. It narrows the jitter bounds with about ± 2 points for 99%. Observing another example illustrated in Fig. 3 for $\gamma_l^- = 9.25625$ and $\gamma_l^+ = 2.61186$,

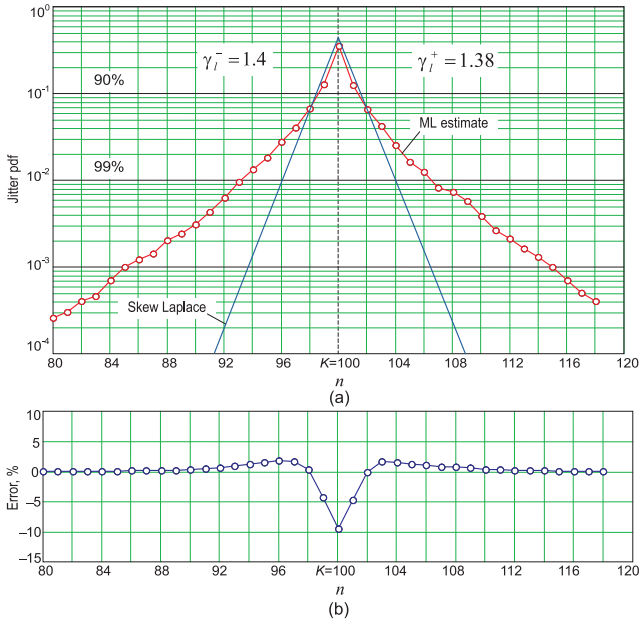


Figure 2: The jitter pdf approximated using the discrete skew Laplace distribution and found experimentally (circled) using a ML estimator over 10^4 runs for $\gamma_l^- = 1.4$ and $\gamma_l^+ = 1.38$: (a) pdfs and (b) approximation errors.

we infer that the Laplace distribution fits the process with very high accuracy if SNR $\gg 1$.

2) The approximation error may be large in the sense of the narrowed jitter bounds if SNR < 1.

3) The jitter bounds commonly cannot be determined correctly for $(\gamma_l^-, \gamma_l^+) \ll 1$.

3 Estimate Bounds

The upper bound (UB) and lower bound (LB) peculiar to the estimate confidential interval can now be found implying segmental white Gaussian noise and accepting the discrete skew Laplace-based jitter distribution in the breakpoints.

Segmental Errors. In white Gaussian noise environment, simple averaging is most efficient between the breakpoints as being optimal in the sense of the minimum produced noise. Provided the estimate $\hat{n}_l$ of the breakpoint location n_l, simple averaging applied on an interval of $N_j = n_j - n_{j-1}$ readings from n_{j-1} to $n_j - 1$ gives the following estimate for the lth segmental change

$$\hat{a}_j = \frac{1}{N_j}\sum_{v=n_{j-1}}^{n_j-1} y_v\,, \tag{25}$$

which mean value is $E\{\hat{a}_j\} = a_j$ and variance is

$$\hat{\sigma}_j^2 = \frac{\sigma_j^2}{N_j}\,. \tag{26}$$

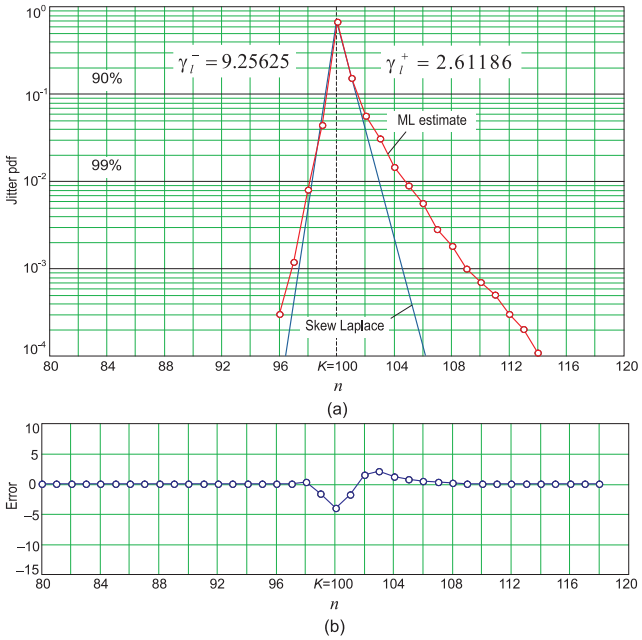


Figure 3: The jitter pdf approximated using the discrete skew Laplace distribution and found experimentally (circled) using a ML estimator over 10^4 runs for $\gamma_l^- = 9.25625$ and $\gamma_l^+ = 2.61186$: (a) pdfs and (b) approximation errors.

The UB for segmental estimates can be formed in the θ-sigma sense as $\hat{a}_j^{\rm UB} = E\{\hat{a}_j\} + \theta\sqrt{\frac{\sigma_j^2}{N_j}}$, where $\theta \geqslant 1$ is commonly integer. However, neither an actual $a_j = E\{\hat{a}_j\}$ nor multiple measurements necessary to approach a_j by averaging are available. We thus specify UB and LB approximately as

$$\hat{a}_j^{\rm UB} \cong \hat{a}_j + \theta\sqrt{\frac{\sigma_j^2}{N_j}}\,, \tag{27}$$

$$\hat{a}_j^{\rm LB} \cong \hat{a}_j - \theta\sqrt{\frac{\sigma_j^2}{N_j}}\,. \tag{28}$$

where $\theta = 1$ guarantees an existence of true changes between UB and LB with the probability of 68.27% or error probability of $\varkappa = 0.3173$ that is 31.73%; $\theta = 2$ of 95.45% or $\varkappa = 0.0555$ that is 5.55% and $\theta = 3$ of 99.73% or $\varkappa = 0.0027$ that is 0.27%.

Jitter Bounds. The jitter left bound (JLB) $J_l^{\rm L}$ and the jitter right bound (JRB) $J_l^{\rm R}$ can be determined with respect to n_l as follows. Because a step is unity with integer k, we specify the jitter probability at the kth point using (5) as

$$P_k(\gamma_l^-, \gamma_l^+) = p[k|d(\gamma_l^-, \gamma_l^+), q(\gamma_l^-, \gamma_l^+)]\,. \tag{29}$$

We then equate (29) to $\varkappa$ and solve it for the right and left jitter to have, respectively,

$$k_l^{\rm R} = \left\lfloor \frac{\nu_l}{\kappa_l} \ln \frac{(1-d_l)(1-q_l)}{\varkappa(1-d_l q_l)} \right\rfloor, \tag{30}$$

$$k_l^{\rm L} = \left\lfloor \nu_l \kappa_l \ln \frac{(1-d_l)(1-q_l)}{\varkappa(1-d_l q_l)} \right\rfloor, \tag{31}$$

where $\lfloor x \rfloor$ means the maximum integer equal to or lower than x. The JLB and JRB can be defined with respect to n_l as $J_l^{\rm L} = n_l - k_l^{\rm L}$ and $J_l^{\rm R} = n_l + k_l^{\rm R}$. Now observe that n_l is unknown and use the estimate $\hat{n}_l$. If it happens that $\hat{n}_l$ lies at the right bound, then the true n_l can be found $k_l^{\rm R}$ points to the left. Otherwise, if $\hat{n}_l$ lies at the left bound, then i_l can be found $k_l^{\rm L}$ points to the right. Approximate JLB and JRB are thus the following

$$J_l^{\rm L} \cong \hat{n}_l - k_l^{\rm R}\,, \tag{32}$$

$$J_l^{\rm R} \cong \hat{n}_l + k_l^{\rm L}\,. \tag{33}$$

Note that $\varkappa$ in (30) and (31) should be specified in the θ-sense as in (27) and (28).

UB and LB Masks and Algorithm. By combining (27), (28), (32), and (33), the UB mask $\mathcal{B}_n^{\rm U}$ and the LB mask $\mathcal{B}_n^{\rm L}$ can now be formed to outline the region for true genomic changes. The relevant algorithm was designed in [41]. Its inputs are measurements y_n, breakpoints estimates $\hat{n}_l$, tolerance parameter θ, number L of the breakpoints, and number of readings M. At the output, the algorithms produces two masks: $\mathcal{B}_n^{\rm U}$ and $\mathcal{B}_n^{\rm L}$.

The UB and LB masks have the following basic applied properties:

- The true CNVs exist between $\mathcal{B}_n^{\rm U}$ and $\mathcal{B}_n^{\rm L}$ with the probability determined in the θ-sigma sense.
- If $\mathcal{B}_n^{\rm U}$ or $\mathcal{B}_n^{\rm L}$ covering two or more breakpoints is uniform, then there is a probability of no changes in this region.
- If both $\mathcal{B}_n^{\rm U}$ and $\mathcal{B}_n^{\rm L}$ covering two or more breakpoints are uniform, then there is a high probability of no changes in this region.

We notice again that the jitter bounds in $\mathcal{B}_n^{\rm U}$ and $\mathcal{B}_n^{\rm L}$ may have enough accuracy if $(\gamma_l^-, \gamma_l^+) > 1$. They may be considered in the lower bound sense if $(\gamma_l^-, \gamma_l^+) < 1$. However, the approximation error is commonly large if $(\gamma_l^-, \gamma_l^+) < 0.5$. For details, see Section 2.2.

