

NEW DEVELOPMENTS in PURE and APPLIED MATHEMATICS

**Proceedings of the International Conference on Mathematical
Methods, Mathematical Models and Simulation in Science and
Engineering (MMSSE 2015)**

**Proceedings of the International Conference on Pure Mathematics -
Applied Mathematics (PM-AM 2015)**

**Vienna, Austria
March 15-17, 2015**

NEW DEVELOPMENTS in PURE and APPLIED MATHEMATICS

Proceedings of the International Conference on Mathematical Methods, Mathematical Models and Simulation in Science and Engineering (MMSSE 2015)

Proceedings of the International Conference on Pure Mathematics - Applied Mathematics (PM-AM 2015)

**Vienna, Austria
March 15-17, 2015**

Copyright © 2015, by the editors

All the copyright of the present book belongs to the editors. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the editors.

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 | 42

ISSN: 2227-4588

ISBN: 978-1-61804-287-3

NEW DEVELOPMENTS in PURE and APPLIED MATHEMATICS

**Proceedings of the International Conference on Mathematical
Methods, Mathematical Models and Simulation in Science and
Engineering (MMSSE 2015)**

**Proceedings of the International Conference on Pure Mathematics -
Applied Mathematics (PM-AM 2015)**

**Vienna, Austria
March 15-17, 2015**

Organizing Committee

Editors:

Professor Nikos E. Mastorakis, Technical University of Sofia, Bulgaria
Professor Panos M. Pardalos, University of Florida, USA
Professor Ravi P. Agarwal, Department of Mathematics, Texas A&M University

Program Committee:

Prof. Lotfi Zadeh (IEEE Fellow, University of Berkeley, USA)
Prof. Leon Chua (IEEE Fellow, University of Berkeley, USA)
Prof. Michio Sugeno (RIKEN Brain Science Institute (RIKEN BSI), Japan)
Prof. Dimitri Bertsekas (IEEE Fellow, MIT, USA)
Prof. Demetri Terzopoulos (IEEE Fellow, ACM Fellow, UCLA, USA)
Prof. Georgios B. Giannakis (IEEE Fellow, University of Minnesota, USA)
Prof. George Vachtsevanos (Georgia Institute of Technology, USA)
Prof. Abraham Bers (IEEE Fellow, MIT, USA)
Prof. Brian Barsky (IEEE Fellow, University of Berkeley, USA)
Prof. Aggelos Katsaggelos (IEEE Fellow, Northwestern University, USA)
Prof. Josef Sifakis (Turing Award 2007, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland)
Prof. Hisashi Kobayashi (Princeton University, USA)
Prof. Kinshuk (Fellow IEEE, Massey Univ. New Zeland),
Prof. Leonid Kazovsky (Stanford University, USA)
Prof. Narsingh Deo (IEEE Fellow, ACM Fellow, University of Central Florida, USA)
Prof. Kamisetty Rao (Fellow IEEE, Univ. of Texas at Arlington, USA)
Prof. Anastassios Venetsanopoulos (Fellow IEEE, University of Toronto, Canada)
Prof. Steven Collicott (Purdue University, West Lafayette, IN, USA)
Prof. Nikolaos Paragios (Ecole Centrale Paris, France)
Prof. Nikolaos G. Bourbakis (IEEE Fellow, Wright State University, USA)
Prof. Stamatios Kartalopoulos (IEEE Fellow, University of Oklahoma, USA)
Prof. Irwin Sandberg (IEEE Fellow, University of Texas at Austin, USA),
Prof. Michael Sebek (IEEE Fellow, Czech Technical University in Prague, Czech Republic)
Prof. Hashem Akbari (University of California, Berkeley, USA)
Prof. Yuriy S. Shmaliy, (IEEE Fellow, The University of Guanajuato, Mexico)
Prof. Lei Xu (IEEE Fellow, Chinese University of Hong Kong, Hong Kong)
Prof. Paul E. Dimotakis (California Institute of Technology Pasadena, USA)
Prof. M. Pelikan (UMSL, USA)
Prof. Patrick Wang (MIT, USA)
Prof. Wasfy B Mikhael (IEEE Fellow, University of Central Florida Orlando, USA)
Prof. Sunil Das (IEEE Fellow, University of Ottawa, Canada)
Prof. Panos Pardalos (University of Florida, USA)
Prof. Nikolaos D. Katopodes (University of Michigan, USA)
Prof. Bimal K. Bose (Life Fellow of IEEE, University of Tennessee, Knoxville, USA)
Prof. Janusz Kacprzyk (IEEE Fellow, Polish Academy of Sciences, Poland)
Prof. Sidney Burrus (IEEE Fellow, Rice University, USA)
Prof. Biswa N. Datta (IEEE Fellow, Northern Illinois University, USA)
Prof. Mihai Putinar (University of California at Santa Barbara, USA)
Prof. Wlodzislaw Duch (Nicolaus Copernicus University, Poland)
Prof. Tadeusz Kaczorek (IEEE Fellow, Warsaw University of Tehcnology, Poland)
Prof. Michael N. Katehakis (Rutgers, The State University of New Jersey, USA)
Prof. Pan Agathoklis (Univ. of Victoria, Canada)
Prof. P. Demokritou (Harvard University, USA)
Prof. P. Razelos (Columbia University, USA)
Dr. Subhas C. Misra (Harvard University, USA)
Prof. Martin van den Toorn (Delft University of Technology, The Netherlands)

Prof. Malcolm J. Crocker (Distinguished University Prof., Auburn University, USA)
Prof. S. Dafermos (Brown University, USA)
Prof. Urszula Ledzewicz, Southern Illinois University, USA.
Prof. Dimitri Kazakos, Dean, (Texas Southern University, USA)
Prof. Ronald Yager (Iona College, USA)
Prof. Athanassios Manikas (Imperial College, London, UK)
Prof. Keith L. Clark (Imperial College, London, UK)
Prof. Argyris Varonides (Univ. of Scranton, USA)
Prof. S. Furfari (Direction Generale Energie et Transports, Brussels, EU)
Prof. Constantin Udriste, University Politehnica of Bucharest, ROMANIA
Prof. Patrice Brault (Univ. Paris-sud, France)
Prof. Jim Cunningham (Imperial College London, UK)
Prof. Philippe Ben-Abdallah (Ecole Polytechnique de l'Universite de Nantes, France)
Prof. Photios Anninos (Medical School of Thrace, Greece)
Prof. Ichiro Hagiwara, (Tokyo Institute of Technology, Japan)
Prof. Andris Buikis (Latvian Academy of Science, Latvia)
Prof. Akshai Aggarwal (University of Windsor, Canada)
Prof. George Vachtsevanos (Georgia Institute of Technology, USA)
Prof. Ulrich Albrecht (Auburn University, USA)
Prof. Imre J. Rudas (Obuda University, Hungary)
Prof. Alexey L Sadovski (IEEE Fellow, Texas A&M University, USA)
Prof. Amedeo Andreotti (University of Naples, Italy)
Prof. Ryszard S. Choras (University of Technology and Life Sciences Bydgoszcz, Poland)
Prof. Remi Leandre (Universite de Bourgogne, Dijon, France)
Prof. Moustapha Diaby (University of Connecticut, USA)
Prof. Brian McCartin (New York University, USA)
Prof. Elias C. Aifantis (Aristotle Univ. of Thessaloniki, Greece)
Prof. Anastasios Lyrintzis (Purdue University, USA)
Prof. Charles Long (Prof. Emeritus University of Wisconsin, USA)
Prof. Marvin Goldstein (NASA Glenn Research Center, USA)
Prof. Costin Cepisca (University POLITEHNICA of Bucharest, Romania)
Prof. Kleantlis Psarris (University of Texas at San Antonio, USA)
Prof. Ron Goldman (Rice University, USA)
Prof. Ioannis A. Kakadiaris (University of Houston, USA)
Prof. Richard Tapia (Rice University, USA)
Prof. F.-K. Benra (University of Duisburg-Essen, Germany)
Prof. Milivoje M. Kostic (Northern Illinois University, USA)
Prof. Helmut Jaberg (University of Technology Graz, Austria)
Prof. Ardeshir Anjomani (The University of Texas at Arlington, USA)
Prof. Heinz Ulbrich (Technical University Munich, Germany)
Prof. Reinhard Leithner (Technical University Braunschweig, Germany)
Prof. Elbrous M. Jafarov (Istanbul Technical University, Turkey)
Prof. M. Ehsani (Texas A&M University, USA)
Prof. Sesh Commuri (University of Oklahoma, USA)
Prof. Nicolas Galanis (Universite de Sherbrooke, Canada)
Prof. S. H. Sohrab (Northwestern University, USA)
Prof. Rui J. P. de Figueiredo (University of California, USA)
Prof. Hiroshi Sakaki (Meisei University, Tokyo, Japan)
Prof. K. D. Klaes, (Head of the EPS Support Science Team in the MET Division at EUMETSAT, France)
Prof. Emira Maljevic (Technical University of Belgrade, Serbia)
Prof. Kazuhiko Tsuda (University of Tsukuba, Tokyo, Japan)
Prof. Milan Stork (University of West Bohemia, Czech Republic)
Prof. Lajos Barna (Budapest University of Technology and Economics, Hungary)
Prof. Nobuoki Mano (Meisei University, Tokyo, Japan)

Prof. Nobuo Nakajima (The University of Electro-Communications, Tokyo, Japan)
Prof. Victor-Emil Neagoe (Polytechnic University of Bucharest, Romania)
Prof. P. Vanderstraeten (Brussels Institute for Environmental Management, Belgium)
Prof. Annaliese Bischoff (University of Massachusetts, Amherst, USA)
Prof. Virgil Tiponut (Politehnica University of Timisoara, Romania)
Prof. Andrei Kolyshkin (Riga Technical University, Latvia)
Prof. Fumiaki Imado (Shinshu University, Japan)
Prof. Sotirios G. Ziavras (New Jersey Institute of Technology, USA)
Prof. Constantin Volosencu (Politehnica University of Timisoara, Romania)
Prof. Marc A. Rosen (University of Ontario Institute of Technology, Canada)
Prof. Alexander Zemliak (Puebla Autonomous University, Mexico)
Prof. Thomas M. Gatton (National University, San Diego, USA)
Prof. Leonardo Pagnotta (University of Calabria, Italy)
Prof. Yan Wu (Georgia Southern University, USA)
Prof. Daniel N. Riahi (University of Texas-Pan American, USA)
Prof. Alexander Grebennikov (Autonomous University of Puebla, Mexico)
Prof. Bennie F. L. Ward (Baylor University, TX, USA)
Prof. Guennadi A. Kouzaev (Norwegian University of Science and Technology, Norway)
Prof. Eugene Kindler (University of Ostrava, Czech Republic)
Prof. Geoff Skinner (The University of Newcastle, Australia)
Prof. Hamido Fujita (Iwate Prefectural University(IPU), Japan)
Prof. Francesco Muzi (University of L'Aquila, Italy)
Prof. Les M. Sztandera (Philadelphia University, USA)
Prof. Claudio Rossi (University of Siena, Italy)
Prof. Sergey B. Leonov (Joint Institute for High Temperature Russian Academy of Science, Russia)
Prof. Arpad A. Fay (University of Miskolc, Hungary)
Prof. Lili He (San Jose State University, USA)
Prof. M. Nasseh Tabrizi (East Carolina University, USA)
Prof. Alaa Eldin Fahmy (University Of Calgary, Canada)
Prof. Paul Dan Cristea (University "Politehnica" of Bucharest, Romania)
Prof. Gh. Pascovici (University of Koeln, Germany)
Prof. Pier Paolo Delsanto (Politecnico of Torino, Italy)
Prof. Radu Munteanu (Rector of the Technical University of Cluj-Napoca, Romania)
Prof. Ioan Dumitrache (Politehnica University of Bucharest, Romania)
Prof. Corneliu Lazar (Technical University Gh.Asachi Iasi, Romania)
Prof. Miquel Salgot (University of Barcelona, Spain)
Prof. Amaury A. Caballero (Florida International University, USA)
Prof. Maria I. Garcia-Planas (Universitat Politecnica de Catalunya, Spain)
Prof. Petar Popivanov (Bulgarian Academy of Sciences, Bulgaria)
Prof. Alexander Gegov (University of Portsmouth, UK)
Prof. Lin Feng (Nanyang Technological University, Singapore)
Prof. Colin Fyfe (University of the West of Scotland, UK)
Prof. Zhaohui Luo (Univ of London, UK)
Prof. Mikhail Itskov (RWTH Aachen University, Germany)
Prof. George G. Tsympkin (Russian Academy of Sciences, Russia)
Prof. Wolfgang Wenzel (Institute for Nanotechnology, Germany)
Prof. Weilian Su (Naval Postgraduate School, USA)
Prof. Phillip G. Bradford (The University of Alabama, USA)
Prof. Ray Hefferlin (Southern Adventist University, TN, USA)
Prof. Gabriella Bognar (University of Miskolc, Hungary)
Prof. Hamid Abachi (Monash University, Australia)
Prof. Karlheinz Spindler (Fachhochschule Wiesbaden, Germany)
Prof. Josef Boercsoek (Universitat Kassel, Germany)
Prof. Eyad H. Abed (University of Maryland, Maryland, USA)

Prof. Robert K. L. Gay (Nanyang Technological University, Singapore)
Prof. Andrzej Ordys (Kingston University, UK)
Prof. Harris Catrakis (Univ of California Irvine, USA)
Prof. T Bott (The University of Birmingham, UK)
Prof. Petr Filip (Institute of Hydrodynamics, Prague, Czech Republic)
Prof. T.-W. Lee (Arizona State University, AZ, USA)
Prof. Le Yi Wang (Wayne State University, Detroit, USA)
Prof. John K. Galitos (Houston Community College, USA)
Prof. Oleksander Markovskyy (National Technical University of Ukraine, Ukraine)
Prof. Suresh P. Sethi (University of Texas at Dallas, USA)
Prof. Hartmut Hillmer (University of Kassel, Germany)
Prof. Bram Van Putten (Wageningen University, The Netherlands)
Prof. Alexander Iomin (Technion - Israel Institute of Technology, Israel)
Prof. Roberto San Jose (Technical University of Madrid, Spain)
Prof. Minvydas Ragulskis (Kaunas University of Technology, Lithuania)
Prof. Arun Kulkarni (The University of Texas at Tyler, USA)
Prof. Joydeep Mitra (New Mexico State University, USA)
Prof. Vincenzo Niola (University of Naples Federico II, Italy)
Prof. S. Y. Chen, (Zhejiang University of Technology, China and University of Hamburg, Germany)
Prof. Duc Nguyen (Old Dominion University, Norfolk, USA)
Prof. Tuan Pham (James Cook University, Townsville, Australia)
Prof. Jiri Klima (Technical Faculty of CZU in Prague, Czech Republic)
Prof. Rossella Cancelliere (University of Torino, Italy)
Prof. L.Kohout (Florida State University, Tallahassee, Florida, USA)
Prof. Dr-Eng. Christian Bouquegneau (Faculty Polytechnique de Mons, Belgium)
Prof. Wladyslaw Mielczarski (Technical University of Lodz, Poland)
Prof. Ibrahim Hassan (Concordia University, Montreal, Quebec, Canada)
Prof. Stavros J.Baloyannis (Medical School, Aristotle University of Thessaloniki, Greece)
Prof. James F. Frenzel (University of Idaho, USA)
Prof. Vilem Srovnal, (Technical University of Ostrava, Czech Republic)
Prof. J. M. Giron-Sierra (Universidad Complutense de Madrid, Spain)
Prof. Walter Dosch (University of Luebeck, Germany)
Prof. Rudolf Freund (Vienna University of Technology, Austria)
Prof. Erich Schmidt (Vienna University of Technology, Austria)
Prof. Alessandro Genco (University of Palermo, Italy)
Prof. Martin Lopez Morales (Technical University of Monterey, Mexico)
Prof. Ralph W. Oberste-Vorth (Marshall University, USA)
Prof. Vladimir Damgov (Bulgarian Academy of Sciences, Bulgaria)
Prof. P.Borne (Ecole Central de Lille, France)

Additional Reviewers

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

Table of Contents

Generalized Least-Squares Regressions V: Multiple Variables <i>Nataniel Greene</i>	17
New Computational Methods for Spectrometer Signal Analysis <i>Petra Perner</i>	26
Inter-Firm Transactional Relationship in Yokokai using IDE Spatial Model: An Empirical Investigation <i>Takao Ito, Rajiv Mehta, Tsutomu Ito, Makoto Sakamoto, Satoshi Ikeda, Seigo Matsuno, Yasuo Uchida</i>	32
Informetric Models for Citation Frequency Data: An Empirical Investigation <i>Lucio Bertoli-Barsotti, Tommaso Lando</i>	37
Bounds on the Generalized Krein Parameters of an Association Scheme <i>Vasco Moco Mano, Luis Almeida Vieira</i>	40
Ridge Regression and Bootstrapping in Asthma Prediction <i>Ioannis I. Spyroglou, Eleni A. Chatzimichail, E. N. Spanou, E. Paraskakis, Alexandros G. Rigas</i>	44
Bayesian Multivariate Growth Curve Models <i>Steward H. Huang</i>	49
Effects of Dry Friction on Linear Electromechanical Actuators: A New Prognostic Method based on Simulated Annealing Algorithm <i>Matteo D. L. Dalla Vedova, Paolo Maggiore, Lorenzo Pace</i>	54
Math-Statistical Models of Income Distribution: L-moments and TL-moments and Their Estimations <i>Diana Bílková</i>	63
Stability Breakdown along a Line of Equilibria in Nonlinear Circuits with Memristors <i>Ricardo Rianza</i>	79
Modelling of Exoskeleton Movement in Verticalization Process <i>Sergey Jatsun, Sergei Savin, Petr Bezmen</i>	83
Multiobjective Genetic Algorithm-Based for Time-Cost Optimization <i>Jorge Magalhães-Mendes</i>	88
Normalizations of the Proposal Density in Markov Chain Monte Carlo Algorithms <i>Antoine E. Zambelli</i>	96

Modeling the Virtual Machine Allocation Problem <i>Zoltán Ádám Mann</i>	102
Mathematical Modeling of Incheon Bridge, Structural Monitoring <i>Gheorghe M. T. Radulescu, Corina M. Radulescu, Adrian T. Radulescu</i>	107
Systems Optimization Prospected from Torus Cyclic Groups <i>Volodymyr V. Riznyk</i>	115
A Special Constant Acceleration Curve Equation <i>Mehmet Pakdemirli, İhsan Timuçin Dolapçı</i>	119
Application of Statistic Complexity Metrics to Detection of Malware Threats in Autonomic Component Ensembles <i>A. Prangishvili, O. Shonia, I. Rodonaia, V. Rodonaia</i>	124
Numerical Calculation of the Magnetic Field Produced by a Multi-Conductor Power Cable <i>Dumitru Toader, Iulia Cata, Constantin Blaj, Alina Lihaciu</i>	130
Quasi-Conformal Harmonic Mappings Related to the Janowski Starlike Functions <i>Melike Aydogan, Yasar Polatoglu, H. Esra Ozkan Ucar, Arzu Yemisci, Yasemin Kahramaner</i>	135
EH-WSNs Optimizing Technique <i>Vladimir Shakhov</i>	139
On the Financial Applications of Multivariate Stochastic Orderings <i>Sergio Ortobelli, Tomas Tichy, Tommaso Lando, Filomena Petronio</i>	142
Mathematical Model of Two-Links Mechanism Movement at Discrete Control Actions <i>Sergey Jatsun, Sergei Savin, Petr Bezmen</i>	146
Navier-Stokes Equations-Millennium Prize Problems <i>Asset Durmagambetov, Leyla Fazilova</i>	150
On a Subclass of p-Valent Starlike Functions Associated with a Generalized Hypergeometric Differential Operator <i>Entisar El-Yagubi, Maslina Darus, Melike Aydogan</i>	159
Point Triangulation using Convex Layers <i>V. Tereshchenko, Y. Tereshchenko</i>	163
Onboard Electromechanical Actuators Affected by Motor Static Eccentricity: A New Prognostic Method based on Spectral Analysis Techniques <i>Dario Belmonte, Matteo D. L. Dalla Vedova, Paolo Maggiore</i>	166

Optimal Problems with Control-State Constraints in a Regional Economy Model Identification	173
<i>Vasily V. Dikumar, Nicholas N. Olenov, Marek Wojtowicz</i>	
The Decision Making Process in the System of Product Design and Planning based on Kansei Engineering	181
<i>Kai-Shuan Shen</i>	
Longitudinal Dispersion Coefficient as Sensitivity Parameter in Water Quality Simulation Model	191
<i>Yveta Velísková, Marek Sokáč</i>	
The Tactical Model based on a Multi-Depot Vehicle Routing Problem	196
<i>P. Stodola, J. Mazal</i>	
New Results on Stability of Hybrid Stochastic Systems	202
<i>Manlika Rajchakit</i>	
Vision-Based Navigation and System Identification of Unmanned Underwater Survey Vehicle	208
<i>Seda Karadeniz Kartal, M. Kemal Leblebiciođlu</i>	
The Video Game as Practice for Developing Virtual Reality Sports Jumping Skills in Children 5 Years. Case Study of Innovative Practices in Educational Institutions of Bogotá, Colombia	215
<i>J. Lopez, L. Coy, J. Caviativa, Y. Guzman, A. Gutierrez</i>	
On Selection of Efficient Fuzzy Models Incorporated with Multi-Objective Reactive Power Control	222
<i>Ragab A. El Sehiemy</i>	
Application of the Orthogonal Invariants of Three-Dimensional Operators in some Hydrodynamic Problems and Hubble Expansion Law	228
<i>Iliia R. Lomidze</i>	
Features Gas Explosion in a Cylindrical Tube with a Hole on the Side	232
<i>Iurii H. Polandov, Vitaly A. Babankov, Sergei A. Dobrikov</i>	
Double Check of Optimization Results Using Neural Network and Statistical Methods	237
<i>Natalja Fjodorova, Marjana Novič</i>	
On the Use of Conditional Expectation Estimators	244
<i>Sergio Ortobelli, Tommaso Lando</i>	
LQR Control of a Quadrotor Helicopter	247
<i>Demet Canpolat Tosun, Yasemin Işık, Hakan Korul</i>	

Molecular Dynamics Simulations for Lithographic Production of Carbon Nanotube Structures from Graphene	253
<i>D. Fülep, I. Zsoldos, I. László</i>	
Swarm Optimization-Based Personalization of Interactive Systems	257
<i>Alexander Nikov, Stefka Stoeva, Tricia Rambharose</i>	
PMSG Wind System Control for Time-Variable Wind Speed by Imposing the DC Link Current	264
<i>Ciprian Sorandaru, Sorin Musuroi, Gheza-Mihai Erdodi, Doru-Ionut Petrescu</i>	
Dual Approach to Complex Ecological System Analysis and Modeling	270
<i>Migdat Hodzic, Mirsad Hadzikadic, Ted Carmichael, Suvad Selman</i>	
Performance Estimate for a Proton Exchange Membrane Fuel Cell: Sensitivity Analysis Aimed to Optimization	276
<i>Enrico Testa, Paolo Maggiore, Lorenzo Pace, Matteo D. L. Dalla Vedova</i>	
The Mathematical Model of Reflection of Plane Waves in a Transversely Isotropic Magneto-Thermoelastic Medium under Rotation	282
<i>Abo-El-Nour N. Abd-alla, Fatimah Alshaikh</i>	
Prediction and Evaluation of Response to Breast Cancer Chemotherapy by Use of Multifractal Analysis	290
<i>Jelena Vasiljevic, Jelena Pribic, Ksenija Kanjer, Wojtek Jonakowski, Jelena Sopta, Dragica Nikolic Vukosavljevic, Marko Radulovic</i>	
Exponential Stability of Linear Hybrid Systems with Interval Time-Varying Delays	295
<i>Grienggrai Rajchakit</i>	
Modeling of a Small Unmanned Aerial Vehicle	300
<i>Ahmed Elsayed Ahmed, Hossam Eldin Hussein Ahmed, Ashraf Hafez, Hala Mohamed Abd-Elkader, A. N. Ouda</i>	
Gradient-Statistical Algorithm for Calculating Critical Points of Density Probability of Gaussian Mixture	309
<i>N. N. Aprausheva, V. V. Dikusar, S. V. Sorokin</i>	
Hall Current Effect on MHD Free Convection Flow an Inclined Porous Plate with Constant Heat Flux	313
<i>G. Venkata Ramana Reddy</i>	
An Optimal Manner of Distribution of Drinking Water using Heuristic Method	325
<i>Abdullah Al-Hossain, Said Bourazza</i>	

Certain Integrable Cases in Dynamics of a Multi-Dimensional Rigid Body in a Nonconservative Field	328
<i>Maxim V. Shamolin</i>	
Fuzzy-Multi Agent Hybrid System for Decision Support of Consumers of Energy from Renewable Sources	343
<i>Otilia Dragomir, Florin Dragomir, Eugenia Minca</i>	
A Hybrid System for Identification of Elastic, Isotropic Thin Plate Parameters Applying Lamb Waves and Artificial Neural Networks	349
<i>Zenon Waszczyszyn, Ewa Pabisek</i>	
Adaptive Spline Processing of Discrete Flow	355
<i>Irina Burova, Yu. K. Dem'yanovich</i>	
Introduction and Simulation of a New Model of Phantom by Monte Carlo to Obtain Depth Dose	359
<i>Seyed Alireza Mousavi Shirazi</i>	
Simulation Study of Using Shift Registers Based on 16th Degree Primitive Polynomials	363
<i>Mirella Amelia Mioc</i>	
Topological Optimization of Lake Aeration Process	370
<i>Mohamed Abdelwahed</i>	
Rigid Non-Archimedean Spaces and Applications	375
<i>Nikolaj Glazunov</i>	
Math Modeling of Underground Water Infiltration in Exhausted Gas Deposit	379
<i>Irina N. Polshkova</i>	
Mathematical Modelling of Groundwater Flow Coupled with Internal Flow in Drainage Pipe Situates in a Bounded Shallow Aquifer	384
<i>I. David, C. Grădinaru, C. Gabor, I. Vlad, C.Stefanescu</i>	
Generalized Real Numbers Pendulums and Transport Logistic Applications	388
<i>A. P. Buslaev, A. G. Tatashev</i>	
Scorpion Envenomation in Naama, Algeria	393
<i>Schehrazad Selmane</i>	
Mathematical Modelling of Groundwater Flow in Aquifers which Contain Extraction/infiltration Cavity of Arbitrary Shape, using the Theory of Functions of a Complex Variable	400
<i>I. David, C. Ştefănescu, C. Grădinaru, I. Vlad, C. Gabor</i>	

One Approach to the Design of TS Fuzzy Fault Detection Filters	406
<i>Dusan Krokavec, Anna Filasova, Vratislav Hladky</i>	
Modeling and Simulation of a 12kW Direct Driven PM Synchronous Generator of Wind Power	412
<i>A. Senthil Kumar, Thomas Cermak, Stanislav Misak</i>	
Some Problems of Fuzzy Modeling of Telecommunications Networks	418
<i>Kirill Garbuzov, Alexey S. Rodionov</i>	
A Mathematical Model of Hierarchical Organization	423
<i>Satoshi Ikeda, Takao Ito, Makoto Sakamoto</i>	
Advance Trends of Hybrid Electric Vehicles	433
<i>Shahram Javadi</i>	
Study of a Neutron Transport Problem by the Variational Iteration Method	441
<i>Olga Martin</i>	
Authors Index	449

Generalized Least-Squares Regressions V: Multiple Variables

Nataniel Greene

Abstract—The multivariate theory of generalized least-squares is formulated here using the notion of generalized means. The multivariate generalized least-squares problem seeks an m dimensional hyperplane which minimizes the average generalized mean of the square deviations between the data and the hyperplane in $m + 1$ variables. The numerical examples presented suggest that a multivariate generalized least-squares method can be preferable to ordinary least-squares especially in situations where the data are ill-conditioned.

Keywords—Generalized least-squares, geometric mean regression, least-squares, multivariate regression, multiple regression, orthogonal regression.

I. OVERVIEW

ORDINARY least-squares regression in multiple variables x_0, x_1, \dots, x_m suffers from a fundamental lack of symmetry. It begins with the choice of one variable, x_0 , as the dependent variable and x_1, \dots, x_m as the independent variables. It then minimizes the distance between the data and the regression hyperplane in the x_0 variable alone. However, the regression hyperplane formed by minimizing the distance between the data and the hyperplane in the variable x_k is not the same as solving for x_k in the hyperplane formed by minimizing the distance in the variable x_j for $j \neq k$.

For each of the variables x_k , minimizing the distance between the data and the hyperplane in the variable x_k is called ordinary least-squares (OLS) $x_k | \{x_0, \dots, x_m\} \setminus \{x_k\}$ regression. To predict the value of x_k based on the data using OLS, one must use the OLS $x_k | \{x_0, \dots, x_m\} \setminus \{x_k\}$ regression hyperplane. It is not valid to take the OLS $x_j | \{x_0, \dots, x_m\} \setminus \{x_j\}$ hyperplane and solve for x_k when $j \neq k$.

The fact that there are $m + 1$ OLS hyperplanes to model a single set of data in $m + 1$ variables is problematic. One wishes to have a single linear model for the data, for which it is valid to solve for any one of the variables for prediction purposes. Multivariate generalized least-squares solves this problem by seeking to minimize the average generalized mean of the square deviations between the data and the hyperplane in all the variables simultaneously. For the resulting regression hyperplane, it is valid to solve for any of the variables for prediction purposes.

N. Greene is with the Department of Mathematics and Computer Science, Kingsborough Community College, City University of New York, 2001 Oriental Boulevard, Brooklyn, NY 11235 USA (phone: 718-368-5929; e-mail: ngreene.math@gmail.com).

II. MULTIVARIATE REGRESSIONS

The theory of generalized least-squares was already described by this author for the case of two variables [4]–[7]. The extension of this theory to multiple variables is now begun.

A. The Explicit Error Formula and Solution for Ordinary Least-Squares

The multivariate ordinary least-squares problem is defined as follows.

Definition 1: (Multivariate Ordinary Least-Squares Problem) An m dimensional hyperplane

$$x_0 = b_0 + b_1x_1 + b_2x_2 + \dots + b_mx_m \quad (1)$$

is sought which minimizes the error function defined by

$$E = \frac{1}{N} \sum_{i=1}^N (\Delta x_{0i})^2 \quad (2)$$

where

$$\Delta x_{0i} = b_0 + b_1x_{1i} + b_2x_{2i} + \dots + b_mx_{mi} - x_{0i}. \quad (3)$$

This is called OLS $x_0 | \{x_1, \dots, x_m\}$ regression. In general, OLS $x_k | \{x_0, \dots, x_m\} \setminus \{x_k\}$ regression seeks a hyperplane which minimizes

$$E = \frac{1}{N} \sum_{i=1}^N (\Delta x_{ki})^2 \quad (4)$$

where

$$\Delta x_{ki} = \left(\frac{1}{b_k} x_{0i} - \frac{b_0}{b_k} - \frac{b_1}{b_k} x_{1i} - \dots - \frac{b_{k-1}}{b_k} x_{(k-1)i} - \frac{b_{k+1}}{b_k} x_{(k+1)i} - \dots - \frac{b_m}{b_k} x_{mi} \right) - x_{ki} \quad (5)$$

The deviation Δx_{ki} at the i th data point $(x_{0i}, x_{1i}, \dots, x_{mi})$ is the difference between the hyperplane solved in terms of the variable x_k and evaluated at the data point and the data value x_{ki} . Standard subscript notations are employed for dealing with means, standard deviations, correlation coefficients and covariances in multiple variables. The i th data value for the k th variable x_k is denoted by x_{ki} , which is short for $(x_k)_i$. The means, standard deviations, and correlation coefficients are denoted as follows: $\mu_k = \mu_{x_k}$, $\sigma_k = \sigma_{x_k}$, $\rho_{jk} = \rho_{x_j, x_k}$. The covariance notation $\sigma_{jk} = \rho_{jk} \sigma_j \sigma_k$ is preferred in this paper because it makes many of the complex formulas presented here more manageable. The notation $y = x_0$ and $\Delta y_i = \Delta x_{0i}$ can also be used. However, denoting the y -variable always using the zero subscript x_0 allows one to

easily identify it when there are any number of variables and naturally fits with the subscript convention just described.

The notation Δx_{ki} greatly simplifies working with the error function. The next lemma describes a fundamental relation between the regression coefficient b_k , Δx_{ki} , which is the deviation of the variable x_k from the hyperplane at the i th data value and Δx_{0i} , which is the deviation of x_0 from the hyperplane at the i th data value.

Lemma 2: (Fundamental Relation)

$$b_k = -\frac{\Delta x_{0i}}{\Delta x_{ki}} \quad (6)$$

or

$$\Delta x_{ki} = -\frac{1}{b_k} \cdot \Delta x_{0i} \quad (7)$$

The proof is straightforward algebra. This lemma will play a fundamental role further on in allowing one to always extract a weight function from any generalized least-squares error expression.

The explicit bivariate formula for the ordinary least-squares error described by Ehrenberg [3] has a known generalization using covariance notation. It is written here as a matrix-vector equation and also explicitly.

Theorem 3: (Explicit Multivariate Error Formula) Let $\mathbf{b} = (b_1, \dots, b_m)^T$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_m)^T$, $\mathbf{s}_0 = (\sigma_{10}, \dots, \sigma_{m0})^T$ and $\mathbf{S} = [\sigma_{jk}]_{m \times m}$. Then the multivariate ordinary least-squares error written in matrix-vector notation is

$$E = \mathbf{b}^T \mathbf{S} \mathbf{b} - 2\mathbf{b}^T \mathbf{s}_0 + \sigma_{00} + (b_0 - \mu_0 + \mathbf{b}^T \boldsymbol{\mu})^2 \quad (8)$$

and explicitly it is

$$E = \sum_{j=1}^m \sum_{k=1}^m \sigma_{jk} b_j b_k - 2 \sum_{k=1}^m \sigma_{k0} b_k \quad (9)$$

$$+ \sigma_{00} + \left(b_0 - \mu_0 + \sum_{k=1}^m b_k \mu_k \right)^2. \quad (10)$$

Alternatively write

$$E = \sum_{k=1}^m \sigma_{kk} b_k^2 + 2 \sum_{j < k} \sigma_{jk} b_j b_k - 2 \sum_{k=1}^m \sigma_{k0} b_k + \sigma_{00} + \left(b_0 - \mu_0 + \sum_{k=1}^m b_k \mu_k \right)^2. \quad (11)$$

Proof: Begin with the error expression and manipulate as follows.

$$\begin{aligned} E &= \frac{1}{N} \sum_{i=1}^N (b_0 + b_1 x_{1i} + \dots + b_m x_{mi} - x_{0i})^2 \\ &= \frac{1}{N} \sum_{i=1}^N (b_1 (x_{1i} - \mu_1) + \dots + b_m (x_{mi} - \mu_m) \\ &\quad - (x_{0i} - \mu_0) + (b_0 - \mu_0 + b_1 \mu_1 + \dots + b_m \mu_m))^2 \end{aligned}$$

Square the summand and distribute the summation onto each term. To simplify, utilize the covariance notation $\sigma_{jk} = \frac{1}{N} \sum_{i=1}^N (x_{ji} - \mu_j)(x_{ki} - \mu_k)$ and utilize the fact that

$\sum_{i=1}^N (x_{ji} - \mu_j) = 0$ for all $j = 0 \dots m$. The result is then obtained. ■

Corollary 4: The explicit error formula written in covariance notation for two variables x_1 and x_0 is

$$E = \sigma_{11} b_1^2 - 2\sigma_{10} b_1 + \sigma_{00} + (b_0 - \mu_0 + b_1 \mu_1)^2. \quad (12)$$

This is equivalent to Ehrenberg's formula. For three variables x_1, x_2 and x_0 , the formula is

$$E = \sigma_{11} b_1^2 + \sigma_{22} b_2^2 + 2\sigma_{12} b_1 b_2 - 2\sigma_{10} b_1 - 2\sigma_{20} b_2 + \sigma_{00} + (b_0 - \mu_0 + b_1 \mu_1 + b_2 \mu_2)^2 \quad (13)$$

For four variables x_1, x_2, x_3 and x_0 , the formula is

$$\begin{aligned} E &= \sigma_{11} b_1^2 + \sigma_{22} b_2^2 + \sigma_{33} b_3^2 \\ &\quad + 2\sigma_{12} b_1 b_2 + 2\sigma_{13} b_1 b_3 + 2\sigma_{23} b_2 b_3 \\ &\quad - 2\sigma_{10} b_1 - 2\sigma_{20} b_2 - 2\sigma_{30} b_3 \\ &\quad + \sigma_{00} + (b_0 - \mu_0 + b_1 \mu_1 + b_2 \mu_2 + b_3 \mu_3)^2 \end{aligned} \quad (14)$$

The explicit formula for the multivariate OLS regression coefficients is now written simply in matrix-vector form.

Theorem 5: (OLS explicit solution) The vector \mathbf{b} of OLS $x_0 | \{x_1, \dots, x_m\}$ regression coefficients is given explicitly by

$$\mathbf{b} = \mathbf{S}^{-1} \mathbf{s}_0 \quad (15)$$

and

$$b_0 = \mu_0 - \mathbf{b}^T \boldsymbol{\mu} \quad (16)$$

Proof: Let $\nabla = (\partial/\partial b_1, \dots, \partial/\partial b_m)^T$ denote the gradient operator. Take the gradient of the error with respect to \mathbf{b} and set it equal to zero: $\nabla E = \mathbf{0}$. Use the matrix-vector form of the error and distribute the gradient onto each term.

$$\nabla \left(\mathbf{b}^T \mathbf{S} \mathbf{b} - 2\mathbf{b}^T \mathbf{s}_0 + \sigma_{00} + (b_0 - \mu_0 + \mathbf{b}^T \boldsymbol{\mu})^2 \right) = \mathbf{0}$$

$$2\mathbf{S} \mathbf{b} - 2\mathbf{s}_0 - 2\boldsymbol{\mu} (b_0 - \mu_0 + \mathbf{b}^T \boldsymbol{\mu}) = \mathbf{0}$$

Set $b_0 = \mu_0 - \mathbf{b}^T \boldsymbol{\mu}$, and obtain

$$\mathbf{S} \mathbf{b} = \mathbf{s}_0$$

$$\mathbf{b} = \mathbf{S}^{-1} \mathbf{s}_0.$$

■

B. The Hessian Matrix

Since it is already known that the ordinary least-squares solution vector \mathbf{b} minimizes the error function, the Hessian matrix \mathbf{H} of second-order partial derivatives of E must be positive definite. Recall that \mathbf{H} is positive definite when $\det \mathbf{H} > 0$ and when the determinants of all the upper-left submatrices of \mathbf{H} are positive. Alternatively, \mathbf{H} is positive definite when all the eigenvalues of \mathbf{H} are positive. It is instructive here to compute the Hessian matrix.

Theorem 6: The Hessian matrix is given by

$$\mathbf{H} = 2 \begin{bmatrix} 1 & \mu_1 & \cdots & \mu_m \\ \mu_1 & \sigma_{11} + \mu_1^2 & \cdots & \sigma_{1m} + \mu_1 \mu_m \\ \vdots & \vdots & \ddots & \vdots \\ \mu_m & \sigma_{m1} + \mu_1 \mu_m & \cdots & \sigma_{mm} + \mu_m^2 \end{bmatrix}. \quad (17)$$

Proof: Form all second order partial derivatives of the error function

$$H_{(j+1)(k+1)} = \frac{\partial^2 E}{\partial b_k \partial b_j} \Big|_{(b_0, b_1, \dots, b_m)}$$

for $j, k = 0 \dots m$. Verify that

$$\begin{aligned} H_{11} &= 2 \\ H_{1(k+1)} &= H_{(k+1)1} = 2\mu_k \\ H_{(j+1)(k+1)} &= H_{(k+1)(j+1)} = 2(\sigma_{jk} + \mu_j \mu_k). \end{aligned}$$

Theorem 7: The Hessian determinant is given by

$$\det \mathbf{H} = 2^{m+1} \det \mathbf{S}$$

Proof: Taking a multiple of one row and adding it to another does not affect the determinant. For $k = 1 \dots m$, multiply the first row by $-\mu_k$ and add it to row $k + 1$. After this has been done to every row, perform a cofactor expansion along the first column and obtain the result.

$$\begin{aligned} \det \mathbf{H} &= 2^{m+1} \begin{vmatrix} 1 & \mu_1 & \cdots & \mu_m \\ \mu_1 & \sigma_{11} + \mu_1^2 & \cdots & \sigma_{1m} + \mu_1 \mu_m \\ \vdots & \vdots & \ddots & \vdots \\ \mu_m & \sigma_{m1} + \mu_1 \mu_m & \cdots & \sigma_{mm} + \mu_m^2 \end{vmatrix} \\ &= 2^{m+1} \begin{vmatrix} 1 & \mu_1 & \cdots & \mu_m \\ 0 & \sigma_{11} & \cdots & \sigma_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \sigma_{m1} & \cdots & \sigma_{mm} \end{vmatrix} \\ &= 2^{m+1} \begin{vmatrix} \sigma_{11} & \cdots & \sigma_{1m} \\ \vdots & \ddots & \vdots \\ \sigma_{m1} & \cdots & \sigma_{mm} \end{vmatrix} \end{aligned}$$

Theorem 8: The Hessian matrix \mathbf{H} and the covariance matrix \mathbf{S} are both positive definite.

Proof: Since $\det \mathbf{H} = 2^{m+1} \det \mathbf{S}$, the Hessian matrix \mathbf{H} is positive definite if and only if the covariance matrix \mathbf{S} is positive definite. However, it is already known that the OLS solution vector \mathbf{b} minimizes the error function. Therefore, conclude that \mathbf{H} and \mathbf{S} are both positive definite. ■

C. Generalized Means

Definition 9: A function $M(x_0, x_1, \dots, x_m)$ defines a generalized mean for all $x_i > 0$ if it satisfies Properties 1-5 below. If it satisfies Property 6 it is called a homogenous generalized mean. The properties are:

1. (Continuity) $M(x_0, x_1, \dots, x_m)$ is continuous in each variable.
2. (Monotonicity) $M(x_0, x_1, \dots, x_m)$ is non-decreasing in each variable.
3. (Symmetry)

$$M(x_0, x_1, \dots, x_m) = M(x_{s(0)}, x_{s(1)}, \dots, x_{s(m)})$$

where $s(i)$ is any permutation of the indices 0 through m .

4. (Identity)

$$M(x, x, \dots, x) = x.$$

5. (Intermediacy)

$$\min(x_0, \dots, x_m) \leq M(x_0, \dots, x_m) \leq \max(x_0, \dots, x_m)$$

6. (Homogeneity)

$$M(tx_0, tx_1, \dots, tx_m) = tM(x_0, x_1, \dots, x_m)$$

for all $t > 0$.

All the special multivariate means are included in this definition. Note that $m + 1$ variables are used in this definition in order that it share the same form as the $m + 1$ regression variables. XMR notation is used here to name generalized regressions: if 'X' is the letter used to denote a given generalized mean, then XMR is the corresponding generalized mean square regression.

Example 10: The multivariate harmonic mean is given by

$$H(x_0, \dots, x_m) = \frac{m + 1}{\frac{1}{x_0} + \dots + \frac{1}{x_m}}. \quad (18)$$

This mean generates multivariate orthogonal regression (HMR).

Example 11: The multivariate geometric mean is given by

$$G(x_0, \dots, x_m) = (x_0 \cdot \dots \cdot x_m)^{1/(m+1)}. \quad (19)$$

This mean generates multivariate geometric mean regression (GMR).

Example 12: The multivariate arithmetic mean is given by

$$A(x_0, \dots, x_m) = \frac{1}{m + 1} (x_0 + \dots + x_m). \quad (20)$$

This mean generates multivariate arithmetic mean regression (AMR).

Example 13: The selection mean is given by

$$S^{(k)}(x_0, \dots, x_m) = x_k \quad (21)$$

after x_0, \dots, x_m are arranged in increasing order. This mean generates OLS $x_k | \{x_0, \dots, x_m\} \setminus \{x_k\}$ regression.

Regressions based on these special cases are used in this work further on. The generalized means in the next examples have free parameters which can be used to parameterize these special cases.

Example 14: The power mean of order p is given by

$$M_p(x_0, \dots, x_m) = \left(\frac{1}{m + 1} (x_0^p + \dots + x_m^p) \right)^{1/p}. \quad (22)$$

Example 15: The weighted arithmetic mean with positive weights satisfying $\alpha_0 + \alpha_1 + \dots + \alpha_m = 1$, is given by

$$M_{(\alpha_0, \alpha_1, \dots, \alpha_m)}(x_0, x_1, \dots, x_m) = \alpha_0 x_0 + \alpha_1 x_1 + \dots + \alpha_m x_m \quad (23)$$

after x_0, \dots, x_m are arranged in increasing order.

Example 16: The weighted geometric mean with positive weights satisfying $\beta_0 + \beta_1 + \dots + \beta_m = 1$, is given by

$$M_{(\beta_0, \beta_1, \dots, \beta_m)}(x_0, x_1, \dots, x_m) = x_0^{\beta_0} x_1^{\beta_1} \dots x_m^{\beta_m} \quad (24)$$

after x_0, \dots, x_m are arranged in increasing order.

D. Two Generalized Least-Squares Problems and the Equivalence Theorem

The multivariate symmetric least-squares problem is formulated as follows.

Definition 17: (The Multivariate Symmetric Least-Squares Problem) An m dimensional hyperplane

$$x_0 = b_0 + b_1x_1 + b_2x_2 + \dots + b_mx_m \quad (25)$$

is sought which minimizes the error function defined by

$$E = \frac{1}{N} \sum_{i=1}^N M \left((\Delta x_{0i})^2, (\Delta x_{1i})^2, \dots, (\Delta x_{mi})^2 \right) \quad (26)$$

where $M(x_0, x_1, \dots, x_m)$ is any generalized mean.

A more general related problem is the weighted ordinary least-squares problem.

Definition 18: (The Weighted Ordinary Least-Squares Problem) An m dimensional hyperplane

$$x_0 = b_0 + b_1x_1 + b_2x_2 + \dots + b_mx_m \quad (27)$$

is sought which minimizes the error function defined by

$$E = g(b_1, \dots, b_m) \cdot \frac{1}{N} \sum_{i=1}^N (\Delta x_{0i})^2 \quad (28)$$

or simply $E = g \cdot E_{OLS}$.

The next theorem states that every multivariate symmetric least-squares problem is equivalent to a weighted multivariate ordinary least-squares problem with weight function $g(b_1, \dots, b_m)$.

Theorem 19: (Equivalence Theorem) Every general symmetric least-squares error function can be written equivalently as

$$E = g(b_1, \dots, b_m) \cdot \frac{1}{N} \sum_{i=1}^N (\Delta x_{0i})^2 \quad (29)$$

or simply $E = g \cdot E_{OLS}$ with

$$g(b_1, \dots, b_m) = M \left(1, \frac{1}{b_1^2}, \dots, \frac{1}{b_m^2} \right) \quad (30)$$

and E_{OLS} expressed using the explicit error formula.

Proof: Write

$$\begin{aligned} E &= \frac{1}{N} \sum_{i=1}^N M \left((\Delta x_{0i})^2, (\Delta x_{1i})^2, \dots, (\Delta x_{mi})^2 \right) \\ &= \frac{1}{N} \sum_{i=1}^N M \left((\Delta x_{0i})^2, \frac{1}{b_1^2} (\Delta x_{0i})^2, \dots, \frac{1}{b_m^2} (\Delta x_{0i})^2 \right) \\ &= \frac{1}{N} \sum_{i=1}^N (\Delta x_{0i})^2 M \left(1, \frac{1}{b_1^2}, \dots, \frac{1}{b_m^2} \right) \end{aligned}$$

where the fundamental relation

$$(\Delta x_{ki})^2 = \frac{1}{b_k^2} (\Delta x_{0i})^2$$

is used. Let $g(b_1, \dots, b_m) = M \left(1, \frac{1}{b_1^2}, \dots, \frac{1}{b_m^2} \right)$ and factor it out from the summation. ■

Example 20: The weight function corresponding to HMR is given by

$$g(b_1, \dots, b_m) = \frac{m+1}{1+b_1^2+\dots+b_m^2}. \quad (31)$$

Example 21: The weight function corresponding to GMR is given by

$$g(b_1, \dots, b_m) = (b_1 \cdot \dots \cdot b_m)^{-2/(m+1)}. \quad (32)$$

Example 22: The weight function corresponding to AMR is given by

$$g(b_1, \dots, b_m) = \frac{1}{m+1} \left(1 + \frac{1}{b_1^2} + \dots + \frac{1}{b_m^2} \right). \quad (33)$$

Example 23: The weight function corresponding to the k th selection mean is given by

$$g(b_1, \dots, b_m) = \frac{1}{b_k^2}. \quad (34)$$

Example 24: The weight function corresponding to the power mean is given by

$$g_p(b_1, \dots, b_m) = \left(\frac{1}{m+1} \left(1 + b_1^{-2p} + \dots + b_m^{-2p} \right) \right)^{1/p}. \quad (35)$$

Example 25: The weight function corresponding to the weighted arithmetic mean is

$$g(b_1, \dots, b_m) = \alpha_0 + \alpha_1 b_1^{-2} + \dots + \alpha_m b_m^{-2}. \quad (36)$$

Example 26: The weight function corresponding to the weighted geometric mean is given by

$$g(b_1, \dots, b_m) = b_1^{-2\beta_1} \cdot \dots \cdot b_m^{-2\beta_m}. \quad (37)$$

E. Solving for the Generalized Regression Coefficients

The fundamental practical question of multivariate generalized regression is how to solve for the coefficients b_1, \dots, b_m . The next theorem describes the procedure in general. The procedure is applied further on to produce the specific regression equations in three and four variables for several cases of interest.

Theorem 27: (System of Equations for Generalized Regression Coefficients) Let E denote the ordinary least-squares error function, $g_k = \partial g / \partial b_k$, $\nabla g = (g_1, \dots, g_m)^T$ and

$$F = \mathbf{b}^T \mathbf{S} \mathbf{b} - 2\mathbf{b}^T \mathbf{s}_0 + \sigma_{00}. \quad (38)$$

Then the vector $\mathbf{b} = (b_1, \dots, b_m)^T$ of regression coefficients is obtained by solving the nonlinear matrix-vector equation

$$F \nabla g + 2g(\mathbf{S} \mathbf{b} - \mathbf{s}_0) = \mathbf{0} \quad (39)$$

for \mathbf{b} and $b_0 = \mu_0 - \mathbf{b}^T \boldsymbol{\mu}$. Explicitly, one solves the nonlinear system

$$g_k F + g F_k = 0 \quad (40)$$

for b_1, \dots, b_m where $k = 1 \dots m$,

$$F = \sum_{j=1}^m \sum_{k=1}^m \sigma_{jk} b_j b_k - 2 \sum_{k=1}^m \sigma_{k0} b_k + \sigma_{00} \quad (41)$$

and $F_k = \partial F / \partial b_k$ is given by

$$F_k = 2 \sum_{j=1}^m \sigma_{jk} b_j - 2\sigma_{k0}. \quad (42)$$

Proof: Let $E_g = gE$ be the generalized regression error where $E = E_{OLS}$. Use the matrix-vector form of the error, take the gradient of the error with respect to \mathbf{b} and set it equal to zero.

$$\begin{aligned}\nabla(gE) &= \mathbf{0} \\ E\nabla g + g\nabla E &= \mathbf{0}\end{aligned}$$

Substitute

$$E = \mathbf{b}^T \mathbf{S} \mathbf{b} - 2\mathbf{b}^T \mathbf{s}_0 + \sigma_{00} + (b_0 - \mu_0 + \mathbf{b}^T \boldsymbol{\mu})^2$$

and

$$\nabla E = 2(\mathbf{S} \mathbf{b} - \mathbf{s}_0) - 2\boldsymbol{\mu}(b_0 - \mu_0 + \mathbf{b}^T \boldsymbol{\mu}).$$

Set $b_0 = \mu_0 - \mathbf{b}^T \boldsymbol{\mu}$, and obtain

$$2g(\mathbf{S} \mathbf{b} - \mathbf{s}_0) + F\nabla g = \mathbf{0} \quad (43)$$

Alternatively, take the partial derivative of E_g with respect to b_k for $k = 1 \dots m$ and set the resulting expressions equal to zero. ■

F. The Hessian Matrix and Determinant

In order for the regression coefficients b_0, \dots, b_m to minimize the error function and be admissible, the Hessian matrix of second order partial derivatives must be positive definite when evaluated at b_0, \dots, b_m . The general Hessian matrix is calculated next. As in the bivariate case, certain combinations of g and its first and second partial derivatives appear in the matrix. One combination is denoted here by J_{jk} and another is denoted by G_{jk} . They are called indicative functions.

Definition 28: Define the indicative functions

$$J_{jk} = \frac{g_{jk}}{g} - \frac{2g_j g_k}{g^2} \quad (44)$$

and

$$G_{jk} = \frac{2g_j}{g} - \frac{g_{jk}}{g_k}. \quad (45)$$

The two indicative functions are related by the equation $G_{jk} F_k = J_{jk} F$.

Theorem 29: (Hessian matrix) The Hessian matrix \mathbf{H} of second order partial derivatives of the error function given by

$$H_{(j+1)(k+1)} = \left. \frac{\partial^2}{\partial b_k \partial b_j} (gE) \right|_{(b_0, b_1, \dots, b_m)} \quad (46)$$

for $j, k = 0 \dots m$. It is computed explicitly as follows.

$$H_{11} = 2g \quad (47)$$

$$H_{1(k+1)} = H_{(k+1)1} = 2g\mu_k \quad (48)$$

$$H_{(j+1)(k+1)} = H_{(k+1)(j+1)} = g(J_{jk} F + 2\sigma_{jk} + 2\mu_j \mu_k) \quad (49)$$

Alternatively,

$$H_{(j+1)(k+1)} = g(G_{jk} F_k + 2\sigma_{jk} + 2\mu_j \mu_k). \quad (50)$$

Proof: Take the second order partial derivative

$$\begin{aligned}\frac{\partial^2}{\partial b_k \partial b_j} (gE) &= \frac{\partial}{\partial b_k} (g_j E + g E_j) \\ &= g_{jk} E + g_j E_k + g_k E_j + g E_{jk}.\end{aligned}$$

Since

$$g_k E + g E_k = 0$$

substitute $E_k = -\frac{g_k}{g} E$ and $E_j = -\frac{g_j}{g} E$ into the two middle terms, simplify, and obtain

$$\frac{\partial^2}{\partial b_k \partial b_j} (gE) = g \left(\left(\frac{g_{jk}}{g} - \frac{2g_j g_k}{g^2} \right) E + E_{jk} \right)$$

which is the first form of the Hessian. Now substitute $E = -\frac{g}{g_k} E_k$ and obtain the second form

$$\frac{\partial^2}{\partial b_k \partial b_j} (gE) = g \left(\left(\frac{2g_j}{g} - \frac{g_{jk}}{g_k} \right) E_k + E_{jk} \right).$$

As before, upon substituting for b_0 , $E = F$, $E_k = F_k$ and $E_{jk} = F_{jk} = 2\sigma_{jk} + 2\mu_j \mu_k$. ■

Theorem 30: (Hessian determinant) The Hessian determinant is given by

$$\det \mathbf{H} = g^{m+1} \det (F\mathbf{J} + 2\mathbf{S}) \quad (51)$$

where $\mathbf{J} = [J_{jk}]_{m \times m}$ and explicitly by

$$\det \mathbf{H} = g^{m+1} \begin{vmatrix} FJ_{11} + 2\sigma_{11} & \cdots & FJ_{m1} + 2\sigma_{1m} \\ \vdots & \ddots & \vdots \\ FJ_{m1} + 2\sigma_{1m} & \cdots & FJ_{mm} + 2\sigma_{mm} \end{vmatrix}. \quad (52)$$

Alternatively,

$$\det \mathbf{H} = g^{m+1} \det (\mathbf{K} + 2\mathbf{S}) \quad (53)$$

where $\mathbf{K} = [G_{jk} F_k]_{m \times m}$ and explicitly by

$$\det \mathbf{H} = g^{m+1} \begin{vmatrix} G_{11} F_1 + 2\sigma_{11} & \cdots & G_{1m} F_m + 2\sigma_{1m} \\ \vdots & \ddots & \vdots \\ G_{m1} F_1 + 2\sigma_{1m} & \cdots & G_{mm} F_m + 2\sigma_{mm} \end{vmatrix}. \quad (54)$$

Proof: Begin with the $(m+1) \times (m+1)$ determinant of \mathbf{H} and reduce it to an equivalent $m \times m$ determinant. The $(m+1) \times (m+1)$ determinant is given by

$$\det \mathbf{H} = g^{m+1} \begin{vmatrix} 2 & & 2\mu_1 \\ 2\mu_1 & G_{11} F_1 + 2\sigma_{11} + 2\mu_1^2 & \\ \vdots & \vdots & \\ 2\mu_m & G_{1m} F_1 + 2\sigma_{1m} + 2\mu_1 \mu_m & \\ \cdots & 2\mu_m & \\ \cdots & G_{1m} F_m + 2\sigma_{1m} + 2\mu_1 \mu_m & \\ \vdots & \vdots & \\ \cdots & G_{1m} F_m + 2\sigma_{mm} + 2\mu_m^2 & \end{vmatrix} \quad (55)$$

For GMR the following system of equations is obtained.

$$\begin{cases} 2\sigma_{11}b_1^2 - \sigma_{22}b_2^2 + \sigma_{12}b_1b_2 - \sigma_{10}b_1 + 2\sigma_{20}b_2 - \sigma_{00} = 0 \\ \sigma_{11}b_1^2 - 2\sigma_{22}b_2^2 - \sigma_{12}b_1b_2 - 2\sigma_{10}b_1 + \sigma_{20}b_2 + \sigma_{00} = 0 \end{cases} \quad (65)$$

For AMR the following system of equations is obtained.

$$\begin{cases} \sigma_{11}b_1^4b_2^2 + \sigma_{12}b_1^3b_2^3 - \sigma_{10}b_1^3b_2^2 + \sigma_{11}b_1^4 - \sigma_{22}b_2^4 + \sigma_{12}b_1^3b_2 \\ -\sigma_{12}b_2^3b_1 + 2\sigma_{20}b_2^2 - \sigma_{10}b_1^3 + \sigma_{10}b_2^2b_1 - \sigma_{00}b_2^2 = 0 \\ \sigma_{22}b_1^2b_2^4 + \sigma_{12}b_1^3b_2^3 - \sigma_{20}b_1^2b_2^2 - \sigma_{11}b_1^4 + \sigma_{22}b_2^4 + \sigma_{12}b_1b_2^3 \\ -\sigma_{12}b_1^3b_2 + 2\sigma_{10}b_1^3 - \sigma_{20}b_2^2 + \sigma_{20}b_1^2b_2 - \sigma_{00}b_1^2 = 0 \end{cases} \quad (66)$$

For hybrid LVR the following system of equations is obtained.

$$\begin{cases} \sigma_{11}b_1^2 - \sigma_{22}b_2^2 + 2\sigma_{20}b_2 - \sigma_{00} = 0 \\ \sigma_{11}b_1^2 - \sigma_{22}b_2^2 - 2\sigma_{10}b_1 + \sigma_{00} = 0 \end{cases} \quad (67)$$

I. Specific Regression Equations for the Case of Four Variables

The general formula for the regression coefficients is applied here to the problem of determining the coefficients in the equation

$$x_0 = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \quad (68)$$

for certain special cases. In all cases, $b_0 = \mu_0 - b_1\mu_1 - b_2\mu_2 - b_3\mu_3$.

For OLS $x_0 | \{x_1, x_2, x_3\}$ regression, which is standard ordinary least-squares, the following system of equations is obtained.

$$\begin{cases} \sigma_{11}b_1 + \sigma_{12}b_2 + \sigma_{13}b_3 = \sigma_{10} \\ \sigma_{12}b_1 + \sigma_{22}b_2 + \sigma_{23}b_3 = \sigma_{20} \\ \sigma_{13}b_1 + \sigma_{23}b_2 + \sigma_{33}b_3 = \sigma_{30} \end{cases} \quad (69)$$

For OLS $x_1 | \{x_2, x_3, x_0\}$ regression, the following system of equations is obtained.

$$\begin{cases} \sigma_{22}b_2^2 + \sigma_{33}b_3^2 + \sigma_{12}b_1b_2 + 2\sigma_{23}b_2b_3 + \sigma_{13}b_1b_3 \\ -\sigma_{10}b_1 - 2\sigma_{20}b_2 - 2\sigma_{30}b_3 + \sigma_{00} = 0 \\ \sigma_{12}b_1 + \sigma_{22}b_2 + \sigma_{23}b_3 - \sigma_{20} = 0 \\ \sigma_{13}b_1 + \sigma_{23}b_2 + \sigma_{33}b_3 - \sigma_{30} = 0 \end{cases} \quad (70)$$

For OLS $x_2 | \{x_1, x_3, x_0\}$ regression, the following system of equations is obtained.

$$\begin{cases} \sigma_{11}b_1 + \sigma_{12}b_2 + \sigma_{13}b_3 - \sigma_{10} = 0 \\ \sigma_{11}b_1^2 + \sigma_{33}b_3^2 + \sigma_{12}b_1b_2 + \sigma_{23}b_2b_3 + 2\sigma_{13}b_1b_3 \\ -2\sigma_{10}b_1 - \sigma_{20}b_2 - 2\sigma_{30}b_3 + \sigma_{00} = 0 \\ \sigma_{13}b_1 + \sigma_{23}b_2 + \sigma_{33}b_3 - \sigma_{30} = 0 \end{cases} \quad (71)$$

For OLS $x_3 | \{x_1, x_2, x_0\}$ regression, the following system of equations is obtained.

$$\begin{cases} \sigma_{11}b_1 + \sigma_{12}b_2 + \sigma_{13}b_3 - \sigma_{10} = 0 \\ \sigma_{12}b_1 + \sigma_{22}b_2 + \sigma_{23}b_3 - \sigma_{20} = 0 \\ \sigma_{11}b_1^2 + \sigma_{22}b_2^2 + 2\sigma_{12}b_1b_2 + \sigma_{23}b_2b_3 + \sigma_{13}b_1b_3 \\ -2\sigma_{10}b_1 - 2\sigma_{20}b_2 - \sigma_{30}b_3 + \sigma_{00} = 0 \end{cases} \quad (72)$$

For HMR the following system of equations is obtained.

$$\begin{cases} (\sigma_{22} - \sigma_{11})b_1b_2^2 + \sigma_{12}b_1^2b_2 + \sigma_{33}b_1b_2^3 - \sigma_{13}b_2^2b_3 \\ -\sigma_{11}b_1b_2^3 + \sigma_{13}b_1^2b_3 - \sigma_{12}b_2b_2^3 + 2\sigma_{23}b_1b_2b_3 \\ -\sigma_{12}b_2^3 - \sigma_{13}b_3^3 - \sigma_{10}b_1^2 + \sigma_{10}b_2^2 \\ +\sigma_{10}b_2^3 - 2\sigma_{20}b_1b_2 - 2\sigma_{30}b_1b_3 + (\sigma_{00} - \sigma_{11})b_1 \\ -\sigma_{12}b_2 - \sigma_{13}b_3 + \sigma_{10} = 0 \\ -\sigma_{12}b_1b_2^2 - \sigma_{33}b_2b_2^3 - 2\sigma_{13}b_1b_2b_3 - \sigma_{23}b_2^2b_3 \\ + (\sigma_{22} - \sigma_{11})b_1^2b_2 + \sigma_{22}b_2b_2^3 + \sigma_{12}b_1b_2^3 + \sigma_{23}b_1^2b_3 \\ +\sigma_{12}b_1^3 + \sigma_{23}b_3^3 - \sigma_{20}b_1^2 + \sigma_{20}b_2^2 \\ -\sigma_{20}b_2^3 + 2\sigma_{30}b_2b_3 + 2\sigma_{10}b_1b_2 + \sigma_{12}b_1 \\ + (\sigma_{22} - \sigma_{00})b_2 + \sigma_{23}b_3 - \sigma_{20} = 0 \\ -\sigma_{13}b_1b_2^3 + (\sigma_{33} - \sigma_{11})b_1^2b_3 - \sigma_{23}b_2b_2^3 - \sigma_{22}b_2^2b_3 \\ -2\sigma_{12}b_1b_2b_3 + \sigma_{23}b_3^3 + \sigma_{13}b_1^3 + \sigma_{33}b_2^2b_3 \\ +\sigma_{23}b_1^2b_2 + \sigma_{13}b_1b_2^2 - \sigma_{30}b_1^2 - \sigma_{30}b_2^2 \\ +\sigma_{30}b_2^3 + 2\sigma_{20}b_2b_3 + 2\sigma_{10}b_1b_3 \\ +\sigma_{13}b_1 + \sigma_{23}b_2 + (\sigma_{33} - \sigma_{00})b_3 - \sigma_{30} = 0 \end{cases} \quad (73)$$

For GMR, the following system of equations is obtained.

$$\begin{cases} 3\sigma_{11}b_1^2 - \sigma_{22}b_2^2 - \sigma_{33}b_3^2 + 2\sigma_{12}b_1b_2 - 2\sigma_{23}b_2b_3 \\ +2\sigma_{13}b_1b_3 - 2\sigma_{10}b_1 + 2\sigma_{20}b_2 + 2\sigma_{30}b_3 - \sigma_{00} = 0 \\ \sigma_{11}b_1^2 - 3\sigma_{22}b_2^2 + \sigma_{33}b_3^2 - 2\sigma_{12}b_1b_2 - 2\sigma_{23}b_2b_3 \\ +2\sigma_{13}b_1b_3 - 2\sigma_{10}b_1 + 2\sigma_{20}b_2 - 2\sigma_{30}b_3 + \sigma_{00} = 0 \\ \sigma_{11}b_1^2 + \sigma_{22}b_2^2 - 3\sigma_{33}b_3^2 + 2\sigma_{12}b_1b_2 - 2\sigma_{23}b_2b_3 \\ -2\sigma_{13}b_1b_3 - 2\sigma_{10}b_1 - 2\sigma_{20}b_2 + 2\sigma_{30}b_3 + \sigma_{00} = 0 \end{cases} \quad (74)$$

For AMR the following system of equations is obtained.

$$\begin{cases} \sigma_{11}b_1^4b_2^2b_3^2 + \sigma_{12}b_1^3b_2^3b_3^2 + \sigma_{13}b_1^3b_2^2b_3^3 - \sigma_{10}b_1^3b_2^2b_3^2 \\ +\sigma_{12}b_1^2b_2b_2^3 + \sigma_{13}b_1^2b_2^2b_3 - \sigma_{12}b_1b_2^2b_3^2 - \sigma_{13}b_1b_2^2b_3^3 \\ \sigma_{11}b_1^4b_2^3 - \sigma_{22}b_2^4b_3^2 - \sigma_{33}b_2^4b_3^3 - 2\sigma_{23}b_2^3b_3^3 + \sigma_{13}b_1^3b_3^3 \\ +\sigma_{11}b_1^4b_2^2 + \sigma_{12}b_1^3b_2^3 + \sigma_{10}b_1b_2^2b_3^2 + 2\sigma_{20}b_2^2b_3^2 \\ +2\sigma_{30}b_2^2b_3^3 - \sigma_{10}b_1^3b_2^3 - \sigma_{10}b_1^3b_2^2 - \sigma_{00}b_2^2b_3^2 = 0 \\ \sigma_{22}b_1^2b_2^2b_3^2 + \sigma_{23}b_1^2b_2^3b_3^3 + \sigma_{12}b_1^2b_2^3b_3^2 - \sigma_{20}b_1^2b_2^3b_3^2 \\ -\sigma_{12}b_1^2b_2b_2^3 - \sigma_{23}b_1^2b_2b_3^3 + \sigma_{23}b_1^2b_2^2b_3 + \sigma_{12}b_1b_2^3b_3^2 \\ -\sigma_{11}b_1^4b_2^3 + \sigma_{22}b_2^4b_3^2 + \sigma_{23}b_2^3b_3^3 + \sigma_{12}b_1^2b_2^3 - 2\sigma_{13}b_1^3b_3^3 \\ -\sigma_{33}b_1^2b_3^4 + \sigma_{22}b_1^2b_2^4 + \sigma_{20}b_1^2b_2b_2^3 - \sigma_{20}b_3^3b_3^2 + 2\sigma_{10}b_1^3b_2^3 \\ +2\sigma_{30}b_1^2b_3^3 - \sigma_{20}b_1^2b_2^2 - \sigma_{00}b_1^2b_2^2 = 0 \\ \sigma_{13}b_1^3b_2^2b_3^3 + \sigma_{23}b_1^2b_2^3b_3^3 + \sigma_{33}b_1^2b_2^2b_3^3 - \sigma_{30}b_1^2b_2^2b_3^3 \\ -\sigma_{13}b_1^2b_2^2b_3 + \sigma_{13}b_1b_2^2b_3^3 - \sigma_{23}b_1^2b_2^2b_3 + \sigma_{23}b_1^2b_2b_3^3 \\ +\sigma_{33}b_2^2b_3^4 + \sigma_{23}b_2^3b_3^3 + \sigma_{13}b_1^3b_3^3 - \sigma_{11}b_1^4b_2^2 - 2\sigma_{12}b_1^2b_2^2 \\ +\sigma_{33}b_1^2b_3^4 - \sigma_{22}b_1^2b_2^4 - \sigma_{30}b_2^2b_3^3 + 2\sigma_{10}b_1^3b_2^2 - \sigma_{30}b_1^2b_3^3 \\ +2\sigma_{20}b_1^2b_2^2 + \sigma_{30}b_1^2b_2^2b_3 - \sigma_{00}b_1^2b_2^2 = 0 \end{cases} \quad (75)$$

For hybrid LVR the following system of equations is obtained.

$$\begin{cases} \sigma_{11}b_1^2 - \sigma_{22}b_2^2 - \sigma_{33}b_3^2 - 2\sigma_{23}b_2b_3 \\ +2\sigma_{20}b_2 + 2\sigma_{30}b_3 - \sigma_{00} = 0 \\ \sigma_{11}b_1^2 - \sigma_{22}b_2^2 + \sigma_{33}b_3^2 + 2\sigma_{13}b_1b_3 \\ -2\sigma_{10}b_1 - 2\sigma_{30}b_3 + \sigma_{00} = 0 \\ \sigma_{11}b_1^2 + \sigma_{22}b_2^2 - \sigma_{33}b_3^2 + 2\sigma_{12}b_1b_2 \\ -2\sigma_{10}b_1 - 2\sigma_{20}b_2 + \sigma_{00} = 0 \end{cases} \quad (76)$$

III. NUMERICAL EXAMPLES

Example 34: (Cement Data) This example is taken from Hald's statistics text [8] (p. 636). The data describe the heat evolved in the curing of cement as a function of the percentage in weight of certain compounds in the mixture. The y variable, called here x_0 , is the heat measured in calories per gram.

The variables x_1 and x_2 are the percentages by weight of two different cement compounds. The data are as follows.

x_1	7	1	11	11	7	11	3	1	2	21	1	11	10
x_2	26	29	56	31	52	55	71	31	54	47	40	66	68
x_0	78.5	74.3	104.3	87.6	95.9	109.2	102.7	72.5	93.1	115.9	83.8	113.3	109.4

The reader can verify that $\mu_0 = 95.42308$, $\sigma_{00} = 208.90485$,

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 7.46154 \\ 48.15385 \end{bmatrix}, \quad (77)$$

$$\mathbf{s}_0 = \begin{bmatrix} \sigma_{10} \\ \sigma_{20} \end{bmatrix} = \begin{bmatrix} 59.68935 \\ 176.38106 \end{bmatrix}, \quad (78)$$

and

$$\mathbf{S} = \begin{bmatrix} 31.94082 & 19.31361 \\ 19.31360 & 223.51479 \end{bmatrix}. \quad (79)$$

The data are well-conditioned: $\text{cond } \mathbf{S} = 7.51147$. Also $\det \mathbf{S} = 6.76623 \times 10^3$. The standard OLS $x_0 | \{x_1, x_2\}$ regression plane is given by

$$x_0 = 52.5773 + 1.4683x_1 + 0.6623x_2. \quad (80)$$

The OLS $x_1 | \{x_2, x_0\}$ regression plane is given by

$$x_0 = 52.2466 + 1.5685x_1 + 0.6536x_2. \quad (81)$$

The OLS $x_2 | \{x_1, x_0\}$ regression plane is given by

$$x_0 = 51.1917 + 1.4491x_1 + 0.6940x_2. \quad (82)$$

The HMR plane is given by

$$x_0 = 52.2110 + 1.5271x_1 + 0.6607x_2. \quad (83)$$

The GMR plane is given by

$$x_0 = 52.0069 + 1.4950x_1 + 0.6700x_2. \quad (84)$$

The AMR plane is given by

$$x_0 = 51.7164 + 1.4699x_1 + 0.6799x_2. \quad (85)$$

The hybrid LVR plane is given by

$$x_0 = 51.7166 + 1.5085x_1 + 0.6739x_2. \quad (86)$$

This example suggests that all the methods yield regression planes that are reasonably close to each other when the data are well-conditioned. The next example illustrates that when the data are ill-conditioned, ordinary least-squares can perform poorly while generalized least-squares methods can still perform well.

Example 35: This example is chosen from the work of Tofallis [10], [11] where the data are used to compare least-volume regression to ordinary least-squares. The data are from a model problem in Belsley's book on collinearity [1] (p. 5) and are ill-conditioned.

Suppose an underlying linear model of some physical relationship is known and given by

$$x_0 = 1.2 - 0.4x_1 + 0.6x_2 + 0.9x_3 + \varepsilon \quad (87)$$

where ε has a normal distribution with mean 0 and variance 0.01. Suppose two persons A and B wish to estimate the

linear relationship for themselves. Suppose they both share the same x_0 data but they take independent measurements of variables x_1, x_2 , and x_3 . Their results are presented in a table.

A								
x_1	-3.138	-0.297	-4.582	0.301	2.729	-4.836	0.065	4.102
x_2	1.286	0.25	1.247	0.498	-0.28	0.35	0.208	1.069
x_3	0.169	0.044	0.109	0.117	0.035	-0.094	0.047	0.375
x_0	3.3979	1.6094	3.7131	1.6767	0.0419	3.3768	1.1661	0.4701
B								
x_1	-3.136	-0.296	-4.581	0.300	2.730	-4.834	0.064	4.103
x_2	1.288	0.251	1.246	0.498	-0.281	0.349	0.206	1.069
x_3	0.170	0.043	0.108	0.118	0.036	-0.093	0.048	0.376
x_0	3.3979	1.6094	3.7131	1.6767	0.0419	3.3768	1.1661	0.4701

The goal is to try and recover the actual coefficients b_0, b_1, b_2 and b_3 from the data using regression.

The reader can verify that for Person A, $\mu_0 = 1.93150$, $\sigma_{00} = 1.73426$,

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} -0.70700 \\ 0.57850 \\ 0.10025 \end{bmatrix}, \quad (88)$$

$$\mathbf{s}_0 = \begin{bmatrix} \sigma_{10} \\ \sigma_{20} \\ \sigma_{30} \end{bmatrix} = \begin{bmatrix} -3.86706 \\ 0.40169 \\ -0.048495 \end{bmatrix}, \quad (89)$$

and

$$\mathbf{S} = \begin{bmatrix} 9.33584 & -0.55747 & 0.20635 \\ -0.55747 & 0.27862 & 0.04081 \\ 0.20635 & 0.04081 & 0.01607 \end{bmatrix}. \quad (90)$$

The data suffer from multicollinearity, which is the near linear dependence of one of the variables on the remaining variables. This is evidenced by the high condition number of the covariance matrix: $\text{cond } \mathbf{S} = 3.54374 \times 10^7$. Also note that the determinant is nearly singular: $\det \mathbf{S} = 6.35047 \times 10^{-7}$.

For Person B, $\mu_0 = 1.93150$, $\sigma_{00} = 1.73426$,

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} -0.70625 \\ 0.57825 \\ 0.10075 \end{bmatrix}, \quad (91)$$

$$\mathbf{s}_0 = \begin{bmatrix} \sigma_{10} \\ \sigma_{20} \\ \sigma_{30} \end{bmatrix} = \begin{bmatrix} -3.86644 \\ 0.40204 \\ -0.04886 \end{bmatrix}, \quad (92)$$

and

$$\mathbf{S} = \begin{bmatrix} 9.33332 & -0.55747 & 0.20721 \\ -0.55747 & 0.27918 & 0.04078 \\ 0.20721 & 0.04078 & 0.01609 \end{bmatrix}. \quad (93)$$

Again the covariance matrix is ill-conditioned and nearly singular: $\text{cond } \mathbf{S} = 1.28134 \times 10^7$ and $\det \mathbf{S} = 1.75906 \times 10^{-6}$.

The standard OLS $x_0 | \{x_1, x_2, x_3\}$ regression planes for the data of Person A and Person B are

$$\text{A : } x_0 = 1.2546 + 0.9741x_1 + 9.0219x_2 - 38.4400x_3 \quad (94)$$

$$\text{B : } x_0 = 1.2752 + 0.2470x_1 + 4.5116x_2 - 17.6486x_3. \quad (95)$$

The discrepancy between the OLS regression coefficients and the model coefficients is striking. The OLS regression

coefficients disagree with the model coefficients in both sign and magnitude. Therefore OLS regression does not appear to be useful in a case such as this. In comparison, several generalized regression methods presented next appear to do a much better job in recovering the model coefficients, agreeing with the model coefficients in both sign and magnitude.

The AMR planes are

$$A : x_0 = 1.2479 - 0.4069x_1 + 0.5151x_2 + 0.9767x_3 \quad (96)$$

$$B : x_0 = 1.2478 - 0.4070x_1 + 0.5151x_2 + 0.9766x_3. \quad (97)$$

The GMR planes are

$$A : x_0 = 1.2440 - 0.4345x_1 + 0.3438x_2 + 1.8095x_3 \quad (98)$$

$$B : x_0 = 1.2433 - 0.4348x_1 + 0.3437x_2 + 1.8100x_3. \quad (99)$$

The hybrid LVR planes are

$$A : x_0 = 1.2206 - 0.4370x_1 + 0.3617x_2 + 1.9225x_3 \quad (100)$$

$$B : x_0 = 1.2200 - 0.4372x_1 + 0.3613x_2 + 1.9231x_3. \quad (101)$$

The LVR planes are computed by Tofallis as

$$A : x_0 = 1.20 - 0.43x_1 + 0.37x_2 + 1.97x_3 \quad (102)$$

$$B : x_0 = 1.20 - 0.43x_1 + 0.37x_2 + 1.98x_3. \quad (103)$$

In this example, AMR comes the closest to recovering the model coefficients. GMR, hybrid LVR and LVR appear to perform comparably well. The calculations of the three remaining OLS regressions and HMR require further study. They appear to have a negative Hessian determinant, making them inadmissible. The hybrid LVR coefficients obtained here are in good agreement with the LVR coefficients presented by Tofallis, differing only in the hundredths place. This suggests that hybrid LVR can be a useful least-squares alternative to LVR.

IV. SUMMARY

The extension of the bivariate theory of generalized least-squares to multivariate regression is begun in this paper. The multivariate symmetric least-squares problem in $m + 1$ variables seeks an m dimensional hyperplane which minimizes the average generalized mean of the square deviations between the data and hyperplane in each of the variables. The weighted multivariate ordinary least-squares problem in $m + 1$ variables is also defined using an explicit formula for the ordinary least-squares error. As in the bivariate case, every symmetric least-squares problem is shown to be equivalent to a weighted ordinary least-squares problem. The weight function $g(b_1, \dots, b_m)$ characterizes the regression method. The multivariate generalized least-squares error is then a product of the weight

function g and the explicit multivariate error function. Partial derivatives of this analytic expression for the error are then taken with respect to each of the coefficients b_1, \dots, b_m and set equal to zero. The result is a nonlinear system of equations in b_1, \dots, b_m involving only the covariances σ_{jk} which can then be solved to yield the regression coefficients for any generalized least-squares method. In order for the solution to minimize the error function and be admissible, the Hessian matrix must be computed and found to be positive definite.

The specific system of equations for the regression coefficients is presented for OLS, HMR, GMR, and AMR, for three and four variables. A related least-squares alternative to Tofallis' least-volume regression (LVR) called hybrid LVR is presented here as well.

Numerical evidence suggests that when the data are ill-conditioned ordinary least-squares regressions may not succeed in uncovering an underlying linear model. In comparison, certain generalized least-squares methods can come closer to uncovering the model coefficients.

REFERENCES

- [1] D. A. Belsley, *Conditioning Diagnostics*, New York: John Wiley & Sons, 1991.
- [2] N. R. Draper and Y. Yang, "Generalization of the mean functional relationship," *Computational Statistics and Data Analysis*, vol. 23, pp. 355-372, 1997.
- [3] S. C. Ehrenberg, "Deriving the Least-Squares Regression Equation," *The American Statistician*, vol. 37, p.232, Aug. 1983.
- [4] N. Greene, "Generalized Least-Squares Regressions I: Efficient Derivations," in *Proceedings of the 1st International Conference on Computational Science and Engineering (CSE'13)*, Valencia, Spain, 2013, pp. 145-158.
- [5] N. Greene, "Generalized Least-Squares Regressions II: Theory and Classification," in *Proceedings of the 1st International Conference on Computational Science and Engineering (CSE '13)*, Valencia, Spain, 2013, pp. 159-166.
- [6] N. Greene, "Generalized Least-Squares Regressions III: Further Theory and Classification," in *Proceedings of the 5th International Conference on Applied Mathematics and Informatics (AMATHI '14)*, Cambridge, MA, 2014, pp. 34-38.
- [7] N. Greene, "Generalized Least-Squares Regressions IV: Theory and Classification Using Generalized Means," in *Mathematics and Computers in Science and Industry*, Varna, Bulgaria, 2014, pp. 19-35.
- [8] A. Hald, *Statistical Theory with Engineering Applications*, New York: John Wiley & Sons, 1952.
- [9] C. Tofallis, "Model Fitting Using the Least Volume Criterion," in *Algorithms for Approximation IV*, J.C. Mason and J. Levesly, Eds., University of Huddersfield Press, 2002.
- [10] C. Tofallis, "Model Fitting for Multiple Variables by Minimizing the Geometric Mean Deviation," in *Total least squares and errors-in-variables modelling: algorithms analysis and applications*, S. Van Huffel and P. Lemmerling, Eds., Dordrecht: Kluwer Academic Publishers, 2002.
- [11] C. Tofallis, "Multiple Neutral Data Fitting," *Annals of Operations Research*, vol. 124, pp. 69-79, 2003.

New Computational Methods for Spectrometer Signal Analysis

Petra Perner

Abstract—Different spectrometer methods exist that have been developed over time to practical applicable systems. Researchers in different fields try to apply these methods to different applications especially in the chemical and biological area. One of these methods is RAMAN spectroscopy for protein crystallization or Mid-Infrared spectroscopy for biomass identification. For the applications are required robust and machine learnable automatic signal interpretation methods. These methods should take into account that not so much spectrometer data about the application are available from scratch and that these data need to be learnt while using the spectrometer system. We propose to represent the spectrometer signal by a sequence of 0/1 characters obtained from a specific Delta Modulator. This prevents us from a particular symbolic description of peaks and background. The interpretation of the spectrometer signal is done by searching for a similar signal in a constantly increasing data base. The comparison between the two sequences is done based on a syntactic similarity measure. We describe in this paper how the signal representation is obtained by Delta Modulation, the similarity measure for the comparison of the signals and give results for searching the data base.

Keywords— Computational Methods, Delta Modulation, Feature Extraction, Incremental Knowledge Acquisition, Spectrometer signal analysis, Similarity-based Signal Interpretation

I. INTRODUCTION

Different spectrometer methods exist that have been developed over time to practical applicable systems. Researchers in different fields try to apply these methods to different applications especially in the chemical and biological area. One of these methods is RAMAN spectroscopy for protein crystallization [1], [2] or Mid-Infrared spectroscopy for biomass identification [3].

Databases that allow the extraction of the chemical and biological compounds are built up based on different spectrometer methods [4]. With the rapid increase in accessible data from prior experiments and the development of spectrometer control software that supports large inclusion lists for targeted analyses, the use of search and matching strategies

This work has been sponsored by the Federal Ministry of Economic Affairs BMWI under the grant title „Marker Free Raman-Screening for the Molecular Investigation of biological Interactions MARAS” grant number 16IN0477.

Petra Perner is the director of the Institute of Computer Vision and Applied Computer Sciences IBaI, Kohlenstrasse 2, 04107 Leipzig Germany (phone +49 341 8612273; fax: +49 341 8612275; e-mail: pperner@ibai-institut.de)

can be expected to increase [5]. Peak detection in spectrometer signal data can be done based on statistics such as on the Welch’s t-test [6], on Fuzzy logic [7], on curve fitting based on the Levenberg–Marquardt algorithm [8], or on correlation [9]. Several heuristic and probabilistic algorithms for peak detection are described and evaluated in [10]. The results show that there is no unique algorithm for peak detection.

For the applications are required robust and machine learnable automatic signal interpretation methods. These methods should take into account the sparse available data for the application and that new data need to be acquired while using the spectrometer system. We propose a novel spectrometer analysis method based on Delta Modulation and similarity determination. We represent the spectrometer signal by a sequence of 0/1 characters obtained from a specific Delta Modulator. While doing this we preprocess the signal by smoothing at the same time. This prevents us from the extraction of a specific symbolic description of peaks and background from the basic spectrometer signal based on signal-theoretic methods [11]. The interpretation of the spectrometer signal is done by searching for a similar signal in a constantly increasing data base. The two 0/1 sequences of the spectrometer signal are compared based on a syntactic similarity measure.

The proposed new method has been tested on RAMAN spectrometer signals for screening of bio-molecular interactions but the method can be used for all kinds of spectrometer signals. With the aid of Raman spectroscopy, the vibrational spectrum of molecules can be examined. Functional groups like amino, carboxyl or hydroxyl groups can be identified through characteristic vibrational frequencies.

In this paper the architecture of the spectrometer-signal analysis system is described in Section II. The calculation of the signal representation obtained by Delta Modulation is explained in Section III for three different kinds of delta modulation methods. Then we describe three different syntactical dissimilarity measures used for this study in Section IV. Finally, we give results in Section V for the three Delta Modulation methods and select the best one. This method is used for further studying the best dissimilarity measure. We show how good these measures can group similar spectra. At the end we use a prototype-based classifier to show how good we can classify the spectra based on the chosen representation and with the three different dissimilarity measures. In Section VI we give

conclusion.

II. ARCHITECTURE OF THE AUTOMATED SPECTROMETER SIGNAL PROGRAM

The architecture of the automatic spectrometer identification system is shown in Fig. 1.

After preprocessing the spectrometer-signal, the signal is coded into a 0/1 sequence by the delta modulator. While doing that the signal is step-wise smoothed by a linear function. The representation makes it unnecessary to develop special high-level features that describe all interesting properties of the spectra. The sequence itself can be interpreted in different ways. It can be ask for identity, similarity of the whole sequence or for partial identity or similarity. That allows identifying part-spectra, special single peaks or peak combinations within spectra.

This sequence is compared to sequences of reference spectra stored in a memory. The name of the spectrum where the coded sequence gives the highest similarity is given as output to the user. A side effect of the coding is also that the spectra is not stored with its real values but instead it is stored as 0/1 sequence. This saves memory capacity and makes it possible to implement the method into a special purpose processor.

When there is no similar sequence in the data base the input spectrum is stored into the data base after it has been coded by the delta modulator. The spectrum is labeled manually after it has been checked by other method what the spectrum is about. This data collection is necessary since the appearance of the spectra for different proteins is not known yet.

The pre-processing of the RAMAN spectra is in this special case a baseline correction [12], a Fourier transformation to eliminate the influence of the special system device and its parts [13], and the calculation of the difference between the spectrum of the buffer and the liquid in the buffer.

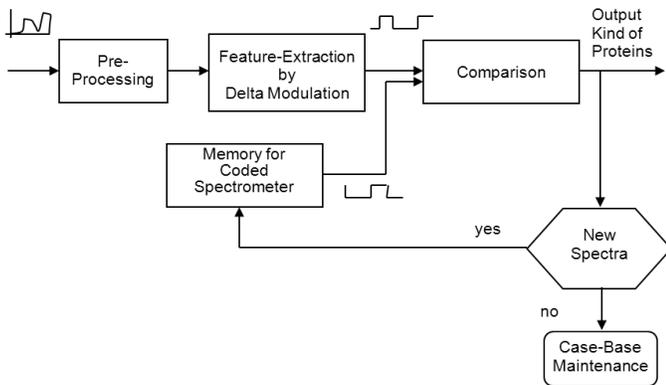


Fig. 1. Architecture of the Spectrum Interpretation System

III. REPRESENTATION OF THE SPECTRA BY DELTA MODULATION

The delta modulator compares the actual signal value $s(i)$ with an estimated signal value $r(i)$ of the coder. The difference $e(i)$ between these two signals is coded by only one bit. It mainly represents if the signal was increased or decreased by a certain constant. Three different methods exist to estimate actual signal

value: Linear Delta Modulation (LDM) [14], Constant Factor Delta Modulation (CFDM) [15], and Continuously Variable Slope Delta Modulator (CVSD) [16], [17], [18].

A. Linear Delta Modulation

In case of the Linear Delta Modulation, the difference $e(i)$ between the actual signal value $s(i)$ and the estimated signal value $r(i)$ at sampling point i is calculated, see Fig. 2:

$$e(i) = s(i) - r(i) \quad (1)$$

If the difference is positive then the code D is equal "1" and D is equal "0" if the difference is negative. This binary signal D_n is stored in the memory. At the same time the magnitude of the signal to be expected at the next sampling point i is estimated from it. The corresponding rule is:

$$s(i) > r(i). D_n = 0. r = r(i - 1) + \Delta u \quad (2)$$

$$s(i) \leq r(i). D_n = 1. r = r(i - 1) - \Delta u \quad (3)$$

The incremental size Δu is a constant value which has to be selected in function of the standard-deviation δ_Δ of the first-order difference signal: $\Delta(i) = s(i) - s(i - 1)$.

On the reproduction side (which is not necessary here since we do not want to reconstruct the signal) an inversely functioning decoder then generates the original curve by means of the binary signal stored in the memory. This approximated signal is $s'(i)$. The difference between the original signal $s(i)$ and the approximated signal $s'(i)$ is the approximation error $\varepsilon(i) = s(i) - s'(i)$, see Fig. 3.

When process dynamics change, the linear delta modulator is not adjusted optimally anymore and the reconstruction error is increasing strongly. The adaptive delta modulators compensate this disadvantage. They dispose of a function block which takes over the control of the incremental size Δu in accordance with process dynamics. In the literature different adaptive delta modulators are known, two of which are presented in the following section.

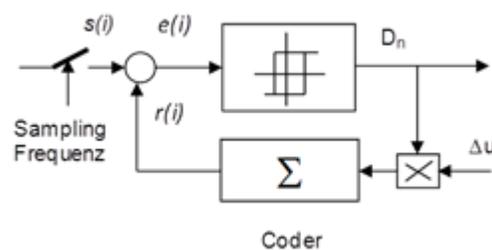


Fig. 2. Linear Delta Modulator

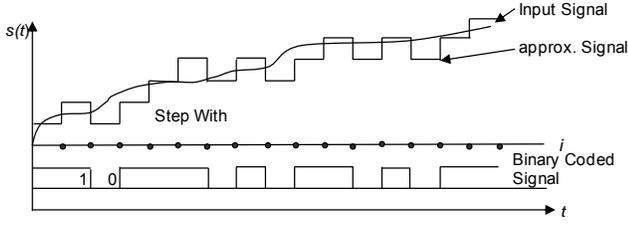


Fig. 3. Diagram with Input Signal, approximated Signal, and Binary Coded Signal

B. Constant Factor Delta Modulation

The instantaneous-value compander, also called “Constant Factor Delta Modulator” (CFDM), changes its increment size at each sampling point.

An adaptation-logic decides based on the input signal sequence $\langle D_n, D_{n-1} \rangle$ by which factor k the preceding increment size has to be multiplied:

$$\Delta u_i = \Delta u_{i-1} * k \quad (4)$$

with $D_n = D_{n-1}$ then $k = P$ and $D_n \neq D_{n-1}$ then $k = Q$.

It needs to be $P * Q = 1$, in order to observe the stability condition. For speech signals are the values $P = 1.5$ and $Q = 0.66$ known from the literature that have also been shown good performance in case of the application presented in this paper.

C. Continuously Variable Slope Delta Modulator CVSDM

The syllable compander, also called Continuously Variable Slope Delta Modulator (CVSDM), pursues, in contrast to the instantaneous-value compander, the tendency of the signal. Only when the same state has been recorded three times in a coincidence-register $\langle D_n = D_{n-1} = D_{n-2} \rangle$, the syllable compander increases its increment size. It is therefore more inert than the instantaneous-value compander. The rule for syllable companding is:

$$3 \text{ bit coincidence } k = 1; \Delta u_i = \Delta u_{i-1} + 1 \quad (5)$$

$$\text{no coincidence } k = 0; \Delta u_i = \Delta u_{i-1} - 1 \text{ until } \Delta u_i = 0 \quad (6)$$

As Δu must not become zero, a minimum increment size u_{min} larger than l needs to be added. As standard value u_{min} can be assumed as $u_{min} = \sqrt{\delta_\Delta}$.

IV. SIMILARITY DETERMINATION BETWEEN TWO SPECTRA

The spectra are represented by 0/1 sequences. To compare different spectra we need a distance measure that can work on such kind of representation. Different measures are known from text comparison and DNA sequence analysis. We choose for this work the Hamming distance [19], the Levenstein distance [20] and the Levenstein-Damerau distance [21].

A. Hamming Distance

The representation of a spectrum A and spectrum B is illustrated in Table I.

TABLE I Representation of two Spectra, Sampling Points, and XOR connection

Spectrum A	1	0	0	0	1	1	0	0	0	...
Spectrum B	1	1	1	0	1	1	0	0	0	...
Sampling Points i	1	2	3	4	5	6	7	8	9	...
A XOR B	1	0	0	1	1	1	1	1	1	...

We assume that all spectra have the same length n and that the peaks are stable at their position (wavelength) in the spectra.

Then we have to compare two sequences A and B . The distance d between these two binary representations is the number of bits in which the two vectors are different. That is the well-known Hamming Distance:

$$d(A, B) = \|A - B\| = \sum_{i=1}^n |A_i - B_i| \quad (7)$$

B. Levenshtein Distance

Let $d_L(A, B) = D_{m,n}/n$ be the Levenshtein-Distance between the two 0/1 sequence A and B with $m = |A|$ and $n = |B|$. The Levenshtein distance is defined as the minimum number of modifications needed to transform the sequence A into B . The allowed operations are substitutions, insertions, and deletions. The dissimilarity in $D_{0,0}$ should be $D_{0,0} = 0$. Then the dissimilarity is calculated as follows:

$$D_{i,0} = i, 1 \leq i \leq m$$

$$D_{0,j} = j, 1 \leq j \leq n$$

$$D_{i,j} = \min \begin{cases} D_{i-1,j-1} + 0 & \text{if } A_i = B_j \\ D_{i-1,j-1} + 1 & \text{Substitution} \\ D_{i,j-1} + 1 & \text{Insertion} \\ D_{i-1,j} + 1 & \text{Deletion} \end{cases} \quad (8)$$

for $1 \leq i \leq m, 1 \leq j \leq n$.

C. Damerau-Levenshtein-Distance

Let $D_{DL}(A, B) = D_{m,n}/n$ be the Damerau-Levenshtein-distance between the two 0/1 sequences A and B with $m = |A|$ and $n = |B|$. The Damerau-Levenshtein distance is defined as the minimum number of modifications needed to transform the sequence A into B . Besides substitution, insertion, and deletion of a single character are allowed exchange of two adjacent single characters. The dissimilarity in $D_{0,0}$ should be $D_{0,0} = 0$. Then the dissimilarity is calculated as follow:

$$D_{i,0} = i, 1 \leq i \leq m$$

$$D_{0,j} = j, 1 \leq j \leq n$$

$$D_{i,j} = \min \begin{cases} D_{i-1,j-1} + 0 & \text{if } A_i = B_j \\ D_{i-1,j-1} + 1 & \text{Substitution} \\ D_{i,j-1} + 1 & \text{Insertion} \\ D_{i-1,j} + 1 & \text{Deletion} \end{cases} \quad (9)$$

for $(1 \leq i \leq 2, 1 \leq j \leq n)$ or $(1 \leq i \leq m, 1 \leq j \leq 2)$

$$D_{i,j} = \min \begin{cases} D_{i-1,j-1} + 0 & \text{if } A_i = B_j \\ D_{i-1,j-1} + 1 & \text{Substitution} \\ D_{i,j-1} + 1 & \text{Insertion} \\ D_{i-1,j} + 1 & \text{Deletion} \\ D_{i-2,j-2} + c & \text{Exchange if} \\ & A_i = B_{j-1} \text{ and } A_{i-1} = B_j \end{cases} \quad (10)$$

for $3 \leq i \leq m, 3 \leq j \leq n$.

V. EVALUATION AND RESULTS

We have a data set of 30 different spectrometer signals. Each of the spectrometer signals have been preprocessed according to the methods described in Section II, and afterwards processed and coded based on the delta modulation (see Sect. III). The final outcome is a 0/1 sequence. The achieved results for the representation are presented in Section V.A.

We calculated the pairwise distances between the thirty signals based on three distance measures: Hamming distance, Levenshtein distance, and the Damerau-Levenshtein distance. We used the single-linkage clustering method to evaluate the goodness of the measures in Section V.B.

A. Representation of the Spectrometer Signal by Delta-Modulation

The representation of the real signal by the approximated signal of the delta modulator is exemplary shown in Fig. 4 for Linear Delta Modulation and in Fig. 5 for Constant Factor Delta Modulation. The binary coded signal for both methods is shown in TABLE II. It can be seen that the coded signal is different depending on the used delta modulation method. TABLE III shows the mean and maximum approximation error between the input signal and the approximated signal by the delta modulator.

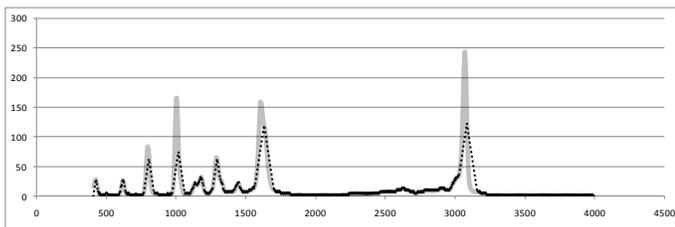


Fig. 4. Representation of Benzoic acid using LDM

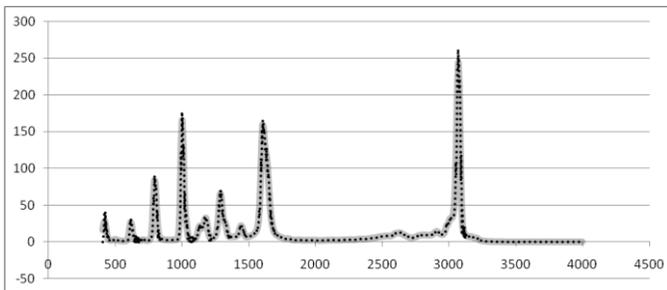


Fig. 5. Representation of Benzoic acid using CFDM

TABLE II Binary Representation of the Spectrum of Benzoic acid

Name of Compander	Sequence of Spectrum
LDM	... 10010100101010101010101010101010101010101001 ...
CFDM	... 0100100100100101001010110011110111011100 ...

TABLE III. Mean and Maximum Approximation Error between Input Signal and approx. Signal

Substance	Name of Delta Modulator			
	Linear Delta Modulator		CFDM	
	mean ϵ	max ϵ	mean ϵ	max ϵ
Acetone	1,74955958	4,642862	0,45178963	5,275991
Ascorbic acid	2,19114882	13,715031	1,06728145	10,356945
Benzamide	1,7514159	4,282954	0,40339027	2,977274
Benzoic acid	16,8602393	147,708368	4,78830105	45,56784
...
mean	5,6380909	42,5873038	1,6776906	16,045

As expected the CFMD method shows the best result. The mean error is 1.677 increments and the maximum error is 16.04 increments. In the recent settings the CVSDM gave the worst results. It is left for further work to improve this method.

In this study, we chose the CFMD method for the representation of the spectrometer signal

B. Results for Similarity between two Spectra

It has been shown in Section V.A that the CFDM delta modulator gives the best result for calculating the 0/1 sequence of the signal. The dendrogram for the different similarity measures between the thirty different spectra are shown in Fig. 6-8. The Hamming distance shows the highest differences in similarity but does not represent the similar groups well (see Fig. 6). The similarity measure will be sensitive to small changes in the spectra that might be caused by noise. Much better are represented similar groups in case of the Levenshtein (Fig. 7) and Damerau-Levenshtein similarity (Fig. 8).

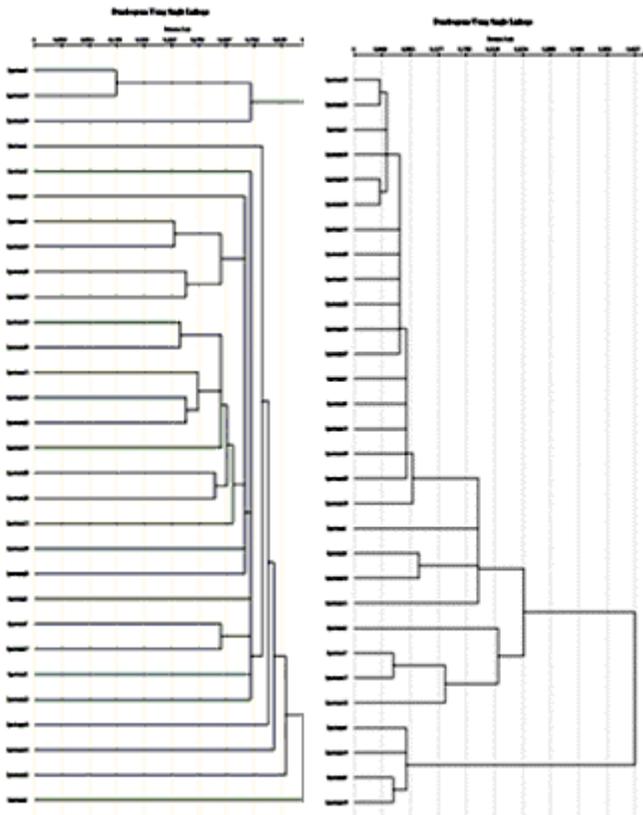


Fig. 6. Dendrogram using Hamming Distance using CFDM

Fig. 7. Dendrogram using Levenshtein Distance using CFDM

Both dendrograms show similarity in the group structure. They only slightly differ in the representation of the large group at the top of the dendrogram but in general the group structure is preserved.

C. Accuracy of the Classification

We enlarged the data base by ten samples from the same spectrum. The final data base consists of three hundred samples. Our prototype-based classifier PROTOCLASS [22] was used for classification where 299 samples were the prototypes and one sample was classified against the 299 samples by searching for the three nearest neighbors. Cross-validation was used for calculating the error rate. The results are show in TABLE IV.

TABLE IV. Accuracy of prototype-based classifier for the different similarity measures

Distance Measure	Accuracy in %
Hamming	85,2
Levenshtein	90,5
Damerau-Levenshtein	91,2

The best results we have got for the Damerau-Levenshtein distance followed by the Levenshtein distance. The worst result we have got for the Hamming distance.

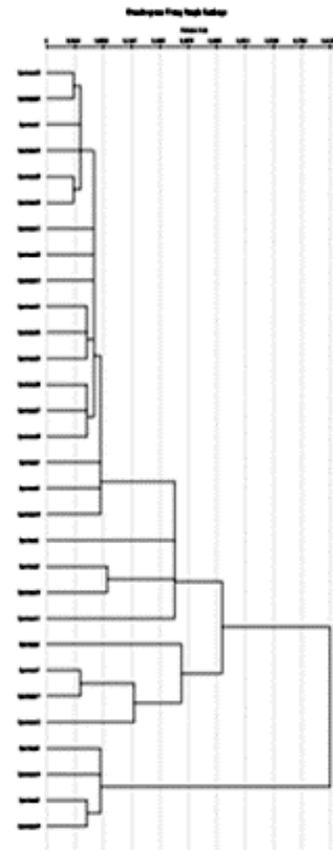


Fig. 8. Dendrogram using Damerau-Levenshtein Distance using CFDM

VI. CONCLUSION

The representation of the spectra by a 0/1 sequence is a good representation for a spectrometer signal. While coding the signal in a 0/1 sequence it also smoothing the signal by a step-wise linear function. To keep the approximation error between the original signal and the coded signal small an adaptive delta modulator has to be selected. In the experiment above we used the CFDM delta modulation method instead of the linear delta modulator. A better method than this might be the continuously variable slope delta modulator. To construct such a modulator for this kind of signals is left for further work.

Three different similarity measures have been used: Hamming distance, Levenshtein distance, and the Damerau-Levenshtein distance. While the Hamming distance is very fast and simple to calculated, the latter two distances seem to represent the similar groups of spectra's very well. However, these two distance measures are more computationally expensive as the Hamming distance. The advantage of the Levenshtein and the Damerau-Levenshtein distance is that this distance can compare strings with different number of bits and since these measures can delete and substitute bits small differences in the sequence caused by noise or the behavior of the delta modulation can be eliminated. Finally, we tested the methods with our prototype-based classifier. We obtained good

classification results for the Damerou-Levenshtein distance and the Levenshtein distance. The worst result we obtained for the Hamming distance.

In general we can say that the proposed novel method is a good method to represent spectrometer signals and that the similarity-based classification works very well. The proposed method allows us to extend our database of spectrometer signals very easily in the timely sequence the spectrometer signals occur and at the same time immediately to use the new acquired spectra for classification in daily work without going into a heavy update of the system parameters and functions.

We have tested it on data from RAMAN spectroscopy. However the method is not only applicable to RAMAN spectra. The method can be used for other spectra as well.

REFERENCES

- [1] Janzen Chr, Delbrück H., Perner P., MARAS – Marker Free RAMAN Screening for Molecular Investigation of Biological Interactions, Project Report 2006
- [2] Altose M. D., Zheng Y., Dong J., Palfey B. A., Carey P. R. „Comparing protein–ligand interactions in solution and single crystals by Raman spectroscopy“, Proceedings of the national Academy of Science, Vol. 98, 6, 3006 – 3011 (2001).
- [3] Rammal A., Perrin E., Chabbert B., Bertrand I., Mihai G., Vrabie V., Optimal Preprocessing of Mid InfraRed spectra. Application to classification of lignocellulosic biomass: maize roots and miscanthus internodes, In: P. Perner (Eds.) *Advances in Mass Data Analysis of Images and Signals in Medicine, Biotechnology, Chemistry and Food Industry*, ibai-publishing 2013, p. 66-76, ISBN 978-3-942952-21-7.
- [4] Desiere et al., Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry, *Genome Biology* 2004, 6:R9 doi:10.1186/gb-2004-6-1-r9.
- [5] Domon E. and Aebersold R., *Mass Spectrometry and Protein Analysis*, Science 312, 212-217 (2006)
- [6] Ketterlinus R., Sen-Yung Hsieh, Shih-Hua Teng, Lee H., and Pusch W., Fishing for biomarkers: analyzing mass spectrometry data with the new ClinProTools™ software, *BioTechniques* 38 (2005):S37-S40.
- [7] Perez-Pueyo R., Soneira M.J. and Ruiz-Moreno S., A fuzzy logic system for band detection in Raman spectroscopy, *Journal of Raman Spectroscopy*, Special Issue: Raman spectroscopy in Art and Archaeology, Volume 35, Issue 8-9, pages 808–812, August - September 2004.
- [8] Sadezky A., Muckenhuber H., Grothe H., Niessner R., Pöschl U., Raman microspectroscopy of soot and related carbonaceous materials: Spectral analysis and structural information, *Carbon* 43 (2005) 1731–1742.
- [9] Sobron P., Sobron F., Sanz A., and Rull F., Raman Signal Processing Software for Automated Identification of Mineral Phases and Biosignatures on Mars, *Applied Spectroscopy*, Vol. 62, Issue 4, pp. 364-370 (2008).
- [10] Kapp E. A. et al., An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: Sensitivity and specificity analysis, *Proteomics* 2005, 5, 3475–3490.
- [11] Bleghith A., Collet Ch., Armspach J.-P., A Unified Framework for Peak Detection and Alignment: Application to HR-MAS 2D NMR Spectroscopy, In: P. Perner (Eds.) *Advances in Mass Data Analysis of Images and Signals in Medicine, Biotechnology, Chemistry and Food Industry*, ibai-publishing 2011, p. 106-118, ISBN 978-3-942952-02-6.
- [12] Savitzky, A., Golay, M.J.E.: Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry* 36(8), 1627-1639 (1964)
- [13] Zhao, J., Carrabba, M.M., Allen, F.S.: Automated Fluorescence Rejection Using Shifted Excitation Raman Difference Spectroscopy. *Applied Spectroscopy* 56(7), 834-845 (2002)
- [14] Un, C.K., Lee, H.S.: A Study of Comparative Performance of Adaptive Delta Modulation Systems, *IEEE Trans. on Communications* 28 (1), 96-101 (1980)
- [15] Jayant, N.S.: Adaptive Delta Modulation with one-bit Memory, *The Bell System Technical Journal* 49(3), 76-80 (1970)
- [16] Jayant, N.S.: Adaptive Delta Modulation with one-bit Memory, *The Bell System Technical Journal* 49(3), 76-80 (1970)
- [17] Tazaki, S., Osawa, H., Shigematsy, Y.: A Useful Analytical Method for Discrete Adaptive Delta Modulation. *IEEE Trans. on Communications* 25 (2), 195-199 (1977)
- [18] Perner, P. *Datenreduktionsverfahren für technologische Industrierobotersteuerungen mit direkter Teach-in-Programmierung*. 2nd, unrevised edition, ISBN: 978-3-940501-16-5, ibai-publishing, Leipzig (2010)
- [19] Hamming, R. W.: Error detecting and error correcting codes. *Bell System Technical Journal* 29 (2), 147–160 (1950)
- [20] Levenshtein, V. I.: Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8), 707–710 (1966)
- [21] Damerou, F.: A technique for computer detection and correction of spelling errors. *Communications of the ACM* 7(3), 171–176 (1964)
- [22] Perner, P., *Prototype-Based Classification*, *Applied Intelligence* 28(3): 238-246 (2008)

Author Petra Perner (IAPR Fellow) is the director of the Institute of Computer Vision and Applied Computer Sciences IBaI. She received her Diploma degree in electrical engineering and her PhD degree in computer science for the work on “Data Reduction Methods for Industrial Robots with Direct Teach-in-Programming”. Her habilitation thesis was about “A Methodology for the Development of Knowledge-Based Image-Interpretation Systems”.

She has been the principal investigator of various national and international research projects. She received several research awards for her research work and has been awarded with 3 business awards for her work on bringing intelligent image interpretation methods and data mining methods into business.

Her research interest is image analysis and interpretation, data mining, machine learning, image mining and case-based reasoning. Recently, she is working on various medical, chemical and biomedical applications, information management applications, technical diagnosis and e-commerce applications. Most of the developments are protected by legal patent rights and can be licensed to qualified industrial companies. She has published numerous scientific publications and patents and is often requested as a plenary speaker in distinct research fields as well as across disciplines. Her vision is to build intelligent flexible and robust data-interpreting systems that are inspired by the human case-based reasoning process.

Inter-firm Transactional Relationship in Yokokai Using IDE Spatial Model: An Empirical Investigation

Takao Ito, Rajiv Mehta, Tsutomu Ito, Makoto Sakamoto, Satoshi Ikeda, Seigo Matsuno, Yasuo Uchida

Abstract—This paper discusses recent fundamental changes in the Japanese alliance networks known as keiretsu, and reports the findings of an empirical investigation on the relationship between these changes and corporate performance. More specially, the performance of Japanese auto manufacturers, such as Toyota, Mazda and Nissan, among others, has significantly improved due to sophisticated production system technologies, highly productive workers, and recurring transaction relationship with other partners in their network organization. One possible determinant of their success could be due to their unique organization forms –the keiretsu– which provides a strong platform to forge their strategic alliance relationship with their parts suppliers as well as collaboration in research and development with other automobile makers.

After the Lehman Brothers bankruptcy in 2008, the strong ties between automobile makers and their supplier partners experienced significant changes, which are known as “external influence”. Consequently, what is the status quo of automotive keiretsus? Does a transactional relationship in keiretsu still culminate in improving corporate performance? To answer these questions, this paper reports the results of a study that collected data on transaction to shed light on the relationship between inter-firm transactional relationship and corporate performance. The findings of this empirical investigation reveal that: (1) Keiretsu is a flexible, highly adaptive organizational form; its scale changes in response to economic situations; (2) Transactional relationship is still a significant determinant of

increasing profits for keiretsu partners even in the aftermath of the Lehman crash in 2008.

Keywords—Corporate performance, Keiretsu, The IDE spatial model, Transactional network, Yokokai.

I. INTRODUCTION

JAPANESE automobile manufacturers still show signs of performing at a significantly higher level than their global counterparts. This could possibly be due to the sophisticated technologies deployed for their production systems, highly productive employees, and continuous transaction relationships with other member-partners in the keiretsu network. Possibly, one explanatory factor contributing to their success could be their unique organization forms –the keiretsu– which provides a strong platform to forge strategic alliances with their parts suppliers, as well as collaboration in research and development with other automobile makers. In the aftermath of the 1990s economic bubble, the strong interrelationships between car producers and their automotive parts suppliers in the keiretsu network underwent a significant transition referred to as “keiretsu loosening”. Moreover, the 2008 financial crisis had a strong impact on keiretsu.

Thus, it is necessary to determine the current status quo of keiretsus. More specifically, does a transaction relationship in keiretsu, still conduce to higher levels of corporate performance? To find answers to this and related questions, this paper reviews the extant literature on keiretsu to propose a new approach known as the IDE spatial model that sheds light on the interrelationship between transaction and corporate performance.

This manuscript is organized as follows: Section 2 reviews the relevant literature associated with keiretsu networks. Section 3 describes the data collection process and the new network model. Based upon the findings, the managerial implications are discussed in section 4. In section 5, the study limitations are identified and section 6 proffers avenues of future research.

This work was supported in part by the JSPS KAKENHI Grant Number 24510217.

Takao Ito is with the Graduate School of Engineering, Hiroshima University, 1-4-1, Kagamiyama, Higashihiroshima, 739-8527 Japan (corresponding author to provide phone: 082-424-5594; fax: 082-424-5594; e-mail: jtotakao@hiroshima-u.ac.jp).

Rajiv Mehta is with the School of Management, New Jersey Institute of Technology, University Heights, Newark, NJ 07102-1982 USA (e-mail: mehta@njit.edu).

Tsutomu Ito is with the Hamura Factory, Hino Motors, Ltd., 3-1-1, Midorigaoka, Hamura, 205-8660 Japan (e-mail: fw.eldorado.500cuin@gmail.com).

Makoto Sakamoto is with the Faculty of Engineering, University of Miyazaki, 1-1 Kibanadai-nishi, Gakuen, 889-2192 Japan (e-mail: sakamoto@cs.miyazaki-u.ac.jp).

Satoshi Ikeda is with the Faculty of Engineering, University of Miyazaki, 1-1 Kibanadai-nishi, Gakuen, 889-2192 Japan (e-mail: bisu@cs.miyazaki-u.ac.jp).

Seigo Matsuno is with the Department of Business Administration, Ube National College of Technology, 2-14-1 Tokiwadai, Ube, 755-8555 Japan (e-mail: matsuno@ube-k.ac.jp).

Yasuo Uchida is with the Department of Business Administration, Ube National College of Technology, 2-14-1 Tokiwadai, Ube, 755-8555 Japan (e-mail: uchida@ube-k.ac.jp).

II. BACKGROUND

Keiretsu have become successful model of inter-firm collaboration. Keiretsu involves any type of relationship between one or more companies attempting to pursue individual and joint corporate and market related goals that each firm alone could not easily attain. It is based on the notion that it is difficult for a firm to “go it alone” and excel in performing all business functions. Keiretsu are formidable organizational forms owing to their global reach and lower investment costs. Cooperation among partner firm forms the “heart” of keiretsu alliances.

Consequently, it is crucial to shed light on the essential principles of rational inter-firm alliances not only based on theoretical research, but also grounded in quantitative methods.

Although many quantitative methods have been developed, an effective mathematical tool is graph theory. As a network organization, the interrelationships among member partners in keiretsu should be calculated from the viewpoint of factors, such as centrality, density, effective size, and influence, among others. To find new approaches, many studies have been published on keiretsu. Fukuoka et al. calculated correlation ratio between transaction and cross shareholdings data and found a positive relationship between the correlation ratio and corporate performance after comparing Nissan and Toyota [1]. Ito et al. discovered a relationship between network indices such as centrality and capacity and corporate performance in Mazda’s Yokokai [2, 3]. Moreover, Tagawa et al. uncovered the relationship between organizational structure and corporate performance such as sales and profits, in Mazda’s Yokokai [4]. And more recently, Ito et al. uncovered the relationship between organizational structure and corporate performance in Mazda’s Yokokai using IDE spatial model [5]. All these studies support the theory that mutual assistance and access to stable financing are equally important determinants that leverage the performance of manufacturing firms.

After 2008 economic downturn, the strong ties between automobile manufacturers and their suppliers in keiretsu underwent significant changes, which are known as “external influence”. McGulre and Dow indicated that the four characteristics that underscore the evolution of keiretsu ties are (1) diminished bank debt; (2) reduced cross-holdings; (3) reduced buyer-supplier ties (vertical keiretsu); and (4) diminished inter-firm exchanges of board and personnel [6]. Thus, review of the literature readily reveals that many scholars have found results consistent to those obtained by McGulre and Dow.

Because of the importance of keiretsu, the following questions should be investigated: What is the status quo of present-day keiretsu? Is transactional relationship in keiretsu, still a statistically significant predictor of corporate performance? To our best knowledge, no research provides answers to these questions.

III. DATA COLLECTION AND VARIABLES SELECTION

To shed light on these issues and to examine the network relationship between transaction and corporate performance, data were collected from Mazda’ Yokokai keiretsu. Mazda’s

keiretsu is composed of three sub- organizations: Nishi-Nihon Yokokai, Kanto Yokokai and Kansai Yokokai.

A. Data Collection

Data were collected for 2006, 2007, 2008, 2010, 2011 and 2012 fiscal years to establish the status quo of keiretsu and ascertain changes in its structure in the aftermath of the Lehman crash.

The relevant information about the Yokokai is shown in Table 1.

Table 1 Yokokai Network Data with Singletons

	Suppliers	Car makers	Total Number
2006	190	11	179
2007	189	11	178
2008	188	11	177
2010	172	11	161
2011	183	11	172
2012	183	11	172

Table 1 also includes data on singletons, which refers to a partner firm in the keiretsu that has no relationship with other member firms. Singletons were removed from the data-set because singletons have no impact on the calculation of network indexes. The revised data is shown in Figure 1.

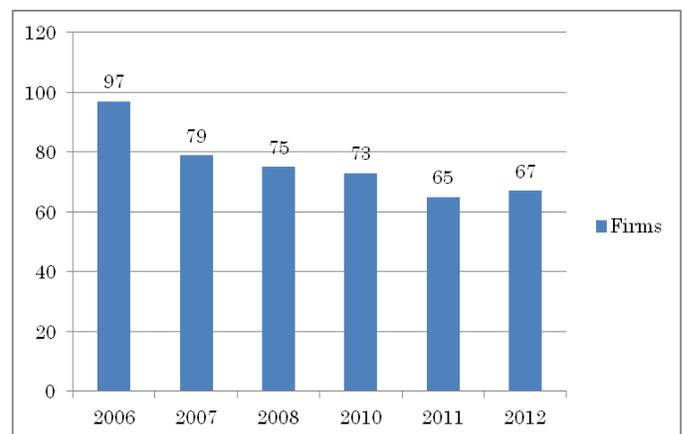


Figure 1 Yokokai network data without singletons

Transactional relationship in Yokokai, which was also collected, refers to the percent of the parts one company purchases from network partners.

Table 2 shows the transactional data in Yokokai. For instance, Hikari Seiko (No. 157) sells 12.3 percent parts to Denso (No. 145). As such, the cell between Hikari Seiko and Denso is 12.3 percent. In other words, Denso purchases parts from Hikari Seiko and it occupy 12.3 percent of Hikari Seiko’s total sales.

The inter-firm transactional relationship in 2006 is illustrated in Figure 2.

Table 2 Yokokai Network Matrix Data in 1985

		115	116	117	119	120	121	122	130	133	136	139	145

113	...	0	0	0	0	0	0	0	0	0	0	0	0
114	...	0	0	0	0	0	0	0	0	0	0	0	3.5
115	...	0	0	0	0	0	0	0	0	0	0	0	0
116	...	0	0	0	0	0	0	0	0	0	0	0	0
117	...	0	0	0	0	0	0	0	0	0	0	0	0
119	...	0	0	0	0	0	0	0	0	0	0	0	0
120	...	0	0	0	0	0	0	0	0	0	0	0	0
121	...	0	0	0	0	0	0	0	0	0	0	0	0
122	...	0	4.3	0	0	0	0	0	0	0	0	0	0
130	...	0	0	0	0	0	0	0	0	0	0	0	0
133	...	0	0	0	0	0	0	0	0	0	0	0	0
136	...	0	0	0	0	0	0	0	0	0	0	0	0
139	...	0	0	0	0	0	0	0	0	0	0	0	0
145	...	0	0	0	0	0	0	0	0	0	0	0	0
147	...	0	0	0	0	0	0	0	0	0	0	0	0
150	...	0	0	0	0	0	0	0	0	0	0	0	0
154	...	0	0	0	0	0	0	0	0	0	0	0	0
155	...	0	0	0	0	0	0	0	0	0	0	0	37.7
157	...	3.1	5.5	0	0	0	0	0	0	8.7	0	0	12.3
159	...	0	0	0	0	0	0	0	0	0	0	0	0
162	...	0	0	0	0	0	0	0	0	0	0	0	0

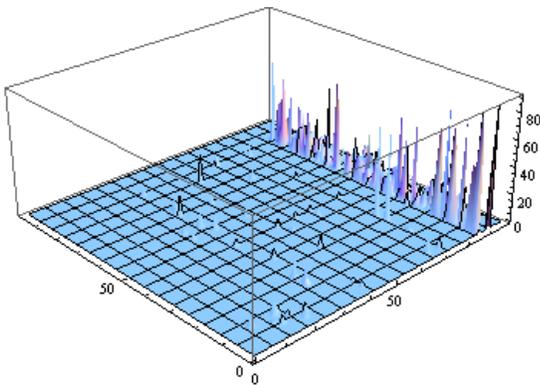


Figure 2 Transactional networks in Yokokai

B Model and Measurement

As previously noted, many structural indices of network analysis have been developed, but this study selected degree, influence and effective size of the firms included in Yokokai to analyze the relationship between those indices and corporate performance as these interrelationships have not been previously investigated.

Degree is an index of a firm’s potential communication activity. Degree is calculated as below.

$$C_D(p_k) = \sum_{i=1}^n a(p_i, p_k) \quad (1)$$

$i = 1, 2, \dots, n; \quad k = 1, 2, \dots, n$

where

$a(p_i, p_k) = 1$; if and only if p_i and p_k are connected by a line
 $= 0$; otherwise

Influence reflects the power to influence or have an impact on

other member firms directly and indirectly in a network. Suppose that A is the matrix of the direct network, and A^n means the indirect influence from one firm to another firm by n steps. Then influence is calculated as follows:

$$T = A + R = A + A^2 + A^3 + \dots + A^n$$

$$= A(I - A)^{-1} \quad (2)$$

where

- T: Total influence;
- A: direct influence;
- R: indirect influence;
- I: Identity matrix.

Effective size of the network refer to the number of alters that ego has, minus the average number of ties that each alter has to other alters. As suggested by Borgatti [7], it can be calculated as follows:

$$ES(p_k) = (n - 1) - \frac{\sum x_{pk}}{n - 1} \quad (3)$$

where

- n: number of ego network;
- x_{pk} : node k’s connection lines in k’s ego network.

A three dimension is composed of a set of network indexes: degree, influence and effective size. The position of each firms located in the three dimension will be considered as one factors of its performance. Accordingly, the following hypothesis is posited:

H1: Distance between each firm and Mazda will be negatively associated with sales.

IV. RESULTS AND DISCUSSIONS

Euclidean distance based upon degree, influence and effective size are calculated in this paper. In order to calculate the relationship between Euclidean distance and its corporate performance, the data of corporate performance such as sales are collected. The results of the regression analysis are reported in Table 3.

Table 3 Regression results of the Euclidean distances and corporate performance

	Correlation Coefficient	Probability	Degree of Freedom	DW Ratio
2006	-0.2442	0.0301	1, 77	1.5969
2007	-0.242	0.0364	1, 73	1.5017
2008	-0.4158	0.0005	1, 65	2.0185
2010	-0.6801	0	1, 66	2.1245
2011	-0.6578	0	1, 59	1.952
2012	-0.7361	0	1, 57	2.2231

The results reveal that all of the correlation coefficients are statistically significant. Table 3 shows all correlation coefficients are negative, which means that longer distance from Mazda are associated with lower sales. Thus, there is support for H1 in a transactional network. Moreover, compared with the results in 2006 and 2007, the correlation coefficients are seemingly higher after 2008. According to McGuire and Dow's study, keiretsu, as one of the vertical organization, is oversighted by a core firm, which encourages transactions with its network partners with long-term perspective. All of the basic functions are still working even in the aftermath of economic downturn of 2008.

V. CORPORATE MANAGEMENT IMPLICATIONS

Based on the findings of this empirical study, some corporate management implications can be gleaned. Thus, to augment corporate efficiency, the following suggestions for managing the interrelationships among keiretsu members are offered. First, Euclidean distance could be considered as a new measure for improving corporate performance. Second, position in three-dimension space should be observed as it is an important factor for determining corporate behavior.

VI. LIMITATIONS OF THE STUDY

Although this study makes a contribution to the extant literature, there are some drawbacks that may temper the findings to be held tentative. First, the results of this study should be compared with other factors, such as capital relationship, work flow relationship and friend relationship for identifying the antecedents of corporate management. Second, it is suggested that additional time series data over a period of 10 or 20 years or more should be gathered to longitudinally analyze position and distance trends and changes over time. This will provide a better picture of whether these relationships are stable or vary over time, perhaps owed to economic conditions. Owed to these limitations, the findings of the study should be treated with caution.

VII. DIRECTIONS FOR FUTURE RESEARCH

The findings of this investigation should be viewed in light of the above-noted limitations that are suggestive of future research efforts. First, the interrelationships among the constructs should be verified using data gathered from a sample of keiretsu comprised in different industrial sectors that include machinery, steel, shipbuilding and electronic products. Second, data drawn over a period of six years is a starting point, but insufficient in providing a comprehensive understanding on the real behavior of the firms. Third, three dimension space is only one perspective for analyzing the behavior. Additional indexes should be identified as possible antecedents of corporate performance in network organizations. For example, future studies should investigate the linkage between degree, influence and effective size as determinants of corporate performance.

VIII. CONCLUSIONS

This paper applied IDE spatial model and calculated correlation coefficients between Euclidean distances and corporate performance. The relationship between Euclidean distances and sales is supported. But sales are not dependent with only Euclidean distance. Additional investigations, such as the association between corporate performance and degree, influence and effective size should be tested.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper.

REFERENCES

- [1] Fukuoka S., Ito T., Passerini K., and Sakamoto M. (2006) An Analysis between Transaction and Cross Shareholdings in the Keiretsu of Nissan, Proceedings of the 6th International Business Information Management Association, International Conference, *Managing Information in Digital Economy*, pp. 163-169, June 19-21, 2006, Bonn, Germany.
- [2] Ito T., Takida R., Matsuno S., Mehta R., Ishida Y., and Sakamoto M. (2011) An analysis of a firm's capacity in Mazda's Keiretsu, *Artificial Life and Robotics*, Volume 16, Number 2, pp. 248-252, Springer Japan, 2011.
- [3] Ito T., Niki E., Takida R., Mehta R., Passerini K., and Sakamoto M. (2011) Transactions and cross shareholdings in Mazda's Keiretsu: a centrality analysis, *Artificial Life and Robotics*, Volume 16, Number 3, pp.297-300, Springer Japan, 2011.
- [4] Tagawa S., Ito T., Mehta R., Passerini K., Voges K., and Sakamoto M. (2012) Organizational structure of Mazda's Keiretsu: A graph theoretic analysis, *Artificial Life and Robotics*, Volume 16, Number 4, pp.455-459, Springer Japan, 2012.
- [5] Takao Ito, Makoto Sakamoto, Rajiv Mehta, Tsutomu Ito, Satoshi Ikeda, An Empirical Examination of Inter-firm Capital Relationships in Mazda's Yokokai using the IDE Spatial Model, Proceedings of The 2014 International Conference on ARTIFICIAL LIFE AND ROBOTICS, pp. 288-291, Jan. 11-13, 2014, Oita Japan.
- [6] McGulre J, Dow S., (2009) Japanese keiretsu: Past, present, future, http://download.springer.com/static/pdf/983/art%253A10.1007%252Fs10490-008-9104-5.pdf?auth66=1399794791_ce53db11bd93d8098b0304e16daae06f&ext=.pdf retrieved May 9, 2014
- [7] Borgatti S. P., (1997) Structural Holes: Unpacking Burt's Redundancy Measures, <http://www.analytictech.com/>

[connections/v20\(1\)/holes.htm](http://connections.v20(1)/holes.htm), retrieved November 25, 2013

Takao Ito (Doctor of Economics, Kyoto University; Ph.D. of Engineering, Miyazaki University) is Professor of Management of Technology (MOT) in Graduate School of Engineering at Hiroshima University. Previously, he served on the faculty of Ube National College of Technology (UNCT, Japan) where he teaches courses in Information System Theories and Corporate Organization Theories. He is serving concurrently as Professor of Harbin Institute of Technology (HIT) at Weihai, China. He has published numerous papers in refereed journals and proceedings, particularly in the area of corporate management, IT strategy, and computer science. He has published more than 8 academic books including a book on Network Organizations and Information (Japanese Edition). His current research interests include automaton theory, the MOT and quantitative analysis of supply chain relationships using graph theory, theoretical analysis of the relationship between IT and corporate strategy, and engineering analysis of organizational structures using complex system theory. He was one of the winners of the Best Paper Award in the International Symposium on Artificial Life and Robotics (AROB) in January 2006. Dr. Ito is one of the associate editors of the International Journal of Advances in Information Sciences and Service Science, associate editor of International Journal Robotics, Networking and Artificial Life, and the member of editorial review board of International Journal of Data Mining, Modeling and Management and Information and Communication Technologies for the Advanced Enterprise. And he also is the Assistant General Chair of the International Conference on Artificial Life and Robotics. Dr. Ito earned his Doctor degree of Economics from Kyoto University and Ph. D. of Engineering from Miyazaki University, Japan. Dr. Ito is a visiting research professor of School of Management, New Jersey Institute of Technology (NJIT) from March 29 2008 to March 28, 2009.

Rajiv Mehta (Ph.D., Drexel University) is Professor of Marketing at the School of Management, New Jersey Institute of Technology. Previously, he served on the faculty of Loyola University New Orleans and, prior to entering academia, worked in sales and marketing for an international manufacturer of steel wire ropes and cables.

He has co-authored two university-level textbooks: Sales Management: Building Customer Relationships and Partnerships (2009), Houghton-Mifflin (now CENGAGE Learning) and Personal Selling: Building Customer Relationships and Partnerships, (2014), 3rd Edition, Kendall Hunt.

Additionally, his research has been widely published in major academic journals and presented at national and international academic conferences. Focusing on the areas of selling and sales management, domestic and global marketing channels, Japanese Keiretsus, and the cross-border management of international strategic distribution alliances, over 30 of Dr. Mehta's research manuscripts have appeared in outlets, such as Journal of Business Research, Industrial Marketing Management, Journal of Personal Selling and Sales Management, Business Horizons, European Journal of Marketing, International Marketing Review, Journal of Business to Business Marketing, Journal of Business and Industrial Marketing, Journal of Marketing Channels, Journal of Global Marketing, International Journal of Physical Distribution and Logistics Management, Journal of Managerial Issues, Journal of Services Marketing, Management Bibliographies and Reviews, International Journal of Quality and Reliability Management, International Journal of Technology, Knowledge & Society, Total Quality Management, Artificial Life and Robotics, and Journal of Shopping Center Research.

In addition to winning the 2001 award for an outstanding journal article in International Marketing Review, Dr. Mehta's teaching contributions were recognized by NJIT alumni when he has awarded the prestigious university-wide Robert W. Van Houten Award for Teaching Excellence in 2005—a recognition that is given to only one Institute faculty member annually. He also received the University Award for Excellence in the Category of Teaching in Upper Division Undergraduate Instruction in 2005 and was bestowed the Master Teacher Award in 2006 for having demonstrated the highest level of teaching excellence.

Tsutomu Ito (Master of Engineering, Yamaguchi University) is a researcher of Production Management at Hamura Factory, Hino Motors, Ltd., Japan

Makoto Sakamoto (Ph.D., Yamaguchi University) is Associate Professor of image processing and automaton language studies in Faculty of engineering at

University of Miyazaki. Previously, he served on the faculty of Oshima National College of Maritime Technology (ONCTMT, Japan). He has published numerous papers in refereed journals and proceedings, particularly in the area of Turing machine, automaton theory and image processing.

Satoshi Ikeda (Ph.D., Hiroshima University) is Associate Professor of mathematics and graph theory in Faculty of engineering at University of Miyazaki. He has published numerous papers in refereed journals and proceedings, particularly in the area of application of graph theory.

Seigo Matsuno (Ph.D., Kyushu University) is Professor of e-Commerce and Supply chain management in Department of Business Administration at Ube National College of Technology, Japan.

Yasuo Uchida (Ph.D., University of Miyazaki) is Professor of database and Computer Network in Department of Business Administration at Ube National College of Technology, Japan.

Informetric models for citation frequency data: an empirical investigation

Lucio Bertoli-Barsotti, Tommaso Lando

Abstract— This paper tries to answer to the question “Which is the best model for representing the citation frequency curve?”. We consider 131 practical cases of physicists who were applicants for a full professorship in the specific area of Condensed Matter Physics, and we estimate a set of four possible different types of distribution for size-frequency data. The most remarkable result is that the well-known Lotkaian model is not the best fitting among all we have considered. From our data we conclude that the geometric distribution can provide a valid alternative to more traditional models.

Keywords—Citation analysis, geometric distribution, Lotka’s law, Kullback-Leibler estimation, size-frequency data.

I. INTRODUCTION

CITATION counting techniques are used for evaluating scientific activities. Number of citations received by article, or individual, are frequently used as a measure of “quality” in science. In citation analysis, there are two possible ways to interpret citation distributions and, accordingly, there are also two possible ways to fit citation distributions with probabilistic models. We may consider, for example, the number of citations (for a given author) of each paper as observations that constitute a sample. In this case, the frequency of an observation c represents the number of articles with c citations, and we speak of size-frequency distribution.

Following a second approach, we may observe the rank of a paper and the frequency of its citations. Here, the frequency of an observation, say r , is the number of citations of the paper ranked at the r -th position and we speak of rank-frequency distribution. In other words, we could interpret the citations as empirical observations (first approach) or frequencies (second approach). In this paper we consider the problem on the point of view of the first approach, the size-frequency analysis.

This work has supported by the Italian funds ex MURST 60% 2014, 2015 and the project Opportunity for young researchers, reg. no. CZ.1.07/2.3.00/30.0016 (to T.L.), supported by Operational Programme Education for Competitiveness and co-financed by the European Social Fund and the state budget of the Czech Republic.

L.B.B. Author is with University of Bergamo, via dei Caniana, 2, Bergamo, Italy; e-mail: lucio.bertoli-barsotti@unibg.it.

T.L. Author is with University of Bergamo, via dei Caniana, 2, Bergamo, Italy; and VŠB -TU Ostrava, Sokolská trida 33, Ostrava, Czech republic; e-mail: tommaso.lando@unibg.it.

II. MODEL DISTRIBUTION

We assume that, for a fixed author, the number of papers with c citations received is given by a formula $n(c, \alpha)$, $c = 1, 2, \dots$, where alpha is a (possibly vectorial) parameter. The unknown parameter alpha can be determined via a fitting procedure. By definition, the function n must satisfy the constraint $\sum_{c=1}^{\infty} n(c, \alpha) = N$, where N represents the author’s total number of publications with at least one citation. We might also give a probabilistic interpretation of n by noting that this function must be proportional to a probability mass function. Indeed

$$\sum_{c=1}^{\infty} f(c, \alpha) = 1$$

where $f(c, \alpha) = N^{-1} \cdot n(c, \alpha)$. In this form, f represents the standard form of a probability mass function (p.m.f.), with support $\{1, 2, 3, \dots\}$.

In this paper we shall consider the following models of p.m.f.

A. Zeta distribution

The p.m.f. is $f(c, \alpha) = \frac{c^{-\alpha}}{\zeta(\alpha)}$, $\alpha > 1$, where $\zeta(\alpha) = \sum_{c=1}^{\infty} c^{-\alpha}$ denotes the Riemann zeta function ([1], p.527). This distribution is also referred to as *discrete Pareto distribution* and, depending on the context, it is also called *Zipf distribution* (see e.g. [2]). In the bibliometric literature, this formula is also known to as *power-law distribution*. When α is set equal to 2, $\zeta(2) = \frac{\pi^2}{6}$, we obtain the Lotka distribution ([1], p.527, [3], [4]). Somewhat strangely, in the literature the term Lotka’s law is frequently used, in a more general sense than that used by Lotka [3], to refer to the above formula $c^{-\alpha} \zeta(\alpha)^{-1}$ as a size-frequency density function expressing the number/proportion of articles with exactly c citations ([5]-[6]-[7]-[8]-[9]-[10]). In applications, the “Lotka’s law” is probably the simplest and the most used model for the analysis of citation frequency data.

B. Geometric distribution.

We consider the p.m.f. of the geometric distribution in the following form: $f(c, \alpha) = \alpha(1 - \alpha)^c$, $0 < \alpha < 1$.

C. Logarithmic distribution

The p.m.f. is

$$f(c, \alpha) = -[\log(1 - \alpha)]^{-1} c^{-1} \alpha^c, \quad 0 < \alpha < 1$$

(see [1], p.302).

D. Pareto distribution

Let $f(c, \alpha) = \int_{c-0.5}^{c+0.5} g(y, \alpha) dy$, where $g(y, \alpha) = (\alpha - 1)0.5^{\alpha-1}y^{-\alpha}$, $y \geq 0.5$, $\alpha > 1$

(see [1] p.574). The Pareto distribution is a continuous variant of the above zeta distribution. For this reason, g is also known as continuous Lotka function.

Denote by c_i the number of citations gained by the i -th paper, $i = 1, \dots, N$. Let $C = \sum_{i=1}^N c_i$ the total number of citations (of an author). And let n_j the number of papers with exactly j citations. Let $F_N^*(t)$ be the empirical distribution function, defined as $F_N^*(t) = \sum_{j \leq t} \frac{n_j}{N}$, for every $t \in \mathbb{R}$.

Since in our context it is hard to assume the independence between observations, we rely on the estimation approach given by the minimum distance (MD) method (see [11], p.65-67), by adopting the Kullback-Leibler distance. Remember that the minimum distance estimate of the parameter α , with respect to the Kullback-Leibler distance, is the value of α for which

$$-\sum_{j=1}^{c_{max}} n_j \log f(j, \alpha) = \min_{\alpha} \left(-\sum_{j=1}^{c_{max}} n_j \log f(j, \alpha) \right).$$

(Note that, under the independence assumption, the minimum Kullback-Leibler estimator coincides with the maximum likelihood estimator). Otherwise said, we search for the point $\hat{\alpha}$ for which the function $\sum_{j=1}^{c_{max}} n_j \log f(j, \alpha)$ attains its absolute maximum value, given the set of observed pairs (j, n_j) .

III. DATA

We considered 131 datasets, also analyzed elsewhere for a comparative study concerning 13 different bibliometric indices [12]. The publication and citation data considered were obtained from Scopus in January 2014 and refer to a sample of 131 physicists who were applicants in the 2012 ASN (Abilitazione Scientifica Nazionale; the Italian National Scientific Qualification for the recruitment of academic staff in Italy) for a full professorship in the specific area of Condensed Matter Physics. Table 1 summarizes some of the most important citation metrics concerning the considered datasets.

	C	N	C/N	MC	h
Mean	2206	85	25	359	21.6
Min	18	5	3	5	2
Max	13916	328	102	3068	53
Q1	1156	57	16	104	18
Q2	1786	77	21	177	22
Q3	2740	107	31	330	27
SD	1935	51	16	543	8.7

Table 1. Characteristics of the 131 datasets analyzed in the present study. $MC=c_{max}$; C = total number of citations;

N =total number of cited papers; C/N =average number of citations per paper; h =h-index

IV. EXPERIMENTAL RESULTS

As said above, our goal was to obtain an estimate of the number of papers with c citations, for every $c \in \{1, 2, \dots, c_{max}\}$, using a theoretical model distribution. Based on the empirical observations, we estimate each one of the considered models and we compute a Kolmogorov-Smirnov (K-S) distance as a discrepancy measure (between observed and fitted data) for goodness-of-fit purposes. Indeed, the K-S distance can be used (here only for descriptive and comparative purposes) to compare the different distributional assumptions and to identify the model which better complies with observed citation frequency data, among those considered. Remember that the K-S statistic, D_N , is defined as the maximum (vertical) distance between the empirical and the estimated theoretical distribution function, say $F_{\hat{\alpha}}(t)$, that is, in symbols, $D_N = \sup_t |F_{\hat{\alpha}}(t) - F_N^*(t)|$. For taking into account the sample dimension, we also compute the statistic $D_N^* = D\sqrt{N}$.

We observe that, frequently, the sets of citations contain outliers, that is, some author may have one (or few) article(s) which has been cited an outstandingly high number of times, compared to all his (or her) other papers. This may be due to several reasons (“age” of the paper, number of co-authors, etc.). We note that the presence of outliers has a negative influence on the geometric model. Table 2 reports the correlation coefficients between the K-S distance and the most important bibliometric indicators, namely: the h -index; the total number of citations C ; the number of papers with at least one citation, N ; the average number of citations per paper C/N and the number of citations of the most cited paper MC . From data reported in Table 2, we can observe a quite strong dependence between the K-S distance between empirical and geometric distribution and the maximum number of citations (MC), and thereby the average number of papers (C/N). This result suggest that, for the geometric model, the goodness-of-fit could be enhanced by excluding from the sample the highest observed values, which can actually be considered as outliers. In particular, we find a satisfactory improvement of the goodness of fit by trimming the 5% of the highest observations in each sample dataset (note that, however, the K-S distance is evaluated over the whole sample). As the MD estimator for the geometric distribution is the reciprocal of the sample mean, then we estimate α by the reciprocal of a trimmed (or truncated) sample mean, which is less sensitive to outliers and is especially suitable for dealing with heavy tailed distributions. However, we observed that this technique has a negative effect on the other models, as also confirmed by table 2 (interestingly, the logarithmic model yields even a better fit for higher values of MC). For these reasons we report the values produced by the trimming method, just for the geometric distribution.

	zeta	geo	log	par	geo(t)
h	0.17	0.10	0.02	-0.06	-0.36

<i>C</i>	0.14	0.31	-0.07	-0.08	-0.29
<i>n</i>	0	-0.07	-0.13	-0.22	-0.46
<i>C/n</i>	0.28	0.50	0.08	0.06	0.05
<i>MC</i>	0.16	0.60	-0.11	-0.05	-0.03

Table 2. Correlations between the K-S distance and the some of the most important citation metrics; *geo(t)* refers to the geometric model, with the estimation based on the trimmed sample.

From Table 2 we observe that the geometric distribution, estimated with the trimming method, is substantially insensitive to *MC* and *C/N*, but is positively influenced by the bibliometric indices of “productivity” (*h*, *C* and *N*). In particular, the K-S distance is reduced for larger samples (a sort of consistency), especially with the (trimmed) geometric distribution.

Table 3 reports the average value of the K-S statistic *D*, over the 131 datasets, and the number of cases when it is smaller than 0.1, 0.15 and 0.2. The geometric model shows better fit compared to the other distributions, especially when the parameter is estimated via trimming method. And finally, Table 4 summarizes the basic statistics regarding *D** for the whole sample of datasets.

	<i>zeta</i>	<i>geo</i>	<i>log</i>	<i>par</i>	<i>geo(t)</i>
<i>M(D)</i>	0.26	0.18	0.18	0.25	0.12
#(<i>D</i> <0.10)	1 (1%)	19 (14%)	10 (7%)	0 (0%)	52 (40%)
#(<i>D</i> <0.15)	2 (1%)	49 (37%)	44 (33%)	0 (0%)	105 (80%)
#(<i>D</i> <0.20)	12 (9%)	83 (63%)	80 (61%)	56 (43%)	123 (94%)

Table 3. In the first row we report the average values of the K-S distance *D*. In rows 2,3,4 we count the number of cases when *D*<0.10,0.15,0.20 (respectively).

<i>D*</i>	<i>zeta</i>	<i>geo</i>	<i>log</i>	<i>par</i>	<i>geo(t)</i>
<i>Mean</i>	2.26	1.60	1.58	1.81	1.00
<i>Min</i>	0.29	0.26	0.17	0.53	0.32
<i>Max</i>	4.59	4.89	3.34	3.31	2.49
<i>SD</i>	0.79	0.84	0.65	0.51	0.35
<i>Q1</i>	1.76	1.02	1.09	1.50	0.77
<i>Q2</i>	2.30	1.47	1.58	1.81	0.95
<i>Q3</i>	2.74	2.06	2.02	2.10	1.18

Table 4. Summary statistics concerning $D_N^* = D\sqrt{N}$ (*SD* = Standard Deviation, $Q_i = i$ -th quartile ($i=1,2,3$))

V. CONCLUSION

In this paper we considered four different types of distributions, suitable for describing citation frequency data. All these models are used for fitting the citation frequency curves of a relatively homogeneous group of physicists. The investigated persons can be considered as “average authors”

(the average value of the *h*-index was about 21), then this case study can be considered less typical than would be expected from a standard informetric analysis (that frequently focuses on very prominent persons).

Overall, our study provides sufficient evidence of the fact that the (perhaps) most popular model for the analysis of citation frequency data -i.e. the model known as the Lotka’s power law- is not always the best candidate for the representation of the citation frequency curve. Indeed, as far as concerns the size-frequency data at hand, the geometric distribution can provide a valid alternative to more traditional models.

REFERENCES

- [1] N.L. Johnson, A.W. Kemp and S. Kotz., *Univariate discrete distributions*, Wiley Series in Probability and Statistics (third ed.) John Wiley, New York, 2005.
- [2] N. L. Johnson, S. Kotz and N. Balakrishnan, *Continuous univariate distributions*, Vol. 1, 2nd Edition. John Wiley, New York, 1994.
- [3] P. T. Nicholls, “Empirical validation of Lotka’s law,” *Information Processing and Management*, 22, pp. 417-419, 1986.
- [4] A. J. Lotka, “The frequency distribution of scientific productivity,” *Journal of the Washington Academy of Sciences*, vol. 16, no. 12, pp. 317-323, 1926.
- [5] R. C. Coile, “Lotka’s frequency distribution of scientific productivity,” *Journal of the American Society for Information Science*, vol.28, no.6, pp.366-370, 1977.
- [6] L. Egghe, *Power laws in the information production process: Lotkaian informetrics*, London: Academic Press, 2005.
- [7] L. Egghe, “Relations between the continuous and the discrete Lotka power function,” *Journal of the American Society for Information Science and Technology*, vol. 56, no. 7, pp. 664–668, 2005.
- [8] L. Egghe, R. Rousseau, “An informetric model for the Hirsch-index,” *Scientometrics*, vol. 69, no. 1, pp. 121–129, 2006.
- [9] L. Egghe, “Lotkaian informetrics and applications to social networks”, *Bulletin of the Belgian Mathematical Society-Simon Stevin*, vol. 16, no. 4, pp. 689–703, 2009.
- [10] L. Egghe, “A new short proof of Naranan’s theorem, explaining Lotka’s law and Zipf’s law,” *Journal of the American Society for Information*, vol. 61, no. 12, pp. 2581-2583, 2010.
- [11] B. Rousseau, R. Rousseau, “LOTKA: A program to fit a power law distribution to observed frequency data,” *Cybermetrics*, vol. 4, no.1.
- [12] A. A. Borovkov, *Mathematical Statistics*, Amsterdam: Gordon and Breach Science Publishers, 1998.
- [13] T. Lando, L. Bertoli-Barsotti, “A New Bibliometric Index Based on the Shape of the Citation Distribution,” *PLoS ONE*, vol. 9, no. 12: e115962, 2014.



Lucio Bertoli-Barsotti is an associate professor of Statistics at the University of Bergamo, Faculty of Economics; Management, Economics and Quantitative Methods, Via dei Caniana 2, 24127, Bergamo (Italy). Before moving to Bergamo, he was a professor at the University of Torino (1998-2002). Prior to this, he was a researcher at the Catholic University of the Sacred Heart of Milan (1987-1998). His current research interests include Item Response Theory models, Rasch Analysis, Latent variable modeling, treatment of missing data, stochastic orderings.

Bounds on the Generalized Krein Parameters of an Association Scheme

Vasco Moço Mano and Luís Almeida Vieira.

Abstract—In this paper we generalize the Krein parameters of a symmetric association scheme and obtain some bounds on these parameters and, consequently, on the classical Krein parameters of an association scheme, taking into account the properties of its eigenmatrix and dual eigenmatrix.

Keywords—Association scheme, matrix analysis, strongly regular graph.

I. INTRODUCTION

THIS paper is organized in three sections. In the first one we will present the basic definitions and properties of symmetric association schemes which are necessary for our work. All the concepts presented are described in detail, for instance, in [1]. In Section II we generalize the Krein parameters of an association scheme and establish some bounds for these generalizations. Finally, in Section III, we present some conclusions examples.

A symmetric association scheme, Ω , with d associate classes on a finite set X is a partition of $X \times X$ into sets R_0, R_1, \dots, R_d , which are relations on X satisfying the following axioms: (i) $R_0 = \{(x, x) : x \in X\}$; (ii) if $(x, y) \in R_i$, then $(y, x) \in R_i$, for all x, y in X and i in $\{0, 1, \dots, d\}$; (iii) for all i, j, l in $\{0, 1, \dots, d\}$ there is an integer p_{ij}^l such that, for all (x, y) in R_l

$$|\{z \in X : (x, z) \in R_i \text{ and } (z, y) \in R_j\}| = p_{ij}^l.$$

The numbers p_{ij}^l are called the *intersection numbers* of Ω . It is usual to observe the intersection numbers as the entries of the so called *intersection matrices* L_0, L_1, \dots, L_d , with $(L_i)_{lj} = p_{ij}^l$, where $L_0 = I_n$.

This definition is due to Bose and Shimamoto, [2], and by axiom (ii) the relations R_i are all symmetric. A more general definition of non necessarily symmetric association schemes can be seen in [4]. Along this text we will only consider symmetric association schemes.

One can describe the associate classes R_0, R_1, \dots, R_d of a symmetric association scheme, Ω , by their adjacency matrices A_0, A_1, \dots, A_d , where each A_i is a matrix of order n defined by $(A_i)_{xy} = 1$, if $(x, y) \in R_i$, and $(A_i)_{xy} = 0$, otherwise. We also have the corresponding axioms for these matrices: (a) $A_0 = I_n$; (b) $\sum_{i=0}^d A_i = J_n$; (c) $A_i = A_i^T, \forall i \in \{0, 1, \dots, d\}$; (d) $A_i A_j = \sum_{l=0}^d p_{ij}^l A_l, \forall i, j \in \{0, 1, \dots, d\}$. Regard that I_n and J_n stand for the identity matrix and the all ones matrix of order n , respectively, and A^T denotes

the transpose of A . Note that equality (b) implies that the matrices $A_i, i \in 0, 1, \dots, d$, are linearly independent. It is also well known (see [1, Lemma 1.3]) that the symmetry of the scheme asserts that $p_{ij}^l = p_{ji}^l$ and thus $A_i A_j = A_j A_i$, for all $i, j \in \{0, 1, \dots, d\}$.

We can acknowledge A_1, A_2, \dots, A_d as adjacency matrices of undirected simple graphs G_1, G_2, \dots, G_d , with common vertex set V . Each graph G_i is regular with valency n_i . The matrices A_0, A_1, \dots, A_d of a symmetric association scheme generate a commutative algebra, \mathcal{A} , with dimension $d + 1$, of symmetric matrices with constant diagonal. This algebra is called the *Bose-Mesner algebra* of the scheme because it was firstly studied by these two mathematicians in [3]. Note that \mathcal{A} is an algebra with respect to the usual matrix product as well as to the *Hadamard* (or *Schur*) *product*, defined for two matrices A, B of order n as the componentwise product: $(A \circ B)_{ij} = A_{ij} B_{ij}$. The algebra \mathcal{A} is commutative and associative relatively to this product with unit J_n .

An element E in \mathcal{A} is an *idempotent* if $E^2 = E$. Two idempotents E and F in \mathcal{A} are orthogonal if $EF = 0$. The Bose-Mesner algebra \mathcal{A} has a unique basis of minimal orthogonal idempotents $\{E_0, \dots, E_d\}$ such that $E_i E_j = \delta_{ij} E_i, \sum_{i=0}^d E_i = I_n$, where $\delta_{ij} = 1$, if $i = j$ and $\delta_{ij} = 0$, otherwise, for any i, j natural numbers. Let \mathcal{A} be an association scheme with d classes. If $A_j \in \mathcal{A}, j \in \{0, 1, \dots, d\}$ has $d + 1$ distinct eigenvalues, namely $\lambda_0, \lambda_1, \dots, \lambda_d$, the idempotents E_i can be obtained as the projectors associated to the matrix A_j through the equality:

$$E_i = \prod_{l=0, l \neq i}^d \frac{A_j - \lambda_l I_n}{\lambda_i - \lambda_l}. \quad (1)$$

Along this paper we will denote the rank of each E_i by $\mu_i, i \in \{0, 1, \dots, d\}$.

Besides the intersection numbers already introduced in the beginning of the section each association scheme contains three more families of parameters: the eigenvalues, the dual eigenvalues and the Krein parameters. In fact, there are scalars $p_i(j)$ and $q_i(j)$ such that, for all $i \in 0, 1, \dots, d$, we have

$$A_i = \sum_{j=0}^d p_i(j) E_j \text{ and} \quad (2)$$

$$E_i = \sum_{j=0}^d q_i(j) A_j, \quad (3)$$

where the numbers $p_i(j)$ and $q_i(j)$ are the *eigenvalues* and the *dual eigenvalues* of the scheme, respectively. We also define the *eigenmatrix*, $P = (P_{ij})$, and the *dual eigenmatrix*, $Q =$

Vasco Moço Mano and Luís Vieira are with Department of Civil Engineering of Faculty of Engineering of University of Porto, Portugal e-mail: vascomocoman@gmail.com and lvieira@fe.up.pt.

Manuscript received February 20, 2015; revised March 20, 2015.

(Q_{ij}) , each with dimension $(d+1) \times (d+1)$, as $P_{ij} = p_j(i)$ and $Q_{ij} = q_j(i)$, respectively. From (2) and (3) one can deduce that $PQ = I_n$. As a consequence, the dual eigenvalues are determined by the eigenvalues of \mathcal{A} .

Finally, the *Krein parameters* discovered by Scott, [6], of an association scheme with d classes are the numbers $q_{(i,j;1,1)}^l$, with $i, j, l \in \{0, 1, \dots, d\}$, such that

$$E_i \circ E_j = \sum_{l=0}^d q_{(i,j;1,1)}^l E_l. \quad (4)$$

This notation will become clear later, in Section II, with the introduction of the generalized Krein parameters of an association scheme. These parameters can be seen as dual parameters of the intersection numbers and they are determined by the eigenvalues of the scheme. Also, the Krein parameters can be considered as the entries of the matrices $L_0^*, L_1^*, \dots, L_d^*$, such that $(L_i^*)_{lj} = q_{ij}^l$, which are called the *dual intersection matrices* of the scheme.

Now we will emphasize some properties of the matrices P and Q that we will use in the proofs of some of the theorems that we will present in this paper.

$$Q(i, j)Q(i, k) = \sum_{l=0}^2 q_{(j,k;1,1)}^l Q(i, l); \quad (5)$$

$$|Q(i, j)| \leq \frac{\mu_j}{n}; \quad (6)$$

$$|P(i, j)| \leq n_j; \quad (7)$$

$$\sum_{i=0}^d n_i Q(i, j)Q(i, k) \leq \frac{\mu_j}{n} \delta(j, k). \quad (8)$$

II. GENERALIZED KREIN PARAMETERS AND SOME BOUNDS

In what follows we generalize the Krein parameters of an association scheme. Let A_0, A_1, \dots, A_d be the adjacency matrices of an association scheme with d classes, Ω , on a finite set of order n , \mathcal{A} the underlying Bose-Mesner algebra and $\mathcal{S} = \{E_0, E_1, \dots, E_d\}$ be the associated unique basis of minimal orthogonal idempotents. Let p be a natural number and denote by $\mathcal{M}_n(\mathbb{R})$ the set of square matrices of order n with real entries. Then, for $B \in \mathcal{M}_n(\mathbb{R})$, we denote by $B^{\circ p}$ the *Hadamard power* of order p of B , with $B^{\circ 1} = B$.

Now, we introduce the following compact notation for the Hadamard powers of the elements of \mathcal{S} . Let x, y, α and β be natural numbers such that $0 \leq \alpha, \beta \leq d$. Then we define $E_{\alpha}^{\circ x} = (E_{\alpha})^{\circ x}$ and $E_{\alpha, \beta}^{\circ x, y} = (E_{\alpha})^{\circ x} \circ (E_{\beta})^{\circ y}$. Note that, when $\alpha = \beta$, there is a connection between the two notations: $E_{\alpha, \alpha}^{\circ x, y} = E_{\alpha}^{\circ x+y}$.

Since the Bose-Mesner algebra \mathcal{A} , that is generated by the adjacency matrices of Ω , is closed under the Hadamard product, then there exist real numbers $q_{(\alpha, \beta; x, y)}^i$ such that

$$E_{\alpha, \beta}^{\circ x, y} = \sum_{i=0}^d q_{(\alpha, \beta; x, y)}^i E_i. \quad (9)$$

We call the parameters $q_{(\alpha, \beta; x, y)}^i$, $i \in \{0, 1, \dots, d\}$, the *generalized Krein parameters* of the association scheme Ω , since for $x = y = 1$ we obtain the ‘‘classical’’ Krein parameters

already presented in (4). With this notation, the greek letters are used as idempotent indices and the latin letters are used as exponents of Hadamard powers.

Next we present a formula to compute the generalized Krein parameters by making use of just the entries of the matrices P and Q .

Theorem 1: Let Ω be an association scheme with d classes and let $i, j, s \in \{0, 1, \dots, d\}$. Then the generalized Krein parameters of Ω , defined in (9), satisfy the equality

$$q_{(i,j;m,n)}^s = \sum_{t=0}^d (Q(t, i))^m (Q(t, j))^n P(s, t). \quad (10)$$

Proof: We have $E_{i,j}^{\circ n, m} = E_i^{\circ n} \circ E_j^{\circ m} = \sum_{t=0}^d (Q(t, i))^n A_t \circ \sum_{t=0}^d (Q(t, j))^m A_t$. It follows that $E_i^{\circ n} \circ E_j^{\circ m} = \sum_{t=0}^d (Q(t, i))^n (Q(t, j))^m A_t$. But then, from (2)-(3) one can write $E_i^{\circ n} \circ E_j^{\circ m} E_s = \sum_{t=0}^d (Q(t, i))^n (Q(t, j))^m A_t E_s$. This is $q_{(i,j;n,m)}^s E_s = \sum_{t=0}^d (Q(t, i))^n (Q(t, j))^m P(s, t) E_s$. Therefore (10) follows. ■

From Theorem 1 we obtain the following consequence.

Corollary 1: Consider an association scheme Ω with d classes and let $j, k, l \in \{0, 1, \dots, d\}$. Then, the classical Krein parameters of Ω satisfy

$$q_{(j,k;1,1)}^l = \sum_{i=0}^d Q(i, j)Q(i, k)P(l, i).$$

Now we present some bounds on the generalized Krein parameters. They can be obtained by making use of the properties (5)-(8).

Theorem 2: Consider an association scheme Ω with d classes. Then, for all natural numbers i, x, y, α and β such that $0 \leq i, \alpha, \beta \leq d$, we have

$$0 \leq q_{(\alpha, \beta; x, y)}^i \leq 1.$$

The following result presents another upper-bound on the generalized Krein parameters associated to only one idempotent.

Theorem 3: Let Ω be an association scheme with d classes on a finite set of order n . Let i, s, x and y be natural numbers such that $0 \leq i, s \leq d$. Then the generalized Krein parameter $q_{(i,i;x,y)}^s$, with $x + y = m$, satisfies $q_{(i,i;x,y)}^s \leq (\frac{\mu_i}{n})^{m-1}$.

Proof: Since $q_{im}^s = \sum_{t=0}^d (Q(t, i))^m P(s, t)$ and $|Q(t, i)| \leq \frac{\mu_i}{n}$, and $|P(s, t)| \leq n_t$ see (i)-(iv), we conclude that

$$\begin{aligned}
 q_{im}^s &= \sum_{t=0}^d (Q(t, i))^m P(s, t) \\
 &= \left| \sum_{t=0}^d (Q(t, i))^m P(s, t) \right| \\
 &\leq \sum_{t=0}^d |(Q(t, i))^m| |P(s, t)| \\
 &\leq \sum_{t=0}^d |Q(t, i)|^{m-2} (Q(t, i))^2 n_t \\
 &\leq \sum_{t=0}^d \left(\frac{\mu_i}{n} \right)^{m-2} (Q(t, i))^2 n_t \\
 &\leq \left(\frac{\mu_i}{n} \right)^{m-2} \sum_{t=0}^d (Q(t, i))^2 n_t \\
 &\leq \left(\frac{\mu_i}{n} \right)^{m-2} \frac{\mu_i}{n} \\
 &= \left(\frac{\mu_i}{n} \right)^{m-1}.
 \end{aligned}$$

Proceeding in an analogous manner as we have done in the proof of Theorem 3, we obtain the following result.

Theorem 4: Let Ω be an association scheme with d classes on a finite set of order n . Let i, j, k, s be natural numbers such that $0 \leq i, j, s \leq d$ and $i < j$. Then, for any natural $k \geq 1$, the generalized Krein parameter $q_{(i,j;k,k)}^s$ satisfies the inequality

$$q_{(i,j;k,k)}^s \leq \left(\frac{1}{2} \right)^k \left(\frac{2(\max\{\mu_i, \mu_j\})^2}{n^2} \right)^{k-1}.$$

From Theorems 3 and 4 we conclude the following corollary that states the above bounds for the classical Krein parameters of association schemes.

Corollary 2: Let Ω be an association scheme with d classes on a finite set of order n . Let i, k, l be natural numbers such that $0 \leq i, j, l \leq d$ and $i < j$. Then:

- (i) $q_{(i,i;1,1)}^l \leq \frac{\mu_i}{n}$;
- (ii) $q_{(i,j;1,1)}^l \leq \frac{1}{2}$.

III. CONCLUSIONS

In this paper we have generalized the Krein parameters of an association scheme and through this generalization we have obtained new conditions over the classical Krein parameters. In fact, the results obtained in Corollary 2 present upper-bounds on the classical Krein parameters that we will show that cannot be improved, providing suitable examples.

Example 1: In this example we consider association schemes with two classes which are equivalent to strongly regular graphs.

- (a) Let us consider the family of strongly regular graphs known as the conference graphs. A member of this family of order n satisfies $\mu_0 = 1, \mu_1 = \frac{n-1}{2}$ and $\mu_2 = \frac{n-1}{2}$. Also, we have: $q_{(1,1;1,1)}^0 = 1/2 - 1/2n = \mu_1/n$. Therefore, the upper-bound presented in (i) of Corollary 2 is attained.
- (b) Now, we consider the family of strongly regular graphs known as the cocktail party graphs. For a member of this family of order $2l$ we have:

$$q_{(1,2;1,1)}^1 = \frac{l-1}{2l},$$

and therefore the upper-bound presented in (ii) of Corollary 2 is asymptotically attained.

Example 2: In this example we present a family of association schemes with three classes constructed from symmetric designs. This family has an infinite number of elements and it is presented and studied in [7], where the following definition can be seen.

Let \mathcal{P} be a set of points and \mathcal{B} be a set of blocks, where a *block* is a subset of \mathcal{P} . Then, the ordered pair $(\mathcal{P}, \mathcal{B})$ is a *symmetric design* with parameters (n, k, c) , with $c < k$, if it satisfies the following properties:

- (i) \mathcal{B} is a subset of the power set of \mathcal{P} ;
- (ii) $|\mathcal{P}| = |\mathcal{B}| = n$;
- (iii) $\forall b \in \mathcal{B}, |b| = k$;
- (iv) $\forall p \in \mathcal{P}, |\{b \in \mathcal{B} : p \in b\}| = k$;
- (v) $\forall p_1, p_2 \in \mathcal{P}, p_1 \neq p_2, |\{b \in \mathcal{B} : p_1, p_2 \in b\}| = c$;
- (vi) $\forall b_1, b_2 \in \mathcal{B}, b_1 \neq b_2, |\{p \in \mathcal{P} : p \in b_1 \wedge p \in b_2\}| = c$.

Given a symmetric design with parameters (n, k, c) , we build a three class association scheme, as in [7], in the following manner. Let $X = \mathcal{P} \cup \mathcal{B}$. We define the following relations in $X \times X$:

$$\begin{aligned}
 R_0 &= \{(x, x) : x \in X\}; \\
 R_1 &= \{(x, y) \in \mathcal{P} \times \mathcal{B} : x \in y\} \cup \{(y, x) \in \mathcal{B} \times \mathcal{P} : x \in y\}; \\
 R_2 &= \{(x, y) \in \mathcal{P} \times \mathcal{P} : x \neq y\} \cup \{(x, y) \in \mathcal{B} \times \mathcal{B} : x \neq y\}; \\
 R_3 &= \{(x, y) \in \mathcal{P} \times \mathcal{B} : x \notin y\} \cup \{(y, x) \in \mathcal{B} \times \mathcal{P} : x \notin y\}.
 \end{aligned}$$

Through the axioms (i) – (vi) of a symmetric design it is proved that R_0, R_1, R_2, R_3 constitute an association scheme with three classes over X . From the relations above we compute the intersection matrices of the association scheme, given by $L_0 = I_4$,

$$\begin{aligned}
 L_1 &= \begin{pmatrix} 0 & k & 0 & 0 \\ 1 & 0 & k-1 & 0 \\ 0 & c & 0 & k-c \\ 0 & 0 & k & 0 \end{pmatrix}, \\
 L_2 &= \begin{pmatrix} 0 & 0 & n-1 & 0 \\ 0 & k-1 & 0 & n-k \\ 1 & 0 & n-2 & 0 \\ 0 & k & 0 & n-k-1 \end{pmatrix}, \\
 L_3 &= \begin{pmatrix} 0 & 0 & 0 & n-k \\ 0 & 0 & n-k & 0 \\ 0 & k-c & 0 & n-2k+c \\ 1 & 0 & n-k-1 & 0 \end{pmatrix}.
 \end{aligned}$$

Now, using axioms (a) – (d) of the matrices of the Bose-Mesner algebra, $\mathcal{A} = \{A_0, A_1, A_2, A_3\}$, we obtain:

- $A_0 \times A_i = A_i \times A_0 = A_i$, for $i \in \{0, 1, 2, 3\}$;
- $A_1 \times A_1 = kA_0 + cA_2$;
- $A_1 \times A_2 = A_2 \times A_1 = (k - 1)A_1 + kA_3$;
- $A_1 \times A_3 = A_3 \times A_1 = (k - c)A_2$;
- $A_2 \times A_2 = (n - 1)A_0 + (n - 2)A_2$;
- $A_2 \times A_3 = A_3 \times A_2 = (n - k)A_1 + (n - k - 1)A_3$;
- $A_3 \times A_3 = (n - k)A_0 + (n - 2k + c)A_2$.

Now we can calculate the powers of A_1 to obtain the following polynomial:

$$p_{A_1}(\lambda) = \lambda^4 + (-k^2 - k + c)\lambda^2 + k^2(k - c), \quad (11)$$

such that $p_{A_1}(A_1) = \mathcal{O}_n$, where \mathcal{O}_n denotes the n dimensional null matrix. Then A_1 has four distinct eigenvalues and therefore the least natural number such that the set $\{I_n, A_1, A_1^2, \dots, A_1^k\}$ is linear dependent is 4. Then, we conclude that the polynomial (11) is the minimal polynomial of A_1 .

Applying formula (1) to the matrix A_1 , considering the eigenvalues of the polynomial (11), $\lambda_0 = k$, $\lambda_1 = -k$, $\lambda_2 = \sqrt{k - c}$ and $\lambda_3 = -\sqrt{k - c}$, and taking into account the equality

$$(n - 1)c = k(k - 1), \quad (12)$$

satisfied by these symmetric designs with parameters (n, k, c) , see [5], we obtain the elements of the unique basis of minimal orthogonal idempotents of \mathcal{A} :

$$\begin{aligned} E_0 &= \frac{A_0 + A_1 + A_2 + A_3}{2n} = \frac{J_n}{2n}; \\ E_1 &= \frac{A_0 - A_1 + A_2 - A_3}{2n}; \\ E_2 &= \frac{(n - 1)\sqrt{k - c}A_0 + (n - k)A_1 - \sqrt{k - c}A_2 - kA_3}{2n\sqrt{k - c}}; \\ E_3 &= \frac{(n - 1)\sqrt{k - c}A_0 - (n - k)A_1 - \sqrt{k - c}A_2 + kA_3}{2n\sqrt{k - c}}. \end{aligned}$$

Now we apply equalities (2) and (3) to compute the matrices P and Q , respectively:

$$\begin{aligned} P &= \begin{pmatrix} 1 & k & n - 1 & n - k \\ 1 & -k & n - 1 & k - n \\ 1 & \sqrt{k - c} & -1 & -\sqrt{k - c} \\ 1 & -\sqrt{k - c} & -1 & \sqrt{k - c} \end{pmatrix}, \\ Q &= \frac{1}{2n} \begin{pmatrix} 1 & 1 & n - 1 & n - 1 \\ 1 & -1 & -\frac{k - n}{\sqrt{k - c}} & \frac{k - n}{\sqrt{k - c}} \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -\frac{k}{\sqrt{k - c}} & \frac{k}{\sqrt{k - c}} \end{pmatrix}. \end{aligned}$$

Finally, we obtain the dual intersection matrices of this association scheme by applying formula (10) from Proposition 1

and taking into account equality (12): $L_0^* = I_4/2n$,

$$\begin{aligned} L_1^* &= \frac{1}{2n} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \\ L_2^* &= \frac{1}{2n} \begin{pmatrix} 0 & 0 & n - 1 & 0 \\ 0 & 0 & 0 & n - 1 \\ 1 & 0 & \frac{n - 2}{2} + \frac{n - 2k}{2\sqrt{k - c}} & \frac{n - 2}{2} - \frac{n - 2k}{2\sqrt{k - c}} \\ 0 & 1 & \frac{n - 2}{2} - \frac{n - 2k}{2\sqrt{k - c}} & \frac{n - 2}{2} + \frac{n - 2k}{2\sqrt{k - c}} \end{pmatrix}, \\ L_3^* &= \frac{1}{2n} \begin{pmatrix} 0 & 0 & 0 & n - 1 \\ 0 & 0 & n - 1 & 0 \\ 0 & 1 & \frac{n - 2}{2} - \frac{n - 2k}{2\sqrt{k - c}} & \frac{n - 2}{2} + \frac{n - 2k}{2\sqrt{k - c}} \\ 1 & 0 & \frac{n - 2}{2} + \frac{n - 2k}{2\sqrt{k - c}} & \frac{n - 2}{2} - \frac{n - 2k}{2\sqrt{k - c}} \end{pmatrix}. \end{aligned}$$

From the dual intersection matrices presented above, it is possible to extract some evidence of the optimality of the upper bound 1/2, for the Krein parameters q_{ij}^l , with $i \neq j$, presented in Corollary 2, (ii). In fact, we can observe that

$$q_{23}^0 = (L_2^*)_{03} = \frac{n - 1}{2n}$$

and this value converges to 1/2, when n tends to infinity.

We are also able to provide other examples of association schemes with a number of classes greater than two for which the Krein parameters converge to our upper-bounds presented in Theorems 3 and 4, by making use of the Kronecker product of association schemes.

ACKNOWLEDGMENT

Luís Vieira research partially funded by the European Regional Development Fund through the program COMPETE and by the Portuguese Government through the FCT - Fundação para a Ciência e a Tecnologia under the project PEest-C/MAT/UI0144/2013.

REFERENCES

- [1] R. A. Bailey, Association Schemes, Designed Experiments, Algebra and Combinatorics, Cambridge University Press, Cambridge, 2004.
- [2] R. C. Bose and T. Shimamoto, Classification and analysis of partially balanced incomplete block designs with two associate classes, J. Am. Statist. Assoc. **47**, 151–184, 1952.
- [3] R. C. Bose and D. M. Mesner, On linear associative algebras corresponding to association schemes of partially balanced designs, Ann. Math. Statist. **30**, 21–38, 1959.
- [4] Ph. Delsarte, An algebraic approach to the association schemes of coding theory, Philips Res. Rep. Suppl. **10**, 1973.
- [5] J. H. v. Lint, A Course in Combinatorics, Cambridge University Press, Cambridge, 2006.
- [6] L. L. Scott Jr., A condition on Higman’s parameters, Notices of Amer. Math. Soc. **20** A-97, 721-20-45, 1973.
- [7] G. Shakan and Y. Xin, Q -Polynomial Association Schemes with Irrational Eigenvalues, A Major Qualifying Project submitted to the Faculty of the Worcester Polytechnic Institute, 2012.

Ridge regression and bootstrapping in asthma prediction

Ioannis I. Spyroglou, Eleni A. Chatzimichail, E.N. Spanou, E. Paraskakis, and Alexandros G. Rigas

Abstract—Asthma persistence prediction accuracy is a very important matter, as the most important issue about this chronic disease is the identification during the early ages. The early detection of the preschoolers whose asthma persists after the age of five could lead into better treatment of asthma for the next years of a human life. In this case, the use of generalized linear models is proposed for asthma prediction. In particular, the presence of multicollinearity among the data leads to the use of the penalized likelihood function and more specific to Logistic Ridge Regression. Furthermore, a test for the evaluation of the fitted model is presented based on the randomized quantile residuals which follow a Gaussian distribution. The QQ-plot is used with the addition of the 5% rejection regions of the randomized quantile residuals with the help of a proper bootstrap.

Keywords—Asthma outcome, Logistic Ridge regression, 10-fold cross validation, Randomized Quantile Residuals, Bootstrap

I. INTRODUCTION

ASTHMA is a disease with polymorphic phenotype affected by several environmental and genetic factors which both play a key role in the development and persistence of the disease [1]. Among these factors seasonal symptoms, wheezing episodes during childhood and several prenatal and environmental factors are included [2].

Most children who suffer from asthma develop their first symptoms before the 5th year of age. For the diagnosis of asthma a detailed medical history and physical examination along with a lung function test is usually required. On the other hand, lung function test is hard to be performed in children younger than five years old, so most of the times the diagnosis of the disease is mainly based on the findings of the physical examination and the presence of characteristic personal and family medical history. Although the majority of

Ioannis Spyroglou is with the Electrical Engineering Department, Democritus University of Thrace, Xanthi, CO 67100 GRRECE (corresponding author to provide phone: +306955954849 ; e-mail: ispyrogl@ee.duth.gr).

Eleni Chatzimichail is with the Electrical Engineering Department, Democritus University of Thrace, Xanthi, CO 67100 GRRECE (e-mail: echatzim@ee.duth.gr).

E.N. Spanou is with the Electrical Engineering Department, Democritus University of Thrace, Xanthi, CO 67100 GRRECE (e-mail: ispanou@ee.duth.gr).

Emmanouil Paraskakis, Prof., is with the Medical School, Democritus University of Thrace, Alexandroupolis, CO 68100 GRRECE (e-mail: eparaska@med.duth.gr).

Alexandros Rigas, Prof., is with the Electrical Engineering Department, Democritus University of Thrace, Xanthi, CO 67100 GRRECE (e-mail: rigas@ee.duth.gr).

the preschool children with asthma overcome their disease by the school age a substantial number of preschoolers exhibit a persistence of symptoms requiring early identification and treatment [3].

In preventive medicine, the value of a test lies in its ability to identify those individuals who are at high risk of an illness and who therefore require intervention while excluding those who do not require such intervention. The accuracy of the risk classification is of particular relevance in the case of asthma disease. Early identification of patients at high risk for asthma disease progression may lead to better treatment opportunities and hopefully better disease outcomes in adulthood [3].

II. MATERIALS AND METHODS

A. Clinical Data

Data from 148 patients were collected from the Pediatric Department of the University Hospital of Alexandroupolis, Greece during the period from 2008 to 2010. A group of 148 patients who were diagnosed for asthma were studied prospectively from the 7th to the 14th year of age. From this sample, 36 patients were removed because of missing data. The history of each case was obtained by questionnaire. A second group of 33 children was used for validation of the efficacy of the constructed model in real life. In this group of preschool children the proposed model was used to predict asthma persistence in school age. At mean age (\pm SD) of 9.2 ± 2.7 years these children were re-evaluated. The new dataset has 18 available predictors which are going to be used in the logistic model. The 18 used prognostic factors have been derived by previous studies [1-3] and they are described in Table I. The encoding of the prognostic factor “seasonal symptoms” is presented in Table II.

TABLE I

Category	Prognostic Factors
Demographic	Age, height, weight, waist's perimeter
Bronchiolitis episodes	Until 3 rd year, between 3 rd – 5 th year
Symptoms	Wheezing, cough, allergic rhinitis, allergic conjunctivitis, dyspnea, congestion, runny nose, seasonal symptoms
Pharmaceutical therapy	Antileukotriene, antihistamine, corticosteroids inhaled
Asthma	Diagnosis of asthma (dependent variable), Treatment

The 18 used prognostic factors.

TABLE II

1 (none)	2 (Winter)	3 (Autumn)	4 (Spring)	5 (Summer)	6 (>2seasons)
-------------	---------------	---------------	---------------	---------------	------------------

The encoding of “seasonal symptoms”.

B. Multicollinearity

Generalized Linear Models and Regression Analysis are two of the most important and popular statistical approaches used in biomedical research [4]. In many cases it has been observed that medical data exhibit strong correlations between the predictor variables, a condition known as multicollinearity. Multicollinearity was introduced as a concept by Frisch [5], in order to illustrate a situation, where the variables are subject to two or more correlations.

One of the main consequences of multicollinearity is that the least squares estimates often do not make any sense, and the standard errors of the parameter estimates are very large or the t-ratios are very low. Therefore multicollinearity could lead into inaccurate results. For example, when the null hypothesis that the parameters of the model are zero is rejected, but none of the estimated parameters have a p-value less than 0.5. One of several methods that have been used in order to overcome the multicollinearity problem is the Ridge Regression method which was introduced in [6]. When multicollinearity appears, the ridge estimator has a smaller total Mean Square Error (MSE) than the maximum likelihood estimator. Eigenvalues of the correlation matrix of the independent variables near zero indicate multicollinearity.

Ridge Regression (RR) is an alternative estimation method of the unknown parameters of the linear regression models and belongs to the category of biased regression methods [7-8]. This method introduces a bias in the regression equation in order to reduce the variance of the parameter estimates. This bias is entered with the ridge parameter, which determines the extent of the shrinkage of the least squares estimates. Also in [9] the ridge estimator was introduced for Logistic Regression, which is one of the most popular methods used for binary data modeling. Generally, this method is differentiated from the maximum likelihood as a penalty term is added, which includes the ridge parameter.

C. Ridge Regression

Let $y_i, i = 1, \dots, n$ be the binary responses of n random variables Y_i , where $Y_i \sim B(1, p_i)$, and x_i a vector of explanatory variables which consist of covariates (numerical or binary) and dummy variables corresponding to factor levels.

The logistic regression model is given by:

$$p_i = \frac{\exp(bx_i)}{\{1 + \exp(bx_i)\}}, \tag{1}$$

where b is the parameter vector [10-11]. This model is implemented without the use of a constant term.

Now, the maximum likelihood estimates of the parameters $b_j, j=1, \dots, k$ and from them the probabilities p_i are obtained by maximizing the following likelihood function

$$L(b|y) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}, \quad y_i = 0, 1, \tag{2}$$

or by maximizing the log – likelihood function using a Newton – Raphson algorithm which is:

$$l(b|y) = \log L(b|y), \tag{3}$$

$$l(b|y) = \sum_{i=1}^n \left[y_i \log \left[\frac{1}{1 + \exp(-x_i^T b)} \right] + (1 - y_i) \log \left[1 - \frac{1}{1 + \exp(-x_i^T b)} \right] \right] \tag{4}$$

As it was mentioned before when multicollinearity exists, in order to obtain more stable parameter estimates the logistic ridge regression is used. In ridge logistic regression the penalized maximum likelihood is used and is given by [12]:

$$l^\lambda(b|y) = l(b|y) - \lambda \|b\|^2 = l(b|y) - \lambda R, \tag{5}$$

and is known as restricted maximum likelihood function, whereas $l(b|y)$ is the unrestricted maximum likelihood and R is a penalty term of the following form [13]:

$$R = \sum_{j=0}^{k-1} (b_{j+1} - b_j)^2. \tag{6}$$

Generally the difference between this approach, and the approach of maximum likelihood function is the use of the penalty term which includes the ridge parameter. The ridge parameter is a positive number and its main role is the regulation of the significance of the penalty term R [13]. Therefore it is obvious that when $\lambda = 0$ the estimates produced are the same as the ones obtained by the unrestricted maximum likelihood function. The computational procedure of the penalized parameter estimates \hat{b}^λ is based on the Newton – Raphson algorithm. However, a transformation of the linear estimates of the unrestricted logistic regression model is required, since the term R given by (6), should be in the form of (5). Therefore:

$$\begin{aligned} b_1 x_{i1} + b_2 x_{i2} + \dots + b_k x_{ik} &= b_1 z_{i1} + (b_2 - b_1) z_{i2} + \dots + (b_k - b_{k-1}) z_{ik} \\ &= \gamma_1 z_{i1} + \gamma_2 z_{i2} + \dots + \gamma_k z_{ik}, \end{aligned} \tag{7}$$

where

$$\gamma_1 = b_1, \dots, \gamma_j = b_j - b_{j-1}, \quad j = 2, \dots, k \tag{7.1}$$

and $z_{ij} = \sum_{u=j}^k x_{iu}$ (7.2). Thus, the penalized maximum likelihood becomes as follows:

$$l^\lambda(\gamma|y) = l(\gamma|y) - \lambda \|\gamma\|^2. \tag{8}$$

The first derivative of equation (8) is now:

$$U^\lambda(\gamma) = \sum_{i=1}^n z_i \{y_i - p_i\} - 2\lambda \gamma = U(\gamma) - 2\lambda \gamma. \tag{9}$$

Then, calculating the negative second derivative we get:

$$\Omega^\lambda(\gamma) = \Omega(\gamma) + 2\lambda I, \tag{10}$$

where $\Omega(\gamma) = z^T W z$ and W is the $n \times n$ weight matrix which is diagonal with elements $W_{ii} = p_i(1 - p_i)$.

Applying the Taylor series expansion in the first derivative of the penalized maximum likelihood function, the properties which are valid for large sample can be obtained. Consequently:

$$U^\lambda(\gamma^\lambda) = U^\lambda(\gamma_0) - (\hat{\gamma}^\lambda - \gamma_0)\Omega^\lambda(\gamma_0) + o(\|\hat{\gamma}^\lambda - \gamma_0\|). \quad (11)$$

Using equations (9) and (10) and setting (11) equal to 0 it leads to:

$$\hat{\gamma}^\lambda = \{\Omega(\gamma) + 2\lambda I\}^{-1}\{U(\gamma_0) + \gamma_0\Omega(\gamma_0)\}. \quad (12)$$

D. Choosing the ridge parameter

The most difficult task in RR is to determine the ridge parameter. In this study, we chose the value of the ridge parameter that minimized the Mean Squared Error through 10-fold cross validation[12].

$$MSEcv = \frac{1}{n} \left(\sum_i \{Y_i - \hat{p}_i(X_i)\}^2 \right) \quad (13)$$

The average value of the MSE was considered the overall cross-validation error of the model. We selected the ridge parameter as the one with the minimum cross-validation error.

E. Residuals and bootstrapping

After fitting the model to the observed data, it is necessary to check if the fitted model is valid. A usual technique which is used for validity examination of the model is based on the residuals. In the case of logistic regression with binary response, the distributions of Pearson residuals which are defined by $r_{p,i} = (y_i - \hat{p}_i) / \sqrt{\hat{p}_i(1 - \hat{p}_i)}$, $i = 1, \dots, n$ and of deviance residuals which are defined by, $r_D = \text{sign}\{y_i - \hat{p}_i\}$ are far from normal. In addition, plots of the residuals against the explanatory variables, which are usually used in generalized linear models for model checking, are uninformative in a binary case and are not recommended. More details about the residuals are given in [14].

Let $F(y_i; p_i) = P(Y_i \leq y_i) = \sum_{m=0}^{\lfloor y_i \rfloor} p_i^m (1 - p_i)^{1-m}$ be the cumulative binomial distribution of the i th binary response, and $\lfloor y_i \rfloor$ is the greatest integer less than or equal to y_i , i.e. the ‘floor’ under y_i . Then the randomized quantile residuals for a logistic regression model are defined by

$$r_{rq,i} = \Phi^{-1}\{u\}, \quad (14)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal, and u_i is a uniform random variable on the interval

$$\begin{aligned} (a_i, b_i] &= \left(\lim_{y \uparrow y_i} F(y; \hat{p}_i), F(y; \hat{p}_i) \right) \\ &\approx [F(y_i - 1; \hat{p}_i), F(y; \hat{p}_i)] \end{aligned}$$

The randomized quantile residuals defined by (14) follow exactly the standard normal distribution, apart from sampling variability in \hat{p}_i . These residuals [15] can be used for any

discrete distributed response. Thus, the validity of the model can now be tested by using goodness of fit tests for the normality of $r_{rq,i}$. A very strong method to test the null hypothesis that the randomized quantile residuals follow a standard normal distribution i.e. $r_{rq} \sim N(\mathbf{0}, I)$ that is commonly used to check if a data sample comes from a normal distribution is the Anderson – Darling test[19].

Also the Q-Q plot of the randomized quantile residuals has been proposed by Dunn and Smyth [15] as a mean for checking the validity of the model. Here a method for constructing pointwise a $\times 100\%$ rejection regions around the Q-Q plot of any random sample is proposed by using bootstrapping [16-17]. Because of the large number of the estimated parameters, the additional uncertainty due to the estimation of the regression parameters must be taken into consideration. Therefore a proper bootstrap of the randomized quantile residuals must be used in order to take the above into account. Residual resampling is known to be an appropriate bootstrap process for studying the properties of the estimates [22-23]. Moreover this bootstrap is very important since the standard errors of the ridge estimated parameters can be obtained as it was mentioned before. The bootstrap is implemented with the next steps:

- Step 1: Obtain estimates of p_i , and randomized quantile residuals with the use of logistic ridge regression.
- Step 2: Bootstrapping 2000 times the randomized quantile residuals obtained by the logistic ridge model. So now we have $r_{rq,1}^T, \dots, r_{rq,2000}^T$. We use the randomized quantile residuals because they have unit variance as they approximate standard normal distribution [15-20].
- Step 3: Apply logistic ridge regression 2000 times using as response the summations $\hat{p}^T + r_{rq,t}^T$, $t = 1, \dots, 2000$, where \hat{p}^T are the estimated probabilities from Step 1. In this step if a sum $\hat{p}_i + r_{rq,i} > 1$ then this becomes 1. Also if sum < 0 then it becomes 0 and finally we round to the nearest integer if $0 < \text{sum} < 1$ [20-21]. Moreover 2000 samples of \hat{b}^λ and \hat{p} can be obtained.
- Step 4: The standard errors of the estimated parameters \hat{b}^λ can be obtained by finding the standard deviation of the 2000 bootstrapped samples $\hat{b}_1^\lambda, \dots, \hat{b}_{23}^\lambda$.
- Step 5: From the 2000 sets of estimated response variables \hat{p}_t , $t = 1, \dots, 2000$, we calculate 2000 new sets of randomized quantile residuals which allows us to construct a $\times 100\%$ rejection regions around the Q-Q plot of the randomized quantile residuals.

III. RESULTS

The correlations between some variables are very strong and statistically significant, indicating the presence of multicollinearity. As a first step it is necessary to transform the categorical variables with more than two categories into dummy variables. For the detection of multicollinearity we may use the Condition Indices, by calculating the eigenvalues

of the correlation matrix and other similar procedures as in linear regression models [17-19].

Another problem caused by the multicollinearity is the large values of the standard errors of the estimated parameters, which makes the model unstable. Moreover, while the model according to the F-test seems to be statistically significant against the null hypothesis ($b_1 = b_2 = \dots = b_{23} = 0$), the p-values of the individual terms are all greater than 0.05 which suggests that none of the variables is statistically significant.

Thus the logistic ridge regression is applied to generate an improved model with more stable parameter estimates for a ridge parameter $\lambda=0$ to $\lambda=0.5$. Furthermore when collinearity exists there is always a model for $\lambda>0$ for which the MSE is less than the MSE of the unrestricted model [8-12].

For the calculation of p – values the following statistic is used:

$$T_\lambda = \frac{\hat{b}_j^\lambda}{se(\hat{b}_j^\lambda)} \quad (15)$$

The standard errors were obtained by the bootstrap procedure that was described in section E. Thereafter we assume that under the null hypothesis $T_\lambda \sim N(0,1)$ to test the significance of the ridge coefficients [24].

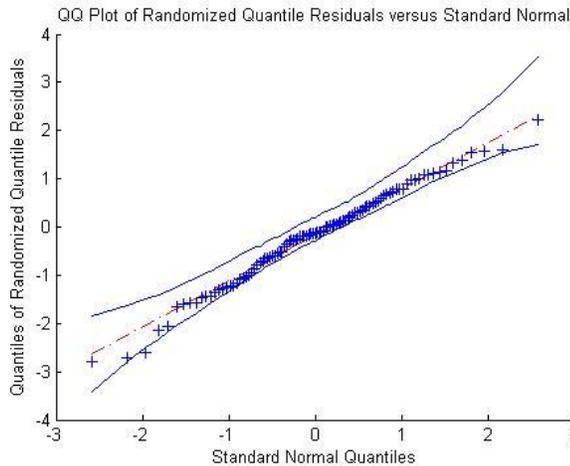


Fig. 1 Q-Q Plot of randomized quantile residuals versus Standard Normal

For $\lambda=0.0261$ the minimum MSE_{cv} is derived which is equal to 0.034. The parameter estimates of the logistic ridge model are shown in Table III. It now becomes clear that the prognostic factors Waist’s perimeter, Congestion, Cough, Wheezing, Dyspnea, Bronchiolitis episodes until 3rd year are statistically significant. Moreover it is obvious that the estimates now are now much more stable than the estimates of the initial unrestricted model and this is indicated from the two first columns of Table III.

We now proceed to provide a test in order to check the validity of the model. Figure 1 shows the Q-Q plot of the randomized quantile residuals of the fitted logistic ridge model denoted with +. The 5% rejection regions were computed by the procedure described in Section 5 after 2000 bootstrap simulations. Only 1 (0.99%) of the 101 residuals lie outside the 5% rejection regions and generally the Q-Q plot does not present serious deviations from normality.

Here it is important to mention that if the percentage of the randomized quantile residuals that are outside the rejection regions is greater than 5%, then the model should be rejected.

In addition, the powerful Anderson-Darling test gives the value 0.3456 with a p-value 0.4810. Therefore, the null hypothesis that the randomized quantile residuals follow an approximate standard normal distribution cannot be rejected, which suggests that the fitted model is valid.

Finally we would like to examine the performance of the proposed model in new real data. These data were collected in a period after 2010 and refer to 33 new patients.

Based on the equation:

$$\hat{p}_{ridge} = \frac{1}{1 + \exp(-X_{new} * \hat{b}_{ridge})}$$

a prediction for the diagnosis of a new patient can be found. The positive predicted value, the negative predicted value and the accuracy of this model are estimated using false positive (FP), false negative (FN), true positive (TP), and true negative (TN) values. The test set consists of the new 33 patients and the 11 patients which were used for the cross – validation test.

$$\begin{aligned} \text{Positive Pred. Value} &= \frac{N_{TP}}{N_{TP} + N_{FN}} \times 100, \\ \text{Negative Pred. Value} &= \frac{N_{TN}}{N_{TN} + N_{FP}} \times 100, \\ \text{Accuracy} &= \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}} \times 100. \end{aligned} \quad (16)$$

All the above are statistical measures of the performance of a binary classification test [25].

IV. CONCLUSION

In this paper, a new intelligent method based on the Logistic Ridge Regression for asthma persistence prediction has been validated in preschool patients in real time data. The proposed model predicted the persistence of asthma at the approximate age of nine years with an accuracy of 93.18%, positive predictive value of 96.15% and negative predictive value of 88.89%. A comparison of this work against [25] could be carried out. In [25] neural networks are used for the prediction of asthma outcome. To conclude, the proposed method exhibits high accuracy in asthma persistence prediction and shows the importance priority of each factor in asthma persistence. A better prediction rate will be possible by increasing the patient data and further clinical evaluation may enhance the implications of the present study. Finally, for future research, we could collect data from different regions, with different environmental and climatic factors, to examine if asthma prediction is affected by them.

Table III

Covariates	Estimates			
	Parameter Estimates	Standard Errors	T_λ	p-values
Age	0,059821	0,1221	0,4900	0,6241
Treatment	0,531238	0,4817	1,1028	0,2701
Corticosteroids inhaled	0,889768	0,4942	1,8002	0,0718
Antileukotriene	-0,32763	0,5650	-0,5799	0,5620
Antihistamine	0,111097	0,6674	0,1665	0,8678
Height	0,600211	0,7353	0,8163	0,4143
Weight	-0,01082	0,0276	-0,3925	0,6947
Waist's perimeter	-0,06579	0,0216	-3,0455	0,0023
Allergic rhinitis	-0,06907	0,5733	-0,1205	0,9041
Allergic conjunctivitis	-0,72709	0,5819	-1,2494	0,2115
Runny nose	0,429069	0,6068	0,7071	0,4795
Congestion	1,096703	0,5424	2,0220	0,0432
Cough	1,640577	0,5871	2,7942	0,0052
Wheezing	1,719255	0,5874	2,9271	0,0034
Dyspnea	1,18429	0,5962	1,9865	0,0470
Seasonal symptoms (none)	1,26928	0,7360	1,7246	0,0846
Seasonal symptoms (winter)	1,148824	0,8505	1,3507	0,1768
Seasonal symptoms (autumn)	0,273617	0,8385	0,3263	0,7442
Seasonal symptoms (spring)	0,856916	0,9088	0,9429	0,3457
Seasonal symptoms (summer)	0,216933	0,8830	0,2457	0,8059
Seasonal symptoms (>2 seasons)	1,15655	0,8174	1,4149	0,1571
Bronchiolitis episodes until 3 rd year	-0,21639	0,1089	-1,9866	0,0470
Bronchiolitis episodes b/w 3 rd - 5 th year	0,132402	0,0932	1,4212	0,1553

REFERENCES

[1] C. Porsbjerg, M.L. von Linstow, C. Ulrik, S. Nepper-Christensen, V. Backer, "Risk factors for onset of asthma: a 12-year prospective follow-up study," *Chest*, vol. 129, no. 2, pp. 309–16, 2006.

[2] N. N. Hansel, E. C. Matsui, R. Rusher, M. C. McCormack, J. Curtin-Brosnan, R. D. Peng, D. Mazique, P. N. Breyse, G. B. Diette, "Predicting future asthma morbidity in preschool inner-city children," *Journal of Asthma*, vol. 48, no.8, pp. 797–803, 2011.

[3] A. Bush, "Diagnosis of asthma in children under five," *Prim Care Respir J*, vol. 16, pp. 7–15, 2007.

[4] R. K. Jain, "Ridge Regression and its Application to Medical Data," *Computers and Biomedical Research*, vol. 18, pp. 363–368, 1984.

[5] R. Frisch, *Statistical Confluence Analysis by Means of Complete Regression Systems*. Oslo: University Institute of Economics, Publication no. 5, 1934.

[6] A. E. Hoerl, "Application of ridge analysis to regression problems," *Chemical Engineering Progress*, vol. 58, no.3, pp. 54–59, 1962.

[7] A.E. Hoerl and R. W. Kennard, "Ridge Regression: Applications to nonorthogonal problems," *Technometrics*, vol. 12, no. 6, 1970.

[8] A.E. Hoerl and R.W. Kennard, "Ridge Regression: Biased estimates for nonorthogonal problems," *Technometrics*, vol. 12, no.55, pp.55–67 1970.

[9] R. L. Schaefer, L.D. Roi, and R.A. Wolfe, "A ridge logistic estimator," *Communications in Statistics - Theory and Methods*, vol. 13, no. 1, pp. 99–113, 1984.

[10] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. London: Chapman & Hall, 1989.

[11] Annette J. Dobson, *An Introduction To Generalized Linear Models*, 2nd ed.: Chapman & Hall, 2002.

[12] S. Le Cessie and J. C. Van Houwelingen, "Ridge Estimators in Logistic Regression," *Journal of the Royal Statistical Society*, vol. 41, no. 1, pp. 191–201, 1992.

[13] D.R. Brillinger, K.A. Lindsay, and J.R. Rosenberg, "Combining frequency and time domain approaches to systems with multiple spike train input and output," *Biological Cybernetics*, vol. 100, pp. 459–474, 2009.

[14] D.A. Pierce and D.W. Schafer, "Residuals in Generalized Linear Models," *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 977–986, 1986.

[15] P. Dunn and G. K. Smyth, "Randomized Quantile Residuals," *J. Computat. Graph. Statist*, vol. 5, pp. 236–244, 1996.

[16] B. Efron and R.J. Tibshirani, "An Introduction to the Bootstrap," (Chapman & Hall, New York), 1993.

[17] E. Lesaffre, E. and B.D. Marx, "Collinearity in Generalized Linear Regression," *Communications in Statistics – Theory and Methods*, vol. 22, no. 7, pp. 1933–1952, 1993.

[18] B. D. Marx and E. P. Smith, "Weighted Multicollinearity in Logistic Regression: Diagnostics and Biased Estimation Techniques with an Example From Lake Acidification," *Canadian Journal of Fisheries and Aquatic Sciences*, vol.47, no. 6, pp. 1128–1135, 1990.

[19] T.W. Anderson and D.A. Darling, "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes," *The Annals of Mathematical Statistics*, vol. 23, no. 2, pp. 193–212, 1952.

[20] H. Friedl and N. Tilg, "Variance estimates in logistic regression using the bootstrap," *Communications in Statistics - Theory and Methods*, vol. 24, no. 2, pp. 473–486, 1995.

[21] D. Firth, J. Glosup, and D.V. Hinkley, "Model Checking with nonparametric curves," *Biometrika*, vol. 78(2), pp. 245-252, 1991.

[22] B. Efron, Bootstrap methods: "Another look at the jackknife," *Ann.Statist.*, vol. 7, pp. 1–26, 1979.

[23] D.A. Freedman, "Bootstrapping regression models," *Ann. Statist.*, vol. 9, pp. 1218–1228, 1981.

[24] Cule et al., "Significance testing in ridge regression for genetic data," *BMC Bioinformatics*, 12:372, 2011.

[25] E. Chatzimichail, E. Paraskakis, and A. Rigas, "Predicting Asthma Outcome Using Partial Least Square Regression and Artificial Neural Networks," *Advances in Artificial Intelligence*, vol. 2013, Article ID 435321, 7 pages, 2013. doi:10.1155/2013/435321

Bayesian Multivariate Growth Curve Models

Steward H. Huang

Abstract– Growth curve models have been widely studied and extensively applied in many areas because they are useful in situations when time (an important factor) is involved. Researchers have considered growth curves (mainly linear) in conjunction with different covariance structures for numerous applications. In this paper, the interest is to use some commonly used nonlinear growth curves to describe (in terms of time) each variable in a multivariate dataset in the presence of random error covariance structures with autocorrelation dependence. No similar attempts have been found in the literature because under this complex scenario, the models become too complicated for classical analysis without making a lot of compromising assumptions. It is shown that under a Bayesian formulation, by judicious choice of priors, one can obtain the full conditionals and this allows one to conveniently implement the Metropolis Hastings algorithm to sample/generate observations from the conditional (posterior) observations. This makes Bayesian approach a simpler but useful alternative to classical analysis. An intrauterine growth retardation in rats data set is used as an illustrative example for our model.

Keywords– Multivariate, Growth curve, Bayesian analysis, Autocorrelation

I. Introduction

The motivation for building multivariate models in this research is that we can study the effect of several variables acting simultaneously. This gives a closer resemblance to our intuition as well as better understanding about the relationship between the variables. When more variables are analyzed simultaneously, greater statistical power will be obtained and we gain easier visualization and interpretation of the data through graphical measures, such as scatter plots or higher dimensional plots (e.g. 3D plots). So our focus is also spontaneously shifted from individual or isolated factors to the relationships among several variables of interest in a data set.

Growth curve models, which are useful especially for studying growth behavior of short time series in economics, biology, medical research and epidemiological problems [1], [2], have a long history. Their initiation may be attributed to Potthoff and Roy [3], who introduced their formulation and then studied the growth curve problems. Then subsequently, Rao [4], Khatri [5], Geisser [6] and von Rosen [7]-[9] became the primary researchers in analyzing the growth curve models. However, it took nearly a decade before the Bayesian approach (including predictive problem from a Bayesian perspective) was applied to the analysis of growth curve models and different assumptions about covariance matrices were also made accordingly. Lindley and Smith [10] and Geisser [11] assumed that covariance matrices were known, Fearn [12] assumed that they were identity matrices with unknown variances. Barry [13] gave a different

treatment of the problem under Bayesian approach but also assumed identity matrix for covariances.

In this research, similar general multivariate growth problems are studied by assuming that the multivariate dependent variables (such as weight, height, etc.) can be described by some commonly used nonlinear growth curves in terms of the independent variable (time) with a certain correlation (dependence) relationship in the covariance matrix. So the multivariate growth curve models proposed in this paper will include nonlinear growth curves with autocorrelated errors in their covariance structures. The classical analysis for these types of realistic models becomes either too complicated to obtain analytical solutions or may require a lot of simplifying assumptions, thus becoming unrealistic. Bayesian analysis, including experts' opinions, can help us computationally to get to the estimates of the parameters for growth curve models and thus become more appealing, as well as important to researchers. No similar models which consider such complex scenarios are available in the literature.