4 Applications

In this section, we test some CNVs measurements and estimates by the UB and LB masks computed

ISBN: 978-1-61804-240-8

in the three-sigma sense, $\theta = 3$, using the algorithm [41–43]. Because the algorithm can be applied to any CNVs data with supposedly known breakpoints, we choose the 1st chromosome measured using the HR-CGH array in [28] and available from [44].

The CNVs structure has 34 segments and 33 breakpoints. Most of the segments have the SNRs exceeding unity meaning that the UB and LB masks will have enough accuracy. The SNRs in segments $\hat{a}_{18}$ and $\hat{a}_{21}$ range between 0.5 and unity which means that real jitter can be here about twice larger. The remaining segments $\hat{a}_{23}$, $\hat{a}_{28}$, $\hat{a}_{31}$ and $\hat{a}_{32}$ demonstrate the SNR below 0.5 that means that the jitter bounds cannot be estimated with sufficient accuracy. We just may say that jitter can be much larger in the relevant breakpoints.

Let us consider the CNVs measurements and estimates in more detail following Fig. 4. As can be seen, there are two intervals with no measurements between the breakpoints $\hat{i}_{15}$ and $\hat{i}_{16}$ and the breakpoints $\hat{i}_{28}$ and $\hat{i}_{29}$. A part of measurements covering the breakpoints from $\hat{i}_5$ to $\hat{i}_{14}$ is shown in Fig. 5a. Its specific is that the segmental SNRs are all larger than unity and the masks thus can be used directly for practical applications. The masks suggest that errors in all of the segmental estimates reach tens of percents. In fact, $\hat{a}_5$ and $\hat{a}_{10}$ are estimated with error of about 50%. Error exceeds 30% in the estimates $\hat{a}_7$, $\hat{a}_9$, $\hat{a}_{12}$, and $\hat{a}_{13}$. A similar problem can be observed in the estimates of almost all of the breakpoints in which left and right jitter reaches several points.

A situation even worse with a part of the chromosome covering the breakpoints from $\hat{i}_{17}$ to $\hat{i}_{26}$. The segmental errors exceed 50% here over almost all segments. Furthermore, the UB is placed above LB around $\hat{i}_{17}$, $\hat{i}_{20}$, and $\hat{i}_{22}$. That means that there is a probability that these breakpoints do not exist. On the other hand, estimates in the part covering $\hat{i}_{24}$–$\hat{i}_{26}$ are not reliable. Thus there is a probability of no changes in this region as well.

5 Conclusions

Effect of measurement noise on the HR-CGH array-based estimates of the CNVs naturally results in segmental errors and jitter in the breakpoints due to typically low SNRs. Errors can be so large that medical expert would hardly be able to arrive at correct conclusions about real CNVs structures irrespective of the estimator used. Two rules of thumb for designers of measurement equipment are thus the following: *the higher probe resolution the more segmental accuracy* and *the larger segmental SNRs the lower jitter in the breakpoints*.

Because of large noise, estimates of the CNVs may bring insufficient information to experts and must be tested by UB and LB masks. To form such masks, the jitter distribution must be known. We have shown that jitter in the breakpoints can be modeled using the discrete skew Laplace distribution if the segmental SNRs exceed unity. Otherwise, the approximation errors can be large and more profound investigations of jitter will be required. The UB and LB masks proposed in this paper in the θ-sigma sense outline the region within which the true changes exist with a high probability (99.73% in the three-sigma sense). Provided the masks, information about CNVs is more complete and sometimes can be crucial for medical experts to make a correct decision about true structure. Testing some measurements and estimates by the UB and LB masks has revealed large errors exceeding $(30...50)\%$ in many segments. It was also demonstrated that jitter in some breakpoints is redundantly large for making any decision about their true locations. We finally notice that further investigations must be focused on the jitter statistics at low SNR values that is required to sketch a more correct probabilistic picture of the CNVs.

References:

[1] International Human Genome Sequencing Consortium, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, pp. 931-945, Oct. 2004.

[2] P. Stankiewicz and J. R. Lupski, "Structural Variation in the Human Genome and its Role in Disease," *Annual Review of Medicine*, vol. 61, pp. 437-455, Feb. 2010.

[3] A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, and C. Lee, "Detection of large-scale variation in the human genome," *Nature Genetics*, vol. 36, no. 9, pp. 949-951, Sep. 2004.

[4] J. R. Pollack, T. Sorlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A. L. Borresen-Dale, and P. O. Brown, "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors," *Proc. Natl Acad. Sci (PNAS)*, vol. 99, no. 20, pp. 12963-12968, Oct. 2002.

[5] R. Pique-Regi, A. Ortega, A. Tewfik, and S. Asgharzadeh, "Detection changes in the DNA copy number," *IEEE Signal Process. Mgn.*, vol. 29, no. 1, pp. 98-107, Jan. 2012.

[6] A. F. Cockburn, M. J. Newkirk, and R. A. Firtel, "Organization of the ribosomal RNA genes of dictyostelium discoideum: mapping of the nontrascribed spacer regions," *Cell*, vol. 9, no. 4, pp. 605-613, Dec 1976.

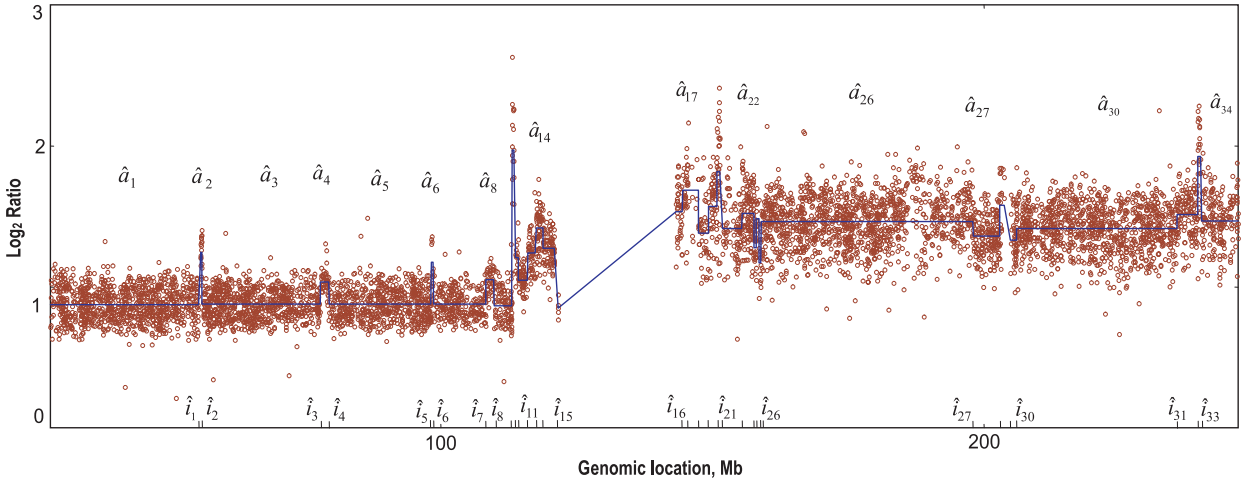


Figure 4: Measurements and estimates of the 1st chromosome changes taken from archive "159A–vs–159D–cut" available in [44].

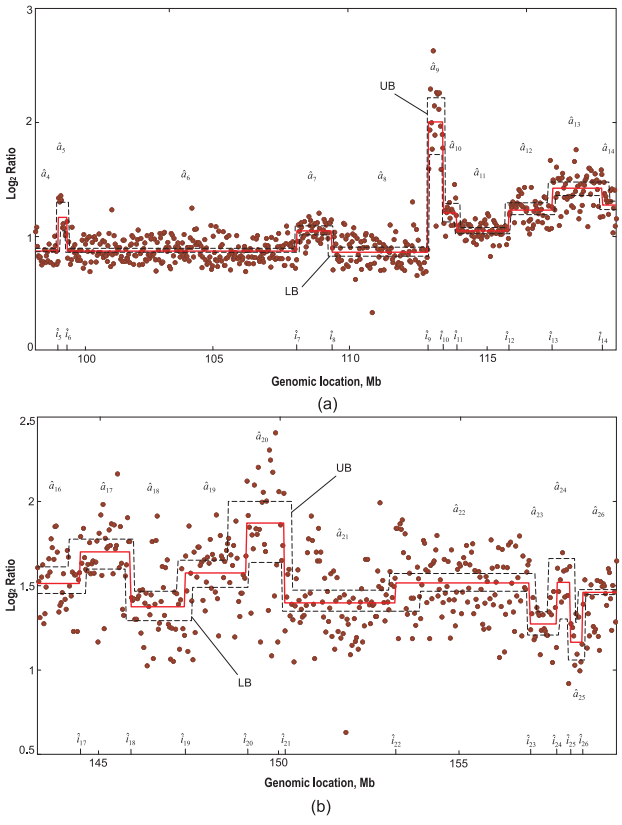


Figure 5: Parts of chromosomal changes (Fig. 4) tested by UB and LB masks: (a) genomic location from about 97Mb to 120Mb and (b) genomic location from 143Mb to 159Mb. Jitter in $\hat{i}_5$, $\hat{i}_6$, $\hat{i}_9$, and $\hat{i}_{10}$ is moderate and these breakpoints are well detectable. Breakpoints $\hat{i}_{17}$, $\hat{i}_{22}$, $\hat{i}_{27}$, $\hat{i}_{30}$, and $\hat{i}_{31}$ cannot be estimated correctly owing to low SNRs. There is a probability that the breakpoints $\hat{i}_{17}$, $\hat{i}_{19}$, $\hat{i}_{20}$, $\hat{i}_{22}$, and $\hat{i}_{24}$ do not exist.