The model formulation will be presented in the following first two section with the Gompertz growth curve as an example. Then in the last two sections, the applications of the models using a bivariate growth data set will be demonstrated. The simulation results will be discussed in the conclusion.

II. Model Formulation

Let's consider a single subject of n observations. Y_j , for $j = 1, \dots, n$, is a vector of p -variate correlated dependent variables. If we let $W = (w_1, \dots, w_n)$ be a vector of the independent variable time and $\Theta = (\theta_1, \dots, \theta_p)$, where θ_k , $k = 1, \dots, p$ is a vector of coefficients (parameters) for growth curves and q is the number of coefficients for the specific growth curve in that model (e.g., $q = 3$ in a Gompertz curve). Also let $f(W|\theta_k)$, $k = 1, \dots, p$ be the growth curve then our model can be defined as follows:

$Y = M + E$, where $E \sim N_p(0, \Omega)$, Ω is a $p \times p$ variance covariance matrix,

$$Y_{(p \times n)} = \begin{pmatrix} y'_1 \\ \vdots \\ y'_p \end{pmatrix}_{(p \times n)}, \text{ where } y_k = \begin{pmatrix} y_{1k} \\ \vdots \\ y_{nk} \end{pmatrix} \text{ for } k = 1, \dots, p,$$

and

$$M_{(p \times n)} = \begin{pmatrix} \mu'_1 \\ \vdots \\ \mu'_p \end{pmatrix}_{(p \times n)} = [f(W|\Theta)_{n \times p}]', \text{ where } \mu_k = \begin{pmatrix} \mu_{1k} \\ \vdots \\ \mu_{nk} \end{pmatrix}_{n \times 1} = f(W|\theta_k)$$

for $k = 1, \dots, p$.

This model considers a covariance structure between weight and length in that, under normal conditions, the lengthier the subject

grows, the weightier it becomes and vice versa. Assume that Y follows a $p \times n$ matrix normal distribution, which is actually a special case of the pn -variate multivariate normal distribution when the covariate matrix is separable. Then denote a pn -variate normal distribution with pn -dimensional mean μ and $pn \times pn$ covariance matrix Ω , the p.d.f. function will be as follows:

$$g(y|\mu, \Omega) = (2\pi)^{-\frac{np}{2}} |\Omega|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y - \mu)' \Omega^{-1} (y - \mu)\right\} \quad (1),$$

where

$$y = \underset{(pn \times 1)}{\text{vect}}(Y') = (y'_1, \dots, y'_p)',$$

$$\mu = \underset{(pn \times 1)}{\text{vect}}(M') = (\mu'_1, \dots, \mu'_p)',$$

in which the operator $\text{vect}(\cdot)$ stacks the columns of its matrix argument from left to right in a single vector. The separable matrix $\Omega = \Sigma \otimes \Phi$, where \otimes is the Kronecker product which multiplies every entry of its first matrix argument by its entire second matrix argument, can be written as:

$$\Sigma \otimes \Phi = \begin{pmatrix} \sigma_{11}\Phi & \dots & \sigma_{1p}\Phi \\ \vdots & & \vdots \\ \sigma_{p1}\Phi & \dots & \sigma_{pp}\Phi \end{pmatrix}.$$

Also we know that $\Omega^{-1} = (\Sigma \otimes \Phi)^{-1} = \Sigma^{-1} \otimes \Phi^{-1}$ and $|\Sigma \otimes \Phi|^{-\frac{1}{2}} = |\Sigma|^{-\frac{p}{2}} |\Phi|^{-\frac{n}{2}}$. Then we have

$$g(y|\mu, \Sigma, \Phi) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{p}{2}} |\Phi|^{-\frac{n}{2}} \cdot$$

$$\exp\left\{-\frac{1}{2}(y - \mu)' (\Sigma \otimes \Phi)^{-1} (y - \mu)\right\}.$$

Note that also with the matrix identity, we have

$$(y - \mu)'_{1 \times np} (\Sigma \otimes \Phi)^{-1}_{np \times np} (y - \mu)_{np \times 1} = \text{tr} \Sigma^{-1}_{p \times p} (Y - M)_{p \times n} \Phi^{-1}_{n \times n} (Y - M)'_{n \times p},$$

$$g(Y|M, \Sigma, \Phi) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{p}{2}} |\Phi|^{-\frac{n}{2}} \cdot$$

$$\exp\left\{-\frac{1}{2} \text{tr} \Sigma^{-1}_{p \times p} (Y - M)_{p \times n} \Phi^{-1}_{n \times n} (Y - M)'_{n \times p}\right\} \quad (2).$$

So Y is a random variable that follows a $p \times n$ matrix normal distribution and can be denoted as:

$$Y|M, \Sigma, \Phi \sim N_{p \times n}(M, \Sigma \otimes \Phi),$$

where (M, Σ, Φ) parameterize the above distribution with $Y \in \mathbb{R}^{p \times n}$, $M \in \mathbb{R}^{p \times n}$ and $\Sigma, \Phi > 0$ (Σ and Φ are commonly referred to as the within and between covariance matrices). Recall that M is a function of Θ and assume that Φ is a function of correlation coefficient ρ and that, for simplicity, Θ , Σ and Φ are independent and adopt vague prior distributions for (Θ, Φ, Σ) . Then we have $h(\Theta, \Phi, \Sigma) = h(\Theta)h(\Phi)h(\Sigma)$ and because Φ is a function of ρ , their prior distribution assumptions are as follows: $h(\Theta) \propto \text{constant}$, $\rho \propto (1 + \rho)^{\tilde{\alpha}-1} (1 - \rho)^{\tilde{\beta}-1}$ for $-1 < \rho < 1$ (i.e., $(1 + \rho)/2 \text{ Beta}(\tilde{\alpha}, \tilde{\beta})$, where $\tilde{\alpha}$ and $\tilde{\beta}$ can be chosen such that the mean $\tilde{\alpha}/(\tilde{\alpha} + \tilde{\beta})$ is consistent with the empirical value for ρ) and $h(\Sigma) \propto \frac{1}{|\Sigma|^{(p+1)/2}}$. So the prior distribution is $h(\Theta, \rho, \Sigma) \propto \frac{(1+\rho)^{\tilde{\alpha}-1} (1-\rho)^{\tilde{\beta}-1}}{|\Sigma|^{(p+1)/2}}$, and the joint posterior distribution of the parameters follows:

$$\pi(\Sigma, \Phi(\rho), \Theta|W, Y) = (2\pi)^{-\frac{np}{2}} |\Sigma|^{-\frac{n+p+1}{2}} |\Phi|^{-\frac{p}{2}} (1 + \rho)^{\tilde{\alpha}-1} \cdot$$

$$(1 - \rho)^{\tilde{\beta}-1} \exp\left\{-\frac{1}{2} \text{tr} \Sigma^{-1}_{p \times p} (Y - M)_{p \times n} \Phi^{-1}_{n \times n} (Y - M)'_{n \times p}\right\} \quad (3).$$

Let $G = (Y - M)\Phi^{-1}(Y - M)'$ then (3) becomes

$$\pi(\Sigma, \Phi(\rho), \Theta|W, Y) \propto \left[|\Sigma|^{-\frac{n+p+1}{2}} \exp\left\{-\frac{1}{2} \text{tr} \Sigma^{-1} G\right\}\right] \cdot |\Phi|^{-\frac{p}{2}} (1 + \rho)^{\tilde{\alpha}-1} (1 - \rho)^{\tilde{\beta}-1}.$$

This can be reduced to the joint distribution of Φ and Θ and become $\pi(\Phi(\rho), \Theta|W, Y)$ if we integrate out Σ .

The integration can be worked out by recognizing that if Σ^{-1}

follows a Wishart distribution [14] then it can be written as:

$$\pi(\Theta, \Phi(\rho)|W, Y) \propto \frac{|\Phi|^{-\frac{p}{2}} (1+\rho)^{\tilde{\alpha}-1} (1-\rho)^{\tilde{\beta}-1}}{|G|^{p/2}} \quad (4).$$

Assume an autocorrelation matrix for Φ with correlation coefficient ρ as follows:

$$\Phi = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix} \quad (5),$$

where ρ is the correlation coefficient. Then we can substitute the results that

$|\Phi| = (1 - \rho^2)^{n-1}$, into (4) and get the posterior function

$$\pi(\Theta, \Phi(\rho)|W, Y) \propto \frac{(1+\rho)^{\tilde{\alpha}-1} (1-\rho)^{\tilde{\beta}-1} (1-\rho)^{(n-1)-p(n-1)/2}}{|G|^{p/2}} \quad (6).$$

III. Gompertz Curve as An Illustrative Example

If we take the Gompertz curve as an illustrative example in fitting a bivariate data set which has weight and length as the variables and the following priors for

$$\Theta = \{\theta_1 \theta_2\} = \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \\ c_1 & c_2 \end{pmatrix}, \text{ where } a_1 \sim \text{Expon}\left(\frac{1}{\tilde{a}_1}\right), a_2 \sim \text{Expon}\left(\frac{1}{\tilde{a}_2}\right),$$

$$\text{Expon}\left(\frac{1}{\tilde{a}_2}\right),$$

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \tilde{b}_1 \\ \tilde{b}_2 \end{pmatrix}, \tilde{\Sigma}_b\right), c_1 \sim \text{Expon}\left(\frac{1}{\tilde{c}_1}\right) \text{ and } c_2 \sim \text{Expon}\left(\frac{1}{\tilde{c}_2}\right).$$

Let $\tilde{\Theta} = (\tilde{a}_1, \tilde{a}_2, \tilde{b}_1, \tilde{b}_2, \tilde{c}_1, \tilde{c}_2, \tilde{\Sigma}_b)$ be a set of empirical Bayes estimates of the coefficients which can be estimated through nonlinear least square regression method. MATLAB *nlinfit* function can be used to fit *nonlinear* Jenss, Gompertz and Richards curves and *polyfit* function to fit *polynomial* curves (MATLAB Help and [15]). Depending on the data, although sometimes we could get estimates of $\tilde{\Sigma}_b$, most of the time we have to assume them to be equal to some proper value for our Bayesian analysis. So the prior distributions are:

$$h(\Theta|\tilde{\Theta}) \propto \frac{1}{\tilde{a}_1 \tilde{a}_2 \tilde{c}_1 \tilde{c}_2 |\tilde{\Sigma}_b|^{1/2}} \cdot$$

$$\exp\left\{-\frac{1}{2} \begin{pmatrix} b_1 - \tilde{b}_1 \\ b_2 - \tilde{b}_2 \end{pmatrix}' \tilde{\Sigma}_b^{-1} \begin{pmatrix} b_1 - \tilde{b}_1 \\ b_2 - \tilde{b}_2 \end{pmatrix} - \left(\frac{a_1}{\tilde{a}_1} + \frac{a_2}{\tilde{a}_2} + \frac{c_1}{\tilde{c}_1} + \frac{c_2}{\tilde{c}_2}\right)\right\} \quad (7).$$

Let (7) be substituted into (6), then it becomes

$$\pi(\Theta, \Phi(\rho)|W, Y) \propto \frac{(1 + \rho)^{\tilde{\alpha}-1} (1-\rho)^{\tilde{\beta}-1} (1-\rho)^{(n-1)-p(n-1)/2}}{|(Y - M)\Phi^{-1}(Y - M)'|^{p/2}} \cdot$$

$$\exp\left\{-\left(\frac{a_1}{\tilde{a}_1} + \frac{c_1}{\tilde{c}_1} + \frac{a_2}{\tilde{a}_2} + \frac{c_2}{\tilde{c}_2}\right) - \frac{1}{2} \begin{pmatrix} b_1 - \tilde{b}_1 \\ b_2 - \tilde{b}_2 \end{pmatrix}' \tilde{\Sigma}_b^{-1} \begin{pmatrix} b_1 - \tilde{b}_1 \\ b_2 - \tilde{b}_2 \end{pmatrix}\right\} \quad (8).$$

Then we get the full conditionals of the parameters as follows:

$$\pi(\rho|\cdot) \propto \frac{(1+\rho)^{\tilde{\alpha}-1} (1-\rho)^{\tilde{\beta}-1} (1-\rho)^{(n-1)-p(n-1)/2}}{|(Y-M)\Phi^{-1}(Y-M)'|^{p/2}} \quad (9),$$

$$\pi(a_1|\cdot) \propto \frac{\exp\{-a_1/\tilde{a}_1\}}{|(Y-M)\Phi^{-1}(Y-M)'|^{p/2}} \quad (10),$$

$$\pi(a_2|\cdot) \propto \frac{\exp\{-a_2/\tilde{a}_2\}}{|(Y-M)\Phi^{-1}(Y-M)'|^{p/2}} \quad (11),$$

$$\pi(b_1|\cdot) \propto \frac{1}{|(Y-M)\Phi^{-1}(Y-M)'|^{p/2}} \cdot$$

$$\exp\left\{-\frac{1}{2} \begin{pmatrix} b_1 - \tilde{b}_1 \\ b_2 - \tilde{b}_2 \end{pmatrix}' \tilde{\Sigma}_b^{-1} \begin{pmatrix} b_1 - \tilde{b}_1 \\ b_2 - \tilde{b}_2 \end{pmatrix}\right\} \quad (12),$$

$$\pi(b_2|\cdot) \propto \frac{1}{|(Y-M)\Phi^{-1}(Y-M)'|^{p/2}} \cdot$$

$$\exp \left\{ -\frac{1}{2} \begin{pmatrix} b_1 - \tilde{b}_1 \\ b_2 - \tilde{b}_2 \end{pmatrix}' \tilde{\Sigma}_b^{-1} \begin{pmatrix} b_1 - \tilde{b}_1 \\ b_2 - \tilde{b}_2 \end{pmatrix} \right\} \quad (13),$$

$$\pi(c_1|\cdot) \propto \frac{\exp\{-c_1/\tilde{c}_1\}}{|(Y-M)\Phi^{-1}(Y-M)'\tilde{c}_1|^{n/2}} \quad (14),$$

$$\pi(c_2|\cdot) \propto \frac{\exp\{-c_2/\tilde{c}_2\}}{|(Y-M)\Phi^{-1}(Y-M)'\tilde{c}_2|^{n/2}} \quad (15).$$

Regarding the MH Algorithm:

Let's take the sampling of a_2 in Gompertz curve as an example. To define the algorithm, let $\varphi(a_2^{(old)}, a_2^{(new)})$ denote a source density for a candidate draw $a_2^{(new)}$ given the current value $a_2^{(old)}$ in the sampled sequence. The density $\varphi(a_2^{(old)}, a_2^{(new)})$ is referred to as the proposal or candidate generating density. Then, the MH algorithm is defined by two steps: a first step in which a proposal value is drawn from the candidate generating density and a second step in which the proposal value is accepted as the next iterate in the Markov chain according to the probability:

$$\alpha(a_2^{(old)}, a_2^{(new)}) = \min \left\{ \frac{\pi_2(a_2^{(new)})\varphi(a_2^{(old)}, a_2^{(new)})}{\pi_2(a_2^{(old)})\varphi(a_2^{(new)}, a_2^{(old)})}, 1 \right\},$$

if $\pi_2(a_2^{(old)})\varphi(a_2^{(old)}, a_2^{(new)}) > 0$ (otherwise $\alpha(a_2^{(old)}, a_2^{(new)}) = 1$).

If the proposal value is rejected, then the next sampled value is taken to be the current value. Let's follow this recursive procedure:

Metropolis-Hastings Algorithm:

- 1) Specify an initial value $a_2^{(0)}$.
- 2) Repeat for $j = 1, 2, \dots, M$:
 - a) Propose $a_2^{(new)} \sim \varphi(a_2^{(j)}, \cdot)$, and
 - b) Let $a_2^{(j+1)} = a_2^{(new)}$ if $U(0, 1) \leq \alpha(a_2^{(j)}, a_2^{(new)})$ otherwise $a_2^{(j+1)} = a_2^{(j)}$.
- 3) Return the values $a_2^{(1)}, a_2^{(2)}, \dots, a_2^{(M)}$.

Then follow the above Metropolis-Hastings Algorithm in taking samples of Θ and ρ by using (9)-(15) through the following steps:

- 1) Set $j = 0$ and select a set of initial parameter values for $\Theta^{(0)}$, $B^{(0)}$ and $\rho^{(0)}$.
- 2) Sample $\rho^{(j+1)}$ from (9) (using MH algorithm).
- 3) Sample $\Theta^{(j+1)}$ from (10)-(15) (using MH algorithm).
- 4) Replace $\rho^{(j)}$ by $\rho^{(j+1)}$, $\Theta^{(j)}$ by $\Theta^{(j+1)}$ and $B^{(j)}$ by $B^{(j+1)}$.
- 5) Set $j = j + 1$ and repeat steps 2 through 4.

Drop the initial burn-in sets and retain the rest of the data for marginal distribution analysis. This analysis includes highest density regions for the estimated parameters. In addition to this analysis of parameters, we can also generate 90% Credible Intervals (CIs or HDR's, Highest Density Regions) for the best-fit growth curve under this Bayesian formulation by using the 5% and 95% percentiles of y calculated by substituting the M samples of Θ at a given w_j .

IV. Example: Using Intrauterine Growth Retardation in Rats Data

An intrauterine growth retarded rats data set [16] in this section as an example to demonstrate how to apply our approach to Bayesian analysis of multivariate growth curve model in a bivariate data setting (weight and length). In their experiment, in [16], they chose fifty female rats that were mated overnight with ten adult males and then divided the pregnant female rats into three groups: control group, intrauterine growth control group

and sham-operated group. They then measured body weight, body length, and other facial characteristics of the rats that were in those three groups, respectively, every four days for twenty days. For illustrative purposes and for simplifying our analysis, the control group has been chosen and only use the bivariate body weight and body length in our growth curve model. The data set for rats growth is in Table 1. Four classic growth curve models explicitly for this specific example are presented below:

- 1) Jenss growth curve: $f(w) = a + bw - \exp(c + dw)$.
- 2) Gompertz growth curve: $f(w) = a \exp[-\exp(b + cw)]$.
- 3) Richards growth curve: $f(w) = a \{1 + b \exp[c(d - w)]\}^{-1/b}$.
- 4) Polynomial growth curve: $f(w) = a + bw + cw^2 + dw^3$.

In that data set, assume that $\tilde{\Sigma}_b$ is equal to $s^2 \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}$ for Gompertz curve. Similar assumption has been made for the other growth curves for comparison. s^2 can be quite small if prior knowledge is reliable. The value assumed here would allow some moderate correlation relationship between the covariates length and weight. The results of Bayesian estimates are displayed in Table 2. Using BIC, in conjunction with the graphs and CIs, it seems natural to say that the Cubic growth curve is the model of selection for this specific bivariate intrauterine growth retarded rats data.

Regarding diagnostic testing for the model, careful consideration has been given to the useful approach presented by Franses [17] for residual autocorrelation in growth curve models. However, it's natural to concur with his own conclusion that there are obvious drawbacks in applying his method to small sample sizes, other growth curve models and to various model selection criteria as well. Needless to mention that this research is dealing with a multivariate scenario.

V. Conclusion

The 90% Credible Intervals, the fitted curves, and the estimates of the model parameters for the four growth curves in Figures 1-4 and Table 1-2 have been presented in this paper. There, we can see that for weight versus time (Figure 1) and length vs. time (Figure 2), the 90% CI of Cubic curve is the narrowest among all four curves when time is small but diverges like a funnel shape as time increases to approximately more than 15 days; for weight vs. time in Figure 1, Jenss and Gompertz curves both have relatively narrow 90% CIs, whereas for length vs. time in Figure 2, Jenss curve has smaller 90% CI. In Figures 3-4, we observed that on the one hand, the data display a positive trend that as length increases, the rate of change in weight also increases; on the other hand, when weight increases, the rate of change in length decreases.

Although all four curves fit the data reasonably well, the Cubic curve is apparently the best fit curve among them. In addition, as time increases approximately before the fifteenth day, the rate of change in length and in weight both increase as weight and length increase. It's obvious that Cubic curve appears to be the best fit curve for the given data. Besides, BIC is useful criteria in model selection because the smaller value it is, the better curve fitting it will be, and this is consistent with our observations in those Figures.

In summary, the Bayesian multivariate growth curve models in this study provide a formulation for generating Bayesian estimates as well as describing the dependence relationship

between variables with a certain autocorrelation relationship under consideration. Further research topic may include better diagnostic testing methods for more growth curve models as well as smaller multivariate sample size data using various selection criteria.

References

[1] J. E. Grizzle and D. M. Allen, "Analysis of growth curves and responses," *Biometrics*, 25, 1969, pp. 357-381.

[2] J. C. Lee and S. Geisser, "Growth curve prediction," *Sankhya A*, 34, 1975, pp. 393-412.

[3] R. F. Potthoff and S. N. Roy, "A generalized multivariate analysis of variance model useful especially for growth curve problems," *Biometrika*, 51, 1964, pp. 313-326.

[4] C. R. Rao, "The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves," *Biometrika*, 52, 1965, pp. 447-458.

[5] C. G. Khatri, "A note on a MANOVA model applied to problems in growth curves," *Annals of Institute of Statistical Mathematics*, 18, 1966, pp. 75-86.

[6] S. Geisser, "Bayesian analysis of growth curves," *Sankhya A*, 32, 1970, pp. 53-64.

[7] D. von Rosen, "Maximum likelihood estimates in multivariate linear normal model," *Journal of Multivariate Analysis*, 31, 1989, pp. 187-200.

[8] D. von Rosen, "Moments for a multivariate linear normal models with application to growth curve model," *Journal of Multivariate Analysis*, 35, 1990, pp. 243-259.

[9] D. von Rosen, "The growth curve model: a review," *Communication in Statistics: Theory and Methods*, 20, 1991, pp. 2791-2882.

[10] D. V. Lindley and A. F. M. Smith, "Bayes estimates for the linear model," *Journal of the Royal Statistical Society B*, 34, 1972, pp. 1-41.

[11] S. Geisser, "Growth curve analysis," *Handbook of Statistics*. (P.R. Krishnaiah ed.), North-Holland, Amsterdam, 1980, pp. 89-115.

[12] T. Fearn, "A Bayesian approach to growth curves," *Biometrika*, 62, 1975, pp. 89-100.

[13] D. Barry, "A Bayesian model for growth curves analysis," *Biometrics*, 51, 1995, pp. 639-655.

[14] S. J. Press, "*The Subjective and Objective Bayesian Statistics Principles, Models and Applications* (2nd ed.)," John Wiley & Sons, Sections 12.3.5-6, 2002.

[15] G. A. F. Seber and C. J. Wild, "*Nonlinear Regression*," John Wiley & Sons, 1989.

[16] E. E. Oyhenart, B. Orden, M. C. Fucini, M. C. Muiçee, H. M. Pucciarelli, "Sexual dimorphism and postnatal growth of intrauterine growth retarded rats," *Growth, Development and Aging*. 67(2), 2003, pp. 73-83.

[17] P. H. Franses, "Testing for residual autocorrelation in growth curve models," *Technological Forecasting and Social Change*, 69, 2002, p. 195-2.

Dr. Steward H. Huang is a faculty member of the University of Arkansas - Fort Smith, United States. Dr. Huang may be reached at through his email at stewardhuang@gmail.com.

Age (Days)	Weight (g)	Length (mm)
1	6.6	54.5
5	10.4	65.6
9	16.3	77.2
13	23.2	87.5
17	28.6	94.6
21	38.4	110.4

TABLE 1
Rats Growth Data

(Units for ages: Days; Weight: Grams and Length: mm)

Estimates	Jensz		Gompertz		Richards		Cubic Polynomial	
	Length vs. Time	Weight vs. Time	Length vs. Time	Weight vs. Time	Length vs. Time	Weight vs. Time	Length vs. Time	Weight vs. Time
a	32.290	21.536	273.221	146.221	966.688	747.626	50.126	3.323
b	2.674	1.573	0.301	1.167	-0.744	-0.399	3.998	1.021
c	0.215	2.908	0.027	0.041	0.005	0.02	-0.162	0.012
d	-0.561	0	-	-	37.438	76.668	0.005	0.001
β	-	-0.336	-	-0.361	-	-0.36	-	-0.369
BIC	-	7.73	-	5.74	-	3.59	-	3.57

TABLE 2
Bayesian Estimates of Parameters and BIC

Note: Take the numbers in the two columns under Gompertz as example: they are the estimates of the parameters (coefficients) of the bivariate growth curves (for length and weight, respectively), where

$$Length(w) = 275.231 \exp[-\exp(0.501 + 0.027w)],$$

$$Weight(w) = 146.321 \exp[-\exp(1.167 + 0.041w)].$$

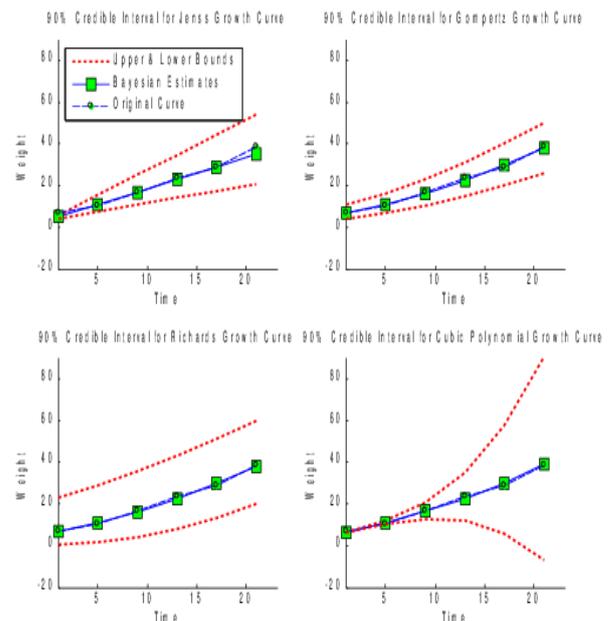


Figure 1 Weight vs. Time Credible Intervals for the Four Different Growth Curves

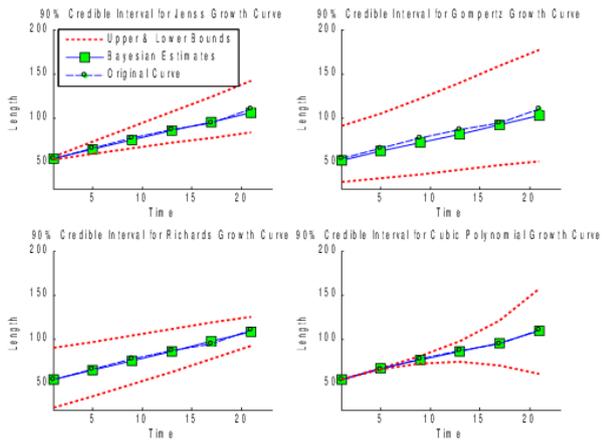


Figure 2 Length vs. Time Credible Intervals for the Four Different Growth Curves

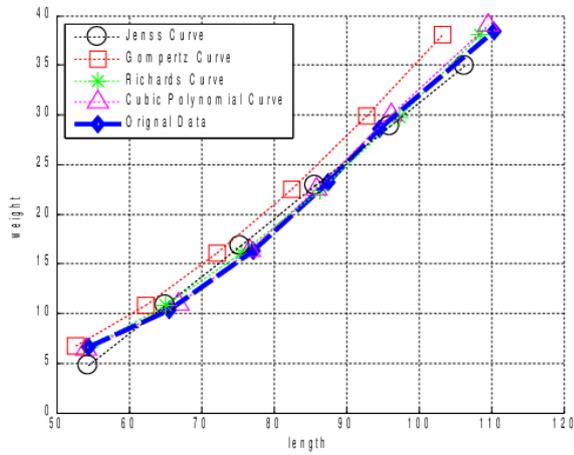


Figure 3 Weight vs. Length for the Four Different Growth Curves

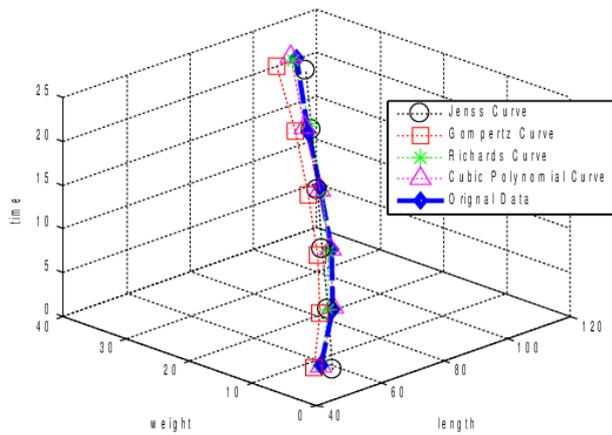


Figure 4 Three Dimensional Plot (Time, Weight and Length)

Effects of Dry Friction on Linear Electromechanical Actuators: A New Prognostic Method based on Simulated Annealing Algorithm

Matteo D. L. Dalla Vedova, Paolo Maggiore and Lorenzo Pace

Abstract—Several approaches can be employed in prognostics, to detect incipient failures of primary flight command electromechanical actuators (EMA), caused by progressive wear. The development of a prognostic algorithm capable of identifying the precursors of an electromechanical actuator failure is beneficial for the anticipation of the incoming failure: a correct interpretation of the failure degradation pattern, in fact, can trig an early alert of the maintenance crew, who can properly schedule the servomechanism replacement. The research presented in this paper proposes a prognostic technique, based on approaches derived from optimization methods, able to identify symptoms of an EMA degradation before the actual exhibition of the anomalous behavior; in this case friction failures are considered. An experimental test bench was developed: results show that the method exhibit adequate robustness and a high degree of confidence in the ability to early identify an eventual fault, minimizing the risk of false alarms or not annunciated failures.

Keywords— Dry Friction, Electromechanical Actuator, Prognostics, Simulated Annealing Algorithm.

I. INTRODUCTION

ACTUATORS are devices capable of operate conversion of mechanical, electrical, hydraulic, or pneumatic power into mechanical power. In aircraft, actuators are commonly used for flight control surfaces and various utility systems. Flight control systems are considered flight critical and, although highly redundant, must meet reliability requirements of less than one catastrophic failure per 10^5 flight hours for the F-18 strike fighter and one per 18×10^6 flight hours for F-35AB [1]. Unanticipated and extreme operating scenarios are a major cause of unscheduled maintenance events, which may result into serious operational issues in terms of safety, mission completion, and cost.

P. Maggiore is with the Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino, Corso Duca degli Abruzzi, 24 - 10129 - Torino, ITALY. (e-mail: paolo.maggiore@polito.it).

L. Pace is with the Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino, Corso Duca degli Abruzzi, 24 - 10129 - Torino, ITALY. (e-mail: lorenzo.pace@polito.it).

M. D. L. Dalla Vedova is with the Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino, Corso Duca degli Abruzzi, 24 - 10129 - Torino, ITALY. (corresponding author to provide phone: +390110906850; e-mail: matteo.dallavedova@polito.it).

Typically, when a monitor registers a fault, there is no information regarding the real cause and effect relationship between the failure mode and failure itself. All that is known is that a failure has occurred. Therefore, the identified need is for a robust health management solution capable of accurate and reliable early fault detection and failure prediction, covering multiple failure modes for flight control actuators (PHM, Prognosis and Health Management system) [2]. PHM is easier to implement on the electric actuators since no additional sensors are required, as the same sensors used to the control scheme and system monitors are also used in many PHM algorithms [2]. Therefore, electro-mechanical actuation systems are going to be considered in this study, "All-electric-aircraft" perspective.

Enormous economic (maintenance and logistics) benefit is expected with the advance of the state of fault detection to failure prognosis for actuator systems, as high Can Not Duplicate (CND - inability to replicate field failures during lower level maintenance assessment) rates still plague many aircrafts. From collected field analyses, CND failures can make up more than 85% of all observed field failures in avionics and account for more than 90% of all maintenance costs. These statistics can be attributed to a limited understanding of root cause failure characteristics of complex systems, inappropriate means of diagnosing the condition of the system, and the inability to duplicate the field conditions in the lower level test environment [3]. Economically speaking, Impact Technologies estimated that CND occurrences result in about a \$30M/yr incurred cost for one particular aircraft. Moreover, their study showed that a \$7M investment to develop a complete PHM solution would produce a ROI (Return On Investment) of 4 to 1 for a 5-year period considering just a 20% CND reduction. In other words, this means saving \$30M in 5 years by spending just \$7M initially. Moreover, in case the technology could produce a further CND cost reduction (e.g. 40%), the ROI would be of almost 10 to 1 in 5 years [2].

Since the prognostic activities typically involve systems having a complex non-linear multidisciplinary nature, the failure detection/evaluation strategies proposed in literature are various and extremely different each other.

For instance, during these years have been proposed model-based techniques based upon the direct comparison between real and monitoring system [4], on the spectral analysis of well-defined system behaviors (typically performed by Fast Fourier Transform FFT) [5], on appropriate combinations of the first two methods [6] or on Artificial Neural Networks [7]. The present work reports the first results of a wider research activity focused on the diagnosis model-based approach and, in particular, on the parametric estimation task, having as a primary objective the design of a modern and fast damage estimator routine for a simple (but real) electromechanical actuation system, in order to prove its accuracy and reliability.

This is done through the following steps:

- 1) define the optimization algorithm to be used for the parameter estimation task;
- 2) set up a real actuation system meeting as much as possible the aeronautical requirements and being capable of responding to different types of signals (step, sinusoidal, random sequence, ramp) as well as recording significant data (velocities, position, current);
- 3) build and validate a dedicated Matlab-Simulink numerical model of the considered actuation system (it must be noted that the aforesaid model, having to be run several times in the process of identification and evaluation of faults, must represent a compromise between the most reduced calculation effort and a satisfying representativeness of the actual EMA behaviors);
- 4) simulate different fault conditions on the real actuation system (without damaging it);
- 5) test the damage estimator to evaluate its speed and reliability on the simulated faulty response of the system.

In particular, this paper shows the results obtained applying the proposed fault detection/evaluation method in case of EMA subjected to dry friction phenomena.

II. AIMS OF WORK

The aims of this work are:

- 1) the proposal of a numerical algorithm able to perform the simulations of the dynamic behavior of a typical electromechanical servomechanism evaluating the global effects due to dry frictions acting on actuation system (e.g. rotor bearings, gear reducer, ball screw, gaskets);
- 2) the proposal of an innovative fault detection/evaluation method, based on techniques derived from optimization methods, able to detect the EMA failure precursors and evaluate the corresponding failure entity.

To assess the robustness of the proposed techniques, an appropriate experimental test environment was developed; the effects due to the said failures on the EMA behavior have been evaluated by means of several tests (related to different values of dry friction). These results have been compared with the ones provided by a corresponding numerical simulation model, in order to evaluate the differences and, by a proper algorithm based on simulating annealing, timely identify the failures and evaluate their magnitudes.

III. OPTIMIZATION ALGORITHM

Different optimization techniques are commonly used also for model parameter estimation tasks. They can be divided into two main groups: deterministic (direct or indirect) and probabilistic (stochastic, as Monte Carlo method, simulated annealing and genetic algorithms). Most methods, are local minima search algorithms and often do not find the global solution. As a result, they are highly dependent on good initial guesses. While this is a viable solution in an off-line scenario, where initial guesses can be reiterated, these approaches are not suitable for an on-line automated identification process because a good initial guess for one data set may not be for the next identification. These approaches would not be robust and may provide a false indication of parameter changes in an on-line system. Alternatively, global search methods, such as genetic algorithms and simulated annealing, are much better options for on-line model identification. However, similar to simplex methods, genetic algorithms do not always find the global minima [8]. Simulated annealing methods are more effective at finding the global minima, but at the cost of many more iterations [2]. The simulated annealing method originates, as the name suggests, from the study of thermal properties of solids (Metropolis et al. 1953 [9]). The Metropolis procedure was then an exact copy of the physical process which could be used to simulate a collection of atoms in thermodynamic equilibrium at a given temperature. In fact, the abstraction of this method in order to allow arbitrary problem spaces is straightforward. There is a significant correlation between the terminology of thermodynamic annealing process (the behavior of systems with many degrees of freedom in thermal equilibrium at a finite temperature) and combinatorial optimization (finding global minimum of a given function based on many parameters). A detailed analogy of annealing in solids provides frame work for optimization. As reported in [12], Table 1 shows the key terms which are related with thermodynamic annealing and its association with optimization process.

Table 1: Association between thermodynamic simulation and combinatorial optimization

Thermodynamic Annealing	Combinatorial Optimization
System State	Feasible Solutions
Energy of a State	Cost of Solution ¹
Change of state	Neighbor solution ²
Temperature	Control parameter ³
Minimum Energy	Minimum Cost

¹ The cost of a solution represents the corresponding objective function value (i.e. the function that the optimization algorithm attempts to minimize in order to identify the optimal solution).

² A new system solution calculated by the optimization algorithm and evaluated, with respect to the previous one, using the said cost functions.

³ The system parameters iteratively modified by the optimization process so as to minimize its objective function.

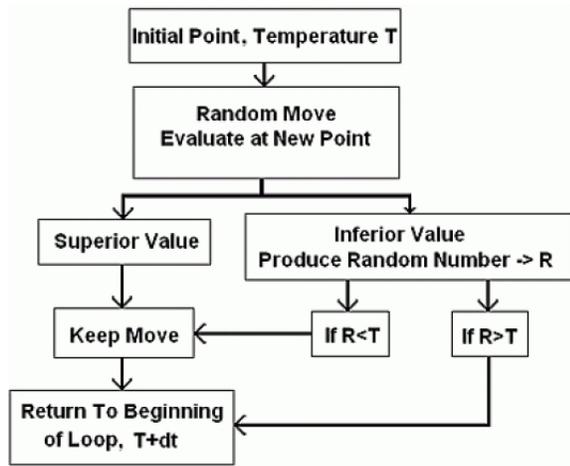


Fig. 1 Operating Logic of Simulated Annealing Method

At a given temperature and energy, a new nearby geometry $i + 1$ is generated in each iteration as a random displacement from the current geometry i . The energy of the resulting new geometry is then computed and the **energetic difference** E is determined with respect to preceding energy as:

$$\Delta E = E_{i+1} - E_i \quad (1)$$

The probability that this new geometry will be accepted is:

$$P(\Delta E) = \begin{cases} e^{-\Delta E/(k_b T)} & \text{if } \Delta E > 0, \\ 1 & \text{if } \Delta E < 0 \end{cases} \quad (2)$$

This means that, if the new nearby geometry has a lower energy level (successful iteration), the transition is accepted. Otherwise (unsuccessful iteration), a uniformly distributed random number more or equal than 0 and less than 1 is drawn and the step will only be accepted in the simulation if it is less or equal the Boltzmann probability factor, i.e. $r \leq P(E)$. After a certain number of steps at the same temperature T , the latter is decreased following the specified cooling schedule. It is worth noticing that the temperature does not take part directly to the optimization itself, but it acts merely as an exploration parameter. As at high temperatures T the factor $P(E)$ is very close to 1, most likely many up-hill steps are accepted, even if they are unsuccessful.

In this way, a wide exploration of the search space can be performed (this is the main feature of this algorithm). Subsequently, as the temperature falls off, the search is confined in a more limited space since Boltzmann factor $P(E)$ collapses to very low values, thus decreasing the acceptance probability in case of $E > 0$ (the algorithm becomes more selective). Finally, the global optimum should be found as soon as the temperature reaches its minimum value but, in practice, reannealing is performed, raising the temperature after a certain number of new points have been accepted so that the search starts again at the higher temperature. Basically, reannealing avoids getting caught at local minima.

In particular, the authors performed the abovementioned optimization analysis by means of the MATLAB Optimization Tool; in this case, the main annealing parameters are the following:

Annealing function: specifies the function used to generate new points for the next iteration:

- Fast annealing takes random steps, with size proportional to temperature
- Boltzmann annealing takes random steps, with size proportional to the square root of temperature using multivariate normal distribution.

Reannealing interval: is the number of points to accept before reannealing. Reannealing sets the annealing parameters to lower values than the iteration number, thus raising the temperature in each dimension.

The annealing parameters depend on the values of the estimated gradients of the objective function in each dimension. The basic formula is:

$$k_i = \log\left(\frac{T_0 \max(s_j)}{T_i s_j}\right) \quad (3)$$

Where:

k_i = annealing parameter for component i .

T_0 = initial temperature of component i .

T_i = current temperature of component i .

s_j = gradient of objective in direction i times difference of bounds in direction i .

Temperature update function: specifies how the temperature will be decreased.

Initial temperature: is the temperature at the beginning of the run.

The **acceptance criterion** evaluates the change in function values between the current point and new point to determine whether the new point is accepted or not, according to the abovementioned statistical mechanics criteria (i.e. using the Boltzmann probability density distribution).

Specifically MATLAB uses the following function:

$$P(\Delta E) = \frac{1}{1 + e^{\Delta E/T}} \quad (4)$$

which ranges from 0 to 0.5 and therefore differs from (2).

IV. ACTUATION SYSTEM

Until a few years ago, the actuators mainly used in aeronautical applications were generally hydraulic and precisely hydro-mechanical or, more recently, electrohydraulic. This kind of actuator, because of its great accuracy, high specific power and very high reliability, is often equipped on current aircrafts, even if on more modern airliners electro-hydrostatic actuators (EHA) or electro-mechanical actuators (EMA) are installed. Especially in the last years, the trend towards the all-electric aircrafts brought to an extensive application of novel optimized electrical actuators, such as the electromechanical ones (EMA).

To justify the fervent scientific activity in this field and the great interest shown by the aeronautical world, it must be noticed that, compared to the electrohydraulic actuations, the EMAs offer many advantages: overall weight is reduced, maintenance is simplified and hydraulic fluids, which is often contaminated, flammable or polluting, can be eliminated.

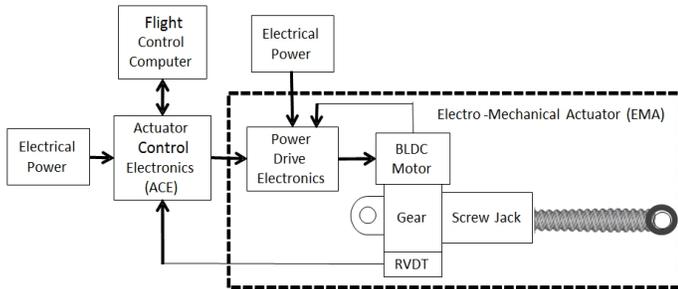


Fig. 2 Electromechanical Actuator Scheme

As shown in Fig. 2, a typical electromechanical actuator used in a primary flight control is composed by:

- 1) an **actuator control electronics** (ACE) that closes the feedback loop, by comparing the commanded position (FBW) with the actual one, elaborates the corrective actions and generates the reference current I_{ref} ;
- 2) a Power Drive Electronics (PDE) that regulates the three-phase electrical power;
- 3) an electrical motor, often BLDC (BrushLess Direct Current) type;
- 4) a gear reducer having the function to decrease the motor angular speed (RPM) and increase its torque to desired values;
- 5) a system that transforms rotary motion into linear motion: ball screws or roller screws are usually preferred to acme screws because, having a higher efficiency, they can perform the conversion with lower friction;
- 6) a network of sensors used to close the feedback rings (current, angular speed and position) that control the whole actuation system (reported in Fig. 2, as RVDT).

In order to evaluate the behavior of the proposed prognostic method in case of EMA progressive failures, a proper experimental test-bench has been conceived.

After a tradeoff among available actuators, controllers and power supplies, the following components have been chosen to compose the case study shown in Fig. 3:

- MecVel ALI-2 (version M01) actuator, powered by a brushed DC (BDC) electrical motor and equipped with 24 VDC brake and encoder (Fig. 4);
- RoboteQ AX1500 controller (with encoder module);
- Acopian unregulated power supply (220 AC - 24 VDC, 23 A);
- RS-232 to USB converter.

Subsequently to a proper stage of setup and calibration of the EMA control logic (selection of the proper PID gains and anti-windup filters), the actuation system was fully ready to operate.



Fig. 3 Complete EMA actuation system

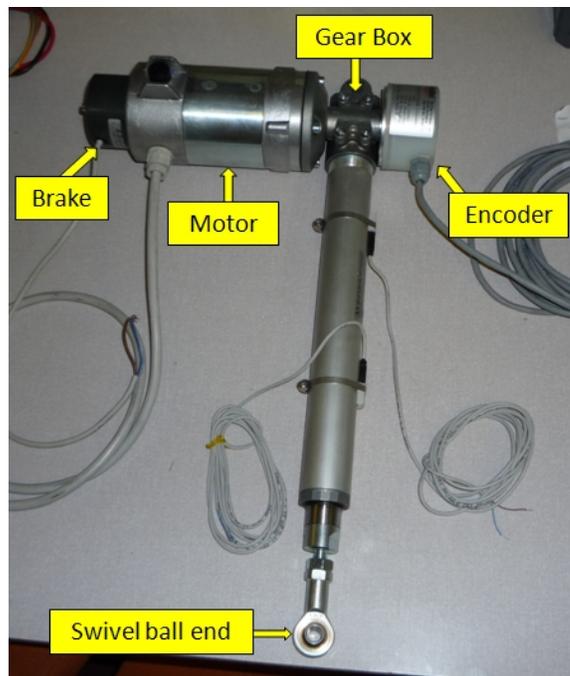


Fig. 4 Considered EMA actuator

The abovementioned controller logic closes the control loops feeding the EM actuator with various type of input meaningful for the parameter estimation process (sinusoidal with/without linear frequency sweep, ramp, step and external commands, all of them both in open and closed loop). Every significant datum (RPM, rod position, controller current, motor power level, PID actions) could have been recorded and exported to Microsoft Excel or even to Matlab.

V. EMA NUMERICAL MODEL

As previously reported, the subsequent step was to build an adequate Simulink model of the actuation system to be used as core of the damage estimator thus making it capable of recognizing the most representative actuator's failure modes according to some faulty experimental data achieved by the aforementioned software.

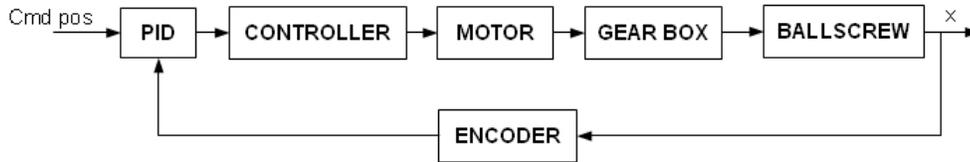


Fig. 5 Conceptual model's scheme

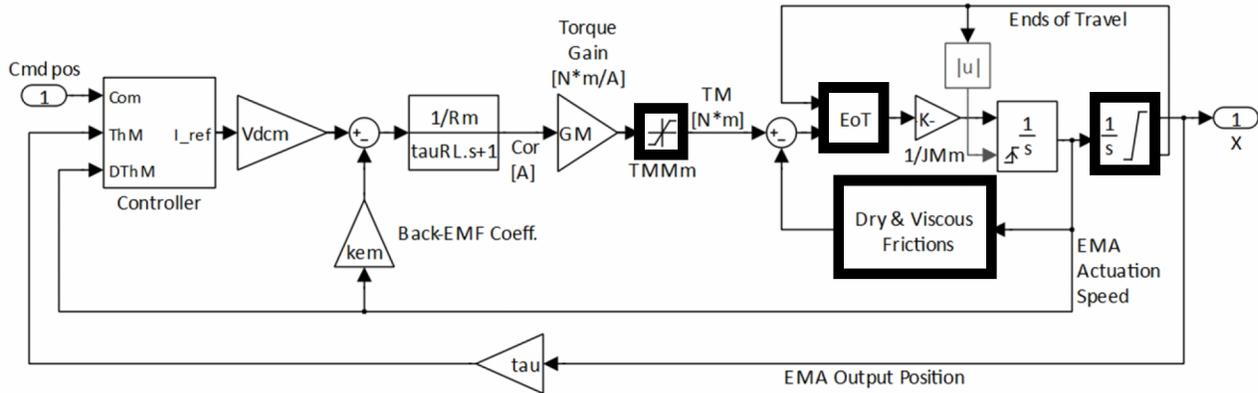


Fig. 6 Block Diagram of EMA numerical model: the blocks that implement the nonlinearities considered (saturation of the motor torque, friction phenomena and ends-of-travels acting on the final ballscrew actuator) are highlighted in the diagram by bold border.

In order to build an efficient model, two important (and often antithetical) aspects must be considered: the execution speed of the algorithm and the level of accuracy of the simulated results (with respect to the real ones).

In the present work, a parameter estimation task is involved (as shown in previous sections) meaning that the numerical model will go through an optimization problem and thus the speed aspect must be privileged. The proposed numerical model is composed of six blocks representing the different, physical or functional, components of the actual EMA. Hence, the Simulink model includes following blocks:

- PID Control Logic (i.e. PID controller with saturated output and anti-windup);
- Controller (simulating the RoboteQ AX1500 controller behaviors);
- Motor (simplified electro-magneto-mechanical model of the considered DC motor);
- Gear box;
- Ball screw;
- Encoder.

As shown in [7], every block has been modelled starting from its basic electromechanical equations, but since the objective is to achieve a model capable of recognizing also some significant actuator failure modes (e.g. dry friction), it was decided to model in a suitably simplified way the electromagnetic aspects and focus instead on mechanical ones.

The considered numerical model is developed from the monitoring model conceived by the authors for an EMA model-based prognostic application [4].

The electro-magneto-mechanical dynamics of the BDC motor is simulated by means of a classic resistive-inductive (RL) numerical model.

In particular, it is a 1st order linear model capable of calculating the moving torque TM as a function of the motor torque gain GM , of its power supply voltage ($V_{dcm} * I_{ref}$), of the back *emf*, of the dynamic characteristics of the RL circuit and of the saturation of magnetic induction flux.

The dynamics of the mechanical actuation system (rotor of BCD motor, gear box and ball screw) is represented by a simplified 1 degree-of-freedom system (obtained assuming an ideal rigid transmission without elastic deformations or backlashes). According to [6], it is modelled by means of a 2nd order non-linear numerical model able to simulate the EMA behavior taking into account the global effects due to inertia, viscous damping, ball screw ends-of-travel and dry frictions.

The dry friction torques acting on the actuation system are simulated by a numerical algorithm which implements the classical Coulomb's model; in particular, the proposed algorithm has been developed by means of a lumped parameter model based on the Karnopp friction model [10] and suitably modified as shown in [11].

VI. PROPOSED PROGNOSTIC ALGORITHM

The outlined nonlinear third-order model can simulate the system response, taking into account the effects due to Coulomb friction, being then potentially able to reproduce seizure due to ball return jamming or bearing binding/sticking.

Subsequently, its execution speed was tested in order to verify its suitability for optimization purposes. It must be noted that, despite being a relatively simplified numerical model, it shows a good accuracy, guaranteeing a satisfying correspondence with the experimental data (as reported in the following sections). The proposed prognostic algorithm performs the failure detection/evaluation by means of an optimization process implemented by means of simulated annealing algorithm; this process aims to minimize the value of appropriate objective functions (typically related to the magnitude of the error $E(t)$ calculated comparing together experimental and numerical data) by acting on well-defined parameters of the numerical model. In particular, by means of simulated annealing algorithm, the optimization process modifies the parameter CSJ, representative of the dry frictions acting on the EMA numerical model, up to identify its value that minimizes the abovementioned objective functions. It is clear that, in this case, the objective function of the optimization problem is the error generated, for a well-defined command input (*Cmd pos*), between the experimental data and the corresponding model output. Before verifying the actual ability of the proposed prognostic method to identify and evaluate friction precursors, the calibration of the numerical model parameters has been performed. The ideal values of these parameters have been identified by comparing the dynamic response of the real system in nominal conditions (NC: e.g. nominal dry friction level and no other failures) with that generated by the numerical model, then, identifying the corresponding objective function (E_{int}) and, at last, applying the proposed optimization process to the above parameters.

For instance, in Fig. 7 and 8 the experimental response of the EMA test bench is compared with the corresponding dynamic behaviors of the numerical model, putting clearly in evidence the best match that occurs (between experimental and simulated data) following of this calibration.

The model thus conceived and calibrated in NC was then used to estimate the global amount of the dry friction torques acting on the real EMA. The dynamic response of the real EM actuation system (subjected to a well-defined system of frictional actions) is compared with that produced by the simulation model and, by means of the abovementioned optimization method, it is calculated the value of the parameter CSJ that minimizes the error between real and simulated.

It should be noted that this parameter, dimensionally expressed in Newton-metre (Nm), is a global coefficient representing the equivalent static frictional torque acting on the whole EMA. The Simulated Annealing method used by the proposed prognostic routine to perform the fault estimation is implemented by means of **Matlab Optimization Tool**.

It must be noted that these optimizations have been carried out in condition of unloaded actuator since, within an operational scenario, these kinds of tests could be performed on the ground, without any aerodynamic loads, but rather just with the control surface weight, which is usually negligible compared to the actuator's capabilities.

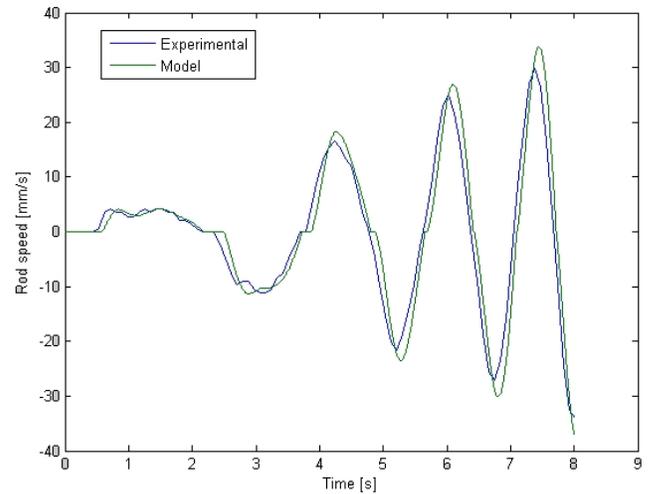


Fig. 7 EMA actuation speed before optimization

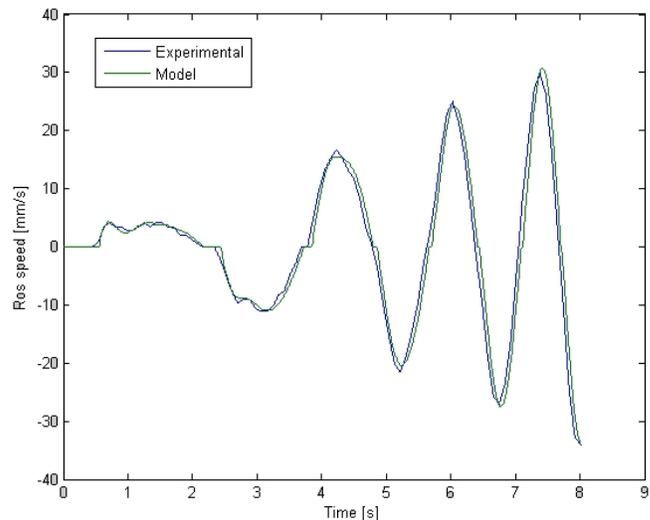


Fig. 8 EMA actuation speed after optimization

The problem of what type of signal should have been used to test the optimization algorithm has not a precise solution and depends strongly by the system's application. In the case here examined, a sinusoidal linear frequency sweep wave was chosen as standard input position signal for the parameter estimation process. In fact, such a signal allows testing, at one time, a wide range of system response frequencies.

For instance, in the low frequency range the stick-slip motion could be highlighted, enabling the optimization algorithm to finely tune the friction coefficients of the model and, at the same time, adapt the other parameters according also to the high frequency range, representing more significantly the system dynamic response. A simple step or ramp response could not comply with this necessity.

In order to obtain accurate results and assure a quick algorithm convergence, the static friction coefficient CSJ (varying during the optimization process in order to minimize the error between experimental data and corresponding numerical simulations) has been limited between a lower and an upper bound (respectively LB and UB).

In particular, CSJ can assume values from 0.01 Nm to 1.5 Nm which represent a quite large band given that its initial value is assumed equal to 0.12 Nm (NC dry frictional torque provided by motor datasheet) and the corresponding peak torque of the motor is worth 1.48 Nm.

In order to test the performance of the proposed method, different experimental tests have been conducted (with different time-history input and different levels of failure) that were then used as input to the optimization process performing the failure analysis.

For instance, the following section shows the results gained by the authors in case of experimental system affected by a friction torque of 0.779 Nm⁴. In this case, the considered input is a position command evolving like a sinusoidal linear frequency sweep wave.

Figures 9 and 10, comparing the dynamic response of the actual EMA with the corresponding numerical simulation before and after the optimization, allow evaluating immediately the effects that this optimization produces on the behavior of the simulated system. As shown in Fig. 9, before optimization the numerical model is not able to simulate the dynamic response of the real "faulty" system (i.e. affected by the fault in question) but, provided that the calibration of the model parameters is correct, approximates the real "healthy" system (i.e. in nominal conditions). The difference between the real position of the test case and the corresponding non-optimized numerical model is shown in Fig. 11 (blue curve).

Fig. 10 compares the trend of the instantaneous position of the experimental "faulty" system (blue line) with that of the optimized numerical model (green line). Compared to Fig. 9, Fig. 10 puts in evidence how the optimization process, realized by means of the Simulated Annealing algorithm, has significantly reduced the error between experimental and simulated data (as show in Fig. 11), increasing the accuracy of the numerical model with respect to the performance of the "faulty" test-bench. This means that the value of friction torque CSJ estimated at the end of the optimization process is reasonably close to the corresponding real and that, at least for the considered typology of fault, this approach can be satisfactorily used to identify / evaluate the failure. Finally, Fig. 12 shows the diagram of the temperature of the Simulated Annealing process concerning to the just described case.

⁴ Of course, the actuator could not be damaged to quickly obtain "real" faulty data and even less there was enough time to carry out tests aimed at causing wear or seizure in the system. Hence, an expedient had to be found. As the MecVel actuator was equipped with an electric power-out safety brake mounted on the motor's shaft, the authors decided to lower its nominal voltage (24 V, i.e. brake not engaged) in order to increase the braking torque applied to the motor and obtain a sort of "seizure" simulation (i.e. a constant friction torque applied to the motor shaft which can alter its response). In fact, supplying the brake with 11 V by means of a separate power supply, a partial braking action is generated and, as a consequence, the BDC motor exhibits large sticking zones and a much higher delay in response. In this condition the braking torque applied should be theoretically equal to $11 \text{ V} / 24 \text{ V} \cdot 1.7 \text{ Nm} = 0.779 \text{ Nm}$, given that the maximum brake torque is 1.7 Nm. These behaviors could then be assimilated to an incoming failure like a bearing binding which implies an increase of the friction coefficients.

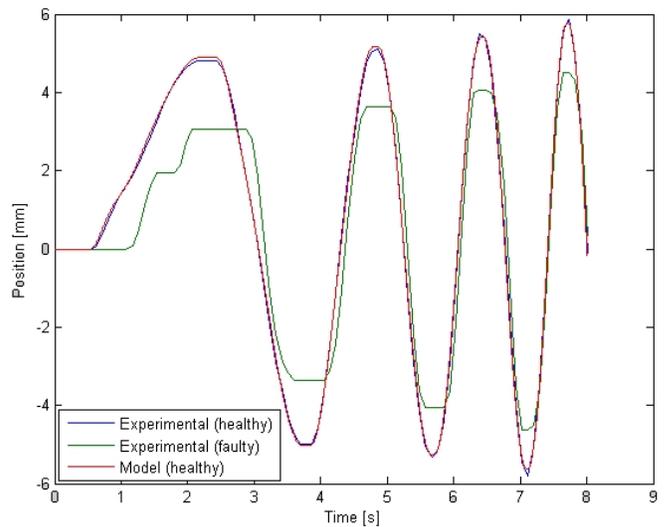


Fig. 9 Experimental vs simulated EMA position before optimization

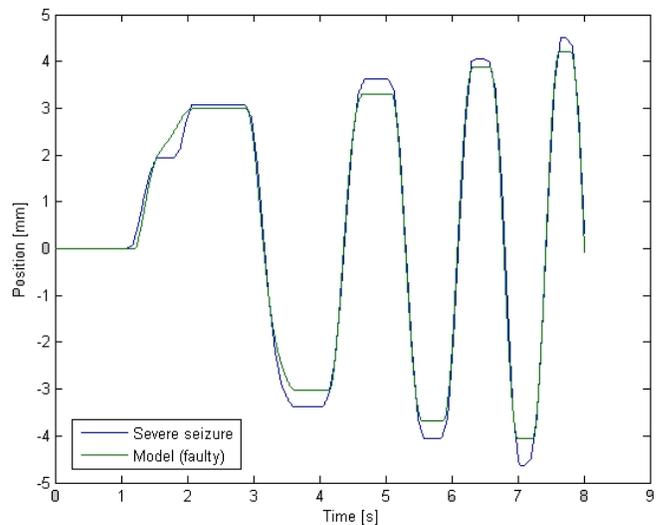


Fig. 10 Experimental vs simulated EMA position after optimization

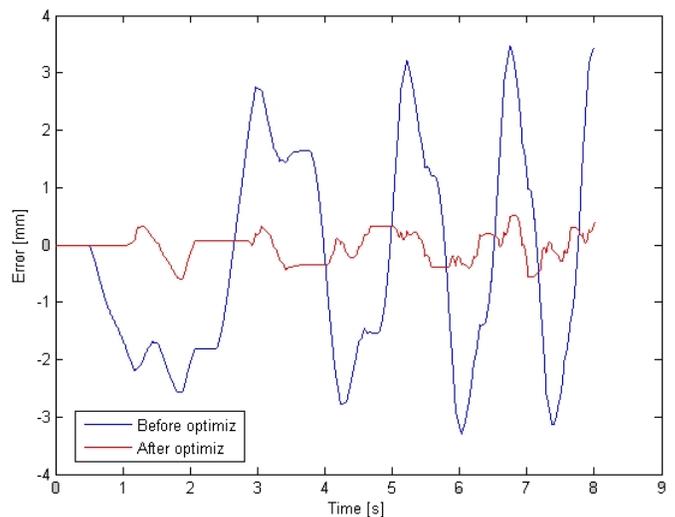


Fig. 11 Experimental vs simulated EMA position residuals before and after optimization

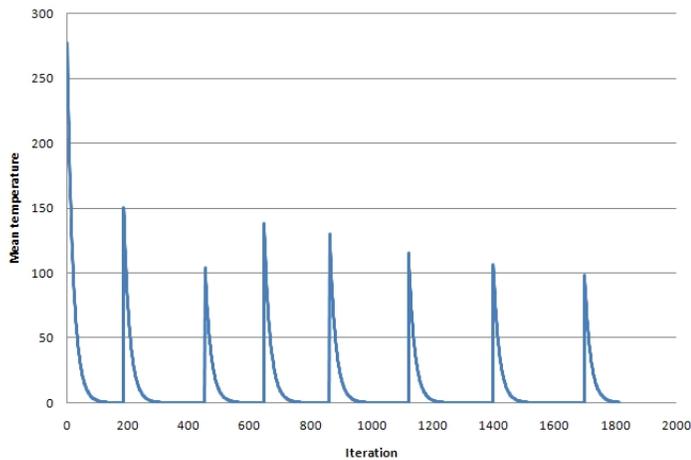


Fig. 12 Simulated annealing process temperature diagram

Comparing the results obtained with the proposed method it is possible to notice how, in this case, the Simulated Annealing algorithm has found a good solution, estimating a static friction coefficient CSJ equal to 0.7352 Nm (and, therefore, very close to the assumed experimental value of 0.779 Nm) and, therefore, a 5.6% of accuracy⁵. It is clear that the brake is heavily affecting the motor's motion causing a significant increase of the friction torque. This can be considered as a "limit" situation. Nevertheless, the simulated annealing algorithm has found a good solution even if starting from a totally different initial point. It can be further observed that the backlash increase is negligible with respect to CSJ and that the faulty CSJ is approximately half of the nominal brake torque (1.7 Nm). Specifically, the resultant 5.6% of accuracy has to be intended just as a rough estimate since it has been calculated supposing that the brake was applying 0.779 Nm but it is not an accurate datum. Finally, the dynamic friction coefficient has decreased by -53%. Additional investigations, performed taking into account also the effects due to electrical noises, analog to digital conversion (ADC) problems, signal transducers affected by offsets or electrical drifts or (reasonable) variations of the boundary conditions, have put in evidence the robustness and the accuracy of this algorithm.

VII. CONCLUSIONS

A model-based damage estimator for an electromechanical actuation system has been developed and tested under different operational conditions using the Simulated Annealing optimization algorithm with a MATLAB Simulink model capable of reproducing the effects of progressive growth of friction acting on mechanical devices (this is simulated properly modifying the static friction coefficient CSJ).

The experimental data useful to demonstrate the damage estimator capabilities have been achieved by means of an electromechanical system developed for this purpose.

⁵ It must be noted that the resultant 5.6% of accuracy, has to be intended just as a rough estimate since it has been calculated supposing that the brake was applying 0.779 Nm, but it is not an accurate datum because, at the moment, this value is only estimated.

This test-bench is able to feed the physical system with different type of signals (i.e. step, ramp, sinusoidal and generic external commands, both in open and closed loop mode), acquiring the position/speed response to a sinusoidal frequency sweep input which showed to be effective within the damage estimation process. The Simulated Annealing proved to be very effective, as its execution times were fairly acceptable (a few minutes) for an operational scenario. However, this method showed a strong dependence of the results on its initialization settings (i.e. initial temperature, function tolerance, reannealing interval) and also on the variables bounds which have to be chosen carefully, making, for example, some considerations regarding their physical limits. In view of the achieved results, this kind of damage estimator can be considered a very powerful tool for PHM applications. Hence its developing should be further improved.

ACKNOWLEDGMENT

In conclusion, the authors wish to extend a heartfelt thanks to Professor Lorenzo Borello for his precious role in the definition of the concepts that have allowed the realization of this work and to Eng. Davide Lauria for his essential support.

REFERENCES

- [1] DoD Panel to Review the V-22 Program, Report of the Panel to Review the V-22 Program, April 2001.
- [2] C. S. Byington, M. Watson, D. Edwards & P. Stoelting, *A Model-Based Approach to Prognostics and Health Management for Flight Control Actuators*, IEEE Aerospace Conference Proceedings, USA, 2004.
- [3] Diagnostics and Prognostics Terms Related to Integrated Systems Diagnostics Design, <http://prognosticshhealthmanagement.com>
- [4] L. Borello, M. D. L. Dalla Vedova, G. Jacazio, M. Sorli, *A Prognostic Model for Electrohydraulic Servovalves*, Proceedings of the Annual Conference of the Prognostics and Health Management Society, 2009.
- [5] M. D. L. Dalla Vedova, P. Maggiore, L. Pace, *Proposal of Prognostic Parametric Method Applied to an Electrohydraulic Servomechanism Affected by Multiple Failures*. WSEAS Transactions on Environment and Development, ISSN: 1790-5079, pp. 478-490
- [6] P. Maggiore, M. D. L. Dalla Vedova, L. Pace, A. Desando, *Definition of parametric methods for fault analysis applied to an electromechanical servomechanism affected by multiple failures*, Proceedings of the Second European Conference of the Prognostics and Health Management Society, 08-10 July 2014, pp. 561-571
- [7] M. Battipede, M.D.L. Dalla Vedova, P. Maggiore, and S. Romeo, *Model based analysis of precursors of electromechanical servomechanisms failures using an artificial neural network*, Proceedings of the AIAA SciTech Modeling and Simulation Technologies Conference, Kissimmee, Florida, 5-9 January 2015.
- [8] M. Pirlot, *General Local Search Methods*, European Journal of Operational Research, vol. 92, pp. 493-511, 1996.
- [9] N. Metropolis, A. N. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equation of state calculation by fast computing machines*, Journal of Chemical Physics, Vol 2, No.6, pp. 1087-1092, 1953.
- [10] D. Karnopp, *Computer simulation of stick-slip friction in mechanical dynamic systems*, Journal of Dynamic Systems, Measurement, and Control, Vol. 107, No. 1, pp. 100-103, 1985.
- [11] L. Borello, and M. D. L. Dalla Vedova, *A dry friction model and robust computational algorithm for reversible or irreversible motion transmission*, International Journal of Mechanics and Control (JoMaC), Vol. 13, No. 02, December 2012, pp. 37-48, ISSN: 1590-8844.
- [12] K.K. Vishwakarma, H.M. Dubey, M. Pandit, and B.K. Panigrahi, *Simulated annealing approach for solving economic load dispatch problems with valve point loading effects*, International Journal of Engineering, Science and Technology, Vol. 4, No. 4, 2012.

Matteo D. L. Dalla Vedova received the M.Sc. and the Ph.D. from the Politecnico di Torino in 2003 and 2007, respectively. He is currently assistant researcher at the Department of Mechanics and Aerospace Engineering. His research activity is mainly focused on the aeronautical systems engineering and, in particular, is dedicated to design, analysis and numerical simulation of on board systems, study of secondary flight control system and conception of related monitoring strategies and developing of prognostic algorithms for aerospace servomechanism.

Paolo Maggiore is a professor at the Mechanical and Aerospace Engineering Department of Politecnico di Torino, that joined in 1992, where he teaches aerospace general systems engineering. Currently his students are involved in projects ranging from hydrogen fuel cell powered airplanes and UAVs, and health monitoring of flight controls, to multi-disciplinary design optimization of aerospace systems design

Lorenzo Pace graduated in Aerospace Engineering at Politecnico di Torino in 2008. Since 2008 to 2011, he worked as an assistant researcher, following studies about system experimental testing and modelization in the aerospace field, with a focus to energy saving techniques. Since 2012 to 2014 he completed a PhD in Aerospace Engineering at Politecnico di Torino, with the contribution of Thales Alenia Space, focused on the application of Model Based System Engineering to verification in the space industry.

Math-statistical Models of Income Distribution: L-moments and TL-moments and Their Estimations

Diana Bílková

Abstract—This paper deals with the application of such robust methods of point parameter estimation, as the methods of L-moments and TL-moments on economic data. The advantages of these highly robust parametric estimation methods are aware when applied to small data sets, especially in the field of hydrology, meteorology and climatology, in particular considering extreme precipitation. The main aim of this contribution is to use these methods on large datasets, and comparison the accuracy of these two methods of parametric estimation with the accuracy of the method of maximum likelihood, especially in terms of efficiency of parametric estimation. The study is divided into a theoretical part, in which mathematical and statistical aspects are described, and an analytical part, during which the results of the use of three robust parametric estimation methods are presented. Total 168 income distributions of the years from 1992 to 2007 in the Czech Republic (distribution of net annual income per capita in CZK) were analyzed. There are a total income distribution for all households of the Czech Republic together and further the income distributions broken down by gender, job classification, classification of economic activities, age and educational attainment. Three-parametric lognormal curves represent the basic theoretical probability distribution. For all analyzed income distributions parameters of this model distribution were estimated using the method of TL-moments, method of L-moments and maximum likelihood method simultaneously and accuracy of these methods were then compared.

Keywords—L-moments and TL-moments of probability distribution, order statistics, quantile function, sample L-moments and TL-moments.

I. INTRODUCTION

THE advantages of these highly robust parametric estimation methods are aware when applied to small data sets, especially in the field of hydrology, meteorology and climatology, in particular considering extreme precipitation. The main aim of this contribution is to use these methods on large datasets, and comparison the accuracy of these two methods of parametric estimation with the accuracy of the

method of maximum likelihood, especially in terms of efficiency of parametric estimation.

There are a total income distribution for all households of the Czech Republic together and further the income distributions broken down by gender, job classification, classification of economic activities, age and educational attainment. Three-parametric lognormal curves represent the basic theoretical probability distribution. For all analyzed income distributions parameters of this model distribution were estimated using the method of TL-moments, method of L-moments and maximum likelihood method simultaneously and accuracy of these methods were then compared.

L-moments form the basis for a general theory, which includes the summarization and description of the theoretical probability distributions, summarization and description of obtained sample data sets, parameter estimation of theoretical probability distributions and hypothesis testing of parameter values for the theoretical probability distributions. The theory of L-moments includes such established methods such as the use of order statistics and Gini middle difference and this leads to some promising innovations in the area of measuring skewness and kurtosis of the distribution and it provides the relatively new methods of parameter estimation for individual distribution. L-moments can be defined for any random variable whose expected value exists. The main advantage of L-moments over conventional moments is that the L-moments can be estimated by linear functions of sample values and they are more resistant to the influence of sample variability. L-moments are more robust than conventional moments to the existence of outliers in the data and they allow better conclusions obtained on the basis of the small samples of basic probability distribution. L-moments sometimes bring even more efficient parameter estimations of parametric distribution than the estimations acquired using maximum likelihood method, particularly for small samples.

L-moments have certain theoretical advantages over conventional moments resting in the ability to characterize a wider range of distribution and they are more resistant to the - compared with conventional moments, L-moments are less prone to estimation bias and approximation by asymptotic normal distribution is more accurate in finite samples.

Let X be a random variable having a distribution with distribution function $F(x)$ and quantile function $x(F)$, and let

This paper was subsidized by the funds of institutional support of a long-term conceptual advancement of science and research number IP400040 at the Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic.

Diana Bílková is with the Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, Sq. W. Churchill 1938/4, 130 67 Prague 3, Czech Republic (e-mail: bilkova@vse.cz).

X_1, X_2, \dots, X_n is a random sample of sample size n from this distribution. Then $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ are the order statistics of random sample of sample size n , which comes from the distribution of random variable X .

L-moments are analogous to conventional moments. They can be estimated based on linear combinations of sample order statistics, i.e. L-statistics. L-moments are an alternative system describing the shape of the probability distribution.

II. THEORY AND METHODS

A. L-moments of Probability Distribution

The problem of L-moments is discussed for example in [1] or [2]. Let X be a continuous random variable that has a distribution with distribution function $F(x)$ and the quantile function $x(F)$. Let $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ be the order statistics of random sample of sample size n , which comes from the distribution of random variable X . L-moment of the r -th order of random variable X is defined

$$\lambda_r = \frac{1}{r} \sum_{j=0}^{r-1} (-1)^j \binom{r-1}{j} \cdot E(X_{r-j:r}), \quad r=1, 2, \dots \quad (1)$$

The expected value of the r -th order statistic of random sample of sample size n has the form

$$E(X_{r:n}) = \frac{n!}{(r-1)! \cdot (n-r)!} \int_0^1 x(F) \cdot [F(x)]^{r-1} \cdot [1-F(x)]^{n-r} dF(x). \quad (2)$$

If we substitute equation (2) into equation (1), we obtain after adjustments

$$\lambda_r = \int_0^1 x(F) \cdot P_{r-1}^* [F(x)] dF(x), \quad r=1, 2, \dots, \quad (3)$$

where

$$P_r^* [F(x)] = \sum_{j=0}^r p_{r,j}^* \cdot [F(x)]^j \quad \text{and} \quad p_{r,j}^* = (-1)^{r-j} \binom{r}{j} \binom{r+j}{j}, \quad (4)$$

and $P_r^* [F(x)]$ is the r -th shifted Legendre polynomial. We also obtain substituting expression (2) into expression (1)

$$\lambda_r = \frac{1}{r} \sum_{j=0}^{r-1} (-1)^j \binom{r-1}{j} \cdot \frac{r!}{(r-j-1)! \cdot j!} \int_0^1 x(F) \cdot [F(x)]^{r-j-1} \cdot [1-F(x)]^j dF(x), \quad r=1, 2, \dots \quad (5)$$

The letter „L“ in title „L-moments“ stresses that the r -th L-moment λ_r is a linear function of the expected value of a certain linear combination of order statistics. Own estimation of the r -th L-moment λ_r based on the obtained data sample is then a linear combination of order data values, i.e. L-statistics. The first four L-moments of the probability distribution are now defined

$$\lambda_1 = E(X_{1:1}) = \int_0^1 x(F) dF(x), \quad (6)$$

$$\lambda_2 = \frac{1}{2} E(X_{2:2} - X_{1:2}) = \int_0^1 x(F) \cdot [2F(x) - 1] dF(x), \quad (7)$$

$$\lambda_3 = \frac{1}{3} E(X_{3:3} - 2X_{2:3} + X_{1:3}) = \int_0^1 x(F) \cdot [6[F(x)]^2 - 6F(x) + 1] dF(x), \quad (8)$$

$$\lambda_4 = \frac{1}{4} E(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}) = \int_0^1 x(F) \cdot [20[F(x)]^3 - 30[F(x)]^2 + 12[F(x)] - 1] dF(x). \quad (9)$$

The probability distribution can be specified by its L-moments, even if some of its conventional moments do not exist, but the opposite is not true. It can be proved that the first L-moment λ_1 is the location characteristic, the second L-moment λ_2 is the variability characteristic. It is often desirable to standardize higher L-moments λ_r , $r \geq 3$, so that they are independent on specific units of the random variable X . The ratio of L-moments of the r -th order of random variable X is defined

$$\tau_r = \frac{\lambda_r}{\lambda_2}, \quad r=3, 4, \dots \quad (10)$$

We can also define a function of L-moments, which is analogous to the classical coefficient of variation, i.e. the so called L-coefficient of variation

$$\tau = \frac{\lambda_2}{\lambda_1}. \quad (11)$$

The ratio of L-moments τ_3 is the skewness characteristic and the ratio of L-moments τ_4 is the kurtosis characteristic of the corresponding probability distribution. Main properties of the probability distribution are very well summarized by the following four characteristics: L-location λ_1 , L-variability λ_2 , L-skewness τ_3 and L-kurtosis τ_4 . L-moments λ_1 and λ_2 , L-coefficient of variation τ and ratios of L-moments τ_3 and τ_4 are the most useful characteristics for summarization of probability distribution. Their main properties are: existence (if the expected value of the distribution exists, then all its L-moments exist) and uniqueness (if the expected value of the distribution exists, then L-moments define the only one distribution, i.e. no two distributions have the same L-moments). Using equations (6)–(9) and equation (10) we obtain the expressions for L-moments, respectively for the ratios of L-moments for the case of chosen probability distributions, see Table I.

B. Sample L-moments

L-moments are usually estimated by random sample obtained from an unknown distribution. Since the r -th L-moment λ_r is a function of the expected values of order statistics of a random sample of sample size r , it is natural to estimate it using the so-called U-statistic, i.e. the corresponding function of sample order statistics (averaged over all subsets of sample size r , which may be formed from the obtained random sample of sample size n).

Table I. Formulas for the Distribution Function or Quantile Function, and for L-Moments and Ratios of L-Moments of Chosen Probability Distributions

Distribution	Distribution function $F(x)$ or quantile function $x(F)$	L-moments and ratios of L-moments
Uniform	$x(F) = \alpha + (\beta - \alpha) \cdot F(x)$	$\lambda_1 = \frac{\alpha + \beta}{2}$ $\lambda_2 = \frac{\beta - \alpha}{6}$ $\tau_3 = 0$ $\tau_4 = 0$
Exponential	$x(F) = \xi - \alpha \cdot \ln[1 - F(x)]$	$\lambda_1 = \xi + \alpha$ $\lambda_2 = \frac{\alpha}{2}$ $\tau_3 = \frac{1}{3}$ $\tau_4 = \frac{1}{6}$
Gumbel	$x(F) = \xi - \alpha \cdot \ln[-\ln F(x)]$	$\lambda_1 = \xi + e \cdot \alpha$ $\lambda_2 = \alpha \cdot \ln 2$ $\tau_3 = 0,1699$ $\tau_4 = 0,1504$
Logistic	$x(F) = \xi + \alpha \cdot \ln \frac{F(x)}{1 - F(x)}$	$\lambda_1 = \xi$ $\lambda_2 = \alpha$ $\tau_3 = 0$ $\tau_4 = \frac{1}{6}$
Normal	$F(x) = \Phi \left[\frac{x(F) - \mu}{\sigma} \right]$	$\lambda_1 = \mu$ $\lambda_2 = \pi^{-1} \cdot \sigma$ $\tau_3 = 0$ $\tau_4 = 30 \cdot \pi^{-1} \cdot (\tan \sqrt{2})^{-1} - 9 = 0,1226$
Generalized Pareto	$x(F) = \xi + \alpha \cdot \frac{1 - [1 - F(x)]^k}{k}$	$\lambda_1 = \xi + \frac{\alpha}{1+k}$ $\lambda_2 = \frac{\alpha}{(1+k) \cdot (2+k)}$ $\tau_3 = \frac{1-k}{3+k}$ $\tau_4 = \frac{(1-k) \cdot (2-k)}{(3+k) \cdot (4+k)}$
Generalized extreme value	$x(F) = \xi + \alpha \cdot \frac{1 - [-\ln F(x)]^k}{k}$	$\lambda_1 = \xi + \alpha \cdot \frac{1 - \Gamma(1+k)}{k}$ $\lambda_2 = \alpha \cdot \frac{(1-2^{-k}) \cdot \Gamma(1+k)}{k}$ $\tau_3 = \frac{2 \cdot (1-3^{-k})}{1-2^{-k}} - 3$ $\tau_4 = \frac{1-6 \cdot 2^{-k} + 10 \cdot 3^{-k} - 5 \cdot 4^{-k}}{1-2^{-k}}$
Generalized logistic	$x(F) = \xi + \alpha \cdot \frac{1 - \left[\frac{1-F(x)}{F(x)} \right]^k}{k}$	$\lambda_1 = \xi + \alpha \cdot \frac{1 - \Gamma(1+k) \cdot \Gamma(1-k)}{k}$ $\lambda_2 = \alpha \cdot \Gamma(1+k) \cdot \Gamma(1-k)$ $\tau_3 = -k$ $\tau_4 = \frac{1+5k^2}{6}$
Lognormal	$F(x) = \Phi \left\{ \frac{\ln[x(F) - \xi] - \mu}{\sigma} \right\}$	$\lambda_1 = \xi + \exp \left(\mu + \frac{\sigma^2}{2} \right)$ $\lambda_2 = \exp \left(\mu + \frac{\sigma^2}{2} \right) \cdot \operatorname{erf} \left(\frac{\sigma}{2} \right)$ $\tau_3 = 6 \pi^{\frac{1}{2}} \cdot \frac{\int_0^{\frac{\sigma}{2}} \operatorname{erf} \left(\frac{x}{\sqrt{3}} \right) \cdot \exp(-x^2) dx}{\operatorname{erf} \left(\frac{\sigma}{2} \right)}$

Source: [13]; own research

Table I. Continuation

Distribution	Distribution function $F(x)$ or quantile function $x(F)$	L-moments and ratios of L-moments
Gamma	$F(x) = \frac{\beta^{-\alpha}}{\Gamma(\alpha)} \cdot \int_0^{x(F)} t^{\alpha-1} \cdot \exp\left(-\frac{t}{\beta}\right) dt$	$\lambda_1 = \alpha \cdot \beta$ $\lambda_2 = \pi^{\frac{1}{2}} \cdot \beta \cdot \frac{\Gamma\left(\alpha + \frac{1}{2}\right)}{\Gamma(\alpha)}$ $\tau_3 = 6 I_{\frac{1}{3}}(\alpha, 2\alpha) - 3$ ¹⁾

Source: [13]; own research

Let x_1, x_2, \dots, x_n is the sample and $x_{1:n} \leq x_{2:n} \leq \dots \leq x_{n:n}$ is order sample. Then the r -th sample L-moment can be written as

$$l_r = \binom{n}{r}^{-1} \sum_{1 \leq i_1 < i_2 < \dots < i_r \leq n} \frac{1}{r} \sum_{j=0}^{r-1} (-1)^j \binom{r-1}{j} \cdot x_{i_{r-j}:n}, \quad r=1, 2, \dots, n. \quad (12)$$

Hence the first four sample L-moments have the form

$$l_1 = \frac{1}{n} \cdot \sum_i x_i, \quad (13)$$

$$l_2 = \frac{1}{2} \cdot \binom{n-1}{2}^{-1} \cdot \sum_{i>j} (x_{i:n} - x_{j:n}), \quad (14)$$

$$l_3 = \frac{1}{3} \cdot \binom{n-1}{3}^{-1} \cdot \sum_{i>j>k} (x_{i:n} - 2x_{j:n} + x_{k:n}), \quad (15)$$

$$l_4 = \frac{1}{4} \cdot \binom{n-1}{4}^{-1} \cdot \sum_{i>j>k>l} (x_{i:n} - 3x_{j:n} + 3x_{k:n} - x_{l:n}). \quad (16)$$

U-statistics are widely used especially in nonparametric statistics. Their positive properties are: absence of bias, asymptotic normality and some slight resistance due to the influence of outliers.

When calculating the r -th sample L-moment it is not necessary to repeat the process over all sub-sets of sample size r , but this statistic can be expressed directly as a linear combination of order statistics of a random sample of sample size n .

If we thing the estimation of $E(X_{r:r})$ obtained using U-statistics, this estimation can be written as $r \cdot b_{r-1}$, where

$$b_r = \frac{1}{n} \cdot \binom{n-1}{r}^{-1} \cdot \sum_{j=r+1}^n \binom{j-1}{r} \cdot x_{j:n}, \quad (17)$$

namely

$$b_0 = \frac{1}{n} \cdot \sum_{j=1}^n x_{j:n}, \quad (18)$$

$$b_1 = \frac{1}{n} \cdot \sum_{j=2}^n \frac{(j-1)}{(n-1)} \cdot x_{j:n}, \quad (19)$$

$$b_2 = \frac{1}{n} \cdot \sum_{j=3}^n \frac{(j-1) \cdot (j-2)}{(n-1) \cdot (n-2)} \cdot x_{j:n}, \quad (20)$$

therefore generally

¹⁾ $I_x(p, q)$ is incomplete beta function

$$b_r = \frac{1}{n} \sum_{j=r+1}^n \frac{(j-1) \cdot (j-2) \cdot \dots \cdot (j-r)}{(n-1) \cdot (n-2) \cdot \dots \cdot (n-r)} \cdot x_{j:n} \quad (21)$$

The first four sample L-moments can be therefore written as

$$l_1 = b_0, \quad (22)$$

$$l_2 = 2b_1 - b_0, \quad (23)$$

$$l_3 = 6b_2 - 6b_1 + b_0, \quad (24)$$

$$l_4 = 20b_3 - 30b_2 + 12b_1 - b_0. \quad (25)$$

Table II. Formulas for Estimations of Parameters Taken by the Method of L-Moments of Chosen Probability Distributions

Distribution	Parameter estimation
Exponential	(ξ known) $\hat{\alpha} = l_1$
Gumbel	$\hat{\alpha} = \frac{l_2}{\ln 2}$ $\hat{\xi} = l_1 - e \cdot \hat{\alpha}$
Logistic	$\hat{\alpha} = l_2$ $\hat{\xi} = l_1$
Normal	$\hat{\sigma} = \pi^{\frac{1}{2}} \cdot l_2$ $\hat{\mu} = l_1$
Generalized Pareto	(ξ known) $\hat{k} = \frac{l_1}{l_2} - 2$ $\hat{\alpha} = (1 + \hat{k}) \cdot l_1$
Generalized extreme value	$z = \frac{2}{3+t_3} - \frac{\ln 2}{\ln 3}$ $\hat{k} = 7,8590z + 2,9554z^2$ $\hat{\alpha} = \frac{l_2 \cdot \hat{k}}{(1-2^{-\hat{k}}) \cdot \Gamma(1+\hat{k})}$ $\hat{\xi} = l_1 + \hat{\alpha} \cdot \frac{\Gamma(1+\hat{k})-1}{\hat{k}}$
Generalized logistic	$\hat{k} = -t_3$ $\hat{\alpha} = \frac{l_2}{\Gamma(1+\hat{k}) \cdot \Gamma(1-\hat{k})}$ $\hat{\xi} = l_1 + \frac{l_2 - \hat{\alpha}}{\hat{k}}$
Lognormal	$z = \frac{\sqrt{8}}{\sqrt{3}} \cdot \Phi^{-1}\left(\frac{1+t_3}{2}\right)$ $\hat{\sigma} = 0,999281z - 0,006118z^3 + 0,000127z^5$ $\hat{\mu} = \ln \frac{l_2}{\text{erf}\left(\frac{\hat{\sigma}}{2}\right)} - \frac{\hat{\sigma}^2}{2}$ $\hat{\xi} = l_1 - \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right)$
Gamma	(ξ known) $t = \frac{l_2}{l_1}$ if $0 < t < \frac{1}{2}$, then: $z = \pi \cdot t^2$ $\hat{\alpha} \approx \frac{1 - 0,3080z}{z - 0,05812z^2 + 0,01765z^3}$ if $\frac{1}{2} \leq t < 1$, then: $z = 1 - t$ $\hat{\alpha} \approx \frac{0,7213z - 0,5947z^2}{1 - 2,1817z + 1,2113z^2}$ $\hat{\beta} = \frac{l_1}{\hat{\alpha}}$

Source: [13]; own research

We can therefore write generally

$$l_{r+1} = \sum_{k=0}^r {}^* p_{r,k} \cdot b_k, \quad r=0,1,\dots,n-1, \quad (26)$$

where

$${}^* p_{r,k} = (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} = \frac{(-1)^{r-k} \cdot (r+k)!}{(k!) \cdot (r-k)!} \quad (27)$$

Use of sample L-moments is similar to the use of sample conventional L-moments. Sample L-moments summarize the basic properties of the sample distribution, which are the location (level), variability, skewness and kurtosis. Thus, sample L-moments estimate the corresponding properties of the probability distribution from which the sample comes, and they can be used in estimating the parameters of the relevant probability distribution. L-moments are often preferred over conventional moments within such applications, since, as the linear functions of sample values, sample L-moments are less sensitive to sample variability than conventional moments, or to measurement errors in the case of extreme observations. L-moments therefore lead to more accurate and robust estimations of parameters or characteristics of the basic probability distribution.

Sample L-moments have been used previously in the statistics, although not as part of a unified theory. The first sample L-moment l_1 is a sample L-location (sample average), the second sample L-moment l_2 is a sample L-variability. Natural estimation of the ratio of L-moments (10) is the sample ratio of L-moments

$$t_r = \frac{l_r}{l_2}, \quad r = 3, 4, \dots \quad (28)$$

Hence t_3 is a sample L-skewness and t_4 is a sample L-kurtosis. Sample ratios of L-moments t_3 and t_4 may be used as characteristics of skewness and kurtosis of the sample data set.

Gini middle difference relates to the sample L-moments and it has the form

$$G = \binom{n}{2}^{-1} \cdot \sum_{i>j} (x_{i:n} - x_{j:n}), \quad (29)$$

and Gini coefficient, which depends only on a single parameter σ in the case of two-parametric lognormal distribution, but it depends on the values of all three parameters in the case of three-parametric lognormal distribution. Table II presents the expressions for parameter estimations of chosen probability distributions obtained using the method of L-moments. For more details see for example [3]– [14].

²⁾ $\Phi^{-1}(\cdot)$ is quantile function of standardized normal distribution

C. TL-moments of Probability Distribution

An alternative robust version of L-moments will be introduced now. This modification of L-moments is called the „trimmed L-moments“ and it is noted TL-moments. In this modification of L-moments the expected values of order statistics of random sample in the definition of L-moments of probability distributions are replaced with expected values of order statistics of larger random sample and the sample size grows in such a way that it corresponds to the total size of the adjustment, as shown below.

TL-moments have certain advantages over conventional L-moments and central moments. TL-moment of probability distribution may exist even if the corresponding L-moment or central moment of this probability distribution does not exist, as it is the case of Cauchy distribution. Sample TL-moments are more resistant to outliers in the data. Method of TL-moments is not intended to replace the existing robust methods, but rather as their supplement, especially in situations where we have outliers in the data.

In this alternative robust modification of L-moments the expected value $E(X_{r:j;r})$ is replaced by the expected value $E(X_{r+t_1-j : r+t_1+t_2})$. Thus, for each r we increase the sample size of random sample from the original r to $r + t_1 + t_2$ and we work only with the expected values of these r modified order statistics $X_{t_1+1:r+t_1+t_2}, X_{t_1+2:r+t_1+t_2}, \dots, X_{t_1+r:r+t_1+t_2}$ by trimming the t_1 smallest and the t_2 largest from the conceptual sample. This modification is called the r -th trimmed L-moment (TL-moment) and it is marked $\lambda_r^{(t_1,t_2)}$. Thus, TL-moment of the r -th order of random variable X is defined

$$\lambda_r^{(t_1,t_2)} = \frac{1}{r} \cdot \sum_{j=0}^{r-1} (-1)^j \cdot \binom{r-1}{j} \cdot E(X_{r+t_1-j : r+t_1+t_2}), \quad r = 1, 2, \dots \quad (30)$$

It is evident from the expressions (30) and (1) that TL-moments reduce to L-moments where $t_1 = t_2 = 0$. Although we can also consider applications where the adjustment values are not equal, i.e. $t_1 \neq t_2$, we focus here only on the symmetric case $t_1 = t_2 = t$. Then the expression (30) can be rewritten

$$\lambda_r^{(t)} = \frac{1}{r} \cdot \sum_{j=0}^{r-1} (-1)^j \cdot \binom{r-1}{j} \cdot E(X_{r+t-j : r+2t}), \quad r = 1, 2, \dots \quad (31)$$

Thus, for example $\lambda_1^{(t)} = E(X_{1+t:1+2t})$ is the expected value of the median of the conceptual random sample of sample size $1 + 2t$. It is necessary to note here that $\lambda_1^{(t)}$ is equal to zero for distributions that are symmetrical around zero.

The first four TL-moments have the form for $t = 1$

$$\lambda_1^{(1)} = E(X_{2:3}), \quad (32)$$

$$\lambda_2^{(1)} = \frac{1}{2} E(X_{3:4} - X_{2:4}), \quad (33)$$

$$\lambda_3^{(1)} = \frac{1}{3} E(X_{4:5} - 2X_{3:5} + X_{2:5}), \quad (34)$$

$$\lambda_4^{(1)} = \frac{1}{4} E(X_{5:6} - 3X_{4:6} + 3X_{3:6} - X_{2:6}). \quad (35)$$

Measurements of location, variability, skewness and kurtosis of the probability distribution analogous to conventional L-moments (6)–(9) are based on $\lambda_1^{(1)}, \lambda_2^{(1)}, \lambda_3^{(1)}$ a $\lambda_4^{(1)}$.

Expected value $E(X_{r:n})$ can be written using the formula (2). Using equation (2) we can re-express the right side of equation (31)

$$\lambda_r^{(t)} = \frac{1}{r} \cdot \sum_{j=0}^{r-1} (-1)^j \cdot \binom{r-1}{j} \cdot \frac{(r+2t)!}{(r+t-j-1)! \cdot (t+j)!} \cdot \int_0^1 x(F) \cdot [F(x)]^{r+t-j-1} \cdot [1-F(x)]^{t+j} dF(x), \quad r = 1, 2, \dots \quad (36)$$

It is necessary to notify here that $\lambda_r^{(0)} = \lambda_r$ represents normal the r -th L-moment with no adjustment.

Expressions (32)–(35) for the first four TL-moments ($t = 1$) may be written in an alternative manner

$$\lambda_1^{(1)} = 6 \cdot \int_0^1 x(F) \cdot [F(x)] \cdot [1-F(x)] dF(x), \quad (37)$$

$$\lambda_2^{(1)} = 6 \cdot \int_0^1 x(F) \cdot [F(x)] \cdot [1-F(x)] \cdot [2F(x)-1] dF(x), \quad (38)$$

$$\lambda_3^{(1)} = \frac{20}{3} \cdot \int_0^1 x(F) \cdot [F(x)] \cdot [1-F(x)] \cdot [5[F(x)]^2 - 5F(x) + 1] dF(x), \quad (39)$$

$$\lambda_4^{(1)} = \frac{15}{2} \cdot \int_0^1 x(F) \cdot [F(x)] \cdot [1-F(x)] \cdot [14[F(x)]^3 - 21[F(x)]^2 + 9[F(x)] - 1] dF(x). \quad (40)$$

Distribution may be identified by its TL-moments, although some of its L-moments and conventional moments do not exist. For example $\lambda_1^{(1)}$ (the expected value of median of conceptual random sample of sample size three) exists for Cauchy distribution, although the first L-moment λ_1 does not exist.

TL-skewness $\tau_3^{(t)}$ and TL-kurtosis $\tau_4^{(t)}$ can be defined analogously as L-skewness τ_3 and L-kurtosis τ_4

$$\tau_3^{(t)} = \frac{\lambda_3^{(t)}}{\lambda_2^{(t)}}, \quad (41)$$

$$\tau_4^{(t)} = \frac{\lambda_4^{(t)}}{\lambda_2^{(t)}}. \quad (42)$$

D. Sample TL-moments

Let x_1, x_2, \dots, x_n is the sample and $x_{1:n} \leq x_{2:n} \leq \dots \leq x_{n:n}$ is order sample. The expression

$$\hat{E}(X_{j+l+1:j+l+1}) = \frac{1}{\binom{n}{j+l+1}} \cdot \sum_{i=1}^n \binom{i-1}{j} \cdot \binom{n-i}{l} \cdot x_{i:n} \quad (43)$$

is considered to be an unbiased estimation of the expected value of the $(j + 1)$ -th order statistic $X_{j+1:j+l+1}$ in the conceptual random sample of sample size $(j + l + 1)$. Now we assume that in the definition of the TL-moment $\lambda_r^{(t)}$ in (31) we replace the expression $E(X_{r+t-j:r+2t})$ by its unbiased estimation

$$\hat{E}(X_{r+t-j:r+2t}) = \frac{1}{\binom{n}{r+2t}} \sum_{i=1}^n \binom{i-1}{r+t-j-1} \binom{n-i}{t+j} \cdot x_{i:n}, \quad (44)$$

which is obtained by assigning $j \rightarrow r+t-j-1$ a $l \rightarrow t+j$ in (43). Now we get the r -th sample TL-moment

$$l_r^{(t)} = \frac{1}{r} \sum_{j=0}^{r-1} (-1)^j \binom{r-1}{j} \cdot \hat{E}(X_{r+t-j:r+2t}), \quad r=1, 2, \dots, n-2t, \quad (45)$$

i.e.

$$l_r^{(t)} = \frac{1}{r} \sum_{j=0}^{r-1} (-1)^j \binom{r-1}{j} \cdot \frac{1}{\binom{n}{r+2t}} \sum_{i=1}^n \binom{i-1}{r+t-j-1} \binom{n-i}{t+j} \cdot x_{i:n}, \quad r=1, 2, \dots, n-2t, \quad (46)$$

which is an unbiased estimation of the r -th TL-moment $\lambda_r^{(t)}$. Note that for each $j = 0, 1, \dots, r-1$ the values $x_{i:n}$ in (46) are not equal to zero only for $r+t-j \leq i \leq n-t-j$ relative to the combination numbers. Simple adjustment of equation (46) provides an alternative linear form

$$l_r^{(t)} = \frac{1}{r} \sum_{i=r+t}^{n-t} \left[\frac{\sum_{j=0}^{r-1} (-1)^j \binom{r-1}{j} \binom{i-1}{r+t-j-1} \binom{n-i}{t+j}}{\binom{n}{r+2t}} \right] \cdot x_{i:n}. \quad (47)$$

For example, we obtain for the first sample TL-moment for $r = 1$

$$l_1^{(t)} = \sum_{i=t+1}^{n-t} w_{i:n}^{(t)} \cdot x_{i:n}, \quad (48)$$

where the weights are given by

$$w_{i:n}^{(t)} = \frac{\binom{i-1}{t} \binom{n-i}{t}}{\binom{n}{2t+1}}. \quad (49)$$

The above results can be used to estimate TL-skewness $\tau_3^{(t)}$ and TL-kurtosis $\tau_4^{(t)}$ by simple ratios

$$t_3^{(t)} = \frac{l_3^{(t)}}{l_2^{(t)}}, \quad (50)$$

$$t_4^{(t)} = \frac{l_4^{(t)}}{l_2^{(t)}}. \quad (51)$$

We can choose $t = n\alpha$ representing the amount of adjustment from each end of the sample, where α is a certain ratio, where $0 \leq \alpha < 0,5$.

Table III contains the expressions for TL-moments and ratios of TL-moments and expressions for parameter estimations of chosen probability distributions obtained using

the method of TL-moments ($t = 1$), more see for example in [15].

Table III. Formulas for TL-Moments and Ratios of TL-Moments and Formulas for Estimations of Parameters Taken by the Method of TL-Moments of Chosen Probability Distributions ($t = 1$)

Distribution	TL-moments and ratios of TL-moments	Parameter estimation
Normal	$\lambda_1^{(1)} = \mu$ $\lambda_2^{(1)} = 0,297 \sigma$ $\tau_3^{(1)} = 0$ $\tau_4^{(1)} = 0,062$	$\hat{\mu} = l_1^{(1)}$ $\hat{\sigma} = \frac{l_2^{(1)}}{0,297}$
Logistic	$\lambda_1^{(1)} = \mu$ $\lambda_2^{(1)} = 0,500 \sigma$ $\tau_3^{(1)} = 0$ $\tau_4^{(1)} = 0,083$	$\hat{\mu} = l_1^{(1)}$ $\hat{\sigma} = 2l_2^{(1)}$
Cauchy	$\lambda_1^{(1)} = \mu$ $\lambda_2^{(1)} = 0,698 \sigma$ $\tau_3^{(1)} = 0$ $\tau_4^{(1)} = 0,343$	$\hat{\mu} = l_1^{(1)}$ $\hat{\sigma} = \frac{l_2^{(1)}}{0,698}$
Exponential	$\lambda_1^{(1)} = \frac{5\alpha}{6}$ $\lambda_2^{(1)} = \frac{\alpha}{4}$ $\tau_3^{(1)} = \frac{2}{9}$ $\tau_4^{(1)} = \frac{1}{12}$	$\hat{\alpha} = \frac{6l_1^{(1)}}{5}$

Source: [12]; own research

E. Maximum Likelihood Method

Let the random sample of sample size n comes from three-parametric lognormal distribution with probability density function

$$f(x; \mu, \sigma^2, \theta) = \frac{1}{\sigma \cdot (x-\theta) \cdot \sqrt{2\pi}} \cdot \exp\left[-\frac{[\ln(x-\theta) - \mu]^2}{2\sigma^2}\right], \quad x > \theta, \quad (52)$$

$$= 0, \quad \text{else,}$$

where $-\infty < \mu < \infty$, $\sigma^2 > 0$, $-\infty < \theta < \infty$ are parameters. Three-parametric lognormal distribution is described in detail for example in [8], [9], [11] or [12].