[7] H. Ren, W. Francis, A. Boys, A. C. Chueh, N. Wong, P. La, L. H. Wong, J. Ryan, H. R. Slater, and K. H. Choo, "BAC-based PCR fragment microarray: high-resolution detection of chromosomal deletion and duplication breakpoints," *Human Mutation*, vol. 25, no. 5, pp. 476-482, May 2005.

[8] A. E. Urban, J. O. Korbel, R. Selzer, T. Richmond, A. Hacker, G. V. Popescu, J. F. Cubells, R. Green, B. S. Emanuel, M. B. Gerstein, S. M. Weissman, and M. Snyder, "High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays," *Proc. Natl. Acad. Sci. (PNAS)*, vol. 103, no. 12, pp. 4534-4539, Mar. 2006.

[9] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Research*, vol. 30, no. 4, pp. 1–10, Feb 2002.

[10] P.J. Campbell, P.J. Stephens, E.D. Pleasance, S. OMeara, H. Li, T. Santarius, L.A Stebbings, C. Leroy, S. Edkins, C. Hardy, J.W. Teague, A. Menzies, I. Goodhead, D.J. Turner, C.M. Clee, M.A. Quail, A. Cox, C. Brown, R. Durbin, M.E. Hurles, P.A.W Edwards, G.R. Bignell, M.R. Stratton, and P.A. Futreal, "Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing", *Nature Genetics*, vol. 40, no. 6, pp. 722-729, Jun. 2008.

[11] L. Hsu, S.G. Self, D. Grove, T. Randolph, K. Wang, J.J. Delrow, L. Loo, and P. Porter, "Denoising array-based comparative genomic hybridization data using wavelets", *Biostatistics*, vol. 6, no. 2, pp. 211–226, 2005.

[12] E. Ben-Yaacov and Y.C. Eldar, "A fast and flexible method for the segmentation of aCGH data", *Biostatistics*, vol. 24, no. 16, pp. i139–i145, 2008.

[13] V. Katkovnik and V.G. Spokoiny, "Spatial adaptive estimation via fitted local likelihood techniques," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 873–886, 2008.

[14] A. Goldenshluger and A. Nemirovski, "Adaptive de-noising of signals satisfying differential inequalities," *IEEE Trans. Inf. Theory*, vol. 43, no. 3, pp. 872–889, 1997.

[15] I.E. Frank and J.H. Friedman, "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, vol. 35, pp. 109-148, 1993.

[16] A.E. Hoerl and R.W. Kennard, R.W., "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, pp. 55-67, 1970.

[17] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of Royal Statist. Soc.*, ser. B, vol. 58, pp. 267-288, 1996.

[18] J. Fridlyand, A.M. Snijders, D. Pinkel, D.G. Albertson, and A.N. Jain, "Hidden Markov models approach to the analysis of array CGH data", *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 132-153, 2004.

[19] J. Chen and Y.-P. Wang, "A statistical change point model approach for the detection of DNA copy number variations in array CGH data", *IEEE/ACM Trans. on Comput. Biology and Bioinform.*, vol. 6, no. 4, pp. 529–541, 2009.

[20] S. H. Chung and R. A. Kennedy, "Forward-backward non-linear filtering technique for extracting small biological signal from noise," *J Neuroscience Meth*, vol. 40, no. 1, pp. 71–86, Nov 1991.

[21] O. Vite-Chavez, R. Olivera-Reyna, O. Ibarra-Manzano, Y. S. Shmaliy, and L. Morales-Mendoza, "Time-variant forward-backward FIR denoising of piecewise-smooth signals," *Int. J. Electron. Commun. (AEU)*, vol. 67, no. 5, pp. 406–413, May 2013.

[22] J. Muñoz-Minjares, O. Ibarra-Manzano, and Y. S. Shmaliy, "Maximum likelihood estimation of DNA copy number variations in HR-CGH arrays data," In *Proc. 12th WSEAS Int. Conf. on Signal Process., Comput. Geometry and Artif. Vision (ISCGAV'12), Proc. 12th WSEAS Int. Conf. on Systems Theory and Sc. Comput. (ISTASC'12)*, Istanbul (Turkey), pp. 45-50, 2012.

[23] J. Huang, W. Wei, J. Zhang, G. Liu, G. R. Bignell, M. R. Stratton, P. A. Futreal, R. Wooster, K. W. Jones, and M. H. Shapero, "Whole genome DNA copy number changes identified by high density oligonucleotide arrays," *Hum. Genomics*, vol. 1, no. 4, pp. 287–299, May 2004.

[24] C. Xie and M. T. Tammi, "CNV-seq, a new method to detect copy number variantion using high-throughput sequencing," *BMC Bioinform.*, vol. 10, no. 80, pp. 1–9, Mar 2009.

[25] A. K. Alqallaf and A. H. Teqfik, "DNA copy number detection and Sigma filter," In *Proc. GENSIPS*, pp. 1–4, 2007.

[26] S. Ivakhno, T. Royce, A. J. Cox, D. J. Evers, R. K. Cheetham, and S. Tavare, "CNAseg–A novel framework for identification of copy number changes in cancer from second-generation sequencing data," *Bioinform.*, vol. 26, no. 24, pp. 3051–3058, Dec. 2010.

[27] J. Chen and A. K. Gupta, *Parametric Statistical Change Point Analysis with Applications to Genetics, Medicine, and Finance*, 2nd Ed., Springer, 2012.

[28] R. Lucito, J. Healy, J. Alexander, A. Reiner, D. Esposito, M. Chi, L. Rodgers, A. Brady, J. Sebat, J. Troge, J. A. West, S. Rostan, K. C. Nquyen, S. Powers, K. Q. Ye, A. Olshen, E. Venkatraman, L. Norton, and M. Wigler, "Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation," *Genome Research*, vol. 13, no. 10, pp. 2291–2305, Oct. 2003.

[29] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin, "A statistical approach for array CGH data analysis," *BMC Bioinformatics*, vol. 6, no. 1, pp. 27–37, 2005.

[30] A. B. Olshen, E. S. Venkatraman, R. Lucito, M. Wigner, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics*, vol. 5, no. 4, pp. 557–572, Oct. 2004.

[31] J. T. Simpson, R. E. Mclntyre, D. J. Adams, and R. Durbin, "Copy number variant detection in inbred strains from short read sequence data," *Bioinformatics*, vol. 26, no. 4, pp. 565–567, Feb. 2010.

[32] L. Wang, A. Abyzov, J. O. Korbel, M. Snyder, and M. Gerstein, "MSB: A mean-shift-based approach for the analysis of structural variation in the genome," *Genomic Res.*, vol. 19, no. 1, pp. 106–117, Jan 2009.

[33] V. Boeva, A. Zinovyev, K. Bleakley, J. P. Vert, I. Janoueix-Lerosey, O. Delattre, and E. Barillot, "Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization," *Bioinformatics*, vol. 27, no. 2, pp. 268–269, Jan. 2011.

[34] R. Tibshirani and P. Wang, "Spatial smoothing and hot spot detection for CGH data using the fused lasso," *Biostatistics*, vol. **9**, no. 1, pp. 18–29, Jan. 2008.

[35] R. Pique-Regi, J. Monso-Varona, A. Ortega, R. C. Seeger, T. J. Triche, and S. Asgharzadeh, "Sparse representation and Bayesian detection of genome copy number alterations from microarray data," *Bioinformatics*, vol. 24, no. 3, pp. 309–318, Mar. 2008.

[36] X. Gao and J. Huang, "A robust penalized method for the analysis of noisy DNA copy number data," *BMC Genomics*, vol. 11, no. 517, pp. 1–10, Sep. 2010.

[37] O. M. Rueda and R. Diaz-Uriante, "RJaCGH: Bayesian analysis of a aCGH arrays for detecting copy number changes and recurrent regions," *Bioinformatics*, vol. 25, no. 15, pp. 1959–1960, Aug. 2009.

[38] Y. Yuan, C. Curtis, C. Caldas, and F. Markowetz, "A sparse regulatory network of copy-number driven gene expression reveals putative breast cancer oncogenes," *IEEE Trans. Comput. Biology and Bioinformatics*, vol. 9, no. 4, pp. 947–954, Jul.-Aug. 2012.

[39] J. Muñoz-Minjares, J. Cabal-Aragon, and Y. S. Shmaliy, "Jitter probability in the breakpoints of discrete sparse piecewise-constant signals," *Proc. 21st European Signal Process. Conf. (EUSIPCO-2013)*, 2013.

[40] T. J. Kozubowski and S. Inusah, "A skew Laplace distribution on integers," *Annals of the Inst. of Statist. Math.*, vol. 58, no. 3, pp. 555-571, Sep. 2006.

[41] J. Muñoz-Minjares, J. Cabal-Aragon, Y. S. Shmaliy, "Probabilistic bounds for estimates of genome DNA copy number variations using HR-CGH microarrays," *Proc. 21st European Signal Process. Conf. (EUSIPCO-2013)*, 2013.

[42] J. Muñoz-Minjares, Y. S. Shmaliy, J. Cabal-Aragon, "Confidence limits for genome DNA copy number variations in HR-CGH array measurements," *Biomedical Signal Processing & Control,* vol. 10, pp. 166–173, Mar. 2014.