The likelihood function then has the form

$$L(x; \mu, \sigma^2, \theta) = \prod_{i=1}^n f(x_i; \mu, \sigma^2, \theta) =$$

$$= \frac{1}{(\sigma^2)^{n/2} \cdot (2\pi)^{n/2} \cdot \prod_{i=1}^n (x_i - \theta)} \cdot \exp\left\{-\sum_{i=1}^n \frac{[\ln(x_i - \theta) - \mu]^2}{2\sigma^2}\right\}. \quad (53)$$

We determine the natural logarithm of the likelihood function

$$\ln L(x; \mu, \sigma^2, \theta) = \sum_{i=1}^n \left[-\frac{[\ln(x_i - \theta) - \mu]^2}{2\sigma^2} - \frac{n}{2} \cdot \ln \sigma^2 - \frac{n}{2} \cdot \ln(2\pi) - \sum_{i=1}^n \ln(x_i - \theta) \right]. \quad (54)$$

We put the first partial derivations of the logarithm of the likelihood function according to μ and σ^2 in the equality to zero. We obtain a system of likelihood equations

$$\frac{\partial \ln L(\mathbf{x}; \mu, \sigma^2, \theta)}{\partial \mu} = \frac{\sum_{i=1}^n [\ln(x_i - \theta) - \mu]}{\sigma^2} = 0, \quad (55)$$

$$\frac{\partial \ln L(\mathbf{x}; \mu, \sigma^2, \theta)}{\partial \sigma^2} = \frac{\sum_{i=1}^n [\ln(x_i - \theta) - \mu]^2}{2\sigma^4} - \frac{n}{2\sigma^2} = 0. \quad (56)$$

After adjustment we obtain maximum likelihood estimations of the parameters μ and σ^2 for the parameter θ

$$\hat{\mu}(\theta) = \frac{\sum_{i=1}^n \ln(x_i - \theta)}{n}, \quad (57)$$

$$\hat{\sigma}^2(\theta) = \frac{\sum_{i=1}^n [\ln(x_i - \theta) - \hat{\mu}(\theta)]^2}{n}. \quad (58)$$

If the value of the parameter θ is known, we get maximum likelihood estimations of the remaining two parameters of three-parametric lognormal distribution using equations (57) and (58). However, if the value of the parameter θ is unknown, the problem is more complicated. It can be proved that if the parameter θ closes to $\min\{X_1, X_2, \dots, X_n\}$, then the maximum likelihood approaches to infinity. The maximum likelihood method is also often combined with Cohen method, where we put the smallest sample value to be equal to the $100 \cdot (n + 1)^{-1}$ -percentage quantile

$$x_{\min}^V = \hat{\theta} + \exp(\hat{\mu} + \hat{\sigma} \cdot u_{(n+1)^{-1}}). \quad (59)$$

Equation (59) is then combined with a system of equations (57) and (58).

For solving of maximum likelihood equations (57) and (58) it is also possible to use $\hat{\theta}$ satisfying the equation

$$\sum_{i=1}^n (x_i - \hat{\theta}) + \frac{\sum_{i=1}^n z_i^{\frac{1}{2}}}{\hat{\sigma}(\hat{\theta})} = 0, \quad (60)$$

where

$$z_i = \frac{\ln(x_i - \hat{\theta}) - \hat{\mu}(\hat{\theta})}{\hat{\sigma}(\hat{\theta})}, \quad (61)$$

where $\hat{\mu}(\hat{\theta})$ and $\hat{\sigma}(\hat{\theta})$ satisfy equations (57) and (58) with the parameter θ replaced by $\hat{\theta}$. We may also obtain the limits of variances

$$n \cdot D(\hat{\theta}) = \frac{\sigma^2 \cdot \exp(2\mu)}{\omega \cdot [\omega \cdot (1 + \sigma^2) - 2\sigma^2 - 1]}, \quad (62)$$

$$n \cdot D(\hat{\mu}) = \frac{\sigma^2 \cdot [\omega \cdot (1 + \sigma^2) - 2\sigma^2]}{\omega \cdot (1 + \sigma^2) - 2\sigma^2 - 1}, \quad (63)$$

$$n \cdot D(\hat{\sigma}) = \frac{\sigma^2 \cdot [\omega \cdot (1 + \sigma^2) - 1]}{\omega \cdot (1 + \sigma^2) - 2\sigma^2 - 1}. \quad (64)$$

III. DISCUSSION AND RESULTS

In the past L-moments were mainly used in hydrology, climatology and meteorology in the research of extreme precipitation, see for example [14]. There are mainly small data sets in this cases. This study presents an application of L-moments and TL-moments on large sets of economic data, Table IV presents the sample sizes of obtained sample sets of households.

The researched variable is the net annual household income per capita (in CZK) within the Czech Republic (nominal income). The data obtained come from a statistical survey Microcensus – years 1992, 1996, 2002, and statistical survey EU-SILC (The European Union Statistics on Income and Living Conditions) – the period 2004-2007, from the Czech Statistical Office. Total 168 income distributions were analyzed this way, both for all households of the Czech Republic together and also broken down by gender, country (Bohemia and Moravia (including Silesia), see Fig. 1), social groups, municipality size, age and the highest educational attainment, while households are classified into different subsets according to the head of household, which is man in the vast majority of households. Sharply smaller sample sizes for women than for men in Table IV correspond to this fact. Head of household is always a man in two-parent families of the type the husband and wife or two partners, regardless of the economic activity. In single-parent families of the type only one parent with children and in non-family households, where persons are not related by marriage or by union partner, nor parent-child relationship, the first decisive criterion for determining the head of household is the economic activity and the second aspect is the amount of money income of individual household members. This criterion also applies in the case of more complex types of households, such as the case of joint management of more than two-parent families.

The parameters of three-parametric lognormal curves were estimated simultaneously using three robust methods of parametric estimation, namely the method of TL-moments, the method of L-moments and the maximum likelihood method and accuracy of these methods were compared with each other using the familiar test criterion

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n \pi_i)^2}{n \pi_i}, \quad (65)$$

where n_i are the observed frequencies in individual income intervals, π_i are the theoretical probabilities of belonging of statistical unit to the i -th interval, $n \cdot \pi_i$ are the theoretical frequencies in individual income intervals, $i = 1, 2, \dots, k$, n is the total sample size of the corresponding statistical set and k is the number of intervals.

Table IV. Sample sizes of income distributions broken down by relatively homogeneous categories

Gender	Set	Year						
		1992	1996	2002	2004	2005	2006	2007
Gender	Men	12,785	21,590	5,870	3,203	5,456	7,151	8,322
	Women	3,448	6,558	2,103	1,148	2,027	2,524	2,972
Country	Czech Republic	16,233	28,148	7,973	4,351	7,483	9,675	11,294
	Bohemia	9,923	22,684	5,520	2,775	4,692	6,086	7,074
	Moravia	6,310	5,464	2,453	1,576	2,791	3,589	4,220
Social group	Lower employee	4,953	4,963	1,912	1,068	1,880	2,385	2,811
	Self-employed	932	1,097	740	391	649	802	924
	Higher employee	3,975	4,248	2,170	1,080	1,768	2,279	2,627
	Pensioner with s EA	685	594	278	178	287	418	493
	Pensioner without EA	4,822	4,998	2,533	1,425	2,577	3,423	4,063
	Unemployed	189	135	172	131	222	258	251
Municipality size	0–999 inhabitants	2,458	3,069	999	727	1,164	1,607	1,947
	1,000–9,999 inhabitants	4,516	4,471	2,300	1,233	2,297	3,034	3,511
	10,000–99,999 inhabitants	5,574	5,755	2,401	1,508	2,655	3,347	3,947
	100,000 and more inhabitants	3,685	2,853	2,273	883	1,367	1,687	1,889
Age	To 29 years	1,680	2,809	817	413	627	649	827
	From 30 to 39 years	3,035	4,718	1,398	716	1,247	1,620	1,655
	From 40 to 49 years	3,829	6,348	1,446	738	1,249	1,609	1,863
	From 50 to 59 years	2,621	5,216	1,642	919	1,581	2,051	2,391
	From 60 years	5,068	9,057	2,670	1,565	2,779	3,746	4,558
Education	Primary	9,302	15,891	3,480	553	940	1,183	1,385
	Secondary	4,646	3,172	2,493	3,186	5,460	7,168	8,371
	Complete secondary	1,951	6,356	1,129	118	282	266	319
	Tertiary	334	2,729	871	494	801	1,058	1,219

Source: Own research

However, the question of the appropriateness of the model curve for income distribution is not quite common mathematical and statistical issue in which we test the null hypothesis

H_0 : The sample comes from the assumed theoretical distribution

against the alternative hypothesis

H_1 : non H_0 ,

because in goodness of fit tests in the case of income distribution we meet frequently with the fact that we work with large sample sizes and therefore the tests would almost always lead to the rejection of the null hypothesis. This results not only from the fact that with such large sample sizes the power of the test is so high at the chosen significance level that the test uncovers all the slightest deviations of the actual income distribution and a model, but it also results from the principle of construction of the test. But practically we are not interested in such small deviations, so only gross agreement of the model with reality is sufficient and we so called “borrow” the model (curve). Test criterion χ^2 can be used in that direction only tentatively. When evaluating the suitability of the model we proceed to a large extent subjective and we rely on experience and logical analysis.

Method of TL-moments provided the most accurate results in almost all cases, with the negligible exceptions. Method of

L-moments results as the second in more than half of the cases, although the differences between the method of L-moments and maximum likelihood method are not distinctive enough to turn in the number of cases where the method of L-moments came out better than maximum likelihood method. Table V is a typical representative of the results for all 168 income distributions. This table provides the results for the total household sets in the Czech Republic. It contains the estimated values of the parameters of three-parametric lognormal distribution, which were obtained simultaneously using the method of TL-moments, method of L-moments and maximum likelihood method, and the value of test criterion (65). This is evident from the values of the criterion that the method of L-moments brought more accurate results than maximum likelihood method in four of seven cases. The most accurate results were obtained using the method of TL-moments in all seven cases.

The estimation of the value of the parameter θ (beginning of the distribution, theoretical minimum) obtained using the maximum likelihood method is negative in 1992 and 2005–2007. This mean that three-parametric lognormal curve gets into negative values initially its course in terms of income. Since at first the curve has very tight contact with the horizontal axis, it does not matter good agreement of model with real distribution.

Table V. Parameter estimations of three-parametric lognormal curves obtained using three various robust methods of point parameter estimation and the value of χ^2 criterion

Year	Method of TL-moments			Method of L-moments			Maximum likelihood method		
	μ	σ^2	θ	μ	σ^2	θ	μ	σ^2	θ
1992	9.722	0.521	14,881	9.696	0.700	14,491	10.384	0.390	-325
1996	10.334	0.573	25,981	10.343	0.545	25,362	10.995	0.424	52.231
2002	10.818	0.675	40,183	10.819	0.773	37,685	11.438	0.459	73.545
2004	10.961	0.552	39,899	11.028	0.675	33,738	11.503	0.665	7.675
2005	11.006	0.521	40,956	11.040	0.677	36,606	11.542	0.446	-8.826
2006	11.074	0.508	44,941	11.112	0.440	40,327	11.623	0.435	-42.331
2007	11.156	0.472	48,529	11.163	0.654	45,634	11.703	0.421	-171.292
Year	Criterion χ^2			Criterion χ^2			Criterion χ^2		
1992	739.512			811.007			1,227.325		
1996	1,503.878			1,742.631			2,197.251		
2002	998.325			1,535.557			1,060.891		
2004	494.441			866.279			524.478		
2005	731.225			899.245			995.855		
2006	831.667			959.902			1,067.789		
2007	1,050.105			1,220.478			1,199.035		

Source: Own research

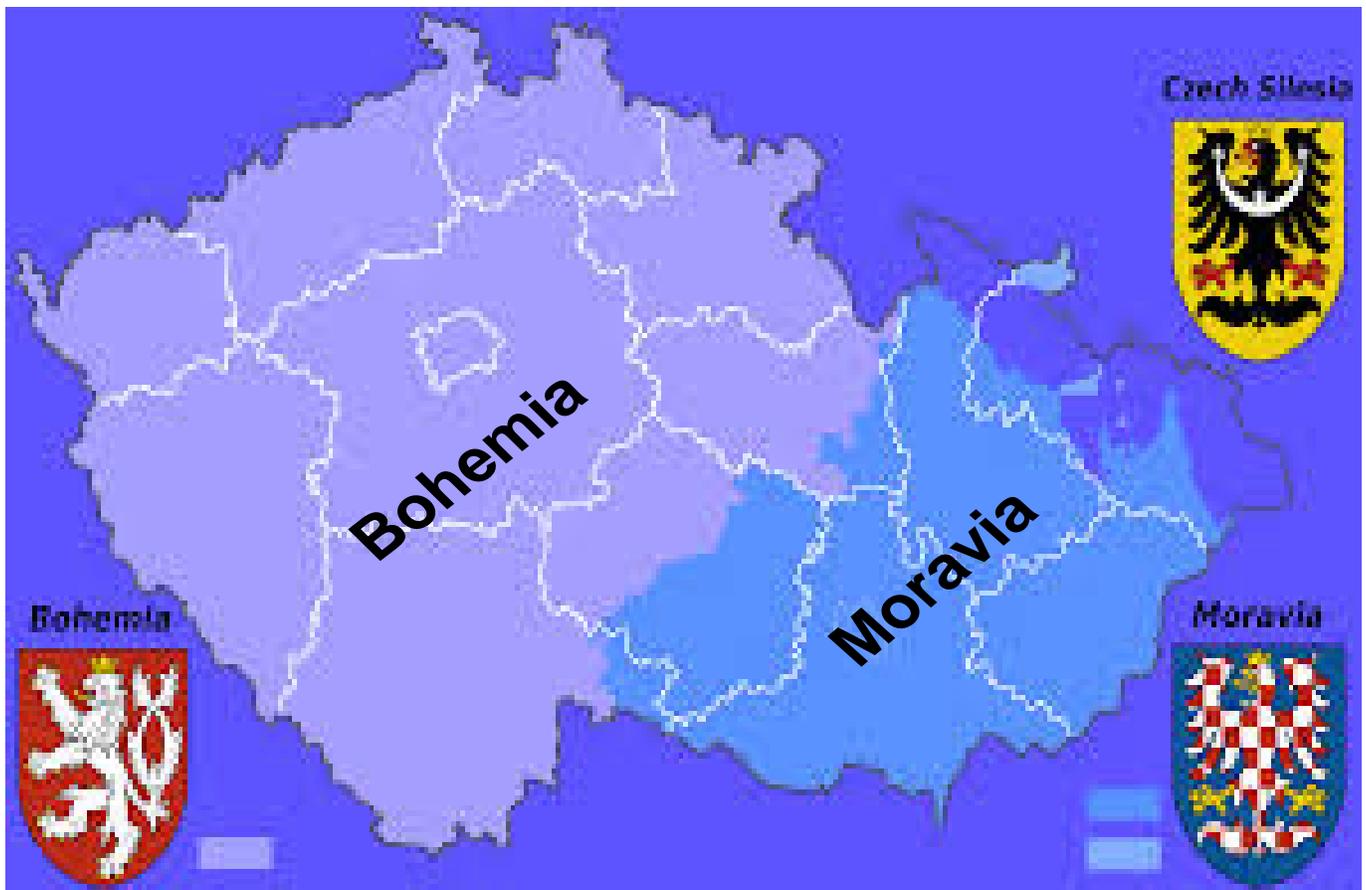


Fig. 1. Map of the Czech Republic (Bohemia and Moravia)

Source: www.google.cz

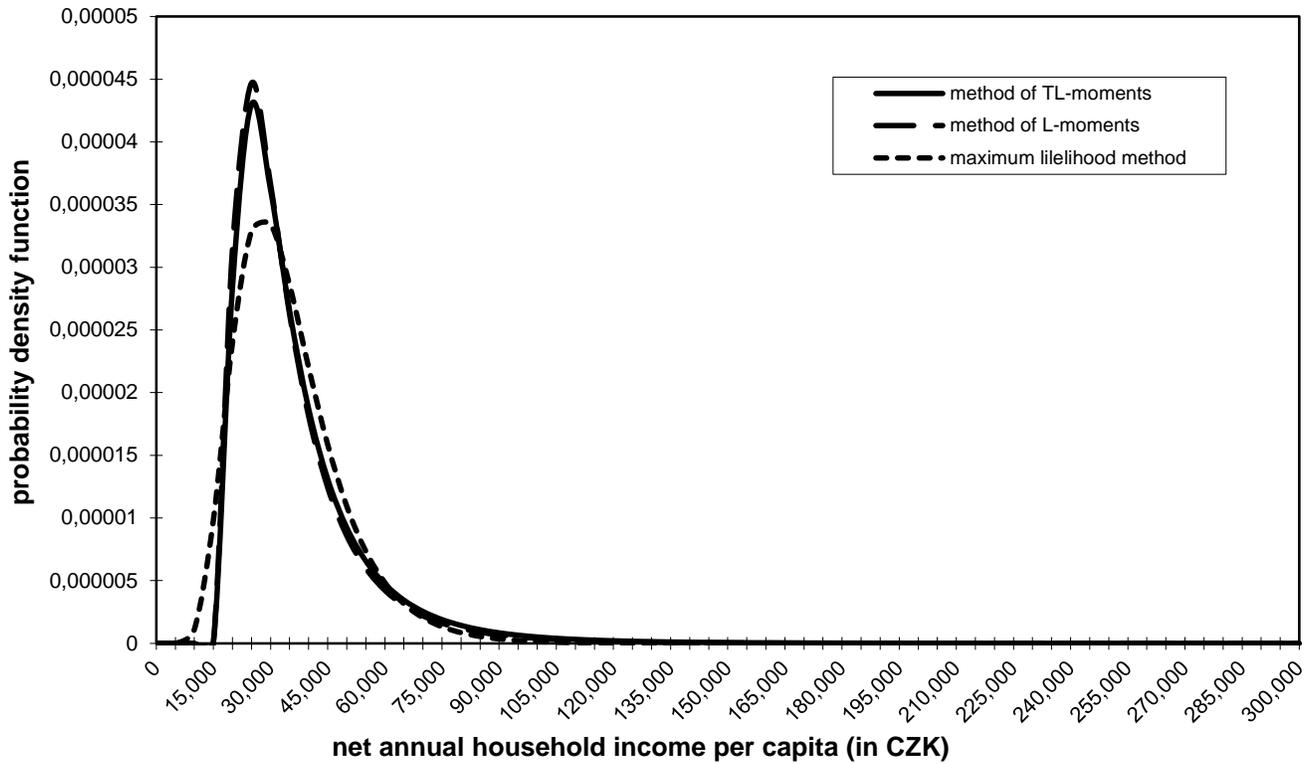


Fig. 2. Model probability density functions of three-parametric lognormal curves in 1992 with parameters estimated using three various robust methods of point parameter estimation

Source: Own research

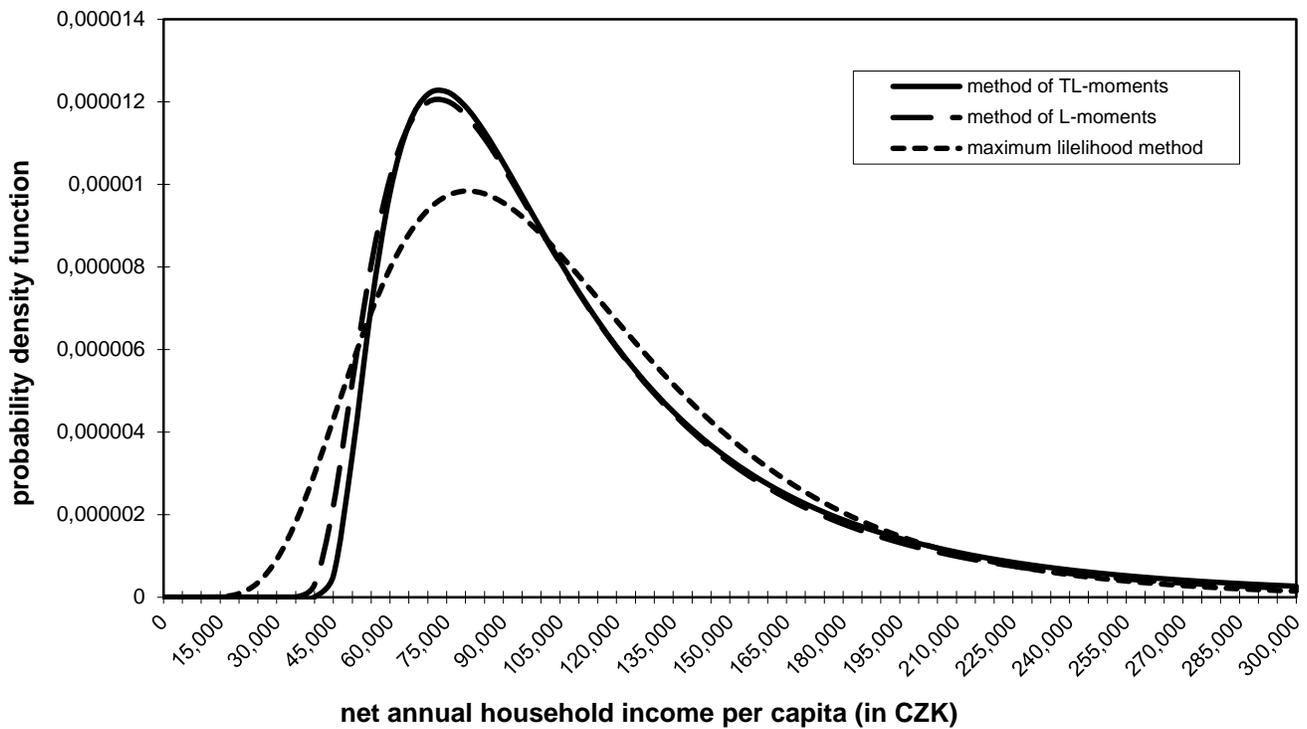


Fig. 3. Model probability density functions of three-parametric lognormal curves in 2004 with parameters estimated using three various robust methods of point parameter estimation

Source: Own research

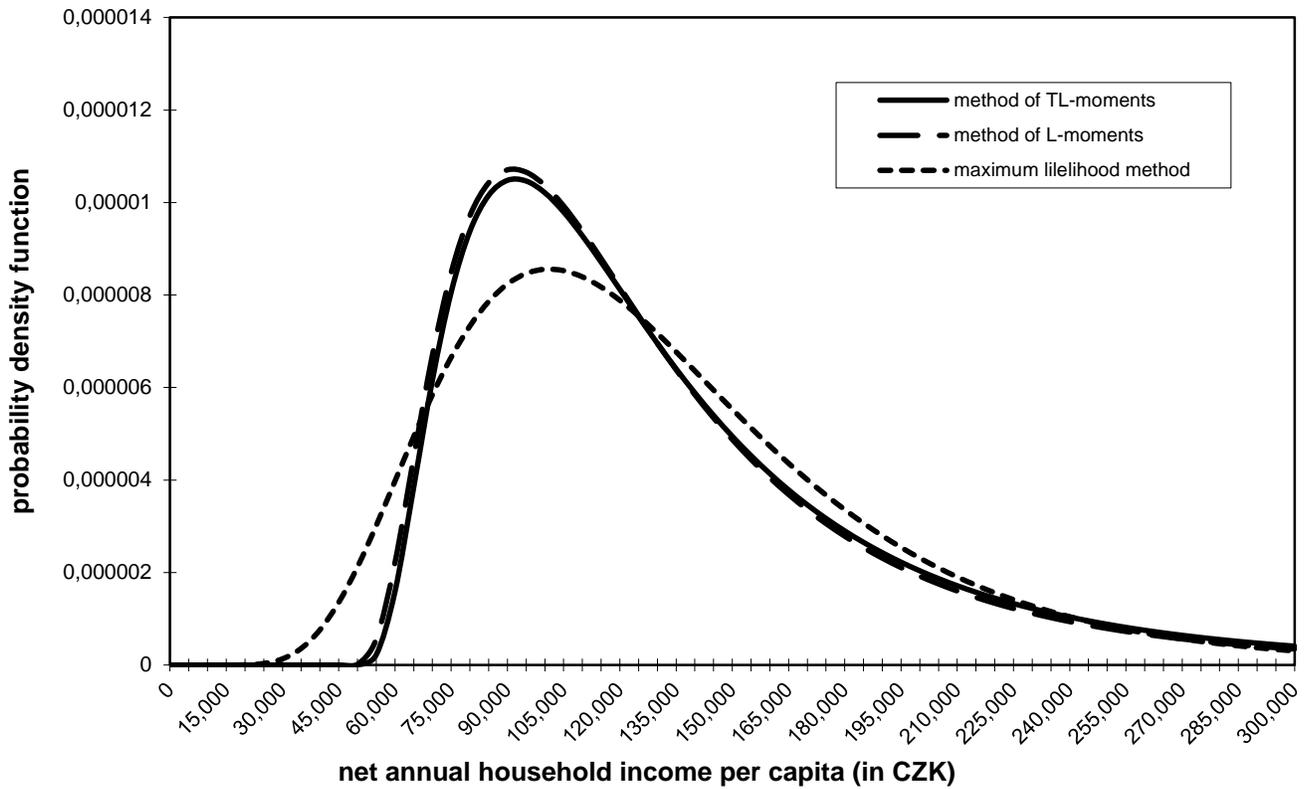


Fig. 4. Model probability density functions of three-parametric lognormal curves in 2007 with parameters estimated using three various robust methods of point parameter estimation

Source: Own research

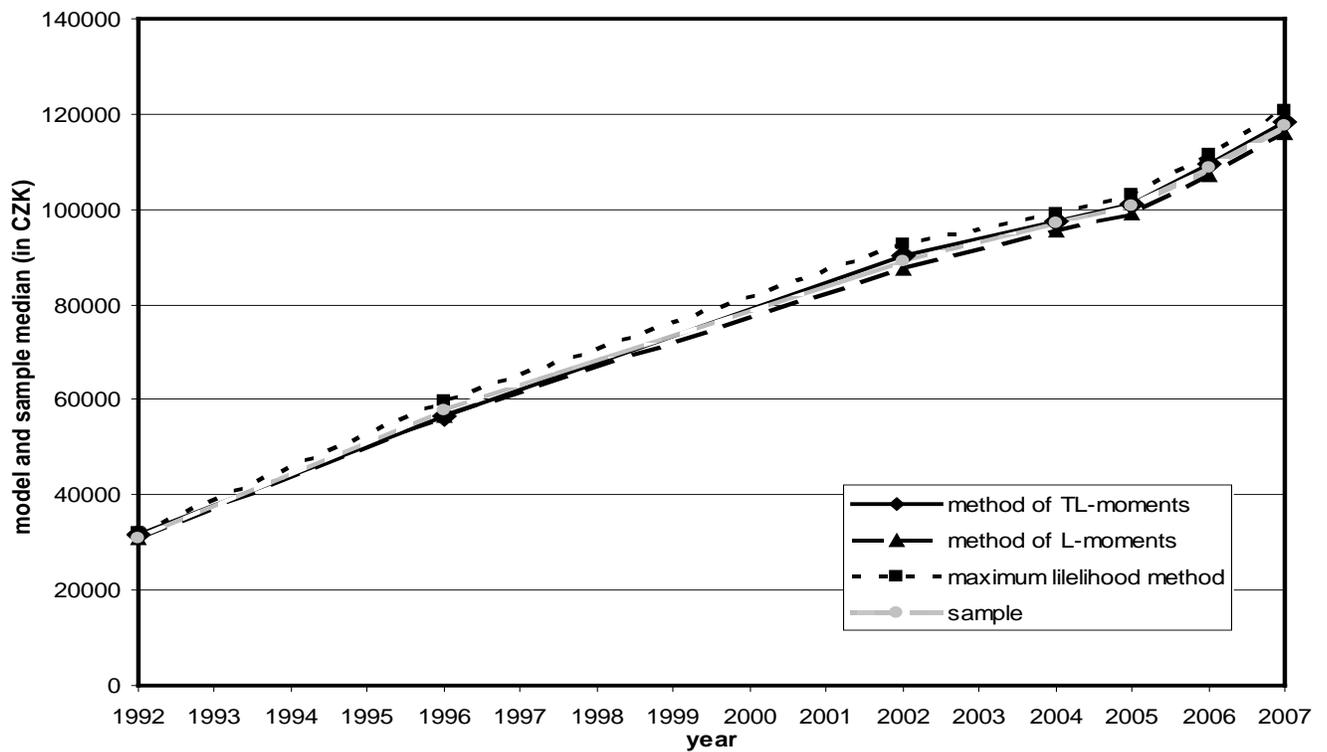


Fig. 5. Development of model and sample median of net annual household income per capita (in CZK)

Source: Own research

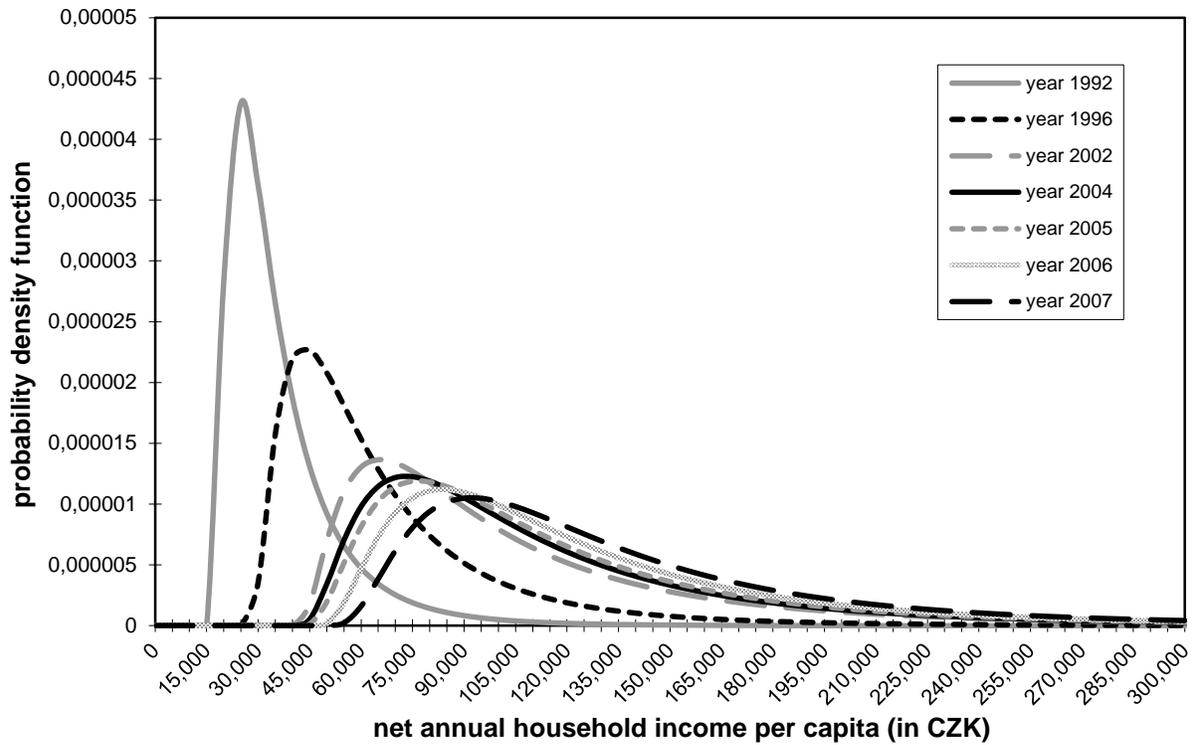


Fig. 6. Development of probability density function of three-parameter lognormal curves with parameters estimated using the method of TL-moments

Source: Own research

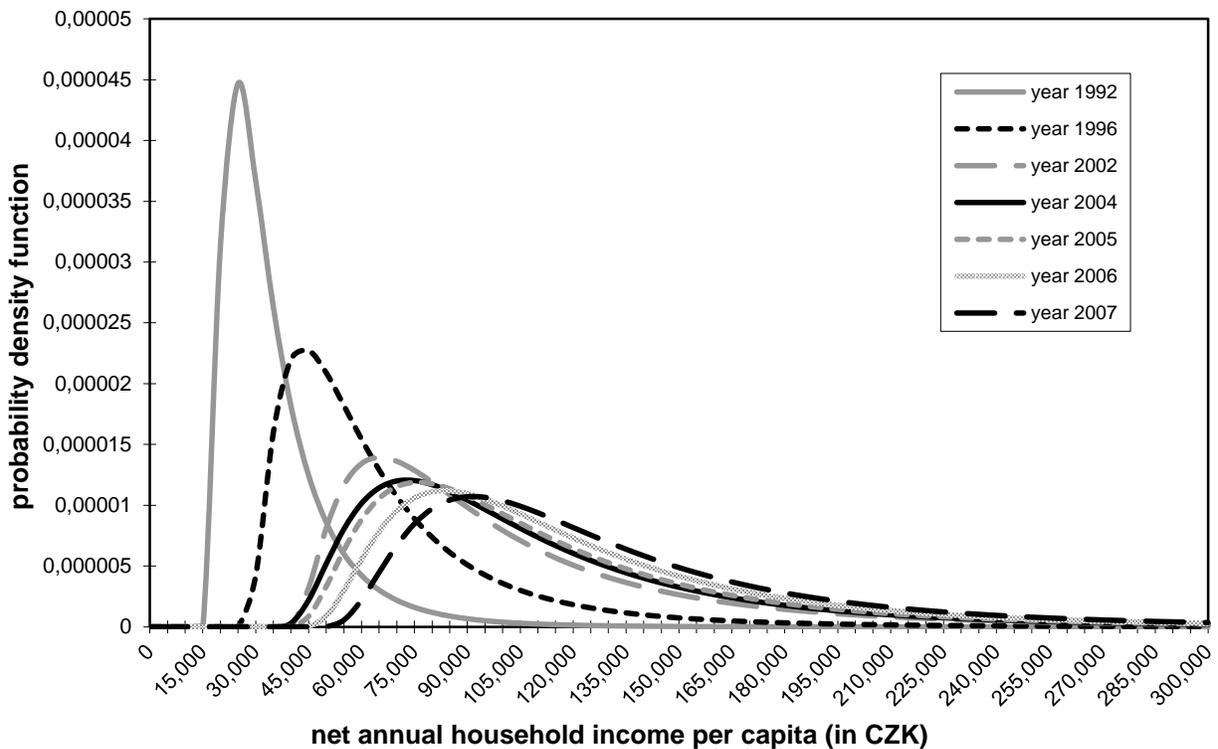


Fig. 7. Development of probability density function of three-parameter lognormal curves with parameters estimated using the method of L-moments

Source: Own research

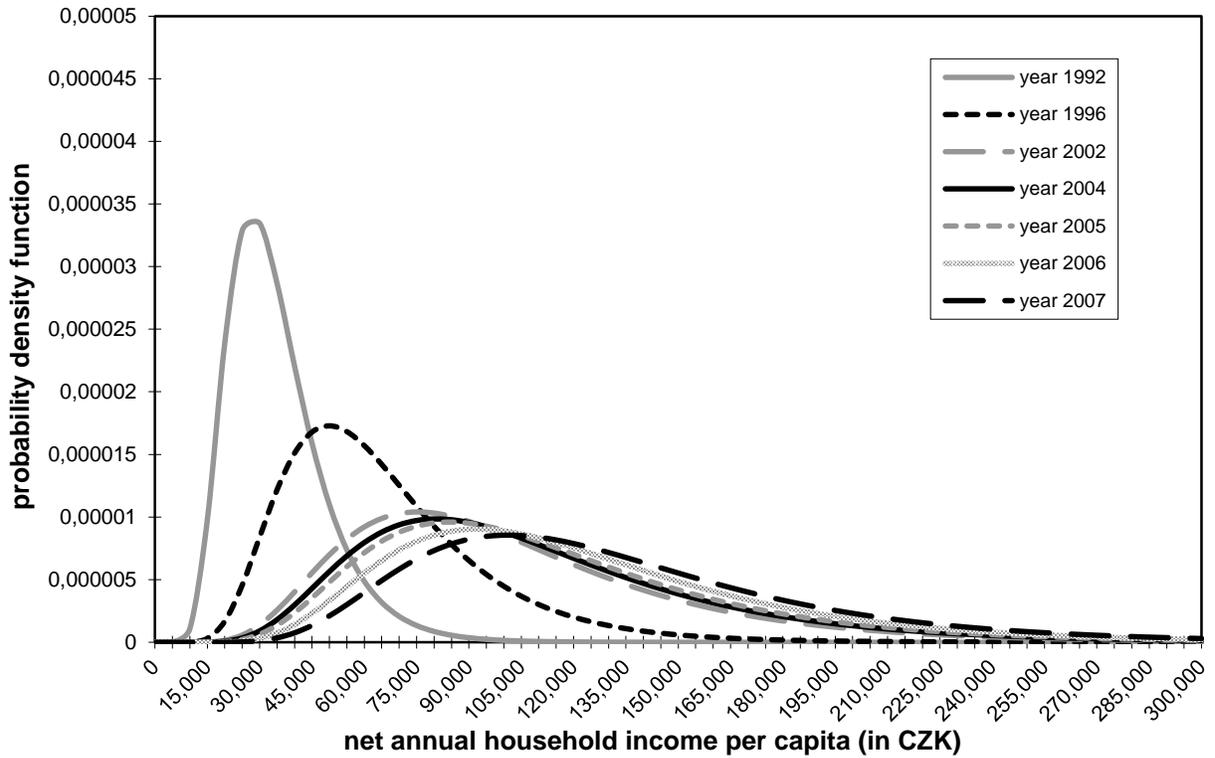


Fig. 8. Development of probability density function of three-parameter lognormal curves with parameters estimated using the maximum likelihood method

Source: Own research

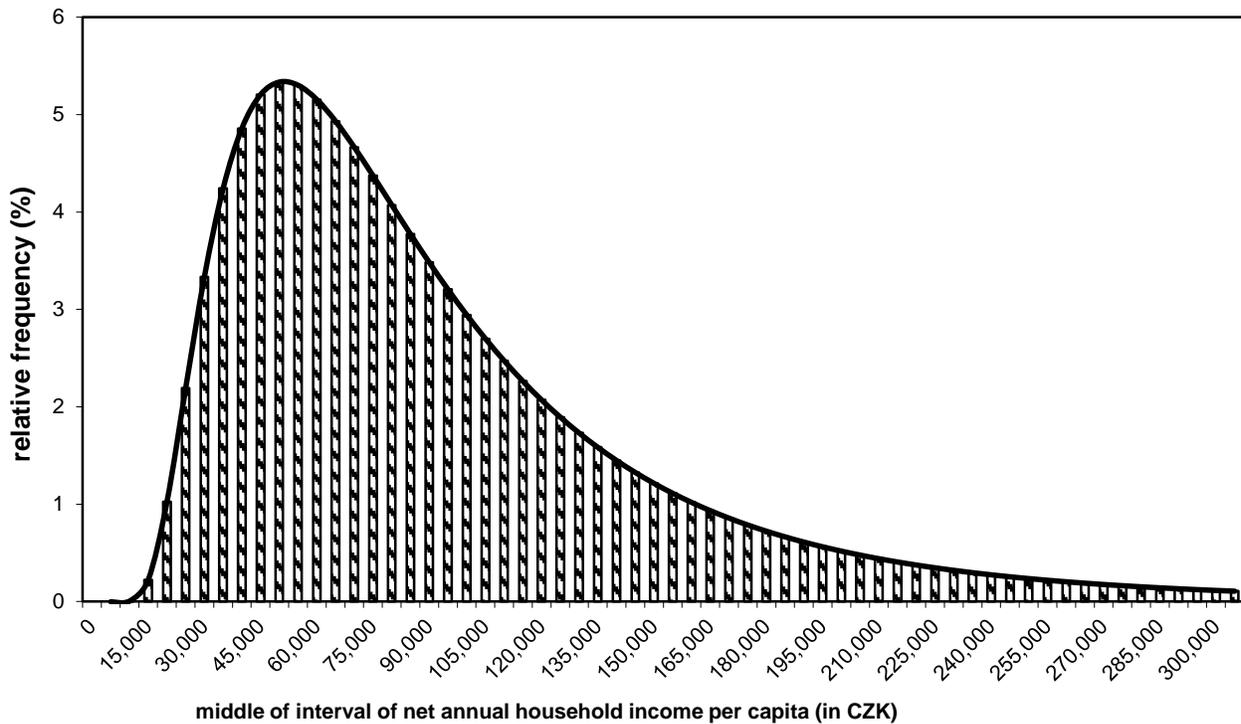


Fig. 9. Model ratios of employees by the band of net annual household income per capita with parameters of three-parametric lognormal curves estimated by the method of TL-moments in 2007

Source: Own research

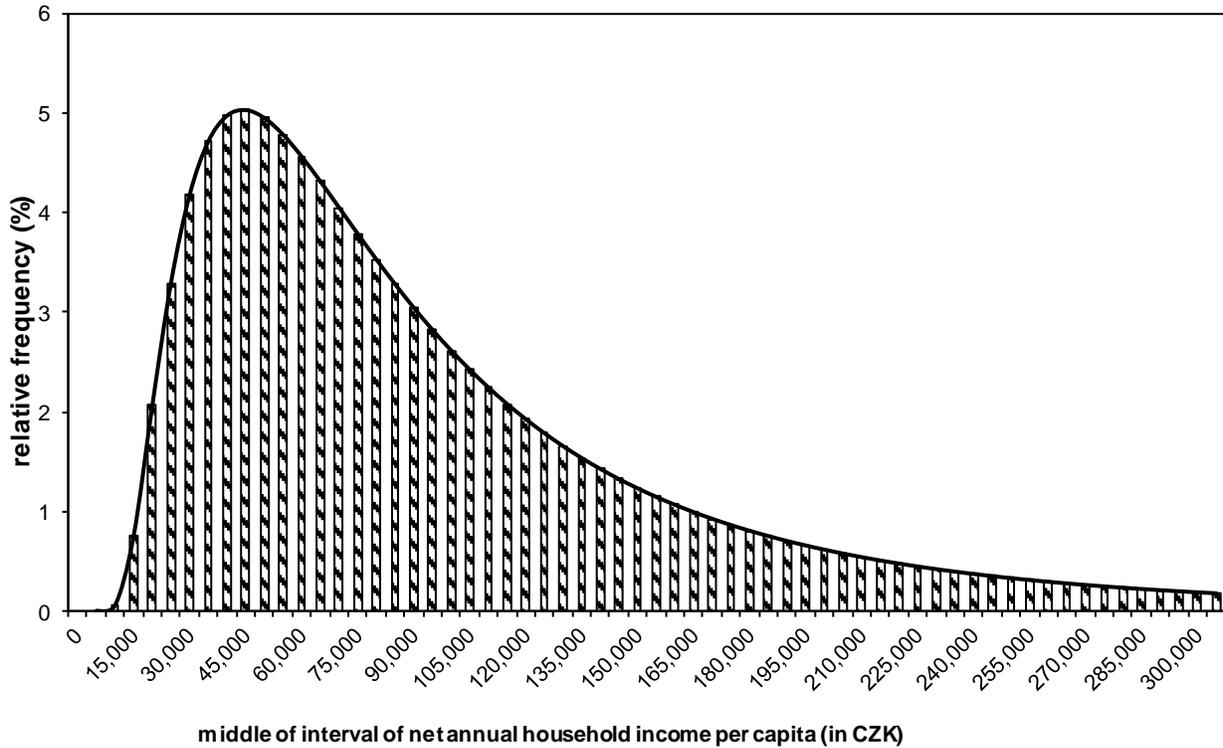


Fig. 10. Model ratios of employees by the band of net annual household income per capita with parameters of three-parametric lognormal curves estimated by the method of L-moments in 2007

Source: Own research

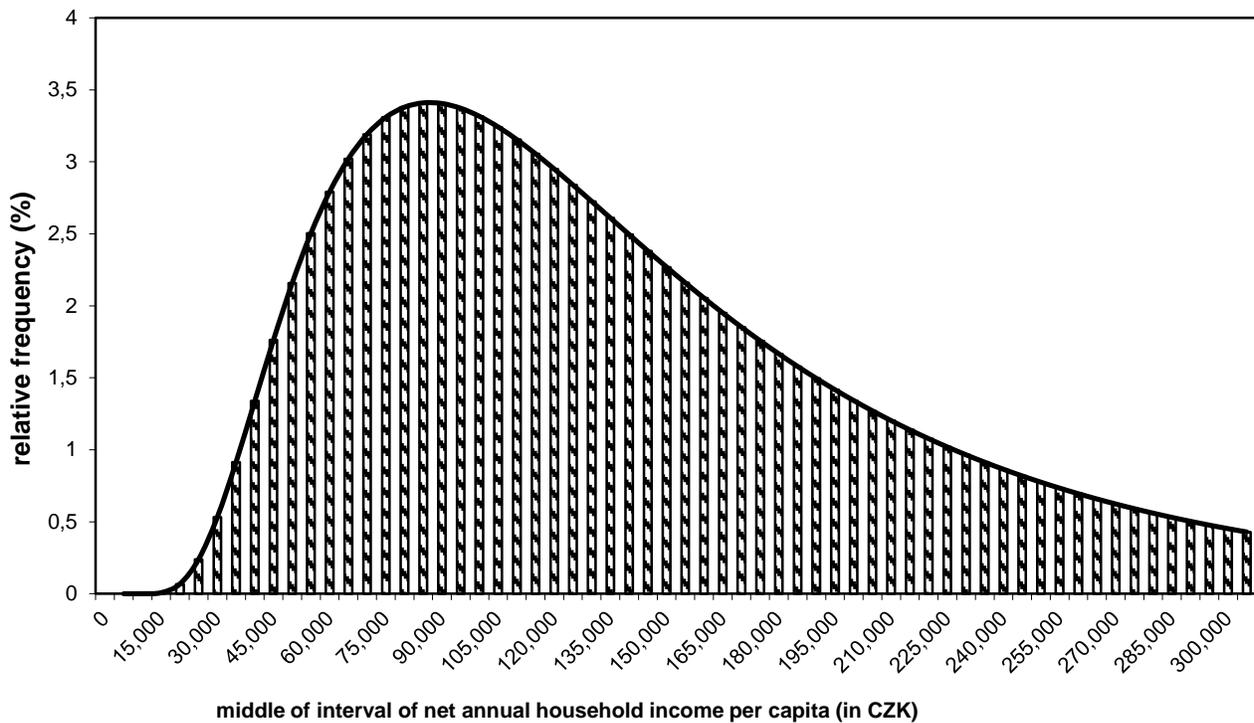


Fig. 11. Model ratios of employees by the band of net annual household income per capita with parameters of three-parametric lognormal curves estimated by the maximum likelihood method in 2007

Source: Own research

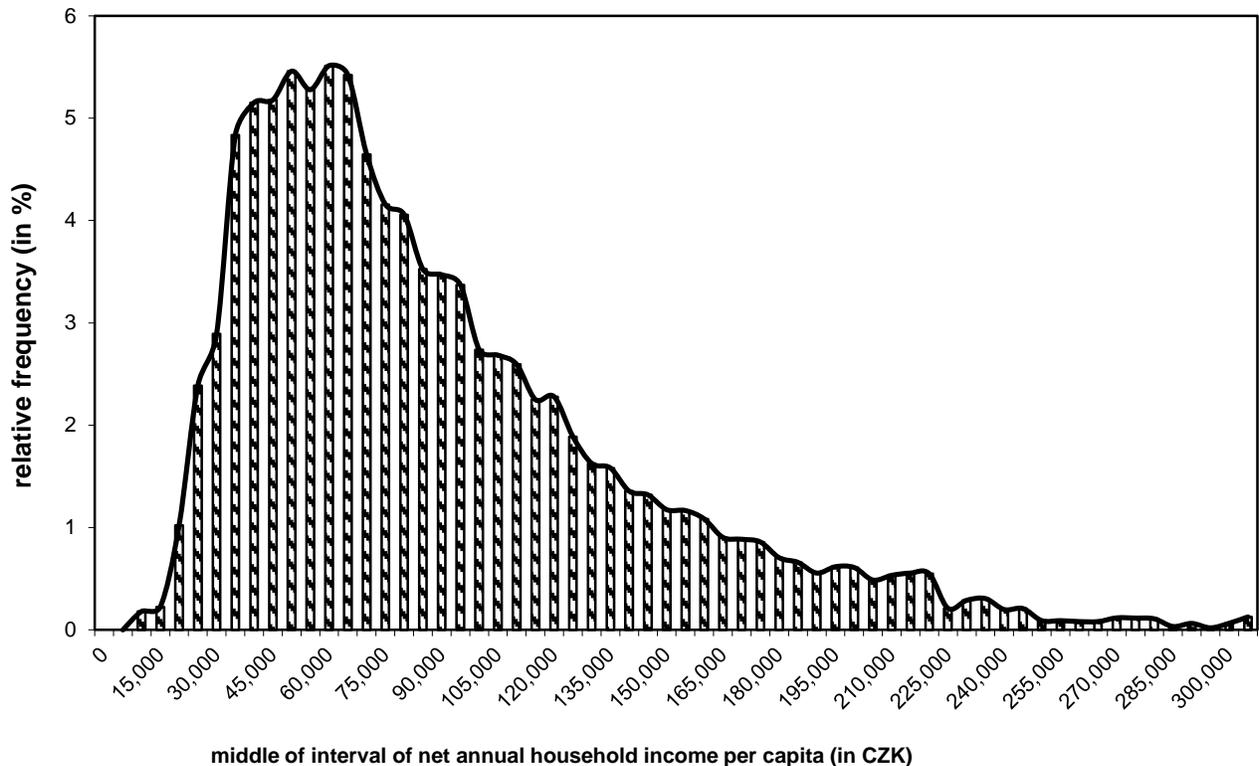


Fig. 12. Sample ratios of employees by the band of net annual household income per capita in 2007

Source: Own research

Figs. 2–4 allow the comparison of these methods in terms of model probability density functions in choosing years (1992, 2004 and 2007) for the total set of households throughout the Czech Republic together. It should be noted at this point that other scale is on the vertical axis in Fig. 2 than in Figs. 3 and 4 for better legibility, because income distribution just after the transformation of the Czech economy from a centrally planned to a marked economy still showed different behaviour (lower level and variability, higher skewness and kurtosis) than the income distribution closer to the present. It is clear from these three figures that the methods of TL-moments and L-moments bring the very similar results, while the probability density function with the parameters estimated by maximum likelihood method is very different from model probability density functions constructed using the method of TL-moments and the method of L-moments.

Fig. 5 also provides some comparison of the accuracy of these three methods of point parameter estimation. It represents the development of the sample median and the theoretical medians of lognormal distribution with parameters estimated using the method of TL-moments, method of L-moments and maximum likelihood method in the researched period again for the total set of households of the Czech Republic. It is also clear from this figure that the curve representing the course of the theoretical medians of lognormal distribution with parameters estimated by methods of TL-moments and L-moments are more tightly to the curve

showing the course of sample median compared with the curve representing the development of theoretical median of lognormal distribution with parameters estimated by maximum likelihood method.

Figs. 6–8 show the development of the model probability density functions of three-parametric lognormal distribution again with parameters estimated using three researched methods of parameter estimation in the analysed period for total set of households of the Czech Republic. Also, in view of these figures income distribution in 1992 shows a strong difference from the income distributions in next years. Also here, we can observe a certain similarity of the results taken using the methods of TL-moments and L-moments and considerable divergence of the results obtained using these two methods of point parameter estimation from the results obtained using the maximum likelihood method.

Figs. 9–11 then represent the model relative frequencies (in %) of employees by the band of net annual household income per capita in 2007 obtained using three-parametric lognormal curves with parameters estimated by the method of TL-moments, method of L-moments and maximum likelihood method. These figures also allow some comparison of the accuracy of the researched methods of point parameter estimation compared with Fig. 12, where are the really observed relative frequencies in individual income intervals obtained from a sample.

IV. CONCLUSION

Relatively new class of moment characteristics of probability distributions were here introduced. There are the characteristics of location (level), variability, skewness and kurtosis of probability distributions constructed using L-moments and TL-moments that are robust extension of L-moments. Own L-moments have been introduced as a robust alternative to classical moments of probability distributions. However, L-moments and their estimations lack some robust features that belong to TL-moments.

Sample TL-moments are linear combinations of sample order statistics, which assign zero weight to a predetermined number of sample outliers. Sample TL-moments are unbiased estimations of the corresponding TL-moments of probability distributions. Some theoretical and practical aspects of TL-moments are still the subject of research or they remain for future research. Efficiency of TL-statistics depends on the choice of α , for example, $l_1^{(0)}, l_1^{(1)}, l_1^{(2)}$ have the smallest variance (the highest efficiency) among other estimations for random samples from normal, logistic and double exponential distribution.

The above methods can be also used for modeling the wage distribution or other analysis of economic data (among other methods, see for example [16] or [17]).

REFERENCES

- [1] K. Adamowski, "Regional Analysis of Annual Maximum and Partial Duration Flood Data by Nonparametric and L-moment Methods," *Journal of Hydrology*, vol. 229, no. 3–4, pp. 219–231, 2000.
- [2] D. Bílková, "Modelling of Wage Distributions Using the Method of L-moments (Published Conference Proceedings style)," in *Proceedings of Conference of AMSE [CD]*, Demänovská Dolina, Slovakia, 2010, pp. 16–30.
- [3] D. Bílková, "Use of the L-Moments Method in Modeling the Wage Distribution (Published Conference Proceedings style)," in *Proceedings of Conference of APLIMAT [CD]*, Bratislava, Slovakia, 2011, pp. 1471–1481.
- [4] D. Bílková, "L-Moments and Their Use in Modeling the Distribution of Income and Wage (Published Conference Proceedings style)," in *Proceedings of Conference of ISI [flashdisk]*, Dublin, Ireland, 2011, pp. 1–6.
- [5] D. Bílková, "Modeling of Income and Wage Distribution Using the Method of L-Moments of Parameter Estimation (Published Conference Proceedings style)," in *Proceedings of Proceedings of Conference of International Days of Statistics and Economics at VŠE [CD]*, Prague, Czech Republic, 2011, pp. 1–10.
- [6] D. Bílková, "Three-Parametric Lognormal Distribution and Estimating Its Parameters Using the Method of L-Moments (Published Conference Proceedings style)," in *Proceedings of Conference of RELIK [CD]*, Prague, Czech Republic, 2011, not pages numbered.
- [7] D. Bílková, "Estimating Parameters of Lognormal Distribution Using the Method of L-Moments," *Research Journal of Economics, Business and ICT*, vol. 4, no. 1, pp. 4–9, 2011.
- [8] D. Bílková, "Modelling of Wage and Income Distributions Using the Method of L-Moments," *Journal of Mathematics and System Science*, vol. 2, no. 1, pp. 13–19, 2012.
- [9] D. Bílková, "Lognormal Distribution and Using L-Moment Method for Estimating Its Parameters," *International Journal of Mathematical Models and Methods in Applied Sciences [online]*, vol. 6, no. 1, pp. 30–44, 2012.
URL: <http://www.naun.org/journals/m3as/17-079.pdf>.
- [10] D. Bílková, "Lognormal Distribution Parameter Estimating Using L-Moments," *Journal of Mathematics and Technology*, vol. 3, no. 1, pp. 33–51, 2012.
- [11] D. Bílková and I. Malá, "Application of the L-Moment Method when Modelling the Income Distribution in the Czech Republic," *Austrian Journal of Statistics*, vol. 41, no. 2, pp. 125–132, 2012.
- [12] E. A. H. Elamir, and A. H. Seheult, "Trimmed L-moments," *Computational Statistics & Data Analysis*, vol. 43, no. 3, pp. 299–314, 2003.
- [13] J. R. M. Hosking, "L-moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics," *Journal of the Royal Statistical Society (Series B)*, vol. 52, no. 1, pp. 105–124, 1990.
- [14] J. Kyselý and J. Píček, "Regional Growth Curves nad Improved Design Value Estimates of Extreme Precipitation Events in the Czech Republic," *Climate research*, vol. 33, no. 3, pp. 243–255, 2007.
- [15] T. Löster and J. Langhamrová, "Analysis of Long-Term Unemployment in the Czech Republic (Published Conference Proceedings style)," in *Proceedings of Conference of International Days of Statistics and Economics at VŠE [CD]*, Prague, Czech Republic, 2011, pp. 228–234.
- [16] L. Marek, "The trend of income distributions in Czech Republic in the years 1995–2008 analysis," *Politická ekonomie*, vol. 58, no. 8, pp. 186–206, 2010.
- [17] T. J. Ulrych, D. R. Velis, A. D. Woodbury and M. D. Sacchi, "L-moments and C-moments," *Stochastic Environmental Research and Risk Assessment*, vol. 14, no. 1, pp. 50–68, 2000.

doc. Ing. Diana Bílková, Dr. was born in May 1969 in Jilemnice, the Czech Republic. She graduated in statistics at the Faculty of Informatics and Statistics of the University of Economics, Prague (1992). Since then she has been teaching at the Department of Statistics and Probability of the same UE faculty. On receiving the doctor's degree in statistics (1996), she was appointed associate professor in 2013. In her research, she focuses on the theory of probability and mathematical statistics, particularly modeling, development analysis and predictions of the wage and income distribution. She has authored or co-authored more than a hundred of research papers both in Czech and foreign journals, compiled university textbooks, reviews and external examiner's reports. She has dozens of citation responses in domestic and foreign publications, those indexed on the Web of Science representing approximately a third of them. She participated in eight research projects, being a research team leader in two of them. She is a member of the Czech Statistical Society and a committee member of "International Days of Statistics and Economics" conference monitored in the "Conference Proceedings Citation Index". She has successfully supervised numerous statistics graduate theses.

Stability breakdown along a line of equilibria in nonlinear circuits with memristors

Ricardo Riaza

Abstract—The design in 2008 of a device with a memristive characteristic has had a great impact in electronics, specially at the nanometer scale. This device, whose existence was predicted by Leon Chua in 1971 for symmetry reasons, is governed in a flux-controlled setting by a relation of the form $i = W(\varphi)v$, and systematically leads to the presence of non-isolated equilibria. In this communication we examine how the stability of such manifolds of equilibria may break down when normal hyperbolicity fails. This phenomenon may be due to the transition of an eigenvalue either through the origin or through infinity. Our approach is a graph-theoretic one, aiming at the analysis of such phenomena in terms of the topology of the digraph underlying the circuit.

Keywords—Nonlinear circuit, memristor, equilibrium, stability, normal hyperbolicity, bifurcation.

I. INTRODUCTION

MEMRISTORS are electronic devices defined by a charge-flux characteristic. Their existence was predicted for symmetry reasons by Leon Chua in 1971, since resistors, capacitors and inductors are defined by voltage-current, charge-voltage and flux-current relations, respectively. The charge-flux characteristic was the only one lacking in this set of relations, since the charge-current and the flux-voltage pairs are related by the electromagnetic laws $q' = i$, $\varphi' = v$. The design of such a *memory-resistor* or *memristor* at the nanometer scale announced by an HP team in 2008 [1] has driven a lot of attention to these devices. The flux-charge relation may have either a charge-controlled form $\varphi = \phi(q)$ or a flux-controlled one $q = \sigma(\varphi)$ [2]. We will focus on the latter but dual results apply to the charge-controlled case.

Applications of memristors and other memory devices are being reported in many fields: see [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] and references therein. In particular, a significant impact in industry is expected to happen in the near future because of the use of memristors in non-volatile memory design. Not only from the point of view of applications but also from a mathematical perspective this device poses challenging problems. In this communication we focus on stability problems related to the systematic presence of manifolds of non-isolated equilibria in circuits including this device. These problems will be addressed in Section III, after compiling some introductory material on Section II. Finally, some concluding remarks can be found on Section II.

R. Riaza is with the Departamento de Matemática Aplicada a las TIC, ETSI Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain. ricardo.riaza@upm.es

II. MEMRISTIVE CIRCUITS

A. The memristor

As indicated above, a flux-controlled memristor is defined by a nonlinear, differentiable relation

$$q = \sigma(\varphi);$$

time derivation yields, by means of the identities $q' = i$, $\varphi' = v$, the current-voltage characteristic

$$i = W(\varphi)v, \quad (1)$$

where $W(\varphi) = \sigma'(\varphi)$ is the *memductance*. The dual case is defined by a flux-charge relation $\varphi = \phi(q)$ which yields a voltage-current characteristic of the form

$$v = M(q)i, \quad (2)$$

where $M(q) = \phi'(q)$ is the so-called *memristance*. Note that (2) is reminiscent of Ohm's law, but the "resistance" $M(q)$ now depends on the charge q , which is the time-integral of the current i ; for this reason the device' characteristic keeps track of its own history. The name *memristor*, which is an abbreviation of *memory-resistor*, comes from this remark [2]. Similar remarks apply to the flux-controlled case defined by (1); this form will be assumed throughout the document.

B. Branch-oriented modelling of memristive circuits

For the sake of simplicity we will focus the attention on a restricted class of memristive circuits, just including flux-controlled memristors, voltage-controlled resistors, and capacitors. We will refer to these as WGC-circuits. Dual devices (charge-controlled memristors, current-controlled resistors, and inductors) as well as voltage and current sources are precluded in order to keep the discussion as simple as possible, but the results may be proved to hold in general. The essential mathematical aspects of the discussion are already present in WGC-circuits. To focus on cases with a one-dimensional manifold (that is, a line) of equilibria we will further assume that the circuit has a unique memristor.

Such circuits can be described by the differential-algebraic model (cf. [18], [19])

$$\varphi'_m = v_m \quad (3)$$

$$C(v_c)v'_c = i_c \quad (4)$$

$$0 = B_m v_m + B_c v_c + B_r v_r \quad (5)$$

$$0 = Q_m W(\varphi_m) v_m + Q_c i_c + Q_r g(v_r). \quad (6)$$

Here the subscripts m , c , r correspond to memristors, capacitors and resistors; $C(v_c)$ is the incremental capacitance matrix,

and $i_r = g(v_r)$ is the current-voltage characteristic of resistors, which is assumed to be differentiable; the incremental conductance matrix is then $G(v_r) = g'(v_r)$. Note that (3)-(6) is a branch-oriented model (cf. [18]) which uses a description of Kirchhoff laws in the form $Bv = 0$, $Qi = 0$ in terms of reduced loop and cutset matrices B and Q . The columns of these matrices are split according to the nature of the different devices, so that $B = (B_m \ B_c \ B_r)$, $Q = (Q_m \ Q_c \ Q_r)$ (find details in [18], [20]).

Working scenario. Both $C(v_c)$ and $G(v_r)$ are assumed to be positive definite everywhere; in circuit-theoretic terms, this means that capacitors and resistors are strictly locally passive. We also assume that $g(0) = 0$, and focus the analysis on the line of equilibria defined by the vanishing of the right-hand side of (3)-(6) when $v_m = i_c = v_c = v_r = 0$, in order to examine the qualitative behavior of the system as the variable φ_m changes along this line. Specifically, we will assume that $W(0) = 0$ and $W'(0) \neq 0$, so that the memristor becomes active as φ_m undergoes the null value. Recall that a memristor is said to be strictly locally passive (resp. active) at a given value of φ if $W(\varphi) > 0$ (resp. $W(\varphi) < 0$).

The vanishing of W may lead to the loss of normal hyperbolicity of the line of equilibria described above. An m -dimensional manifold of equilibrium points in an n -dimensional system is said to be *normally hyperbolic* if $n-m$ eigenvalues of the linearization are away from the imaginary axis [21]: note that m eigenvalues necessarily vanish because of the presence of an m -dimensional manifold of equilibria. In our context, depending on the topology of the circuit and, specifically, on the location of the memristor, the vanishing of the memductance W may result in the loss of normal hyperbolicity and different bifurcation phenomena may follow. Some of these phenomena are addressed in the main results reported in this communication, which can be found in Section III below.

III. STABILITY BREAKDOWN

A. Double zero eigenvalue

The linearization of a dynamical system along a line of equilibria obviously displays a null eigenvalue. When a second eigenvalue undergoes the origin, a *transcritical bifurcation without parameters* occurs generically [22], [23], [24]. If the remaining eigenvalues have negative real parts, this implies that the line of equilibria experiences a loss of stability in the region where the bifurcating eigenvalue becomes positive (find details in the references just cited). We discuss below certain circuit-theoretic conditions which characterize this phenomenon for the set of circuits presented above.

Proposition 1. *Consider the system (3)-(6) in the working scenario described above. If the circuit has a unique WC-cutset which actually includes the memristor, and there are neither C-loops nor C-cutsets, then the null eigenvalue of the linearization of (3)-(6) along the line of equilibria becomes*

a double one at the origin. This corresponds to a second eigenvalue which crosses the origin and becomes positive as W becomes negative (that is, as the memristor becomes strictly locally active) when φ varies. The remaining eigenvalues have negative real parts, and therefore this transition implies the loss of stability of the equilibrium line when W becomes negative.

Both the statement and the proof of this result make use of some notions and properties of digraph theory which we compile in what follows. Find detailed introductions to digraph theory and its use in circuit analysis in [18], [19], [20], [25], [26], [27] A *loop* or *cycle* in a directed graph (or digraph) is the set of branches in a closed path without self-intersections. A *cutset* K is a set of branches whose removal increases the number of connected components of the digraph, and which is minimal with respect to this property, that is, the removal of any proper subset of K does not increase the number of components. In a connected digraph, a cutset is just a minimal disconnecting set of branches. The removal of the branches of a cutset increases the number of connected components by exactly one. We assume that the digraph has neither bridges (cutsets defined by a single branch) nor selfloops (loops formed by a unique branch).

Given a set K of branches, we will denote by B_K (resp. Q_K) the submatrix of B (resp. of Q) defined by the columns which correspond to K -branches. The absence of loops or cutsets including only K -devices can be easily characterized in terms of B_K and Q_K ; specifically, K does not include cutsets if and only if B_K has full column rank (i.e. $\ker B_K = \{0\}$) and, analogously, it does not include loops if and only if Q_K has full column rank.

Proof of proposition 1: The linearization of (3)-(6) at a generic equilibrium is defined by the matrix pencil (cf. subsection III-B below) $\lambda A - J$, where

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & C(0) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad (7)$$

and J is the matrix of partial derivatives of the right-hand side of (3)-(6) with respect to the variables $\varphi_m, v_c, v_m, i_c, v_r$, that is,

$$J = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & I_c & 0 \\ 0 & B_c & B_m & 0 & B_r \\ 0 & 0 & Q_m W(\varphi_m) & Q_c & Q_r G(0) \end{pmatrix}. \quad (8)$$

One can easily check that $\det(\lambda A - J)$ reads as

$$\det \begin{pmatrix} \lambda & 0 & -1 & 0 & 0 \\ 0 & \lambda C(0) & 0 & -I_c & 0 \\ 0 & -B_c & -B_m & 0 & -B_r \\ 0 & 0 & -Q_m W(\varphi_m) & -Q_c & -Q_r G(0) \end{pmatrix}, \quad (9)$$

and, for $\varphi = 0$,

$$\det \begin{pmatrix} \lambda & 0 & -1 & 0 & 0 \\ 0 & \lambda C(0) & 0 & -I_c & 0 \\ 0 & -B_c & -B_m & 0 & -B_r \\ 0 & 0 & 0 & -Q_c & -Q_r G(0) \end{pmatrix},$$

since $W(0) = 0$ because of the working assumptions. In this case, $\lambda = 0$ is indeed a double zero eigenvalue because of the fact that

$$\begin{pmatrix} B_c & B_m & B_r \\ 0 & 0 & Q_r G(0) \end{pmatrix} \quad (10)$$

is a singular matrix with a minimal rank deficiency: this is a consequence of the existence of a unique WC-cutset, which makes the kernel of $(B_c \ B_m)$ non-trivial and, actually, one-dimensional. The positive definiteness of the conductance matrix $G(0)$ transfers this minimal rank deficiency to the matrix (10) and this implies that the null eigenvalue is indeed a double one when $\varphi = 0$.

The fact that this second null eigenvalue actually crosses the origin as φ varies follows from the characterization of the transcritical bifurcation without parameters reported in [22], [23], [24]. Skipping technical details for the sake of brevity, this is specifically a consequence of the assumption $W'(0) \neq 0$; note that, together with $W(0) = 0$, this yields a sign change in $W(\varphi)$ as φ undergoes the null value. Owing to the results in [28], for positive values of W (recall that both $G(0)$ and $C(0)$ are positive definite) all non-vanishing eigenvalues have non-positive real parts, one real eigenvalue actually becoming positive as W takes on negative values. Finally, the fact that the remaining eigenvalues are away from the imaginary axis follows from the results discussed in [29], according to which the absence of inductors is enough to guarantee, under the assumed absence of C-loops and C-cutsets, that no purely imaginary eigenvalues are depicted in the linearized problem. ■

Example 1. Proposition 1 above provides a circuit-theoretic description of the topological reasons supporting the stability loss example discussed in [30]. Indeed, the simplest instance of a circuit verifying the assumptions in Proposition 1 is depicted in Figure 1.

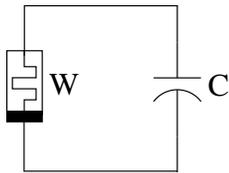


Fig. 1. Example 1

Assuming that the capacitor is a linear one, with positive capacitance C , the circuit equations amount to (cf. [30])

$$\begin{aligned} \varphi'_m &= v \\ C v' &= -W(\varphi_m)v. \end{aligned}$$

It is a simple matter to check that the equilibrium line, defined by $v = 0$ and parameterized by φ_m , becomes unstable as W becomes negative. The two eigenvalues can be checked to read, at a generic equilibrium,

$$\lambda_1 = 0, \quad \lambda_2 = -\frac{W(\varphi_m)}{C},$$

and, assuming $W(0) = 0$, $W'(0) \neq 0$, we have at the origin a double zero eigenvalue with geometric multiplicity one which is indeed responsible for the stability breakdown; note that, indeed, the second eigenvalue becomes positive as W takes on negative values.

In circuit-theoretic terms, this is just a consequence of the fact that the two branches of the circuit define a WC-cutset; together with the absence of C-loops and C-cutsets, this means that Proposition 1 applies.

B. Eigenvalue divergence through $\pm\infty$

It is interesting to note that the dual behavior to the one above may be depicted by divergence of one eigenvalue of the pencil $\lambda A - J$, with A and J given in (7) and (8). Given two matrices A, B in $\mathbb{R}^{n \times n}$ the *matrix pencil* $\{\lambda A, B\}$ is defined as the one-parameter family $\lambda A + B$. If the polynomial (in λ) $\det(\lambda A + B)$ does not vanish identically (that is, if there exists at least one value of λ for which $\det(\lambda A + B) \neq 0$), the matrix pencil is called *regular*. The (finite) eigenvalues of a regular matrix pencil $\{\lambda A, B\}$ are the values of $\lambda \in \mathbb{C}$ for which $\det(\lambda A + B) = 0$. The polynomial $\det(\lambda A + B)$ of a regular pencil has in general a degree $m \leq n$, with $m < n$ when A is a singular matrix; in the latter case we say that the pencil has $n - m$ infinite eigenvalues.

In our setting, provided that the derivative of the right-hand side of (3)-(6) with respect to the variables v_m, i_c, v_r is non-singular, then the pencil $\lambda A - J$ has exactly m eigenvalues, where m is the total number of memristors and capacitors, because of the index-one nature of the differential-algebraic equations modelling the circuit [18], [19]. By contrast, the vanishing of $W(\varphi)$ at a given value of φ may result in the appearance of additional infinite eigenvalues and, again, in a stability breakdown along the line of equilibria; this can be seen as a result of the index jump resulting from the singularity (cf. [30], [31], [32]).

We illustrate this behavior by means of a simple example, obtained after replacing the capacitor in Figure 1 by an inductor, as depicted in Figure 2. The key idea is that, open-circuiting the memristor, the circuit results in an L-cutset, which yields an index-two circuit configuration.

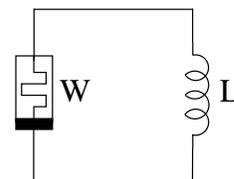


Fig. 2. Example 2

The circuit equations now read as

$$\varphi_m' = v \quad (11)$$

$$Li' = -v \quad (12)$$

$$0 = i - W(\varphi_m)v. \quad (13)$$

Assume $L > 0$. One eigenvalue of (11)-(13) is fixed at the origin, consistently with the existence of the branch of equilibria defined by the identities $i = v = 0$. The second eigenvalue can be checked to read as

$$\frac{-1}{LW(\varphi_m)}$$

and jumps from $-\infty$ to $+\infty$ as W crosses zero and becomes negative. Note that, again, the change of stability occurs along the line of equilibria.

A topological characterization of this phenomenon in memristive circuits, in analogous terms to the ones of Proposition 1, is the object of undergoing research.

IV. CONCLUSION

Many mathematical properties of memristive circuits remain to be solved. Some chaotic phenomena have been explored in [33], [34], [35], but a complete analysis of the analytical properties of manifolds of equilibria in problems with one or several memristors has not yet been fully addressed in the literature. Such results should be relevant in practical applications involving memristors and other mem-devices.

ACKNOWLEDGMENT

Research supported by Project MTM2010-15102 of Ministerio de Ciencia e Innovación, Spain.

REFERENCES

- [1] D. B. Strukov, G. S. Snider, D. R. Stewart and R. S. Williams, The missing memristor found, *Nature* **453** (2008) 80-83.
- [2] L. O. Chua, Memristor – The missing circuit element, *IEEE Trans. Circuit Theory* **18** (1971) 507-519.
- [3] A. Ascoli, T. Schmidt, R. Tetzlaff and F. Corinto, Application of the Volterra series paradigm to memristive systems, in R. Tetzlaff (ed.), *Memristors and Memristive Systems*, pp. 163-191, Springer, 2014.
- [4] D. Birolek, Z. Birolek and V. Biolkova, SPICE modeling of memristive, memcapacitive and meminductive systems, *Proc. Eur. Conf. Circuit Theor. Design 2009*, pp. 249-252, 2009.
- [5] F. Corinto, A. Ascoli and M. Gilli, Analysis of current-voltage characteristics for memristive elements in pattern recognition systems, *Internat. J. Circuit Theory Appl.* **40** (2012) 1277-1320.
- [6] M. Di Ventra, Y. V. Pershin and L. O. Chua, Circuit elements with memory: memristors, memcapacitors and meminductors, *Proc. IEEE* **97** (2009) 1717-1724.
- [7] M. Itoh and L. O. Chua, Memristor oscillators, *Internat. J. Bifur. Chaos* **18** (2008) 3183-3206.
- [8] M. Itoh and L. O. Chua, Memristor Hamiltonian circuits, *Internat. J. Bifur. Chaos* **21** (2011) 2395-2425.
- [9] L. Jansen, M. Matthes and C. Tischendorf, Global unique solvability for memristive circuit DAEs of index 1, *Int. J. Circuit Theory Appl.*, in press, 2014.
- [10] D. Jeltsema and A. J. van der Schaft, Memristive port-Hamiltonian systems, *Math. Comp. Model. Dyn. Sys.* **16** (2010) 75-93.

- [11] D. Jeltsema and A. Doria-Cerezo, Port-Hamiltonian formulation of systems with memory, *Proc. IEEE* **100** (2012) 1928-1937.
- [12] O. Kavehei, A. Iqbal, Y. S. Kim, K. Eshraghian, S. F. Al-Sarawi and D. Abbott, The fourth element: characteristics, modelling and electromagnetic theory of the memristor, *Proc. R. Soc. A* **466** (2010) 2175-2202.
- [13] M. Messias, C. Nespola and V. A. Botta, Hopf bifurcation from lines of equilibria without parameters in memristors oscillators, *Internat. J. Bifur. Chaos* **20** (2010) 437-450.
- [14] Y. V. Pershin and M. Di Ventra, Memory effects in complex materials and nanoscale systems, *Advances in Physics* **60** (2011) 145-227.
- [15] Y. V. Pershin and M. Di Ventra, Neuromorphic, digital and quantum computation with memory circuit elements, *Proc. IEEE* **100** (2012) 2071-2080.
- [16] R. Riaza, Nondegeneracy conditions for active memristive circuits, *IEEE Trans. Circuits and Systems - II* **57** (2010) 223-227.
- [17] R. Tetzlaff (ed.), *Memristors and Memristive Systems*, Springer, 2014.
- [18] R. Riaza, *Differential-Algebraic Systems*, World Scientific, 2008.
- [19] C. Tischendorf, Coupled systems of differential algebraic and partial differential equations in circuit and device simulation. Modeling and numerical analysis, Habilitationsschrift, Inst. Math., Humboldt University, Berlin, 2003.
- [20] L. O. Chua, C. A. Desoer and E. S. Kuh, *Linear and Nonlinear Circuits*, McGraw-Hill, 1987.
- [21] B. Aulbach, *Continuous and Discrete Dynamics near Manifolds of Equilibria*, Lect. Note Math. 1058, Springer-Verlag, 1984.
- [22] B. Fiedler, S. Liescher, and J. C. Alexander, Generic Hopf bifurcation from lines of equilibria without parameters: I. Theory, *J. Differential Equations* **167** (2000) 16-35.
- [23] B. Fiedler and S. Liescher, Generic Hopf bifurcation from lines of equilibria without parameters: II. Systems of viscous hyperbolic balance laws, *SIAM J. Math. Anal.* **31** (2000) 1396-1404.
- [24] S. Liescher, *Bifurcation without Parameters*, Springer, 2015.
- [25] B. Andrásfai, *Introductory Graph Theory*, Akadémiai Kiadó, Budapest, 1977.
- [26] B. Andrásfai, *Graph Theory: Flows, Matrices*, Adam Hilger, 1991.
- [27] B. Bollobás, *Modern Graph Theory*, Springer-Verlag, 1998.
- [28] R. Riaza and C. Tischendorf, Qualitative features of matrix pencils and DAEs arising in circuit dynamics, *Dynamical Systems* **22** (2007) 107-131.
- [29] R. Riaza and C. Tischendorf, The hyperbolicity problem in electrical circuit theory, *Math. Methods Appl. Sci.* **33** (2010) 2037-2049.
- [30] R. Riaza, Manifolds of equilibria and bifurcations without parameters in memristive circuits, *SIAM J. Appl. Math.* **72** (2012) 877-896.
- [31] R. E. Beardmore, The singularity-induced bifurcation and its Kronecker normal form, *SIAM J. Matrix Anal. Appl.* **23** (2001) 126-137.
- [32] V. Venkatasubramanian, H. Schättler and J. Zaborszky, Local bifurcations and feasibility regions in differential-algebraic systems, *IEEE Trans. Aut. Contr.* **40** (1995) 1992-2013.
- [33] B. Bao, Z. Ma, J. Xu, Z. Liu and Q. Xu, A simple memristor chaotic circuit with complex dynamics, *Internat. J. Bifurcation and Chaos* **21** (2011) 2629-2645.
- [34] B. Muthuswamy, Implementing memristor based chaotic circuits, *Internat. J. Bifur. Chaos* **20** (2010) 1335-1350.
- [35] B. Muthuswamy and L. O. Chua, Simplest chaotic circuit, *Internat. J. Bifur. Chaos* **20** (2010) 1567-1580.

Ricardo Riaza (Madrid, Spain, 1972) received the MSc degree in Mathematics from Universidad Complutense de Madrid in 1996, the MSc degree in Electrical and Electronic (Telecommunication) Engineering from Universidad Politécnica de Madrid (UPM) in 1997, and the PhD degree in Mathematics from UPM in 2000. He received the Outstanding PhD Award for his Doctoral Thesis in 2001 and the Research/Technological Development Award for young professors in 2005, both from UPM. Currently, he serves as an Associate Professor at the Departamento de Matemática Aplicada a las Tecnologías de la Información y las Comunicaciones of the ETSI Telecomunicación (UPM). His current research interests are focused on differential-algebraic equations (DAEs) and also on analytical aspects of nonlinear electrical and electronic circuits, including circuits with memristors. He is the author of the book "Differential-Algebraic Systems: Analytical Aspects and Circuit Applications" (World Scientific, 2008), and the first or unique author of nearly forty papers in JCR journals.

MODELLING OF EXOSKELETON MOVEMENT IN VERTICALIZATION PROCESS

Sergey Jatsun, Doctor of science, Professor, Head of the department of mechanics, mechatronics and robotics, Southwest State University

Sergei Savin, Candidate of science, Junior research fellow of the department of mechanics, mechatronics and robotics, Southwest State University

Petr Bezmen, Candidate of science, Associate professor of the department of mechanics, mechatronics and robotics, Southwest State University

Abstract—The present paper focuses on comparison and analysis of different techniques of data processing as related to the problem of acquiring experimental data of human getting up process in such form that it can be used as an input to the control system of an exoskeleton. Use of approximation by trigonometric series, polynomials and spline functions is discussed.

Keywords—Experimental data processing, exoskeleton control inputs.

I. INTRODUCTION

Nowadays, various robotics facilities are systems composed of two main elements – a man and a machine. These systems include objects called the exoskeleton and used to extend the functionality of a person. The interaction of these two components determines the quality of the system as a whole.

Recently, the new generation of exoskeletons came into service and it allows a person to move in space, even in case of damage of the lower extremities. In addition, it becomes possible to significantly expand human capabilities when a person performs tasks that impose high demands on endurance and physical strength of a man.

Obviously, the creation of such devices is possible with a well-developed theory of the functioning of “man-machine” systems, and the particular emphasis should be given to the control. The main aim demands the consideration of interaction between man and machine. The common problems in the theory of walking mechanisms have been developed in [1-4]. The mathematical modelling of elements motion is one of the most important tools in the study of behavior of the systems like an exoskeleton. At the same time it is necessary to process a large number of experimentally obtained movement curves, solve the problem of approximation and obtain the analytical dependences which reflect the change in generalized coordinates describing the position of the mechanism links.

The paper is devoted to the analytical construction of exoskeleton motion trajectories by means of the experimental data characterizing the movement of a person in the getting up process. The solution of this problem will allow adapting the dynamic characteristics exoskeleton to the motion of person.

II. STATEMENT OF THE PROBLEM

The purpose of this paper is the comparative analysis of data processing methods to process information about the person motions during the process of getting up. It is possible to synthesize the control actions for the exoskeleton control system on the basis of these methods. The methods of obtaining the experimental data may be different and are not described here. We assume that the source data is written in the form of numerical sequences which define a person position at each time interval. A person position is determined by the generalized coordinates. In this paper, we consider the case when the system of three generalized coordinates defines the orientation of a shin, a hip and a trunk. These parts of a person body execute the plane-parallel motions but a foot remains stationary. Hence it appears that the person movement can be described as the four-link mechanism motion with one fixed link (figure 1).

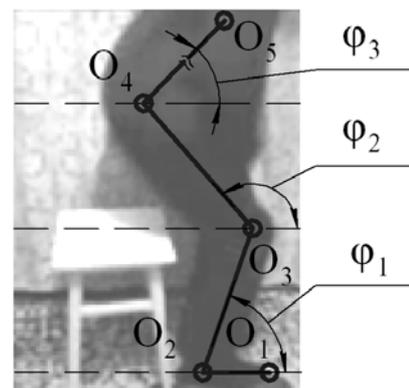


Figure 1 A person in the process of getting up and the four-link mechanism

A shin orientation is determined by the angle φ_1 , a hip orientation – by the angle φ_2 , and a trunk orientation – by the angle φ_3 . The figure 2 shows the human rising process data derived from experiments in the laboratory of the department of mechanics, mechatronics and robotics, Southwest State University. The corresponding numerical dependences are shown in figure 2 (a).

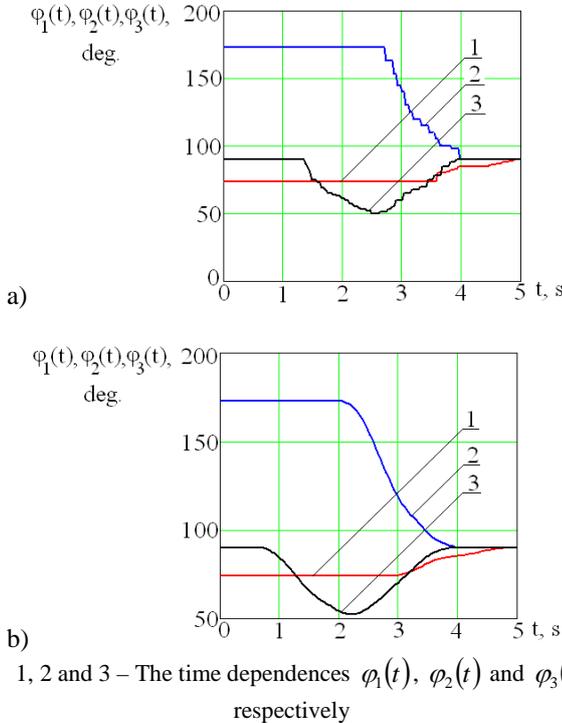


Figure 2 The experimental data: a) – before smoothing process, b) – after smoothing process

The resulting graphs (figure 2 (a)) are largely non-linear, which is due to the inaccuracy of measurement. High-frequency components of the signal give a stepped appearance and do not carry useful information and should be removed before the signal can be used as the input action for the exoskeleton control system. The smoothing method is the simplest way to process these signals. The figure 2 (b) shows the dependences received after data processing by the sliding window method that uses the following expression:

$$\tilde{\varphi}_i(t_j) = \frac{1}{n} \sum_{k=1}^n \varphi_i(t_{j+k}), \quad i = 1, 2, 3, \quad (1)$$

where: $\varphi_i(t_j)$ – the generalized coordinate φ_i value at the time t before data processing, $\tilde{\varphi}_i(t_j)$ – the generalized coordinate φ_i value at the time t after data processing, n – the width of the sliding window.

The use of these dependences has several disadvantages. Smoothed graphs and their derivatives retain high-frequency components described above, (although their amplitude decreased significantly) that can impair the performance of the control system.

Moreover, the smoothing shifts the boundaries of transitions between different stages of the movement.

III. THE APPROXIMATION BY TRIGONOMETRIC SERIES

An approximation is a method of data processing that allows to selectively retain the required information about the motion by eliminating unwanted high-frequency components. We consider an approximation of the original dependences by the trigonometric Fourier series:

$$f_i(t) = \frac{1}{2}a_{i,0} + \sum_{k=1}^n (a_{i,k} \cos(k \cdot t) + b_{i,k} \sin(k \cdot t)) \quad (2)$$

where: f_i – the function, that approximates the experimental dependence φ_i , $a_{i,k}, b_{i,k}$ – the Fourier series coefficients, $i = 1, 2, 3$.

The figure 3 shows the results of approximation at $n = 30$. The series coefficients for the i -th generalized coordinate are chosen by minimizing of the following positive definite function:

$$E_i = \sum_{j=1}^{m_i} \left(\varphi_i(t_j) - \frac{1}{2}a_{i,0} - \sum_{k=1}^{30} (a_{i,k} \cos(k \cdot t_j) + b_{i,k} \sin(k \cdot t_j)) \right)^2, \quad (3)$$

where m_i – the total number of the dependence φ_i points, $i = 1, 2, 3$.

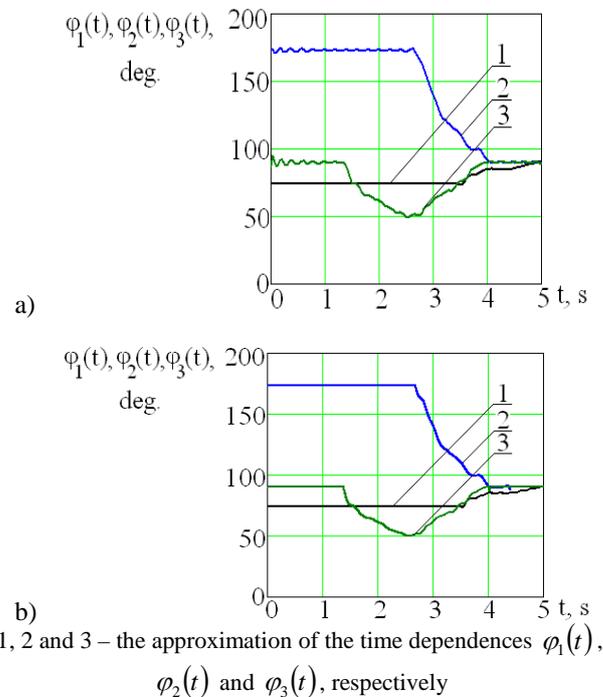


Figure 3 The approximation of the initial data by trigonometric series: a) with using of the formula (3), b) with using of the formula (4)

It is possible to pay attention to the occurrence of oscillations on the “direct” sections of the graphs (where

the first order time-derivative is equal to null, figure 3 (a). The approximation by piecewise-defined function

$$\rho_i(t) = \begin{cases} \varphi_i(t_1) & \text{if } t \in [t_{i1}, t_{i2}] \\ \varphi_i(t_3) & \text{if } t \in [t_{i3}, t_{i4}] \\ \frac{1}{2} a_{i,0} + \sum_{k=1}^n (a_{i,k} \cos(k \cdot t) + b_{i,k} \sin(k \cdot t)) & \text{otherwise,} \end{cases} \quad (4)$$

where: $\rho_i(t)$ – the piecewise-defined function used for the approximation of original dependence φ_i , $[t_{i1}, t_{i2}]$ and $[t_{i3}, t_{i4}]$ – the first and the second sections, respectively, where φ_i has the first order derivative that is equal to null. The figure 3 (b) presents the experimental data approximation results of the piecewise-defined function.

We note that in both cases, the dependences have significant vibrational state, and it is especially conspicuous in the graphs of the time-derivatives. The figure 4 shows the first order time-derivatives plots of the functions f_k in modulo.

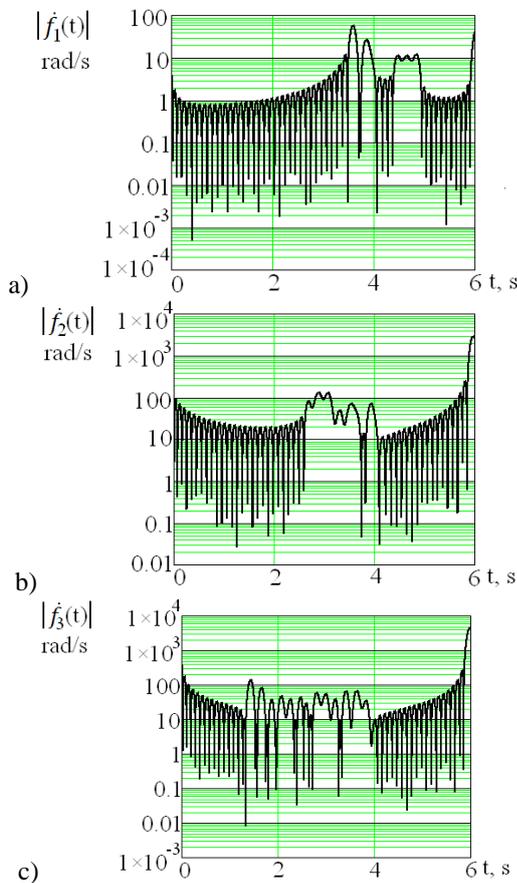


Figure 4 The time dependences a) $|\dot{f}_1|$, b) $|\dot{f}_2|$, c) $|\dot{f}_3|$ in a logarithmic scale

The figures 3 and 4 allow us to conclude that the use of the approximation by trigonometric series make it possible to reliably reproduce the original dependences, but this method leads to additional high-frequency components in the signal spectrum. The use of functions obtained by this way as the input action for the exoskeleton control system can negatively affect the control process.

can avoid these oscillations. We consider the case when the graph has two straight sections:

IV. THE APPROXIMATION BY POLYNOMIAL FUNCTIONS AND SPIELS

The n-th order polynomial can be written in form:

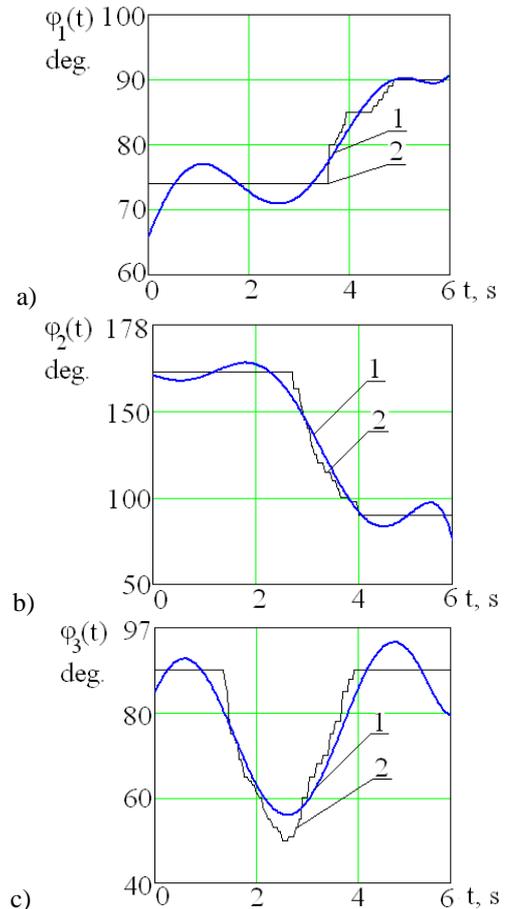
$$P_i(t) = \sum_{k=0}^n c_{i,k} t^k, \quad i = 1, 2, 3, \quad (5)$$

where: $P_i(t)$ – the polynomial functions that are used for the approximation of the original dependence φ_i , $c_{i,k}$ – the polynomial coefficients.

The polynomial coefficients for the i-th generalized coordinate are chosen in the course of minimizing of the following function:

$$E_i = \sum_{j=1}^{m_i} \left(\varphi_i(t_j) - \sum_{k=0}^n c_{i,k} t_j^k \right)^2, \quad i = 1, 2, 3, \quad (6)$$

The figure 5 shows the graphs obtained due to the approximation of the functions φ_i by the sixth order polynomials.



1 – the polynomial functions $P_i(t)$, 2 – the time dependences φ_i

Figure 5 The approximation of the experimental data by the sixth order polynomials for the generalized coordinate: a) φ_1 b) φ_2 , c) φ_3

As well as in the case of approximation by trigonometric series, the use of polynomial functions

leads to errors in straight segment of the dependences φ_i . To obtain the better results, we use the approximation by spline functions. For this purpose we divide the dependences $\varphi_1(t)$ and $\varphi_2(t)$ into three portions, and the dependence $\varphi_3(t)$ into the four sections. Let this partitioning occur at the points corresponding to the time instants: $t_{11} = 3.48$ sec, $t_{12} = 4.93$ sec for the dependence $\varphi_1(t)$, and $t_{21} = 2.69$ sec, $t_{22} = 4.4$ sec for the dependence $\varphi_2(t)$, and $t_{31} = 1.38$ sec, $t_{32} = 2.56$ sec, $t_{33} = 4.03$ sec for the dependence $\varphi_3(t)$. The first section and the last section of each spline are specified by the zero order polynomial, and the rest is defined by the seventh order polynomials.

To calculate the polynomial coefficients, we can write the following conditions:

$$\left\{ \begin{array}{l} S_1(t_{11}) = \varphi_1(t_{11}) \\ S_1(t_{12}) = \varphi_1(t_{12}) \\ \dot{S}_1(t_{11}) = \dot{S}_1(t_{11}) = \ddot{S}_1(t_{11}) = 0 \\ \dot{S}_1(t_{12}) = \dot{S}_1(t_{12}) = \ddot{S}_1(t_{12}) = 0 \end{array} \right\}, \left\{ \begin{array}{l} S_2(t_{21}) = \varphi_2(t_{21}) \\ S_2(t_{22}) = \varphi_2(t_{22}) \\ \dot{S}_2(t_{21}) = \dot{S}_2(t_{21}) = \ddot{S}_2(t_{21}) = 0 \\ \dot{S}_2(t_{22}) = \dot{S}_2(t_{22}) = \ddot{S}_2(t_{22}) = 0 \end{array} \right\}, \left\{ \begin{array}{l} S_3(t_{31}) = \varphi_3(t_{31}) \\ S_3(t_{32}) = \varphi_3(t_{32}) \\ S_3(t_{33}) = \varphi_3(t_{33}) \\ \dot{S}_3(t_{31}) = \dot{S}_3(t_{31}) = \ddot{S}_3(t_{31}) = 0 \\ \dot{S}_3(t_{32}) = \dot{S}_3(t_{32}) = \ddot{S}_3(t_{32}) = 0 \\ \dot{S}_3(t_{33}) = \dot{S}_3(t_{33}) = \ddot{S}_3(t_{33}) = 0 \end{array} \right\}, \quad (7)$$

where $S_1(t), S_2(t), S_3(t)$ are the spline functions used for approximation of the time dependences $\varphi_1(t), \varphi_2(t)$ и $\varphi_3(t)$, respectively.

Due to using of the criterion (7), we can find the desired coefficients to plot splines (figure 6).

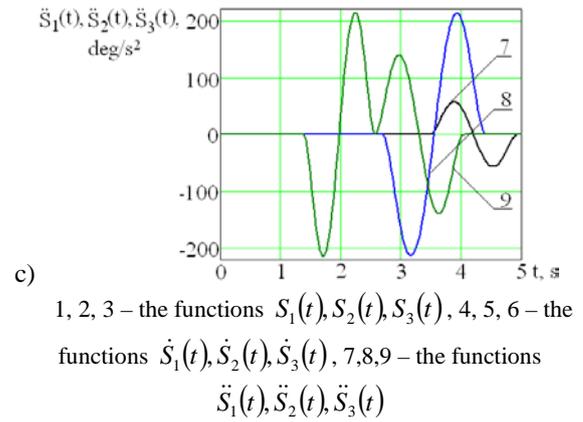
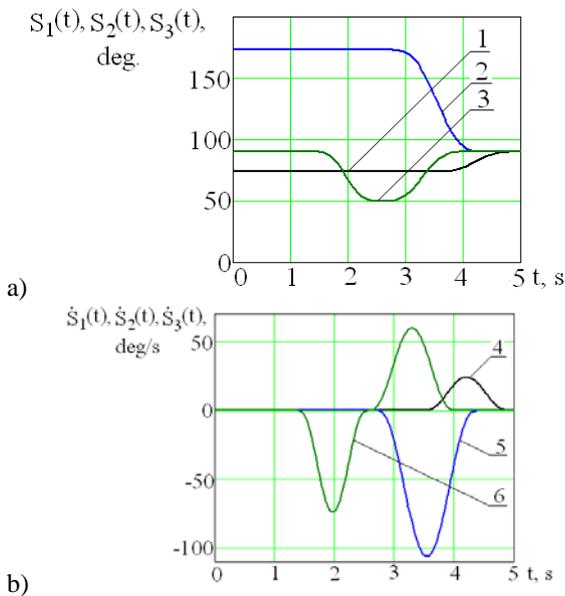


Figure 6 The graphs: a) the spline functions, b) the first order time-derivatives of spline functions, c) the second order time-derivatives of spline functions

Because of the seven order spline functions, it is possible to achieve absence of function discontinuities in the graphs of the time dependences of the generalized velocities and accelerations (figure 6 (b) and (c)). Also, the spline functions eliminate the high-frequency oscillations.

V. THE DETERMINATION OF MOMENTS NEEDED TO IMPLEMENT THE OBTAINED MOVEMENT TRAJECTORIES OF THE MECHANISM

Different approaches with some accuracy allow obtaining the mechanism movement that is determined by certain changes of the generalized coordinates. For example, it is possible to build the automatic control system using negative feedback to control the generalized coordinates. We consider another approach: the moments sequence realizes the desired movement and can be determined by solving the inverse problem of dynamics.

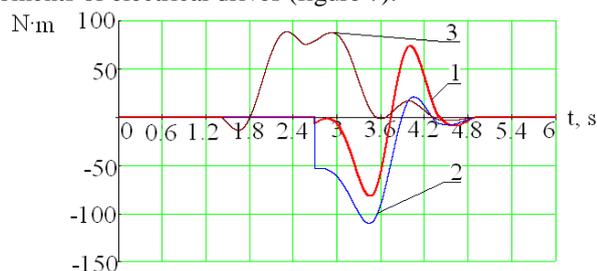
The equations of the flat three-link mechanism dynamics can be found in a number of papers, including [5], we do not give them in the paper. The flat three-link mechanism is a series of connected links by joints. In general terms, the equation of the mechanism dynamics can be written as follows:

$$A(\vec{\varphi}) \cdot \ddot{\vec{\varphi}} + \vec{b}(\vec{\varphi}, \dot{\vec{\varphi}}) + \vec{g}(\vec{\varphi}) = T \cdot \vec{\tau}, \quad (8)$$

where $A(\vec{\varphi})$ – the matrix of kinetic energy, $\vec{\varphi}, \dot{\vec{\varphi}}, \ddot{\vec{\varphi}}$ – the vectors of the generalized coordinates, generalized velocities, and generalized accelerations, respectively, $\vec{b}(\vec{\varphi}, \dot{\vec{\varphi}})$ – the vector bound with the compound centrifugal forces, $\vec{g}(\vec{\varphi})$ – the vector of the generalized potential forces, $\vec{\tau}$ – the vector consists of some elements – the moments which are generated by electrical drives, T – the transition matrix.