[43] J. Muñoz-Minjares, J. Cabal-Aragon, Y. S. Shmaliy, "Effect of noise on estimate bounds for genome DNA structural changes," *WSEAS Trans. on Biology and Biomedicine,* vol. 11, pp. 52–61, Apr. 2014.

[44] Representational oligonucleotide microarray analysis (ROMA), http://Roma.cshl.org.

Fundamentals of a fuzzy inference system for educational evaluation

M.A. Luis Gabriel Moreno Sandoval
William David Peña Peña

Abstract—This paper exposes features a fuzzy inference system General focused on educational evaluation, constructed from a data base relational, with the objectives of, first, generate discussion about fuzzy logic as a new tool to interpret the educational process through information technologies and communication, and second, exposing the virtues and drawbacks of these data structures in the development of models of complex phenomena. First is the introduction with the theoretical bases is exposed, Later, the technical characteristics of the application with the successes and difficulties rose by this, and finally presents the conclusions of the authors.

Keywords—Fuzzy Logic, Fuzzy Inference System, decision making with Multicriteria, alternative assessment in education.

I. Introduction

Education is a system based on the particular properties of its elements and non-linear and interdependent relationships that they have with non-reversible dynamics in time and that remain without any outside control, presenting patterns of order at various scales [1, 2]. That is why education is a complex phenomenon, where training and learning process cannot be weighted from a scale standardized for all contexts and experiences, since the only knowledge can be appropriate, understood and applied in all their dimensions, when each person internalizing it from in their everyday practice.

Current education and its ways of assessment is based in the classical world modern vision as machine, which dismembered the reality in its parts and reset it from a logical order, on the one hand, this has caused that the educational evaluation is developed under a same logical scheme regardless of the person who studies [3].

And on the other hand It has led to underestimate, and even at times ignored, the particularity of student and its context, and the possibilities it offers to exercise and develop studied knowledge, putting the topics above its actual application, where the students are overwhelmed and concealed by concepts, diagrams and formulas that do not relate to its existence [4].

M.A. Luis Gabriel Moreno Sandoval
University "Manuela Beltran"
Bogotá, Colombia
gabrielmoreno10@gmail.com

William David Peña Peña
UniversityCorporation UNIMINUTO
Cundinamarca, Colombia
wpenapena1@hotmail.com

The problem suggests a solution that enables you to relate the specific content with daily practice and perception of personal realization that knowledge gives to each person. Why is proposed an evaluation system with fuzzy logic, that not only focuses on themes, concepts and theories, they represent the connection of multiple variables within the educational process, allowing study it, understand it to improve it permanently, from the reading of each student.

A. Education: A product and a right at the same time

This project is mainly geared towards the educational evaluation for distance mode of studies in higher education, because in this case the tutor, for sharing with the student less than in face-to-face mode, requires greater predictive and assertive, to update and improve permanently their contents and methodologies, projecting to a more personal realization and knowledge to a higher yield in the working life of the students.

The document "Opens and distance learning. Considerations on trends, policies and strategies" of the UNESCO [5], it exposes education distance and open as a product provided by suppliers, but at the time defines the mode of studies as a possibility that all persons have access to the right to education, the academic to the student the broadest spectrum of possibilities through knowledge purposes providing education to isolated populations or with difficulties to access to traditional education methods, and offer academic programs oriented to the training for work [5]. The educational spectrum in Colombian society is characterized by difficulties of access to higher education, economic dependence, segregation, and in some cases violence (social, economic and political) [6], and that requires evaluation methods not only considered the interpretation of a specific subject, but the way in which this subject can contribute to the personal, social and productive development of the person, and not just one of these aspects.

B. Fuzzy logic

Fuzzy logic is a knowledge originated from mathematical statistics, giving an element a weighting of non-deterministic as the one used in the classical logic, because in this one, an element belongs or not to a set, but in the fuzzy logic that element has different ranges of belonging. Loty Admeh Sadeh (1972), creator of fuzzy logic, for example, proposes the case of a classical group called "Tall people", to which belong the people with stature greater than or equal to 1.8 meters, where a person with stature of 1,79 m not belonged to this, however, in fuzzy logic, by means of mathematical functions, pre-

established rules, and other calculations. It would be possible to say that the person with stature of 1,79 m belongs in a range from 0.78 to diffuse all "Tall people", where the minimum membership is 0 and the maximum 1 [7].

II. Fuzzy Inference System (FIS)

A FIS is the set of logical and mathematical procedures through which establishes the membership of an element to a diffuse set [8], understanding the latter as a grouping of open borders, with margins of belonging in its range [0, 1] [7] elements. Parts of a FIS are database, which stores the elements that will be evaluated and the results of their weightings. The basis of the rules, as the set of prepositions through which relate the background variables, and its consequences in certain fuzzy sets. The Fuzzification, as assessment of the background elements on mathematics, called membership functions, trying to represent the variables with diffuse labels (varying linguistic), by means of fuzzy values (linguistic estimations), front of the resulting sets. The implication, which is the relationship between the background variables and functions of the membership of the result set, according to a rule (on the basis of rules) that unites them. Aggregation, which is the union between the different implications of a rule group. And the Defuzzification, which is the procedure that does not diffuse value opposite the input variables and the result set (result) raised [8] (Figure 1).

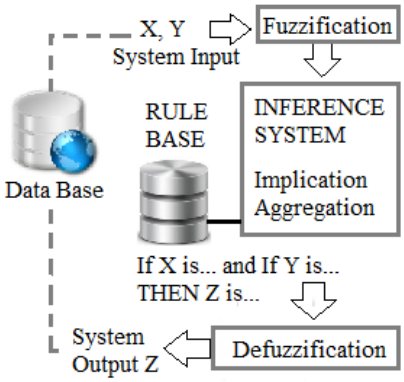


Figure 1.General diagram of a Fuzzy Inference System.

Because of that socio-economic conditions have a direct bearing on the understanding and application of knowledge, by the social projection of the knowledge, and personal fulfillment that this gives the person [1, 9], this system is based on a joint process of evaluation on specific issues and self-assessment of the students about the chances that they observed in their daily life to create and re-create it through knowledge.

Explain the database architecture, its virtues, defects and operation is impossible for reasons of format in this document, for that reason, first, simply shows the model entity-relationship (Figure 2), and second, only reference is made to the tables related to the main procedures of the system.

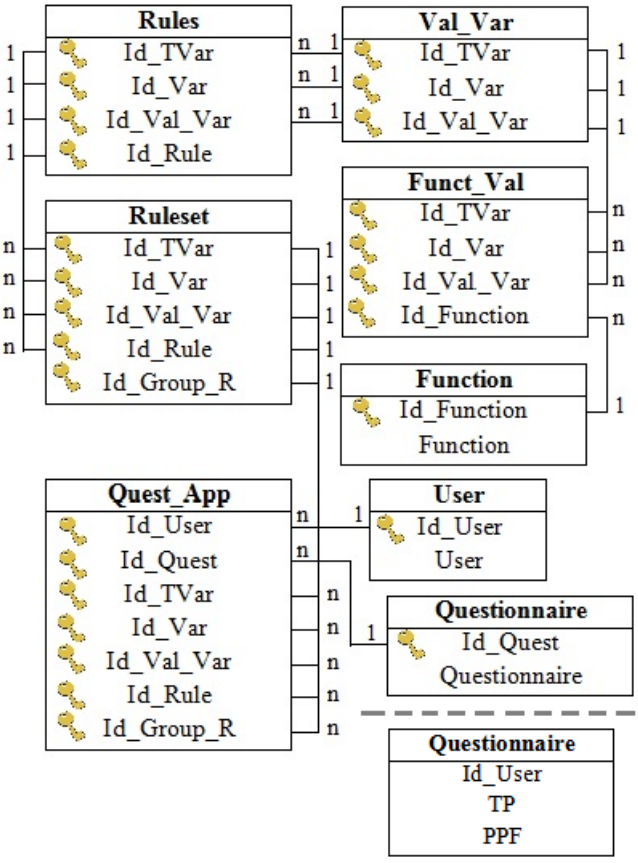


Figure 2. Entity-Relationship Model based on FIS in educational evaluation.

A table is normalized when its attributes have referential integrity, and correspond to a single record of a previous structure that gives them coherence, where for each set of attributes of that table, there is a unique index on a structure of previous data, and are further organized into structures that give them independence and interdependence, none can duplicate the key of another structure , or produce inconsistent dependencies on other attributes that are not related [10] . For example, the Table *Ruleset* prevents an attribute take more than one value (1st Normal Form) with total reliance on primary key attributes other than the primary key (2nd NF) with no transitive dependency (3st NF) so that independent attributes do not cause cyclic redundancy (Boyce - Codd Normal Form), and also avoids the possible cyclic redundancy through the structure that relates *Funct_Val* valuations variables in rules and membership functions (4th NF) [10, 11].The system model is designed following the above conditions, and therefore the application organizes information consistently.

Table *Quest*, abbreviation Questionnaire contains data answered by students, but the structure is not standardized. Therefore there is a disadvantage in moving data to a normalized table (*Quest_App*) because the movement of information depends on the skill of the programmer, who connects through consultation incoming data and primary keys of other Tables.

ISBN: 978-1-61804-240-8

Table *Quest_App*, short for Applied Questionnaire, stores the responses of students in a standard from its relationship with the User Id (User Table), Id Questionnaire (*Id_Quest*), Id Type Variable (*Id_TVar*) structure, Id Variable (*Id_Var*), Id valuation variable (*Id_Val_Var*) and Rule Id (*Id_Rule*) from the input data (Table *Ruleset* fields) (Figure 3). Table *Ruleset* contains all relationships between the antecedent variables and consequences, it inherits all the attributes of the *Rules* Table, and also records the group to which each rule belongs, used in the processes of involvement and aggregation explained later. These relationships are called rules of Implication, and in this case are of the form IF A is a' and IF B is b' THEN C is c' (Mamdani type) which are the most used, unlike the rules type Takagi-Sugeno: IF a is a' and IF B is b' THEN C is f(x) [12].

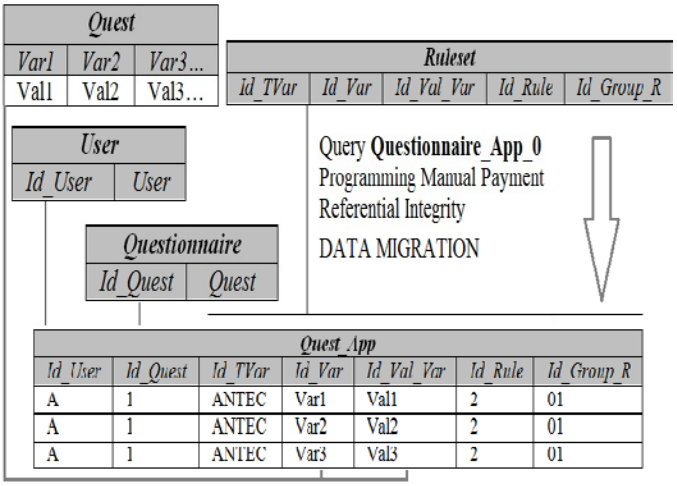


Figure 3. Scheme of hand building through consultation (Questionnaire_App_0) Table Questionnaire_App.

As this is a pilot project, for now the interest is to generate a relationship between evaluation of issues and the perception of personal realization of the student through such content, that is why the background variables considered are theoretical precision (*TP*), which gets an average rating on a specific topic, with ratings by using tags, and perception of personal fulfillment (*PPF*), which is obtained from the auto-evaluation question to the student: in what measure do you think these skills have served to engage more fully and learn about yourself? And the resulting consequence of the involvement of these variables is called Staff-academic achievement (*SAA*). Although each variable has its own identity, its ratings labels are the same for all: low (1), medium-low (2), medium (3), medium-high, (4), high (5), creating the matrix of involvement from these rules. For example in table in., the rule in the position (1,1), reads: IF TP is Low (1) and IF PPF is Low (1) THEN SAA is Low (1) (Table I).

Table *Val_Var* is also the basis of the referential integrity of the structure *Funct_Val* short for Valuation Function, wherein the data of the membership functions that correspond to each evaluation of the background variables and the consequences are stored. The membership functions for the antecedent variables (*TP*, *PPF*) are sigmoidal, because this function has an inflection point at which growth accelerates, from representing that the person has given valuation closer to the issues or more likely to relate their personal fulfillment with knowledge (A. Figure 4.). And the membership function of the consequence (*SAA*) is trapezoidal, because it represents a transition (in belonged = 1), a momentary stage of the relationship between knowledge and perception of embodiment through this (B. Figure 4.).

TABLE I. RULES OF IMPLICATION FOR "SAA".

PPF \ TP	1	2	3	4	5
1	1	2	2	3	3
2	1	2	3	3	4
3	1	2	3	4	4
4	1	2	3	4	5
5	2	3	4	5	5

Note: Rules temporary transition, subject to review by experts.

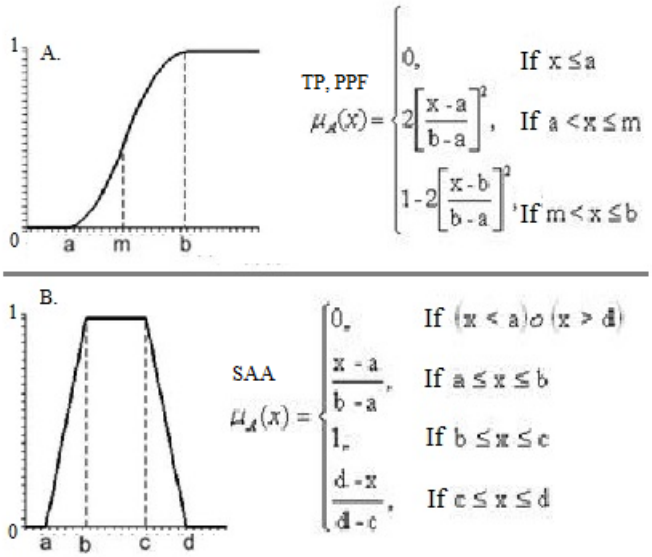


Figure 4. A. Membership sigmoidal function for background variables (TP, PPF), B. Trapezoidal Membership Function for the SAA accordingly (Based in [12]).

A. Fuzzification

Each of the data recorded by the students (made from a query in the Table *Quest_App*), must be assessed in accordance with the membership functions (Table *Funct_Val*). This created a query with SQL's union with each of the sections of the sigmoidal function, and if one wanted the Fuzzification to dynamically generate, should join each of the sections of all types of membership (A. figure 5) function.

ISBN: 978-1-61804-240-8

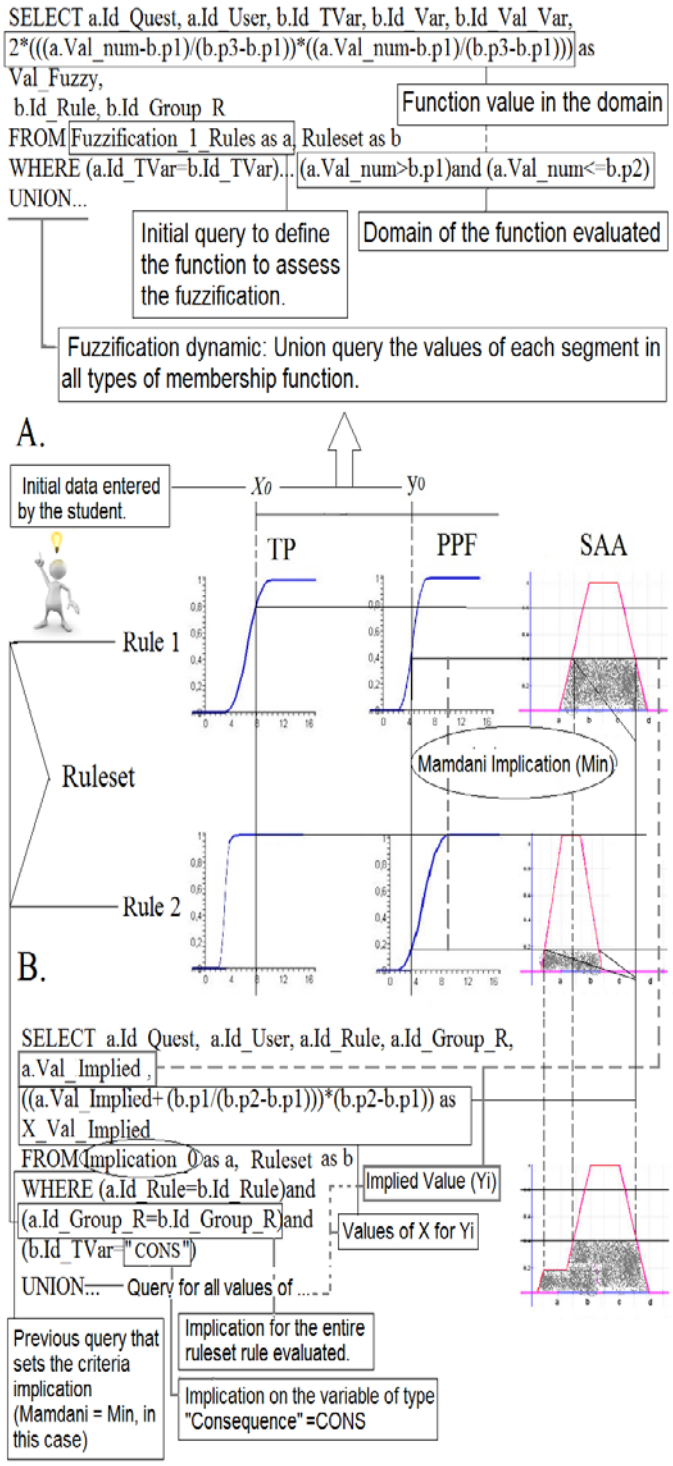


Figure 5. Fuzzification (A) and implication (B)Schemein the FIS by querying.

B. Implication

Held a consultation to assess the results of the Fuzzification from a logical condition called involvement. In this case using the involvement of Mamdani [13] (1).

$$\mu_{Rc}(x, y) = min\,[\mu_{\underline{A}}(x), \mu_{\underline{B}}(y)] \quad (1)$$

Consists in taking the lesser of the two values (for each variable TP and PPF) fusificados, and assess them as f(x)in each membership function of conclusions, according to the rules of involvement (table I.), to find the values of "X" in each one, joining the segments of the membership of the conclusions functions. This consultation generated enough workload to the machine, so it was used a query of inclusion of data of the inquiry of involvement in a new table, to reduce the cost of memory in the following processes (B. Figure 5).

C. Aggregation

Once obtained the ordered pairs built a single geometric shape (from the top), uniting the implications of the entire rule group to which belongs the rule evaluated. This was done through consultation of segregation of rule sets, which generates couples sorted by Group (corresponding to a set, in this case SAA) and not by individual rule, by means of a comparison with the ruleset data structure (A. Figure 6).