The initial data for solving of the inverse problem of dynamics is the law of the generalized coordinates alteration and their first and second order time-derivatives. As a law, we use the functions $S_1(t), S_2(t), S_3(t)$, described in the previous section. In the solving of the inverse problem of

dynamics, we obtained the results which depend on the moments of electrical drives (figure 7).



1, 2, 3 – the moments of electrical drives $\tau_1(t)$, $\tau_2(t)$ и $\tau_3(t)$ mounted in the ankle joint, the knee joint and the coxofemoral joint of exoskeleton, respectively

Figure 7 The time dependence of the moments generated by the exoskeleton drives

Two graphs $\tau_1(t)$ and $\tau_2(t)$ have function discontinuities at $t = 2.69$ sec. (figure 7). Before this time moment $t = 2.69$ sec a hip and a shin were in static equilibrium under the influence of reaction at supports (i.e. a chair and a floor). Thus, the time moment $t = 2.69$ sec is the power up time of first and second electrical drives.

VI. CONCLUSION

This paper discusses various processing methods of experimental data which describe the motion of a person in the getting up process. It is shown that the approximation of initial relationships by trigonometric series provides a sufficient accuracy to reproduce the shape of the original dependences, but adds the high-frequency harmonics in the spectrum of the signal. These harmonics can adversely affect the quality of the control process. The paper demonstrates that this problem can be avoided by using the spline approximation. The solution results of the inverse problem of dynamics are presented. These results were gotten by means of the approximating spline functions and their derivatives. The spline functions make it possible to reduce the peak magnitude of the second order time-derivatives with respect to the original dependences and other types of approximating functions.

REFERENCES

- [1]. Formalskiy A. M. Peremeshcheniye antropomorfnykh mekhanizmov. M.: Nauka, 1982.
- [2]. Beletskiy V. V., Berbyuk V. Ye. Nelineynaya model dvunogogo shagayushchego apparata, snabzhennogo upravlyayemyimi stopami. M.: Nauka, 1982.
- [3]. Beletskiy V. V. Dvunogaya khodba: Model'nyye zadachi dinamiki i upravleniya. M.: Nauka, 1984.
- [4]. Vukobratovich M. K. Shagayushchiye roboty i antropomorfnyye mekhanizmy. M.: Mir, 1976.
- [5]. Vorochaeva L. Yu. Simulation of Motion of a Three Link Robot with Controlled Friction Forces on a Horizontal Rough Surface / L. Yu. Vorochaeva, G. S. Naumov, S. F. Yatsun // Journal of Computer and Systems Sciences International, 2015, Vol. 54, No. 1, pp. 151–164.

Multiobjective Genetic Algorithm-Based for Time-Cost Optimization

Jorge Magalhães-Mendes

Abstract— This paper presents a hybrid genetic algorithm for the time-cost optimization (TCO) problem. The chromosome representation of the problem is based on random keys. The schedules are constructed using a priority rule in which the priorities are defined by the genetic algorithm. Schedules are constructed using a procedure that generates parameterized active schedules. In construction projects, time and cost are the most important factors to be considered. In this paper, a new hybrid genetic algorithm is developed for the optimization of the two objectives time and cost. The results indicate that this approach could assist decision-makers to obtain good solutions for project duration and total cost.

Keywords—Project management, Genetic Algorithms, Time-cost optimization.

I. INTRODUCTION AND BACKGROUND

Construction projects are found throughout business and areas such as manufacturing facilities, infrastructure development and improvement, and residential and commercial building.

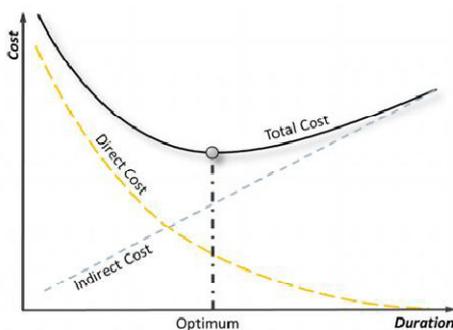


Fig. 1. Project time and cost curve.

In a construction project, there are two main factors, such as project duration and project cost. The activity duration is a function of resources (i.e. crew size, equipments and materials) availability. On the other hand, resources demand direct costs. Therefore, the relationship between project time and direct cost of each activity is a monotonously decreasing

curve. It means if activity duration is compressed then that leads to an increase in resources and so that direct costs. But, project indirect costs increase with the project duration. In general, for a project, the total cost is the sum of direct and indirect costs and exists an optimum duration for the least cost, see Fig.1. Hence, relationship between project time and cost is trade-off [36].

Several approaches to solve the TCO problem have been proposed in the last years: mathematical, heuristic and search methods.

A. Mathematical Methods

Several mathematical models such as linear programming (Kelley [12]; Hendrickson and Au [4]; Pagnoni [2]), integer programming, or dynamic programming (Butcher [33]; Robinson [8]; Elmaghraby [27]; De et al. [25]) and LP/IP hybrid (Liu et al. [21]; Burns et al. [29]), Meyer and Shaffer [31] and Patterson and Huber [14] use mixed integer programming. However, for large number of activity in network and complex problem, integer programming needs a lot of computation effort (Feng et al. [6]).

B. Heuristic Methods

Heuristic algorithms are not considered to be in the category of optimization methods. They are algorithms developed to find an acceptable near optimum solution. Heuristic methods are usually algorithms easy to understand which can be applied to larger problems and typically provide acceptable solutions (Hegazy [30]). However, they have lack mathematical consistency and accuracy and are specific to certain instances of the problem (Fondahl [19]; Prager [32]; Siemens [23] and Moselhi [24]) are some of the research studies that have utilized heuristic methods for solving TCO problems.

C. Search Methods

Some researchers have tried to introduce evolutionary algorithms to find global optima such as genetic algorithm (GA) (Feng et al. [6]; Gen and Cheng [22]; Zheng et al. [10]; Zheng and Ng [9]); the particle swarm optimization algorithm (Yang [11]), ant colony optimization (ACO) (Xiong and

J. Magalhães-Mendes is with the Civil Engineering Department, School of Engineering, Polytechnic of Porto, Porto, Portugal (e-mail: jjm@isep.ipp.pt).

Kuang [34]; Ng and Zhang [29]; Afshar et al. [1]) and harmony search (HS) (Geem [36]).

In this paper it is proposed a hybrid genetic algorithm based on the works [16] and [18], with a new chromosome structure to solve the time-cost optimization problem.

II. MULTIOBJECTIVE OPTIMIZATION

With evolutionary techniques being used for single-objective optimization for over two decades, the incorporation of more than one objective in the fitness function has finally gained popularity in the research [3].

In principle, there is no clear definition of an ‘‘optimum’’ in multiobjective optimization (MOP) as in the case of single-objective issues; and there even does not necessarily have to be an absolutely superior solution corresponding to all objectives due to the incommensurability and conflict among objectives. Since the solutions cannot be simply compared with each other, the ‘‘best’’ solution generated from optimization would correspond to human decision-makers subjective selection from a potential solution pool, in terms of their particulars [10].

The classical methods reduce the MOP to a scalar optimization optimization by using multiobjective weighting (MOW) or a utility function (multiobjective utility analysis). Multiobjective weighting allows decisions makers to incorporate the priority of each objective into decision making. Mathematically, the solutions obtained by equally weighting all objectives may provide the least objective conflicts, but in most cases, each objective is first optimized separately and the overall objective value is evaluated depending on the weighting factors. The weakness of MOW is that the overall optimum is usually at the dominating objective only [6].

In a certain way we can say that the work of Zadeh [20] is the first to advocate the assignment of weights to each objective function and combined them into a single-object function. More recently, Gen and Cheng [22] adopted the adaptive weight approach (AWA) in construction TCO problem (also referred to as GC approach hereafter).

In the GC approach Gen and Cheng [22] proposed the following formulas:

$$Z^+ = \{Z_c^{\max}, Z_t^{\max}\} \quad (1)$$

$$Z^- = \{Z_c^{\min}, Z_t^{\min}\} \quad (2)$$

where,

Z_c^{\max} = maximal value for total cost in the current population;

Z_t^{\max} = maximal value for time in the current population;

Z_c^{\min} = minimal value for total cost in the current

population;

Z_t^{\min} = minimal value for time in the current population.

$$w_c = 1 / (Z_c^{\max} - Z_c^{\min}), \quad w_t = 1 / (Z_t^{\max} - Z_t^{\min}) \quad (3)$$

$$f(x) = w_c (Z_c^{\max} - Z_c) + w_t (Z_t^{\max} - Z_t) \quad (4)$$

In 2004, Zheng et al. [10] proposed the modified weight approach (MAWA) to deal with the multi-objective problem. Under the MAWA, the adaptive weights are formulated through the following four conditions:

1) For Z_t^{\max} is not equal to Z_t^{\min} and Z_c^{\max} is not equal to Z_c^{\min}

$$v_c = \frac{Z_c^{\min}}{Z_c^{\max} - Z_c^{\min}} \quad (5)$$

$$v_t = \frac{Z_t^{\min}}{Z_t^{\max} - Z_t^{\min}} \quad (6)$$

$$v = v_c + v_t \quad (7)$$

$$w_c = v_c / v \quad (8)$$

$$w_t = v_t / v \quad (9)$$

$$w_c + w_t = 1 \quad (10)$$

2) For $Z_t^{\max} = Z_t^{\min}$ and $Z_c^{\max} = Z_c^{\min}$

$$w_c = w_t = 0.5 \quad (11)$$

3) For $Z_t^{\max} = Z_t^{\min}$ and $Z_c^{\max} \neq Z_c^{\min}$

$$w_c = 0.1, \quad w_t = 0.9 \quad (12)$$

4) For $Z_t^{\max} \neq Z_t^{\min}$ and $Z_c^{\max} = Z_c^{\min}$

$$w_c = 0.9, \quad w_t = 0.1 \quad (13)$$

Zheng et al. [10] proposed a fitness formula in accordance with the proposed adaptive weight:

$$f(x) = w_t \frac{(Z_t^{\max} - Z_t) + \gamma}{(Z_t^{\max} - Z_t^{\min}) + \gamma} + w_c \frac{(Z_c^{\max} - Z_c) + \gamma}{(Z_c^{\max} - Z_c^{\min}) + \gamma} \quad (14)$$

where,

γ is a small positive random number between 0 and 1.

Z_c^{\max} = maximal value for total cost in the current population;
 Z_t^{\max} = maximal value for time in the current population;
 Z_c^{\min} = minimal value for total cost in the initial population;
 Z_t^{\min} = minimal value for time in the initial population;
 Z_c = represents the total cost of the x^{th} solution in current population;
 Z_t = represents the time of the x^{th} solution in current population.

This study uses the fitness formula proposed by Gen and Cheng [22] where,

Z_c^{\max} = maximal value for total cost in the current chromosome;
 Z_t^{\max} = maximal value for time in the current chromosome;
 Z_c^{\min} = minimal value for total cost in the initial population;
 Z_t^{\min} = minimal value for time in the initial population;
 Z_c = represents the total cost of the x^{th} solution in current chromosome;
 Z_t = represents the time of the x^{th} solution in current chromosome.

III. THE GA-BASED APPROACH

The approach presented in this paper is based on a genetic algorithm to perform its optimization process. Fig. 2 shows the architecture of approach.

The approach combines a genetic algorithm, a schedule generation scheme and a local search procedure. The genetic algorithm is responsible for evolving the chromosomes which represent the priorities of the activities.

For each chromosome the following four phases are applied:

- 1) *Transition parameters* - this phase is responsible for the process transition between first level and second level;
- 2) *Schedule parameters* - this phase is responsible for transforming the chromosome supplied by the genetic algorithm into the priorities of the activities and delay time;
- 3) *Schedule generation* - this phase makes use of the priorities and the delay time and constructs schedules;

- 4) *Schedule improvement* - this phase makes use of a local search procedure to improve the solution obtained in the schedule generation phase.

After a schedule is obtained, the quality is processed feedback to the genetic algorithm. Fig. 2 illustrates the sequence of phases applied to each chromosome. Details about each of these phases will be presented in the next sections.

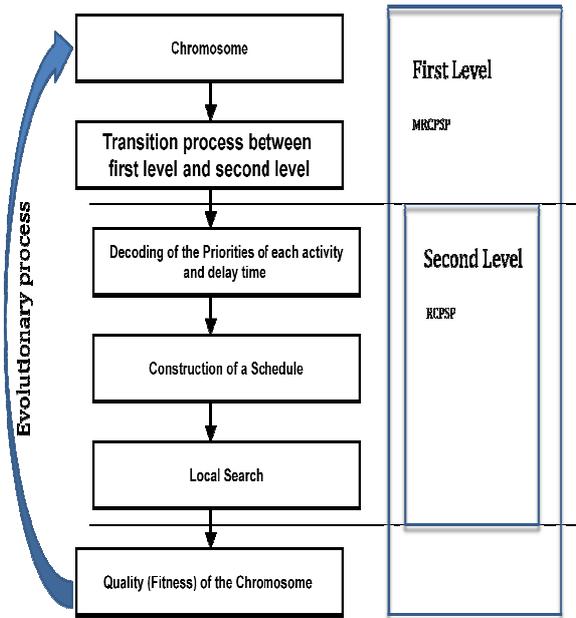


Fig. 2. Architecture of the approach.

A. GA Transition Process

The Genetic Algorithms (GAs) are search algorithms which are based on the mechanics of natural selection and genetics to search through decision space for optimal solutions. One fundamental advantage of GAs from traditional methods is described by Goldberg [7]: in many optimization methods, we move gingerly from a single solution in the decision space to the next using some transition rule to determine the next solution.

First of all, an initial population of potential solutions (individual) is generated randomly. A selection procedure based on a fitness function enables to choose the individual candidate for reproduction. The reproduction consists in recombining two individuals by the crossover operator, possibly followed by a mutation of the offspring. Therefore, from the initial population a new generation is obtained. From this new generation, a second new generation is produced by the same process and so on. The stop criterion is normally based on the number of generations.

The GA based-approach uses a random key alphabet U (0, 1) and an evolutionary strategy identical to the one proposed

by Goldberg [7].

Each chromosome represents a solution to the problem and it is encoded as a vector of random keys (random numbers). Each solution encoded as initial chromosome (first level) is made of $mn+n$ genes where n is the number of activities and m is the number of execution modes, see Fig. 3.

The called first level as the capacity to solving the multi-mode resource constrained project scheduling problem (MRCPS) [16, 18].

In this case of study we do not consider the requirements to the type and number of resources needed for construction mode for each activity as well as the maximum number of available resources.

The transition process between first level and second level consists in choosing the option or construction mode m_j for each activity j . Using this process we obtain the solution chromosome (second level) composed by $2n$ genes, see Fig.4.

The called second level as the capacity to solving the resource constrained project scheduling problem (RCPS) [16, 18].

In this case of study we do not consider the requirements to the type and number of resources needed for each activity as well as the maximum number of available resources.

Activity 1	Mode 1	Gene ₁₁
	Mode 2	Gene ₁₂

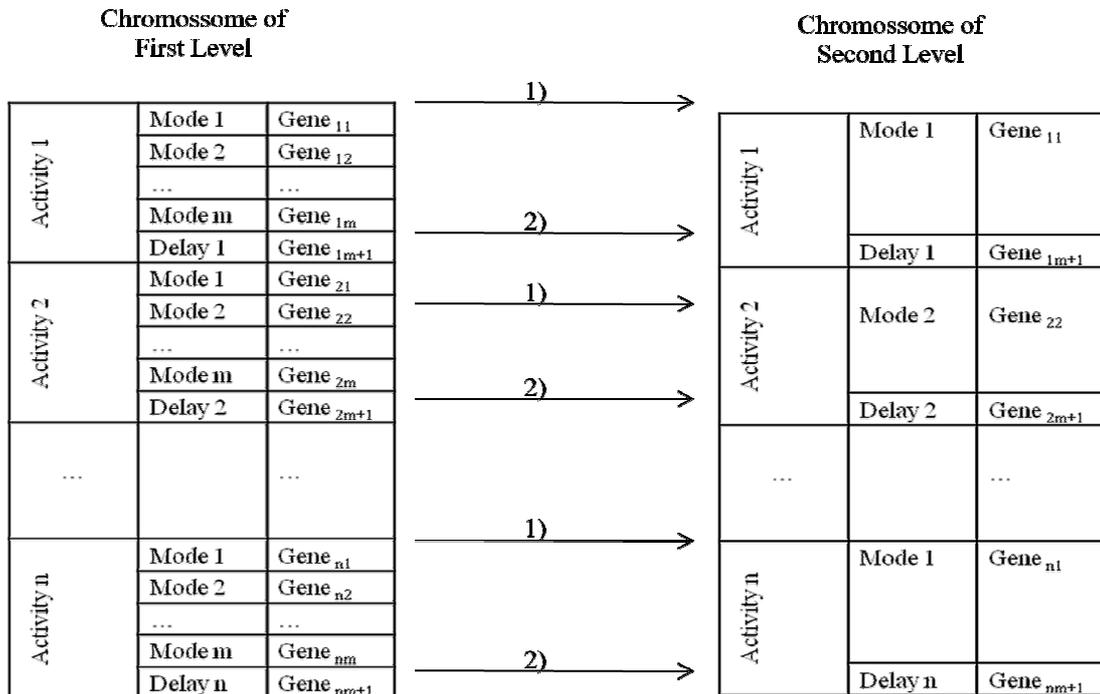
	Mode m	Gene _{1m}
	Delay 1	Gene _{1m+1}
Activity 2	Mode 1	Gene ₂₁
	Mode 2	Gene ₂₂

	Mode m	Gene _{2m}
	Delay 2	Gene _{2m+1}
...		...
Activity n	Mode 1	Gene _{n1}
	Mode 2	Gene _{n2}

	Mode m	Gene _{nm}
	Delay n	Gene _{nm+1}

Fig. 3. Chromosome structure.

After, we evaluate the quality (fitness) of the solution chromosome.



- 1) The gene is chosen by the highest priority
- 2) Automatically carried over to the second level

Fig. 4. Transition process between first and second level.

B. GA Decoding

A real-coded GA is adopted in this paper. Compared with the binary-code GA, the real-coded GA has several distinct advantages, which can be summarized as follows (Y.-Z. Luo et al. [35]):

- It is more convenient for the real-coded GA to denote large scale numbers and search in large scope, and thus the computation complexity is amended and the computation efficiency is improved;
- The solution precision of the real-coded GA is much higher than that of the binary-coded GA;
- As the design variables are coded by floating numbers in classical optimization algorithms, the real-coded GA is more convenient for combination with classical optimization algorithms.

The priority decoding expression uses the following expression:

$$PRIORITY_j = \frac{LLP_j}{LCP} \times \left[\frac{1 + gene_{m_j}}{2} \right] \quad j = 1, \dots, n \quad (15)$$

where,

- [1] LLP_j is the longest length path from the beginning of the activity j to the end of the project;
- [2] LCP is the length along the critical path of the project [15];
- [3] m_j is the gene of the selected mode for activity j .

The gene $jm+1$ is used to determine the delay time when scheduling the activities. The delay time used by each activity is given by the following expression:

$$Delay\ time = gene_{jm+1} \times 1.5 \times MaxDur \quad (16)$$

where $MaxDur$ is the maximum duration of all activities. The factor 1.5 is obtained after some experimental tuning.

A maximum delay time equal to zero is equivalent to restricting the solution space to non-delay schedules and a maximum delay time equal to infinity is equivalent to allowing active schedules. To reduce the solution space is used the value given by formula (16), see Gonçalves et al. [13].

C. Construction of a Schedule

Schedule generation schemes (SGS) are the core of most heuristic solution procedures for project scheduling. SGS start from scratch and build a feasible schedule by stepwise extension of a partial schedule.

There are two different classic methods SGS available. They can be distinguished into activity and time

incrementation. The so called serial SGS performs activity-incrementation and the so called parallel SGS performs time-incrementation.

A third method for schedule generating can be applied: the parameterized active schedules. This type of schedule consists of schedules in which no resource is kept idle for more than a predefined period if it could start processing some activity. If the predefined period is set to zero, then we obtain a non-delay schedule. This type of SGS is used on this work.

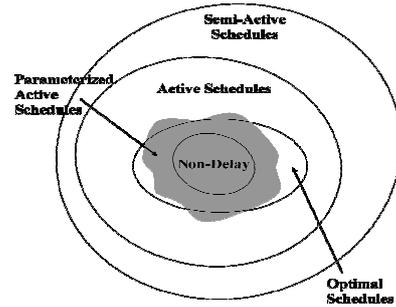


Fig. 5. Types of schedules (adapted from Mendes [18]).

Fig. 5 presents the relationship diagram of various schedules with regard to optimal schedules.

D. Local Search

Local search algorithms move from solution to solution in the space of candidate solutions (the search space) until a solution optimal or a stopping criterion is found. In this paper it is applied backward and forward improvement based on Klein [27].

Initially it is constructed a schedule by planning in a forward direction starting from the project's beginning. After it is applied backward and forward improvement trying to get a better solution. The backward planning consists in reversing the project network and applying the scheduling generator scheme. A detailed example is described by Mendes [15].

E. Evolutionary Strategy

There are many variations of genetic algorithms obtained by altering the reproduction, crossover, and mutation operators. Reproduction is a process in which individual (chromosome) is copied according to their fitness values (makespan). Reproduction is accomplished by first copying some of the best individuals from one generation to the next, in what is called an elitist strategy.

In this paper the fitness proportionate selection, also known as roulette-wheel selection, is the genetic operator for selecting potentially useful solutions for reproduction. The characteristic of the roulette wheel selection is stochastic sampling.

The fitness value is used to associate a probability of

selection with each individual chromosome. If f_i is the fitness of individual i in the population, its probability of being selected is,

$$p_i = \frac{f_i}{\sum_{i=1}^N f_i}, \quad i = 1, \dots, n \quad (17)$$

A roulette wheel model is established to represent the survival probabilities for all the individuals in the population. Then the roulette wheel is rotated for several times [7].

After selecting, crossover may proceed in two steps. First, members of the newly selected (reproduced) chromosomes in the mating pool are mated at random. Second, each pair of chromosomes undergoes crossover as follows: an integer position k along the chromosome is selected uniformly at random between 1 and the chromosome length l . Two new chromosomes are created swapping all the genes between $k+1$ and l , see Mendes [16].

The mutation operator preserves diversification in the search. This operator is applied to each offspring in the population with a predetermined probability. We assume that

the probability of the mutation in this paper is 5%.

F. GA Configuration

Though there is no straightforward way to configure the parameters of a genetic algorithm, we obtained good results with values: population size of $5 \times$ number of activities in the problem; mutation probability of 0.05; top (best) 1% from the previous population chromosomes are copied to the next generation; stopping criterion of 50 generations.

IV. CASE STUDY

In order to compare the proposed RKV-TCO (Random Key Variant for Time-Cost Optimization) approach, a case study of seven activities proposed initially by Liu et al. [21] was used.

A project of seven activities proposed initially by Liu et al. [21] and fitted by Zheng et al. [10] is presented in Table 1 with available activity options and corresponding durations and costs. Indirect cost rate was \$1500/day.

Table 1 Time and cost for each option/mode of activity.

Activity description	Activity number	Precedent activity	Option/ Mode	Duration (days)	Direct cost (\$)
Site preparation	1	-	1	14	23,000
			2	20	18,000
			3	24	12,000
Forms and rebar	2	1	1	15	3,000
			2	18	2,400
			3	20	1,800
			4	23	1,500
			5	25	1,000
Excavation	3	1	1	15	4,500
			2	22	4,000
			3	33	3,200
Precast concrete girder	4	1	1	12	45,000
			2	16	35,000
			3	20	30,000
Pour foundation and piers	5	2, 3	1	22	20,000
			2	24	17,500
			3	28	15,000
			4	30	10,000
Deliver PC girders	6	4	1	14	40,000
			2	18	32,000
			3	24	18,000
Erect girders	7	5, 6	1	9	30,000
			2	15	24,000
			3	18	22,000

The robustness of the new proposed model RKV-TCO in the deterministic situation was compared with two other previous models:

- 1) Gen and Cheng [22] using GC approach;
- 2) Zheng et al. [10] using MAWA with a GA-based approach.

The results of RKV-TCO approach are presented in Table 2. The Table 2 shows the values of time and cost for the first six generations with Gen and Cheng [22] and Zheng et al. [10] approaches. The algorithm RKV-TCO obtains in the third generation a better solution than the works mentioned above. So, the RKV-TCO ends with project time = 63 days and cost = \$225,500 in Table 2.

Additionally we can also state that the RKV-TCO approach produces high-quality solutions quickly once needed only 3 seconds to complete 50 generations.

This computational experience has been performed on a computer with an Intel Core 2 Duo CPU T7250 @2.33 GHz and 1,95 GB of RAM. The algorithm proposed in this work has been coded in VBA under Microsoft Windows NT.

Table 2 Summary of the results.

Approaches	Generation number	Criteria		Calculation Time
		Time	Cost (\$)	
Gen and Cheng [22]	0	83	243,500	Not reported
	1	80	242,400	
	2	80	261,900	
	3	79	256,400	
	4	79	256,400	
	5	79	256,400	
Zheng et al. [10]	0	73	251,500	Not reported
	1	73	251,500	
	2	73	251,500	
	3	66	236,500	
	4	66	236,500	
	5	66	236,500	
This paper	0	73	233,000	3 (two) seconds for 50 generations
	1	68	225,500	
	2	63	225,500	

V. CONCLUSIONS AND FURTHER RESEARCH

A GA based-approach to solving the time-cost optimization problem has been proposed. The project activities have various construction modes, which reflect different ways of performing the activity, each mode having a different impact on the duration and cost of the project. The chromosome representation of the problem is based on random keys. The schedules are constructed using a priority rule in which the priorities are defined by the genetic algorithm. The present approach provides an attractive alternative for the solution of

the construction multi-objective optimization problems.

Further research can be extended to the following directions: extended to more construction project problems to reinforce the results obtained namely expanding the optimization model to consider resource allocation and resource leveling constraints and expanding the number of modes of construction for each activity to turn a more complicated optimization problem.

ACKNOWLEDGMENT

This work has been partially supported by the CIDEM (Centre for Research & Development in Mechanical Engineering). CIDEM is an investigation unit of FCT – Portuguese Foundation for the Science and Technology located at the School of Engineering of Polytechnic of Porto.

REFERENCES

- [1] A. Afshar, A. Ziaraty, A. Kaveh and F. Sharifi, Nondominated Archiving Multicolony Ant Algorithm in Time–Cost Trade-Off Optimization, *J. Constr. Eng. Manage.*, 135(7), 2009, pp. 668-674.
- [2] A. Pagnoni, *Project engineering: computer-oriented planning and operational decision making*. Springer-Verlag, 1990.
- [3] C. C. Coello, An updated survey of GA-based multiobjective optimization techniques, *ACM Comput. Surv.*, 32(2), 2000, pp. 109–143.
- [4] C. Hendrickson and T. Au, *Project management for construction: fundamental concepts for owners, engineers, architects, and builders*, Prentice-Hall International Series in Civil Engineering and Engineering Mechanics, 1989.
- [5] C. Patrick, *Construction Project Planning and Scheduling*, Pearson Prentice Hall, Ohio, 2004.
- [6] C.-W. Feng , L. Liu and S.A. Burns, Using genetic algorithms to solve construction time-cost trade-off problems, *J. Comp. Civ.Engrg.*, ASCE, 11(3), 1997, pp.184-189.
- [7] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley: Reading, MA, 1989.
- [8] D. R. Robinson, A dynamic programming solution to cost-time tradeoff for CPM, *Management Science*, 22(2), 1975, pp.158-166.
- [9] D. X. M. Zheng and S. T. Ng, Stochastic time–cost optimization model incorporating fuzzy sets theory and nonreplaceable front, *Journal of Construction Engineering and Management*, ASCE, 131(2), 2005, pp.176-186.
- [10] D. X.M. Zheng, S.T. Ng and M.M. Kumaraswamy, Applying a genetic algorithm-based multiobjective approach for time-cost optimization, *Journal of Construction Engineering and Management*, 130(2), 2004, pp.168-176.
- [11] I.T. Yang, Using elitist particle swarm optimization to facilitate bicriterion time-cost trade-off analysis, *Journal of Construction Engineering and Management*, ASCE, 133(7), 2007, pp.498-505.
- [12] J. E. Kelley, *Critical-Path Planning and Scheduling: Mathematical Basis*. *Operations Research*, 9(3), 1961, pp. 296-320.
- [13] J.F. Gonçalves, J.J. M. Mendes and M.C.G. Resende. A hybrid genetic algorithm for the job shop scheduling problem. *European Journal of Operational Research*, Vol. 167, 2005, pp.77-95.
- [14] J. H. Patterson and D. Huber, A horizon-varying, zero-one approach to project scheduling, *Management Science*, 20(6), 1974, pp. 990-998.
- [15] J. Magalhaes-Mendes, Project scheduling under multiple resources constraints using a genetic algorithm, *WSEAS Transactions on Business and Economics*, World Scientific and Engineering Academy and Society, USA, Vol. 11, 2008, pp.487-496.
- [16] J. Magalhaes-Mendes, A two-level genetic algorithm for the multi-mode resource-constrained project scheduling problem, *International Journal of Systems Applications, Engineering & Development*, 5(3), 2011, pp. 271-278.

- [17] J. Magalhaes-Mendes, Active, Parameterized Active, and Non-Delay Schedules for Project Scheduling with Multi-Modes, *Proceedings of the 16th WSEAS International Conference on Applied Mathematics*, Montreaux, Switzerland, December 29-31, 2011, pp-134-139.
- [18] J. Magalhaes-Mendes, A hybrid genetic algorithm for the multi-mode resource-constrained project scheduling, *Proceedings of the 6th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2012)*, Vienna University of Technology, Austria, 2012.
- [19] J. W. Fondahl, *A non-computer approach to the critical path method for the construction industry*, Technical Report No. 9, The Construction Institute, Department of Civil Engineering, Stamford University., 1961.
- [20] L. A. Zadeh, Fuzzy sets, *Inf. Control*, 8, 1965, pp.338-353.
- [21] L. Liu, S. Burns and C. Feng, Construction time-cost trade-off analysis using LP/IP hybrid method, *J. Constr. Eng. Manage.*, 121(4), 1995, pp.446-454.
- [22] M. Gen and R. Cheng, *Genetic algorithms & engineering optimization*, Wiley-Interscience, New York, 2000.
- [23] N. Siemens, A simple CPM time-cost trade-off algorithm, *Management Science*, 17(6), 1971, pp. B354-B363.
- [24] O. Moselhi, Schedule compression using the direct stiffness method, *Canadian Journal of Civil Engineering*, 20(1), 1993, pp.65-72.
- [25] P. De, E.J. Dunne, J.B. Ghosh and C. E. Wells, The discrete time-cost trade-off problem revisited, *European Journal of Operational Research*, 81(2), 1995, pp.225-238.
- [26] S. T. Ng and Y. Zhang, Optimizing construction time and cost using ant colony optimization approach, *Journal of Construction Engineering and Management, ASCE*, 134(9), 2008, pp.721-728.
- [27] R. Klein, Bidirectional planning: improving priority rule-based heuristics for scheduling resource constrained projects, *European Journal of Operational Research*, Vol. 127, 2000, pp.619-638.
- [28] S. E. Elmaghraby, Resource allocation via dynamic programming in activity networks, *Eur. J. Operational Res.*, Vol. 64, 1993, pp. 199-215.
- [29] S.A. Burns, L. Liu and C. W. Feng, C.W., The LP/IP hybrid method for construction time-cost trade-off analysis, *Constr. Manag. Econ.*, 14(3), 1996, pp.265-276.
- [30] T. Hegazy, Optimization of construction time-cost trade-off analysis using genetic algorithms, *Canadian Journal of Civil Engineering*, 26(6), 1999, pp. 685-697.
- [31] W. L. Mayer and L.R. Shaffer, Extension of the critical path method through the application of integer programming, *Civ. Engrg. Constr. Res. Ser. 2*, Univ. of Illinois, Urbana, III, 1963.
- [32] W. Prager, A structured method of computing project cost polygons, *Management Science*, 9(3), 394-404, 1963.
- [33] W.S. Butcher, Dynamic programming for project cost-time curve, *Journal of Construction Division, ASCE*, 93(C01), 1967, pp. 59-73.
- [34] Y. Xiong and Y. Kuang, Applying an ant colony optimization algorithm-based multiobjective approach for time-cost trade-off, *Journal of Construction Engineering and Management, ASCE*, 134(2), 2008, pp. 153-156.
- [35] Y.-Z. Luo, G.-J.Tang, Z.-G. Wang and H.-Y. Li, Optimization of perturbed and constrained fuel-optimal impulsive rendezvous using a hybrid approach, *Engineering Optimization*, 38(8), 2006, pp.959-973.
- [36] Z. W. Geem, Multiobjective optimization of time-cost trade-off using harmony search, *Journal of Construction Engineering and Management, ASCE*, Vol. 136, No. 6, 2010, pp.711-716.



J. Magalhães-Mendes was born in Mancelos (Amarante, Portugal) on January 17, 1963.

He has the following academic degrees:

PhD in Mechanical Engineering and Industrial Management by University of Oporto; M.Sc. in Civil Engineering by University of Aveiro; M.Sc. in Systems and Automation by University of Coimbra; Degree in Civil Engineering by Polytechnic of Oporto and Degree in Applied Mathematics by University of Oporto.

He has been Associate Professor of the School of Engineering of Polytechnic of Oporto since January of 2010, where he teaches the courses of organization and management of works and construction management. He has published papers in the European Journal of Operational Research, Computers & Operations Research, Journal of Heuristics, WSEAS/NAUN Journals and several national and international conferences. He has about 350 ISI citations. His research interest includes construction management, project management, genetic algorithms, and operational research and supply chain management.

Normalizations of the Proposal Density in Markov Chain Monte Carlo Algorithms

Antoine E. Zambelli

Abstract—We explore the effects of normalizing the proposal density in Markov Chain Monte Carlo algorithms in the context of reconstructing the conductivity term K in the 2-dimensional heat equation, given temperatures at the boundary points, d . We approach this nonlinear inverse problem by implementing a Metropolis-Hastings Markov Chain Monte Carlo algorithm. Markov Chains produce a probability distribution of possible solutions conditional on the observed data. We generate a candidate solution K' and solve the forward problem, obtaining d' . At step n , with some probability α , we set $K_{n+1} = K'$. We identify certain issues with this construction, stemming from large and fluctuating values of our data terms. Using this framework, we develop normalization terms z_0, z and parameters λ that preserve the inherently sparse information at our disposal. We examine the results of this variant of the MCMC algorithm on the reconstructions of several 2-dimensional conductivity functions.

Keywords—Ill-posed, Inverse Problems, MCMC, Normalization, Numerical Analysis.

I. INTRODUCTION

The idea of an inverse problem is to reconstruct, or retrieve, information from a set of measurements. In many problems, the quantities we measure are different from the ones we wish to study; and this set of d measurements may depend on several elements. Our goal is thus to reconstruct, from the data, that which we wanted to study. In essence, given an effect, what is the cause? For example: If you have measurements of the temperature on a surface, you may want to find the coefficient in the heat equation.

The nonlinearity and ill-posedness of this problem lends itself well to Markov Chain Monte Carlo algorithms. We detail this algorithm in later sections, but we note now that there has been much work done on Metropolis-Hastings MCMC algorithms. However, much of it has been trying to determine optimal proposal densities [3], [5]. Luengo and Martino [3] treat this idea by defining an adaptive proposal density under the framework of Gaussian mixtures. Our work, however, is focused on improving the reconstruction given a proposal density.

We take no views on the optimality of the structure of the proposal density in our case, which we take from [1]. We simply observe possible improvements to this density by normalizing its terms through context-independent formulations. Eventually, we would like to implement the GM-MH algorithm of [3] on our proposal density, and provide a rigorous definition of our construction in an analogous manner to their work.

The paper is structured as follows. We first present the framework of our problem in the subsection below. Section II presents the MHMCMC algorithm and proposal densities along with non-normalized results. The error analysis of those results (in Section III) motivates this work while Sections IV to VI present the new constructions and associated results.

A. Heat Diffusion

In this problem, we attempt to reconstruct the conductivity K in a steady state heat equation of the cooling fin on a CPU. The heat is dissipated both by conduction along the fin and by convection with the air, which gives rise to our equation:

$$u_{xx} + u_{yy} = \frac{2H}{K\delta}u \quad (1)$$

with H for convection, K for conductivity, δ for thickness and u for temperature. The CPU is connected to the cooling fin along the bottom half of the left edge of the fin. We use standard Robin Boundary Conditions with

$$Ku_{normal} = Hu \quad (2)$$

Our data in this problem is the set of boundary points of the solution to (1), which we compute using a standard Crank-Nicolson scheme for an $n \times m$ mesh (here 20×20). We denote the correct value of K by $K_{correct}$ and the data by d . In order to reconstruct $K_{correct}$, we will take a guess K' , solve the forward problem using K' , obtaining d' , and compare those boundary points to d by implementing the Metropolis-Hastings Markov Chain Monte Carlo algorithm (or MHMCMC).

II. METROPOLIS-HASTINGS MCMC

Markov Chains produce a probability distribution of possible solutions (in this case conductivities) that are most likely given the observed data (the probability of reaching the next step in the chain is entirely determined by the current step). The algorithm is as follows (see [1]). Given K_n, K_{n+1} can be found using the following:

- 1) Generate a candidate state K' from K_n with some distribution $g(K'|K_n)$. We can pick any $g(K'|K_n)$ so long as it satisfies
 - a) $g(K'|K_n) = 0 \Rightarrow g(K_n|K') = 0$
 - b) $g(K'|K_n)$ is the transition matrix of the Markov Chain on the state space containing K_n, K' .

2) With probability

$$\alpha(K'|K_n) \equiv \min \left\{ 1, \frac{Pr(K'|d)g(K_n|K')}{Pr(K_n|d)g(K'|K_n)} \right\} \quad (3)$$

set $K_{n+1} = K'$, otherwise set $K_{n+1} = K_n$ (ie. accept or reject K'). Proceed to the next iteration.

More formally, if $\alpha > u \sim U[0, 1]$, then $K_{n+1} = K'$. Using the probability distributions of our example, (3) becomes

$$\alpha(K'|K_n) \equiv \min \left\{ 1, e^{\frac{-1}{2\sigma^2} \sum_{i,j=1}^{n,m} [(d_{ij}-d'_{ij})^2 - (d_{ij}-d_{n_{ij}})^2]} \right\} \quad (4)$$

where d' and d_n denote the set of boundary temperatures from K' and K_n respectively, and $\sigma = 0.1$. To simplify (4), collect the constants and separate the terms relating to K' and K_n :

$$\begin{aligned} \frac{-1}{2\sigma^2} \sum_{i,j=1}^{n,m} [(d_{ij}-d'_{ij})^2 - (d_{ij}-d_{n_{ij}})^2] &= \frac{-1}{2} [f' - f_n] \\ &= -(D_1) \end{aligned}$$

Now, (4) reads

$$\alpha(K'|K_n) \equiv \min \{1, e^{-D_1}\} \quad (5)$$

Note that we are taking this formulation as given, and that the literature mentioned above (most notably Gaussian Mixture based algorithms) would be going from (3) to (4) perhaps differently.

A. Generating K'

To generate our candidate states, we will perturb K_n by a uniform random number $\omega \in [-0.005, 0.005]$. In the simplest case, where we are dealing with a constant K_{correct} , then we could proceed by changing every point in the mesh by ω , and the algorithm converges rapidly.

Looking at non-constant conductivities forces us to change our approach. If we simply choose to change one randomly chosen point at a time, then we have a systemic issue with the boundary points, which exhibit odd behavior and hardly change value. To sidestep this, we will change a randomly chosen grid (2×2) of the mesh at once. Thereby pairing up the troublesome boundary points with the well-behaved inner points.

B. Priors

While a gridwise change enables us to tackle non-constant conductivities, two issues remain. The first is that our reconstructions are still marred with “spikes” of instability. The second, more profound, is that the ill-posedness of the problem means there are in fact infinitely many solutions, and we must isolate the correct one. This brings us to the notion of priors. These can be thought of as weak constraints imposed on our

reconstructions. However, we do not wish to rule out any possibilities, keeping our bias to a minimum. So we define

$$\begin{aligned} T' &= \sum_{j=1}^n \sum_{i=2}^m (K'(i, j) - K'(i-1, j))^2 \\ &\quad + \sum_{i=1}^m \sum_{j=2}^n (K'(i, j) - K'(i, j-1))^2 \end{aligned} \quad (6)$$

$$\begin{aligned} T_n &= \sum_{j=1}^n \sum_{i=2}^m (K_n(i, j) - K_n(i-1, j))^2 \\ &\quad + \sum_{i=1}^m \sum_{j=2}^n (K_n(i, j) - K_n(i, j-1))^2 \end{aligned} \quad (7)$$

let $D_2 = T' - T_n$, and modifying (5), we obtain

$$\alpha_c(K'|K_n) \equiv \min \{1, e^{-\lambda_1 D_1 - \lambda_2 D_2}\} \quad (8)$$

By comparing the smoothness of K' not in an absolute sense, but relative to the last accepted guess, we hope to keep as many solutions as possible open to us, while ensuring a fairly smooth result. We introduce one additional prior, this time imposing a condition on the gradient of our conductivity. The author explores the notion of priors more fully in [7], but much as we take the proposal density as given, the aim of this paper is not to examine priors per se. So we look at the mixed partial derivative of our candidate state and compare it to that of the last accepted guess

$$\begin{aligned} M' &= \sum_{j=1}^n \sum_{i=2}^m (K'_{xy}(i, j) - K'_{xy}(i-1, j))^2 \\ &\quad + \sum_{i=1}^m \sum_{j=2}^n (K'_{xy}(i, j) - K'_{xy}(i, j-1))^2 \end{aligned} \quad (9)$$

$$\begin{aligned} M_n &= \sum_{j=1}^n \sum_{i=2}^m (K_{n_{xy}}(i, j) - K_{n_{xy}}(i-1, j))^2 \\ &\quad + \sum_{i=1}^m \sum_{j=2}^n (K_{n_{xy}}(i, j) - K_{n_{xy}}(i, j-1))^2 \end{aligned} \quad (10)$$

where K'_{xy} and $K_{n_{xy}}$ are computed using central and forward/backward finite difference schemes. We let $D_3 = M' - M_n$ and modify (5) to get

$$\alpha_s(K' | K_n) \equiv \min \{1, e^{-\lambda_1 D_1 - \lambda_3 D_3}\} \quad (11)$$

We now take the acceptance step of our algorithm as

$$\alpha = \max \{\alpha_c, \alpha_s\} \quad (12)$$

So the algorithm seeks to satisfy at least one of our conditions, though not necessarily both. We present some preliminary results in Figure 1 and Figure 2 below. Note that we are clearly on the right path, with the algorithm approaching it's mark, but not to a satisfying degree.

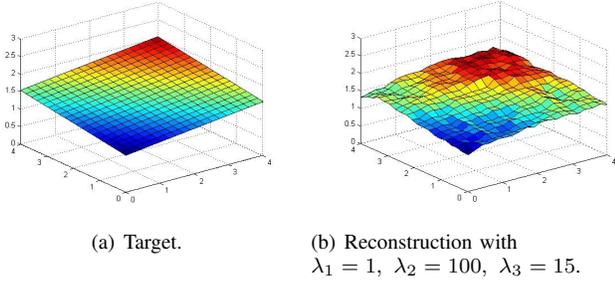


Fig. 1. Reconstruction of a tilted plane with priors, 10 million iterations.

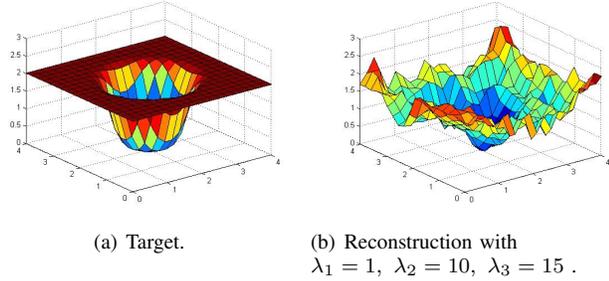


Fig. 2. Reconstruction of a Gaussian well with priors, 10 million iterations.

III. ERROR ANALYSIS

Our work so far has looked at qualitative improvements to our reconstructions, now we seek to quantify those improvements and the performance of the algorithm in general. Several metrics can be used for this purpose, but we will focus our writeup on the following: the difference between the data and the output using our guess (δ), given by

$$\delta = (\delta_1 \cdots \delta_n) \quad , \quad \text{with } \delta_i = \sum (d - d'_i)^2$$

the sum of differences squared between $K_{correct}$ and K_n (β),

$$\beta = \left(\sum (K_{correct} - K_1)^2 \cdots \sum (K_{correct} - K_n)^2 \right)$$

and most importantly, the rate of acceptance of guesses (Γ), where

$$\Gamma_0 = 0 \quad \text{and} \quad \Gamma_i = \begin{cases} \Gamma_{i-1} + 1 & \text{if guess is accepted.} \\ \Gamma_{i-1} & \text{if guess not accepted.} \end{cases}$$

for each subsequent iteration.

The form of Γ is a step function, where accepting every guess would resemble a straight line of slope 1, and accepting none of the guesses results in a slope of 0. The shape of this function should tell us something about when the algorithm is performing best.

A. δ , β , Γ Results

The results of tests involving these parameters reveals some interesting information (see Figure 3). β decreases, as expected, at a decreasing rate over time, slowing down around 6 – 7 million iterations, which seems in line with the

qualitative results.

On the other hand, δ decreases much more rapidly. The difference between the data and simulated temperatures becomes very small starting at as early as 250000 iterations. In a sense, this fits with the problem of ill-posedness, the data is only useful to a certain degree, and it will take much more to converge to a solution (and we have been converging beyond 250k iterations). The most important result, however,

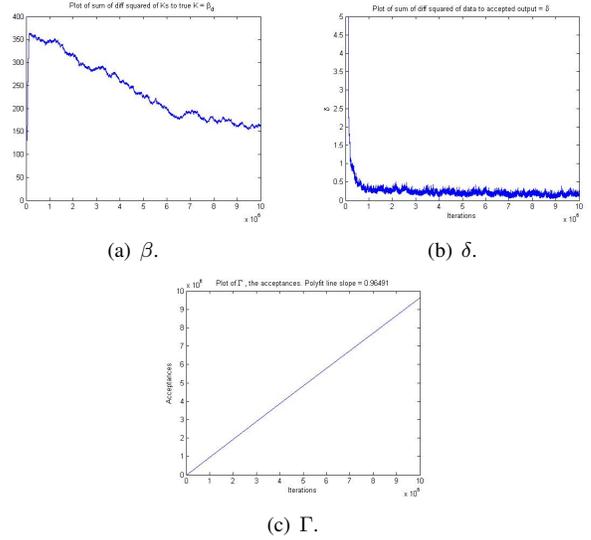


Fig. 3. Plot of the error metrics without normalizations.

comes from Γ . If we fit a line to our step function, we get slopes of 0.95 or more. This means we are accepting nearly every guess. While this could be troubling on its own, the fact that we are accepting at a constant rate as well is indicative of a deeper problem in our method.

Given that Γ is dependent solely on the likelihood of accepting a guess, we take a look at α directly. What we find is that α is evaluating at 1 almost every iteration. The quantities we are looking at within it (comparing data and smoothness) are simply too large. We need to normalize our distribution.

IV. PRELIMINARY STRUCTURE

In the following sections, we examine the impact of normalizations on our data terms, and explore the motivations behind the various constructions. More rigorous data is provided concerning the final form, while the earlier results focus on the concepts that guided their evolution.

One structural change which we will implement is to take equation (12), and change it to be more restrictive. Previously, it was looking for solutions which satisfied at least one of the prior conditions. Here we will instead look for solutions that satisfy all of them at once by setting

$$\alpha(K' | K_n) \equiv \min \left\{ 1, e^{-\left(\sum_{i=1}^3 \lambda_i z_i D_i \right)} \right\} \quad (13)$$

where z_i are as-of-yet undetermined normalization terms.

A. Motivation

We first take a moment to examine the sensitivities λ_i , and impose the following condition: $\lambda_1 > \lambda_2$ and $\lambda_1 > \lambda_3$. Not doing so would mean the algorithm could give us some false positives. This leads us to notice that a key aspect of the MHMCMC method is information. Due to the ill-posed nature of the problem, we need to keep every piece of information that can be gleaned. We will keep this idea in mind throughout the later sections.

As for the normalizations proper, the naive approach to our problem would be to divide each data term by a constant value. In this formulation, our normalization terms would have the form

$$z_i = \frac{1}{c_i} \quad (14)$$

where c_i can be determined by looking at representative values of our data terms.

This approach has one advantage, which is that it retains information very well. The relationships between quantities is affected by a constant factor, and its evolution is therefore preserved across iterations. Unfortunately, this method is very unstable, and is not particularly viable. One can think of the opposite method to this one being dividing each data term by itself. Clearly, this would erase all information contained within our results, but it would successfully normalize it, given a broad enough definition of success.

Concretely, we seek to find a normalization that delivers information about the evolution of our data terms, but bounds the results so that we may control their magnitudes and work with their relative relationships.

V. NORMALIZED WITH INERTIA

We introduce the concept of inertia in this framework. Inertia can be thought of as the weight (call it w) being applied to either previous method. Though we do not want to divide by only a constant, there is merit to letting some information trickle through to us. If we do not bound the quantities we are examining, then we will obtain very small or very large values for α , effectively 0 or 1, which is undoing the work of the MHMCMC. We attempt to bound our likelihood externally. We define α_h such that

$$\alpha(K' | K_n) \equiv z_0 \alpha_h = z_0 e^{-\left(\sum_i \lambda_i z_i D_i\right)} \quad (15)$$

A. Global Normalizations

Even a cursory analysis of our early attempts at solving this heat conductivity problem have revealed a desperate need to correctly normalize our data in order to get meaningful likelihoods. Some issues of note have been the idea that the inertia of the process, the value of previous guesses, contains information which is important to the successful convergence of our algorithm. Another is the fact that the variance of data terms means that we require a strong normalization term,

at the expense, perhaps, of information, if we are to obtain meaningful results.

Addressing the second point, we decide to deviate slightly from one aspect of our method, and use a global result. Computationally, we will only be tracking one variable, and this poses no problem. But note that using a global result in computing α implies that our process is no longer a Markov process, as the probability of reaching the next step is dependent on the past and not just the present.

B. Formulation of $Z^{(1)}$

First, let $\alpha_{h,m} = \max_j \{\alpha_{h,j}\}$, $\forall j$ and $D_{i,m} = \max_j \{D_{i,j}\}$, $\forall j$. We denote $Z^{(1)}$ the normalization

$$z_{0,j}^{(1)} = w_0 \frac{1}{\alpha_{h,j}} + (1 - w_0) \frac{1}{\alpha_{h,m}} \quad (16)$$

$$z_{i,j}^{(1)} = w \frac{1}{|D_{i,j}|} + (1 - w) \frac{1}{|D_{i,m}|} \quad (17)$$

While this effectively bounds our acceptance probability between $[0, 1]$, it does so at the expense of the Markov property of our algorithm. Removing this property exhibits some instability in the evolution of the algorithm. Namely, they appear to converge to false positives, an effect which must be explored more fully.

C. Restricted Random Interval

Examining the values of α that we now produce reveals that we have greatly tightened the spread. Almost all of our values are contained in a narrow band (which changes depending on parameters), say between 0.6 and 0.75. Again, this means we are losing information, as the difference in the values of α are lost by comparing them over the entire $[0, 1]$ interval.

We change the 2nd step in the MHMCMC algorithm, which was $\alpha > u \sim U[0, 1] \Rightarrow K_{n+1} = K'$. We now restrict the interval over which we draw u , taking its lower and upper bounds at the j th iteration to be $[u_{\min}, u_{\max}]$, where for some small constant ζ ,

$$u_{\min} = \min_{i < j} \alpha_i - \zeta \quad \wedge \quad u_{\max} = \max_{i < j} \alpha_i + \zeta \quad (18)$$

While perhaps more restrictive, this formulation also greatly increases the speed at which the algorithm begins to converge by effectively selecting those guesses which are the most promising, relative to the past performance of the algorithm. This method implies that we will not, with probability 1, decide the outcome of a guess, they simply become (as per ζ) extremely unlikely to be accepted or rejected.

VI. LOCALLY FOCUSED NORMALIZATION

We now attempt to modify $Z^{(1)}$ in order to retain the original Markov property of the algorithm. The property was violated in the second term, which unfortunately also guarantees we bound our results.

A. Formulation of $Z^{(2)}$

Denote a new normalization scheme $Z^{(2)}$, given by

$$z_{0,j}^{(2)} = w_0 \frac{1}{\alpha_{h,j}} + (1 - w_0) \frac{1}{\alpha_{h,j-1}} \quad (19)$$

$$z_{i,j}^{(2)} = w \frac{1}{|D_{i,j}|} + (1 - w) \frac{1}{|D_{i,j-1}|} \quad (20)$$

While we have recovered the Markov property, we must now contend with unbounded values for α . We note now that preliminary attempts to use $z_{i,j}^{(2)}$ with $z_{0,j}^{(1)}$ did not yield promising results.

While this formulation provides good results, it does require us to find an empirical bound for α , as it is no longer bounded by z_0 . For the results presented below, we imposed $\alpha \in [0, 1.5]$, setting

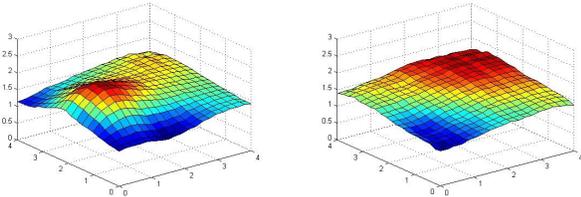
$$\alpha(K' | K_n) = \min \left\{ 1.5, z_0 e^{-\left(\sum_{i=1}^3 \lambda_i z_i D_i\right)} \right\} \quad (21)$$

B. Results

The parameters we have to determine are $\lambda_1, \lambda_2, \lambda_3, w, w_0$ and the cutoff for α as in (21). We have concluded we must set $\lambda_1 > \lambda_i, \forall i > 1$ and we have by definition $w, w_0 \in [0, 1]$. The exact values of the sensitivities and inertia factors are at the moment heuristically chosen to be

$$\begin{aligned} \lambda_1 &= 0.5, \quad \lambda_2 = 0.15, \quad \lambda_3 = 0.45 \\ w_0 &= 0.1, \quad w = 0.75, \quad \alpha_{\text{cutoff}} = 1.5 \end{aligned}$$

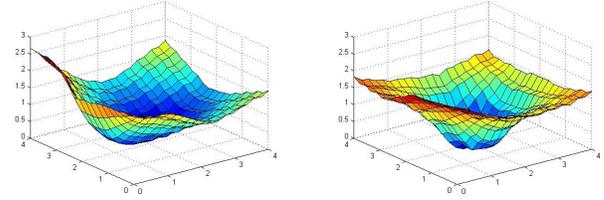
For the tilted plane, we obtain Figure 4. As mentioned



(a) Reconstruction using $Z^{(1)}$. (b) Reconstruction using $Z^{(2)}$.

Fig. 4. $Z^{(1)}$ and $Z^{(2)}$ reconstructions of a tilted plane with priors, 2 million iterations.

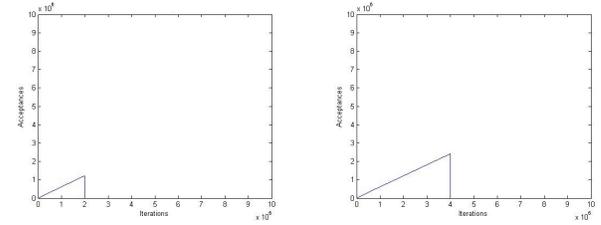
in Section V-B, we have some instability in the form of incorrect convergence for $Z^{(1)}$, which is apparent in Figure 5 as well. On the other hand, $Z^{(2)}$ converges well and produces a smooth reconstruction. We can also note that it achieves slightly better results than the no-normalizations case in only 2 million iterations. The instability in $Z^{(1)}$ is again apparent, and leads us to conclude that the loss of the Markov property in the algorithm may be detrimental to its performance. However, the reconstruction of the Gaussian well has substantially improved when using $Z^{(2)}$. It achieves a smoother reconstruction as without normalizations (see Figure 2), and in $4M$ iterations instead of $10M$.



(a) Reconstruction using $Z^{(1)}$. (b) Reconstruction using $Z^{(2)}$.

Fig. 5. $Z^{(1)}$ and $Z^{(2)}$ reconstructions of a Gaussian well with priors, 4 million iterations.

Going back to our error metric Γ , we see the improvement manifest itself rather clearly, with acceptances being on the order of $\sim 55\%$ instead $\sim 95\%$ as they were before.



(a) $\Gamma_{Z^{(2)}}$ for tilted plane. (b) $\Gamma_{Z^{(2)}}$ for Gaussian well.

Fig. 6. Plots of Γ for $Z^{(2)}$ reconstructions with priors.

VII. CONCLUSION

The need for normalizing factors arose from the variance in the magnitudes of data terms D_i from one iteration to the next. In formulating those factors, we focused on conserving the information contained in D_i while bounding our quantities, and we confirmed the importance of retaining the Markov property in this context. However, by using the $Z^{(2)}$ formulation, we were able to obtain faster and better reconstructions of the conductivity for both the tilted plane and the Gaussian well.

Despite the encouraging results, several avenues need to be explored more fully. The long-run behavior of $Z^{(2)}$ seems to exhibit some stagnation, seemingly having converged as best as it can. In addition, very preliminary results have been obtained for a scheme that lies between $Z^{(1)}$ and $Z^{(2)}$, which updates the $(1 - w)$ terms only when a guess is accepted, has shown competitive performance relative to $Z^{(2)}$.

As the algorithm currently stands α_{cutoff} , the sensitivities λ_i , and the inertia factors w, w_0 must be determined heuristically. It is possible we may be able to dynamically adjust the values as the algorithm runs, through a constrained optimization of the acceptance rate, but that remains to be studied.

Finally, we would like to implement Gaussian-Mixture based MCMC algorithms, that treat the proposal density as an unknown to be approximated, and combine this framework

with our normalization schemes to observe the interaction of the two methods.

REFERENCES

- [1] Fox, C., Nicholls, G., Tan, S. *Inverse Problems, Physics 707*, The University of Auckland, ch. 7-9.
- [2] Hastings, W. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." *Biometrika*, Vol 57, No. 1, (1970), pp. 97-109.
- [3] Luengo, D., Martino, L. "Fully Adaptive Gaussian Mixture Metropolis-Hastings Algorithm" *Proc. ICASSP 2013*, Vancouver (Canada), pp. 6148-6152.
- [4] Metropolis, N., Rosenbluth, A., et. al. "Equations of State Calculations by Fast Computing Machines" *Journal of Chemical Physics*, Vol 21 (1953), pp. 1087-1092.
- [5] Rosenthal, J. "Optimal Proposal Distributions and Adaptive MCMC." Chapter for *MCMC Handbook* (2010), avail. at <http://www.probability.ca/jeff/ftpdir/galinart.pdf>
- [6] Sauer, T., *Numerical Analysis*, Pearson Addison-Wesley, 2006.
- [7] Zambelli, A., "A Multiple Prior Monte Carlo Method for the Backward Heat Diffusion Problem" *Proc. CMMSE 2011*, Benidorm (Spain), Vol 3, pp. 1192-1200.

Modeling the virtual machine allocation problem

Zoltán Ádám Mann

Abstract—Finding the right allocation of virtual machines (VM) in cloud data centers is one of the key optimization problems in cloud computing. Accordingly, many algorithms have been proposed for the problem. However, lacking a single, generally accepted formulation of the VM allocation problem, there are many subtle differences in the problem formulations that these algorithms address; moreover, in several cases, the exact problem formulation is not even defined explicitly. Hence in this paper, we present a comprehensive generic model of the VM allocation problem. We also show how the often-investigated problem variants fit into this general model.

Keywords—Virtual machines, VM placement, VM consolidation, Cloud computing, Data centers

I. INTRODUCTION

Workload allocation in data centers (DCs) has been an important optimization problem for decades [7]. More recently, the wide spread of virtualization technologies and the cloud computing paradigm have established several new possibilities for resource provisioning and workload allocation [4], opening up new optimization opportunities.

Virtualization makes it possible to co-locate multiple applications on the same physical machine (PM) in logically isolated virtual machines (VMs). This way, a high utilization of the available physical resources can be achieved, thus amortizing the capital and operational expenditures associated with the purchase, operation, and maintenance of the DC resources. What is more, *live migration* of VMs makes it possible to move a VM from one PM to another one without noticeable service interruption [2]. This enables the dynamic re-optimization of the allocation of VMs to PMs, reacting to changes in the VMs' workload and the PMs' availability.

Consolidating multiple VMs on relatively few PMs helps not only to achieve good utilization of hardware resources, but also to save energy because unused PMs can be switched off or at least to a low-energy state such as sleep mode. However, too aggressive consolidation may lead to performance degradation. In particular, if the load of some VMs starts to grow, this may result in an overload of the accommodating PM's resources. In many cases, the expected performance levels are laid down in a service level agreement (SLA), defining also penalties if the provider fails to comply. Thus, the provider must find the right balance between the conflicting goals of utilization, energy efficiency, and performance [11].

Beside virtualization and live migration, the most important characteristic of the cloud computing paradigm is the availability of online services with practically unbounded

capacity that can be provisioned elastically as needed. This includes Software-as-a-Service, Platform-as-a-Service, and Infrastructure-as-a-Service [21]. In the latter case, VMs are directly offered to customers; in the first two cases, VMs can be used to provision virtualized resources for the services in a flexible manner. Given the multitude of available public cloud offerings with different capabilities and pricing schemes, it is increasingly difficult for customers to make the best selection for their needs. The problem is further complicated by *hybrid cloud* setups that are increasingly popular in enterprises [5]. In this case, VMs can be either placed on PMs in the own DC(s) or using offerings from external providers, thus further enlarging the search space.

Since the allocation of VMs is an important and challenging optimization problem, several algorithms have been proposed for it. However, as shown in a recent survey, the existing literature includes a multitude of different problem formulations, making the existing approaches hardly comparable [13]. Even worse, some existing works failed to explicitly and precisely define the version of the problem that they are addressing, so that this must be figured out from the algorithm that they proposed or from the way the algorithm was evaluated.

We believe that addressing an algorithmic problem should start with *problem modeling*: a thorough consideration of the problem's characteristics and their importance or non-importance, leading to one or more precisely defined – preferably formalized – problem formulation(s) that capture the important characteristics of the problem. Then and only then should algorithms be proposed if the problem is already well-understood and well-defined. It seems that in the case of the VM allocation problem, this critically important phase was skipped, resulting in a rather chaotic situation where algorithms for “the VM allocation problem” actually address many different problems with sometimes subtle, sometimes serious differences.

The aim of this paper is to remedy this deficiency. Specifically, we devise a rather general formulation of the VM allocation problem that includes most of the problem formulations studied so far in the literature as special cases. We provide a taxonomy of important special cases and analyze their complexity. Section II contains the general problem model and Section III discusses special cases, followed by our conclusions in Section IV.

II. GENERAL PROBLEM MODEL

We consider a *Cloud Provider (CP)* that provides VMs for its customers. For provisioning, the CP can use either its own PMs or external cloud providers (eCPs). The CP

Z.Á. Mann is with the Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Budapest, Hungary

attempts to find the right balance between the conflicting goals of cost-efficiency, energy-efficiency, and performance. In the following, we describe the details of the problem.

A. Hosts

Let D denote the set of data centers available to the CP. For data center $d \in D$, let P_d denote the set of PMs available in d , also including any switched-off PMs. Furthermore, $P = \bigcup\{P_d : d \in D\}$ is the set of all PMs.

Each PM $p \in P$ is characterized by the following numbers:

- $cores(p) \in \mathbb{N}$: number of processor cores
 - $cpu_capacity(p) \in \mathbb{R}_+$: processing power per CPU core, e.g., in MIPS (million instructions per second)
 - $capacity(p, r) \in \mathbb{R}_+$: capacity of resource type $r \in R$.
- For example, R can contain the resource types RAM and HDD, so that the capacity of these resources are given for each PM (e.g., in GB). This should be the net capacity available for VMs, not including the capacity reserved for the OS, the virtualization platform, and other system services

Our approach to model the CPU explicitly and all other resources of a PM through the generic *capacity* function has several advantages. First, this gives maximum flexibility regarding the number of resource types that are taken into account. For instance, also caches, SSD drives, network interfaces, or GPUs can be considered, if relevant. On the other hand, the CPU is quite special, particularly because of multi-core technology. A multi-core processor is not equivalent to a single-core processor of capacity $cores(p) \cdot cpu_capacity(p)$. It is also not appropriate to model each core as a separate resource, because VMs' processing power demand is not specific to each core of the PM, but rather to the set of its cores as a whole. The other reason why it makes sense to model the CPU separately is the impact that the CPU load has on energy consumption.

Each PM $p \in P$ has a set of possible states, denoted by $States(p)$. $States(p)$ always contains the state *On*, in which the PM is capable of running VMs. In addition, $States(p)$ may contain a finite number of low-power states (e.g., *Off* and *Sleep*). Each PM $p \in P$ and $state \in States(p)$ is associated with a static power consumption of $static_power(p, state)$ per time unit. In addition, the *On* state also incurs a dynamic power consumption depending on the PM's load, as defined later. The possible state transitions are given in the form a directed graph $(States(p), Transitions(p))$, where a *transition* $\in Transitions(p)$ is an arc from one state to another. For each *transition* $\in Transitions(p)$, $delay(transition)$ and $energy(transition)$ denote the time it takes to move from the source to the target state and the energy consumption associated with the transition, respectively. (It should be noted that most existing works do not model PM states and transitions in such detail; an exception is the work of Guenter et al. [10].)

Let E denote the set of eCPs from which the CP can lease VMs. For each eCP $e \in E$, $Types(e)$ denotes the set of VM types that can be leased from e , and $Types = \bigcup\{Types(e) : e \in E\}$ is the set of VM types available from at least one eCP.

Each VM type $type \in Types$ is characterized by the same set of parameters as PMs: $cores(type)$, $cpu_capacity(type)$, and $capacity(type, r)$ for all $r \in R$. In addition, for an eCP $e \in E$ and a VM type $type \in Types(e)$, $fee(type, e)$ specifies the fee per time unit for leasing one instance of the given VM type from this eCP. It should be noted that the same VM type may be available from multiple eCPs, potentially for different fees.

Since VMs can be either hosted by a PM or mapped to a VM type of an eCP, let

$$Hosts = P \cup \{(e, type) : e \in E, type \in Types(e)\}$$

denote the set of all possible hosts.

B. VMs

What we defined so far is mostly constant: although sometimes new PMs are installed or existing PMs are taken out of service, eCPs sometimes introduce new VM types or change rental fees, such changes are rare, and can be seen as special events. On the other hand, the load of VMs changes incessantly, sometimes quite quickly. For the purpose of modeling such time-variant aspects, let $Time \subseteq \mathbb{R}$ denote the set of investigated time instances. We make no restriction on $Time$: it can be discrete or continuous, finite or infinite etc.

The set of VMs in time instance $t \in Time$ is denoted by $V(t)$. For each VM $v \in V(t)$, $cores(v)$ is the number of processor cores of v . The CPU load of v in time instance t is a $cores(v)$ -dimensional vector: $vcpu_load(v, t) \in \mathbb{R}_+^{cores(v)}$, specifying the computational load per core, e.g., in MIPS. The load of the other resources is given by $vload(v, r, t) \in \mathbb{R}_+$ for a VM $v \in V(t)$, resource type $r \in R$, and time instance $t \in Time$.

It should be noted that all the cores of a PM's CPU are expected to have the same capacity. In contrast, the cores of the CPU of a VM do not have to have the same load.

C. Mapping VMs to hosts

The CP's task is to maintain a mapping of the VMs to the available hosts. Formally, this is a function

$$Map : \{(v, t) : t \in Time, v \in V(t)\} \rightarrow Hosts.$$

$Map(v, t)$ defines the mapping of VM v in time instance t to either a PM or a VM type of an eCP. Furthermore, if $Map(v, t) = p \in P$, that is, the VM v is mapped to a PM p , then also the mapping of processor cores must be defined, since p may have more cores than v and each core of p may be shared by multiple VM cores, possibly belonging to multiple VMs. Hence in such a case, the function

$$Map_core_v : \{1, \dots, cores(v)\} \times Time \rightarrow \{1, \dots, cores(p)\}$$

defines for each core of v the accommodating core of p , in a given time instance.

Given the mapping of VMs, the load of a PM can be calculated. For a PM $p \in P$ and time instance $t \in Time$, let

$$V(p, t) = \{v \in V(t) : Map(v, t) = p\}$$

be the set of VMs mapped to p in t . The CPU load of p in time instance t is a $cores(p)$ -dimensional vector: $pcpu_load(p, t) \in \mathbb{R}_+^{cores(p)}$, the i th coordinate of which is the sum of the load of the VM cores mapped to the i th core of p , that is:

$$pcpu_load(p, t)_i = \sum_{\substack{v \in V(p, t), \\ Map_core_v(j, t) = i}} vcpu_load(v, t)_j.$$

Similarly, for a resource type $r \in R$, the load of PM p with respect to r in time t is

$$pload(p, r, t) = \sum_{v \in V(p, t)} vload(v, r, t).$$

The dynamic power consumption of a PM p is a monotonously increasing function of its CPU load. This function can be different for each PM. Hence, for a PM $p \in P$, let $dynamic_power_p : \mathbb{R}_+^{cores(p)} \rightarrow \mathbb{R}_+$ define the dynamic power consumption of p per time unit as a function of the load of its cores. This function is monotonously increasing in all of its coordinates. If PM p is in the On state between time instances t_1 and t_2 , then its dynamic energy consumption in this time interval is given by

$$\int_{t=t_1}^{t_2} dynamic_power_p(pcpu_load(p, t)) dt. \quad (1)$$

D. Data transfer

For each pair of VMs, there may be communication between them. The intensity of the communication between VMs $v_1, v_2 \in V$ in time instance $t \in Time$ is denoted by $vcomm(v_1, v_2, t)$, given for example in MB/s. If there is no communication between the two VMs in t , then $vcomm(v_1, v_2, t) = 0$. The communication between a pair of hosts $h_1, h_2 \in H$ is the sum of the communication between the VMs that they accommodate, i.e.,

$$pcomm(h_1, h_2, t) = \sum_{\substack{v_1, v_2 \in V(t), \\ Map(v_1, t) = h_1, \\ Map(v_2, t) = h_2}} vcomm(v_1, v_2, t).$$

For each pair of hosts $h_1, h_2 \in Hosts$, the bandwidth available for the communication between them is $bandwidth(h_1, h_2)$, given for example in MB/s.

E. Live migration

The migration of a VM v from a host h_1 to another host h_2 takes time $mig_time(v, h_1, h_2)$. During this period of time, both h_1 and h_2 are occupied by v . This phenomenon can be modeled by the introduction of an extra VM v' . Let t_{start} and t_{end} denote the time instances in which the migration starts and ends, respectively. Before t_{start} , only v exists, and is mapped to h_1 . Between t_{start} and t_{end} , v continues to occupy h_1 , but starting with t_{start} , also v' appears, mapped to h_2 . In t_{end} , v is removed from h_1 , and only v' remains. Furthermore, data transfer of intensity $mig_comm(v)$ takes place between v and v' during the migration period, which is added to $pcomm(h_1, h_2, t)$.

F. SLA violations

Normally, the load of each resource must be within its capacity. A resource overload, on the other hand, may lead to an SLA violation. Specifically:

- If, for a PM $p \in P$ and one of its processor cores $1 \leq i \leq cores(p)$, $pcpu_load(p, t)_i \geq cpu_capacity(p)$, then this processor core is overloaded, resulting in SLA violation for all VMs using this core, i.e., for each VM $v \in V(p, t)$, for which there is a core of v , $1 \leq j \leq cores(v)$, such that $Map_core_v(j, t) = i$.
- Similarly, if, for a PM $p \in P$ and resource type $r \in R$, $pload(p, r, t) \geq capacity(p, r)$, then this resource is overloaded, resulting in SLA violation for all VMs using this resource, i.e., for each VM $v \in V(p, t)$, for which $vload(v, r, t) > 0$.
- Assume that $Map(v, t) = (e, type)$, where $e \in E$. An SLA violation occurs relating to v , if either $vcpu_load(v, t)_i \geq cpu_capacity(type)$ for some $1 \leq i \leq cores(v)$ or if $vload(v, r, t) \geq capacity(type, r)$ for some $r \in R$.
- If, for a pair of hosts $h_1, h_2 \in Hosts$, $pcomm(h_1, h_2, t) \geq bandwidth(h_1, h_2)$, then the communication channel between the two hosts is overloaded, resulting in SLA violation for all VMs contributing to the communication between these hosts. That is, the set of affected VMs is $\bigcup \{v_1, v_2\} : Map(v_1, t) = h_1, Map(v_2, t) = h_2, vcomm(v_1, v_2, t) > 0\}$.

It should be noted that, in practice, loads will never exceed capacities. However, the loads in the above definitions are calculated as the sum of the loads of the relevant VMs; such a sum can exceed the capacity, and this indeed is a sign of an overload.

In any case, if there is an SLA violation relating to VM v , this leads to a penalty of

$$SLA_fee(v, \Delta t), \quad (2)$$

where Δt is the duration of the SLA violation. The SLA violation fee may be linear in Δt , but it is also possible that longer persisting SLA violations are progressively penalized [9].

In principle, there can be two kinds of SLAs: hard SLAs must be fulfilled in any case, whereas soft SLAs can be violated, but this incurs a penalty. Our above definition allows both: hard SLAs can be modeled with an infinite SLA_fee , whereas soft SLAs are modeled with finite SLA_fee .

G. Optimization objectives

Based on the above definitions, the total power consumption of the CP for a time interval $[t_1, t_2]$ can be calculated as the sum of the following components:

- For each PM p , the interval $[t_1, t_2]$ can be divided into subintervals, in which p remained in the same state. For such a subinterval of length Δt , the static power consumption of p is $static_power(p, state) \cdot \Delta t$. The sum of these values is the total static power consumption of p .

- For each PM p and each state transition of p , $energy(transition)$ is consumed.
- For each PM p and each subinterval of $[t_1, t_2]$ in which p is in state On , the dynamic power consumption is calculated as in Equation (1).

The total monetary cost can be calculated as the sum of the following components:

- The fees to be paid to eCPs. Assume that for $t \in [t_1, t_2]$, $Map(v, t) = (e, type)$, where $e \in E$. This incurs a cost of $(t_2 - t_1) \cdot fee(type, e)$. This must be summed for all VMs mapped to an eCP.
- SLA violation fees, calculated according to Equation 2, for all SLA violations.
- The cost of the consumed power, which is the total power consumption, as calculated above, times the unit power cost.

The objective is to minimize the total monetary costs, by means of optimal arrangement of the Map and Map_core functions and the PMs' states. As a special case, if the other costs are assumed to be 0, the objective is to minimize the overall power consumption of the CP.

It should be noted that there is no need to explicitly constrain or minimize the number of migrations. Rather, the impact of migrations is already contained in the objective function in the form of increased power consumption and potentially SLA violations because of increased system load. (With appropriate costs of migrations and SLA fees, it is possible to also model constraints on migrations, if necessary.)

III. IMPORTANT SPECIAL CASES AND SUBPROBLEMS

The above problem formulation is very general. Most authors investigated simpler problem formulations. We introduced some important special cases and subproblems in [13] and categorized the existing literature on the basis of these problem variants. In the following, we show how these problem variants can be obtained as special cases of our general model. It should be noted that the addressed problem variants are not necessarily mutually exclusive, so that combinations of them are also possible.

A. The Single-DC problem

The subproblem that has received the most attention is the Single-DC problem. In this case, $|D| = 1$ and $|E| = 0$, i.e., the CP has a single DC with a number of PMs, and its aim is to optimize the utilization of these PMs. $|P|$ is assumed to be high enough to serve all customer requests, so that no eCPs are needed. Since all PMs are co-located, $bandwidth$ is usually assumed to be uniform and sufficiently high so that the constraint that it represents can be ignored.

Some representative examples of papers dealing with this problem include [1], [2], [18], [20].

B. The Multi-DC problem

This can be seen as a generalization of the Single-DC problem, in which the CP possesses more than one DC. On the other hand, this is still a special case of our general problem

formulation, in which $|D| > 1$ and $|E| = 0$. An important difference between the Single-DC and Multi-DC problems is that in the latter, communication between DCs is a non-negligible factor. Moreover, the DCs can have different characteristics regarding energy efficiency and carbon footprint. This problem variant, although important, has received relatively little attention [12], [15].

C. The Multi-IaaS problem

In this case, $P = \emptyset$, i.e., the CP does not own any PMs, it uses only leased VMs from multiple IaaS providers. Since there are no PMs, all concerns related to them – states and state transitions, sharing of resources among multiple VMs, load-dependent power consumption – are void. Power consumption plays no role, the only goal is to minimize the monetary costs. On the other hand, $|E| > 1$, so that the choice among the external cloud providers becomes a key question, based on offered VM characteristics and prices. In this case, it is common to also consider the data transfer among VMs.

The Multi-IaaS problem has quite rich literature. Especially popular is the case when communication among the VMs is given in form of a directed acyclic graph (DAG), the edges of which also represent dependencies. Representative examples include [8], [17], [19].

D. Hybrid cloud

This is actually the most general case, in which $|D| \geq 1$ and $|E| \geq 1$. Despite its importance, only few works address it [3], [6].

E. The One-dimensional consolidation problem

In this often-investigated special case, only the computational demands and computational capacities are considered, and no other resources. In our general model, this special case is obtained when the CPU is the only resource considered, and the CPU is taken to be single-core, making the problem truly one-dimensional. That is, $R = \emptyset$ and $cores \equiv 1$.

Whether a single dimension is investigated or also others (e.g., memory or disk), is independent from the number of DCs and eCPs. In other words, all of the above problem variants (Single-DC, Multi-DC, Multi-IaaS, Hybrid cloud) can have a special case of one-dimensional optimization.

F. The On/Off problem

In this case, each PM has only two states: $States(p) = \{On, Off\}$ for each $p \in P$. Furthermore, $static_power(p, Off) = 0$, $static_power(p, On)$ is the same positive constant for each $p \in P$, and $dynamic_power_p \equiv 0$ for each $p \in P$. Between the states On and Off , the transition is possible in both directions, with $delay(transition)$ and $energy(transition)$ both assumed to be 0. As a consequence, the aim is simply to minimize the number PMs that are on. This is an often-investigated special case of the Single-DC problem.

G. Connections to bin-packing

The special case of the Single-DC problem, in which a single dimension is considered, power modeling is reduced to the On/Off problem, all PMs have the same capacity, there is no communication among VMs, migration costs are 0, and hard SLAs are used, is equivalent to the well-known bin-packing problem, since the only objective is to pack the VMs, as one-dimensional objects, into the minimal number of unit-capacity PMs. This has an important consequence: since bin-packing is known to be NP-hard in the strong sense [14], it follows that all variants of the VM allocation problem that contain this variant as special case are also NP-hard in the strong sense.

If multiple dimensions are taken into account, then we obtain a well-known multi-dimensional generalization of bin-packing, the vector packing problem [16].

IV. CONCLUSIONS

In this paper, we attempted to lay a more solid foundation for research on the VM allocation problem. Specifically, we presented a detailed problem formalization that is general enough to capture all important aspects of the problem. We showed how some often-investigated problem variants can be obtained as special cases of our general model. Our work can also be seen as a taxonomy of problem variants, filling the problem modeling gap in the literature between the physical problem and the proposed algorithms. We hope that this will catalyze further high-quality research on VM allocation by showcasing the variety of problem aspects that need to be addressed as well as by defining a set of standardized models to build on. This will hopefully improve the comparability of the proposed algorithms, thus contributing to the maturation of the field.

ACKNOWLEDGMENTS

This work was partially supported by the Hungarian Scientific Research Fund (Grant Nr. OTKA 108947).

REFERENCES

- [1] Anton Beloglazov, Jemal Abawajy, and Rajkumar Buyya. Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future Generation Computer Systems*, 28:755–768, 2012.
- [2] Norman Bobroff, Andrzej Kochut, and Kirk Beaty. Dynamic placement of virtual machines for managing SLA violations. In *10th IFIP/IEEE International Symposium on Integrated Network Management*, pages 119–128, 2007.
- [3] Ruben Van den Bossche, Kurt Vanmechelen, and Jan Broeckhove. Cost-optimal scheduling in hybrid IaaS clouds for deadline constrained workloads. In *IEEE 3rd International Conference on Cloud Computing*, pages 228–235, 2010.
- [4] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic. Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25(6):599–616, 2009.
- [5] Capgemini. Simply. business cloud. http://www.capgemini.com/resource-file-access/resource/pdf/simply_business_cloud_where_business_meets_cloud.pdf (last accessed: February 10, 2015), 2013.
- [6] Emiliano Casalicchio, Daniel A. Mencia, and Arwa Aldhalaan. Automatic resource provisioning in cloud systems with availability goals. In *Proceedings of the 2013 ACM Cloud and Autonomic Computing Conference*, 2013.
- [7] Jeffrey S. Chase, Darrell C. Anderson, Prachi N. Thakar, and Amin M. Vahdat. Managing energy and server resources in hosting centers. In *Proceedings of the 18th ACM Symposium on Operating Systems Principles*, pages 103–116, 2001.
- [8] Thiago A. L. Genez, Luiz F. Bittencourt, and Edmundo R. M. Madeira. Workflow scheduling for SaaS/PaaS cloud providers considering two SLA levels. In *Network Operations and Management Symposium (NOMS)*, pages 906–912. IEEE, 2012.
- [9] Daniel Gmach, Jerry Rolia, Ludmila Cherkasova, and Alfons Kemper. Resource pool management: Reactive versus proactive or let’s be friends. *Computer Networks*, 53(17):2905–2922, 2009.
- [10] Brian Guenter, Navendu Jain, and Charles Williams. Managing cost, performance, and reliability tradeoffs for energy-aware server provisioning. In *Proceedings of IEEE INFOCOM*, pages 1332–1340. IEEE, 2011.
- [11] Gueyoung Jung, Matti A. Hiltunen, Kaustubh R. Joshi, Richard D. Schlichting, and Calton Pu. Mistral: Dynamically managing power, performance, and adaptation cost in cloud infrastructures. In *IEEE 30th International Conference on Distributed Computing Systems (ICDCS)*, pages 62–73, 2010.
- [12] Atefeh Khosravi, Saurabh Kumar Garg, and Rajkumar Buyya. Energy and carbon-efficient placement of virtual machines in distributed cloud data centers. In *Euro-Par 2013 Parallel Processing*, pages 317–328. Springer, 2013.
- [13] Zoltán Ádám Mann. Allocation of virtual machines in cloud data centers – a survey of problem models and optimization algorithms. http://www.cs.bme.hu/~mann/publications/Preprints/Mann_VM_Allocation_Survey.pdf, 2015.
- [14] Silvano Martello and Paolo Toth. *Knapsack problems: algorithms and computer implementations*. John Wiley & Sons, 1990.
- [15] Kevin Mills, James Filliben, and Christopher Dabrowski. Comparing vm-placement algorithms for on-demand clouds. In *Proceedings of the 3rd IEEE International Conference on Cloud Computing Technology and Science*, pages 91–98, 2011.
- [16] Mayank Mishra and Anirudha Sahoo. On theory of vm placement: Anomalies in existing methodologies and their mitigation using a novel vector based approach. In *IEEE International Conference on Cloud Computing*, pages 275–282, 2011.
- [17] Suraj Pandey, Linlin Wu, Siddeswara Mayura Guru, and Rajkumar Buyya. A particle swarm optimization-based heuristic for scheduling workflow applications in cloud computing environments. In *24th IEEE International Conference on Advanced Information Networking and Applications (AINA)*, pages 400–407. IEEE, 2010.
- [18] Shekhar Srikantaiah, Aman Kansal, and Feng Zhao. Energy aware consolidation for cloud computing. *Cluster Computing*, 12:1–15, 2009.
- [19] Johan Tordsson, Rubén S. Montero, Rafael Moreno-Vozmediano, and Ignacio M. Llorente. Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers. *Future Generation Computer Systems*, 28(2):358–367, 2012.
- [20] Akshat Verma, Puneet Ahuja, and Anindya Neogi. pMapper: power and migration cost aware application placement in virtualized systems. In *Middleware 2008*, pages 243–264, 2008.
- [21] Qi Zhang, Lu Cheng, and Raouf Boutaba. Cloud computing: state-of-the-art and research challenges. *Journal of Internet Services and Applications*, 1(1):7–18, 2010.

Mathematical modeling of Incheon Bridge, Structural monitoring

Gheorghe M.T. Radulescu, Corina M. Radulescu and Adrian T. Radulescu

Abstract— Mathematical modeling of structural behavior is extremely important to validate design solutions and check if the construction being analyzed continues to show safety in operating. All data analyzed come from Structural Health Monitoring, an extremely complex and relatively expensive activity, and the current offer of tools, methods and technologies is varied, which can lead to a virtually high number of structural monitoring systems that can be customized for each case. In time, the monitoring of bridges became the engine for the development of SHM tools, methods and technologies, or manager monitoring systems. The case study, in continuous quasi-static condition, was performed on Incheon Grand Bridge, Seoul, South Korea. Tracking the behavior of an objective under the influence of sunshine is performed by VCE Vienna Consulting Engineers ZT GmbH. This paper presents the context in which mathematical modeling fits into the set of activities on checking behavioral hypotheses defined in the design process. This paper presents and the mathematical models of the effect of sunshine on a steel structural element, the 24 lamella front South line, by comparing data pairs that reflect the cause: atmospheric temperature and the effect: the movement of a sensor mounted on the structural element. The analysis was performed using software to achieve a more optimal mathematical model, that has been tested and then validated.

Keywords— Incheon Grand Bridge, Mathematical model, steel structural element, Structural Health Monitoring,

I. INTRODUCTION

A. General considerations of the Structural Health Monitoring

Structural Health Monitoring (SHM) is a non-destructive in-situ structural sensing and evaluation method that uses a variety of sensors attached to, or embedded in, a structure to monitor the structural response, analyze the structural characteristics for the purpose of estimating the severity of damage/deterioration and evaluating the consequences thereof on the structure in terms of response, capacity, and service-life[1, 2, 3, 4]. According to Chang [5], "the goal of structural monitoring is to gain knowledge of the integrity of in-service

F. A. Author is with the Technical University of Cluj Napoca, 400114 Romania, Head of Department of Terrestrial measurements and Cadastre, (corresponding author to provide phone: 0040721942189; fax: 0040262276153; e-mail: gmtradulescu@yahoo.com).

S. B. Author is with the Technical University of Cluj Napoca, 430122 Romania. Head of Department of Economics, (e-mail: corinam.radulescu@gmail.com).

T. C. Author is with the Technical University of Cluj Napoca, 400114 Romania, Department of Terrestrial measurements and Cadastre (e-mail: adrian_r1982@yahoo.com).

structures on a continuous real-time basis". Some of the benefits/advantages of a properly designed SHM are [6]: Real time monitoring with alarms increase the safety for the end-users; Down time reduction; To verify, control, assess, understand the actual behavior of the structure; Calibration of FEM and calculations; Decreased maintenance costs; In general, the activity of SHM during execution differs from its period of service, but some sensors may remain, thus making the overall process less expensive. Structural Monitoring [7] was done using wired systems that collected and monitored data from these structures. This was an expensive and inflexible approach because the system could not be easily redeployed if better data collection points were discovered on the structure. Wireless Sensor Networks became a good way to solve this problem, and thereby meet a major requirement for a viable SM system. Autonomous motes could now be deployed over a field of interest while data was collected at a base station [8]. Real-time data monitoring involves continuous data capture with a very small time margin between data sample blocks[9]. Monitoring of bridges is an adaptation of SHM for this important work of art. Data on structural behavior are collected in the SHM process the and databases are created, which can then be processed by different software, among which are: Excel real-statistics, DataFit 9.1., Statistics Dell software, Table Curve 2D, Table Curve 3D, SimFit and IBM SPSS 21, which helps create different mathematical models of the processes studied.

B. Presentation of the studied structure, Incheon Bridge

The Incheon Bridge (figure 1) is South Korea's longest spanning cable-stayed bridge. At 12.3km long with a main cable stayed span of 800m the new Incheon Bridge will be one of the five longest of its type in the world. Its 33.4m wide steel/concrete composite deck will carry six lanes of traffic 74m above the main shipping route in and out of Incheon port and link the new Incheon International Airport on Yongjŏng Island to the international business district of New Songdo City and the metropolitan districts of South Korea's capital, Seoul. The cable stayed section of the crossing is 1,480m long, made up of five spans measuring 80m, 260m, 800m, 260m and 80m respectively, and the height of the "inverted Y" main towers is 230.5m. The sea crossing bridge section, whose concessionaire is Incheon Bridge Corporation, is funded by the private sector. Korea Expressway Corporation and the Korean Ministry of Land, Transport and Maritime Affairs (MLTM) managed the project [10].



Figure 1. General view 1, Incheon grand bridge South Korea (Source: VCE)

C. Short presentation of the Structural Monitoring Method

From the presentation site we quote: " In order to measure the movement of the cable stayed bridge section and the performance of the modular expansion joints of type LR24, a ROBO@CONTROL remote monitoring system was installed at one at the expansion joint locations. This serves to measure the movements of the first, second and last lamella beams of the joint, as well as the entire gap width and air and structure temperatures. "[10]. Figure 2 shows the ultrasonic sensors, and figure 3 shows the position of the RoboControl box, mounted on the structure. The outstanding feature of the UPK series is its high acoustic power combined with small sensor size. Monitoring system consists of 6 UPK category sensors, i.e. 4 UPK 500 sensors(first lamella, second lamella and 24- last, lamella –South, first lamella -North carriagetaway) and 2 UPK2500 sensors (bridge gap-South and North carriagetaway)[10].



Figure 2. Ultrasonic sensors sonorange UPK (Surse SNT Sensortechnik AG)



Figure 3. The position of the RoboControl box, mounted on the structure(Source: VCE)

2. OVERVIEW OF THE DATA OBTAINED IN THE MONITORING PROCESS, THE OPPORTUNITY FOR THEIR SELECTION

The climate in South Korea is temperate with a lot of rainfall in summer and winters that can get very cold. In Seoul, Incheon bridge area, the average January temperature ranges from -7°C to 1°C, slightly lower in February, and the average

July temperature ranges between 22°C and 29°C, slightly higher in August. The company that performs the monitoring, VCE Vienna Consulting Engineers ZT GmbH, with which our institution has been working with since 2009, provided us with all the data from June 1, 2009, the date of commencement of the action until today, the date of completion of this paper, December 23, 2014, having an initial data volume of 194,881 for nine parameters (similar to those in Table 1.), so a total of 1,753,929 possible pairs of correlations[10], which initially led to the selection of data for the year 2013 and after that the selection of four significant months in 2013. Table 1. shows the extreme values of air temperatures for the period in which the monitoring was made (which continues today) and Table 2 shows the total number of data pairs for different monitoring periods and different intervals between measured data.

	Min.	Max.	Dif.
Momment of Recording; Mounth/ Day/Year; every 15 minutes	7/2/2013 21:28	8/6/2012 5:58	121 Days
Air temperature [°C]	-14.2	36.1	50,3
Movement of 24th lamella from south line [m]	0,256	0,211	0,045
Movement of bridge gap from north line [m]	1,676	0,861	0,815
Movement of bridge gap from south line [m]	1,696	0,853	0,843
Movement of first lamella from north line [m]	0,156	0,101	0,055
Movement of first lamella from south line [m]	0,139	0,083	0,056
Movement of second lamella from south line [m]	0,285	0,168	0,117
Temperature in the ROBO-CONTROL box [°C]	5,9	34,5	28,4
Temperature of steel structure [°C]	-15,0	34,7	49,7

Table 1. The extreme values for recordings made 02.02.2009-23.12.2014, every 15 minutes (Source: Authors)

No .	Period	Interval between measurements	No. of processed data pairs
1.	2009-2014	15 Min.	194.881
2.		1 H	48.720
3.		2 H	24.360
4.		4 H	12.180
5.	2013	15 Min.	35.024
6.		1 H	8756
7.		2 H	4378
8.		4 H	2189
9.	2013, 4 Mounth II, V, VIII, XI	15 Min.	11.616
10.		1 H	2.904
11.		2 H	1.452
12.		4 H	726, Case study

Table 2. The amounts of possible combinations for different monitoring periods at various intervals (Source: Authors)

3. DEFINING THE POSITION AND ROLE OF STRUCTURAL MATHEMATICAL MODELING IN THE HISTORY OF ACHIEVING A CONSTRUCTION

Mathematical modeling has a very important role in achieving a structural objective[1]. In Figure 4 the authors define the position of this stage in the life of a structure.

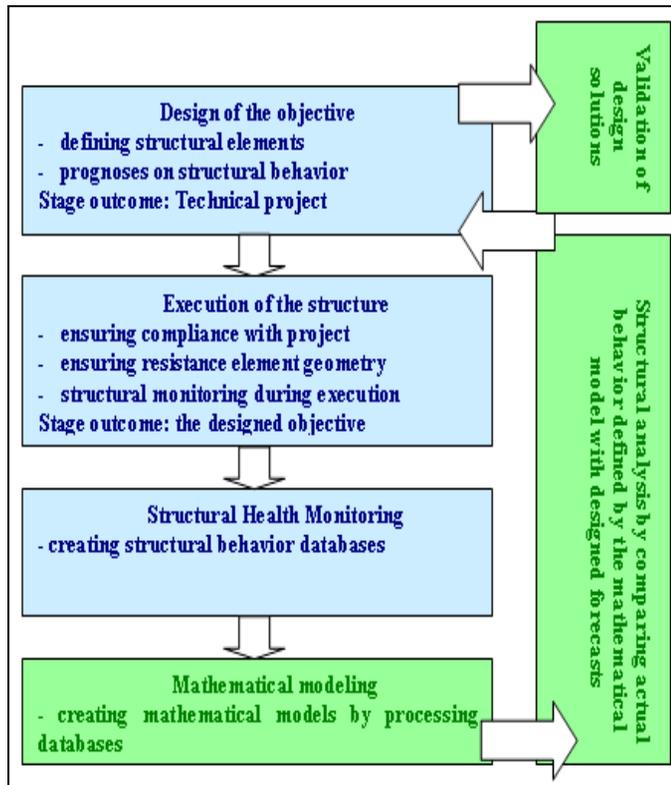


Figure 4. The position of the Mathematical modeling in the Structural live (Source: Authors)

Mathematical modeling of structural behavior complements structural monitoring, a phase which begins during building construction and continues, for special constructions (this includes the bridge which is the subject of this case study), throughout their lifetime[1, 11, 12]. While Structural Health Monitoring aims to create behavioral data banks of the structure for the action of various factors, mathematical modeling, through processing and using various software dedicated to data coming from the previous stage, defines mathematical models of the cause and effect ratio for different strains and resistance elements making up the analyzed objective[13].

4. MAIN STRAINS TO WHICH THE RESISTANCE STRUCTURES OF BRIDGES ARE EXPOSED TO

Bridges are subjected, during the period of service, to the following main strains:

- wind,
- uneven sunshine of structural elements
- variation in the level of the watercourse,
- traffic and its characteristics, number, weight of vehicles, speed, collisions and other events,
- earthquake,
- settlement or compaction of foundation land,
- different physico-chemical factors, one of the most important being corrosion of metallic resistance elements.

The paper analyzes the second aspect, which is considered very important for the monitored objective.

5. MATHEMATICAL MODELS OF THE CORRELATION BETWEEN TEMPERATURE AND SENSOR MOVEMENT.

A. Study of the relevance of mathematical models developed for data taken at intervals of 1, 2 and 4 hours

The first verification was done to see if by reducing the percentage of data, basically increasing the monitoring interval, there are significant changes of the mathematical model created by different mentioned software[14, 15, 16, 17]. We chose the correlation between the temperature of the steel (the material of the bridge deck blades) and the movement of the sensor mounted on them, i.e. lamella 24 from the south line of the bridge. I chose three time intervals, respectively data recorded hourly, every two hours, respectively every four hours.

1. Hourly data

Entering the data in the program we obtain 83 equations which define the relationship between the two sets of data, x , representing the temperature of the steel and $y = f(x)$ the position of the sensor, relative to a predetermined fixed origin. In general, as a mathematical model we chose the first equation, Rank 1 or the closest one to an operable model. In this case, the first equation is:

$$y^{0.5} = a + bx + cx^{20} + dx^3 + ex^4 \quad (1.)$$

Correlation coefficient $r^2 = 0.8821704827$ showing a highly significant correlation between the input data into the program.

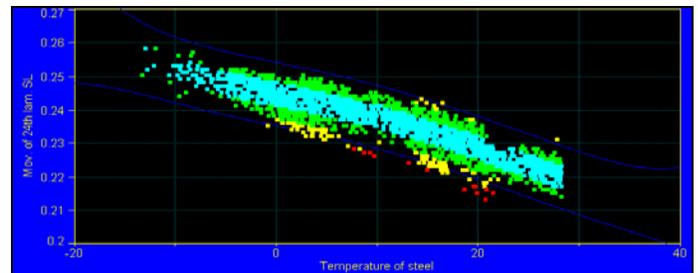


Figure 5. Correlation graph between steel temperature and sensor position on the 24 lamella south line and the data covariance interval (Source: Authors)

For the defined equation the coefficients are shown below, and for the other they have similar values: $a=0.494485022$; $b= -0.00063215$; $c= -3.7835e^{-0.7}$; $d= -7.8368e^{-0.7}$; $e= 1.57508e^{-0.8}$

Figure 5. Shows the graph of the correlation between the temperature of the steel and the position of the sensor on lamella 24 from the south line and the data covariance interval. Figure 6 shows the graph of residual values on the correlation between steel temperature and sensor position on the 24 lamella south line. The vast majority of data falls in the range of -0.005 to $+0.005$ m, reaching a lower percentage in the range of -0.010 to $+0.010$ m.

2. Data taken every two hours

Entering the data in the program we obtain 83 equations which define the relationship between the two sets of data, x , representing the temperature of the steel and $y = f(x)$ the position of the sensor, relative to a predetermined fixed origin.

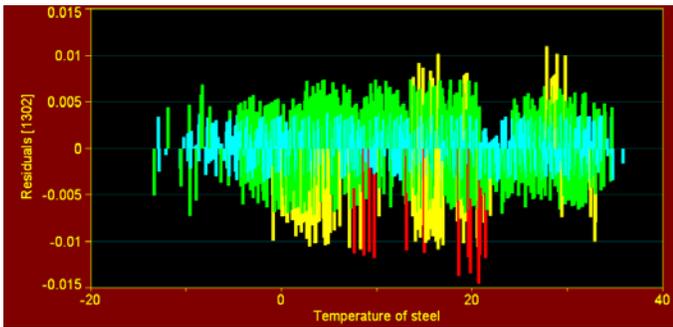


Figure 6. Graph of residual values on the correlation between steel temperature and sensor position (Source: Authors)

In general, as a mathematical model we chose the first equation, Rank 1 or the closest one to an operable model. In this case, the first equation is:

$$y^{0.5} = a + bx + cx^{20} + dx^3 + ex^4 \dots \quad (2)$$

Although the correlation coefficient

$r^2 = 0.8819868406$ decreases slightly, it continues to show a highly significant correlation between the input data into the program. $a = 0.494465456$; $b = -0.00060642$; $c = -9.0723 e^{-0.7}$; $d = -9.4767 e^{-0.7}$; $e = 2.05933 e^{-0.8}$

3. Data taken every four hours

Entering the data in the program we obtain 83 equations which define the relationship between the two sets of data, x , representing the temperature of the steel and $y = f(x)$ the position of the sensor, relative to a predetermined fixed origin. In general, as a mathematical model we chose the first equation, Rank 1 or the closest one to an operable model. In this case, the first equation is:

$$\ln y = a + bx + c \cdot x^{20} + d \cdot x^3 + e \cdot x^4 \quad (3)$$

Although the correlation coefficient $r^2 = 0.9046985517$, surprisingly, increases slightly, it continues to show a highly significant correlation between the input data into the program. For the equation (3) coefficients are shown below, for the rest they have similar values: $a = -1.40691138$; $b = -0.00242608$; $c = -7.2963e^{-0.6}$; $d = -3.6588e^{-0.6}$; $e = 7.50831e^{0.8}$. The following four equations do not change, but have a slight decrease in the correlation coefficient:

$$2. y^{0.5} = a + bx + cx^{20} + dx^3 + ex^4 \quad (4)$$

Correlation coefficient $r^2 = 0.9046910656$

$$3. y = a + bx + cx^2 + dx^3 + ex^4 \quad (5)$$

Correlation coefficient $r^2 = 0.9046597445$

$$4. 1/y = a + bx + cx^2 + dx^3 + ex^4 \quad (6)$$

Correlation coefficient $r^2 = 0.9046423605$

$$5. y^2 = a + bx + cx^{20} + dx^3 + ex^4 \quad (7)$$

Correlation coefficient $r^2 = 0.9045253305$

Figure 7 shows the correlation graph between the temperature of the steel and the position of the sensor on lamella 24 from the south line, recordings made every four hours, and the five equations-mathematical models shown, it becomes evident that the shape of the graphs shown in Figures 5 and 6 and the relationships and indices in the accuracy calculus remain similar; the reduction of the data to a quarter did not

significantly affect the results. Replacing the variable x =steel temperature in the Rank 1 equation leads to values very close to the average ones found in the field, so the correlation is maintained, and it even increases for the density selection every four hours. (Table 3)

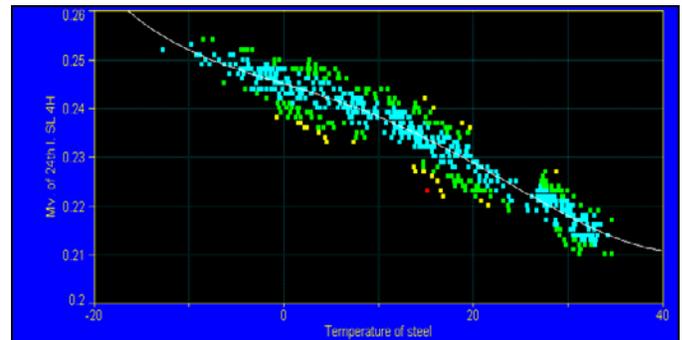


Figure 7. Correlation graph between steel temperature and sensor position, recording every four hours (Source: Authors)

T °C	Rank 1 eq.	Position from terrain measurements			Approx.
		Max.	Min.	Med.	
1	0.244	0,247	0,239	0,243	0,995188
5	0.241	0,248	0,237	0,242	0,995877
10	0.237	0,244	0,235	0,239	0,993319
15	0.237	0,239	0,224	0,231	0,973098
20	0.227	0,226	0,222	0,240	0,988125
25	0.222	0,226	0,216	0,221	0,992810
30	0.216	0,220	0,211	0,215	0,994004

Table 3. Validation of the mathematical model created-The calculation of the approximation accuracy of the mathematical model created on the correlation between steel temperature and sensor position on the 24 lamella south line, for significant temperatures (Source: Authors)

Iteration	Accuracy %	The number of pairs of data removed
Basic data	88,0	22
1	89,2	14
2	89,8	13
3	90,2	8
4	90,5	7
5	90,8	8
6	91,3	5
7	91,4	3
8	91,4	2
9	91,4	1
10	91,5	0

Table 4. Acuity of the mathematical model for each iteration and the number of data pairs removed (Source: Authors)

This comparative analysis shows that for this case analyzing the correlation between the response of the displacement sensor under the effect of temperature change it is sufficient to

study data taken every four hours, which are similar to those taken every two hours or hourly. Further, the analysis of the data and the construction of the other mathematical models will be made at this data capture density, every four hours.