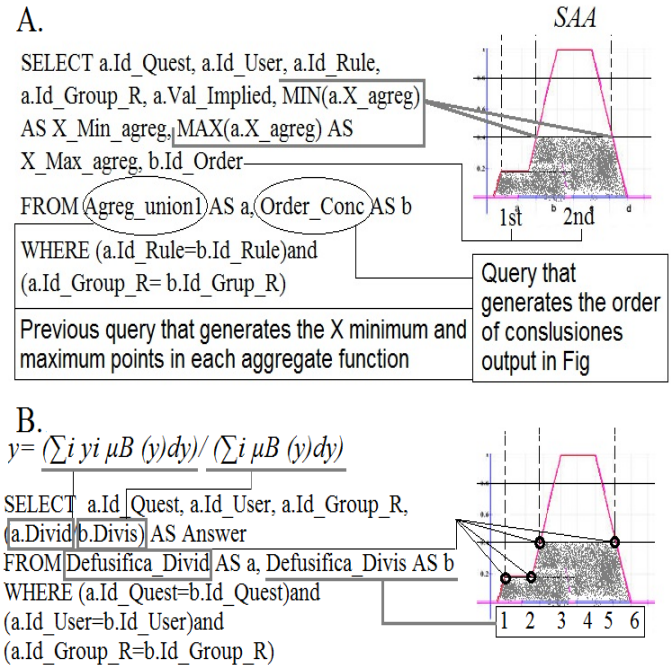


Figure 6. Aggregation (A) and Defuzzification (B) Scheme in FIS by querying

D. Defuzzification

Finally, it is necessary to apply a mathematical procedure to generate a non-diffuse value from the ordered pairs. The most common and used in this application method is the centroid, consistent in finding an average of the area of the figure (B. Figure 5), through the operation [13]:

$$y = (\sum i\, y i \mu_B(y)dy) / (\sum i \mu_B(y)dy) \quad (2)$$

III. Conclusions

1) SQL language has the limitation not to be parameterizable dynamically according to the requirements of the user, in the Fuzzification, the involvement, the aggregation or the Defuzzification, each of the evaluated function segments or each of the mathematical procedures should be performed as a query apart, with which generated great expense of memory

ISBN: 978-1-61804-240-8

on the machine. However, this application proposes as a solution the moving of data from these complex queries joining new tables, releasing the burden of memory of the above procedures.

2) Referential integrity and standardization remain the basis of coherence, strength and operation of data bases. Hence the need to keep in mind these conditions regardless of the type of information structure, since the combination of paradigms in a single computer system, allows to take elements suitable of each, allowing a deeper conceptual, with more ownership and social knowledge, such as which intends to consider this alternative (a pioneer in Colombia) educational evaluation system, from the own perception of the student (in remote mode). That is why that as work to a near future arises the union's database with a JAVA program that supplements the procedures with enough memory load, and even provides a GUI based, to make this program a full framework for Fuzzy Systems.

3) The relational databases, have gradually tended to be replaced by object-oriented database, however this application, checks that the effectiveness of these technologies lies in the ability of abstraction on a problem, coming to represent complex phenomena such as those described by means of fuzzy logic. Taking the most valuable elements of each technology, as for example, some SQL queries of few lines of instruction that can replace many lines of code in advanced programming languages. Sometimes technology makes us see both the future that we seem to wonder traveling on a plane, forgetting the importance of having learned to walk.

References

[1] López, V. “Education as a complex system”. Journal ISLAS, 44 (132). pp. 113-127. April-June 2002.

[2] Perez Martínez A. Stuart Kauffman's work. The problem of the complex order and his philosophical implications. In:Journal Thinking the complexity, 3(6): 21-38, 2009.

[3] Ballester, B. L., Colom, C. A. J. Fuzzy logic: A new epistemology for Education Sciences. Journal of Education, Spain. 340. May-August 2006, pp . 995-1008.

[4] Castro-Gómez, S. Decolonizing the university. The hubris of the zero point knowledge and dialogue. In: The decolonial rotation. Refl ections for epistemic diversity beyond global capitalism. Bogotá, Man Century Publishers. 2007. pp. 79-92.

[5] UNESCO. Open and distance learning . Considerations trends policies and strategies. Division of Higher Education. pp. 11-78. 2002.

[6] Colombian Commission of Jurists.The enjoyment of the right to education in Colombia . Alternate Report to the UN Special Rapporteur on the Right to Education. pp. 23-43, July 2007.

[7] L., Zadeh. Fuzzy sets. Information and Control, 8. pp. 338-353, 1965.

[8] Arango S. D., Conrado A., Serna U. A., Pérez O. G. Business management indicators with fuzzy logic to make decisions.Journal Lámpsakos. “Luis Amigo” University. No. 8. pp. 47-54. July-December, 2012.

[9] Alvarez H., Peña M. Modeling Fuzzy Inference Systems Takagi- Sugeno type.In: Advances in Computer Systems and No. 1. pp. 1–11, Medellíın, July 2004.

[10] S. Jorge. Principles of Relational Databases. Technical and theoretical report. pp. 23-29. 2004.

[11] Date C.J. An Introduction to Database Systems. Pearson Education. Mexico , 2001.

[12] F. Dernoncourt. Introduction to fuzzy logic. Technical and theoretical report.MIT, January 2013.

[13] MSc. Supo H. R. Fuzzy Logic. Technical and theoretical report.Jorge Basadre Grohmann National University. Tacna, Perú, 2003.

About Author (s):

- M.A. Luis Gabriel Moreno Sandoval: Systems Engineer, Master of Science in Information and Communication are the University District "Francisco José de Caldas" (Bogotá, Colombia). Manager and participant in many academic and business initiatives in information science and communication with agent-based applications, natural language processing, and other cutting-edge technologies models.
- William David Peña Peña: Philosopher, researcher in Computer Science. Participant important international events on human and Sciences to technology (Argentina, Uruguay and Cuba). Currently teaching at the university corporation UNIMINUTO in the area of research and developer of multiple p'royectos for academic lasd application of ICT.

The movement equation oh the drivers spine

Raul Miklos Kulcsar, Veronica Argesanu, Ion Silviu Borozan, Inocentiu Maniu, Mihaela Jula, Adrian Nagel

***Abstract*—**The aim of this study is to determine the movement equation of the drivers spine and to simulate the spine's movements in the coronal plane by using the MathCAD and CATIA software. On this line was proposed a methodology to approach the interaction between driver and the vehicle to allow accurate conclusions for the driving activity.

***Keywords*—**Ergonomics, spine, vehicle, simulation.

I. Introduction

RECENT years have brought the current general trend in the design, construction and operation of motor vehicles (research centers of companies producing curricula of universities / departments field, masters that include ergonomics and comfort of road vehicles, European legislation, etc. .), namely the transformation of the interior components of passive elements with uncontrolled reaction to changing human factor elements able to adapt continuously.

In the current scientific and technological conditions is required reconsideration of the research, analysis and design optics of the working places by applying outstanding results obtained recently in some new areas of human activity such as systems theory, cybernetics, information theory, operational research , computer science and ergonomics.

It should be noted that if the configuration of the safety, ventilation, lighting, heating, etc..., interior equipment is analyzed and properly carried out, in terms of scientific evidence (evidence based) on the behavior of the driver's body, especially the spine with associated muscles, still requires further study which continually preoccupied car manufacturing companies.

II. Determining the Movement Equation of the Spinal Column in the Coronal Plane

Using the coordinate system shown in figure 1 it is considered that the center of each vertebra is moving along the arch of circle with center the origin of the coordinate sistem, and radius equal to the height above the x-axis, the seat surface. The length of each arc of circle is depending on the vehicle traveling speed and the radius of the trajectory.

In the coordinate system shown in figure 1, the sinusoidal functions are variations on the x-axis:

$$XC1_i = acp \cdot \cos(f_{\Delta t} \cdot i \cdot \pi) \quad (1)$$

$$XT4_i = aum \cdot \cos(f_{\Delta t} \cdot i \cdot \pi) \quad (2)$$

$$XL1_i = alb \cdot \cos(f_{\Delta t} \cdot i \cdot \pi) \quad (3)$$

Fig. 1. – The spine in the coronal plane related to the coordinate system xOy.

The equations of the arch on what the C1, T4 and L1 vertebras are moving are given by the relations:

$$YC1_i = \sqrt{r_C^2 - XC1_i^2} \quad (4)$$

$$YT4_i = \sqrt{r_T^2 - XT4_i^2} \quad (5)$$

$$YL1_i = \sqrt{r_L^2 - XL1_i^2} \quad (6)$$

In the coronal plane and the coordinate system (fig. 1), the spine shape is a curved line given by the following equation:

ISBN: 978-1-61804-240-8

$$y(x) = a_i \cdot x^3 + b_i \cdot x^2 + c_i \cdot x + d_i \quad (7)$$

In this equation the unknowns are: a_i, b_i, c_i, and d_i. Knowing the coordinates of four points, the unknowns a_i, b_i, c_i, and d_i are determined with the *Cramer* method:

$$\begin{cases} YC1_i = a_i \cdot XC1_i^3 + b_i \cdot XC1_i^2 + c_i \cdot XC1_i + d_i \\ YT4_i = a_i \cdot XT4_i^3 + b_i \cdot XT4_i^2 + c_i \cdot XT4_i + d_i \\ YL1_i = a_i \cdot XL1_i^3 + b_i \cdot XL1_i^2 + c_i \cdot XL1_i + d_i \\ d_i = 0 \end{cases} \quad (8)$$

Due to the fact that the origin is one of the four points results that d_i is equal to 0. So the system of equations (8) is transformed into a system of three equations with three unknowns:

$$\begin{cases} YC1_i = a_i \cdot XC1_i^3 + b_i \cdot XC1_i^2 + c_i \cdot XC1_i \\ YT4_i = a_i \cdot XT4_i^3 + b_i \cdot XT4_i^2 + c_i \cdot XT4_i \\ YL1_i = a_i \cdot XL1_i^3 + b_i \cdot XL1_i^2 + c_i \cdot XL1_i \end{cases} \quad (9)$$

$$\Delta_i = \begin{bmatrix} XC1_i^3 & XC1_i^2 & XC1_i \\ XT4_i^3 & XT4_i^2 & XT4_i \\ XL1_i^3 & XL1_i^2 & XL1_i \end{bmatrix}$$

$$\Delta a_i = \begin{bmatrix} YC1_i & XC1_i^2 & XC1_i \\ YT4_i & XT4_i^2 & XT4_i \\ YL1_i & XL1_i^2 & XL1_i \end{bmatrix}$$

$$\Delta b_i = \begin{bmatrix} XC1_i^3 & YC1_i & XC1_i \\ XT4_i^3 & YT4_i & XT4_i \\ XL1_i^3 & YL1_i & XL1_i \end{bmatrix}$$

$$\Delta c_i = \begin{bmatrix} XC1_i^3 & XC1_i^2 & YC1_i \\ XT4_i^3 & XT4_i^2 & YT4_i \\ XL1_i^3 & XL1_i^2 & YL1_i \end{bmatrix}$$

$$a_i = \frac{\Delta_i}{\Delta a_i}; b_i = \frac{\Delta_i}{\Delta b_i}; c_i = \frac{\Delta_i}{\Delta c_i}$$

Using the Mathcad software has created a program sequence which limits the curve given by the equation (7), that describes the shape of the spine in the coronal plane, between the origin and the center of C1 vertebra. The program sequence is:

$$(x) = \left| \begin{array}{c} (a_i \cdot x^3 + b_i \cdot x^2 + c_i \cdot x + d_i) \text{ if } 0 \le \\ \le (a_i \cdot x^3 + b_i \cdot x^2 + c_i \cdot x + d_i) \le YC1_i \\ (break) \text{ otherwise} \end{array} \right.$$

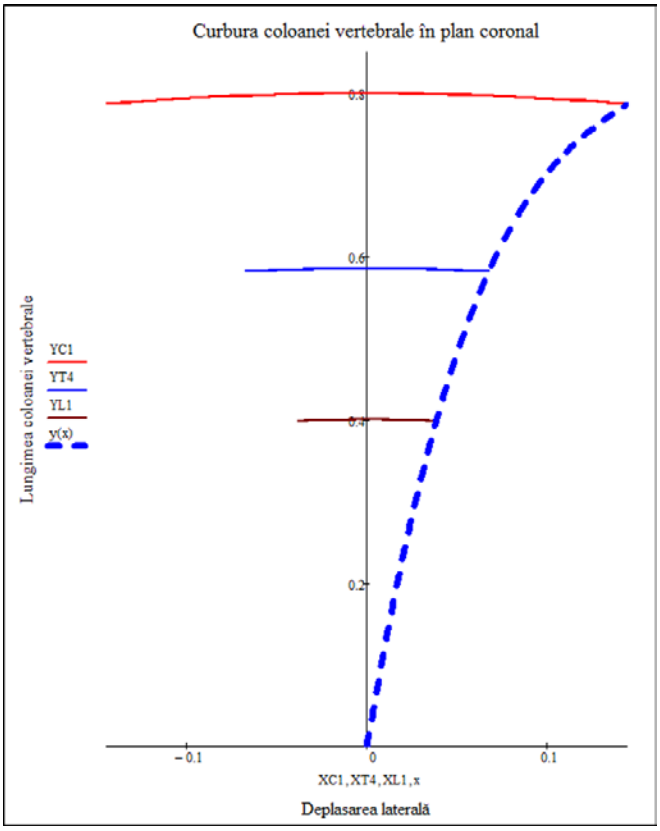


Fig. 2– The shape of spinal colum in coronal plane at the maximum right lateral tilt.

In figure 2 is shown the graph obtained with the Mathcad software in which is represented the shape of the spinal column in coronal plane at the maximum right lateral tilt, given by the equation (7). Also in this graph are the corresponding trajectory arch centers of the C1, T4, and L1 vertebras.

III. SPINE MOTION SIMULATION USING MATHCAD SOFTWARE

Through Animation function, Mathcad software allows changes to a graphic animation using software integrated FRAME variable.

Starting from the equation (7), which describes the shape of the spine in the coronal plane between the origin and the center of C1 vertebra, was created the following sequence of program that animates the graph in figure 2.

$$y(x) = \left| \begin{array}{c} (a_{FRAME} \cdot x^3 + b_{FRAME} \cdot x^2 + c_{FRAME} \cdot x + d_{FRAME}) \text{ if } 0 \le \\ \le (a_{FRAME} \cdot x^3 + b_{FRAME} \cdot x^2 + c_{FRAME} \cdot x + d_{FRAME}) \le YC1_{FRAME} \\ (break) \text{ otherwise} \end{array} \right.$$

In figure 3 are shown some frames from the created animation in Mathcad software.

ISBN: 978-1-61804-240-8

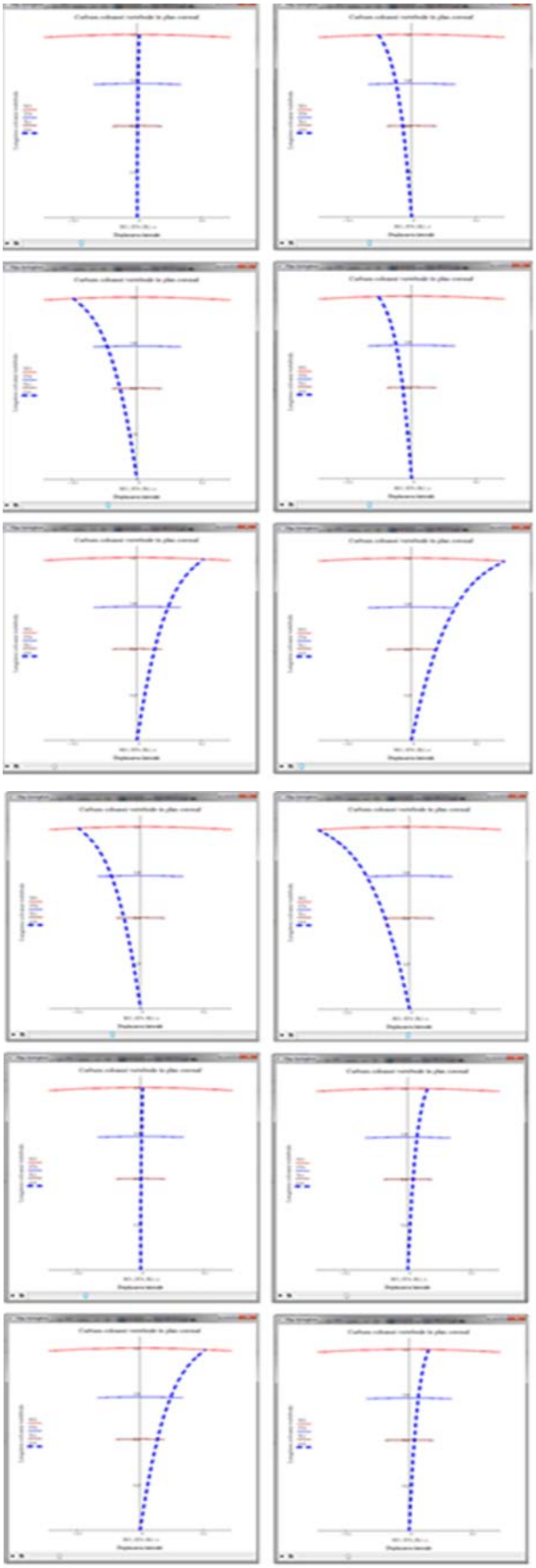


Fig. 3 – Extracted frames from Mathcad simulation.

IV. Spine motion simulation using CAD software CATIA

To achieve spine motion simulation with CAD software CATIA, the DMU Kinematics module is selected. Because the simulation of the spine movement is only in the coronal plane, spine assembly is such that the motion of each vertebra is restricted in the rotation joint at the center of rotation.

Using DMU Kinematics module, the simulation is done with the function Simulation with Laws. The laws of motion are given by the time variation of angle α_i.

The angle α_i is determined by the difference between the two slopes to the curve given by equation (7) in the centers of two consecutive vertebrae. To determine the slope in the center of the vertebrae we should know that center of coordinates must be reported to the reference system. From the 3D model of the whole spine created in CAD software CATIA V5, we extract the values of the distances between the centers of the vertebrae. The values of the distances between the centers of the vertebrae are given in Table 1.

Table. 1 – Distances between the centers of the vertebrae.