B. Data processing using IBM SPSS Statistics software

For the case studied it was important to improve the mathematical model by removing insignificant quantities, through successive iterations. We conducted ten iterations, eliminating the data that the software indicated as irrelevant, and the acuity gradually increased (Table 2.) from the initial 88% to the final value, after ten iterations, of 91.5% . After the tenth iteration the software indicates that no more quantities have to be removed from the processed data, because this percentage of acuity of the mathematical model created is the maximum possible.

It is noted that removing massive amounts of values in the basic data processing stage, and after the first iteration, the second, or fourth, brought the greatest progress in increasing acuity from 1.2% to 0,5 %.

C. Using statistical selection data obtained through IBM SPSS 21 software with the others mathematical software

The author believes that there is no need to resume using the diferents software for obtaining analytical and graphical data. The following are the equations of Rank 1 and the related correlation coefficient for all ten iterations above. For each iteration I used the data set selected after the removal of data pairs specified in the previous iteration.

Raw data.

Correlation coefficient $r^2 = 0.9046985517$,

$$\ln y = a + bx + c \cdot x^{20} + d \cdot x^3 + e \cdot x^4 \tag{8}$$

Iteration 1. Correlation coefficient, $r^2 = 0.8992430098$

$$\ln y = a + bx + c \cdot x^{20} + d \cdot x^3 + e \cdot x^4 \tag{9}$$

Iteration 2. Correlation coefficient,

$$\ln y = a + bx + c \cdot x^{20} + d \cdot x^3 + e \cdot x^4 \tag{10}$$

Iteration 3. Correlation coefficient, $r^2 = 0.9077518531$

$$\ln y = a + bx + c \cdot x^{20} + d \cdot x^3 + e \cdot x^4 \tag{11}$$

Iteration 4. Correlation coefficient, $r^2 = 0.9116396303$,

$$y = a + bx + c \cdot x^{20} + d \cdot x^3 + e \cdot x^{-x} \tag{12}$$

Iteration 5. Correlation coefficient, $r^2 = 0.9138251649$,

$$y = a + bx + c \cdot x^{20} + d \cdot x^3 + e \cdot x^{-x} \tag{13}$$

Iteration 6. Correlation coefficient, $r^2 = 0.9188109224$,

$$y = a + bx + c \cdot x^{20} + d \cdot x^3 + e \cdot x^{-x} \tag{14}$$

Iteration 7. Correlation coefficient, $r^2 = 0.9200105984$,

$$y = a + bx + c \cdot x^{20} + d \cdot x^3 + e \cdot x^{-x} \tag{15}$$

Iteration 8. Correlation coefficient, $r^2 = 0.9198227878$,

$$y = a + bx + c \cdot x^{20} + d \cdot x^3 + e \cdot x^{-x} \tag{16}$$

Iteration 9. Correlation coefficient, $r^2 = 0.9193773591$,

$$y = a + bx + c \cdot x^{20} + d \cdot x^3 + e \cdot x^{-x} \tag{17}$$

Iteration 10. Correlation coefficient, $r^2 = 0.9200420354$

$$y = a + bx + c \cdot x^{20} + d \cdot x^3 + e \cdot x^{-x} \tag{18}$$

Iter.	Coefficients of equations, mathematical models, Rank 1		
Iter.	<i>a</i>	<i>b</i>	<i>c</i>
Raw data	-1.40691138	-0.00242608	-7.2963e ^{-0.6}
1.	-1.40679370	-0.00261402	4.48457 e ^{-0.6}
2.	-1.40556455	-0.00260040	-1.0538 e ^{-0.5}
3.	-1.40465510	-0.00263100	-1.8611 e ^{-0.5}
4.	0.245724404	-0.00063384	-1.5918 e ^{-0.5}
5.	0.245755447	-0.00063776	-1.5727 e ^{-0.5}
6.	0.245877416	-0.00063697	-1.63 e ^{-0.5}
7.	0.245906597	-0.00064837	-1.5721 e ^{-0.5}
8.	0.245935689	-0.00064455	-1.6374 e ^{-0.5}
9.	0.245955161	-0.00064746	-1.6391 e ^{-0.5}
10.	0.245983339	-0.00064442	-1.683 e ^{-0.5}
	<i>d</i>	<i>e</i>	
Raw data	-3.6588e ^{-0.6}	7.50831e ^{-0.8}	
11.	-3.9626 e ^{-0.6}	8.0838 e ^{-0.8}	
12.	-3.2469 e ^{-0.6}	7.24115 e ^{-0.8}	
13.	-2.7267 e ^{-0.6}	6.49408 e ^{-0.8}	
14.	2.25411 e ^{-0.7}	8.63967 e ^{-0.8}	
15.	2.2235 e ^{-0.7}	8.39932 e ^{-0.8}	
16.	2.36081 e ^{-0.7}	8.29797 e ^{-0.8}	
17.	2.28472 e ^{-0.7}	7.79865 e ^{-0.8}	
18.	2.46217 e ^{-0.7}	7.69347 e ^{-0.8}	
19.	2.50703 e ^{-0.7}	7.60461 e ^{-0.8}	
20.	2.60901 e ^{-0.7}	7.7476 e ^{-0.8}	

Table 5. Coefficients of equations, mathematical models - Rank 1, first 10 iterations.(Source: Authors)

For iteration 10, a series of simple equations were found, for example: Equation 28, Correlation coefficient,

$$r^2 = 0.9168929798, y^2 = a + bx, \tag{19}$$

$$a = 0.060575825; b = -0.00042231$$

Equation 32, Correlation coefficient,

$$r^2 = 0.9147585193, y = a + bx \tag{20}$$

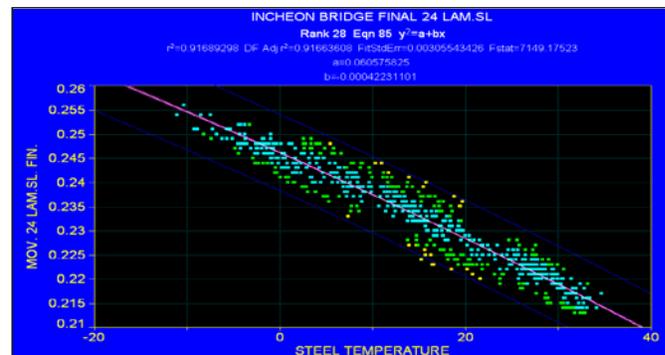


Figure 8. Correlation graph between steel temperature and sensor position on the 24 lamella south line and the data covariance interval (Source: Authors)

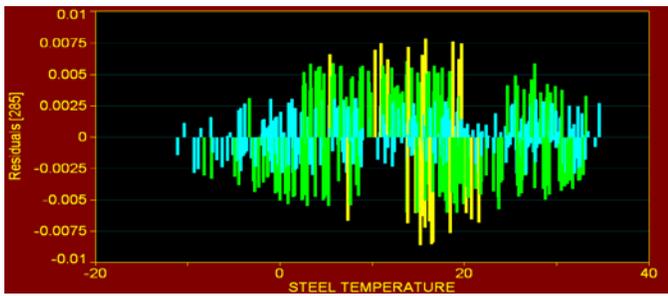


Figure 9. Graph of residual values on the correlation between steel temperature and sensor position on the 24 lamella south line (Source: Authors)

T °C	Rank 1 Eq.	Position from terrain measurements			Approx.
		Max.	Min.	Med.	
-12	0.255	0,256	0,254	0,2550	0,999647
-10	0.252	0,256	0,253	0,2545	0,993556
-5	0.248	0,251	0,249	0,2500	0,995892
0	0.245	0,250	0,238	0,2440	0,993558
5	0.242	0,248	0,237	0,2425	0,997938
10	0.237	0,244	0,235	0,2395	0,992902
15	0.233	0,239	0,224	0,2315	0,984269
20	0.227	0,226	0,222	0,2240	0,983142
25	0.222	0,226	0,216	0,2210	0,991129
30	0.218	0,220	0,211	0,2155	0,987626
35	0.213	0,217	0,210	0,2135	0,998234
40	0.209	-	-	-	-

Table 6. Validation of the mathematical model created - This presents the calculation of the approximation accuracy of the mathematical model created on the correlation between steel temperature and sensor position on the 24 lamella south line, for significant temperatures model created brings. (Source: Authors)

D. Improving the mathematical model by eliminating irrelevant data

The authors continued the selection and deletion of information by analyzing the graphs obtained by sequentially placing data pairs. After five iterations, we reached a correlation coefficient of more than 0.95, which is considered relevant for the response of the structural element to strains, in the present case the way sunshine falls on it. The correlation coefficients and the equations are presented in the relations 21-25, specifying the iteration considered. The coefficients of the equations are in Table 7, and the correlation graphs and residual values graphs are shown in Figures 10 and 11. We also removed a number of 20, 20, 17, 13 and 12 pairs of data after iterations 10-14.

$$11. r^2 = 0.9312735355, 1/y=a+bx+cx^2+dx^3+e x^4 \quad (21)$$

$$12. r^2 = 0.9414311349 1/y=a+bx+cx^2+dx^3+e x^4 \quad (22)$$

$$13. r^2 = 0.9475693596 1/y=a+bx+cx^2+dx^3+e x^4 \quad (23)$$

$$14. r^2 = 0.9530277469 1/y=a+bx+cx^2+dx^3+e x^4 \quad (24)$$

$$15. r^2 = 0.9568250939 y=a+bx+c/x+d x^2+e/x^2+f x^3+g/ x^3+h x^4+i/ x^4 \quad (25)$$

It.	Coefficients of equations, mathematical models, Rank 1		
	a	b	c
11.	4.067898629	0.010819210	0.000108598
12.	4.067432360	0.010399102	8.53709 e ^{-0.5}
13.	4.066422587	0.010104994	7.43603 e ^{-0.5}
14.	4.064486894	0.009927696	.58866 e ^{-0.5}
15.	0.246306624	-0.00062963	0.000105680
15.	f=-3.859 e ^{-0.7}	g=-8.593 e ^{-0.6}	h=9.90611 e ^{-0.9}

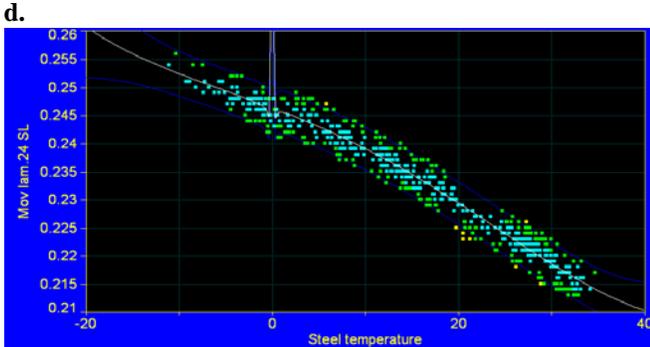
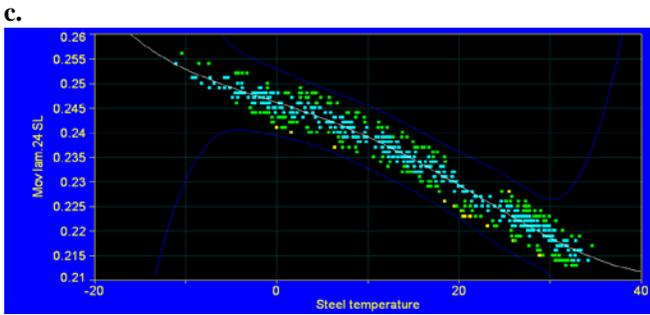
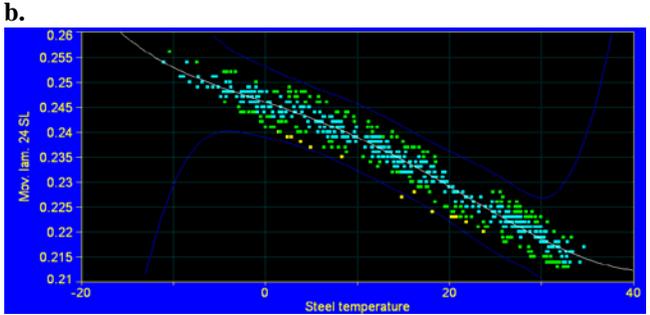
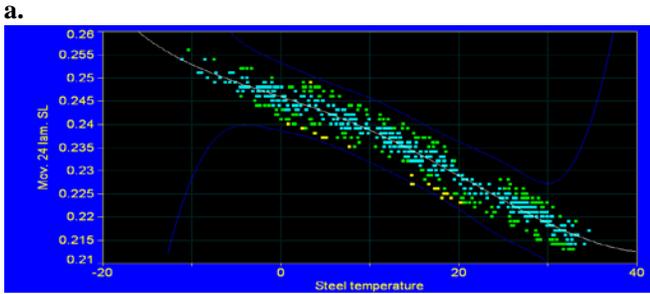
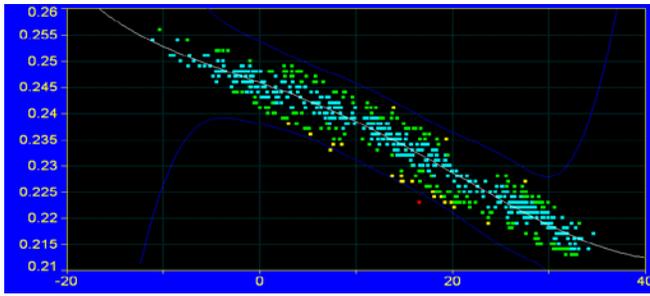
It.	Coefficients of equations, mathematical models, Rank 1	
	d	e
11.	1.09237 e ^{-0.5}	-2.6001 e ^{-0.7}
12.	1.35539 e ^{-0.5}	-3.0487 e ^{-0.7}
13.	1.49076 e ^{-0.5}	-3.2542 e ^{-0.7}
14.	1.42313 e ^{-0.5}	-3.0722 e ^{-0.7}
15.	-7.2178 e ^{-0.6}	-0.00023130
15.	i=4.36782 e ^{-0.6}	-

Table 7. Coefficients of equations, mathematical models - Rank 1, the last 5 iterations.(Source: Authors)

T °C	Rank 1 eq.	Position from terrain measurements			Approx.
		Max.	Min.	Med.	
-12	0.253	0,256	0,254	0,2550	0,992157
-10	0.252	0,256	0,253	0,2545	0,991018
-5	0.249	0,251	0,249	0,2500	0,996000
0	0.246	0,250	0,238	0,2440	0,991870
5	0.242	0,248	0,237	0,2425	0,997938
10	0.238	0,244	0,235	0,2395	0,993737
15	0.234	0,239	0,224	0,2315	0,980316
20	0.229	0,226	0,222	0,2240	0,978166
25	0.223	0,226	0,216	0,2210	0,991031
30	0.218	0,220	0,211	0,2155	0,987626
35	0.213	0,217	0,210	0,2135	0,998234
40	0.210	-	-	-	-

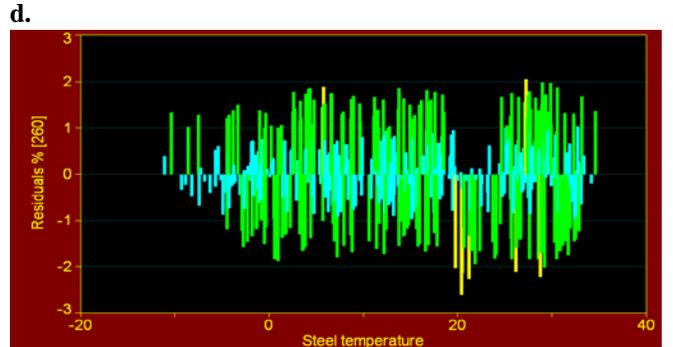
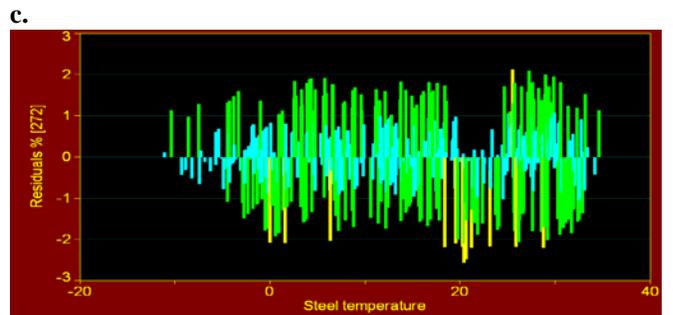
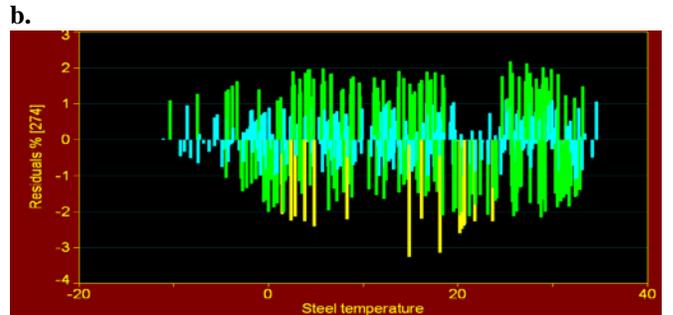
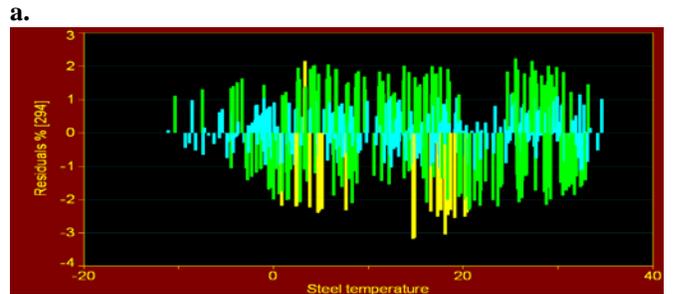
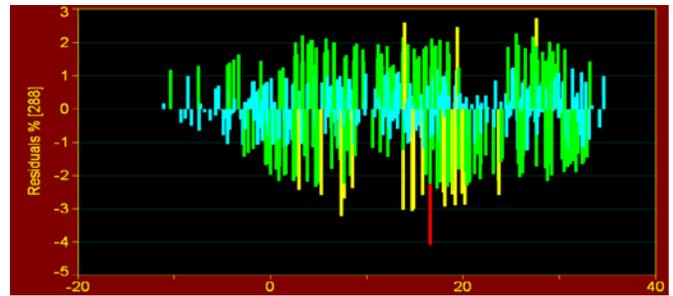
Table 8. Validation of the mathematical model created after 15 iterations.. (Source: Authors)

In Table 8 we validated the mathematical model created. We note that for temperatures between 20-30°C, we have the lowest degree of approximation, which is explained by the slower response of the material of which the structural element analyzed is made of, namely the steel. We can also see that, although the correlation coefficient increased significantly between iterations 10 and 15, not the same can be said about the degree of approximation which decreased for all positions.



a.
b.
c.
d.
e.

Figure 10. Correlation graph between steel temperature and sensor position on the 24 lamella south line and the data covariance interval, iterations a. 11, b. 12, c. 13, d.14, e.15 (Source: Authors)



a.
b.
c.
d.
e.

Figure 11. Graph of residual values on the correlation between steel temperature and sensor position on the 24 lamella south line, iterations a. 11, b. 12, c. 13, d.14, e.15 (Source: Authors)

CONCLUSIONS

The mathematical models of various correlation phenomena which have a correlation coefficient below 0.9 are considered irrelevant. The authors analyzed the possibility, by proper data selection from structural monitoring activity, to improve the mathematical model of the response of the structural element analyzed, from a correlation coefficient of 0.88 to a correlation coefficient greater than 0.95, after one cycle of 15 successive stages of eliminating data considered irrelevant. The model was validated by comparing the values given by the model with those found during operation. Mathematical modeling of phenomena recorded in the structural monitoring activity became extremely useful both for those who make the observations, Structural Health Monitoring specialists, who thus can validate their recorded data, and for the designer of the monitored objective, who can thus validate his chosen design solutions and behavioral patterns predicted in the design phase, but also for the beneficiary who has thus the certainty that the objective works within the design parameters and any accident due to current actions is excluded. This paper is part of the concerns of the collective at the Land Measurements and Cadastre Department of the Technical University of Cluj Napoca, Romania, to implement different software to perform mathematical modeling of structural phenomena studied by Structural Health Monitoring activities. Starting from a model with a small correlation coefficient we succeeded, by a rigorous selection of data, to build a credible model, subsequently validated by comparing the data obtained in-situ.

REFERENCES

- [1] G.M.T. Radulescu (2015) Theoretical and Experimental Research on Structural Geometric Monitoring of Constructions through Surveying and Alternative Methods, *Habilitation Thesis*, Technical University of Cluj Napoca, Romania.
- [2] G. Housner et al.(1997) Developments in the USA in the Field of Structural Control and Monitoring of Civil Infrastructure systems, <http://www.iitk.ac.in/nicee/wcee/article/2217.pdf>
- [3] A. Mufti et al. (2009) SHM Data Interpretation and Structural Condition Assessment of the Manitoba Golden Boy, *Smart Structures and Systems*, 6(1), pp.87-90.
- [4] Y. Dong (2010) Bridges Structural Health Monitoring and Deterioration Detection, *Synthesis of Knowledge and Technology, Alaska University Transportation Center, Fairbanks, AK 99775-5900*, Final Report
- [5] F.K. Chang et al. (1999). Structural Health Monitoring, *Proceedings of the 2nd International Workshop on Structural Health Monitoring*, Stanford, CA, USA.
- [6] E. Merit (2011) Lessons Learned in Structural Health Monitoring of Bridges Using Advanced Sensor Technology, *TRITA-BKN. Bulletin 108*, <http://www.diva-portal.org/smash/get/diva2:456855/pdf>
- [7] A.T.G. Radulescu, G.M.T. Radulescu (2014) Considerations on the Behavior of Incheon Bridge, Seoul, South Korea to the Action of an Uniform Sunshine, Comparison Study to the Extreme Temperature's Exposure Periods, 70402-559, *ISI Proceedings of World Scientific and Engineering Academy and Society Conferences, Recent scientes in applied economics and management, Economic Aspects of Environment, Development, Tourism and Cultural Heritage - Volume 2, Proceedings of the 5th International Conference on Applied Economics, Business and Development (AEBD '13), Chania, Crete Island, Greece*, pp.270-276
- [8] O.Ajiboye(2010) Sensor Computation and Communication for Remote Structural Monitoring, http://etd.library.vanderbilt.edu/available/etd-05272009-111310/unrestricted/Olabode_Ajiboye_Revised.pdf

- [9] C. Anumba (2010) Current trends in the engineering management of subsidence cases, *Structural Survey* ISSN: 0263-080X Vol. 15 Iss: 1, pp.5 – 10,
- [10] VCE-Viena Consulting Engineers (2009-2015) Permanent Structural Health Monitoring Systems, *available at: http://www.brimos.com/DMA/ DMA Frames.*
- [11] A.T.G.Radulescu (2013) Structural Monitoring Today, Modern Surveying Technologies used to Track Behavior over Time of Buildings, *Lap Lambert Academic Publishing.*
- [12] O. Stefan, G. Badescu, M. Gh. Radulescu (2009) Comparative Study Regarding the Accuracy of RTK type GPS in a Classic Geodesic Network, *EUREF ,Symposium Florence,, Italy, www.igmi.org/euref2009/participants.php*
- [13] V. Gikas, M. Sakellariou (2008) Settlement Analysis of the Mornos Earth Dam Evidence from Numerical Modeling and Geodetic Monitoring, *Engineering Structures*. Vol. 30(11),pp.3074–3081
- [14] <http://www.sigmaplot.com/products/tablecurve2d/tablecurve2d.php>
- [15] <http://www.sigmaplot.com/products/tablecurve3d/tablecurve3d.php>
- [16] <http://www-01.ibm.com/software/ro/analytics/spss/>
- [17] <http://www.oakdaleenr.com/webstore.htm>



Professor Gheorghe M.T. Rădulescu was born in Bistrița, Bistrița Năsăud County, Romania in 1950, April, 17. He received the M.S. degree in Geodetic Engineering from the University of Civil Engineering Bukarest, Romania in 1974; M.S. degree in Mathematics from the University Babeș-Bolyai Cluj Napoca, Cluj County, Romania in 1990, the Ph.D. degree in Geodetic Sciences from the University of Civil Engineering Bukarest, Romania in 2003. He is author in 83 papers in national and international journals, 27 books, 57 papers in international conference proceedings. He is Professor and Head of Terrestrial Measurements and Cadastre Department within Technical University of Cluj Napoca, Romania. His research interests includes a) Surveying; b) Engineering Surveying; c) Structural Health Monitoring; d) Mathematical modeling in Structural Health Monitoring; e) GIS.



Corina was born in Cluj Napoca, Cluj County, Romania in 1960, October, 29. She received the M.S. degree in Civil Engineering from the Technical University of Cluj Napoca, Romania in 1985; the Ph.D. degree in Industrial Management from the Technical University of Cluj Napoca, Romania in 2003. She is author in 38 papers in national and international journals, 18 books, 34 papers in international conference proceedings. She is Associate Professor and Head of Economic Department of Technical University within Cluj Napoca, Romania. Her research interests includes a) Strategic Management; b) Industrial Cluster research; c) SME's; d) Mathematical modeling in Economics; e) GIS in Economics and Public area.



Adrian was born in Cluj Napoca, Cluj County, Romania in 1982, July, 17.. He received the M.S. degree in Civil Engineering from the Technical University of Cluj Napoca, Romania in 2006; M.S. degree in Geodetic Engineering from the Technical University of Timișoara, Romania in 2009; the Ph.D. degree in Mining Surveying from the University of Petroșani Romania in 2011. He is author in 21 papers in national and international journals, 6 books, 23 papers in international conference proceedings. He is Assistant Professor at Terrestrial Measurements and Cadastre Department within Technical University of Cluj Napoca, Romania. His research interests includes a) Surveying; b) Engineering Surveying; c) Structural Health Monitoring; d) Mathematical modeling in Structural Health Monitoring.

Systems Optimization Prospected from Torus Cyclic Groups

Volodymyr V. Riznyk

Abstract—This paper relates to techniques for improving the quality indices of engineering devices or systems with respect to performance reliability, transmission speed, positioning precision, and resolving ability, using novel design based on structural perfection and remarkable properties of proposed modification of combinatorial configurations, prospected from fundamental laws of the space dimensionality, namely the concept of Ideal Vector Rings (IVR)s. These design techniques make it possible to configure systems with fewer elements than at present, while maintaining or improving on resolving ability and the other significant operating characteristics of the system.

Keywords—Combinatorial configuration, systems optimization, perfect torus group, monolithic code, vectorial space harmony laws.

I. INTRODUCTION

Combinatorial configurations arise in many problems of pure mathematics, notably in algebra, applied physics, topology, and geometry. Combinatorics also has many applications in mathematical optimization, computer science, and quantum physics. One of the most acceptable parts of combinatorics is systems theory, which also has numerous natural connections to other areas. Combinatorial optimization started as a part of combinatorics and graph theory, but is now viewed as a branch of applied mathematics and computer science, related to coding theory as a part of design theory with combinatorial constructions of error-correcting codes. The main idea of the subject is to design efficient and reliable methods of data transmission. It is a large field of study, part of information theory in systems engineering and data communications. Combinatorial configurations such as cyclic difference sets [1] and Ring Bundles [2], is known, to be of very important in systems engineering for improving the quality indices of devices or systems with non-uniform structure (e.g. arrays of radar systems) with respect to resolving ability [3].

This work is connected in part with Cooperative Grant Program to be provided (no supported financially) by the U.S. Civilian Research & Development Foundation (CRDF). General scientific field of proposed activity: Mathematics, Systems Engineering. Title for the proposal: "Researches and Applications of the Combinatorial Configurations for Innovative Devices and Process Engineering". U.S. co-investigator S.W. Golomb "University Professor" (EE& Math.), University of Southern California; Ukraine co-investigator V.V.Riznyk, D.Sc., Professor, Lvivska Polytechnika State University (1996).

II. OPTIMUM ORDERED COMBINATORIAL SEQUENCES

A. Optimum Chain Ordered Sequences

The "ordered chain" approach to the study of elements and events is known to be of widespread applicability, and has been extremely effective when applied to the problem of finding the optimum ordered arrangement of structural elements in a distributed technological systems.

Let us calculate all S_n sums of the terms in the numerical n -stage chain sequence of distinct positive integers $C_n = \{k_1, k_2, \dots, k_n\}$, where we require all terms in each sum to be consecutive elements of the sequence. Clearly the maximum such sum is the sum $S_c = k_1 + k_2 + \dots + k_n$ of all n elements; and the maximum number of *distinct* sums is

$$S_n = 1 + 2 + \dots + n = n(n+1)/2 \quad (1)$$

If we regard the chain sequence C_n as being *cyclic*, so that k_n is followed by k_1 , we call this a *ring sequence*. A sum of consecutive terms in the ring sequence can have any of the n terms as its starting point, and can be of any length (number of terms) from 1 to $n-1$. In addition, there is the sum S_n of all n terms, which is the same independent of the starting point. Hence the maximum number of distinct sums S_{max} of consecutive terms of the ring sequence is given by

$$S_{max} = n(n-1) + 1 \quad (2)$$

Comparing the equations (1) and (2), we see that the number of sums S_{max} for consecutive terms in the ring topology is nearly double the number of sums S_n in the daisy-chain topology, for the same sequence C_n of n terms.

B. Two-dimensional Vector Rings

Let us calculate all S sums of the terms in the n -stage ring sequence of non-negative integer 2-stage sub-sequences (2D vectors) of the sequence $C_{n2} = \{(k_{11}, k_{12}), (k_{21}, k_{22}), \dots, (k_{n1}, k_{n2}), \dots, (k_{n1}, k_{n2})\}$ as being *cyclic*, so that (k_{n1}, k_{n2}) is followed by (k_{11}, k_{12}) , where we require all terms in each *modular 2D vector sum* to be consecutive elements of the *cyclic sequence*, and a modulo sum m_1 of k_{12} and a modulo sum m_2 of k_{22} are taken, respectively. A *modular 2D vector sum* of consecutive terms in this sequence can have any of the n terms as its starting point, and can be of any length (number of terms)

from 1 to $n-1$. Hence the maximum number of such sums is given by

$$S = n(n-1) \tag{3}$$

If we require all modular vector sum of consecutive terms give us each vector value exactly R times, than

$$S_R = \frac{n(n-1)}{R} \tag{4}$$

Let $n = m_1$, $n - 1 = m_2$, then a space coordinate grid $m_1 \times m_2$ forms a frame of two modular (close-loop) axes modulo m_1 and modulo m_2 , respectively, over a surface of torus as an orthogonal two modulo cyclic axes of the system being the product of two ($t=2$) circles. We call this two-dimensional Ideal Vector Ring (2D IVR), shortly "Vector Ring".

Example: Let $n=3$, $m_1=2$, $m_2=3$, $R=1$, and complete set of the IVRs takes four variants as follows:

- (a) $\{(0,1),(0,2),(1,2)\}$; (b) $\{(0,1),(0,2),(1,1)\}$;
- (c) $\{(0,1),(0,2),(1,0)\}$; (d) $\{(1,0),(1,1),(1,2)\}$.

To see this, we observe that ring sequence $\{(0,1), (0,2), (1,2)\}$ gives the next circular vector sums to be consecutive terms in this sequence:

$$\left. \begin{aligned} (0,1) + (0,2) &= (0,0) \\ (0,2) + (1,0) &= (1,2) \\ (1,2) + (0,1) &= (1,0) \end{aligned} \right\} \pmod{2, \text{ mod } 3} \tag{5}$$

So long as the terms (0,1), (0,2), (1,2) of the three-stage ($n=3$) ring sequence themselves are two-dimensional vector sums also, the set of the modular vector sums ($m_1=2, m_2=3$) forms a set of nodal points of annular reference grid over 2×3 exactly once ($R=1$):

$$\begin{matrix} (0,0) & (0,1) & (0,2) \\ (1,0) & (1,1) & (1,2) \end{matrix}$$

Easy check to see, that the rest of these ring-sequences has the principal property of forming reference grid 2×3 over a torus using only three ($n=3$) two-stage ($t=2$) terms of these ring sequences.

Schematic model of two-dimensional Vector Ring in torus system of reference is given below (Fig.1) as the simplest and well useful for analytic study of two-dimensional Vector Rings.

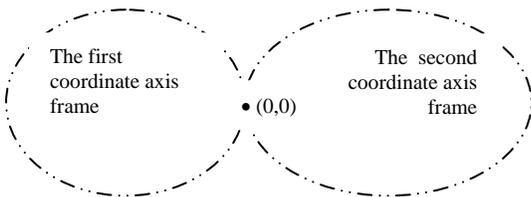


Fig.1. Schematic model of two-dimensional Vector Ring in torus system of coordinates with ground coordinate (0,0).

Easy check to see, that the rest of ring sequences have the principal property of forming reference grid 2×3 over a torus using only three ($n=3$) two-stage ($t=2$) terms of these circular sequences.

III .MULTIDIMENSIONAL VECTOR CYCLIC GROUPS

A. Principal Consideration

To discuss concept of Vector Cyclic Groups (VCG)s let us regard structural model of t -dimensional vector ring as ring n -sequence $C_{nt} = \{K_1, K_2, \dots, K_i, \dots, K_n\}$ of t -stage sub-sequences (terms) $K_i = (k_{i1}, k_{i2}, \dots, k_{it})$ each of them to be completed with nonnegative integers (Fig.2).

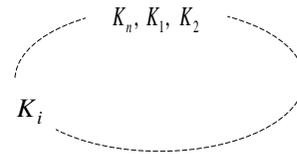


Fig.2. Schematic model of t -dimensional n -stage ring sequence.

Here is an example of 3D Vector Ring with $n= 6$, $m_1=2$, $m_2 =3$, $m_3=5$, and $R=1$ which contains circular 6-stage sequence of 3-stage ($t=3$) sub-sequences $\{K_1, \dots, K_6\}$: $K_1 \Rightarrow (k_{11}, k_{21}, k_{31}) = (0,2,3)$; $K_2 \Rightarrow (k_{12}, k_{22}, k_{32}) = (1,1,2)$; $K_3 \Rightarrow (k_{13}, k_{23}, k_{33}) = (0,2,2)$; $K_4 \Rightarrow (k_{14}, k_{24}, k_{34}) = (1,0,3)$; $K_5 \Rightarrow (k_{15}, k_{25}, k_{35}) = (1,1,1)$; $K_6 \Rightarrow (k_{16}, k_{26}, k_{36}) = (0,1,0)$. The set of all circular sums over the 6-stage sequence, taking 3-tuple ($t=3$) modulo (2,3,5) gives the next result:

$$\begin{aligned} (0,0,1) &= ((0,2,2) + (1,0,3) + (1,1,1)), \\ (0,0,2) &= ((1,1,2) + (0,2,2) + (1,0,3)), \\ (0,0,3) &= ((0,2,3) + (0,1,0)), \\ (0,0,4) &= ((0,2,2) + (1,0,3) + (1,1,1) + (0,1,0) + (0,2,3)), \\ (0,1,1) &= ((0,2,2) + (1,0,3) + (1,1,1) + (0,1,0)), \\ (0,1,2) &= ((1,0,3) + (1,1,1) + (0,1,0) + (0,2,3)), \\ (0,1,3) &= ((1,1,1) + (0,1,0) + (0,2,3) + (1,1,2) + (0,2,2)), \\ (0,1,4) &= ((0,1,3) + (1,1,1)), \\ (0,2,0) &= ((0,2,3) + (1,1,2) + (0,2,2) + (1,0,3)), \\ (0,2,1) &= ((1,1,1) + (0,1,0) + (0,2,3) + (1,1,2)), \\ &\dots \dots \dots \\ (1,2,4) &= ((0,2,3) + (1,1,2) + (1,1,1) + (1,0,3) + (0,1,0)). \end{aligned}$$

Easy to see this verify of the theoretical proposition (6).

$$\prod_1^t m_i = \frac{n(n-1)}{R}, \quad (m_1, m_2, \dots, m_t) = 1 \tag{6}$$

B. Vector Ring Sequences as Cyclic Groups

Next, we consider a set of Vector Rings with $n=3, m_1=2, m_2=3, R=1$ as a cyclic multiplicative group of a finite field. With this aim let us multiply the VR $\{(0,1),(0,2),(1,2)\}$ through by 2D ($t=2$) coefficient (1,2) taking both (mod 2), and (mod 3) as follows: $(0,1) \cdot (1,2) \Rightarrow (0,2), (0,2) \cdot (1,2) \Rightarrow (0,1), (1,2) \cdot (1,2) \Rightarrow (1,1)$. As a result of this transformation we got circular sequence $\{(0,2),(0,1),(1,1)\}$ different from the previous but it is the same as the sequence (b), and the reverse transform by the multiplicative coefficient is true. However, multiplying circular sequences (b) or (c) through by (1,2) no transform them to others variants of the sequences but to themselves as combinations of reflection and cyclic shifting. Hence, the complete set of four VRs with $n=3, m_1=2, m_2=3$, and $R=1$ contains both two isomorphic, and two non-isomorphic variants of the sequences, each of them makes it possible to cover the set of nodal points over torus grid 2×3 exactly once ($R=1$) using only three ($n=3$) basic vectors for configure optimum specify coordinates with respect to torus surface frame of reference. A new type of cyclic groups is among the most perfect Vector Rings which properties hold for the same set of the VRs in varieties permutations of terms in the set (e.g. set of two-dimensional Vector Rings $\{(1,0), (1,1), (1,2), (1,3), (1,4)\}$ and $\{(1,0), (1,2), (1,4), (1,1), (1,3)\}$). We call this class of Vector Rings the “Perfect Torus Group” or “Gloria to Ukraine Stars”. We have found numerous families of the Stars.

Here is the simplest and well useful for analytic study and applications of the underlying properties of Torus and Hypertorus Groups for development of new mathematical, physical and technological results.

IV. OPTIMUM MONOLITHIC VECTOR CODES

A. Useful Properties of Optimum Vector Codes

The remarkable properties of Vector Ring Sequences are that all ring sums of vectors in the sequence exhaust the set of vectors of a finite modular vector space by R ways exactly, which allows on binary encoding of two- and multidimensional vectors as sequences of the same signals or characters in ring code combination length. This makes it possible to use *a priori* maximal number of combinatorial varieties of ring sums for coded design of signals (6). As an example it is chosen the VR sequence $\{(1,1), (0,1), (2,2), (2,1)\}$ with $n=4, m_1=3, m_2=4, R=1$. Here digit weight of the first position is vector value (1,1), the next – (0,1), (2,2), and (2,1) formed a circle. Here is result of the code design (Table 1).

TABLE 1.
Vector code based on the VR $\{(1,1), (0,1), (2,2), 2,1\}$

Vector	Code	Vector	Code	Vector	Code
(0,0)	1110	(1,0)	0111	(2,0)	1011
(0,1)	0100	(1,1)	1000	(2,1)	0001
(0,2)	1000	(1,2)	1100	(2,2)	0010
(0,3)	1101	(1,3)	0011	(2,3)	0110

We can see that sequence $\{(1,1),(1,0),(0,2),(1,1)\}$ forms complete set of ring code combinations on 2D ignorable array 3×4 , and each of its occurs exactly once ($R=1$). Note, each of them forms massive arranged (solid parts of bits) both symbols “1” and of course “0” in the all possible binary circular code combinations. This property makes VR codes useful in applications to coded design of signals for communications, control systems and vector computing with a limited number of bits and improving noise immunity.

B. Definitions of the Ring Monolithic Vector Codes

a) Ring Monolithic Code is a set of ring sequence code combinations which the same characters arranged all together into the code combinations.

b) Numerical Optimum Ring Code is weighed binary Ring Monolithic Code which ring n - sequence of positive integer digit weights forms a set of binary n -digital code combinations of a finite interval $[1,S]$, the sums of connected digit weights taken modulo $S=n(n-1)/R$ enumerate the set of integers $[1,S]$ exactly R -times.

c) Two-dimensional Optimum Ring Code is weighed binary Ring Monolithic Code which set of connected 2-stage modular sums taken modulo m_1 and m_2 , respectively, allows an enumeration of nodal points of reference grid $m_1 \times m_2$ over torus exactly R -times with respect to torus surface frame of axes, $m_1 \cdot m_2 = n(n-1)/R$.

d) Multidimensional Optimum Ring Code is weighed binary Ring Monolithic Code which set of connected t -stage modular sums taken modulo m_1, m_2, \dots, m_t , respectively, allows an enumeration of nodal points of reference grid $m_1 \times \dots \times m_t$ over hypertorus exactly R -times, $m_1 \cdot m_2 \dots \cdot m_t = n(n-1)/R$.

V. CONCLUSION

Concept of the systems optimizations provides, essentially, a new model of technical systems. Moreover, the optimization has been embedded in the underlying combinatorial models. The favorable qualities of the Ideal Vector Rings provide breakthrough opportunities to apply them to numerous branches of science and advanced technology, with direct applications to vector data coding and information technology, signal processing and telecommunications, and other engineering areas. Structural perfection and harmony has been embedded not only in the abstract models but in vectorial laws of the real world [4], [5].

ACKNOWLEDGMENTS

Author thanks to University Professor S.Golomb from University of Southern California for his acceptance of the proposal “Research and Applications of the Combinatorial Configurations for Innovative Devices and Process Engineering” for Cooperative Grants Program from CRDF (U.S.) and to Dr J. Ludvig from Mannheim University for assistance in propagation the concept of Gold Ring Bundles.

REFERENCES

- [1] M. Hall ,Jr. *Combinatorial Theory*. Blaisdell Publ. Comp., Waltham (Massachusetts).Toronto.London, 1967.
- [2] V.Riznyk and O.Bandyrska, "Application of gold ring bundles for innovative non-redundant sonar systems", *Eur. Phys. J.* Springer-Verlag 2008, vol. 154, pp. 183-186, Feb.2008.
- [3] V. Riznyk, "Multidimensional Systems: Problems and Solutions," (Published Conference Proceedings style)," in *Proc. 2nd Colloquium Multidimensional Systems: Problems and Solutions*, IEE, Savoy Place, London, 1998, pp. 5/1-5/5.
- [4] E. P. Wigner, *Symmetries and reflections*. Bloomington-London, Indiana University Press, 1970.
- [5] V. Riznyk, "Multidimensional space hypertorus models" (Received of Nature manuscript style, unpublished work).



Volodymyr V. Riznyk born in Poliss'ke, Kyiv Region, Ukraine, May 21, 1940. PhD Degree, Cybernetics and Theory of Information, Physical and Mechanical Institute, Academy of Sciences, Ukr. SSR, 1980; D Sc, Mathematical Modeling and Mathematical Techniques for Scientific Researches, Vinnytsia State Technical University, 1994. Full professor in Lviv Polytechnic National University, 1995-, University of Technology and Life Sciences Bydgoszcz, Poland, 1996-2012.

Power Station Engineering diploma 1962; Radio Engineering diploma 1967; Power engr, Institution on Rationalization of Electrical Stations, 1962-65, automation engr. 1966-73; fine mechanic Rsch.Lab.U., 1965-66; engr. designer Inst. Automation, 1973-81; tchr., sr. tchr. Lviv Poly.State U., 1981-89, assoc.prof, 1989-95; full prof. Lviv Poly. State U., 1995-, U. Tech. and Life Sci., Bydgoszcz, Poland, 1996-2012. Author: *Synthesis of optimum combinatorial systems*, Franko Lviv State University, 1989; *Combinatorial Models and Techniques of optimization for information problem*, Ministry Educational Office Ukraine, 1991; "Application of the Golden Numerical Rings for Configure Acoustic Systems of Fine Resolution," *Acta Physica Polonica*, vol.119, no.6-A, pp. 1046-1049, June 2011.

Current research interests: The scientific basis of multidimensional optimum distributed systems theory, including the appropriate algebraic structures based on cyclic groups in extensions of Galois fields, and the generalization of these methods and results to optimization of a larger class of technological systems; development of fundamental and applied research in systems engineering for improving such quality indices as reliability, precision, speed, resolving ability, and functionality, using innovative methodologies based on combinatorial techniques; better understanding of the fundamental role of summery and asymmetry relationships in the worldwide harmony laws. Previous research interests: Design and engineering of an improved devices and process engineering for industrial automation of power stations.

Prof. Riznyk. Shewchenko SM Member, Ukraine (1995); Iee Member and Charter Electrical Engineer, UK (1997); WSES Member, USA (1999). DAAD grant, Germany (1997); IEE grant, London (1998); etc. Avocations: stamp and coin collecting, nature and animals.

A Special Constant Acceleration Curve Equation

Mehmet Pakdemirli and İhsan Timuçin Dolapçı

Abstract—An ordinary differential equation describing a curve for which the tangential and normal acceleration components of the object remains constant is derived. The equation and initial conditions are expressed in dimensionless form. In its dimensionless form, the curves are effected only by a parameter which represents the ratio of the tangential acceleration to the normal acceleration. For constant velocity case, a circular arc solution is obtained. For nonzero tangential acceleration, closed form solutions are not available. Using a series solution, the curve is approximated by polynomials. A perturbation solution is also presented. The approximate solutions and the numerical solution are contrasted and within the domain considered, the curves can be successfully approximated by the analytical solutions. Potential application areas can be the design of highway curves, highway exits, railroads, route selection for ships and aircrafts.

Keywords— Curve Design, Highways, Kinematics, Numerical Solution, Perturbation Solution, Series Solution, Vehicle Routes

I. INTRODUCTION

During transportation, aerial, marine and land vehicles cannot travel always in straight routes. Tracking a curved path is inevitable at least for some portion of the travel. To seek for an ideal curve path becomes then a technological problem. Especially at high speeds, smooth transitions in curvatures are needed when entering curved routes. Abrupt changes in the curvatures affect safety and comfort of the travel negatively.

Usually entering to the curves, the velocity should be reduced and the straight path velocity can no longer be maintained. For a constant tangential deceleration, the goal in this study is to seek a specific curve for which the normal acceleration component throughout the curve remains constant.

In curved parts of roads, a special function named clothoid is used [1-3]. The clothoid has the property that its curvature varies linearly with its arc length. Since they are transcendental functions, they have been approximated by polynomials, power series, continued fractions and rational functions [2]. Clothoids are especially useful in transportation engineering, since they can be navigated at constant speed by linear steering and a constant rate of angular acceleration [3]. The curves

derived in this study are not clothoids, since the basic assumption is not a constant velocity with a constant angular acceleration, rather the assumption is a constant tangential and normal (centripetal) acceleration components with respect to the curve.

The equation determining the curve is derived using basic principles of kinematics. Equation and initial conditions are expressed in dimensionless form. The curves depend on a single parameter which is the ratio of the tangential acceleration to the normal acceleration. For vanishing of the parameter, the curve is a circular arc for which constant normal acceleration with constant velocity implies constant radius of curvature. For non-zero parameters, closed form solutions do not exist. The next best choice is to find approximate analytical solutions. A polynomial series solution is constructed to approximate the curve function. Furthermore, the curve parameter is selected as the perturbation parameter and a first order uniform perturbation solution is also presented. Finally, numerical solutions are calculated using a variable step size Runge-Kutta algorithm. It is found that the numerical solutions can be replaced with the approximate solutions in a wide range of the interval.

II. DERIVATION OF THE CURVE EQUATION

Assume that the object enters a curve with initial radius of curvature ρ_0 and velocity v_0 . The object has a constant deceleration a_0 throughout the curve. s is the length coordinate along the curve with $s=0$ representing the entrance and cartesian coordinates are selected as shown in Figure 1.

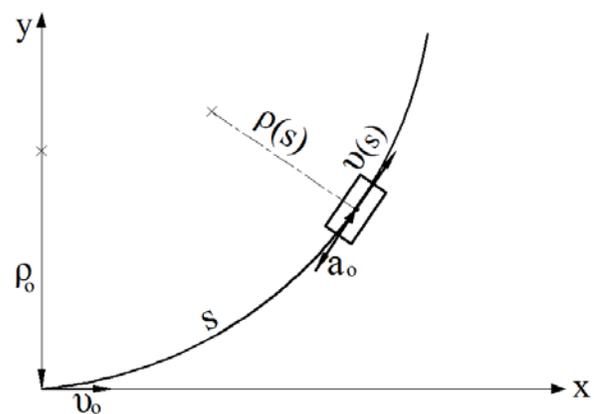


Figure 1- Sketch of the curve

M. Pakdemirli is with the Applied Mathematics and Computation Center, Celal Bayar University, Manisa, TURKEY (corresponding author phone: 90-236-201-2040; fax: 90-236-241-2143; e-mail: mpak@cbu.edu.tr).

İ. T. Dolapçı is with the Department of Mechanical Engineering, Celal Bayar University, Manisa, TURKEY (e-mail: İhsan.dolapci@cbu.edu.tr).

If the normal (centripetal) acceleration [4] remains constant within the curve

$$\frac{v(s)^2}{\rho(s)} = \frac{v_0^2}{\rho_0} \quad (1)$$

where $v(s)$ and $\rho(s)$ represent velocity and radius of curvature at distance s from the entrance. For a constant tangential deceleration component a_0 , the reduced speed at location s is

$$v(s)^2 = v_0^2 - 2a_0s \quad (2)$$

From calculus, the length of a curve and the radius of curvature are given as [5]

$$s = \int_0^x \sqrt{1+y'^2} dx \quad (3)$$

$$\frac{1}{\rho} = \frac{y''}{(1+y'^2)^{3/2}} \quad (4)$$

where prime denotes differentiation with respect to x . Upon substitution of (2)-(4) into (1),

$$\left(v_0^2 - 2a_0 \int_0^x \sqrt{1+y'^2} dx \right) \frac{y''}{(1+y'^2)^{3/2}} = \frac{v_0^2}{\rho_0} \quad (5)$$

solving for the parenthesis, differentiating once to eliminate the integral, and rearranging yields

$$(1+y'^2)y''' - \left(3y' + \frac{2a_0\rho_0}{v_0^2} \right) y'' = 0 \quad (6)$$

which is the differential equation determining a constant normal and tangential acceleration curve. For the specific coordinates chosen, the initial conditions are

$$y(0) = 0, \quad y'(0) = 0, \quad y''(0) = \frac{1}{\rho_0} \quad (7)$$

The first condition is evident from the origin of coordinate location, the second condition requires a tangent slope at the entrance for smooth transition and the last condition is due to the initial curvature of the function. For universality of results, the system is represented in dimensionless form by defining

$$x^* = \frac{x}{\rho_0}, \quad y^* = \frac{y}{\rho_0} \quad (8)$$

and substituting into (6) and (7)

$$(1+y'^2)y''' - (3y' + 2\varepsilon)y'' = 0 \quad (9)$$

$$y(0) = 0, \quad y'(0) = 0, \quad y''(0) = 1 \quad (10)$$

where

$$\varepsilon = \frac{a_0\rho_0}{v_0^2} = \frac{a_0}{v_0^2/\rho_0} \quad (11)$$

For simplicity, the symbol star is not shown on the variables keeping in mind that the variables are all dimensionless. The above differential system defines a constant tangential and normal acceleration curve. The family of curves depend on only one parameter ε which is the ratio of the tangential acceleration to the normal acceleration. Rather than choosing separately the accelerations, radius of curvatures and velocities, it is sufficient to choose ε , the combination of all parameters in the analysis which reduces substantially the calculations and presentations in the form of figures.

III. ANALYTICAL SOLUTIONS

Analytical solutions of the model are presented in this section. The degenerate case of $\varepsilon=0$ can be solved in closed form functions. However, $\varepsilon \neq 0$ case cannot be solved in closed form functions and approximations are inevitable. A series solution as well as a perturbation solution are presented in this section.

For the degenerate case, the equation is

$$(1+y'^2)y''' - 3y'y'' = 0 \quad (12)$$

$$y(0) = 0, \quad y'(0) = 0, \quad y''(0) = 1 \quad (13)$$

A straightforward calculation by employing reduction of order and successive integrations yield

$$y = 1 - \sqrt{1-x^2} \quad (14)$$

which represents a circular arc since $\varepsilon=0$ corresponds to no tangential acceleration and the normal component of the acceleration remains constant only in a circular path if the speed is constant.

Series Solution

Assume a power series solution for the problem with nonzero ε ,

$$y(x) = \sum_{i=0}^{\infty} a_i x^i \quad (15)$$

Initial conditions (13) require

$$a_0 = 0, \quad a_1 = 0, \quad a_2 = \frac{1}{2} \quad (16)$$

Substituting (15) into (9), using (16) yields the polynomial solution

$$y(x) = \frac{1}{2}x^2 + \frac{1}{3}\varepsilon x^3 + \left(\frac{1}{8} + \frac{1}{3}\varepsilon^2 \right) x^4 + \left(\frac{19}{60}\varepsilon + \frac{2}{5}\varepsilon^3 \right) x^5 + \left(\frac{1}{16} + \frac{29}{45}\varepsilon^2 + \frac{8}{15}\varepsilon^4 \right) x^6 + \left(\frac{81}{280}\varepsilon + \frac{388}{315}\varepsilon^3 + \frac{16}{21}\varepsilon^5 \right) x^7 + O(x^8) \quad (17)$$

For vanishing curve parameter, the Taylor expansion of the circular solution (14) is obtained. The original circular solution is an even power polynomial and deformations from this solution with the curve parameter introduces the odd powers also.

Perturbation Solution 1

If the curve parameter is our perturbation parameter, an approximate solution

$$y(x) = y_0(x) + \varepsilon y_1(x) + O(\varepsilon^2) \quad (18)$$

can be constructed. Substituting the expansion into (9) and (10), separating at different orders yields

$$O(1): (1+y_0'^2)y_0''' - 3y_0'y_0'' = 0 \quad (19)$$

$$y_0(0) = 0, \quad y_0'(0) = 0, \quad y_0''(0) = 1$$

$$O(\varepsilon): (1+y_0'^2)y_1''' - 3y_1'y_0'' - 6y_0'y_0''y_1'' + 2y_0'y_1'y_0''' - 2y_0''^2 = 0$$

$$y_1(0) = 0, \quad y_1'(0) = 0, \quad y_1''(0) = 0 \quad (20)$$

The first order solution is the circular arc solution presented before

$$y_0 = 1 - \sqrt{1-x^2} \tag{21}$$

Substituting this solution to the next order

$$y_1''' - \frac{6x}{1-x^2} y_1'' + 3 \frac{2x^2-1}{(1-x^2)^2} y_1' = \frac{2}{(1-x^2)^2} \tag{22}$$

$$y_1(0) = 0, \quad y_1'(0) = 0, \quad y_1''(0) = 0$$

and solving yields

$$y_1 = -\frac{2x-\pi}{\sqrt{1-x^2}} - \frac{2 \operatorname{arccosh}(x)}{\sqrt{x^2-1}} \tag{23}$$

The approximate solution is

$$y(x) = 1 - \sqrt{1-x^2} - \varepsilon \left(\frac{2x-\pi}{\sqrt{1-x^2}} + \frac{2 \operatorname{arccosh}(x)}{\sqrt{x^2-1}} \right) + O(\varepsilon^2) \tag{24}$$

Since the function is not defined near $x=1$, the singularity at this point is unimportant.

Perturbation Solution 2

An alternative perturbation solution can be constructed by assuming the dependent variable to be small. If α is the perturbation parameter, the smallness of the dependent variable is represented by the transformation

$$y(x) = \alpha u(x) \tag{25}$$

and the equation in terms of this transformation becomes

$$(1 + \alpha^2 u'^2) u''' - (3\alpha^2 u' u'' + 2\varepsilon \alpha) u'' = 0 \tag{26}$$

$$u(0) = 0, \quad u'(0) = 0, \quad u''(0) = 1/\alpha \tag{27}$$

The expansion in terms of the perturbation parameter is

$$u(x) = u_0(x) + \alpha u_1(x) + \alpha^2 u_2(x) + O(\alpha^3) \tag{28}$$

Substituting and separation at different orders yields

$$O(1): \quad u_0''' = 0 \tag{29}$$

$$u_0(0) = 0, \quad u_0'(0) = 0, \quad u_0''(0) = 1/\alpha$$

$$O(\alpha): \quad u_1''' - 2\varepsilon u_0'' = 0 \tag{30}$$

$$u_1(0) = 0, \quad u_1'(0) = 0, \quad u_1''(0) = 0$$

$$O(\alpha^2): \quad u_2''' + u_0' u_0'' - 3u_0' u_1'' - 4\varepsilon u_0' u_1' = 0 \tag{31}$$

$$u_2(0) = 0, \quad u_2'(0) = 0, \quad u_2''(0) = 0$$

The equations can be solved consecutively

$$u_0 = \frac{1}{2\alpha} x^2, \quad u_1 = \frac{1}{3\alpha^2} \varepsilon x^3, \quad u_2 = \frac{1}{\alpha^3} \left(\frac{1}{8} + \frac{1}{3} \varepsilon^2 \right) x^4 \tag{32}$$

Substituting into (28) and back transforming to the original variable $y(x)$

$$y(x) = \frac{1}{2} x^2 + \frac{1}{3} \varepsilon x^3 + \left(\frac{1}{8} + \frac{1}{3} \varepsilon^2 \right) x^4 + O(x^5) \tag{33}$$

which is the same solution with the series solution up to the approximation considered.

IV. COMPARISONS WITH THE NUMERICAL SOLUTIONS

The series solution and the first perturbation solution is compared with the numerical solution. Equation (9) and (10) is cast into a suitable form first by defining $y_1 = y, y_2 = y', y_3 = y''$ In terms of the new variables

$$y_1' = y_2$$

$$y_2' = y_3$$

$$y_3' = \frac{(3y_2 + 2\varepsilon)y_3^2}{1 + y_2^2} \tag{34}$$

$$y_1(0) = 0, \quad y_2(0) = 0, \quad y_3(0) = 1$$

the system is reduced to a system with three equations of first order. The above system is solved by employing a variable step size Runge-Kutta algorithm. Figure 2 shows that as the number of terms in the series solution increases, convergence to the numerical solution is achieved. Note that the figure is drawn for a fairly large curve parameter of $\varepsilon=1$. Since the curve parameter is the ratio of tangential acceleration to the normal one, $\varepsilon=1$ corresponds to equal acceleration components.

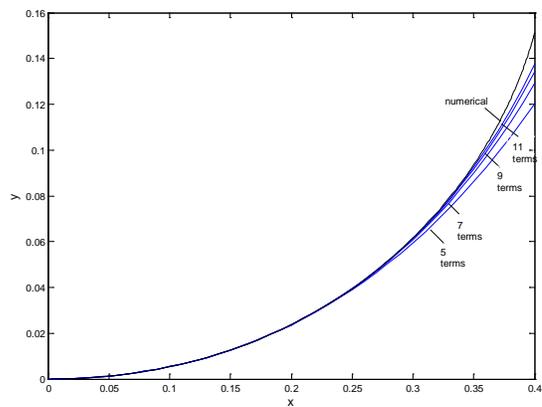


Figure 2- Convergence of series solutions to numerical solution ($\varepsilon=1$)

For $\varepsilon=0.2$, the 7 and 11-term series solution and the perturbation solution is contrasted with the numerical solution in Figure 3. The one correction term perturbation solution (i.e. equation 24) performs slightly better than the 7-term series solution.

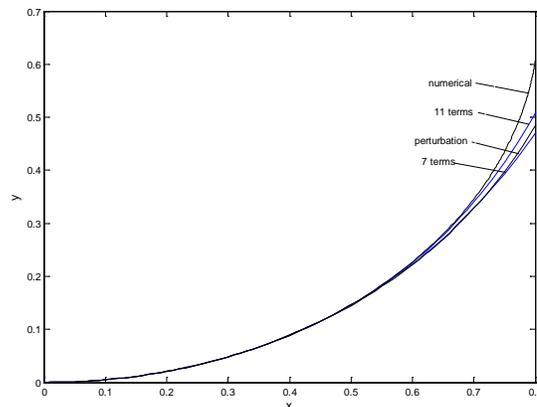


Figure 3- Comparisons of the series, perturbation and numerical solutions ($\varepsilon=0.2$)

Finally an intermediate value of $\varepsilon=0.6$ is considered in Figure 4. Five-term series solution performs slightly better than the perturbation solution. In conclusion, perturbation solution can replace the numerical solution for small curve parameter values. For larger parameter values, the series solution better approximates the numerical solutions.

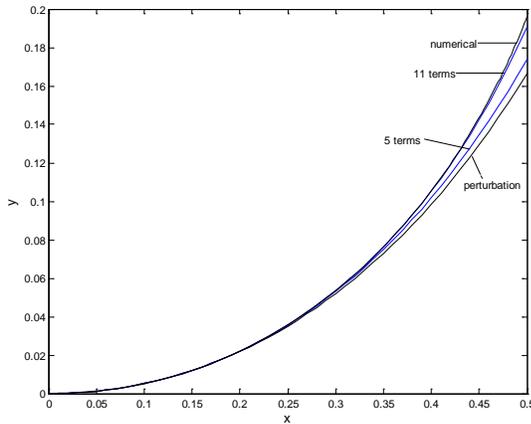


Figure 4- Comparisons of the series, perturbation and numerical solutions ($\varepsilon=0.6$)

In practical calculations, back substitution to dimensional quantities should be done as a final step.

The error analysis is also done for the three curve parameters considered. In Figure 5, the residual error corresponding to $\varepsilon=1$ is presented. As the number of terms increase, the residual error decrease in most of the domain. However, in a narrow region at the right, a reverse behavior is observed and as the number of terms increase, the residual error increases. For larger x values, the higher order polynomial terms added is the reason of this residual error. As can be seen from Figure 2, the absolute error is still smaller for higher term polynomials in this region also.

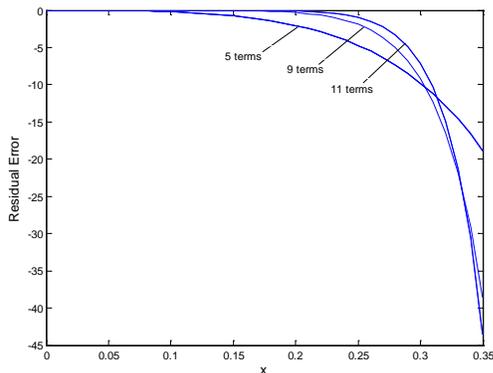


Figure 5- The residual error of polynomial solution ($\varepsilon=1$)

For a smaller value of the curve parameter (i.e. $\varepsilon=0.2$), the residual errors of 7 and 11-term solutions are contrasted. A similar behavior is observed. Adding terms reduces the residual errors in most of the domain except in a narrow region of higher x values.

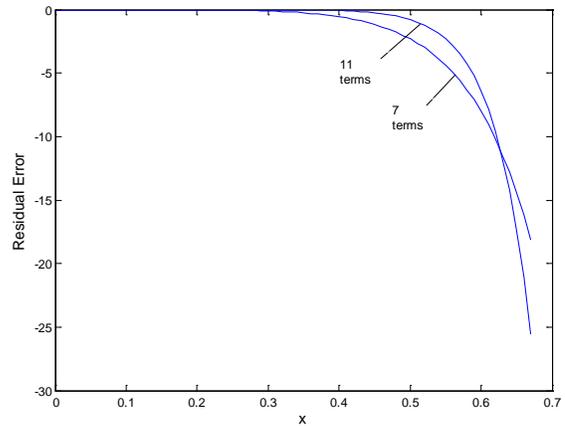


Figure 6- The residual error of polynomial solution ($\varepsilon=0.2$)

Finally, residual error analysis for $\varepsilon=0.6$ is presented in Figure 7. The qualitative behavior is the same with the previous figures.

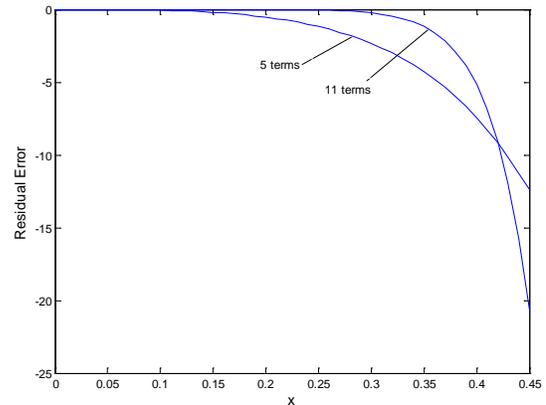


Figure 7- The residual error of polynomial solution ($\varepsilon=0.6$)

In conclusion, to decrease the absolute and residual errors, the series solution should be truncated and not recommended for usage for high x values.

V.CONCLUDING REMARKS

A new family of curves which can be used in highways, routes of marine and aerial vehicles is derived. Throughout the curve, the tangential and normal accelerations remain constant, providing a comfortable transport within the vehicle. Although the curves are derived under the assumption of tangential deceleration, if they are tracked from the reverse, the curves will represent motion with tangential acceleration. An application might be a highway entrance, where the vehicle should adjust its speed to the higher speed of the highway. The curves outlined may find other application areas such as manufacturing engineering in the future.

ACKNOWLEDGMENT

Mehmet Pakdemirli acknowledges the travel and research support of the Turkish Academy of Sciences.

REFERENCES

- [1] D. S. Meek and D. J. Walton, "Clothoid spline transition spirals," *Mathematics of Computation*, vol 59, pp. 117-133, 1992.
- [2] D.S. Meek and D.J. Walton, "An arc spline approximation to a clothoid," *Journal of Computational and Applied Mathematics*, vol. 170, pp. 59–77, 2004.
- [3] J. McCrae and K. Singh, "Sketching Piecewise Clothoid Curves," *EUROGRAPHICS Workshop on Sketch-Based Interfaces and Modeling* (C. Alvarado and M. P. Cani (Editors)), 2008.
- [4] F. P. Beer, E. R. Johnston Jr., D. F. Mazurek, P. J. Cornwell and E. R. Eisenberg, *Vector Mechanics for Engineers: Statics & Dynamics*, New York: The McGraw-Hill Companies, 2010.
- [5] G. Strang, *Calculus*, Wellesley: Wellesley-Cambridge Press, 1991.

Application of statistic complexity metrics to detection of malware threats in autonomic component ensembles

A.Prangishvili, O.Shonia, I.Rodonaia, V.Rodonaia

Abstract— The paper proposes a new technique for detecting malware threats in autonomic component ensembles. The technique is based on the statistic complexity metrics, which relate objects to random variables and (unlike other complexity measures considering objects as individual symbol strings) are ensemble based. This transforms the classic problem of assessing the complexity of an object into the realm of statistics. The proposed technique requires implementation of the process X (which generates ‘healthy’ flows containing no malware threats) and objects generated by the actual (possible infected) process Y . The component flows files are used as objects of the processes X and Y . The result of the proposed procedure gives us the distribution of probabilities of malware infection among autonomic components.

Keywords— autonomic ensemble, complexity measure, statistic complexity, traffic flows, malware

I. INTRODUCTION

THE problem of anomaly detection in autonomic component ensembles was considered in [1], [2], where the following problem was set. A singleton application currently runs on one of the VMs at a Datacenter. During the session the application experiences consistently high CPU load. This increase may be caused either by legitimate traffic overload or by coordinated attacks (DDOS) launched against the PaaS provider. The latter might be wrongly assumed to be legitimate requests and resources would be scaled up to handle them. This would result in an increase in the cost of running the application (because provider will be charged by these extra resources) as well as in violation of SLA (due to increased response times). Hence, it is necessary to distinguish between these two cases, the earlier this distinction is made, the higher is the degree of protection of the application from

A.P. Author is with the Faculty of Informatics, Georgian Technical University, 77 Kostava Str., Tbilisi, 0175, Georgia (e-mail: rectoroffice@gtu.ge).

O. S. Author is with the Faculty of Informatics, Georgian Technical University, 77 Kostava Str., Tbilisi, 0175, Georgia (e-mail: o.shonia@gtu.ge).

I.R. Author is with the Faculty of Informatics, Georgian Technical University, 77 Kostava Str., Tbilisi, 0175, Georgia (e-mail: irodonaia@yahoo.com).

V.R. Author is with the Faculty of Informatics, Georgian Technical University, 77 Kostava Str., Tbilisi, 0175, Georgia (e-mail: vrodona@yahoo.com).

failure and poor performance. To provide this protection, the following security measures were suggested. The traffic flows through the VM_i had to be analyzed using Kolmogorov complexity metrics. During the session the constant monitoring of the metric (by the special probe implemented in the separate module), along with measure of CPU load and available memory size, was being executed. If the traffic satisfied some pre-formulated criteria (indicated that there exist serious DDOS attack threats) then the application rapidly migrated to some other VM_j .

The technique described in [1], [2] implemented Kolmogorov complexity metrics to reveal possible malware attacks and had to deal only with DDOS attacks. Despite its usefulness, Kolmogorov complexity does not capture the intuitive notion of complexity very well. For example, random strings without any regularities, say, strings that are constructed bitwise by repeated tosses of a fair coin, have very large Kolmogorov complexity. However, those strings are not “complex” from an intuitive point of view — those strings are completely random and do not carry any interesting structure at all. Many approaches have been suggested to define some complexity measure that is closer to the intuitive notion of complexity and overcomes the difficulties of Kolmogorov complexity. For example, Kolmogorov complexity is based on algorithmic information theory considering objects as individual symbol strings, whereas the measures *effective measure complexity* (EMC), *excess entropy*, *predictive information*, etc., relate objects to random variables and are *ensemble* (that is, set of interrelated objects –symbol strings) based

The Kolmogorov complexity measures M assigns a complexity value to each individual object x' under consideration. Let's denote it as $C_M(x')$. It is assumed that x' corresponds to a string sequence of a certain length and its components assume values from a certain domain. In [3] *statistic complexity* that is not only different to all other complexity measures introduced so far, but also connects directly to statistics, specifically, to statistical inference, was introduced. More precisely, a complexity measure with the following properties is introduced. First, the measure is bivariate comparing two objects, corresponding to pattern generating processes, on the basis of the *normalized compression distance* (NCD) [4] with each other:

$$NCD(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

where $C(x)$ denotes the compression size of string x and $C(xy)$ the compression size of the concatenated strings x and y .

Second, this measure provides the quantification of an error that could have encountered by comparing samples of finite size from the underlying processes. Hence, the statistic complexity provides a statistical quantification of the statement ‘ X is similarly complex as Y ’. This implies that a fundamental complexity measure needs to be bivariate, $C(X, Y)$, instead of univariate comparing two processes X and Y .

Next, the desirable property of any complexity measure is : a complexity measure should quantify the uncertainty of the complexity value. As motivation for this property we just want to mention that there is a crucial difference between an observed object x' and its generating process X . If the complexity of X should be assessed, based on the observation x' only, this assessment may be erroneous. This error may stem from the limited (finite) size of observations. Also, the possibility of measurement errors would be another source of wrong assessment.

Based on these considerations, the statistic complexity measure, suggested in [3], is defined by the following procedure:

1. Estimate the empirical distribution function \hat{F}_{XX} of the normalized compression distance from n_1 ,
 $S_{X,X}^{n_1} = \{x_i = NCD(x', x'') | x', x'' \prec X\}_{i=1}^{n_1}$, from objects x' and x'' of size m generated by process X (here ‘ \prec ’ means ‘is generated by X ’)
2. Estimate the empirical distribution function \hat{F}_{XY} of the normalized compression distance from n_2 ,
 $S_{X,Y}^{n_2} = \{y_i = NCD(x', y') | x' \prec X, y' \prec Y\}_{i=1}^{n_2}$ from objects x' and y' of size m generated by two different processes X and Y
3. Determine $T = \sup_x \left| \hat{F}_{X,X}(x) - \hat{F}_{X,Y}(x) \right|$ and
 $p = Prob(T \leq t)$
4. Define $C_s(S_{X,X}^{n_1}, S_{X,Y}^{n_2} | X, Y, m, n_1, n_2) := p$ as
statistic complexity

This procedure corresponds to a two-sided, two-sample Kolmogorov-Smirnov (KS) test based on the normalized compression distance [4] obtaining distances among observed objects.

The statistic complexity corresponds to the p-value of the underlying null hypotheses, $H_0 : F_{XX} = F_{XY}$, and, hence, assumes values in $[0,1]$. The null hypothesis is a statement about the null distribution of the test statistic $T = \sup_x \left| \hat{F}_{X,X}(x) - \hat{F}_{X,Y}(x) \right|$, and because the distribution functions are based on the normalized compression distances

among objects x' and x'' , drawn from the processes X and Y , this leads to a statement about the distribution of normalized compression distances. Hence, verbally, H_0 can be phrased as “on average, the compression distance of objects from X to objects from Y equals the compression distance of objects only taken from X ”. If the alternative hypothesis, $H_1 : F_{XX} \neq F_{XY}$ is true, this equality does no longer hold implying differences in the underlying processes X and Y , leading to differences in the NCDs.

Applied to the problem of finding malware threats in the flows between autonomic components CP_i ([1], [2]) the above procedure will look as follows. For *each autonomic component* (AC) of the *autonomic-component ensembles* (ACEs) the processes X and Y are considered as the processes generating objects represented in the form of strings. The strings, in turn, represent traffic flows through these autonomic components. The specific ways of how flows are transformed into strings are considered later in the paper. The process X (‘training process’) is the process generating flows in the conditions when there are no malware threats. So, objects (strings) generated by the process X are ‘healthy’ (they do not contain any patterns of malware). These strings have to be generated preliminary (before actual workload on an autonomic components ensemble). Some fraction of objects (string) have to be generated for situation with unusual (but not malicious) behavior. For randomly taken pairs x' and x'' (the amount of such pairs is n_1) of the generated strings the metric $NCD(x', x'')$ is calculated. The size of samples n_1 has to be sufficient to account for various possible situations and conditions that may occur in the specific autonomic ensemble under consideration. Then the empirical distribution function \hat{F}_{XX} is being built and stored to the specific place.

When the ensemble starts actual operation (receives workload), the process Y (‘production process’) generates objects (strings) y' , which represent actual current traffic between ensemble’s components. Some of these objects may contain malware patterns. The sample of the size n_2 of objects x' (generated preliminary by the ‘training process’ X) and objects y' is being created and the metric $NCD(x', y')$ is calculated for each pair. Then the empirical distribution function \hat{F}_{XY} is being built. Now, by applying the steps 3 and 4 of the above procedure, the values of the *statistic complexity* for *each autonomic component* can be computed.

The obtained numerical value of the statistic complexity can be interpreted in the following sense: in the current conditions the flows of packets through the given autonomic component cannot be regarded as complex flows (with the probability equal to p). That is, the flows may contain some patterns (indicating the possible presence of some malware threats) with the probability p .

It should be pointed out that in production conditions (when the ensemble is under actual workload) the sample size

n_2 cannot be determined in advance. This size depends on actual working conditions: traffic intensity, frequency of creation of objects (strings), actual hardware indices (CPU load, available memory, etc.). As a rule, the number n_2 is less than the number n_1 . This fact can somewhat decrease the precision of the metric, but it requires additional technical consideration. In general, the statistic complexity has the very desirable property that the power reaches asymptotically 1 when $n_1 \rightarrow \infty$ and $n_2 \rightarrow \infty$. This means, for infinite many observations the error of the test to falsely accept the null hypotheses when in fact the alternative is true becomes zero. Formally, this property can be stated as $p \rightarrow 0$ for $n_1 \rightarrow \infty$ and $n_2 \rightarrow \infty$.

Finally, note that despite the fact that statistic complexity is a statistical test, it borrows part of its strength from the NCD and, respectively, Kolmogorov complexity on which this is based on. Hence, it unites various properties from very different concepts.

II. APPLICATION OF STATISTIC COMPLEXITY TO AUTONOMIC COMPONENTS ENSEMBLES.

In the proposed approach to anomaly detection in autonomic component ensembles, an attempt to deal with wide range of malware threats has been made (unlike the techniques described above and in [1], [2], which had to deal only with DDOS attacks).

In autonomic cloud computing datacenters can be considered as autonomic-component ensembles (ACEs) and be represented by constructions of SCEL (Software Component Ensemble Language), a kernel language for programming autonomic computing systems ([1], [5], [6]). Each (virtual) machine is running one instance of the Cloud Platform called Cloud Platform instance (CP_i). Each CP_i is considered to be a service component. Multiple CPs communicate over the Internet (IP protocol), thus forming a cloud and within this cloud one or more service component ensembles. The notions of autonomic components (ACs) and autonomic-component ensembles (ACEs) ([5], [6]) have been put forward as a means to structure a system into well understood, independent and distributed building blocks that interact in specified ways.

The process part of a component (Fig.1) is split into an *autonomic manager* controlling execution of a *managed element*. The autonomic manager monitors the state of the component, as well as the execution context, and identifies relevant changes that may affect the achievement of its goals or the fulfillment of its requirements. It also plans adaptations in order to meet the new functional or non-functional requirements, executes them, and monitors that its goals are achieved, possibly without any interruption. A managed element can be seen as an empty "executor" which retrieves from the knowledge repository the process implementing a required functionality *id* and bounds it to a process variable *Z*, sends the retrieved process for execution and waits until it

terminates. Also actual parameters for the process to be executed can be stored as knowledge items and retrieved by the executor (or by the process itself) when needed.

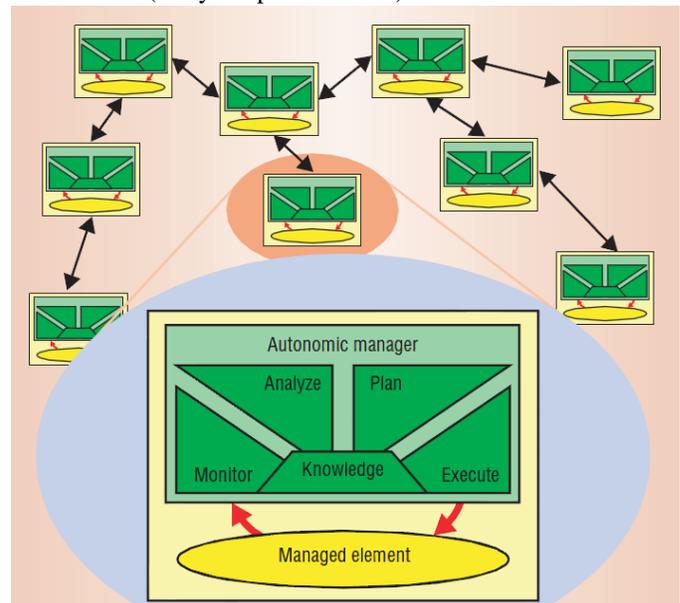


Fig.1 Functional description of a component

In our approach the notions of *netflows*, their *informational-theoretical metrics* and components' *autonomic manager* are essentially leveraged. A network *flow* can be defined in many ways. In a general sense, a flow is a series of packets with some attribute(s) in common. Each packet that is forwarded within a router or switch is examined for a set of IP packet attributes. These attributes are the IP packet identity or fingerprint of the packet and determine if the packet is unique or similar to other packets. All packets with the same source/destination IP address, source/destination ports, protocol interface, and class of service are grouped into a flow and then packets and bytes are labeled. This methodology of fingerprinting or determining a flow is scalable because a large amount of network information is condensed into a database of netflow information called the netflow cache.

A *netflow-enabled device* (*netflow exporter*: router or switch) (see the Fig.2) sends to the *netflow collector* single flow as soon as the relative connection expires. This can happen when 1) when TCP connection reaches the end of the byte stream (FIN flag or RST flag) are set; 2) when a flow is idle for a specific timeout; 3) if a connection exceeds long live terms (30 minutes by default). Packets captured by the netflow collector are stored to a *flow storage*. In our approach the duration of each flow's formation time is unknown in advance and actually is defined by relevant collectors on the basis of the selected connection expiration time criteria.

Flows accumulated at the flow storage, are then subdivided into *component flows*. That is, flows which have the component's IP address as a destination address are grouped and sent to the corresponding component (more exactly, to the *autonomic manager* of a component - these flows are marked with blue arrows in the Fig.2).

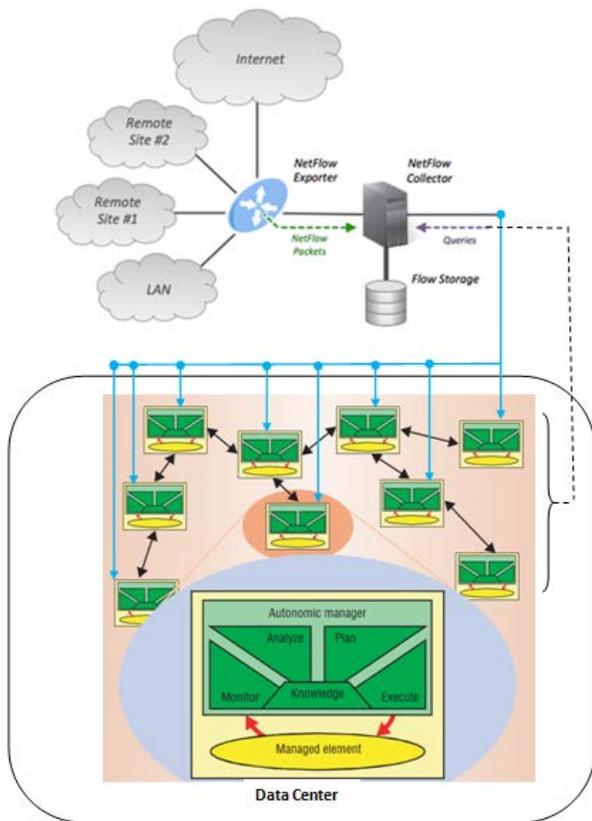


Fig.2 Interaction between netflow devices and autonomic components

After receiving their destined flows, the component’s autonomic manager can start the processing in order to reveal the abnormal behavior of flows in accordance with the following technique.

Application for collecting and processing NetFlow statistics are defined below (Fig.3):

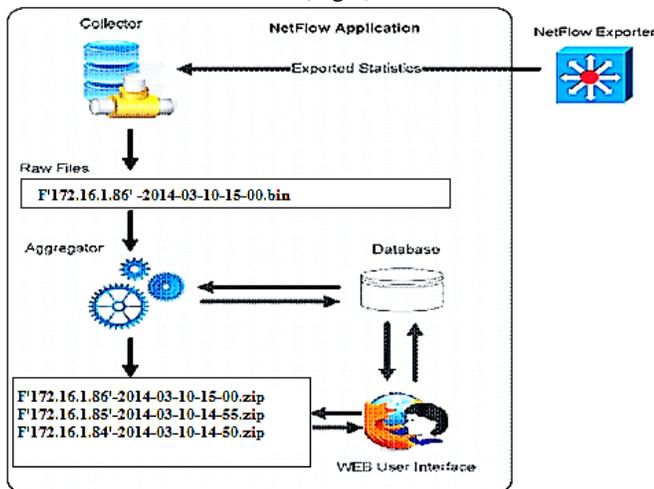


Fig. 3: Components of the NetFlow system for analysis of the statistics

Once the collector populates the raw file, the file is passed on to the second component in the system, which is called an aggregator. The aggregator receives the file from the collector and processes it using predefined information from the database. The data thus processed (aggregated) is stored in the

database. The user interface is a web application that enables us to obtain information on the status of the network, based on the data aggregated in the database. If it is necessary to get more detailed information about a specific communication, the user may open the relevant raw file via the web and filter it according to the desired criteria. The location of the device collecting NetFlow statistics depends on the architecture of the network itself. The amount of NetFlow information exported by network devices is directly dependent on the amount of traffic passing through that device (exporter). Experience has shown that the amount of NetFlow traffic does not exceed 1% of the total amount of traffic through the network, so the “distance” between the server (collector) and the network device exporting the data (exporter) is not relevant. The accessibility and the security of the server are the more important parameters.

In the proposed approach the different files with the particular titles (relevant to the concrete SCP_i’s IP addresses) to store component flows are used. For example, for the component flow to the SCP_i with IP address 172.16.1.86, occurred on 2014/03/16 at 15:00, the files with titles F’171.16.1.86’-2014-03-10-15-00.bin and the F’171.16.1.86’-2014-03-10-15-00.zip will be created.

If we look at known threats in data networks from point of unwanted traffic, we can separate the following groups [7]:

1. Denial of service attacks.
2. Port scans and remote vulnerability searching and virus spread.
3. P2P files exchange networks.
4. Email spam and web popup.
5. Open resources misuse (open DNS, open mail relay, open proxy, Trojan horse, etc.)

In our approach we observe the following traffic flow attributes (8):

- Source/destination IP address and port number
- To measure changes in IP address and port number space we observe a value of Shannon entropy related to these attributes (entropy is used to capture the degree of dispersal or concentration of the distributions for traffic attributes). Entropy values are calculated for separate component flows files (obtained by using the utility *nfdump*,). Different AMs (Autonomic Manager) use various time periods length (see connection expiration time criteria above). The following network variables are used for each component flows files: entropy of source IP address, entropy of destination IP address, entropy of destination port number, entropy of source port number . Duration attributes of each component flow time are different and depend on the traffic conditions and selected connection expiration time criteria.
- Number of bytes and packets
- These values are: bytes received by a host, bytes sent by a host, packets received by a host, packets sent by a host. Again, duration attributes of each component flow files are different
- TCP flags

The attribute TCP_FLAG - a difference between number of SYN packets sent and RST and FIN packets received - is measured in the proposed approach. In normal conditions, in long time observation we should get the mean value of TCP_FLAG near zero. Intrusive actions like system scanning, DoS attacks, may cause the temporal distortion of the mean value of TCP_FLAG

- Duration of the connection

During various types of attacks, this value will be affected and so an anomaly may be detected. For example, worm infection will generate a large number of connections with similar duration. We simply use the value of connections' duration attribute contained in the given component flow file.

- Communication Patterns

Fan-in is the number of nodes that originate data exchange with the current CP_i , while *Fan-out* is the number of hosts to which CP_i initiates conversations. The above patterns are invariant during most time of normal system activity or change in a predictive way. But while attack appears they will change significantly.

As one can see, the component flows files contain the same volume of information (they contain the same amount of attributes of the same size). Hence, we can assume that the size m of a component flow file represents the object (in terms of the statistic complexity procedure) of size m . In general, component flows files are regarded as objects x', x'', x''', \dots generated by the process X ('training process') and $y', y'', y''' \dots$, objects generated by the process Y ('production process').

As it was described, the proposed procedure requires implementation of the 'training process' X (which generates 'healthy' flows containing no malware threats) before starting real 'production' (real-time) process Y . In order to decrease overheads, this process is executed just once with as large value of the sample size n_i as it is possible. The obtained results (the empirical distribution function \hat{F}_{XY}) is stored to each CP_i which can run applications subsequently. When applications are executed on the CPs, the objects $y', y'', y''' \dots$, (corresponding component flows files) are created and the empirical distribution functions \hat{F}_{XY} are calculated on each CP_i . Then, according to the steps 3 and 4 of the procedure, the value of statistic complexity for each autonomic component is calculated.

The result of the proposed procedure gives us the distribution of probabilities of malware infection among autonomic components of the datacenter.

To estimate the statistic complexity's value, which practically indicates real malware threat, numerous simulation experiments were carried out. The well-known simulation tool *CloudSim* - a framework for modeling and simulation of cloud computing infrastructures and services - has been used. As a result of simulation experiments we determined that the statistic complexity's value larger than 0.4 can be practically

regarded as serious malware threat. In this condition the immediate migration of the application from the VM (where the application is being run currently) to another VM (which is to be selected by using the ensemble's components autonomic managers' knowledge base and issuing the special SCCL statement **qry**) is required.

It should be pointed out that detection of malware threats and consequent migration are being executed in real-time scale and thus minimize damage from possible malware threats. This also contributes to maintaining the required SLA.

The time of migration must be taken into account when determining the response time. In general, streams of requests generated by each client (application) may be decomposed into a number of different VMs. In case of more than one VM serving the i^{th} client, requests are assigned probabilistically, i.e., α_{ij} portion of the incoming requests are forwarded to the j^{th} server (host of a VM) for execution.

The exponential distribution function is used to model the service time of the clients in this system. Based on this model, the response time distribution of a VM (placed on server j) is an exponential distribution with mean:

$$\bar{R}_{ij} = \frac{1}{C_j^p \phi_{ij} \mu_{ij} - \alpha_{ij} \lambda_i} \tag{1}$$

where μ_{ij} denotes the service rate of the i -th client on the j -th server when a unit of processing capacity is allocated to the VM of this client. The VM unit is defined as the basic unit of virtual resource, which is associated with a set of physical resources such as CPU time, main memory, storage space, electricity etc. In real cloud systems, any virtual resource a customer can apply should be a multiple of the VM unit.

Migrating a VM between servers causes a downtime in the client's application. Duration of the downtime is related to the migration technique used in the datacenter. The downtime also is the function of the link speed and VM memory size.

Let's assume that an application i had to migrate n_i times during its execution cycle. We introduce the following notations:

- n_i - amount of migration of the i -th application during its execution cycle;
- m_k - the number (index) of VM (CP) on which the application runs in k -th migration period;

SC_{ip} - the value of the statistic complexity obtained for the i -th application running on the p -th VM in the given time period

\bar{R}_{ij} - see (1)

Then the formula (1) must be updated by adding the term representing the expected downtime of the VM_{ij} :

$$\begin{cases} \bar{R}_{im_i} & \text{if } n_i = 0 \\ \sum_{k=1}^{n_i} (SC_{im_k} * (\bar{R}_{im_k} + DT_{im_k} (LinkSpeed))) & \text{otherwise} \end{cases}$$

The obtained estimation of response times is much closer to actual response times (observed in real operational conditions) and thereby contributes to maintaining the required SLA.

III. CONCLUSION

In the paper we presented a new technique for detecting malware threats in autonomic component ensembles. The technique is based on the statistic complexity metrics. Unlike the Kolmogorov complexity, which is based on algorithmic information theory considering objects as individual symbol strings, the statistic complexity relate objects to random variables and are ensemble based. It is a bivariate measure that compares two objects, corresponding to pattern generating processes, on the basis of the normalized compression distance with each other. Besides, this measure provides the quantification of an error that could have been encountered by comparing samples of finite size from the underlying processes. The approach transforms the classic problem of assessing the complexity of an object into the realm of statistics. This may open a wider applicability of this complexity measure to diverse application areas. In particular, the statistic complexity is applied to the problem of detecting malware threats in autonomic component ensembles. The proposed procedure requires implementation of the ‘training process’ X (which generates ‘healthy’ flows containing no malware threats) and objects generated by the actual (possible infected) process Y (‘production process’). The component flows files are used as objects of the processes X and Y . The result of the proposed procedure gives us the distribution of probabilities of malware infection among autonomic components of the datacenter. The proposed procedure of detecting malware threats and consequent migration are being executed in real-time scale and thus minimizes damage from possible malware threats. This also contributes to maintaining the required SLA.

REFERENCES

- [1] A. Prangishvili, O. Shonia, I. Rodonaia, V. Rodonaia. Formal security modeling in autonomic cloud computing environment. WSEAS / NAUN International Conferences, Valencia, Spain, 2013
- [2] A. Prangishvili, O. Shonia, I. Rodonaia, M. Mousa. Formal verification in autonomic-component ensembles, WSEAS / NAUN International Conferences, Salerno, Italy, 2014
- [3] F. Emmert-Streib. Statistic Complexity: Combining Kolmogorov Complexity with an Ensemble Approach, Queen’s University, Belfast, United Kingdom, 2010
- [4] Cilibrasi R, Vitanyi P. Clustering by compression. IEEE Transactions Information Theory 51: 1523–1545. 2005
- [5] ASCENS, P.: <http://www.ascens-ist.eu/> (2010)
- [6] Rocco De Nicola, Michele Loreti, Rosario Pugliese, Francesco Tiezzi. “SCEL- a Language for Autonomic Computing”. ASCENS project, Technical report, January 2013
- [7] Unwanted traffic identification problems. Martins Ekmanis. Department of Telecommunications, Riga Technical University, Azenes iela 12, LV-1048, Riga, Latvia
- [8] G. Kolaczek, K. Juszczyszyn. Attack pattern analysis framework for multiagent intrusion detection system. International Journal of Computational Intelligence Systems, Vol.1, No. 3 (August, 2008), 215 - 224

Numerical calculation of the magnetic field produced by a multi-conductor power cable

Dumitru Toader, Iulia Cata, Constantin Blaj, and Alina Lihaciu

Abstract—The paper presents a numerical model using finite element method to calculate the magnetic field produced by current passing through multi conductor power cable, with helical turns whose thickness is comparable to the winding radius. The software package Vector Field Opera is used in order to optimize the calculation of electric parameters taking into account the radius and the helical shape of the multi conductor wires.

Keywords—finite element method, helical turn, magnetic field, magnetic field.

I. INTRODUCTION

THE magnetic field produced by coils is, usually, made considering the turns circular and filamentous. This paper presents a numerical model for the calculation of the magnetic field of multi conductor power cable.

When the thickness of the coil's conductor is close to the radius of the cylinder the turns are helical not circular. In literature [1] - [9] the analytical model used to calculate the magnetic field for helical turns neglects the conductor thickness. The magnetic field was determined for helical turn considering homogeneous linear and nonlinear medium, with the software package Vector Fields Opera [10].

II. THE FINITE ELEMENTS METHOD FOR THE CALCULATION OF THE MAGNETIC FIELD

A. The functional for the magnetic field

The variational model of the magnetic field for conductive environments at rest, nonlinear, with magnetic anisotropy, inhomogeneous and with permanent magnets is created using the magnetic vector potential. In this case the current density is as given in [11, 12]:

$$\bar{J} = \sigma \cdot \bar{E} = -\sigma \cdot \text{grad}V - \sigma \cdot \frac{\partial \bar{A}}{\partial t} = \bar{J}_a - \sigma \cdot \frac{\partial \bar{A}}{\partial t} \quad (1)$$

Where: \bar{E} - electric field strength; σ - conductivity; t - time; \bar{A} - magnetic vector potential; \bar{J}_a - imposed current density; $\sigma \cdot \frac{\partial \bar{A}}{\partial t}$ - induced current density.

The functional is, [11] - [20],

$$F = \int_V \left\{ \left[\int_0^{\bar{B}} \frac{1}{\mu(B)} (\bar{B} - \bar{B}_R) d\bar{B} \right] - \left[\int_0^{\bar{A}} (\bar{J}_a - \sigma \cdot \frac{\partial \bar{A}}{\partial t}) d\bar{A} \right] \right\} dv - \int_{\Sigma} (\bar{A} \times \bar{H}) \cdot \bar{n}_{\Sigma} d\Sigma - \int_S (\bar{J}_S \cdot \bar{A}) ds \quad (2)$$

where \bar{B} - the magnetic flux density, $\bar{B}_R = \mu_0 \cdot \bar{M}_p$ - the magnetic flux density for permanent magnets, μ_0 - permeability of free space, \bar{M}_p - the permanent magnetization, $\mu(B)$ - is the permeability tensor, Σ - is the boundary of the domain, $d\Sigma$ - is the surface element of the boundary, \bar{n}_{Σ} - is the normal unit vector of the domain boundary, $\bar{A} \times \bar{H}$ - the density of magnetic energy transferred through boundary Σ , \bar{J}_S - the current density of surface S contained in domain V , \bar{A} - magnetic vector potential.

In the case of linear media, homogeneous and without permanent magnetization (missing permanent magnets) relation (2.2) becomes,

$$F = \int_V \left\{ \left[\int_0^{\bar{B}} \frac{1}{\mu} B d\bar{B} \right] - \left[\int_0^{\bar{A}} (\bar{J}_a - \sigma \cdot \frac{\partial \bar{A}}{\partial t}) d\bar{A} \right] \right\} dv - \int_{\Sigma} (\bar{A} \times \bar{H}) \cdot \bar{n}_{\Sigma} d\Sigma - \int_S (\bar{J}_S \cdot \bar{A}) ds \quad (3)$$

B. Three-dimensional finite element

The three-dimensional finite element, at which interpolation polynomial is linear, has 4 nodes ($p=4$), so $i = \overline{1,4}$. In the coordinate system Cartesian energy functional (2.3), for linear environments without permanent magnetization, domain does not contain current density surface and Dirichlet conditions on the boundary becomes,

$$F = F_x + F_y + F_z \quad (4)$$

where

Dumitru Toader is with the Politehnica University of Timisoara, Romania,; e-mail: dumitru.toader@upt.ro.

Iulia Cata is PhD graduated at Politehnica University of Timisoara

Constantin Blaj is with the Politehnica University of Timisoara, Romania,; e-mail: constantin.blaj@upt.ro

Alina Lihaciu is PhD student at Politehnica University of Timisoara.

$$F_x = \int_V \left\{ \frac{1}{2\mu} \left[\left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \right)^2 - J_{ax} \cdot A_x + \sigma \frac{\partial A_x}{\partial t} A_x \right] \right\} dx dy dz$$

$$F_y = \int_V \left\{ \frac{1}{2\mu} \left[\left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x} \right)^2 - J_{ay} \cdot A_y + \sigma \frac{\partial A_y}{\partial t} A_y \right] \right\} dx dy dz$$

$$F_z = \int_V \left\{ \frac{1}{2\mu} \left[\left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \right)^2 - J_{az} \cdot A_z + \sigma \frac{\partial A_z}{\partial t} A_z \right] \right\} dx dy dz$$

The mathematical expression of the vector magnetic potential \vec{A} , for finite element "e" becomes,

$$\vec{A}_e = A_{ex}(x, y, z, t) \cdot \vec{i} + A_{ey}(x, y, z, t) \cdot \vec{j} + A_{ez}(x, y, z, t) \cdot \vec{k} = \sum_{i=1}^4 N_{ei}(x, y, z) \cdot \vec{A}_{ei}(x, y, z, t)$$

$$[M] \cdot [\vec{A}_i] + [C] \cdot \left[\frac{\partial \vec{A}_i}{\partial t} \right] + [\vec{F}_i] = 0 \quad (5)$$

Where $[M]$ – square matrix of linear system, $[C]$ – column matrix of current density induced coefficients, $[\vec{F}_i]$ – column matrix of imposed current density, $[\vec{A}_i]$ – column matrix of vector magnetic potential. The equation (5) allows the calculation of the vector magnetic potential in all nodes.

III. ELECTROMAGNETIC FIELD CALCULATION USING VECTOR FIELDS OPERA

In order to calculate the electromagnetic field was used the software package Opera Vector Field based on the finite element method [10]. It includes programs for solving and analyzing the electromagnetic field in plane (2D) and space (3D).

The calculation of the magnetic field created by current passing through a multi wire helical turn conductor (Fig.1), used in overhead lines [21], with cross section 35 mm² and length h=500mm is done using finite element method implemented in the program Vector Fields Opera.

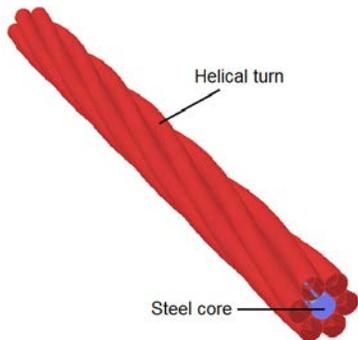


Figure. 1 The 6 wire conductor with iron core

The magnetic field created by currents passing through a multi wire helical turn conductor (Fig.1), used in overhead lines [21], with the cross section

35 mm² and length h = 500 mm is calculated using the finite element method implemented in the program Vector Field Opera.

The domain where is calculated the magnetic field is cylindrical with length $h_1 + h + h_2$ and radius b (Fig.2), where $h_1 = h_2 = 10h$. In figure 2 were not represented the helical turns, being of μ_0 permeability as the surrounding air.