Nr. crt.	Vertebras	Distance [mm]
1.	C2 – C3	15,7
2.	C3 – C4	14
3.	C4 – C5	14,9
4.	C5 – C6	14,5
5.	C6 – C7	15
6.	C7 – T1	16
7.	T1 – T2	18,48
8.	T2 – T3	21
9.	T3 – T4	21,5
10.	T4 – T5	21,5
11.	T5 – T6	22,46
12.	T6 – T7	23,46
13.	T7 – T8	24,47
14.	T8 – T9	25,49
15.	T9 – T10	27
16.	T10 – T11	28
17.	T11 – T12	29,5
18.	T12 – L1	30,98
19.	L1 – L2	31,49
20.	L2 – L3	31,92
21.	L3 – L4	32,45
22.	L4 – L5	32,47

These values will be introduced in Mathcad software under the string of numbers with the note *dv*. Knowing the center of each vertebra is moving in an arc of a circle with its center at the origin of the coordinate system, we determine the radius of these arcs with the relationship:

ISBN: 978-1-61804-240-8

$$r_u = r_c - \left(\sum_{j=0}^{u} dv_j\right) \cdot 10^{-3}$$

Where u is a counter for determining the number of orders according to Table 1.

Each of the vertebrae is moving under a center of an arc of a circle with a time variation given by the sinusoidal function similar to those given by the relations (1), (2) and (3). The amplitude of movement is determined by the intersection of the arc corresponding to each center of the vertebrae, and the curve that describes the shape of the spine extreme point side slope.

To determine the coordinates of these points, the following sequence was created using the software program Mathcad:

$$xiv_{u,k} = \left| \begin{array}{c} h_k \; if \; trunc\left[\left(\sqrt{r_u^2 - h_k^2}\right) \cdot 10^4\right] = \\ = trunc[(a_0 h_k^3 + b_0 h_k^2 + c_0 h_k) \cdot 10^4] \\ 0 \; otherwise \end{array} \right.$$

$$xv_u = max(xiv_{u,k})$$

Where $k = 0 \ldots 145 \cdot 10^6$, $h_k = k \cdot 10^{-6}$.

$$yv_u = \sqrt{r_u^2 - xv_u^2}$$

Knowing the coordinates of the center vertebrae in the coronal plane, the shape of the spine side slope extreme point, can be determined from the slopes of the curve (7) in these points with the following relationship:

$$\gamma_u = arctg\left(\frac{y(xv_u + 10^{-6}) - y(xv_u - 10^{-6})}{(xv_u + 10^{-6}) - y(xv_u - 10^{-6})}\right)$$

Thus the angle α will be:

$$\alpha amp_u = (\gamma_{u+1} - \gamma_u)$$

The values of the angles αamp and γ are returned in radians by the Mathcad calculation software. The values of the angle *αamp* represent the amplitude of the sinusoidal function given by the following relationship (10). The functions describing the variation in time of the angle between the vertebras.

$$\alpha_u = [\alpha amp_u \cdot sin(f_{\Delta t} \cdot i \cdot \pi)] \cdot \frac{180}{\pi} \qquad (10)$$

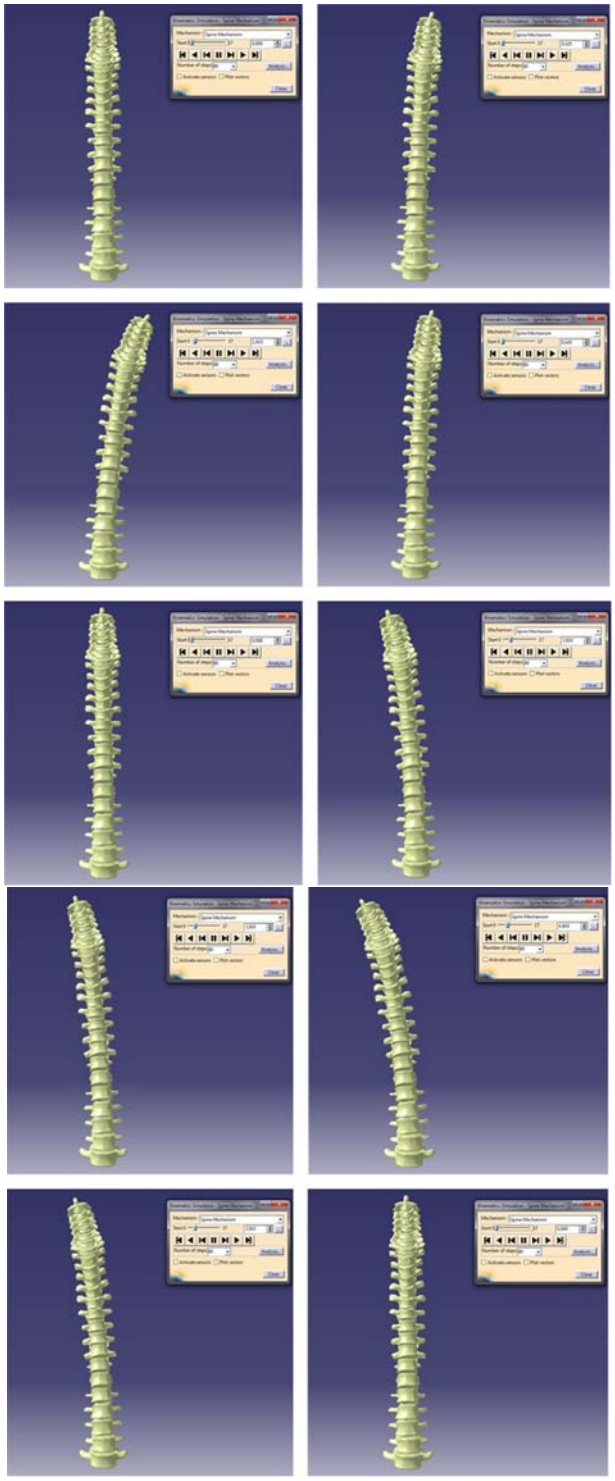


Fig. 4 – Extracted frames realized with CAD software CATIA V5.

V. Conclusion

The amplitude values of the sinusoidal functions describing the variation in time of angles between vertebrae gives an insight into the degree of deformation of intervertebral discs.

According to the literature the maximum tilt in the coronal plane of the lumbar vertebrae are 5 ° for L1-L2; 5 ° for L2-L3; 4.5 ° for L3-L4; 2.2 ° for L4-L5; 1 degrees for L1-S1.

ISBN: 978-1-61804-240-8

This sort of ergonomic analysis over the driver spine involves a study of the effects of vibration on the human body.

As reported above it is evident that we have a maximum interest for understanding the pathogenesis of diseases caused by vibration, to determine the hygienic conditions of operation of vehicles. Such research should focus on framing all operation of machinery, especially motor vehicles, in parameters corresponding to operator health insurance, in this analysis of driver and passengers.

On this line was proposed a methodology to approach the interaction between driver and the vehicle to allow accurate conclusions for the driving activity.

REFERENCES

[1] Adams M. ; Bogduk N. ; Burton K. ; Dolan P. (2006). The Biomechanics of Back Pain Second Edition, Churchill Livingstone Elsevier.

[2] Borozan I. S.; Maniu I.; Kulcsar R. M.; „Ergonomic analysis on driving an Automatic Gearbox equipped vehicle", SACI 2012 IEEE 7th International Symposium on Applied Computational Intelligence and Informatics, May 24-26, 2012, Timisoara, Romania.

[3] Borozan I. S.; Kulcsar R. M.; "Vertebral column bioengineering analysis at bending and torsion", International Conference on Human-Machine Systems, Cyborgs and Enhancing Devices HUMASCEND, Iasi, Romania, June 14-17, 2012.

[4] Goran Devedžić, Saša Ćuković, Vanja Luković, Danijela Milošević, K. Subburaj, Tanja Luković, "ScolioMedIS: web-oriented information system for idiopathic scoliosis visualization and monitoring", Journal of Computer Methods and Programs in Biomedicine, Vol.108, No.-, pp. 736-749, ISSN -, Doi 10.1016/j.cmpb.2012.04.008, 2012.

[5] Hilohi C., Untaru M., Soare I., Druța Gh., „Metode si mijloace de incercare a automobilelor", Editura Tehnică Bucure□ti, 1982.

[6] Hinza B., Seidel H., „The significance of using anthropometric parameters and postures of European drivers as a database for finite-element models when calculating spinal forces during whole-body vibration exposure", International Journal of Industrial Ergonomics, Elsevier, 2008.

[7] Kolich M., „A conceptual framework proposed to formalize the scientific investigation of automobile seat comfort", Applied Ergonomics, Elsevier, 2008.

[8] Kulcsar R. M.; Madaras L.; "Ergonomical study regarding the effects of the inertia and centrifugal forces on the driver", MTM & Robotics 2012, The Joint International Conference of the XI International Conference on Mechanisms and Mechanical Transmissions (MTM) and the International Conference on Robotics (Robotics'12), Clermont-Ferrand, France, June 6-8, 2012 Applied Mechanics and Materials, Vol. 162, Mechanisms, Mechanical Transmissions and Robotics, ISBN-13:978-3-03785-395-5, pp. 84-91.

[9] Muksian R., Nash C.D.Jr. "A model for the response of seated humans to sinusoidal displacements of the seat", J. Biomechanics, vol.7, pp 209-215, Pergamon Press, 1974.

[10] Tae-Yun Koo1, Kee-Jun Park, „A Study on Driver's Workload of Telematics Using a Driving Simulator: A Comparison among Information Modalities", International journal of precision engineering and manufacturing vol. 10, no. 3, pp. 59-63, 2009.

Authors Index

ISBN: 978-1-61804-240-8