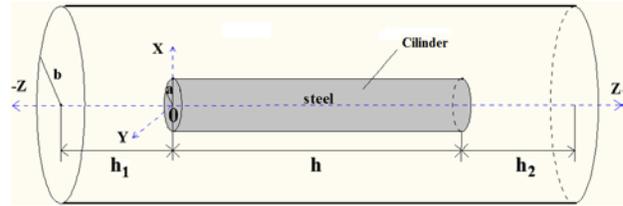


Figure 2. The domain for magnetic field analysis

The software package, using the finite element method, Vector Field Opera allows the calculation of the magnetic field in both, linear and nonlinear, environments (Fig.3). For nonlinear environment it is possible to use magnetization curves $B(H)$ (Fig.4) from the library or the curves can be introduced by the user [10], [17]-[20].

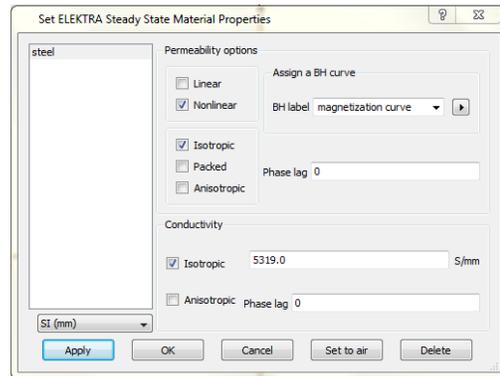


Figure 3. Set material properties

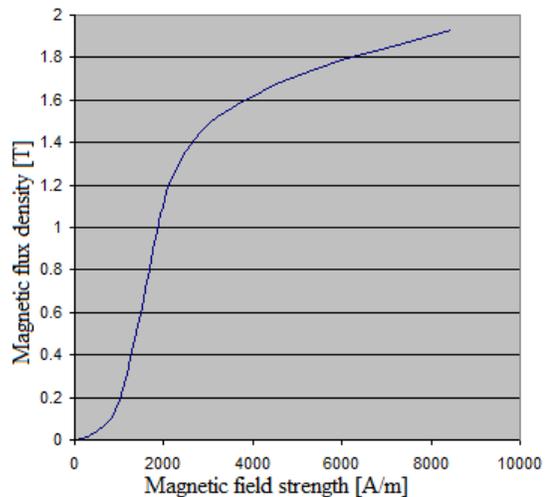


Figure 4. Magnetization curve

IV. NUMERICAL RESULTS

The 6 wire conductor, with iron core (Fig.1) produces the magnetic field and has the following geometric characteristics [22]-[24]: $a = a_1 = 1.35 \text{ mm}$, $h = 500\text{mm}$ and the magnetization curve $B(H)$ from figure 4. In the following calculations were considered effective values of the current such as 150A, 100A, 75A, 50A and 30A. The radius “a” of the steel core (cylinder) was divided into 10 segments of length $a/10$. The magnetic flux density was calculated only inside the steel cylinder, because we were interested only in the magnetic flux. The results are presented in Table 1.

$I[\text{A}]$ \ $r[\text{mm}]$	150A	100A	75A	50A	30A
0	0.737	0.409	0.180	0.047	0.021
0.135	0.896	0.479	0.217	0.068	0.032
0.27	0.862	0.451	0.196	0.060	0.030
0.405	0.869	0.453	0.196	0.061	0.030
0.54	0.873	0.454	0.196	0.061	0.030
0.675	0.876	0.455	0.196	0.061	0.030
0.81	0.879	0.455	0.197	0.061	0.030
0.945	0.883	0.456	0.197	0.061	0.030
1.08	0.887	0.457	0.197	0.061	0.030
1.215	0.892	0.458	0.197	0.061	0.030
1.35	0.873	0.448	0.192	0.058	0.029

For the calculation of the magnetic flux in the cylinder of steel, the expression is

$$\Phi = \sum_{k=0}^{10} \pi(r_{k+1}^2 - r_k^2) \cdot \frac{B_{k+1} + B_k}{2} \quad (6)$$

The results are presented in Table 2.

Table 2. Magnetic flux (calculated)

$I[\text{A}]$	150A	100A	75A	50A	30A
$\Phi[\mu\text{Wb}]$	5.011	2.601	1.121	0.352	0.172

V. EXPERIMENTS

Wiring connections scheme used for experimental determinations is presented in Fig.5.

From the measurements was determined the current through the conductor and the induced voltage into coil with 2700 turns [24], [25].

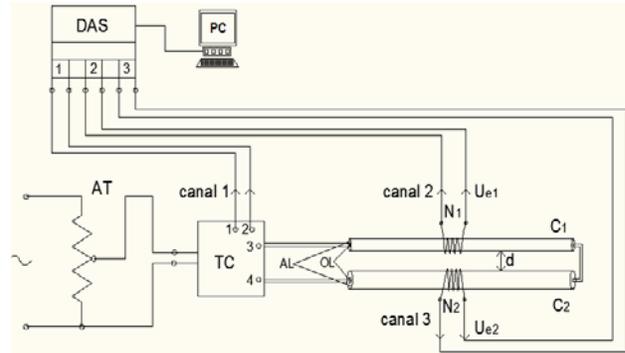


Figure 5. Measurement scheme

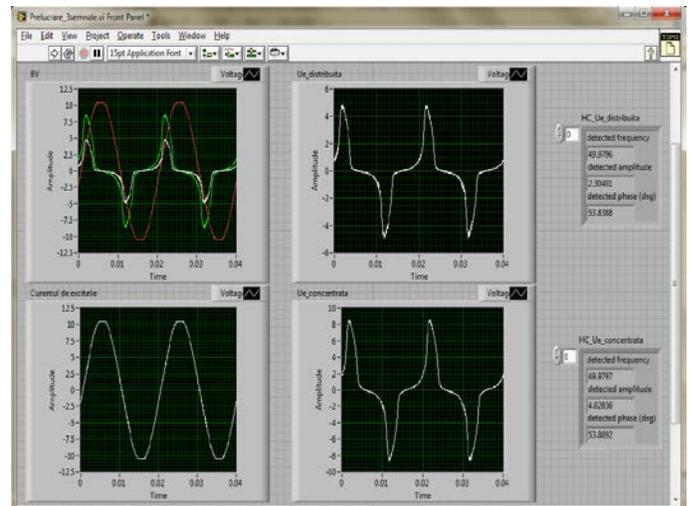


Figure 6. Measurements results

For a current of 150 A, rms value, as shown in figure 7, the voltage induced is shown in figure 8, and the corresponding magnetic flux variation is shown in figure 9. The magnetic flux was obtained by integration of the induced voltage. The results obtained for magnetic flux at different values (rms) of the current in the conductor are presented in Table 3

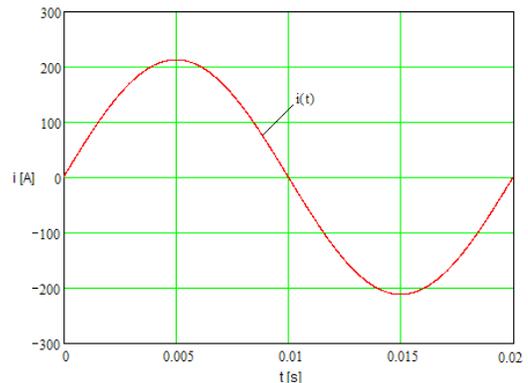


Figure 7 The current flowing in the conductor

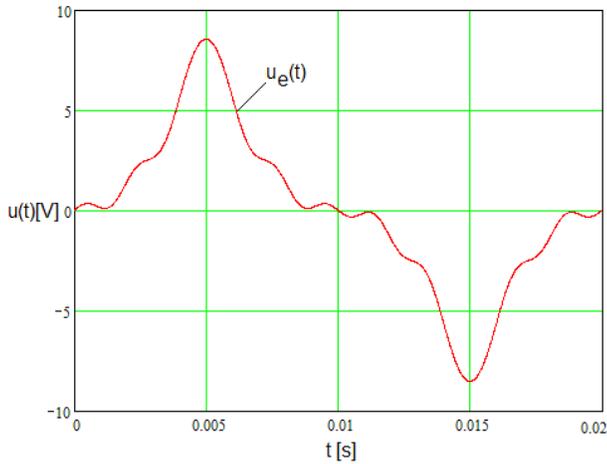


Figure 8. Momentary values of induced voltage

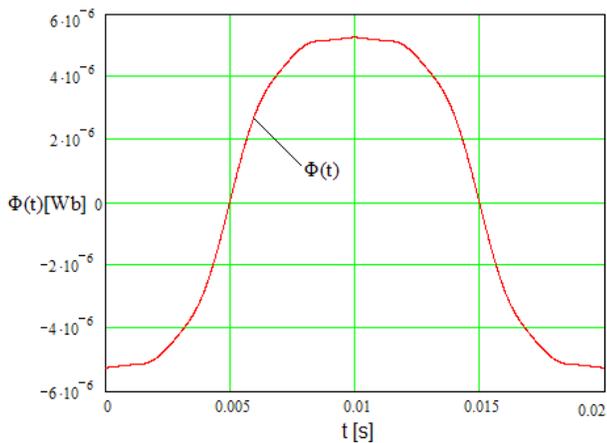


Figure 9. Momentary values of magnetic flux

Table 3. Magnetic flux (measured)

I[A]	150A	100A	75A	50A	30A
Φ[μWb]	5.011	2.601	1.121	0.352	0.172

It was made a comparative calculation of the values for magnetic field strength along the cylinder radius and in the middle of the turn (rotation angle $\tau = 180^\circ$ and $h/2$), using the two modeling programs available in Vector Fields Opera (Geometric Modeler and Pre-Processor) [29]. The Geometric Modeler works with tetraedric finite elements, and the Pre-Processor uses paralelipipedic finite elements.

The outside domain was extended, successively to a cylindrical border placed at $b = 10a$, $b = 20a$, $b = 40a$, $b = 80a$.

The results show that the radial dimension of the boundary it is not necessary to be extended to more than 10 times the cylinder radius.

Next it was considered the field calculation extended up to 10 times the cylinder radius and it was analyzed how the mesh step influences the magnetic field intensity values

For different combinations of mesh steps and length of the domain analyzed the results are shown in figure 10.

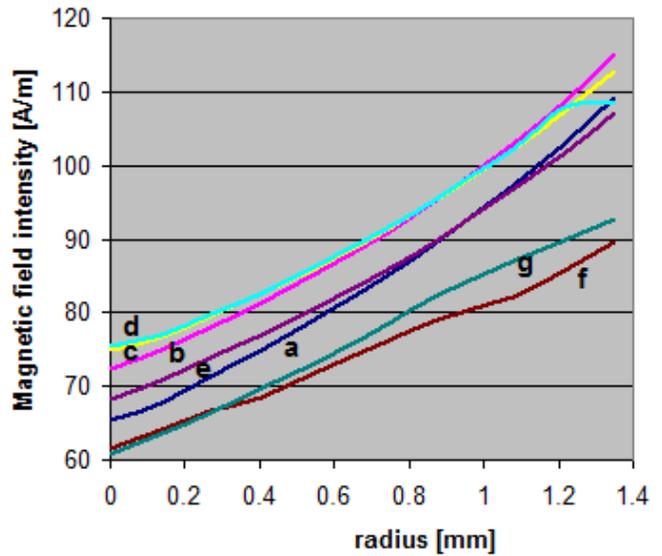


Figure 10. Magnetic field strength

It is noticed that using finite elements of different shapes result differences between magnetic field intensity values obtained in the two types of modeling. It can be concluded that variant g is the best because Geometric Modeler uses the tetrahedral-type finite elements, so modifying the finite element size it changes the three spatial variables (r, z, τ) simultaneously. The variant g is based on the size of finite element (introduced in Geometric Modeler) for the outer domain being twice the radius of the one used inside the steel cylinder.

Further it is analyzed how the length of the cable, h , taken in the modelization, affects the magnetic field intensity values, computed in the mediator plane. In figure 11 are represented the values computed for the length of the cable corresponding to 1 turn, 3 turns and 5 turns, where the notations.

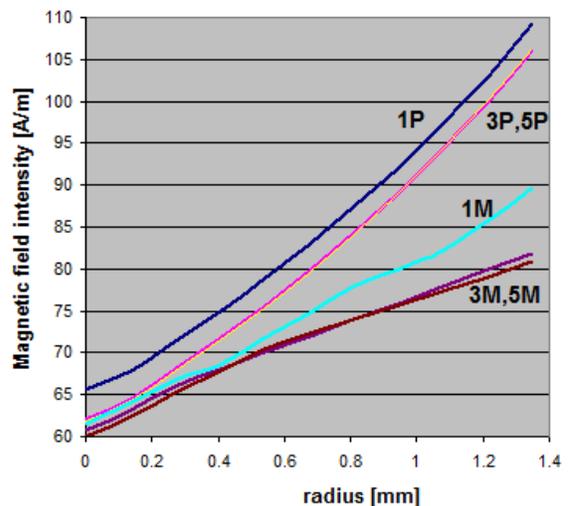


Figure 11.

have the following meanings: 1M, 3M and 5M represent 1 turn, 3 turns and 5 turns obtained with modeling program Geometric Modeler respectively 1P, 3P and 5P represent 1 turn, 3 turns and 5 turns obtained with modeling program Pre-processor.

From Fig. 11 results that for lengths corresponding to more than 3 turns, the computed magnetic field values in the mid-plane, do not change significantly. The analysis of a long conductor, with a large number of turns, it can be made considering a minimum length corresponding to at least 3 turns. In our calculations the length h , in the modelization, was taken of about 5 times the average pitch of the helix. The cables which were analyzed have the width of one complete helix turn (pitch length) of the following values: 81mm, 97.2mm and 113.4mm.

VI. CONCLUSIONS

From the study results the following conclusions:

- Numerical model that uses finite elements form a tetrahedron leads to the best results but require the longest calculation time (about two hours and variant a only 1 minute and the 24 seconds);

- A finer mesh step along the radius ($a/20$ - variant b and e, compared to $a/10$ - variant c and d) leads to results close to those obtained for variant g;

- The numerically analyzed and implemented in the software package Vector Field Opera allows precise calculation of the magnetic field, but this requires appropriate choice of field sizes considered, the mesh step, namely the finite element size and border conditions.

- Numerical model based on finite element software package and implement in the Vector Field Opera can be used with good results in calculating the magnetic field in both linear and nonlinear environments, if properly chosen finite element size and domain.

REFERENCES

- [1] Lucca G., *Integral equations methods for the analysis and design of ELF conductive and magnetic shields*, European Transactions on Electrical Power, Vol. 20, n0. 3, 2010, p. 335-353.
- [2] Yinshun W., Yibo Z., Hongwei L., *A novel approach for design DC HTS cables*, IEEE Transactions on Applied Superconductivity, vol. 21, no. 3, 2011, p. 1042-1045.
- [3] Moro, F.; Turri, R., *Fast Analytical Computation of Power-Line Magnetic Fields by Complex Vector Method*, Power Delivery, IEEE Transactions on, vol. 23, no. 2, 2008, p. 1042-1048.
- [4] Siqun Zhang; Chang Yin; Haojun Xu, *The series calculations model for the magnetic field of coil above arbitrary layer conductive plates*, Mechatronics and Automation (ICMA), 2014 IEEE International Conference on, Tianjin, China, 2014, p. 475-480.
- [5] Keikko T., Isokorpi J., Reivonen S., *Magnetic field measurements and calculations with 20 kV underground power cables*, Proceedings 9th International Conference on Computational Methods and Experimental Measurements, Sorrento, Italy, 1999, p. 27-36..
- [6] Căta, D. Păunescu, D. Toader, *Calculation of magnetic field intensity vector in the axis of a wire cable with helically wound wires*, Scientific Bulletin of the Politehnica, University of Timisoara, Romania, Transactions on Mathematics and Physics, Tom 56(70), Fascicule 2, 2010, p.73-85X2.
- [7] T. Tominaka, *Calculations using the helical filamentary structure for current distributions of a six around one superconducting strand cable and a multifilamentary composite*, J. Appl. Phys. 96 5069–80, 2004.
- [8] T. Tominaka, *Analytical field calculations for various helical conductors*, IEEE Trans. Appl. Supercond. 14 1838–41, 2004.
- [9] Xiao-Bang Xu; Guanghao Liu, *Formulation of a computational model for determining the magnetic field produced by an in-service underground pipe-type cable*, Southeast Conference, 2002. Proceedings IEEE, p. 99-103
- [10] Vector Field Opera 13.0 User Guide..
- [11] Şora C., De Sabata I., Bogoevici N., Heler A., Daba D., Vetreş I., ş.a., *The Foundations of Electrotechnics*, Publisher Polytechnic Timisoara 2008.
- [12] Mîndru G., Rădulescu M., *Numerical analysis of electromagnetic field*, Lithography Institutului Polytechnic Cluj Napoca, 1983.
- [13] Cata I., Toader D., *Finite Element Method for Calculation of Magnetic Field Produced from Helical Turn in Linear and Nonlinear Medium*, Proceedings Mathematical Models and Methods in Modern Science, 2011, p. 100-106.
- [14] J. Donea, A. Huerta, *Finite element methods for flow problems*, John Wiley & Sons, Chichester, 2003.
- [15] E. Madenci, I. Guven, *The finite element method and application in engineering using Ansys*, Springer Science Business media LLC, ISBN 0-387-28289-0, New York, S.U.A., 2006.
- [16] O. C. Zienkiewicz, R. L. Taylor, *The Finite Element Method*, Fifth Edition, Butterworth – Heinemann, ISBN 978-0-7506-5049-6, Oxford, UK, 2000.
- [17] Xiao-Bang Xu; Guanghao Liu, *A finite-element analysis of electromagnetic field produced by ELF sources enclosed by a nonlinear ferromagnetic pipe*, Antennas, Propagation and EM Theory, 2000. Proceedings. ISAPE 2000. 5th International Symposium on, p. 215-218.
- [18] T. Belytschko, W.K. Liu, B. Moran, *Nonlinear finite element for continua and structures*, John Wiley & Sons, 2000.
- [19] T. St. Mănescu, N. L. Zaharia, D. S. Avram, N. M. Ligănasu, M. D. Stroia, *Nonlinear Finite Element Analysis Used at Tank Wagons*, Proceedings of the 4th WSEAS International Conference on Finite Differences - Finite Elements - Finite Volumes - Boundary Elements, Paris, France, April 28-30, 2011, pp. 110-113
- [20] C. Erdönmez1, C. E. Dmrak, *Numerical model for an IWRC bending over sheave problem and its finite element solution*, Proceedings of the International Conference on Applied, Numerical and Computational Mathematics (ICANCM'11), Barcelona, Spain, September 15-17, 2011, pp.199-205.
- [21] IPROEB Bistrita, Overhead bare conductors, 2007.
- [22] Căta I., Toader D., *Power losses computation in steel core of multiple conductors wrapped helically*, Buletin Stiintific al UPT, Seria Matematica-Fizică, Ed. Politehnica, Timisoara, România, Tom 57(71), Fascicula 1, 2012, p.98-109.
- [23] Toader D., Cata I., *Computation of Magnetic Flux in a Helical Multiple Conductors with Finite Element Method*, Proceedings Mathematical Models and Methods in Modern Science, 2011, p. 251-256.
- [24] *** LabView. User's Guide.
- [25] *** International Standard IEC 60404-6 ed 2.0. Magnetic materials – Part 6: Methods of measurement of the magnetic properties of magnetically soft metallic and powder materials at frequencies in the range 20 Hz to 200 kHz by the use of ring specimens. Dec 2008.

Quasi-conformal Harmonic Mappings Related To The Janowski Starlike Functions

Melike Aydoğan¹, Yaşar Polatoğlu², H. Esra Özkan Uçar³, Arzu Yemişçi⁴ and Yasemin Kahramaner⁵

Abstract—Let $f(z) = h(z) + \overline{g(z)}$ be a univalent sense-preserving harmonic mapping in the open unit disc $\mathbb{D} = \{z|z| < 1\}$. If $f(z)$ satisfies the condition $|w(z)| = \left| \frac{g'(z)}{h'(z)} \right| < k$, ($0 \leq k < 1$), then $f(z)$ is called k -quasiconformal harmonic mapping in \mathbb{D} [6]. The class of such mappings is denoted by $S_{H(k)}$.

The aim of this paper is to give some properties of the solution of non-linear partial differential equation $\overline{f_z} = w(z)f(z)$ under the condition $|w(z)| < k$, $w(z) \prec \frac{k^2(b_1-z)}{k^2-b_1z}$, $h(z) \in S^*(A, B)$, where $S^*(A, B)$ is the class of Janowski starlike functions. The proofs of this paper are based on the idea Robinson [7].

Keywords— k -quasiconformal harmonic mapping, distortion theorem, growth theorem.

I. INTRODUCTION

LET Ω be the family of functions $\phi(z)$ regular in the disc \mathbb{D} and satisfying the conditions $\phi(0) = 0$, $|\phi(z)| < 1$ for all $z \in \mathbb{D}$.

Next, for arbitrary fixed real numbers A, B , $-1 \leq B < A \leq 1$, we denote by $P(A, B)$ the family of functions $p(z) = 1 + p_1z + p_2z^2 + p_3z^3 + \dots$ regular in \mathbb{D} and such that $p(z)$ is in $P(A, B)$ if and only if

$$p(z) = \frac{1 + A\phi(z)}{1 + B\phi(z)} \tag{1}$$

for some $\phi(z) \in \Omega$ and every $z \in \mathbb{D}$.

Moreover, let $S^*(A, B)$ denote the family of functions $s(z) = z + c_2z^2 + c_3z^3 + \dots$ regular in \mathbb{D} and such that $s(z)$ is in $S^*(A, B)$ if and only if

$$z \frac{s'(z)}{s(z)} = p(z) \tag{2}$$

for some $p(z)$ is in $P(A, B)$ and all $z \in \mathbb{D}$ [5]. Let $s_1(z) = z + d_2z^2 + \dots$ and $s_2(z) = z + e_2z^2 + \dots$ be analytic functions in the open unit disc in \mathbb{D} . If there exists a function $\phi(z) \in \Omega$ such that $s_1(z) = s_2(\phi(z))$ for all $z \in \mathbb{D}$, then we say that $s_1(z)$ is subordinate to $s_2(z)$ and we write $s_1(z) \prec s_2(z)$ if and only if $s_1(\mathbb{D}) \subset s_2(\mathbb{D})$ and $s_1(0) = s_2(0)$ implies $s_1(\mathbb{D}_r) \subset s_2(\mathbb{D}_r)$, where $\mathbb{D}_r = \{z|z| < r, 0 < r < 1\}$. (Subordination and Lindelof principle [1], [3]).

Finally, a planar harmonic mapping in the open unit

disc \mathbb{D} is a complex-valued harmonic function f , which maps \mathbb{D} onto the some planar domain $f(\mathbb{D})$. Since \mathbb{D} is a simply-connected domain the mapping f has a canonical decomposition $f(z) = h(z) + \overline{g(z)}$, where $h(z)$ and $g(z)$ are analytic in \mathbb{D} and have the following power series expansions

$$h(z) = \sum_{n=0}^{\infty} a_n z^n, \quad g(z) = \sum_{n=0}^{\infty} b_n z^n \tag{3}$$

where $a_n, b_n \in C$, $n = 0, 1, 2, 3, \dots$ as usual we call $h(z)$ the analytic part of $f(z)$ and $g(z)$ is co-analytic part of $f(z)$. An elegant and complete account of the theory of harmonic mappings is given in Duren's monograph [2] proved in 1936 that the harmonic function $f(z)$ is locally univalent in \mathbb{D} if and only if its Jacobian

$$J_f = |h'(z)|^2 - |g'(z)|^2 \tag{4}$$

is different from zero in \mathbb{D} . In view of this result, locally univalent harmonic mappings in the open unit disc \mathbb{D} are either sense-reserving if $|g'(z)| > |h'(z)|$ in \mathbb{D} or sense-preserving if $|g'(z)| < |h'(z)|$ in \mathbb{D} .

Throughout this paper we will restrict ourselves to the study of sense-preserving harmonic mappings. We will also note that $f(z) = h(z) + \overline{g(z)}$ is sense-preserving in \mathbb{D} if and only if $h'(z)$ doesn't vanish in \mathbb{D} and the second dilatation $w(z) = \frac{g'(z)}{h'(z)}$ has the property $|w(z)| < 1$ for all $z \in \mathbb{D}$. Therefore, the class of all sense-preserving harmonic mappings in the open unit disc with $a_0 = b_0 = 0$ and $a_1 = 1$ will be denoted by S_H . Thus S_H contains standard class S of univalent functions. The family of all mappings $f \in S_H$ with the additional property $g'(0) = 0$, i.e. $b_1 = 0$ is denoted by S_H^0 . Hence it is clear that $S \subset S_H^0 \subset S_H$. For the aim of this paper we need the following lemma and theorem.

Lemma 1.1 ([4]) Let $\phi(z)$ be a non-constant and analytic function in the unit disc \mathbb{D} with $\phi(0) = 0$. If $|\phi(z)|$ attains its maximum value on the circle $|z| = r$ at the point z_0 , then $z_0\phi'(z_0) = m\phi(z_0)$, $m \geq 1$.

Theorem 1.2 ([5]) If $s(z) \in S^*(A, B)$, then for $|z| = r$, $0 < r < 1$

$$F(r, -A, -B) \leq |s(z)| \leq F(r, A, B) \tag{5}$$

$$F(r, A, B) = \begin{cases} r(1 + Br)^{\frac{A-B}{B}} & \text{for } B \neq 0, \\ re^{Ar} & \text{for } B = 0. \end{cases} \tag{6}$$

M. Aydoğan is with the Department of Mathematics, Işık University, Meşrutiyet Köyü, Campus of Şile, Istanbul, Turkey, e-mail: melike.aydogan@isikun.edu.tr.

H.E. Özkan, A. Yemişçi and Y. Polatoğlu are with Istanbul Kültür University.

Y. Kahramaner is with Istanbul Commerce University.

These bounds are sharp, being attained at the point $z = re^{i\theta}$, $0 \leq \theta \leq 2\pi$ by

$$s(z) = \begin{cases} z(1 + Be^{-i\theta z})^{\frac{A-B}{B}} & \text{for } B \neq 0, \\ ze^{Ae^{-i\theta z}} & \text{for } B = 0. \end{cases} \quad (7)$$

II. MAIN RESULTS

Theorem 2.1 Let $f(z) = h(z) + \overline{g(z)}$ be an element of S_H and $h(z)$ be an element of $S^*(A, B)$, then the solution of the differential equation $\overline{f_z} = w(z)f_z$ under the conditions $|w(z)| < k$, $0 \leq k < 1$ and $w(z) \prec \frac{k^2(b_1 - z)}{k^2 - \overline{b_1}z}$ is $\frac{g(z)}{h(z)} = \frac{k^2(b_1 - \phi(z))}{k^2 - \overline{b_1}\phi(z)}$, where $\phi(z) \in \Omega$.

Proof. We consider the linear transformation $w(z) = \frac{k^2(b_1 - z)}{k^2 - \overline{b_1}z}$ this transformation maps $|z| < k$ onto itself, i.e.

$$\Delta = \{w||w| < k\} = w(\mathbb{D}) = \{z||z| < k\} \quad (8)$$

On the other hand

$$w(z) = \frac{g'(z)}{h'(z)} = \frac{(b_1z + b_2z^2 + \dots)'}{(z + a_2z^2 + \dots)'} = \frac{b_1 + 2b_2z + \dots}{1 + 2a_2z + \dots}$$

$$\Rightarrow w(0) = b_1$$

Therefore the function

$$\phi(z) = \frac{k^2(b_1 - w(z))}{k^2 - \overline{b_1}w(z)}$$

satisfies the conditions of Schwarz lemma then we have

$$w(z) = \frac{g'(z)}{h'(z)} \prec \frac{k^2(b_1 - z)}{k^2 - \overline{b_1}z} \quad (9)$$

and the transformation $w(z) = \frac{k^2(b_1 - z)}{k^2 - \overline{b_1}z}$ maps $|z| = r$ onto the disc with the centre

$$C(r) = \left(\frac{k^2(1 - r^2)Reb_1}{k^2 - |b_1|^2r^2}, \frac{k^2(1 - r^2)Imb_1}{k^2 - |b_1|^2r^2} \right)$$

and the radius $\rho(r) = \frac{k(k^2 - |b_1|^2)}{k^2 - |b_1|^2r^2}$. Using the subordination principle, then we can write

$$w(\mathbb{D}_r) = \left\{ \frac{g'(z)}{h'(z)} \mid \left| w(z) - \frac{k^2(1 - r^2)b_1}{k^2 - |b_1|^2r^2} \right| \leq \frac{k(k^2 - |b_1|^2)r}{k^2 - |b_1|^2r^2} \right\} \quad (10)$$

Now we define the function $\phi(z)$ by

$$\frac{g(z)}{h(z)} = \frac{k^2(b_1 - \phi(z))}{k^2 - \overline{b_1}\phi(z)}, \quad (11)$$

then $\phi(z)$ is analytic and $\phi(0) = 0$. If we take the derivative from the (10) and after brief calculations we get

$$w(z) = \frac{g'(z)}{h'(z)} = \frac{k^2(b_1 - \phi(z))}{k^2 - \overline{b_1}\phi(z)} + \frac{k^2(|b_1|^2 + k^2 - 2b_1\phi(z))z\phi'(z)}{(k^2 - \overline{b_1}\phi(z))^2} \frac{1 - \phi(z)}{1 + \phi(z)} \quad (12)$$

Now it is easy to realize that the subordination

$$\frac{g(z)}{h(z)} = \frac{k^2(b_1 - \phi(z))}{k^2 - \overline{b_1}\phi(z)} \Leftrightarrow \frac{g(z)}{h(z)} \prec \frac{k^2(b_1 - z)}{k^2 - \overline{b_1}z} \quad (13)$$

is equivalent to $|\phi(z)| < 1$ for all $z \in \mathbb{D}$. Indeed we assume the contrary; then there is a $z_0 \in \mathbb{D}_r$ such that $|\phi(z_0)| = 1$. So by I.S.Jack's lemma (Lemma 1.1) $z_0\phi'(z_0) = m\phi(z_0)$, $m \geq 1$ and for such z_0 we have

$$w(z_0) = \frac{g'(z_0)}{h'(z_0)} = \frac{k^2(b_1 - \phi(z_0))}{k^2 - \overline{b_1}\phi(z_0)} + \frac{k^2(|b_1|^2 + k^2 - 2b_1\phi(z_0))m\phi(z_0)}{(k^2 - \overline{b_1}\phi(z_0))^2} \frac{1 - \phi(z_0)}{1 + \phi(z_0)} \notin w(\mathbb{D}_r)$$

but this contradicts with (9). So our assumption is wrong, i.e., $|\phi(z)| < 1$ for every $z \in \mathbb{D}$.

Corollary 2.2 Let $\frac{g(z)}{h(z)} = \frac{k^2(b_1 - \phi(z))}{k^2 - \overline{b_1}\phi(z)}$ be the solution of the non-linear partial differential equation $\overline{f_z} = w(z)f_z$ under the condition $|w(z)| = \left| \frac{g'(z)}{h'(z)} \right| < k$ ($0 \leq k < 1$),

$w(z) \prec \frac{k^2(b_1 - z)}{k^2 - \overline{b_1}z}$ and $h(z) \in S^*(A, B)$, then

$$\begin{cases} rG(A, B, k, |b_1|, -r) \leq |g(z)| \leq rG(A, B, k, |b_1|, r), \\ rH(A, k, |b_1|, -r) \leq |g(z)| \leq rG(A, B, k, |b_1|, r) \end{cases} \quad (14)$$

where

$$G(A, B, k, |b_1|, r) = \frac{k(|b_1| + kr)}{k + |b_1|r} (1 + Br)^{\frac{A-B}{B}}$$

$$H(A, B, k, |b_1|, r) = \frac{k(|b_1| + kr)}{k + |b_1|r} e^{Ar}$$

Proof. Using Theorem 2.1 we can write

$$F(k, |b_1|, -r) \leq \left| \frac{g(z)}{h(z)} \right| \leq F(k, |b_1|, r)$$

$$F(k, |b_1|, r) = \frac{k(|b_1| + kr)}{k + |b_1|r} \quad (15)$$

then we have

$$F(k, |b_1|, -r)|h(z)| \leq |g(z)| \leq F(k, |b_1|, r)|h(z)| \quad (16)$$

Using Theorem 1.2 in the inequality (15) we get (13).

Corollary 2.3 Let $f(z) = h(z) + \overline{g(z)}$ be an element of S_H . If $f(z)$ satisfies the non-linear partial differential equation $\overline{f_z} = w(z)f_z$ under the condition $|w(z)| = \left| \frac{\overline{f_z}}{f_z} \right| < k$,

($0 \leq k < 1$), $w(z) \prec \frac{k^2(b_1 - z)}{k^2 - \overline{b_1}z}$ and $h(z) \in S^*(A, B)$ then

$$\begin{cases} G_1(A, B, k, |b_1|, -r) \leq |g'(z)| \leq G_1(A, B, k, |b_1|, r), B \neq 0 \\ G_2(A, k, |b_1|, -r) \leq |g'(z)| \leq G_2(A, k, |b_1|, r), B = 0 \end{cases} \quad (17)$$

where

$$G_1(A, B, k, |b_1|, r) = \frac{k(|b_1| + kr)}{k + |b_1|r} (1 + Ar)(1 + Br)^{\frac{A-2B}{B}}$$

$$G_2(A, k, |b_1|, r) = \frac{k(|b_1| + kr)}{k + |b_1|r} (1 + Ar)e^{Ar}$$

Proof. Since $h(z) \in S^*(A, B)$, then we have

$$\left| z \frac{h'(z)}{h(z)} - \frac{1 - AB r^2}{1 - B r^2} \right| \leq \frac{(A - B)r}{1 - B^2 r^2}$$

$$\Rightarrow \frac{1 - Ar}{1 - Br} \leq \left| z \frac{h'(z)}{h(z)} \right| \leq \frac{1 + Ar}{1 + Br}$$

$$\Rightarrow \begin{cases} F_1(A, B, -r) \leq |h'(z)| \leq F_1(A, B, r), B \neq 0 \\ F_2(A, -r) \leq |h'(z)| \leq F_2(A, r), B = 0 \end{cases} \quad (18)$$

where

$$F_1(A, B, r) = (1 + Ar)(1 + Br)^{\frac{A-2B}{B}}, \quad B \neq 0$$

$$F_2(A, r) = (1 + Ar)e^{Ar}, \quad B = 0$$

On the other hand since $\frac{g'(z)}{h'(z)} \prec \frac{k^2(b_1 - z)}{k^2 - \bar{b}_1 z}$ then we have (using subordination principle)

$$\begin{cases} F(k, b_1, -r)F_1(A, B, -r) \leq |g'(z)| \\ \leq F(k, b_1, r)F_1(A, B, r), B \neq 0 \\ F(k, b_1, -r)F_2(A, -r) \leq |g'(z)| \\ \leq F(k, b_1, r)F_2(A, r), B = 0 \end{cases} \quad (19)$$

where $F(k, |b_1|, r)$ is given in (15). Therefore we have (16).

Corollary 2.4 Let $f(z) = h(z) + \overline{g(z)}$ be an element of S_H . If $f(z)$ satisfies the non-linear partial differential equation $\overline{f_z} = w(z)f_z$ under the condition $|w(z)| = \left| \frac{\overline{f_z}}{f_z} \right| < k$,

$(0 \leq k < 1)$, $w(z) \prec \frac{k^2(b_1 - z)}{k^2 - \bar{b}_1 z}$ and $h(z) \in S^*(A, B)$ then

$$\begin{cases} (1 - Ar)(1 - Br)^{\frac{2(A-2B)}{B}} F_2(k, |b_1|, r) \leq J_f \\ \leq (1 + Ar)^2(1 + Br)^{\frac{2(A-2B)}{B}} F_1(k, |b_1|, r), \quad B \neq 0 \\ (1 - Ar)^2 e^{-2Ar} F_2(k, |b_1|, r) \leq J_f \\ \leq (1 + Ar)^2 e^{2Ar} F_1(k, |b_1|, r), \quad B = 0 \end{cases} \quad (20)$$

Proof. Using Theorem 2.1 we can write

$$F_2(k, |b_1|, r) \leq (1 - |w(z)|^2) \leq F_1(k, |b_1|, r),$$

where

$$\begin{cases} F_1(k, |b_1|, r) = \frac{[(k+k|b_1|) - (|b_1|+k^2)r][(k-k|b_1|) - (|b_1|-k^2)r]}{(k-|b_1|r)^2} \\ F_2(k, |b_1|, r) = \frac{[(k+k|b_1|) + (|b_1|+k^2)r][(k-k|b_1|) - (|b_1|-k^2)r]}{(k+|b_1|r)^2} \end{cases} \quad (21)$$

On the other hand we have

$$J_f = |h'(z)|^2 - |g'(z)|^2 = |h'(z)|^2(1 - |w(z)|^2) \Rightarrow$$

$$|h'(z)|^2(1 - |w(z)|^2)J_f \leq |h'(z)|^2 F(k, |b_1|, r) \quad (22)$$

If we use Theorem 1.2 in the inequality (21) we get (19).

Corollary 2.5 Let $\frac{g(z)}{h(z)} = \frac{k^2(b_1 - \phi(z))}{k^2 + \bar{b}_1 \phi(z)}$, $\phi(z) \in \Omega$ be the solution of the non-linear partial differential equation $\overline{f_z} = w(z)f_z$ under the condition $f(z) = h(z) + g(z) \in S_H$, $|w(z)| < k$, $w(z) \prec \frac{k^2(b_1 - z)}{k^2 + \bar{b}_1 z}$, $0 \leq k < 1$, $h(z) \in S^*(A, B)$ then,

$$\begin{cases} \int_0^r F(A, B, -\rho) \frac{(k - k|b_1|) + (|b_1| - k^2)\rho}{k + |b_1|\rho} d\rho \leq |f| \\ \leq \int_0^r F(A, B, \rho) \frac{(k + k|b_1|) + (|b_1| + k^2)\rho}{k + |b_1|\rho} d\rho, B \neq 0 \\ \int_0^r (1 - A\rho)e^{-A\rho} \frac{(k - k|b_1|) + (|b_1| - k^2)\rho}{k + |b_1|\rho} d\rho \leq |f| \\ \leq \int_0^r (1 + A\rho)e^{A\rho} \frac{(k + k|b_1|) + (|b_1| + k^2)\rho}{k + |b_1|\rho} d\rho, B = 0 \end{cases} \quad (23)$$

Proof. Using Theorem 2.1 we obtain

$$\frac{(k|b_1| + k) - (|b_1| + k^2)r}{k - |b_1|r} \leq (1 + |w(z)|)$$

$$\leq \frac{(k|b_1| + k) + (|b_1| + k^2)r}{k + |b_1|r} \quad (24)$$

$$\frac{(k - k|b_1|) + (|b_1| - k^2)r}{k + |b_1|r} \leq (1 - |w(z)|)$$

$$\leq \frac{(k - k|b_1|) - (|b_1| - k^2)r}{k - |b_1|r} \quad (25)$$

On the other hand we have

$$(|h'(z)| - |g'(z)|)|dz| \leq d|f| \leq (|h'(z)| + |g'(z)|)|dz| \Rightarrow$$

$$|h'(z)|(1 - |w(z)|^2)|dz| \leq d|f| \leq |h'(z)|(1 - |w(z)|)|dz| \quad (26)$$

Using (23), (24) and (25) and integrating we get (22).

III. ACKNOWLEDGEMENTS

The work presented here was partially supported by Isik University Scientific Research Funding Agency under Grant Number: BAP-14B102.

REFERENCES

- [1] Duren P., Univalent Functions, Springer Verlag, (1983).
- [2] Duren P., Harmonic Mappings in the plane, vol. 156 of Cambridge Tracts in Mathematics, Cambridge University press Cambridge U.K. (2004).
- [3] Goodman A. W., Univalent Functions Vol I and Vol II. Mariner Publishing, Tampa Florida, (1983).
- [4] Jack I. S. , Functions starlike and convex of order alpha, J. Lond. Math. Soc. (2), 3, 469-474, (1971).
- [5] Janowski W. , Some extremal problems for certain families of analytic functions, I. Ann. Polon. Math 28, 297-326, (1973).
- [6] Kalaj D. , Quasiconformal Harmonic Mappings and Close-to-Convex Domains, Filomat, 24, 1, 63-68, (2010).
- [7] Robertson R. M. , Univalent Majorants, Trans. Amer. Math. Soc. 61, 1-35, (1947).

EH-WSNs Optimizing Technique

Vladimir Shakhov

Institute of Computational Mathematics and Mathematical Geophysics
Siberian Branch of Russian Academy of Sciences
Novosibirsk, Russia
e-mail: shakhov@rav.sccc.ru

Abstract—Nowadays, networked wireless sensors can outnumber traditional electronic appliances. They will enable a plethora of new applications in industrial automation, asset management, environmental monitoring, medical and transportation business, and in a variety of safety and security scenarios. Wireless sensor networks have got a wide range of important and vital applications. Performance of wireless sensor becomes one of the most important issues. The goal of this work is to provide an approach for the efficient estimation of network throughput. The node performance model is offered and discussed. Next, a wireless sensor network is modeled as undirected probabilistic graph. We consider the networks which carry on work acceptably even if some amount of nodes fails. We define the throughput of network as the probability that sink nodes are connected and can collect data from other nodes. The corresponding calculating method is obtained.

Keywords- wireless sensor networks, network reliability, random graph connectivity

I. INTRODUCTION

The applications of wireless sensor networks (WSNs) have evolved over the last years from a stage where these networks were designed in a technology-dependent manner to a stage where some broad conceptual understanding results now exist. An essential progress in electro-mechanical and digital electronics technologies leads to development of low-power, low-cost, multifunctional sensors, which are attractive for customers.

Small sensors, which consist of sensing, data processing, and communicating modules, are combined by wireless channels into wireless sensor networks [1]. These networks are intended to be context aware, self-governing, flexible and reliable. WSNs have a wide range of potential applications, including industrial process monitoring, health care [2], military surveillance, agricultural monitoring [3], fire detection [4], smart home [5] etc.

Nowadays, low-power sensor network performance increasing is an important issue. However, this theme had not been considered in previous works. Thus, it needs to offer the corresponding mathematical and simulation models. It helps us to develop practical Energy Harvesting WSNs (EH-WSNs).

Probabilistic graph models have been used extensively in the literature for studying network reliability problems,

especially in the case of unreliable edges [6,7]. In this paper we offer a novel concept of ad hoc network reliability, which has not been discussed in the previous works. We consider ad hoc networks with imperfect nodes and perfectly reliable links. Nodes unavailability can be caused by scuffing or intrusions. An operational probability is associated with every node. It is assumed that the node failures are statistically independent. At the same time, if any two operable nodes are within a communication range then the nodes communicate with each other without any losses.

The rest of this short report is organized as follows. In section 2 the basic notations are presented. Sections 3 describes the approach of reliability calculation. In Section 3 the mathematical models for estimating of WSNs node performance are considered as well. Finally, we conclude the paper in Section 4.

II. SYSTEM MODEL

We model the ad hoc network by an undirected probabilistic graph $G = (V;E)$ whose vertices represent the nodes and whose edges represent the links. We assume that each node succeeds or fails independently with an associated probability. Further on we refer to this probability as node reliability. We suppose that the links are perfectly reliable. We use following notations for the number of network elements: $|V| = N$, $|E| = M$.

Let us define EH-WSNs reliability as the probability of EH-WSNs structure connectivity. For structural optimization of a network the reliability polynomials are used. This polynomial shows dependency of a reliability index on reliabilities of a graph components..

A sensor node randomly comes to a restoration mode and returns to a working mode. Therefore, a node is randomly available. Let p be the probability of the node reliability (availability). In the considered case, if p increases then the transmission range is reduced and the number of links between sensor nodes decreases. And vice-versa, if the sensors transmission rage is reduced then the energy harvesting period can be reduced and p is in

Assume, that an asynchronous MAC protocol is used in the system. Our motivation is as follows. The synchronous protocols require time synchronization, which causes control message overhead and makes sensor nodes more complex and expensive. Hence, the system reliability is

degraded. In asynchronous duty cycle MAC protocols, each sensor node wakes up and sleeps independently. Thus, time synchronization is not necessary.

Most asynchronous duty cycle MAC protocols adopt a random wake-up interval in order to avoid repeated collisions. Given that sensor nodes wake up at different times with random wakeup intervals, it is necessary to ensure that a sender and its intended receiver are active at the same time period to transmit data. To do this, preamble-based protocols were proposed as B-MAC, Wise-MAC, Wise-MAC and X-MAC. PW-MAC and TA-MAC are asynchronous MAC protocol based on asynchronous duty cycling. Ones minimize sensor energy consumption by enabling senders to predict receiver wakeup times.

We describe EH-WSNs topology by random graph (random vertices). Let us make the following designations.

$G(n; m; p_1, \dots, p_n)$ is a non-oriented graph with m edges (all links in WSNs) and n nodes (the number of sensors in WSNs);

V and E are the set of nodes and edges of graph G correspondingly;

p_i is the reliability of the vertex v_i ; if all nodes are homogeneous then the designation p is used.

$R(G)$ is the reliability polynomial for the graph G , in other words it is the connectivity probability for G .

For $R(G)$ calculation the factoring method can be used. It is described in the paper [8].

Let us designate the sensor working time as T and the total sensor restoration time as S . Thus, the duration of sensor active state is $T - S$. Assume that a sensor is randomly switched from the active state to the restoration state and inversely.

III. RESULTS

The Generally, a model of the sensor's behavior is based on Continuous Time Discrete States Markov process with an absorbing state. It is assumed that all nodes of WSNs are unreliable. In this paper we do not consider any effects related to the sensor repairing. A node of WSNs can get the stages as follows. Attack stage (A) – a sensor is active and transmits packets. Next, the stage (D) – a battery exhausting is detected and a energy harvesting mechanism is activated. The detection technique is generally based on change-point detection methods. OFF stage – the sensor is failed. For example, in this case the sensor battery is fully exhausted, the node buffer is overfilled etc. If a sensor is in energy harvesting mode then an intensive traffic can lead to sensor failure. The corresponding state diagram is shown on Fig.1.

The states diagram is described below. Let the intensity of the traffic be λ . The intensity of battery exhausting is d . The intensity of battery restoration is μ . Let us make the following designation: p_A, p_D, p_{OFF} are the probabilities of active state, energy harvesting state and failure state correspondingly. The index of probability corresponds to the state designation.

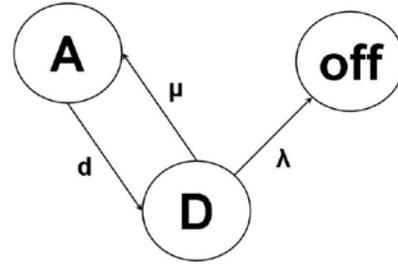


Fig. 1. The states diagram.

Using probabilities formulas, we get

$$\begin{aligned}
 P_A(t + \Delta t) &= P_D(t)\mu\Delta t + P_A(t)(1 - d\Delta t), \\
 P_D(t + \Delta t) &= P_A(t)d\Delta t + P_D(t)(1 - (d + \lambda)\Delta t), \\
 P_{off}(t + \Delta t) &= P_{off}(t) + P_D(t)\lambda\Delta t.
 \end{aligned}$$

Therefore, the probabilities of the sensor states are described by the following system of independent differential equations:

$$\begin{aligned}
 \frac{dp_A(t)}{dt} &= p_D(t)\mu - p_A(t)d, \\
 \frac{dp_D(t)}{dt} &= -p_D(t)(\lambda + \mu) + p_A(t)d, \\
 \frac{dp_{off}(t)}{dt} &= p_D(t)\lambda.
 \end{aligned}$$

Using the normalization condition we get,

$$p_{OFF} = 1 - p_D - p_A.$$

The system of homogeneous linear differential equations can be solved by using the standard methods. $p_{off}(t)$ is the probability that a sensor has failed at or before time t . Thus, reliability of a sensor is given as,

$$S(t) = 1 - p_{OFF}(t).$$

Hence, the average sensor lifetime LT is given by,

$$\int_0^{\infty} S(t)dt.$$

It is reasonable to input the following boundary conditions,

$$p_A(0) = 1, p_D(0) = 0, p_{OFF}(0) = 0.$$

Using the obtained solutions, we can control node throughput and get a reasonable performance of EH-WSNs. From the normalization condition we get,

$$\frac{d}{dt} p_D + \frac{d}{dt} p_A + \frac{d}{dt} p_{OFF} = 0.$$

So, we get

$$\begin{aligned} \frac{d}{dt} p_{OFF} &= \lambda p_D, \\ \frac{d}{dt} p_A &= -p_D(1 + \lambda). \end{aligned}$$

Using the differential equation from the previous section, we receive

$$\begin{aligned} p_A &= \exp(-qt), \\ q &= 1 + \frac{\eta}{1 + \lambda}. \end{aligned}$$

Now, the probabilities of other states can be derived and average sensor lifetime is calculated. Let us remark that the sensor transmit traffic in the stage A. Thus, the function p_A gives as an estimation of sensor throughput.

Thus, we get the method to estimate node reliability.

If we focus on energy consumption then the choice of transmission range is equivalent to determination of energy harvesting schedule (network operating strategy).

Our goal is network reliability optimizing. The corresponding mathematical statement is as follows

$$R(G(n, m(p), p(S))) \rightarrow \max_s .$$

Taking into account the arguments above, The optimization problem for EH-WSNs operation strategy is reduced to the following statement

$$r = \arg \max_{r \in \Omega_r} R(G(n, m(r), p(r))). \quad (1)$$

Here, Ω_r is a set of admissible distances.

In some practical cases it is reasonable to constrain the number of neighboring sensors owing to MAC protocols specificity or interference problems. Hence, the problem (1) has to be reformulated as follows

$$\begin{aligned} R(G(n, m(p), p(S))) &\rightarrow \max_s, \\ \deg(v_i) &\leq a, \forall i \in V. \end{aligned}$$

Here, $a + 1$ is a maximal admissible number of neighbors.

Let us make a few propositions. The transmission range takes a value from a set of admissible distances,

$$r \in \Omega_r.$$

Remark, the growth of r is not necessary leads the growth of m , i.e. it is possible

$$r_1 > r_2, m(r_1) = m(r_2).$$

And,

$$p(r_1) < p(r_2).$$

Hence, for a fixed m

$$R(G(n, m, r_1)) < R(G(n, m, r_2)),$$

if

$$r_1 < r_2.$$

Therefore,

$$\sup_{r \in \Omega_r} r \leq \max |v_i, v_j|, \quad v_i, v_j \in V.$$

We can limit our consideration by a finite discrete set of admissible distances (Ω_r).

Remark, the calculation of $R(G)$ can be hard computational problem. However, in some practical cases (tree-topology, cycles etc.) a polynomial time algorithm can be applied. In general case, proper approximation technique has to be used.

ACKNOWLEDGMENT

This research was supported by the Basic Research Program of the Presidium of the Russian Academy of Sciences.

- [1] Akyildiz I. A survey on sensor networks. IEEE Communications Magazine, vol. 40, no 8, 102 – 114 (2002).
- [2] Jovanov E., Lords A., Raskovic D., Cox P., Adhami R., Andrasik F. Stress monitoring us-ing a distributed wireless intelligent sensor system. IEEE Engineering in Medicine and Biolo-gy Magazine, 22 (3) (2003), pp. 49–55.
- [3] Zhu Y., Song J., Dong F.Applications of wireless sensor network in the agriculture envi-ronment monitoring. Procedia Engineering, 16 (2011), 608-614..
- [4] A. Bayo, D. Antolín, N. Medrano, B. Calvo, S. Celma. Early detection and monitoring of forest fire with a wireless sensor network system. Procedia Engineering, 5 (2010), 248-251.
- [5] Byun J., Jeon B., Noh J., Kim Y., Park S. An intelligent self-adjusting sensor for smart home services based on ZigBee communications. IEEE Trans. Consum. Electr., 58 (2012), 591-596.
- [6] Rodionov, A.S., Migov, D.A., Rodionova, O.K.: Improvements in the Efficiency of Cumulative Updating of All-Terminal Network Reliability. IEEE Trans. on Reliability. 61(2), 460–465 (2012).
- [7] Petingi L.: Introduction of a New Network Reliability Model to Evaluate the Performance of Sensor Networks. International Journal of Mathematical Models and Methods in Applied Sciences. 5(3), 577–585 (2011).
- [8] Moore, E., Shannon, C. Reliable Circuits Using Less Reliable Relays. J. Franklin Inst., 262, n. 4b (1956) 191-208.

On the Financial Applications of Multivariate Stochastic Orderings

Sergio Ortobelli, Tomas Tichy, Tommaso Lando, Filomena Petronio

Abstract—The paper proposes a multivariate comparison among different financial markets, using risk/variability measures consistent with investors' preferences. First of all, we recall a recent classification of multivariate stochastic orderings and properly define the selection problem among different financial markets. Then, we propose an empirical financial application, using multivariate stochastic orderings consistent with the non-satiable and risk averse investors' preferences. For the empirical analysis we examine two different approaches; first, we assume that the return are normally distributed; second, we deal with the more general assumption that the returns' distribution follow a stable sub-Gaussian law.

Keywords—Financial Market comparison, Multivariate preferences, Stochastic Dominance.

I. INTRODUCTION

This paper focuses on the investors' preferences related to the portfolio selection problem. Thus we introduce multivariate stochastic orderings consistent with investors' preferences and show how we can use multivariate risk measures and orders (in terms of probability functionals) to determine dominant sectors/markets in different financial contexts.

We define the dominance among financial markets generalizing the concept of univariate FORS orderings, risk and reward measures in the multivariate framework (see Ortobelli et al. in [2], [3] and [4]). FORS probability functionals and orderings generalize those found in the literature (see [1] and [9]) and are strictly related to the theory of choice under uncertainty and to the theory of probability functionals and metrics (see [6] and [10]). While the new orderings can be used to further characterize and specify the

investors' choices and preferences, the new risk measures should be used either to minimize the risk or to minimize its distance from a given benchmark.

The main contribution of this paper is to use multivariate ordering consistent with investors' preferences to define the dominance among financial markets/sectors. Thus we propose two different approaches: the first one is based on a generalization, of the mean-variance approach. The second one takes into account the possibility of heavy tailed distributions. In this last case, the conditions for the multivariate dominance are based on a comparison between: i) means; ii) dispersion indices and iii) stability indices. Therefore we propose an ex-ante empirical application of multivariate orderings, to evaluate the possible dominance among different financial stock markets (USA, China, Japan and Germany).

The paper is organized as follows. In Section 2 we introduce multivariate FORS orderings and the definition of orderings among markets. Section 3 introduces a preliminary empirical analysis.

II. MULTIVARIATE DOMINANCE

We recall that the most important property that characterizes any *probability functional* associated with a choice problem is the consistency with a stochastic order.

We say that a functional $\mu : \Lambda \times \Lambda \rightarrow R$ is consistent with a preferences orderings \succ anytime that X dominates Y (with respect to a given order of preferences \succ), implies that $\mu(X, Z) \leq \mu(Y, Z)$ for a fixed arbitrary benchmark Z (where $X, Y, Z \in \Lambda$, that is a non-empty space of real valued random variables defined on $(\Omega, \mathfrak{F}, P)$). A univariate FORS measure induced by a given order of preferences \succ can be any probability functional $\mu : \Lambda \times \Lambda \rightarrow R$ which is consistent with \succ . Hence we can similarly define multivariate FORS measures.

Definition 1 We call FORS measure induced by a preference order \succ , any probability functional $\mu : \Lambda \times \Lambda \rightarrow R^s$ (where Λ a non-empty set of real-valued n -dimensional random vectors defined on the probability space $(\Omega, \mathfrak{F}, P)$) that is consistent with a given order of preferences \succ (that is, if X dominates Y with respect to a given order of preferences \succ implies $\mu(X, Z) \leq \mu(Y, Z)$ for a fixed arbitrary benchmark Z where the vectorial inequality is considered for each component i.e., $\mu_i(X, Z) \leq \mu_i(Y, Z)$ for any $i = 1, \dots, s$).

As for the FORS measures we can easily extend the definition of multivariate FORS ordering developed in [2] and [3].

Definition 2 Let $\rho_X : A \rightarrow \bar{R}^s$ (with compact and convex

This paper has been supported by the Italian funds ex MURST 60% 2014 and MIUR PRIN MISURA Project, 2013–2015. The research was also supported through the Czech Science Foundation (GACR) under project 15-23699S and through SP2015/5, an SGS research project of VSB-TU Ostrava, and furthermore by the European Social Fund in the framework of CZ.1.07/2.3.00/20.0296 (first and second author) and CZ.1.07/2.3.00/30.0016 (third and fourth author).

S. Ortobelli is with the Department of Mathematics, Statistics, Informatics and Applications, University of Bergamo, Via Dei Caniana 2, 24127 (BG), Italy (corresponding author e-mail: sergio.ortobelli@unibg.it).

T. Tichy is with the Department of Finance, VŠB-Technical University of Ostrava, Sokolská třída 33, 70121 Ostrava, Czech Republic, (e-mail: tomas.tichy@vsb.cz).

T. Lando is with the Department of Finance, VŠB-Technical University of Ostrava, Sokolská třída 33, 70121 Ostrava, Czech Republic, (e-mail: tommaso.lando@vsb.cz).

F. Petronio is with the Department of Finance, VŠB-Technical University of Ostrava, Sokolská třída 33, 70121 Ostrava, Czech Republic. (e-mail: filomena.petronio@vsb.cz).

$A \subseteq \bar{R}^n$) be a bounded variation function, for every n -dimensional random vector X belonging to a given class Λ . Assume that $\in \Lambda, \rho_X = \rho_Y$, a.e. on A iff $X \stackrel{d}{=} Y$. If, for any fixed $\lambda \in A$, $\rho_X(\lambda)$ is a FORS measure induced by an ordering $>$, then we call FORS orderings induced by $>$ the following new class of orderings defined $\forall X, Y \in \Lambda_\alpha = \left\{ X \in \Lambda : \left| \int_A \prod_{i=1}^n |t_i|^{\alpha_i-1} d\rho_X(t_1, \dots, t_n) \right| < \infty \right\}$ for every $(\alpha_1, \dots, \alpha_n)$ with $\alpha_i \geq 1$ we say that X dominates Y in the sense α -FORS ordering induced by $>$, in symbols:

$$X \text{ FORS } Y_{>, \alpha} \text{ if and only if } \rho_{X, \alpha}(u) \leq \rho_{Y, \alpha}(u) \forall u \in A \tag{1}$$

where

$$\rho_{X, \alpha}(u_1, \dots, u_n) = \begin{cases} \frac{1}{\prod_{i=1}^n \Gamma(\alpha_i)} \int_{\alpha_1}^{u_1} \dots \int_{\alpha_n}^{u_n} \prod_{i=1}^n (u_i - t_i)^{\alpha_i-1} d\rho_X(t_1, \dots, t_n) \\ \rho_X(t_1, \dots, t_n) \text{ if } \alpha_i = 1; i = 1, \dots, n \end{cases} \tag{2}$$

and the integral is a vector applied for each component of the vector $d\rho_X = [d\rho_{(1)X}, \dots, d\rho_{(s)X}]$, whose components are the differential of components of vector $\rho_X = [\rho_{(1)X}, \dots, \rho_{(s)X}]$.

This expression generalizes the one proposed in [5]. Besides, we call ρ_X FORS measure associated with the FORS ordering of random vectors belonging to Λ . We say that ρ_X generates the FORS ordering. Multivariate orderings can have several applications in economics and finance. In this paper we discuss a possible application in ordering financial markets by the point of view of investors. With this aim we need to give some possible alternative definitions of orderings among financial markets/sectors.

Let us assume there are two markets: market A with n assets, and market B with s assets. Assume that the vector of the positions taken by an investor in the n risky assets of market A is denoted by $x = [x_1, \dots, x_n]'$ and similarly the vector of the positions taken by an investor in the m risky assets of market B is denoted by $y = [y_1, \dots, y_s]'$. We assume that no short sales are allowed.

Definition 3 We say that a market/sector A with n assets strongly dominates another market/sector B with s assets with respect to a multivariate FORS ordering if for any vector of returns Y_B of $t \leq u = \min(s, n)$ assets of market/sector B there exists a vector X_A of market/sector A such that X_A FORS Y_B . Similarly we say that a market/sector A with n assets weakly dominates another market/sector B with s assets with respect to the FORS ordering if for any given portfolio of gross returns $y'Y_B$ of market/sector B there exists a portfolio $x'X_A$ of the market/sector A such that $x'X_A$ FORS $y'Y_B$.

Example 1. Suppose that the return distributions of markets A and B are jointly elliptically distributed. Suppose the markets have the same number of assets n , vectors of averages Q_B is negative semidefinite. Then market A strongly dominates market B with respect to the increasing concave

multivariate order (see [1]). Moreover market A weakly dominates market B with respect to the concave order since portfolio $x'\mu_A \geq x'\mu_B$ and $x'Q_A x \geq x'Q_B x$ for any vector $x \geq 0$. Observe that the weak dominance between the markets is also known in ordering as the increasing positive linear concave multivariate order (see [1]).

Example 1 can be used in financial applications. In particular, if we assume that the returns of different markets are jointly elliptically distributed and they are uniquely determined by a risk measure and a reward measure, we can order the markets in a reward-risk framework. On the other hand, if we assume that the distribution does not have finite variance, the mean-variance approach is not appropriate. In this paper we propose a sub-Gaussian distributional assumption, which is quite more suitable for dealing with financial problems (see [7] and the references therein). In particular, we denote the univariate Pareto-Lévy stable distribution by $S_\alpha(\sigma, \beta, \mu)$, where $\alpha \in (0, 2)$ is the so-called stability index, which specifies the asymptotic behavior of the tails, $\sigma > 0$ is the dispersion parameter, $\beta \in [-1, 1]$ is the skewness parameter and $\mu \in \mathbb{R}$ is the location parameter. We consider the same notion used in [8].

A quite easy way to deal with stable distributions is to assume that the vector of returns follows a sub-Gaussian distribution. All components of a sub-Gaussian distribution are α -stable distributions, obtained by setting the skewness parameter $\beta = 0$, i.e. they are symmetric α -stable distributions. Thus, we propose to base the comparison between markets on: i) the vectors of averages; ii) the matrices of dispersion and iii) the stability indices. The motivation is that empirical evidence leads us to strongly suspect that, in the univariate case, a distribution with heavier tails cannot dominate, at the second order (SSD), a distribution with higher expectation, inferior dispersion but heavier tails. Figure 1 and Figure 2 actually show that, on fixed values of σ, μ , the distribution with heavier tails is dominated by the distribution with lighter tails.

This concept can be applied in a multivariate context, generalizing the multivariate mean-variance approach described in Example 1, by taking into account the asymptotic behavior of the tail distributions. This yields the following definition of asymptotic multivariate dominance among financial markets.

Definition 4. Assume that the markets A and B have an equal number of assets n . Assume that the markets A and B are stable sub-Gaussian distributed with stability indices α_A and α_B , vectors of averages μ_A and μ_B , and dispersion matrixes Q_A and Q_B . We say that market/sector A dominates market/sector B with respect to the asymptotic increasing concave multivariate order if $\alpha_A > \alpha_B$, $\mu_A \geq \mu_B$ and $(Q_A - Q_B)$ is negative semi-definite. We say that market/sector A weakly dominates market/sector B with respect to the asymptotic increasing concave multivariate order if $\alpha_A > \alpha_B$, and, for any vector $x \geq 0$, $x'\mu_A \geq x'\mu_B$ and $x'Q_A x \leq x'Q_B x$.

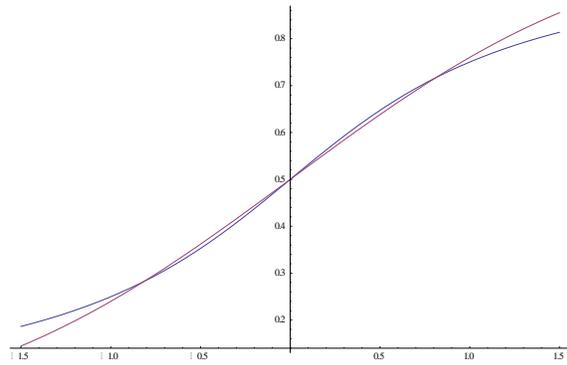


Fig.1. Distribution function of $X \sim S_{1,9}(1,0,0)$ and $Y \sim S_{1,1}(1,0,0)$

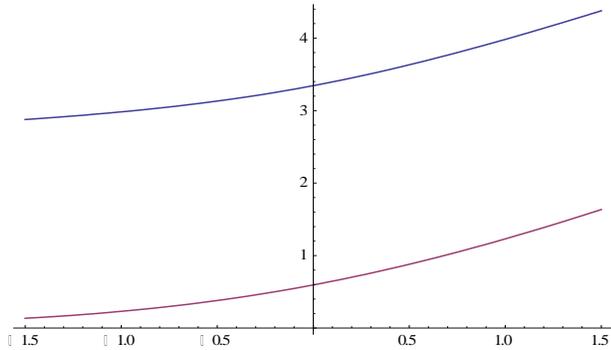


Fig.2. The integral values of the distribution functions of $X \sim S_{1,9}(1,0,0)$ and $Y \sim S_{1,1}(1,0,0)$; $\int_{-\infty}^t F_X(u)du \leq \int_{-\infty}^t F_Y(u)du$ (SSD)

In the following empirical analysis definition 4 to determine in practice if there exists the asymptotic *increasing concave* weak dominance among some equity markets.

III. DOMINANCE COMPARISON AMONG THE US, CHINESE, JAPANESE AND GERMAN STOCK MARKETS

In this section, we evaluate the weak asymptotic multivariate dominance among the US, Chinese, Japanese and German stock markets. In particular, we consider the stocks of: NYSE and NASDAQ (US); Shanghai, Shenzhen and Honk Kong stock exchanges (China); Tokyo, Nagoya and Osaka stock exchanges (Japan); Frankfurt and Berlin stock exchanges (Germany).

First of all, we examine the statistical characteristics of the returns of each market. Then, we verify the dominance among stock markets during the decade 2004-2014. Since, in practical contexts, it is not easy to obtain the strong stochastic dominance among markets, then we verify if the conditions for the asymptotic weak dominance hold, under the implicit assumption the vector of returns of each market is i) normally distributed; ii) alpha stable sub-Gaussian distributed.

- i) We assume that the returns of each country follows a Gaussian distribution with vector of means μ_A and variance-covariance matrix Q_A . For each couple of countries, we determine the mean-variance efficient frontier, as suggested in Example 1.

- ii) We assume that the returns of each country is in the domain of attraction of an α_A stable sub-Gaussian distribution with vector of means μ_A and dispersion matrix Q_A . For each couple of countries, we determine the so called alpha-mean-dispersion efficient frontier, computing the portfolio with minimum dispersion $x'Q_Ax$, for any fixed mean $x'\mu_A$, and finally we compare the efficient frontiers verifying if the conditions $\alpha_A > \alpha_B$, $x'\mu_A \geq x'\mu_B$ and $x'Q_Ax \leq x'Q_Bx$ hold.

The results of the two approaches are summarized in the Table 1.

Tab. 1 Number of trimesters (January 2004- December 2014) when dominance among markets holds.

Mean-Variance comparison				
	USA	Germany	China	Japan
USA	//	1	2	10
Germany	0	//	0	3
China	8	5	//	20
Japan	0	0	0	//
Alpha-Mean-Dispersion comparison				
	USA	Germany	China	Japan
USA	//	0	0	0
Germany	0	//	0	0
China	2	1	//	0
Japan	0	0	0	//

Table 1 reports the number of times (trimesters) when a market dominates another, in terms of reward-risk analysis, during the decade January 2004–December 2014. First of all, we observe that there exists a strong difference between the comparison based on mean-variance efficiency and the alpha-mean-dispersion efficiency. We observe that US and Chinese markets dominate the other two more frequently in the mean-variance framework. On the other hand, using the alpha-mean-dispersion criterion, only the Chinese market dominates few times the German and US markets. Moreover, we observe that the Japanese market never dominates the others and it is often dominated in terms of mean-variance. However, Japanese market is never dominated in terms of alpha-mean-dispersion efficient frontier, because it presents lower kurtosis and smaller tails, as also observed in Table 1. Therefore, from this analysis, the most performing market is the Chinese emerging market.

IV. CONCLUSION

We introduced a methodology aimed at comparing different financial markets/sectors from the point of view of a non-satiating risk-averse investor, the method is applied to four stock markets (US, German, Japan and China). The method could be very useful for investors who want to optimize their international portfolio, in particular, this analysis can be generally applied to preselect the “best” markets to invest in. In section 2, we proposed a definition of multivariate dominance among different markets and evaluate it with empirical comparison between markets, assuming that the returns are in i) normally distributed; ii) sub-Gaussian

distributed. We observe that the mean-variance dominance (approach i) among different markets is verified several times, although we generally do not observe the asymptotic dominance (approach ii), except in few cases. In particular, while the Japanese stock market appears to be dominated in terms of mean-variance, it is never asymptotically dominated since it presents an index of stability generally higher compared to the other countries. This result suggests that the big losses observed during the crisis have a stronger impact in the US, China and German stock markets.

REFERENCES

- [1] A. Muller and D. Stoyan, D., *Comparison Methods for Stochastic Models and Risks*, Wiley Series in Probability and Statistics, 2002.
- [2] S. Ortobelli, S. Rachev, H. Shalit, and F. Fabozzi, "Orderings and Risk Probability Functionals in Portfolio Theory", *Probability and Mathematical Statistics*, Vol. 28, No. 2, pp. 203-234, 2008.
- [3] S. Ortobelli, S. Rachev, H. Shalit, and Fabozzi, F., "Orderings and Probability Functionals Consistent with Preferences", *Applied Mathematical Finance*, vol. 16, No. 1, pp. 81-102, 2009.
- [4] S. Ortobelli, S. Rachev, H. Shalit, "Portfolio Selection Problems Consistent with Given Preference Orderings", *International Journal of Theoretical and Applied Finance*, Vol. 16, No. 5, 2013.
- [5] F. Petronio, S. Ortobelli and T. Tichy, "Multivariate stochastic orderings consistent with preferences and their possible applications", in *Proc. of Mathematical Methods in Economics*, University of Jihlava (VSPJ), 2013, pp. 724-729.
- [6] S. Rachev, *Probability Metrics and the Stability of Stochastic Models*. John Wiley & Sons, Chichester, 1991.
- [7] S. Rachev, and S. Mittnik. *Stable paretian models in finance*, Wiley, New York, 2000.
- [8] G. Samorodnitsky and M.S. Taggu: *Stable Non-Gaussian Random Processes*. Chapman & Hall, New York, 1994.
- [9] M. Shaked and G. Shanthikumar, *Stochastic orders and their applications*. Academic Press Inc. Harcourt Brace & Company, New York, 1993.
- [10] S. Stoyanov, S. Rachev, S. Ortobelli and F. Fabozzi, "Relative deviation metrics and the problem of strategy replication", *Journal of Banking and Finance*, Vol. 32, No. 2, pp. 199-206, 2008.

Mathematical model of two-links mechanism movement at discrete control actions

Sergey Jatsun, Doctor of science, Professor, Head of the department of mechanics, mechatronics and robotics, Southwest State University

Sergei Savin, Candidate of science, Junior research fellow of the department of mechanics, mechatronics and robotics, Southwest State University

Petr Bezmen, Candidate of science, Associate professor of the department of mechanics, mechatronics and robotics, Southwest State University

Abstract—In this paper the problem of modelling dynamics of a two link mechanism with discrete control system is discussed. The influence that discrete nature of different parts of the control system has on the controlled motion of the mechanism is analyzed. It is shown that using model of the control system with discrete components can produce qualitative different results when studying stability of the two link mechanism.

Keywords—Discrete control system, two link mechanism, dynamics.

I. INTRODUCTION

Mathematical modeling of motion of a mobile robot is an inalienable step of its designing. It is necessary to consider the dynamics of nonlinear equations of the mechanism, and the discrete nature of the elements of its automatic control system (CS).

A typical example is an exoskeleton that provides a controlled change in the orientation of its body. The dissimilarity of functioning of the real regulators which are realized on the basis of modern digital electronics against their linear models, can change the dynamic characteristics of the exoskeleton sufficiently to result in instability or incorrect behavior of the object. This is especially important to consider in design of control systems for rehabilitation apparatuses, based on exoskeletons, as in this case, significant errors in the CS may lead to injury of a patient.

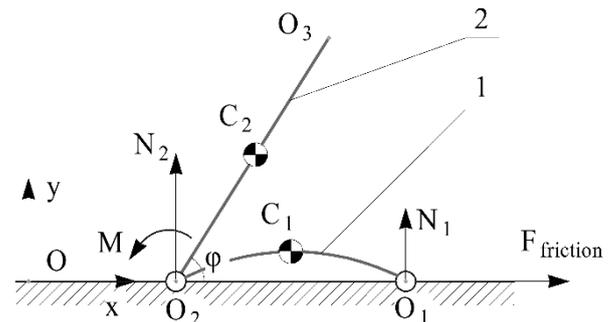
While there are many publications that discuss dynamic behavior of a multilink mechanical systems (for example see [1-5]), as well as publications and text books that deal with discrete control (for example see [6-7]), there are relatively few works dedicated to study of the dynamics of an electromechanical system taking into account discrete nature of elements that form the control system.

In this paper we pose the problem of the development of a mathematical model of the automatic control system for the process of controlled changing of the exoskeleton body orientation. In this mode, the exoskeleton may be regarded as a two-linked mechanism with one fixed link (exoskeleton foot). It is important to ensure the immobility of the foot, because foot slipping, rotation or

detachment from the surface are not part of the normal mode of operation.

II. THE OBJECT OF THE RESEARCH

In this paper, we consider the motion of two-linked mechanism shown in Figure 1. The first link relies on the surface at two points, and we assume that the friction force acts only in one of the reference points. In joint connecting the two link, there is a motor that generates torque M . The masses of the links are m_1, m_2 , the lengths of the links are l_1, l_2 .



1, 2 – first and second links
Figure 1 The scheme of the investigated mechanism

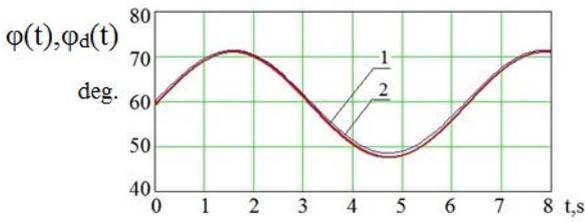
The figure 1 shows the fixed coordinate system Oxy , the normal reaction forces N_1, N_2 , the friction force $F_{friction}$. The points C_1, C_2 are the mass centers of the first and second link. For the specified mechanism in the case when the first link is motionless we can write the equations of dynamics in the form:

$$J\ddot{\varphi} = M - \frac{1}{2}l_2 \cos(\varphi)m_2g, \quad (1)$$

where J is the moment of inertia of the second section relative to point O_2 , g is the gravitational acceleration constant, M is the torque between the first and the second links.

To determine the value M we use the fact that the electric motor torque is proportional to the current in the windings of its armature is $M = iC_e$ [8]. To calculate current we can write the Kirchhoff equation for the armature circuit:

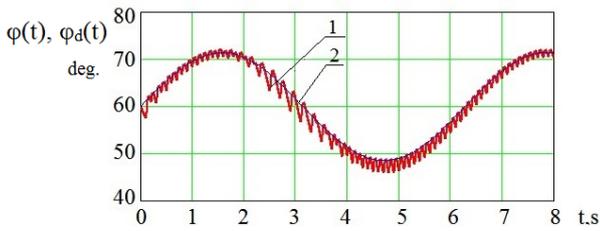
$$u - Ri - \dot{\varphi}C_\omega = 0, \quad (2)$$



1, 2 – The time dependences $\varphi(t)$ and $\varphi_d(t)$

Figure 3 The movement of the system without discrete elements in ACS

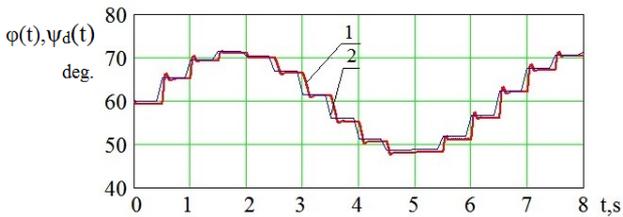
We consider the effect of the period of the PWM signal to the motion of the system. At values of T_p equal to or less than 10 msec, the obtained dependences visually identical to that shown in Figure 3. Since the period of the PWM signal is less than 10 msec (in practice), it can be stated that the effect of the nonlinearity is relatively small and can be neglected in the study of the dynamics of a controlled mechanical system. In cases where the period of the PWM signal is chosen much larger than 10 msec, it can make a significant distortion in the nature of the system movement. This distortion manifests in the appearance of high-frequency oscillations (see figure 4).



1, 2 – The time dependences $\varphi(t)$ and $\varphi_d(t)$

Figure 4 Movement of the system at $T_p = 0.1$ sec

Time period of the control action update T_d begins to have a significant visual influence on the dynamics of the system at $T_d > 50$ msec. This manifests itself in the stepwise nature of the dependence $\varphi(t)$ and the presence of an overshoot at the transition between the “steps” of the graph.

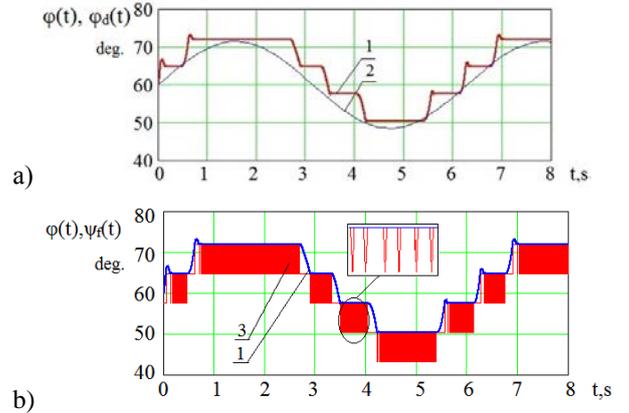


1, 2 – The time dependences $\varphi(t)$ and $\psi_d(t)$

Figure 5 Movement of the system at $T_d = 0.05$ sec

The form of the obtained dependence can be interpreted as follows. The abrupt change of control action value makes the system execute in transient conditions, which leads to oscillations associated with overshoot. After the end of the transitional process, the system executes the control action value that does not change during time T_d . It leads to the appearance of horizontal sections of dependence.

We consider the influence of the measuring device discreteness in the feedback loop on the CS work process. The influence of nonlinearity, produced by the angle sensor, begins to manifest itself visually at $\Delta\varphi > 0.3$ degree (the example is shown in figure 6 (a)).



1, 2, 3 – The time dependences $\varphi(t)$, $\varphi_d(t)$ and $\psi_f(t)$

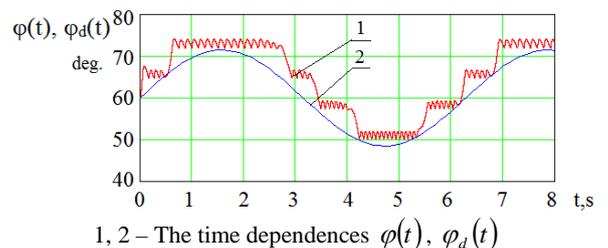
Figure 6 Movement of the system at $\Delta\varphi = 7.2$ degree

The form of obtained dependence $\varphi(t)$ is similar one that is shown in the figure 5. In this case, the control action value changes continuously, i.e. nature of the horizontal sections appearance cannot be explained in the same way as for the dependences shown in the figure 5. To illustrate the horizontal sections occurrence on the graph $\varphi(t)$ we can construct the dependence graph $\psi_f(t)$ (figure 6 (b)). In horizontal sections $\varphi(t)$ the dependence $\psi_f(t)$ graph switches at a high frequency (switching period is comparable to the integration step of the selected numerical method) between two discrete values. Such behavior of the function $\psi_f(t)$ does not reflect the behavior of real angle sensors. In order to eliminate the high-frequency switching process shown in the figure 6 (b), we introduce a model of the sensor so-called “dead zone”:

$$\psi_f(t) = \begin{cases} m \cdot \Delta\varphi & \text{if } |\varphi - m \cdot \Delta\varphi| > k_z \Delta\varphi \\ \psi_f(t - \Delta t) & \text{otherwise} \end{cases} \quad \text{for } \varphi \in [m \cdot \Delta\varphi, (m+1) \cdot \Delta\varphi] \quad (9)$$

where $\psi_f(t - \Delta t)$ is the function value in the previous step of integration, k_z is the factor what determines the width of the “dead” band (as a measurement step part of angle sensor).

The use of derived functions $\psi_f(t)$ allows to obtain the dependences shown in the figure 7.



1, 2 – The time dependences $\varphi(t)$, $\varphi_d(t)$

Figure 7 Movement of the system at $\Delta\varphi = 7.2$ degree (the sensor model with “dead zone” ($k_z = 10\%$))

The figure 7 pays attention to the appearance of high-frequency oscillations on the graph $\varphi(t)$. It is shown that the amplitude and frequency of these oscillations depends on the value k_z .

We note the described nonlinearities influence on the mechanical system performance. Let one of the system quality criteria be immobility of the first section during some period of time. The condition of the first link movement beginning is the state when one of the normal reactions applied at the points O_1 and O_2 equal to zero. We construct the normal reaction N_2 time dependence in the case where the discrete components is not included in CS (see figure 8 (a)), and in the case where the signal in the feedback loop formed by the formula (9) (see figure 8 (b)).

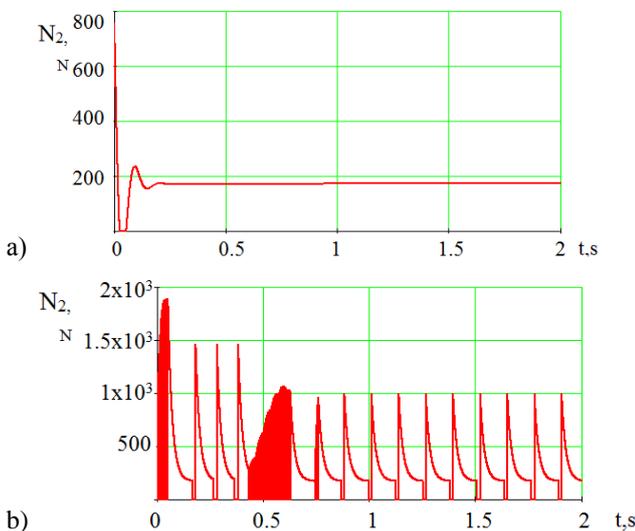


Figure 8 The time dependence of the normal reaction N_2

We must take into account that the discrete elements of CS can give a significantly larger number of the normal reaction zeroing moments. We do not set the task to determine whether the device could be overturned.

Conclusions

As the result of the mathematical simulation of the dynamic behavior of the two link mechanism. It is shown that discrete nature of input signal, signal in the feedback loop and PWM signal that is used to control the motor each has different influence on the form of the trajectory of the system, as expressed in the change law of its generalized coordinate. In particular it is shown that high update time of the input signal (T_d) can result in system behaving as in transition process, which results in overshoots. For models with discrete joint sensor models similar results were obtained, and additional high frequency oscillations were observed after introducing dead zone to the sensor model. In general, it was shown that models that take into account discrete nature of different components of the control system are likely to

produce high frequency oscillations of the controlled parameter. Also as evidenced by the results shown on the figure 8, discrete model of a mechanical system can produce a very different results from its non-discrete analog when studying its stability.

REFERENCES

1. Yatsun, S.F. Simulation of Motion of a Multilink Jumping Robot and Investigation of Its Characteristics / S.F. Yatsun, L.Yu. Volkova // Journal of Computer and Systems Sciences International, 2013, Vol. 52, No. 4, pp. 637–649.
2. Vorochaeva L.Yu. Simulation of Motion of a Three-Link Robot with Controlled Friction Forces on a Horizontal Rough Surface / L.Yu. Vorochaeva, G.S. Naumov, S.F. Yatsun // Journal of Computer and Systems Sciences International, 2015, Vol. 54, No. 1, pp. 151–164.
3. Yatsun, S.F. Modelirovaniye dvizheniya mnogozvennogo prygayushchego robota i issledovaniye yego kharakteristik / S.F. Yatsun, L.YU. Volkova // Izvestiya RAN. Teoriya i sistemy upravleniya. 2013. - № 4. - S. 137–149.
4. Volkova, L.YU. Izucheniye vliyaniya polozheniya tochki zakrepleniya nogi prygayushchego robota v korpuse na kharakter dvizheniya ustroystva / L.YU. Volkova, S.F. Yatsun // Nelineynaya dinamika. 2013. T. 9. № 2. S. 327–342.
5. Vorochayeva L.YU. Modelirovaniye dvizheniya trekhzvennogo robota s upravlyayemyimi silami treniya po gorizontальной sverokhvatoy poverkhnosti / L.YU. Vorochayeva, G.S. Naumov, S.F. Yatsun // Izvestiya RAN. Teoriya i sistemy upravleniya. 2015. № 1. S. 156–170.
6. Phillips, Charles L., and H. Troy Nagle. Digital control system analysis and design. Prentice Hall Press, 2007.
7. Franklin, Gene F., J. David Powell, and Michael L. Workman. Digital control of dynamic systems. Vol. 3. Menlo Park: Addison-wesley, 1998.
8. Besekersky, V., and E. Popov. "Automatic control system theory." Moscow, Russia: Professiya (2004).

Navier-Stokes equations-Millennium Prize Problems

ASSET DURMAGAMBETOV
NATIONAL DEFENSE UNIVERSITY
Scientific Research Institute

72 Turan ave., Astana
KAZAKHSTAN
aset.durmagambet@gmail.com

LEYLA FAZILOVA
Karaganda State University
Department of Applied Mathematics and
Computer Science
28 University Street, 100028 Karaganda
KAZAKHSTAN
leyla.fazilova@gmail.com

Abstract: In this work we present final solving Millennium Prize Problems formulated Clay Math. Inst., Cambridge. A new uniform time estimation of the Cauchy problem solution for the Navier-Stokes equations is provided. Describes the loss of smoothness of classical solutions for the Navier-Stokes equations.

Key-Words: Schrödinger's equation, potential, scattering amplitude, Cauchy problem, Navier-Stokes equations, Fourier transform, the global solvability and uniqueness of the Cauchy problem, the loss of smoothness, The Millennium Prize Problems \LaTeX

1 Introduction

In this work we present final solving Millennium Prize Problems formulated Clay Math. Inst., Cambridge in [1] Before this work we already had first results in [2]-[4]. The Navier-Stokes existence and smoothness problem concerns the mathematical properties of solutions to the Navier-Stokes equations. These equations describe the motion of a fluid in space. Solutions to the Navier-Stokes equations are used in many practical applications. However, theoretical understanding of the solutions to these equations is incomplete. In particular, solutions of the Navier-Stokes equations often include turbulence, which remains one of the greatest unsolved problems in physics. Even much more basic properties of the solutions to Navier-Stokes have never been proven. For the three-dimensional system of equations, and given some initial conditions, mathematicians have not yet proved that smooth solutions always exist, or that if they do exist, they have bounded energy per unit mass. This is called the Navier-Stokes existence and smoothness problem. Since understanding the Navier-Stokes equations is considered to be the first step to understanding the elusive phenomenon of turbulence, the Clay Mathematics Institute in May 2000 made this problem one of its seven Millennium Prize problems in mathematics.

In this paper, we introduce important explanations results presented in the previous studies in [2]-[4]. We therefore reiterate the basic provisions of the preceding articles to clarify understanding them. First, we consider some ideas for the potential in the

inverse scattering problem, and this is then used to estimate of solutions of the Cauchy problem for the Navier-Stokes equations.

A similar approach has been developed for one-dimensional nonlinear equations [5]-[8], but to date, there have been no results for the inverse scattering problem for three-dimensional nonlinear equations. This is primarily due to difficulties in solving the three-dimensional inverse scattering problem.

This paper is organized as follows: first, we study the inverse scattering problem, resulting in a formula for the scattering potential. Furthermore, with the use of this potential, we obtain uniform time estimates in time of solutions of the Navier-Stokes equations, which suggest the global solvability of the Cauchy problem for the Navier-Stokes equations.

Essentially, the present study expands the results for one-dimensional nonlinear equations with inverse scattering methods to multi-dimensional cases. In our opinion, the main achievement is a relatively unchanged projection onto the space of the continuous spectrum for the solution of nonlinear equations, that allows to focus only on the behavior associated with the decomposition of the solutions to the discrete spectrum. In the absence of a discrete spectrum, we obtain estimations for the maximum potential in the weaker norms, compared with the norms for Sobolev' spaces.

Consider the operators

$$H = -\Delta_x + q(x), \quad H_0 = -\Delta_x$$

defined in the dense set $W_2^2(\mathbb{R}^3)$ in the space $L_2(\mathbb{R}^3)$, and let q be a bounded fast-decreasing function. The

operator H is called Schrödinger's operator.

We consider the three-dimensional inverse scattering problem for Schrödinger's operator: the scattering potential must be reconstructed from the scattering amplitude. This problem has been studied by a number of researchers [9], [11], [12] and references therein.

2 Results

Consider Schrödinger's equation:

$$-\Delta_x \Psi + q\Psi = |k|^2 \Psi, \quad k \in C \quad (1)$$

Let $\Psi_+(k, \theta, x)$ be a solution of (1) with the following asymptotic behavior:

$$\begin{aligned} \Psi_+(k, \theta, x) &= e^{ik\theta x} + \\ &+ \frac{e^{i|k||x|}}{|x|} A(k, \theta', \theta) + o\left(\frac{1}{|x|}\right), \quad |x| \rightarrow \infty, \end{aligned} \quad (2)$$

where $A(k, \theta', \theta)$ is the scattering amplitude and $\theta' = \frac{x}{|x|}$, $\theta \in S^2$ for $k \in \bar{C}^+ = \{Imk \geq 0\}$

$$\begin{aligned} A(k, \theta', \theta) &= \\ &= -\frac{1}{4\pi} \int_{R^3} q(x) \Psi_+(k, \theta, x) e^{-ik\theta' x} dx. \end{aligned} \quad (3)$$

Let us also define the solution $\Psi_-(k, \theta, x)$ for $k \in \bar{C}^- = \{Imk \leq 0\}$ as

$$\Psi_-(k, \theta, x) = \Psi_+(-k, -\theta, x).$$

As is well known [9]:

$$\begin{aligned} \Psi_+(k, \theta, x) - \Psi_-(k, \theta, x) &= \\ &= -\frac{k}{4\pi} \int_{S^2} A(k, \theta', \theta) \Psi_-(k, \theta', x) d\theta', \quad k \in R. \end{aligned} \quad (4)$$

This equation is the key to solving the inverse scattering problem, and was first used by Newton [10], [11] and Somersalo et al. [12].

Equation (4) is equivalent to the following:

$$\Psi_+ = S\Psi_-, \quad (5)$$

where S is a scattering operator with the kernel $S(k, t)$ and

$$S(k, t) = \int_{R^3} \Psi_+(k, x) \Psi_-^*(t, x) dx.$$

The following theorem was stated in [9]:

Theorem 1 (The energy and momentum conservation laws) Let $q \in \mathbf{R}$. Then, $SS^* = I$, $S^*S = I$, where I is a unitary operator.

Definition 2 The set of measurable functions \mathbf{R} with the norm, defined by

$$\|q\|_{\mathbf{R}} = \int_{R^6} \frac{q(x)q(y)}{|x-y|^2} dx dy < \infty$$

is recognized as being of Rollnik class.

As shown in [13], $\Psi_{\pm}(k, x)$ is an orthonormal system of H eigenfunctions for the continuous spectrum. In addition to the continuous spectrum there are a finite number N of H negative eigenvalues, designated as $-E_j^2$ with corresponding normalized eigenfunctions $\psi_j(x, -E_j^2)$ ($j = \overline{1, N}$), where

$$\psi_j(x, -E_j^2) \in L_2(R^3).$$

We present Povzner's results [13] below:

Theorem 3 (Completeness) For both an arbitrary $f \in L_2(R^3)$ and for H eigenfunctions, Parseval's identity is valid.

$$\|f\|_{L_2}^2 = (P_D f, P_D f) + (P_{Ac} f, P_{Ac} f).$$

$$P_D f = \sum_{j=1}^N f_j \psi_j(x, -E_j).$$

$$P_{Ac} f = \int_0^\infty \int_{S^2} s^2 \bar{f}(s) \Psi_+(s, \theta, x) d\theta ds, \quad (6)$$

where \bar{f} and f_j are Fourier coefficients for the continuous and discrete cases.

Theorem 4 (Birman-Schwinger estimation). Let $q \in R$. Then, the number of discrete eigenvalues can be estimated as:

$$N(q) \leq \frac{1}{(4\pi)^2} \int_{R^3} \int_{R^3} \frac{q(x)q(y)}{|x-y|^2} dx dy. \quad (7)$$

This theorem was proved in [14].

Let us introduce the following notation:

$$NA = \int_{S^2} A(k, \theta', \theta) d\theta, \quad \text{for } f = f(k, \theta', x),$$

$$Df = k \int_{S^2} A(k, \theta', \theta) f(k, \theta', x) d\theta', \quad (8)$$

$$\phi_0(\sqrt{z}, \theta, x) = e^{i\sqrt{z}\theta x},$$

$$Phi(\sqrt{z}, \theta', x) = (\Psi_+(\sqrt{z}, \theta, x) - e^{i\sqrt{z}\theta x}) \Delta, \quad (9)$$

where $\Delta = \prod_{j=1}^N (k + iE_j)/(k - iE_j)$. We define the operators T_{\pm}, T for $f \in W_2^1(R)$ as follows:

$$T_+ f = \frac{1}{2\pi i} \lim_{Imz \rightarrow 0} \int_{-\infty}^{\infty} \frac{f(s)}{s-z} ds, \quad Im z > 0,$$

$$T_-f = \frac{1}{2\pi i} \lim_{Imz \rightarrow 0} \int_{-\infty}^{\infty} \frac{f(s)}{s-z} ds, \quad Im z < 0, \quad (10)$$

$$Tf = \frac{1}{2}(T_+ + T_-)f. \quad (11)$$

Consider the Riemann problem of finding a function Φ , that is analytic in the complex plane with a cut along the real axis. Values of Φ on the sides of the cut are denoted as Φ_+ , Φ_- . The following presents the results of [15]:

Lemma 5

$$TT = \frac{1}{4}I, \quad TT_+ = \frac{1}{2}T_+, \quad TT_- = -\frac{1}{2}T_-,$$

$$T_+ = T + \frac{1}{2}I, \quad T_- = T - \frac{1}{2}I. \quad (12)$$

Theorem 6 Let $q \in \mathbf{R}$, $g = (\Phi_+ - \Phi_-)$. Then,

$$\Phi_{\pm} = T_{\pm}g. \quad (13)$$

The proof of the above follows from the classic results for the Riemann problem.

Lemma 7 Let $q \in \mathbf{R}$, $g_+ = g(\sqrt{z}, \theta, x)$, $g_- = g(\sqrt{z}, -\theta, x)$. Then,

$$\Psi_+(\sqrt{z}, \theta, x)\Delta = (T_+g_+ + e^{i\sqrt{z}\theta x}),$$

$$\Psi_-(\sqrt{z}, \theta, x)\Delta = (T_-g_- + e^{-i\sqrt{z}\theta x}). \quad (14)$$

The proof of the above follows from the definitions of g , Φ_{\pm} , Ψ_{\pm} .

Lemma 8 Let $q \in \mathbf{R}$,

$$A_+ = A(\sqrt{z}, \theta, x), \quad A_- = A(\sqrt{z}, -\theta, x). \quad (15)$$

Then

$$A(k, \theta', \theta)\Delta = T_+(A_+\Delta - A_-\Delta).$$

The proof of the above again follows from the definitions of the functions g , Φ_{\pm} , Ψ_{\pm} .

Lemma 9 Let $q \in \mathbf{R}$. Then,

$$NA_+\Delta = NT_+(DA_-\Delta). \quad (16)$$

The proof of the above follows from the definitions of g , Φ_{\pm} , Ψ_{\pm} and Theorem 1.

Lemma 10 Let $q \in \mathbf{R}$. Then,

$$|NT(A_+)| \leq 2|NA_+|. \quad (17)$$

The proof of the above follows from the definitions of g , Φ_{\pm} , Ψ_{\pm} and Lemma 9 and dispersions relations for analytics functions.

Definition 11 Denote by TA the set of functions $f(k, \theta, \theta')$ with the norm

$$\|f\|_{TA} = \sup_{\theta, k, \theta'} (|Tf| + |f|) < \infty.$$

Definition 12 Denote by $R_{(I-T_-D)}$ the set of functions g such that

$$g = (I - T_-D)f$$

for any $f \in TA$.

Lemma 13 Suppose $\|A\|_{TA} < \alpha < 1$. Then, the operator $(I - T_-D)$, defined on the set TA has an inverse defined on $R_{(I-T_-D)}$.

The proof of the above follows from the definitions of D , T_- and the conditions of Lemma 13.

Lemma 14 Let $q \in \mathbf{R}$, and assume that $(I - T_{\pm}D)^{-1}$ exists. Then,

$$g = T_+g - T_-g, \quad T_-g_- = (I - T_-D)^{-1}T_-D\phi_0,$$

$$\Psi_- = \frac{1}{\Delta}(I - T_-D)^{-1}T_-D\phi_0 + \phi_0. \quad (18)$$

The proof of the above follows from the definitions of g , Φ_{\pm} , Ψ_{\pm} and equation (4).

Lemma 15 Let $q \in \mathbf{R}$, and assume that $(I - T_{\pm}D)^{-1}$ exists. Then,

$$\Psi_- = \sum_{i=1}^{\infty} \left(-\frac{1}{\Delta}T_-D\right)^i \phi_0 + \phi_0.$$

$$\frac{1}{\Delta}T_-D + \overline{\frac{1}{\Delta}T_-D} =$$

$$= \frac{1}{\Delta}T_-D \overline{\frac{1}{\Delta}T_-D} + \overline{\frac{1}{\Delta}T_-D} \frac{1}{\Delta}T_-D + Q_3, \quad (19)$$

where Q_3 represents terms of highest order of T_-D .

Proof: Using

$$\int_{R^3} \Psi_-(x, k) * \overline{\Psi_-(x, l)} dx = \delta(k - l),$$

$$\int_{R^3} \phi_0(x, k) * \overline{\phi_0(x, l)} dx = \delta(k - l)$$

and (18) we get proof. □

Lemma 16 Let $q \in \mathbf{R}$. Then,

$$q = \lim_{z \rightarrow 0} H_0 \Psi_- / \Psi_- . \quad (20)$$

The lemma can be proved by substituting Ψ_{\pm} into equation (1).

Lemma 17 Let $q \in \mathbf{R}$, and assume that $(I - T_- D)^{-1}$ exists. Then,

$$q = \lim_{z \rightarrow 0} \frac{\frac{1}{\Delta} N(I - T_- D)^{-1} T_- D H_0 \phi_0}{\frac{1}{\Delta} N(I - T_- D)^{-1} T_- D \phi_0 + N \phi_0} . \quad (21)$$

The proof of the above follows from the definitions of N , Ψ_{\pm} and Lemma 14.

Lemma 18 Let $q \in \mathbf{R}$. Then $\|D\| \leq 2$.

The proof of the above follows from the definition of D and the unitary nature of S .

Lemma 19 Let $q \in \mathbf{R} \cap L_4(\mathbf{R}^3)$. Then,

$$E_j^2 \leq \int_{\mathbf{R}^3} |q(x)| |\psi_j|^2 dx, \quad (22)$$

$$\max_x |\psi_j(x)| \leq 2 \|q \psi_j\|_{L_2(\mathbf{R}^3)}. \quad (23)$$

The proof of the above follows from the definitions of E_j^2 , ψ_j and (1).

Lemma 20 Let $q \in \mathbf{R} \cap L_2(\mathbf{R}^3)$, and

$$\|A\|_{TA} < \alpha < 1.$$

Then,

$$[\Psi_{\pm} |q|]_{x=0} \leq \sum_{i=1}^{\infty} (C_0 |NA|_{|k|=0})^i. \quad (24)$$

To prove this result, one should calculate Ψ_{\pm} using (18)

$$\Psi_{\pm} q = \Delta \Psi_{\pm}. \quad (25)$$

Using the notation that:

$$\tilde{q}(k) = \int_{\mathbf{R}^3} q(x) e^{i(k,x)} dx,$$

$$\tilde{q}(k-l) = \int_{\mathbf{R}^3} q(x) e^{i(k-l,x)} dx,$$

$$Qq = \int_{\mathbf{R}^3} q(x) e^{i(k-l,x)} dx,$$

$$Q_E q = \int_{\mathbf{R}^3} q(x) e^{i(k-l,x)} dx_{|k|=|l|},$$

$$NQq = \int_{S^2} Qq(k, \theta', \theta) d\theta, \text{ for } f = f(k, \theta', x),$$

$$Df = k \int_{S^2} A(k, \theta', \theta) f(k, \theta', x) d\theta'. \quad (26)$$

Lemma 21 Let $q \in \mathbf{R} \cap L_2(\mathbf{R}^3)$ and

$$\|TA\|_{TA} < \alpha < 1.$$

Then,

$$\|TNA\|_{L_2} < C \|TQq\|_{L_2}.$$

$$\|NA\|_{L_2} < C \|Qq\|_{L_2}. \quad (27)$$

To prove this result, one should use Ψ_{\pm} using Lemma 14

$$q = \Delta \Psi_{\pm} / \Psi_{\pm}. \quad (28)$$

Lemma 22 Let $q \in \mathbf{R} \cap L_2(\mathbf{R}^3)$, and

$$\|A\|_{TA} < \alpha < 1.$$

Then,

$$\Psi_{\pm} |q|_{x=0} \leq \sum_{i=1}^{\infty} (C_0 |TNA|_{|k|=0})^i. \quad (29)$$

$$\Psi_{\pm} |q|_{x=0} \leq \sum_{i=1}^{\infty} (C_0 |TQq|_{|k|=0})^i. \quad (30)$$

To prove this result, one should calculate A using Lemma 14.

Lemma 23 Let $q \in \mathbf{R}$, $\max_k |\tilde{q}| < \infty$. Then,

$$\int_{\mathbf{R}^3} \int_{\mathbf{R}^3} \frac{q(x)q(y)}{|x-y|^2} dx dy \leq C (\|q\|_{L_2} + \max_k |\tilde{q}|)^2. \quad (31)$$

A proof of this lemma can be obtained using Plancherel's theorem.

Lemma 24 Let $q \in \mathbf{R} \cap L_2(\mathbf{R}^3)$, and $\|q\|_{L_2} + \max_k |\tilde{q}(k)| < \alpha < 1$. Then,

$$\Psi_{\pm}|_{x=0} > 1 - \alpha / (1 - \alpha) \quad (32)$$

$$|q|_{x=0} \leq \sum_{i=1}^{\infty} (C_0 |TQq|_{|k|=0})^i. \quad (33)$$

To prove this result, one should calculate $\Psi_{\pm}|_{x=0}$.

3 Cauchy problem for the Navier-Stokes equation

Numerous studies of the Navier-Stokes equations have been devoted to the problem of the smoothness of its solutions. A good overview of these studies is given in [16]-[20]. The spatial differentiability of the solutions is an important factor, this controls their evolution. Obviously, differentiable solutions do not

provide an effective description of turbulence. Nevertheless, the global solvability and differentiability of the solutions has not been proven, and therefore the problem of describing turbulence remains open. It is interesting to study the properties of the Fourier transform of solutions of the Navier-Stokes equations. Of particular interest is how they can be used in the description of turbulence, and whether they are differentiable. The differentiability of such Fourier transforms appears to be related to the appearance or disappearance of resonance, as this implies the absence of large energy flows from small to large harmonics, which in turn precludes the appearance of turbulence. Thus, obtaining uniform global estimations of the Fourier transform of solutions of the Navier-Stokes equations means that the principle modeling of complex flows and related calculations will be based on the Fourier transform method. The authors are continuing to research these issues in relation to a numerical weather prediction model; this paper provides a theoretical justification for this approach. Consider the Cauchy problem for the Navier-Stokes equations:

$$q_t - \nu \Delta q + (q, \nabla q) = -\nabla p + f(x, t), \quad (34)$$

where

$$\begin{aligned} \operatorname{div} q &= 0, \\ q|_{t=0} &= q_0(x) \end{aligned} \quad (35)$$

in the domain $Q_T = R^3 \times (0, T)$, where :

$$\operatorname{div} q_0 = 0. \quad (36)$$

The problem defined by (34), (35), (36) has at least one weak solution (q, p) in the so-called Leray–Hopf class [16].

The following results have been proved [17]:

Theorem 25 *If*

$$q_0 \in W_2^1(R^3), \quad f \in L_2(Q_T), \quad (37)$$

there is a single generalized solution of (34), (35), (36) in the domain Q_{T_1} , $T_1 \in [0, T]$, satisfying the following conditions:

$$q_t, \nabla^2 q, \quad \nabla p \in L_2(Q_T). \quad (38)$$

Note that T_1 depends on q_0 and f .

Lemma 26 *Let $q_0 \in W_2^1(R^3)$, $f \in L_2(Q_T)$. Then,*

$$\begin{aligned} \sup_{0 \leq t \leq T} \|q\|_{L_2(R^3)}^2 + \int_0^t \|\nabla q\|_{L_2(R^3)}^2 d\tau &\leq \\ &\leq \|q_0\|_{L_2(R^3)}^2 + \|f\|_{L_2(Q_T)}. \end{aligned} \quad (39)$$

Our goal is to provide global estimations for the Fourier transforms of derivatives of the Navier–Stokes equations’ solutions (34), (35), (36) without the that the smallness of the initial velocity and force are small. We obtain the following uniform time estimation.

Proposition 27 *The solution of (34), (35), (36) according to Theorem 25 satisfies:*

$$\tilde{q} = \tilde{q}_0 + \int_0^t e^{-\nu|k|^2(t-\tau)} ([(q, \tilde{\nabla})q] + \tilde{F}) d\tau, \quad (40)$$

where $F = -\nabla p + f$.

This follows from the definition of the Fourier transform and the theory of linear differential equations.

Proposition 28 *The solution of (34), (35), (36) satisfies:*

$$\tilde{p} = \sum_{i,j} \frac{k_i k_j}{|k|^2} q_i \tilde{q}_j + i \sum_i \frac{k_i}{|k|^2} \tilde{f}_i \quad (41)$$

and the following estimations:

$$\|p\|_{L_2(R^3)} \leq 3 \|\nabla q\|_{L_2(R^3)}^{\frac{3}{2}} \|q\|_{L_2(R^3)}^{\frac{1}{2}}, \quad (42)$$

$$|\nabla \tilde{p}| \leq \frac{|\tilde{q}^2|}{|k|} + \frac{|\tilde{f}|}{|k|^2} + \frac{1}{|k|} |\nabla \tilde{f}| + 3 |\nabla \tilde{q}^2|. \quad (43)$$

This expression for p is obtained using *div* and the Fourier transform presentation.

Lemma 29 *Let*

$$Qq_0 \in W_2^1(R^3), \quad Qf \in L_2(Q_T).$$

Then, the solution of (34), (35), (36) in Theorem 25 satisfies the following inequalities:

$$\begin{aligned} \sup_{0 \leq t \leq T} \|NQq\|_{L_2(R^3)}^2 + \int_0^t \|k^2 NQq\|_{L_2(R^3)}^2 d\tau &\leq \\ &\leq \|NQq_0\|_{L_2(R^3)}^2 + \|Qf\|_{L_2(Q_T)}. \end{aligned} \quad (44)$$

Proof this follows from the a priori estimation of Lemma 26 and conditions of Lemma 29.

Lemma 30 *Let $Qq_0 \in W_2^1(R^3)$, $f \in L_2(Q_T)$. Then, the solution of (34), (35), (36) in Theorem 25 satisfies 2 the following inequalities:*

$$\begin{aligned} \sup_{\theta} \sup_{0 \leq t \leq T} [\|Qq\|_{L_2(R^3)}^2 + \int_0^t \|k^2 Qq\|_{L_2(R^3)}^2 d\tau] &\leq \\ &\leq \sup_{\theta} [\|Qq_0\|_{L_2(R^3)}^2 + \|f\|_{L_2(Q_T)}]. \end{aligned} \quad (45)$$

Proof this follows from the a priori estimation of Lemma 26 and conditions of Lemma 30

Lemma 31 *The solution of (34), (35), (36) in Theorem 25 satisfies the following inequalities:*

$$\int_{R^3} |x|^2 |q|^2 dx + \int_0^t \int_{R^3} |x|^2 |\nabla q|^2 dx d\tau \leq const,$$

$$\int_{R^3} |x|^4 |q|^2 dx + \int_0^t \int_{R^3} |x|^4 |\nabla q|^2 dx d\tau \leq const \quad (46)$$

or

$$\|\nabla \tilde{q}\|_{L_2(R^3)} + \int_0^t \int_{R^3} |k|^2 |\tilde{\nabla} q|^2 dk d\tau \leq const,$$

$$\|\nabla^2 \tilde{q}\|_{L_2(R^3)} + \int_0^t \int_{R^3} |k|^2 |\tilde{\nabla}^2 q|^2 dk d\tau \leq const. \quad (47)$$

Proof this follows from the a priori estimation of Lemma 26, conditions of Lemma 31, the Navier–Stokes equations.

Lemma 32 *The solution of (34), (35), (36) satisfies the following inequalities:*

$$\max_k |\tilde{q}| \leq \max_k |\tilde{q}_0| + \frac{T}{2} \sup_{0 \leq t \leq T} \|q\|_{L_2(R^3)}^2 + \int_0^t \|\nabla q\|_{L_2(R^3)}^2 d\tau, \quad (48)$$

$$\max_k |\nabla \tilde{q}| \leq \max_k |\nabla \tilde{q}_0| + \frac{T}{2} \sup_{0 \leq t \leq T} \|\nabla \tilde{q}\|_{L_2(R^3)} + \int_0^t \int_{R^3} |k|^2 |\tilde{\nabla} q|^2 dk d\tau, \quad (49)$$

$$\max_k |\nabla^2 \tilde{q}| \leq \max_k |\nabla^2 \tilde{q}_0| + \frac{T}{2} \sup_{0 \leq t \leq T} \|\nabla^2 \tilde{q}\|_{L_2(R^3)} + \int_0^t \int_{R^3} |k|^2 |\nabla^2 \tilde{q}|^2 dk d\tau. \quad (50)$$

Proof this follows from the a priori estimation of Lemma 26, conditions of Lemma 32, the Navier–Stokes equations.

Lemma 33 *The solution of (34), (35), (36) according to Theorem 25 satisfies $C_i \leq const$, ($i = 0, 2, 4$), where:*

$$C_0 = \int_0^t |\tilde{F}_1|^2 d\tau, \quad F_1 = (q, \nabla)q + F,$$

$$C_2 = \int_0^t |\nabla \tilde{F}_1|^2 d\tau, \quad C_4 = \int_0^t |\nabla^2 \tilde{F}_1|^2 d\tau. \quad (51)$$

Proof this follows from the a priori estimation of Lemma 26, the Navier–Stokes equations.

Lemma 34 *Weak solution of problem (34), (35), (36) from Theorem 25 satisfies the following inequalities*

$$|NQq| \leq zM_1, \quad \left| \frac{\partial NQq}{\partial z} \right| \leq zM_2, \quad \left| \frac{\partial^2 NQq}{\partial z^2} \right| \leq zM_3,$$

where M_1, M_2, M_3 are limited.

Let us prove the first estimate. These inequalities

$$|Qq(z, t)| \leq \frac{z}{2} \int_0^\pi \int_0^{2\pi} |\tilde{q}(z(e_k - e_p), t)| de_p \leq \leq 2\pi z \max_k |\tilde{q}| \leq zM_1,$$

where $M_1 = const$.

Proof now this follows from the a priori estimation of Lemma 26, conditions of Lemma 34, the Navier–Stokes equations.

The rest of estimates are proved similarly.

Lemma 35 *Suppose that $q \in R$, $\max_k |\tilde{q}| < \infty$, then*

$$\int_{R^3} \int_{R^3} \frac{q(x)q(y)}{|x - y|^2} dx dy \leq C(|q|_{L_2} + \max_k |\tilde{q}|)^2.$$

Proof: Using Plancherel’s theorem, we get the statement of the lemma. This proves Lemma 35. \square

Lemma 36 *Weak solution of problem (34), (35), (36) from Theorem 25 satisfies the following inequalities*

$$|Qq| \leq |Qq_0| + \left(\frac{1}{2\nu}\right)^{\frac{1}{2}} \frac{C_0^{\frac{1}{2}}}{z|e_k - e_\lambda|}, \quad (52)$$

where $C_0 = \int_0^t |\tilde{F}_1|^2 d\tau, F_1 = (q, \nabla)q + F$.

Proof: From (40) we get

$$|Qq| \leq |Qq_0| + \left| \int_0^t e^{-\nu z^2 |e_k - e_\lambda|^2 (t-\tau)} \tilde{F}_1(z(e_k - e_\lambda), \tau) d\tau \right|, \quad (53)$$

where $F_1 = (q, \nabla)q + F$. Using the denotation

$$I = \left| \int_0^t e^{-\nu z^2 |e_k - e_\lambda|^2 (t-\tau)} \tilde{F}_1(z(e_k - e_\lambda), \tau) d\tau \right|,$$

taking into account Holder's inequality in I we obtain

$$I \leq \left(\int_0^t |e^{-\nu z^2 |e_k - e_\lambda|^2 (t-\tau)}|^p d\tau \right)^{\frac{1}{p}} \left(\int_0^t |F_1|^q d\tau \right)^{\frac{1}{q}},$$

where p, q satisfies the equality $\frac{1}{p} + \frac{1}{q} = 1$. Suppose $p = q = 2$. Then

$$I \leq \left(\frac{1}{2\nu} \right)^{\frac{1}{2}} \frac{\left(\int_0^t |\tilde{F}_1|^2 d\tau \right)^{\frac{1}{2}}}{z|e_k - e_\lambda|}.$$

Taking into consideration the estimate I in (53), we obtain the statement of the lemma.

This proves Lemma 36. \square

Lemma 37 *Weak solution of problem (34), (35), (36), from Theorem 25 satisfies the following inequalities*

$$\left| \frac{\partial Qq}{\partial z} \right| \leq \left| \frac{\partial Qq_0}{\partial z} \right| + 4\alpha \left(\frac{1}{\nu} \right)^{\frac{1}{2}} \frac{C_0^{\frac{1}{2}}}{z^2 |e_k - e_\lambda|} + \left(\frac{1}{2\nu} \right)^{\frac{1}{2}} \frac{C_2^{\frac{1}{2}}}{z |e_k - e_\lambda|}, \quad (54)$$

where

$$C_2 = \int_0^t \left| \frac{\partial \tilde{F}_1}{\partial z} \right|^2 d\tau.$$

Proof: The underwritten inequalities follows from representation (40)

$$\left| \frac{\partial Qq}{\partial z} \right| \leq \left| \frac{\partial Qq_0}{\partial z} \right| + 2\nu z |e_k - e_\lambda|^2 \times \left| \int_0^t (t - \tau) e^{-\nu z^2 |e_k - e_\lambda|^2 (t-\tau)} \tilde{F}_1(z(e_k - e_\lambda), \tau) d\tau \right| + \left| \int_0^t e^{-\nu z^2 |e_k - e_\lambda|^2 (t-\tau)} \frac{\partial \tilde{F}_1}{\partial z}(z(e_k - e_\lambda), \tau) d\tau \right|.$$

Let us introduce the following denotation

$$I_1 = 2\nu z |e_k - e_\lambda|^2 \times \left| \int_0^t (t - \tau) e^{-\nu z^2 |e_k - e_\lambda|^2 (t-\tau)} \tilde{F}_1(z(e_k - e_\lambda), \tau) d\tau \right|,$$

$$I_2 = \left| \int_0^t e^{-\nu z^2 |e_k - e_\lambda|^2 (t-\tau)} \frac{\partial \tilde{F}_1}{\partial z}(z(e_k - e_\lambda), \tau) d\tau \right|,$$

then

$$\left| \frac{\partial Qq}{\partial z} \right| \leq \left| \frac{\partial Qq_0}{\partial z} \right| + I_1 + I_2.$$

Estimate I_1 by means of

$$\sup_t |t^m e^{-t}| < \alpha,$$

where $m > 0$ we obtain

$$I_1 \leq \frac{4\alpha}{z} \left| \int_0^t e^{-\nu z^2 |e_k - e_\lambda|^2 \frac{t-\tau}{2}} \tilde{F}_1(z(e_k - e_\lambda), \tau) d\tau \right|.$$

On applying Holder's inequality, we get

$$I_1 \leq \frac{4\alpha}{z} \left(\int_0^t |e^{-\nu z^2 |e_k - e_\lambda|^2 \frac{t-\tau}{2}}|^p d\tau \right)^{\frac{1}{p}} \times \left(\int_0^t |F_1|^q d\tau \right)^{\frac{1}{q}},$$

where p, q satisfy the equality $\frac{1}{p} + \frac{1}{q} = 1$.

For $p = q = 2$ we have

$$I_1 \leq 4\alpha \left(\frac{1}{\nu} \right)^{\frac{1}{2}} \frac{C_0^{\frac{1}{2}}}{z^2 |e_k - e_\lambda|},$$

$$I_2 \leq \left(\frac{1}{2\nu} \right)^{\frac{1}{2}} \frac{C_2^{\frac{1}{2}}}{z |e_k - e_\lambda|}, \quad C_2 = \int_0^t \left| \frac{\partial \tilde{F}_1}{\partial z} \right|^2 d\tau.$$

Inserting I_1, I_2 in to $\left| \frac{\partial \tilde{q}}{\partial z} \right|$, we obtain the statement of the lemma. This completes the proof of Lemma 37. \square

Lemma 38 *Weak solution of problem (34), (35), (36), from Theorem 25 satisfies the following inequalities*

$$\begin{aligned} |NQq| &\leq C, \quad |TNQq| \leq C, \quad |Qq| \leq C, \\ |TQq| &\leq C, \quad |NQEq| \leq C, \quad |TNQEq| \leq C, \\ |QEq| &\leq C, \quad |TQEq| \leq C. \end{aligned} \quad (55)$$

Lemma 39 Let $q \in R$, $\max_k |\tilde{q}| < \infty$. Then,

$$N(q) \leq \int_{R^3} \int_{R^3} \frac{q(x)q(y)}{|x-y|^2} dx dy \leq C(|q|_{L_2} + \max_k |\tilde{q}|)^2. \quad (56)$$

A proof of this lemma can be obtained using Plancherels theorem.

We now obtain uniform time estimations for Rollnik’s norms of the solutions of (34), (35), (36). The following (and main) goal is to obtain the same estimations for $\max_x |q|$ – velocity components of the Cauchy problem for the Navier–Stokes equations.

Let’s consider the influence of the following large scale transformations in Navier-Stokes’ equation on

$$K = \frac{\nu^{\frac{1}{2}}}{\nu^{\frac{1}{2}} - 4\pi CC_0^{\frac{1}{2}}}. t' = tA, \nu' = \frac{\nu}{A}, F'_0 = \frac{F_0}{A^2}.$$

Proposition 40 Let

$$A = \frac{4}{\nu^{\frac{1}{3}}(CC_0 + 1)^{\frac{2}{3}}}, \text{ then } K \leq \frac{8}{7}.$$

Proof: By the definitions C and C_0 , we have

$$K = \left(\frac{\nu}{A}\right)^{\frac{1}{2}} \left(\left(\frac{\nu}{A}\right)^{\frac{1}{2}} - \frac{4\pi CC_0}{A^2}\right)^{-1} = \nu^{\frac{1}{2}} \left(\nu^{\frac{1}{2}} - \frac{4\pi CC_0}{A^{\frac{3}{2}}}\right)^{-1} < \frac{8}{7}.$$

This proves Proposition. □

Theorem 41 Let

$$q_0 \in W_2^2(R^3), \nabla^2 \tilde{q}_0 \in L_2(R^3), f \in L_2(Q_T),$$

$$\tilde{f} \in L_1(Q_T) \cap L_2(R^3), \nabla^2 \tilde{f} \in L_1(Q_T) \cap L_2(R^3)$$

and

$$\max_k \|Qq_0\|_{L_2} < const, \max_k \|Qf\|_{L_2} < const,$$

$$\max_k \|QEq_0\|_{L_2} < const, \max_k \|QEf\|_{L_2} < const.$$

Then, there ex satisfying the folloists a unique generalized solution of (34), (35), (36) wing inequality:

$$\max_t \sum_{i=1}^3 \max_x |q_i| \leq const, \text{ where the value of const depends only on the conditions of the theorem.}$$

Proof: It suffices to obtain uniform estimates of the maximum velocity components q_i , which obviously follow from $\max_x |q_i|$, because uniform estimates allow us to extend the local existence and uniqueness theorem over the interval in which they are valid. To estimate the velocity components, Lemma 30 can be used:

$$v_i = q_i / (\int_0^T \|q_x\|_{L_2(R^3)}^2 dt + A_0 + 1),$$

$$A_0 = 4/(\nu^{\frac{1}{3}}(CC_0 + 1)^{\frac{2}{3}}).$$

Using Lemmas (36)–(39) for

$$v_i = q_i / (\int_0^T \|q_x\|_{L_2(R^3)}^2 dt + A_0 + 1),$$

we can obtain $\|A_i\|_{TA} < \alpha < 1$, where A_i is the amplitude of potential q_i and $N(q_i) < 1$. That is, discrete solutions are not significant in proving the theorem, so its assertion follows the conditions of Theorem 41, which defines uniform time estimations for the maximum values of velocity components.

$$\|\nabla q\|_{L_2(R^3)} + \int_0^t \int_{R^3} |\nabla q|^2 dk d\tau \leq const +$$

$$+ \max |q| \int_0^t \|\nabla q\|_{L_2(R^3)} \|\nabla^2 q\|_{L_2(R^3)} d\tau. \quad (57)$$

□

Theorem 41 asserts the global solvability and uniqueness of the Cauchy problem for the Navier–Stokes equations.

Theorem 42 Let

$$q_0 \in W_2^2(R^3), \nabla^2 \tilde{q}_0 \in L_2(R^3),$$

$$f \in L_2(Q_T), \tilde{f} \in L_1(Q_T) \cap L_2(R^3),$$

$$\lim_{t \rightarrow t_0} \|\nabla q\|_{L_2(R^3)} = \infty. \quad (58)$$

Then, there exists i, j, x_0

$$\lim_{t \rightarrow t_0} \psi_j(x_0, t) = \infty \text{ or } \lim_{t \rightarrow t_0} N(q_i) = \infty \quad (59)$$

Proof: A proof of this lemma can be obtained using $q_i = P_{Ac}q_i + P_Dq_i$ and uniform estimates $P_{Ac}q_i$. □

Theorem 42 Describes the loss of smoothness of classical solutions for the Navier–Stokes equations. Theorem 42 describes the time blow up of the classical solutions for the Navier-Stokes equations arises, and complements the results of Terence Tao [20].

4 Conclusions

New uniform global estimations of solutions of the Navier–Stokes equations indicate that the principle modeling of complex flows and related calculations can be based on the Fourier transform method.

Acknowledgements: We are grateful to the Ministry of Education and Science of the Republic of Kazakhstan for a grant, and to the System Research "Factor" Company for combining our efforts in this project. The work was performed as part of an international project, "Joint Kazakh-Indian studies of the influence of anthropogenic factors on atmospheric phenomena on the basis of numerical weather prediction models WRF (Weather Research and Forecasting)", commissioned by the Ministry of Education and Science of the Republic of Kazakhstan.

References:

- [1] Charles L. Fefferman, Existence and Smoothness of the Navier-Stokes Equation. The Millennium Prize Problems, *Clay Math. Inst.* Cambridge, MA, 2006, pp. 57–67.
- [2] A.A. Durmagambetov and L.S. Fazilova, Global Estimation of the Cauchy Problem Solutions Fourier Transform Derivatives for the Navier-Stokes Equation, *International Journal of Modern Nonlinear Theory and Application*, Online at Scientific Research Publishing, 2, No.4, December, 2013.
- [3] A.A. Durmagambetov and L.S. Fazilova, Global Estimation of the Cauchy Problem Solutions the Navier-Stokes Equation, *Journal of Applied Mathematics and Physics*, Online at Scientific Research Publishing, 2, No.4, March, 2014.
- [4] A.A. Durmagambetov and L.S. Fazilova, Existence and Blowup Behavior of Global Strong Solutions Navier-Stokes, *International Journal of Engineering Science and Innovative Technology*, 3, Issue 3, May, 2014, pp. 679–687.
- [5] J.S. Russell, *Report on Waves: Report of the fourteenth meeting of the British Association for the Advancement of Science*, York, September, 1844 (London 1845), Plates XLVII-LVII, pp. 311–390.
- [6] J.S. Russell, *Report of the committee on waves, Report of the 7th Meeting of British Association for the Advancement of Science*, John Murray, London, 1838, pp. 417–496.
- [7] Mark J. Ablowitz and Harvey Segur, *Solitons and the Inverse Scattering Transform*, SIAM, 1981, p. 435.
- [8] N.J. Zabusky and M.D. Kruskal, Interaction of solitons in a collisionless plasma and the recurrence of initial states, *Phys.Rev.Lett.* 15, 1965, pp. 240–243.
- [9] L.D. Faddeev, The inverse problem in the quantum theory of scattering. II, *Itogi Nauki i Tekhniki. Ser. Sovrem. Probl. Mat.* 3, VINITI, Moscow, 1974, pp. 93-180.
- [10] R.G. Newton, New result on the inverse scattering problem in three dimensions, *Phys. rev. Lett.* vol. 43, 8, 1979, pp. 541–542.
- [11] R.G. Newton, Inverse scattering Three dimensions, *Jour. Math. Phys.* 21, 1980, pp. 1698–1715.
- [12] E. Somersalo et al., *Inverse scattering problem for the Schrodinger's equation in three dimensions: connections between exact and approximate methods*, 1988.
- [13] A.Y. Povzner, On the expansion of arbitrary functions in characteristic functions of the operator $-\Delta u + cu$, *Mat. Sb. (N.S.)* 32(74):1, 1953, pp. 109-156.
- [14] M.C. Birman, On the spectrum of singular boundary-value problems, *Mat. Sb. (N.S.)* 55(97), 1961, pp. 125-174.
- [15] H. Poincare, *Lecons de mecanique celeste*, 3, 1910.
- [16] J. Leray, "Sur le mouvement d'un liquide visqueux emplissant l'espace", *Acta Mathematica* 63: 193248. doi:10.1007/BF02547354, 1934.
- [17] O.A. Ladyzhenskaya, *Mathematic problems of viscous incompressible liquid dynamics*, Moscow, Science, 1970, p. 288.
- [18] V.A. Solonnikov, Estimates solving nonstationary linearized systems of Navier-Stokes' Equations, *Transactions Academy of Sciences USSR* 70, 1964, pp. 213–317.
- [19] Huang Xiangdi, Li Jing and Wang Yong, Serrin-Type Blowup Criterion for Full Compressible Navier-Stokes System, *Archive for Rational Mechanics and Analysis*, 2013, pp. 303–316.
- [20] Terence Tao, Finite time blowup for an averaged three-dimensional Navier-Stokes equation, *-arXiv:1402.0290 [math.AP]*.

On a Subclass of p -valent Starlike Functions Associated with a Generalized Hypergeometric Differential operator

Entisar El-Yagubi¹, Maslina Darus² and Melike Aydoğan³

Abstract—The object of the present paper is to introduce a new subclass of p -valent starlike functions with negative coefficients in the open unit disc which is defined by a generalized derivative operator $\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)$. We obtain coefficient inequalities, growth and distortion theorems, and extreme points for the subclass of p -valent functions.

Keywords— p -valent functions, starlike functions, derivative operator.

I. INTRODUCTION

HYPERGEOMETRIC functions theory found to be of common interests in the real case. However, it started to be flourish among the complex analysts ever since de Branges [?] used it to prove the Bieberbach conjecture. Thereon, the theory of hypergeometric functions becomes the favourite topics of discussion among the mathematicians. Many interesting subclasses of analytic functions associated with the generalized hypergeometric functions have been investigated and studied by many researchers, for example, Kumar et al. [?], Gangadharan et al. [?], Liu [?], El-Ashwah [?] and of course many others. In this paper, we shall use the generalized hypergeometric functions to define a new derivative operator. Moreover, we investigate some interesting properties on a subclass of p -valent starlike functions with negative coefficients.

Let \mathcal{A}_p denote the class of functions of the form

$$f(z) = z^p + \sum_{n=p+1}^{\infty} a_n z^n, \quad (z \in \mathbb{U}, p \in \mathbb{N}), \quad (1)$$

which are p -valent functions in the open unit disc \mathbb{U} .

Also let \mathcal{T}_p denote a subclass of \mathcal{A}_p consisting of p -valent functions which can be expressed in the form

$$f(z) = z^p - \sum_{n=p+1}^{\infty} |a_n| z^n, \quad (z \in \mathbb{U}, p \in \mathbb{N}). \quad (2)$$

If $f(z)$ and $g(z)$ belong to \mathcal{A}_p , then the Hadamard product $f * g$ is defined by

$$f(z) * g(z) = z^p + \sum_{n=p+1}^{\infty} a_n b_n z^n, \quad p \in \mathbb{N}.$$

E. El-Yagubi and M. Darus are School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, e-mail: maslina@ukm.edu.my .

M. Aydoğan is with Işık University.

Dziok and Srivastava [?] studied the following p -valent function, which defined by generalized hypergeometric functions

$$H_p(a_i, b_q) = z^p + \sum_{n=p+1}^{\infty} \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p}} \cdot \frac{z^n}{(n-p)!}, \quad p \in \mathbb{N},$$

where $a_i \in \mathbb{C}$, $b_q \in \mathbb{C} \setminus \{0, -1, -2, \dots\}$, ($i = 1, \dots, r, q = 1, \dots, s$), and $r \leq s + 1$; $r, s \in \mathbb{N}_0$, and $(x)_n$ is the Pochhammer symbol defined by

$$(x)_n = \frac{\Gamma(x+n)}{\Gamma(x)} = \begin{cases} 1, & n = 0, \\ x(x+1) \cdots (x+n-1), & n = \{1, 2, 3, \dots\}. \end{cases}$$

Let $\mathcal{L}_{\lambda_1, \lambda_2, p}^{m, b} \in \mathcal{A}_p$ is defined by

$$\mathcal{L}_{\lambda_1, \lambda_2, p}^{m, b} = z^p + \sum_{n=p+1}^{\infty} \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m z^n, \quad p \in \mathbb{N},$$

where $m, b \in \mathbb{N}_0 = \mathbb{N} \cup \{0\}$, $\lambda_2 \geq \lambda_1 \geq 0$. Corresponding to $H_p(a_i, b_q)$, $\mathcal{L}_{\lambda_1, \lambda_2, p}^{m, b}$ and using the Hadamard product, we define a new generalized differential operator $D_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)$ as follows:

Definition 1.1 Let $f \in \mathcal{A}_p$, then a generalized differential operator $D_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z) : \mathcal{A}_p \rightarrow \mathcal{A}_p$ is given as

$$\begin{aligned} D_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z) &= (H_p(a_i, b_q) * \mathcal{L}_{\lambda_1, \lambda_2, p}^{m, b}) * f(z) \\ &= z^p + \sum_{n=p+1}^{\infty} \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p}} \cdot \frac{a_n z^n}{(n-p)!}. \end{aligned} \quad (3)$$

It follows from the above definition that

$$\begin{aligned} &(p + \lambda_2(n-p) + b)D_{\lambda_1, \lambda_2, p}^{m+1, b}(a_i, b_q)f(z) \\ &= (p + \lambda_2(n-p) - p\lambda_1 + b)D_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z) \\ &\quad + \lambda_1 z(D_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z))'. \end{aligned} \quad (4)$$

Remark 1.1 It should be remarked that the linear operator $\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z)$ is a generalization of many operators considered earlier. Let us see some of the examples:

- For $\lambda_2 = b = 0$, the operator $\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f$ reduces to the operator was given by Selvaraj and Karthikeyan [?].
- For $m = 0$, the operator $\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f$ reduces to the operator was given by Dziok and Srivastava [?].
- For $\lambda_2 = b = 0$ and $p = 1$, we get the operator studied by Selvaraj and Karthikeyan [?].
- For $m = 0, r = 2, s = 1$ and $p = 1$, we obtain the operator which was given by Hohlov [?].
- For $r = 1, s = 0, a_1 = 1, \lambda_1 = 1, \lambda_2 = b = 0$ and $p = 1$, we get the Sălăgean derivative operator [?].
- For $r = 1, s = 0, a_1 = 1, \lambda_2 = b = 0$ and $p = 1$, we get the generalized Sălăgean derivative operator introduced by Al-Oboudi [?].
- For $m = 0, r = 1, s = 0, a_1 = \delta + 1$ and $p = 1$, we obtain the operator introduced by Ruscheweyh [?].
- For $r = 1, s = 0, a_1 = \delta + 1$ and $p = 1$, we obtain the operator studied by El-Yagubi and Darus [?].
- For $m = 0, r = 2$ and $s = 1, a_2 = 1$ and $p = 1$, we obtain the operator studied by Carlson and Shaffer [?].
- For $r = 1, s = 0, a_1 = 1, \lambda_2 = 0$ and $p = 1$, we get the operator introduced by Cátás [?].

By making use of the generalized derivative operator $\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z)$, we introduce a new subclass as follows:

Definition 1.2 For f defined by (1), $0 \leq \beta < 1, \lambda_2 \geq \lambda_1 \geq 0, m, b \in \mathbb{N}_0 = \{0, 1, 2, \dots\}, a_i \in \mathbb{C}, b_q \in \mathbb{C} \setminus \{0, -1, -2, \dots\}, (i = 1, \dots, r, q = 1, \dots, s)$, and $r \leq s + 1; r, s \in \mathbb{N}_0$, let $\mathcal{S}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta)$ be the subclass of \mathcal{A}_p consisting of functions f which satisfy

$$\Re \left(\frac{z(\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z))'}{\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z)} \right) > \beta, \quad (5)$$

where $p \in \mathbb{N}$ and $\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z) \neq 0$.

Note that if f given by (2), then we can see that

$$\begin{aligned} \mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z) = & z^p - \sum_{n=p+1}^{\infty} \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \\ & \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p}} \frac{|a_n| z^n}{(n-p)!}, \end{aligned} \quad (6)$$

In addition, we define the class $\mathcal{TS}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta)$ by

$$\mathcal{TS}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta) = \mathcal{S}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta) \cap \mathcal{T}_p. \quad (7)$$

Note that the subclass $\mathcal{TS}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta)$ has been studied by many authors. For example:

when $m = 0, r = 1, s = 0, a_1 = 1, p = 1$, then $\mathcal{TS}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta) \equiv \mathcal{S}_{\mathcal{T}}^*(\beta)$, when $m = 1, \lambda_1 = b = 0, r = 1, s = 0, a_1 = 1, p = 1$, then $\mathcal{TM}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta) \equiv \mathcal{CT}(\beta)$, where the classes $\mathcal{S}_{\mathcal{T}}^*(\beta)$ and $\mathcal{CT}(\beta)$ were studied by Silverman [?].

II. COEFFICIENT INEQUALITIES

We provide a sufficient condition for p -valent functions f in \mathbb{U} , to be in $\mathcal{TS}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta)$.

Theorem 2.1 Let f be defined by (2), then $f \in \mathcal{TS}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta)$ if and only if

$$\sum_{n=p+1}^{\infty} (n-p) \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p} (n-p)!} |a_n| \leq p - \beta, \quad (8)$$

where $0 \leq \beta < 1, \lambda_2 \geq \lambda_1 \geq 0, m, b \in \mathbb{N}_0, p \in \mathbb{N}, a_i \in \mathbb{C}, b_q \in \mathbb{C} \setminus \{0, -1, -2, \dots\}, (i = 1, \dots, r, q = 1, \dots, s)$, and $r \leq s + 1; r, s \in \mathbb{N}_0$.

Proof. Suppose that (8) holds true. It suffices to show that the values $\frac{z(\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z))'}{\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z)}$ lie in a circle centered at $w = p$ with radius $p - \beta$, which is

$$\left| \frac{z(\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z))'}{\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z)} - p \right| \leq p - \beta. \quad (9)$$

So, we can write

$$\begin{aligned} & \left| \frac{z(\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z))'}{\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z)} - p \right| \\ &= \left| \frac{z(\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z))' - p\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z)}{\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z)} \right| \end{aligned}$$

$$= \left| \frac{\sum_{n=p+1}^{\infty} (n-p) \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p} (n-p)!} a_n z^n}{z^p - \sum_{n=p+1}^{\infty} \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p} (n-p)!} a_n z^n} \right|, \quad |z| < 1,$$

$$\begin{aligned} & \sum_{n=p+1}^{\infty} (n-p) \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \\ & \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p} (n-p)!} |a_n| \\ & \leq \frac{\sum_{n=p+1}^{\infty} (n-p) \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p} (n-p)!} |a_n|}{1 - \sum_{n=p+1}^{\infty} \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p} (n-p)!} |a_n|}. \end{aligned} \quad (10)$$

This last expression (10) is bounded by $(p - \beta)$ if the following inequality which is equivalent to (9) holds.

$$\sum_{n=p+1}^{\infty} (n-p) \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p}(n-p)!} |a_n|$$

$$\leq (p - \beta) \left[1 - \sum_{n=p+1}^{\infty} \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p}(n-p)!} |a_n| \right], \quad (11)$$

which is equivalent to

$$\sum_{n=p+1}^{\infty} (n - \beta) \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p}(n-p)!} |a_n| \leq (p - \beta),$$

by using (8). Thus $f \in \mathcal{TS}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta)$. In order to prove the sufficiency, assume that $f \in \mathcal{TS}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta)$ and by condition (5), we have

$$\Re \left(\frac{z(\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z))'}{\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z)} \right)$$

$$= \Re \left\{ \frac{pz^p - \sum_{n=p+1}^{\infty} n \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p}(n-p)!} |a_n| z^n}{z^p - \sum_{n=p+1}^{\infty} \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p}(n-p)!} |a_n| z^n} \right\} > \beta. \quad (12)$$

Choose values of z on real axis so that $\frac{z(\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z))'}{\mathcal{D}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q)f(z)}$ is real. Letting $z \rightarrow 1^-$ through real axis, we have

$$\frac{p - \sum_{n=p+1}^{\infty} n \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p}(n-p)!} |a_n|}{1 - \sum_{n=p+1}^{\infty} \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p}(n-p)!} |a_n|} > \beta. \quad (13)$$

Thus we obtain

$$\sum_{n=p+1}^{\infty} (n - \beta) \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p}(n-p)!} |a_n| \leq p - \beta,$$

which is (8). Hence the proof is complete.

Corollary 2.1 Let f be defined by (2) be in the class $\mathcal{TS}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta)$, then we have

$$|a_n| \leq \frac{p - \beta}{n - \beta} \left[\frac{p + \lambda_2(n-p) + b}{p + (\lambda_1 + \lambda_2)(n-p) + b} \right]^m \cdot \frac{(b_1)_{n-p} \cdots (b_s)_{n-p} (n-p)!}{(a_1)_{n-p} \cdots (a_r)_{n-p}}. \quad (14)$$

The result (14) is sharp with f given by

$$f(z) = z^p - \frac{p - \beta}{n - \beta} \left[\frac{p + \lambda_2(n-p) + b}{p + (\lambda_1 + \lambda_2)(n-p) + b} \right]^m \cdot \frac{(b_1)_{n-p} \cdots (b_s)_{n-p} (n-p)!}{(a_1)_{n-p} \cdots (a_r)_{n-p}} z^n. \quad (15)$$

III. GROWTH AND DISTORTION THEOREMS

We obtain growth and distortion bounds for $f \in \mathcal{TS}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta)$ as follows:

Theorem 3.1 Let f defined by (2) be in the class $\mathcal{TS}_{\lambda_1, \lambda_2, p}^{m, b}(a_i, b_q, \beta)$, then for $0 < |z| = r < 1$, we have

$$r^p - \frac{p - \beta}{p + 1 - \beta} \cdot \left[\frac{p + \lambda_2 + b}{p + \lambda_1 + \lambda_2 + b} \right]^m \cdot \frac{(b_1) \cdots (b_s)}{(a_1) \cdots (a_r)} r^{p+1} \leq$$

$$|f(z)| \leq r^p + \frac{p - \beta}{p + 1 - \beta} \cdot \left[\frac{p + \lambda_2 + b}{p + \lambda_1 + \lambda_2 + b} \right]^m \cdot \frac{(b_1) \cdots (b_s)}{(a_1) \cdots (a_r)} r^{p+1}. \quad (16)$$

and

$$pr^{p-1} - \frac{(p+1)(p-\beta)}{p+1-\beta} \cdot \left[\frac{p + \lambda_2 + b}{p + \lambda_1 + \lambda_2 + b} \right]^m \cdot \frac{(b_1) \cdots (b_s)}{(a_1) \cdots (a_r)} r^p$$

$$\leq |f'(z)|$$

$$\leq pr^{p-1} + \frac{(p+1)(p-\beta)}{p+1-\beta} \cdot \left[\frac{p + \lambda_2 + b}{p + \lambda_1 + \lambda_2 + b} \right]^m \cdot \frac{(b_1) \cdots (b_s)}{(a_1) \cdots (a_r)} r^p. \quad (17)$$

where $0 \leq \beta < 1, \lambda_2 \geq \lambda_1 \geq 0$ and $m, b \in \mathbb{N}_0, a_i \in \mathbb{C}, b_q \in \mathbb{C} \setminus \{0, -1, -2, \dots\}, (i = 1, \dots, r, q = 1, \dots, s), r \leq s + 1; r, s \in \mathbb{N}_0$ and $p \in \mathbb{N}$.

Proof. Since $f \in \mathcal{TM}_{\lambda_1, \lambda_2}^{m, b}(a_i, b_q, \beta)$ by Theorem 2.1, we have

$$\sum_{n=p+1}^{\infty} (n - \beta) \left[\frac{p + (\lambda_1 + \lambda_2)(n-p) + b}{p + \lambda_2(n-p) + b} \right]^m \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p}(n-p)!} |a_n| \leq p - \beta.$$

Now

$$(p+1-\beta) \left[\frac{p + \lambda_1 + \lambda_2 + b}{p + \lambda_2 + b} \right]^m \cdot \frac{(a_1) \cdots (a_r)}{(b_1) \cdots (b_s)} \left(\sum_{n=p+1}^{\infty} |a_n| \right)$$

$$\begin{aligned}
 &= \sum_{n=p+1}^{\infty} (p+1-\beta) \left[\frac{p + \lambda_1 + \lambda_2 + b}{p + \lambda_2 + b} \right]^m \cdot \frac{(a_1) \cdots (a_r)}{(b_1) \cdots (b_s)} |a_n| \\
 &\leq \sum_{n=p+1}^{\infty} (n - \beta) \left[\frac{p + (\lambda_1 + \lambda_2)(n - p) + b}{p + \lambda_2(n - p) + b} \right]^m \\
 &\quad \cdot \frac{(a_1)_{n-p} \cdots (a_r)_{n-p}}{(b_1)_{n-p} \cdots (b_s)_{n-p} (n - p)!} |a_n| \leq p - \beta, \quad p \in \mathbb{N}.
 \end{aligned}$$

Therefore,

$$\sum_{n=p+1}^{\infty} |a_n| \leq \frac{p - \beta}{p + 1 - \beta} \left[\frac{p + \lambda_2 + b}{p + \lambda_1 + \lambda_2 + b} \right]^m \cdot \frac{(b_1) \cdots (b_s)}{(a_1) \cdots (a_r)}. \tag{18}$$

Since

$$f(z) = z^p - \sum_{n=p+1}^{\infty} |a_n| z^n,$$

then we get

$$|f(z)| = \left| z^p - \sum_{n=p+1}^{\infty} |a_n| z^n \right|.$$

Next,

$$\begin{aligned}
 |z|^p - |z|^{p+1} \sum_{n=p+1}^{\infty} |a_n| &\leq |f(z)| \\
 &\leq |z|^p + |z|^{p+1} \sum_{n=p+1}^{\infty} |a_n|,
 \end{aligned}$$

that is

$$r^p - r^{p+1} \sum_{n=p+1}^{\infty} |a_n| \leq |f(z)| \leq r^p + r^{p+1} \sum_{n=p+1}^{\infty} |a_n|.$$

By using the inequality (18), we get the result (16).

Furthermore, we observe that

$$\begin{aligned}
 pr^{p-1} - (p + 1)r^p \sum_{n=p+1}^{\infty} |a_n| &\leq |f'(z)| \\
 &\leq pr^{p-1} + (p + 1)r^p \sum_{n=p+1}^{\infty} |a_n|.
 \end{aligned}$$

By using (18), we easily arrive to the result (17).

IV. ACKNOWLEDGEMENTS

The work presented here was partially supported by AP-2013-009 and by Isik University Scientific Research Funding Agency under Grant Number: BAP-14B102.

REFERENCES

[1] J. Dziok and H. M. Srivastava, Classes of analytic functions associated with the generalized hypergeometric function, *Applied Mathematics and Computation*, 103(1) (1999), 1-13.
 [2] L. de Branges, A proof of the Bieberbach conjecture, *Acta Math.*, 154 (1985), 137-152.
 [3] S. Kumar, H. Taneja, V. Ravichandran. Classes multivalent functions defined by dziok-srivastava linear operator and multiplier transformations, *Kyungpook Mathematical Journal*, 46 (2006), 97-109.
 [4] H. Silverman, Univalent functions with negative coefficients, *Proceedings of the American Mathematical Society*, 51 (1975), 109-116.

[5] J. Dziok and H. M. Srivastava, Classes of analytic functions associated with the generalized hypergeometric function, *Applied Mathematics and Computation*, 103(1) (1999), 1-13.
 [6] A. Gangadharan, T. N. Shanmugam and H. M. Srivastava, Generalized hypergeometric functions associated with k-uniformly convex functions, *Computers Math. Applic.*, 44 (12) (2002), 1515-1526.
 [7] J. -L. Liu, Strongly starlike functions associated with the Dziok-Srivastava operator, *Tamkang J. Math.*, 35, (2004).
 [8] R. M. El-Ashwah, Majorization Properties for Subclass of Analytic p-valent Functions Defined by the Generalized Hypergeometric Function, *Tamsui Oxford Journal of Information and Mathematical Sciences*, 28(4) (2012), 395-405.
 [9] C. Selvaraj and K. R. Karthikeyan, Differential subordination and superordination for certain subclasses of analytic functions, *Far East Journal of Mathematical Sciences*, 29(2) (2008), 419-430.
 [10] C. Selvaraj and K. R. Karthikeyan, Univalence of a general integral operator associated with the generalized hypergeometric function, *Tamsui Oxford Journal of Mathematical Sciences*, 26(1) (2010), 41-51.
 [11] J. E. Hohlov and Operators and operations on the class of univalent functions, *Izv. Vyssh. Uchebn. Zaved. Mat.*, 10(197) (1978), 83-89.
 [12] G. S. Salagean, Subclasses of univalent functions, in *Complex analysis-fifth Romanian-Finnish seminar, Part 1 (Bucharest, 1981)*, 362-372, Lecture Notes in Math., 1013, Springer, Berlin.
 [13] F. M. Al-Oboudi, On univalent functions defined by a generalized Salagean operator, *Int. J. Math. Math. Sci.* 27: (2004), 1429-1436.
 [14] S. Ruscheweyh, New criteria for univalent functions, *Proceedings of the American Mathematical Society*, 49 (1975), 109-115.
 [15] E. El-Yagubi and M. Darus, A new subclass of analytic functions with respect to k-symmetric points, *Far East Journal of Mathematical Sciences*, 82(1) (2013), 45-63.
 [16] B. C. Carlson and D. B. Shaffer, Starlike and prestarlike hypergeometric functions, *SIAM Journal on Mathematical Analysis*, 15(4) (1984), 737-745.
 [17] A. Catas, On certain class of p-valent functions defined by a new multiplier transformations, in *Proceedings Book of the International Symposium G. F. T. A., 2007*, 241-250, Istanbul Kultur University, Istanbul, Turkey.
 [18] S. K. Chatterjea, On starlike functions, *Journal of Pure Mathematics*, 1 (1981), 23-26.
 [19] O. Altintas, On a subclass of certain starlike functions with negative coefficients, *Mathematica Japonica*, 36(3) (1991), 489-495.
 [20] M. Kamali and S. Akbulut, On a subclass of certain convex functions with negative coefficients, *Applied Mathematics and Computation*, 145(2-3) (2003), 341-350.
 [21] O. P. Ahuja, Integral operators of certain univalent functions, *International Journal of Mathematics and Mathematical Sciences*, 8(4) (1985), 653-662.
 [22] K. Al Shaqsi and M. Darus, On certain subclass of analytic univalent functions with negative coefficients, *Applied Mathematical Sciences*, 1(21-24) (2007), 1121-1128.
 [23] H. Kopka and P. W. Daly, *A Guide to L^AT_EX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

Point triangulation using convex layers

V. Tereshchenko, Y. Tereshchenko

Abstract— In this paper, we propose a triangulation method for a set of points in the plane. The method is based on the idea of constructing convex layers by Graham's scan. It allows to develop an algorithm with the optimal complexity of $O(N \log N)$ and an easy implementation. First, convex hulls are constructed for the set S of N points, forming k layers. Then, each layer is triangulated in one scan of the adjacent convex hulls. Algorithm is easily parallelized: each layer can be triangulated independently. The main feature of the proposed algorithm is that it has a very simple implementation and the elements (triangles) of the resulting triangulation are presented in the form of simple and at the same time fast data structures: concatenable triangle queue or triangle tree. This makes the algorithm convenient for solving a wide range of applied problems of computational geometry and computer graphics, including simulation in science and engineering, rendering and morphing.

Keywords—triangulation, convex hull, set of points, convex layers, Graham's scan.

I. INTRODUCTION

THIS paper proposes an optimal triangulation algorithm for a set of points in the plane. The main advantage of triangulation is that from an object, which is potentially very complex, we can move on to more simple polygons (triangles) for their further study.

Relevance. Today, there exist a number of efficient algorithms for solving the problem of triangulation of a set of points [1]. The following groups of triangulation algorithms are distinguished: iterative algorithms (least efficient and quite difficult to implement) [2], algorithms, based on the «divide-and-conquer» strategy (the fastest and relatively easy to implement) [2-5], direct construction algorithms (have good (even linear) average construction time, easy to implement) [6] and two-pass algorithms (most difficult to implement, not very effective) [7].

In my opinion, most effective are triangulation methods, based on the «divide-and-conquer» strategy, that have time complexity $O(N \log N)$ in worst and average cases [2-5]. Among them, algorithms that use concatenable queue data structure at the merge step can be singled out [4, 5]. These algorithms give optimal results in terms of computational complexity – $\Theta(N \log N)$, but it is desirable to have a simpler implementation. So naturally, the question arises – is it

V. M. Tereshchenko is with the Taras Shevchenko National University of Kyiv, Kyiv, Ukraine (corresponding author to provide phone: +38-067-900-2449; fax: +38-044-259-0427; e-mail: v_ter@ ukr.net).

Y. V. Tereshchenko is with the Taras Shevchenko National University of Kyiv, Kyiv, Ukraine (e-mail: y_ter@ ukr.net).

possible to develop an algorithm that would give high efficiency and at the same time would be simple to implement? Once again, we stress that this is important in terms of practical use of the algorithm. An example of such method is Graham's algorithm for construction of convex hulls, that has optimal computation complexity and at the same time is very easy to implement [8]. This method has inspired to develop a triangulation algorithm as efficient and simple in implementation as the algorithm for convex hull construction.

Goal. Develop a triangulation algorithm based on the Graham's method, which would have optimal time complexity ($\Theta(N \log N)$) and at the same time would be easy to implement.

Originality. A new algorithm for triangulation of a set of points is proposed in the paper. The new algorithm is based on the Graham's method and has an optimal complexity of $\Theta(N \log N)$.

II. PROBLEM AND SOLUTION METHOD

Problem formulation. Triangulate the set S of N points in the plane using Graham's scan with the time complexity of $\Theta(N \log N)$.

A. Method of solving the problem

Let S be the set of N points in the plane, Fig. 1.

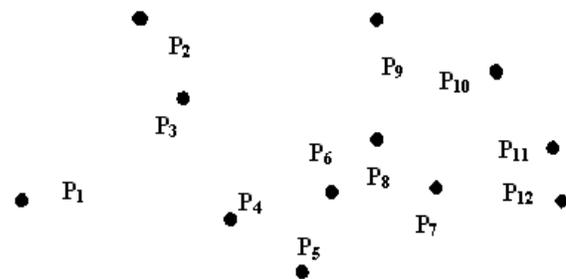


Fig.1. Example of an input set of points.

1. According to the Graham's method, we find centroid q (x_q , y_q) of the first three noncollinear points. For example, in fig. 11 it is the centroid for points P_1 , P_2 , P_3 , with coordinates x_q , y_q , accordingly Fig.2.:

$$x_q = \frac{x_1 + x_2 + x_3}{3}, y_q = \frac{y_1 + y_2 + y_3}{3}.$$

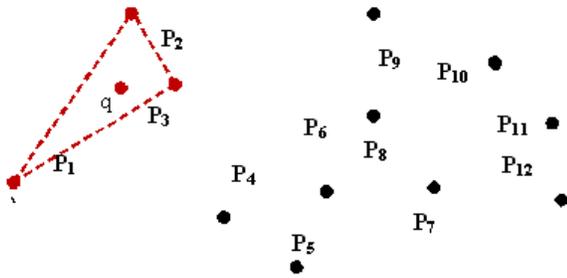


Fig.2. Definition of point q .

2. We sort the given set S of N points by their polar angle (counterclockwise), getting an ordered list U . For our example from fig.1, we will get $U = \{P_4, P_5, P_6, P_7, P_{12}, P_8, P_{11}, P_3, P_{10}, P_9, P_2, P_1\}$.
3. We use the Graham's scan for the list U , as a result obtaining the convex hull for the set S with the boundary $CH(S) = \{P_5, P_{12}, P_{11}, P_{10}, P_9, P_2, P_1\}$, Fig.3.

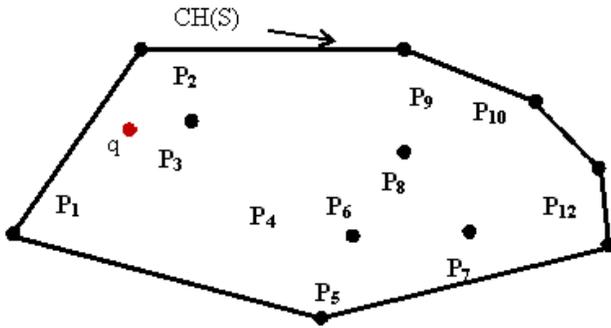


Fig.3. Convex hull for S .

4. For the set of points S_1 , remaining inside the convex hull, we construct convex hull $CH(S_1)$ using Graham's scan. Similarly, we construct convex hulls $CH(S_2), \dots, CH(S_k)$ for the following sets S_2, \dots, S_k , until it is possible, Fig.4.

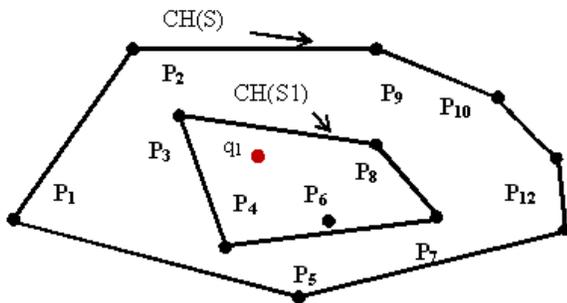


Fig.4.

5. Accordingly we triangulate layers, which are formed by the adjacent convex hull boundaries, Fig. 5. This can be done even during each subsequent scan, and in case of possibility of parallel processing, each layer can be triangulated independently.

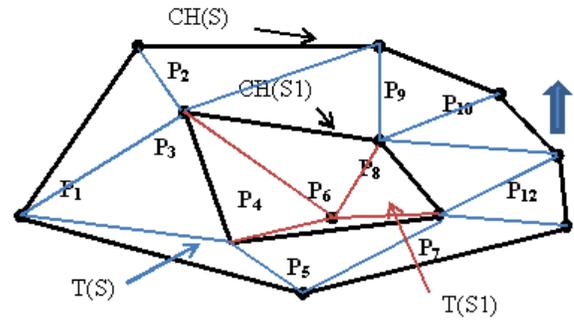


Fig.5.

III. DATA STRUCTURE CONSTRUCTION

The question arises, how to present the resulting triangulation in the form of a certain data structure that could be used for processing and solving the following problems: coloring, finding intersections, morphing, rendering, Boolean operations etc. In this case, the triangulation procedure provides a convenient way for its maintenance, namely:

- 1) During the scan of construction of every new triangle, we assign it a name and add it to the created, cyclically ordered list of such triangles. Fig. 6 shows an appropriate example.

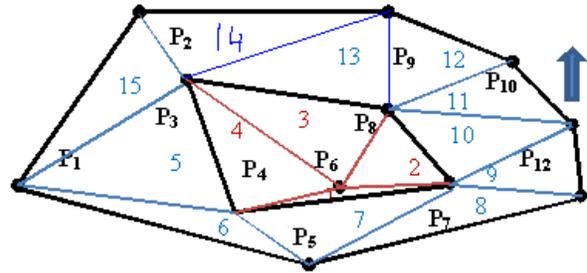


Fig.6. Construction of the list of triangles: $T = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$

- 2) The formed list can be presented in the form of a concatenable queue (Fig. 7), which maintains connectivity in each layer (blue and brown lines) and contains pointers to the links between the edges of the adjacent layers (green lines).

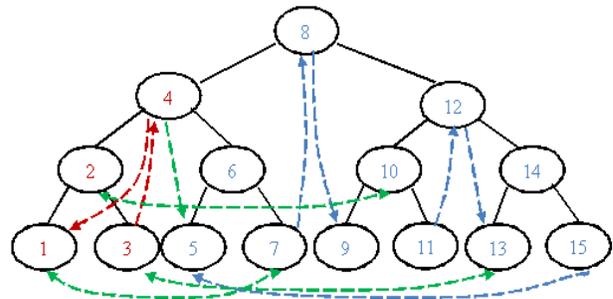


Fig.7. Data structure in the form of a linked queue for fig. 6. First layer edges $\{1,2,3,4\}$ are marked in red, second layer edges $\{5, 6, 7, 8, 9, 10, 11, 12, 13, 14,15\}$ are marked in blue.

This allows to carry out logarithmic search for the triangulation on any layer and in any direction.

3) The actual layered triangulation can be presented in the form of a binary edge tree (Fig. 8), which maintains connectivity.

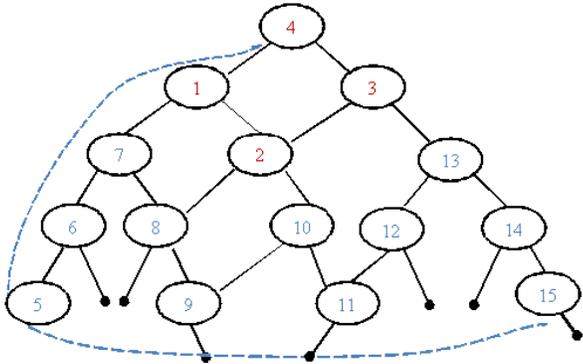


Fig. 8. Data structure in the form of a binary edge tree.

Using this data structure, logarithmic search can be carried out from any triangle.

IV. COMPLEXITY

Theorem. The proposed algorithm has the same time complexity as Graham’s method for convex hull construction – $O(N \log N)$ and uses linear space.

Proof.

1st, 2nd and 3rd steps are steps of the Graham’s algorithm and they require $O(N)$, $O(N \log N)$, $O(N)$ time respectively. 4th step is the Graham’s scan and it requires $O(N)$ time, but for the set of points, left after step 3. 5th step – layer triangulation, which is carried out (considering the ordering of points at the border of each convex hull by their polar angle) in one scan using $O(N)$ time.

V. CONCLUSION

An optimal method for triangulation of a set of points with the complexity of $O(N \log N)$ is proposed in the paper. This algorithm is based on the Graham’s scan for computing the convex hull. First, convex hulls are constructed for the set S of N points, forming k layers. Then, each layer is triangulated in one scan of the adjacent convex hulls. Algorithm is easily parallelized: each layer can be triangulated independently. The main feature of the proposed algorithm is that it has a very simple implementation. The algorithm has application to solving a wide range of applied problems of simulation in science and engineering.

The feature of the proposed method is not only simple process of triangulation constructing, but also convenient representation of its elements for the further use in the form of data structures. These data structures are concatenable queue or a binary faces tree.

REFERENCES

- [1] A.V. Skvorcov. Delaunay Triangulation and Its Application. Izd Tomsk, Gos. Univ., Tomsk, 2002.
- [2] A.B. Skvorcov., Y.I. Kostiuk. Effectivnie algoritmi postroenia trianguliacii Delone. Geoinformatika. Teoria i praktika.- Tomsk: Un-t., 1998, P. 22–47.
- [3] L. Gubias, J. Stolfi. Primitives for the manipulation of general subdivisions and the computation of Voronoi diagrams . *ACM Transactions on Graphics.*- **4**, N 2. – 1985, P. 74–123.
- [4] V. Tereshchenko. The generalized approach of the decision of some problems of computing geometry on the basis of recursively-parallel technology . *Naukovi notatki, Luck-2008.* - **22** , N 2. - , 2008, P. 344-349.
- [5] V.N. Tereshchenko , A.V. Anisimov. Recursion and parallel algorithms in geometric modeling problems. *Journal: Cybernetics and Systems Analysis.*- **46**, N 2.- 2010, P. 173 - 184.
- [6] D.G. Kirpatric. Optimal search in planar subdivisions. *SIAM J. Comput.* – **12**, N 1. –1983, P. 28–35.
- [7] B. Lewis, J. Robinson. Triangulation of planar regions with applications . *The Computer Journal.* – **21**, N 4. – 1978, P. 324–332.
- [8] R. L.Graham . An efficient algorithm for determining the convex hull of a finite planar set. *Info. Proc. Lett.*- 1.- 1972, P. 132-133.



Vasyl Tereshchenko. Professor, Doctor of Phys. - Math. Science, Taras Shevchenko National University of Kyiv, Faculty of Cybernetics. In 1986 graduated from the Mathematics and Mechanics Faculty of Kyiv National Taras Shevchenko University. In 1992 graduated from graduate school and in 1993 defended PhD thesis on the degree C.Sci. (Phys-Math.). Dissertation theme - Non-stationary problems of thermoelasticity for piecewise-homogeneous bodies. In 2011 defended a dissertation for the degree of doctor of Phys. - Math. Science. Dissertation theme - Constructing the common algorithmic environment for solving a set of problems of computational geometry. 01.05.01 - Theoretical bases of computer science and cybernetics. In 1994 - Associate Professor of Faculty of Cybernetics KNTSU. Since 2013 - Full Professor of Faculty of Cybernetics KNTSU. Lecturer in computer graphics and in computational geometry, and also in databases and in the theory of algorithms.



Yaroslav Tereshchenko Since 2011 student of the Faculty of Cybernetics, Kyiv National Taras Shevchenko University. Deals with computational geometry and software engineering.

Onboard Electromechanical Actuators Affected by Motor Static Eccentricity: a New Prognostic Method based on Spectral Analysis Techniques

Dario Belmonte, Matteo D. L. Dalla Vedova, and Paolo Maggiore

Abstract— The proposal of prognostic algorithms able to identifying the precursors of incipient failures of primary flight command electromechanical actuators (EMA) is beneficial for the anticipation of the incoming failure: a correct interpretation of the failure degradation pattern, in fact, can trig an early alert of the maintenance crew, who can properly schedule the servomechanism replacement. In this paper the authors propose an innovative prognostic model-based approach, able to recognize symptoms of an EMA degradation before the actual exhibition of the anomalous behavior. The identification/evaluation of the considered incipient failures is performed analyzing proper critical system operational parameters, able to put in evidence the corresponding degradation path, by means of a numerical algorithm based on spectral analysis techniques. Subsequently, these operational parameters are correlated with the actual health condition of the considered system by means of failure maps created by a reference monitoring model-based algorithm. In the present work, the proposed method has been applied to the case of an actuator having brushless DC motor affected by a progressive increase of the static eccentricity of the rotor. In order to evaluate the performances of the aforesaid prognostic method, a test simulation environment, able to manage different failure modes, has been conceived. This numerical test case simulates the dynamic behaviors of the EMA taking into account nonlinear effects related to different kinds of progressive failures (such as transmission backlash, friction and rotor static eccentricity). Results show that the method exhibit adequate robustness and a high degree of confidence in the ability to early identify an eventual malfunctioning, minimizing the risk of fake alarms or unannounced failures.

Keywords— BLDC Rotor Static Eccentricity, Electromechanical Actuator (EMA), Prognostics and Health Management (PHM), Prognostic Precursors, Progressive Failures, Spectral Analysis

D. Belmonte is with the Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino, Corso Duca degli Abruzzi, 24 - 10129 - Torino, ITALY. (e-mail: dario.belmonte@polito.it).

P. Maggiore is with the Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino, Corso Duca degli Abruzzi, 24 - 10129 - Torino, ITALY. (e-mail: paolo.maggiore@polito.it).

M. D. L. Dalla Vedova is with the Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino, Corso Duca degli Abruzzi, 24 - 10129 - Torino, ITALY. (corresponding author to provide phone: +390110906850; e-mail: matteo.dallavedova@polito.it).

I. INTRODUCTION

ACTUATORS provide mechanical power transforming electrical, pneumatic or hydraulic sources of power, and they are commonly utilized for driving slight control surfaces and utility aircraft subsystems. Flight control systems are critical for reliability indeed it must meet severe reliability requirements of less than one catastrophic failure per 105 flight hours. The reliability of critical components such as actuators used for primary flight controls is designed by conservative safe-life approach which imposes programmed removal of related components after a specific interval of time or operating cycles. Historical records indicate that actual use is often very different from estimated one, because the aforesaid design criterion is not able to evaluate possible initial flaws (occurred during manufacturing) and other impacting factors such as extreme or unanticipated operating scenarios, pilot and flying style in manned systems. Furthermore, statistical based preventive removals are also involved to remove components with significant useful life increasing costs and related inefficiencies. Prognostics is a discipline with the purpose to predict when a certain component loses its functionalities and is not further able to be operative or to meet desired performances. It is based on analysis and knowledge of possible failure modes and on capability to identify incoming faults, due to aging or wear, by monitoring specific operational parameters (called **prognostic precursors**). This discipline is used in other technological fields and could be very useful to condition based maintenance, since it lowers both costs and inspection time. To improve these advantages, a new discipline called **Prognostics and Health Management (PHM)** was born to provide real-time data on current health status of the system and to calculate Remaining Useful Life (RUL) before a fault or failure occurs, when a component becomes unable to perform its features at designed levels. The need for condition based maintenance is clearly recognized, but it is difficult to define robust PHM algorithms due to complex actuators phenomenology. It is necessary to develop a robust health management solution able to perform reliable and acceptable faults detection and failure prediction

analyzing multiple, competitive failures modes by monitoring physically meaningful parameters. Several aircraft use electrohydraulic (EHA) or simply hydraulic actuators for primary flight control system, but incoming Unmanned Autonomous Vehicles (UAVs) will utilize electromechanical actuators (EMA), and several research programs will introduce EMA in future military and civil flight control systems. Typically PHM strategies are easier to implement on EMA since additional sensors are usually not required providing to define the health status of the system. The same sensors framework used for control schemes and systems monitor is also used in many PHM algorithms in a model based approach for health system evaluation. This paper presents a study focused on the development a prognostic technique able to identify failure precursors alerting that degrading performances of a typical aeronautical electromechanical actuator exhibiting an anomalous behavior due to wear phenomena. In particular, three kinds of non-linear physical behaviors are considered: friction, backlash, rotor static eccentricity. To assess the robustness of the proposed techniques, based on a typical **Spectral Analysis** approach, an appropriate simulation test environment has been developed. Simulations have then been run with progressive **Static Rotor Eccentricity** while the EMA model is subjected to different parameters configuration; the algorithms correctly sort out the failure precursors and make a correlation between the actual Static Eccentricity percentage and the calculated operating maps to identify and evaluate incoming failure. Results show that an adequate robustness and confidence has been gained in the ability to early identify the EMA malfunctioning minimizing the risk of fake alarms or unannounced failures.

II. AIMS OF WORK

Aims of this work are:

- 1) The proposal of a numerical algorithm, implemented in MATLAB-Simulink® simulation environment, able to perform the simulations of dynamical systems for EMA taking into account evaluation of rotor static eccentricity due to progressive wear.
- 2) The proposal an innovative prognostic method introducing typical spectral signal analysis techniques able to detect, by specific failure precursors, an accurate health state of flight control systems.

III. PRIMARY FLIGHT CONTROL EMAS

Primary flight controls are typically proportional servomechanisms with continuous activation: they must return a force feedback related to command intensity and a high frequency response. Since their loss is a critical issue, their reliability must be very high. Their purpose is to control the dynamic of the aircraft by generating, by means of the rotation of the corresponding aerodynamic surfaces, unbalanced forces/couples acting on the aircraft itself. These controls are usually conceived to obtain the aircraft rotation around one of the three body axis when one control surface is activated.

This kind of actuator, because of its great accuracy, high specific power and very high reliability, is often equipped on current aircrafts, even if on more modern airliners electro-hydrostatic actuators (EHA) or electro-mechanical actuators (EMA) are installed. In the last years the trend towards the all-electric aircrafts brought to an extensive application of novel optimized electrical actuators, such as the electromechanical ones (EMA). To justify the fervent scientific activity in this field and the great interest shown by the aeronautical world, it must be noticed that, compared to the electrohydraulic actuations, the EMAs offer many advantages: overall weight is reduced, maintenance is simplified and hydraulic fluids, which is often contaminant, flammable or polluting, can be eliminated. For these reasons, as reported in [1], the use of actuation systems based on EMAs is increasing in various fields of aerospace technology.

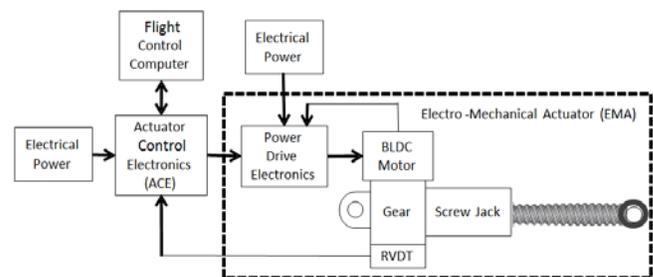


Fig. 1 Electromechanical actuator (EMA) scheme.

As shown in Fig.1, a typical electromechanical actuator used in a primary flight control is composed by:

- 1) An actuator control electronics (ACE) that closes the feedback loop, by comparing the commanded position (FBW) with the actual one, elaborates the corrective actions and generates the reference current I_{ref} ;
- 2) A Power Drive Electronics (PDE) that regulates the three-phase electrical power;
- 3) An electrical motor, often BLDC type.
- 4) A gear reducer having the function to decrease the motor angular speed (RPM) and increase its torque.
- 5) A system that transforms rotary motion into linear motion: ball screws or roller screws are usually preferred to acme screws because, having a higher efficiency, they can perform the conversion with lower friction;
- 6) A network of sensors used to close the feedback rings (current, angular speed and position) that control the whole actuation system (reported in Fig. 1, as RVDT).

IV. PROPOSED ACTUATION SYSTEM NUMERICAL MODEL

As previously mentioned, goal of this research is the proposal of a new a technique to identify precocious symptoms (usually defined failure precursors) of EMA degradations.

In order to assess feasibility performance and robustness of the aforesaid technique, a suitable simulation test environment has been developed in the MATLAB/Simulink® environment. The proposed numerical model (Fig.2), widely described in [2], is consistent with the considered EMA architecture.

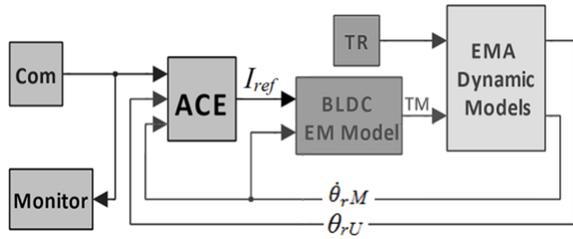


Fig. 2 Proposed EMA block diagram.

As shown in Fig. 2, the propose EMA simulation model is composed by six different subsystems:

- 1) **Com**: input block that generates the different position commands.
- 2) **ACE**: subsystem simulating the actuator control electronics, closing the feedback loops and generating in output the reference current I_{ref} .
- 3) **BLDC EM Model**: subsystem simulating the power drive electronics and the trapezoidal BLDC electromagnetic model, that evaluates the torque developed by the electrical motor as a function of the voltages generated by three-phase electrical regulator.
- 4) **EMA Dynamic Model**: subsystem simulating the EMA mechanical behavior by means of a 2 degree-of-freedom (d.o.f.) dynamic system.
- 5) **TR**: input block simulating the aerodynamic torques acting on the moving surface controlled by the actuator .
- 6) **Monitor**: block simulating the monitoring system.

It must be noted that this numerical model is able simulate the dynamic behavior of the considered EMA servomechanism taking also into account the effects of BLDC motor nonlinearities [3-7], end-of-travels, compliance and backlashes acting on the mechanical transmission [8], analogic to digital conversion of the feedback signals, electrical noise acting on the signal lines and electrical offset of the position transducers [9] and dry friction (e.g. on bearings, gears, hinges and screw actuators) [10].

V. EMA FAILURES AND PERFORMANCE DEGRADATIONS

Only recently EMAs have been employed in aeronautics for flight control systems, so the cumulate flight hours and the on-board installation periods don't permit reliable statistic data about recurring failures. Generally it is possible to classify among four main failures categories:

- 1) Mechanical or structural failures.
- 2) BLDC motor failures.
- 3) Electronics failures.
- 4) Sensor failures.

The present work takes into account effects of mechanical failures due to progressive wear focused on progressive static rotor eccentricity related to bearing wear. Electrical and sensor failures have not a secondary importance, but their evolution is very fast, nearly instantaneous, so corresponding failure precursors are often hard to identify and evaluate; nevertheless it is intention of the authors to study these kind of failure in a next work.

As is known, wear increases the dry friction phenomena that occur between two surfaces in relative motion, increasing both static and dynamic friction coefficients. The driven system requires higher torques to actuate the control surface with the same external load. Even if an increased dry friction does not cause seizure of the entire system, it reduces the corresponding servomechanism accuracy and sometimes it generates dynamic unexpected responses, as stick-slip or limit cycles due to the interaction between PID controller and friction forces

The mechanical wear could also generate backlash in EMA moving parts such as gears, hinges, bearings and especially ball screw actuators; these backlashes, acting on the elements of the mechanical transmission, reduce the EMA accuracy and, as a function the mutual position with respect to the signal transducers, can lead to problems of dynamic stability and controllability of the whole actuator.

The eccentricity fault of a stator and rotor is due mainly to mechanical reasons. In this kind of failures, the axes of symmetry of the stator and of the rotor and the rotational axis of rotor are displaced to each others. This displacement of symmetrical axes can be classified into **static**, **dynamic** and **mixed eccentricities** [11]; in this work, only the first type of eccentricity will be discussed.

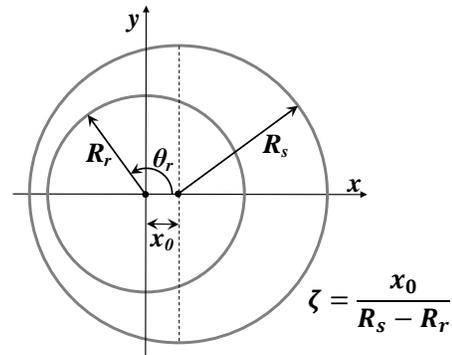


Fig. 3 Reference system for the definition of rotor static eccentricity ζ .

The **static eccentricity** (Fig. 3) consists in a misalignment between the rotor rotation axis and the stator axis of symmetry; the rotor is symmetric and rotates towards its rotation axis, this misalignment initially is mainly due to manufacturing tolerances and imperfections, but, during operational period, increases as a consequence of wear in bearings that support rotor shaft. When this failure occurs in multipolar motor, the rotor generates a magnetic flux that has not cyclic symmetry, since the air gap varies during its 360° degrees turn. Therefore, if a rotor has an evaluable static eccentricity, an additional radial force component arises and its magnitude varies like a sine wave. In condition of Static Eccentricity the air gap is not constant and symmetric along rotor turn (Fig. 3), so the clearance between the rotor and the stator can be mathematically represented by this function:

$$g'(\theta) = g_0 + x_0 \cos(\theta) \tag{1}$$

where g_0 is the initial clearance without misalignment and the second term added represents the sinusoidal air gap variation as a function of the misalignment x_0 . In terms of motor performances, in these conditions the provided torque is lower than in nominal conditions (because of a change in the electromagnetic characteristics of the engine): to simulate eccentricity effects avoiding more complex electromagnetic FEM models, a smart numerical algorithm has been released. Since static eccentricity modifies the magnetic coupling between the stator and the rotor, the proposed algorithm adjusts the angular modulation of the back-EMF coefficients and thereby the related torque constants.

As reported in [1], this algorithm is implemented by means of the functions $f(u)$ contained in the BLDC EM Model block of Fig. 2 and acts on the three back-EMF constants Ce_i (one for each of the three phases) modulating their trapezoidal reference values Ke_i as a function of coil short circuit percentage, static rotor eccentricity ζ and angular position ϑ_r .

$$e_a = Ke_i \cdot Ce_i \cdot (1 + \zeta \cdot \cos(\vartheta_r)) \quad (2)$$

The obtained constants (ke_a, ke_b, ke_c) are then used to calculate the corresponding counter-electromotive forces (ea, eb, ec) to evaluate the mechanical couples (Cea, Ceb, Cec) generated by the three motor phases.

The effects of the aforesaid failures will be briefly analyzed in the next section, in which the related EMA simulations will be examined. With respect to other EM models available in literature, the numerical model shown in the previous sections is able to calculate the instantaneous value of each current phase (Ia, Ib, Ic) also in case of unbalanced electromagnetic system (e.g. partial short circuit on a stator branch or rotor static eccentricity). Then, it is possible correlate the progressive static eccentricity with these currents (used as failure precursors) by means of an algorithm, based on the Fourier spectral analysis, that analyses the filtered phase currents; for this purpose, each phase current is filtered by three low pass signal filter, in order to attenuate noise and disturbances. The Fourier transform is a mathematical instrument that transforms the time domain representation into a frequency domain representation, which has many applications in physics and engineering. The Fourier Transform comes from a study of Fourier series that it represents complicated but periodic functions as infinite sum of sine and cosine functions with different amplitude and phase. It's possible to represent sine and cosine formulas using Euler's Formulas, then the Fourier Series is written by (3).

$$f(x) = \sum_{n=-\infty}^{+\infty} c_n e^{2\pi i \left(\frac{n}{T}\right)x} \quad (3)$$

where c_n is the n-th Fourier coefficient.

The Discrete Fourier Transform (DFT) is the equivalent of Continuous Fourier Transform for a signal known only at N samples time during a finite Time acquisition [12], so we have a finite sequence data considering the signal periodic as:

$$F[n] = \sum_{k=0}^{N-1} f[k] e^{-j\frac{2\pi}{N}nk} \quad (4)$$

The DFT approximates the Fourier Transform since it provides only for a finite set of frequencies during a limited acquisition time. It must be noted that there are two main types of DFT calculation errors: "Aliasing"¹ and "Leakage". According to the Nyquist-Shannon theorem, defined f_M the upper limit of the frequency bandwidth of a signal, in order to avoid "Aliasing" phenomena during DFT calculation, the Sampling Frequency f_s must be defined as:

$$f_s > 2f_M \quad (5)$$

In the present work, in order to avoid Aliasing errors. the performed DFT uses a sampling frequency f_s equal to the inverse of the integration time DT of EMA simulation model: this requisite allows to represent all high frequency components calculated by the model EMA.

The continuous Fourier Transform of a waveform requires the integration to be performed over the interval $-\infty$ to $+\infty$ or over an integer number of cycles of the waveform. If we attempt to calculate the DFT over a non-periodic signal, we have a corrupted frequency transform due to "Leakage" errors. For most waveform of real data is not be possible to reduce "Leakage" without a specific data modification. This used solution is called "Windowing" [14]: a cosine function is applied over the entire signal to taper the samples towards zero at both endpoints without discontinuity with a hypothetical next period. Rather than performing a DFT calculation, a FFT calculation is often preferred to reduce the number of involved multiplications [15-16]. The proposed numerical module uses a FFT with a "Hanning" windows: this type of windowing, often used for general purpose applications in spectral analysis [13], is defined as:

$$Wn = \frac{1}{2} \{1 - \cos[2\pi n/(N - 1)]\}; \quad 0 \leq n \leq (N - 1) \quad (6)$$

Windowing have a side effects because data are reduced and set to zero at the beginning and end of each time record to force the signal to be periodic; it must be noted that there is no loss in amplitude readout accuracy, but a loss in frequency resolution is present.

A processing technique called "overlap" is able to enhance events occurring near the beginning and ending of time record. The entire simulation time is divided in several time section interval, but these intervals are not time sequential but overlapped of 67% from the first time section in order to improve spectral information. Energy information in signal must be preserved during transformation. That is, the energy measured on time signal must equal the energy measured on the frequency representation of that signal.

All the aforesaid spectral techniques (and the related requisites) are merged together and implemented into the proposed numerical module called **EMA Spectral Algorithm**.

¹ We have "Aliasing" DFT calculation errors when the samples are not sufficiently closely spaced to represent high frequency spectral components.

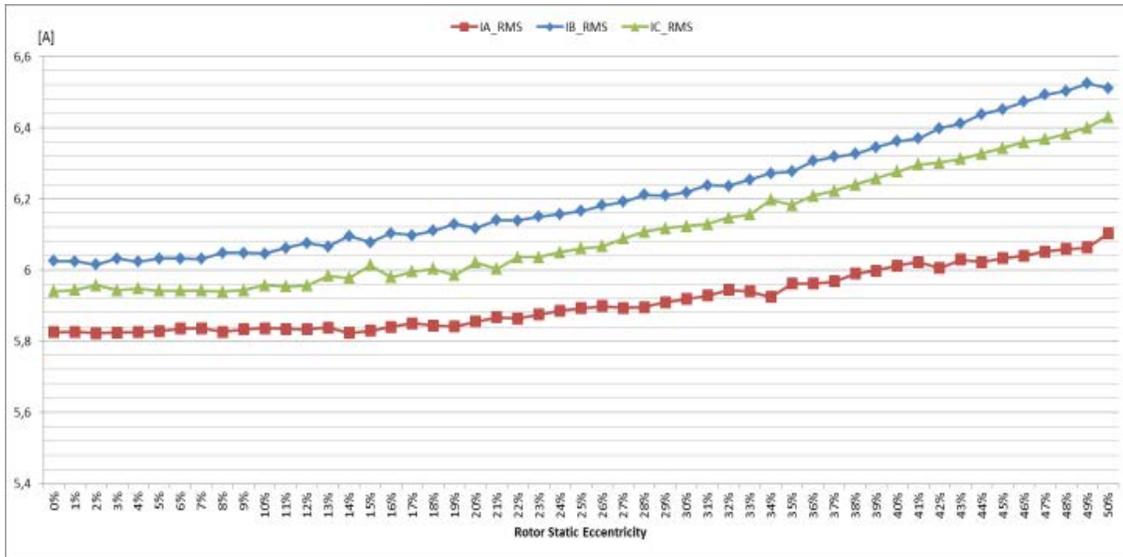


Fig. 4 Evolution of RMS currents as a function of rotor static eccentricity ζ

This module processes all filtered phase currents deriving from each considered value of the rotor static eccentricity and correlates these failures with the corresponding failure precursors, generating a simulated “operating map”.

Once defined, for each type of EMA examined, its own “operating maps” it is possible to conceive dedicated fault detection/evaluation systems (on-board systems or portable devices equipped with embedded versions of the proposed spectral analysis algorithms) able to evaluate static rotor eccentricity during preflight test.

VI. EMA SPECTRAL ALGORITHM AND OPERATING MAPS

Simulation EMA environment works with step command with a very high position value for driving surface to quickly saturate the electromechanical actuator, that after a short transient time, it reaches the max velocity without aerodynamic load. The entire simulation time test amounts to one second and, for each simulated actuation test, all filtered phase currents (I_{fA} , I_{fB} , I_{fC}) are acquired. These filtered current signals, expressed as a function of simulated time, are divided in intervals called “time sections” of 25% of the simulation time and, sequentially, a FFT single sided spectral analysis (according to the Cooley Tukey algorithm [15]) is performed for each time section. To improve the frequency resolution of the algorithm, the time sections are extracted using overlap processing so, during one second of acquired signal, more time sections are post-processed [17].

Thus, a set of FFT spectral diagrams is calculated; the meaningful harmonics are identified by means of a **peak hold function** that, for each frequency line, finds the maximum magnitude peaks between all spectral diagrams that are extracted for each filtered phase current.

The comparison among the sets of max magnitude peaks found in each filtered phase current reveals that from 0% to 33% of static rotor eccentricity the first max magnitude peak has the same constant frequency value called Hz1 (Hz1= 33.568 Hz).

After this percentage of static rotor eccentricity, the first magnitude peak for each filtered phase current switches to another constant frequency value called Hz 2 (Hz 2= 32.805 Hz); this means that, when the static eccentricity is equal to a given percentage of the clearance between rotor and stator, the frequency of the first maximum magnitude peak decreases (Hz1>Hz2).

The Root Mean Square (RMS, also known as the quadratic mean) of a given signal time history is a measure of overall energy and it is often used to extract signal features for prognosis and trending data. In order to avoid numerical problems, the time history of the considered signal must be digitized at a particular sample rate (for a total of N samples), then RMS value can be estimated by:

$$rms = \sqrt{\frac{1}{N} \sum_{i=1}^N x(i)^2}$$

For each filtered phase current (I_{fA} , I_{fB} , I_{fC}) a RMS value is calculated processing the rotor static eccentricity values from 0% to 50% with 1% increasing step; the results are three signals, called I_{A_RMS} , I_{B_RMS} and I_{C_RMS} , evolving as shown in Fig. 4.

The progressive eccentricity causes progressive asymmetry of the magnetic field so the RMS filtered phase current values increase and, therefore, the torque needed to maintain the same actuator velocity increase.

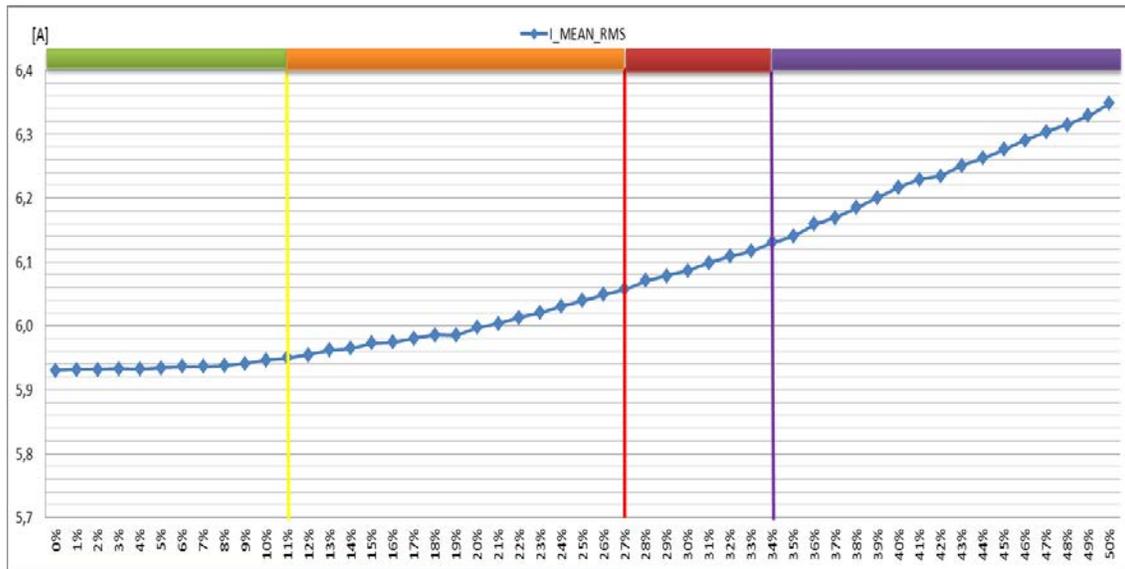


Fig. 5 Evolution of mean value of RMS currents as a function of rotor static eccentricity ζ

As shown in Fig. 5, evaluating the mean value of the RMS of filtered phase currents, calculated for each of the considered eccentricity percentage ζ , it is possible to perform an evaluation of the EMA health conditions, defining several operating intervals related to specific eccentricity percentages:

- 1) **Green Phase:** from 0% to 10% of the rotor static eccentricity (with respect to the stator-rotor air gap); this operating interval corresponds to **Normal Mode** with acceptable actuator performances; it is related to a negligible static eccentricity, mainly due to tolerances of manufacturing and beginnings of mechanical wear. It must be noted that, with respect to reference value mean RMS on 0% static eccentricity, this green phase has a wide interval about 16 mA;
- 2) **Orange Phase:** from 11% to 26% of the rotor static eccentricity percentage; this operating interval corresponds to **Moderate Mode** with actuator performances related to incoming evaluable command degradations. Respect to reference value mean RMS on 11% static eccentricity this orange phase has a wide interval about 100 mA;
- 3) **Red Phase:** from 27% to 33% of the rotor static eccentricity; this operating interval corresponds to **Serious Mode** where actuator performances are degraded and condition based maintenance operations need to be planned. With respect to reference value mean RMS on 27% static eccentricity, the red phase has a wide interval about 60 mA;
- 4) **Violet Phase:** from 34% to 50% of the rotor static eccentricity; this operating interval corresponds to **Extreme Mode**. In this case the actuator performances are very degraded and the based maintenance operations are needed as soon as possible. With respect to reference value mean RMS on 34% static eccentricity, this violet phase has a wide interval about 200 mA;

VII. CONCLUSIONS

The proposed model-based approach allows to calculate specific operating maps for many different EMA models: in fact, modifying defined sets of technical parameters, it is possible to adapt the performances of the numerical system to a given type of EMA and, so, define the corresponding operational map.

Actual EMA failure precursors, directly acquired by on-board maintenance systems, are compared with the corresponding calculated operating map in order to evaluate, during a preflight test, the percentage of static rotor eccentricity avoiding degraded flight command performances.

Once defined the operational maps simulating EMA model, integrated with the authors spectral numerical module, it is possible to have an adequate accuracy to individuate the health state of the actual actuator by performing pre-test flight as indicated in previous paragraphs.

Results are encouraging the extension of the proposed technique to investigate more challenging occurrences, such as the electrical and sensor failures, for which the evolutions are usually very fast, if not instantaneous, and the corresponding failure precursors are often difficult to identify and evaluate.

For this purpose the actuator model should be further detailed and new element should be modelled. Combined failures should also be investigated.

REFERENCES

- [1] M. Battipede, M.D.L. Dalla Vedova, P. Maggiore, and S. Romeo, "Model based analysis of precursors of electromechanical servomechanisms failures using an artificial neural network," *AIAA Modeling and Simulation Technologies Conference*, Kissimmee, Florida, 2015.
- [2] P. Maggiore, M.D.L. Dalla Vedova, L. Pace, and A. Desando, "Definition of parametric methods for fault analysis applied to an electromechanical servomechanism affected by multiple failures," *Second European Conference of the PHM Society 2014 (PHME'14)*, Nantes, France, pp. 561-571, 8-10 July 2014.
- [3] M. Çunkas, and O. Aydoğdu, "Realization of Fuzzy Logic Controlled Brushless DC Motor Drives using Matlab/Simulink", *Mathematical and Computational Applications*, vol. 15, pp. 218-229, 2010.
- [4] A. Halvaei Niasar, H. Moghbelli, and A. Vahedi, "Modelling, Simulation and Implementation of Four-Switch Brushless DC Motor Drive Based On Switching Functions", *IEEE EUROCON 2009*, St.-Petersburg, Russia, pp. 682 – 687, 2009.
- [5] B.K. Lee, and M. Ehsani., "Advanced Simulation Model for Brushless DC Motor Drives", *Electric Power Components and Systems*, Vol. 31, No. 9, pp. 841–868, ISSN: 1532-5008, 2003.
- [6] T. Hemanand, and T. Rajesh, "Speed Control of Brushless DC Motor Drive Employing Hard Chopping PWM Technique Using DSP", *Proceedings of India International Conference on Power Electronics (IICPE 2006)*, 2006.
- [7] T.A. Haskew, D.E. Schinstock, and E.M. Waldrep, "Two-Phase On' Drive Operation in a Permanent Magnet Synchronous Machine Electromechanical Actuator", *IEEE Transactions on Energy Conversion*, vol. 14(2), 1999.
- [8] L. Borello; G. Villero, and M.D.L. Dalla Vedova. "New asymmetry monitoring techniques: effects on attitude control." *Aerospace Science and Technology*, vol. 13(8), pp. 475-487, 2009.
- [9] L. Borello, M.D.L. Dalla Vedova, G. Jacazio, and M. Sorli, "A Prognostic Model for Electrohydraulic Servovalves", *Annual Conference of the Prognostics and Health Management Society*, San Diego, CA, 2009.
- [10] L. Borello, and M.D.L. Dalla Vedova, "A dry friction model and robust computational algorithm for reversible or irreversible motion transmission", *International Journal of Mechanics and Control*, vol. 13(2), pp. 37-48, 2012.
- [11] M. Akar, S. Taskin, S. Seker, and I. Cankaya, "Detection of static eccentricity for permanent magnet synchronous motors using the coherence analysis", *Turkish Journal of Electrical Engineering & Computer Science*, vol. 18(6), pp. 963-974, 2010.
- [12] A. V. Oppenheim, R. W. Schaffer, and J. A. Buck, *Discrete-time signal processing*. Upper Saddle River, N.J.: Prentice Hall. pp. 468–471, 1999.
- [13] F. Harris, Fredric, "On the use of Windows for Harmonic Analysis with the Discrete Fourier Transform". *Proceedings of the IEEE*, vol. 66 (1), pp. 51–83, 1978.
- [14] W. Hongwei, "Evaluation of Various Window Functions using Multi-Instrument", *Virtins technology*, May 2009. <http://www.virtins.com/>
- [15] J.W. Cooley, and J.W. Tukey, "An algorithm for the machine calculation of complex Fourier series", *Mathematics of Computation*, vol. 19 (90), pp. 297–301, 1965.
- [16] D. F. Elliott, and K. R. Rao, "Fast transforms: Algorithms, analyses, applications", New York: Academic Press, 1982.
- [17] F. Ramian, "Implementation of Real-Time Spectrum Analysis", January, 2011.

Dario Belmonte received the M.Sc. at Politecnico di Torino in 2011. Since 2011 he worked as assistant researcher at the Department of Mechanics and Aerospace Engineering and GE Avio Aero srl. His research activity is mainly focused on analysis and numerical simulation of dynamic systems and aerospace servomechanism, integrating correlations between FEM and CAT (experimental data) related to structural vibrational analysis for Gear Box and Propulsor Turbine modules.

Matteo D. L. Dalla Vedova received the M.Sc. and the Ph.D. from the Politecnico di Torino in 2003 and 2007, respectively. He is currently assistant researcher at the Department of Mechanics and Aerospace Engineering. His research activity is mainly focused on the aeronautical systems engineering and, in particular, is dedicated to design, analysis and numerical simulation of on board systems, study of secondary flight control systems and conception of related monitoring strategies, development of prognostic algorithms for aerospace servomechanism and study of innovative primary flight control architectures.

Paolo Maggiore is a professor at the Mechanical and Aerospace Engineering Department of Politecnico di Torino, that joined in 1992, where he teaches aerospace general systems engineering. Currently his students are involved in projects ranging from hydrogen fuel cell powered airplanes and UAVs, and health monitoring of flight controls, to multi-disciplinary design optimization of aerospace systems design

Optimal problems with control-state constraints in a regional economy model identification

Vasily V. Dikusar, Nicholas N. Olenev, and Marek Wojtowicz

Abstract— A multi-sector model of regional economy based on balances for economic agents is considered. The model can be used for the forecast of development for a regional economy if its numerous parameters are identified according to historical data of economic development. A convolution of Theil indexes for time series of statistical and calculated data is used as criterion function. The problem of identification is written in the form of a problem of optimum control with restrictions of a general form. A two-stage method for solution of optimal problems with non-regular control-state constraints is offered. At the first stage the discrete problem on the basis of methods of the factor analysis is solved. This discrete problem consists of subtasks for systems of linear algebraic equations and improper problems of (non)linear programming. Then a hypothesis on geometry of an optimum trajectory is formulated, that is periods of constancy for a set of active restrictions are allocated. At the second stage the formulated hypothesis is checked analytically with use of the Pontryagin Maximum Principle and Dubovitsky-Milyutin's formalism. The offered scheme is applied to a problem of parameter identification for the considered model of regional economy. Methods of distributed computing and GRID-technologies are used in solution of computationally complex problems.

Keywords—Factor analysis, regional economy model, optimal control, parameter identification.

I. INTRODUCTION

Solution of the problem of economy model identification on a uniform mesh of parameters does not require intensive data exchange. So to solve this problem you can use instead of expensive supercomputers the distributed computing and GRID-technologies built on the basis of pooling resources of personal computers. Using mathematical methods, such as numerical methods for optimal control, methods for parameter continuation, and factor analysis, reduces the required time of calculations [1-3].

This paper presents a normative economic model of regional economy [4-6] to demonstrate new mathematical methods to solve the parameter identification problem. This work presents also criteria of closeness for time series of

economic parameters. They use these criteria for indirect estimation of unknown economic model parameters by comparison of calculated macroindexes with their statistical analogues. The simulation model has a lot of unknown parameters so the identification problem is very complicated. The identified model of regional economy enables to receive a quantitative estimation of future dynamics of macroindexes for regional economy.

The presented here normative balance dynamic regional economy model consists of three production sectors. The parameters are identified by statistical data for the Nizhni Novgorod Region economy [5-6]. The model considers state taxation and a shadow turnover. This regional model has eight economic agents: three Production sectors (X , Y , Z), Households (H); Regional bank system (B), Trade intermediary (T), Government (G), and Outside world (O).

Production sectors in the model are presented by three sectors: (X) mining and infrastructure industries, (Y) the manufacturing sector, (Z) services sector, including financial services [5].

The model gives description of formal and informal financial flows. The official stock of money for each production sector grows by new credits, sales of its goods on domestic and foreign markets, by budget transfers, and by flow of shadow money. It decreases by flow of salaries to households, by consumption of intermediate products, and by tax transfers to the consolidated regional budget.

As result the model contains of about hundred equations and more than eighty parameters which cannot be defined directly from the economic statistical data. Moreover, even if all of the necessary statistical data are available, their quality is not very good.

For the model parameter estimation time series of regional economic macroindexes calculated by the model are compared with time series of corresponding statistical data. As criteria of fitting of the time series the Theil index of inequality is used.

The problem of the model identification is formulated as an optimal problem with control-state constraints and can be resolved as this problem. Dubovitskii and Milyutin [1] were the first who formulate and study optimal problems with nonregular control-state constraints. The basic notions and concepts in optimal control theory are related to the maximal principle. The presence of state constraints and control-state constraints of complex nature complicates the maximal principle and its properties. Classical technique does not provide a comprehensive study of such problems. Another approach to the study of an optimal control problem is to

This work was supported in part by the Russian Foundation of Basic Research under Grant 15-07-08952.

V. V. Dikusar is with Dorodnicyn Computing Center of the Russian Academy of Sciences, 40 Vavilova, Moscow, 119333 Russia (phone: +7-499-135-0080; fax: +7-499-135-6159; e-mail: dikussar@yandex.ru).

N. N. Olenev is with Dorodnicyn Computing Center of the Russian Academy of Sciences, 40 Vavilova, Moscow, 119333 Russia (e-mail: nolenev@mail.ru).

M. Wojtowicz is with Kazimierz Pulaski University of Technology and Humanities in Radom, Poland (e-mail mar.wojtowicz@gmail.com).

single out a specific class of problems and study it qualitatively and quantitatively.

II. REGIONAL ECONOMY MODEL

Regional economy production sectors X, Y, and Z use labor, capital and intermediate products of adjacent production sectors. Production sectors deliver product on domestic and foreign market, and on the market of intermediate product. It is considered that prices are formed in each market of each product and change of the prices is in inverse proportion to change of stocks of corresponding products. Households L offer labor and consume final product. Trade intermediary T redistributes material and financial flows. Regional bank system B emits money resources, gives out credits to production sectors. Regional government G accumulates taxes from production sectors and adjusts charges of the budget. The model uses profit tax n_1 , value-added tax n_2 and excises n_3 , uniform social tax n_4 , customs duties on export n_5 and households - the customs duties on import n_6 , the surtax n_7 .

The formal and informal products of each sector differ only that informal one is not taxed. As a result each production sector has two kinds of money – "white" and "black". "Black" money can be washed, and the stock of not washed money is exposed to penal sanctions. All moneys of the consumer are considered "white", and the consumer divides the income by norms of consumption of formal and informal products of all sectors.

In this paper we use the following standard designations for macroindexes and parameters of the model [4]. Superscripts are used to point agents, and subscripts are used to point goods. It is assumed that distribution of a stock of each good is made by norm: a_i^{mm} - a share of a stock of the good i going from economic agent n to economic agent m . It is assumed that distribution of money is made also under a norm: b_i^{mm} - a share of money stock going from agent m to agent n for a product i . It is assumed that capital intensities also are set by some norms: c_i^m - agent m norm of expenses on product i for creation of one unit of capital product. Parameters of production functions of sectors X, Y, and Z are set by constants. For example, sector X output $Y_X(t)$ of product X produced by economic agent X (mining and infrastructure industries complex of the Nizhni Novgorod Region) is described by Cobb-Douglas production function of used production factors (stocks Q): labour L, capital K and intermediate products from sectors Y and Z.

$$Y_X = (a_L^X Q_L^X)^{\delta_L^X} \cdot (a_K^X Q_K^X)^{\delta_K^X} \cdot (a_Y^X Q_Y^X)^{\delta_Y^X} \cdot (a_Z^X Q_Z^X)^{\delta_Z^X}, \quad (1)$$

where $\delta_L^X + \delta_K^X + \delta_Y^X + \delta_Z^X = 1$. Hereinafter, all parameters $Y_X(t)$, $Y_Y(t)$, $Q_L^X(t)$, $Q_K^X(t)$, $Q_Y^X(t)$, $Q_Z^X(t)$, are considered as functions of time t , therefore this argument falls in formulas. All parameters, as a rule, are considered as constants, as here $a_L^X, a_K^X, a_Y^X, a_Z^X, \delta_L^X, \delta_K^X, \delta_Y^X, \delta_Z^X \in (0, 1)$. Agent X produces formal product X and informal product V by the common capital, the common labour and the common stocks of intermediate products, and made product (output) $Y_X(t)$ shares in fix proportion $(1 - q_X) : q_X$ on open ("white") output X and

informal ("black") output V, where q_X is a share of the informal product in the common output of the product $Y_X(t)$. The informal product is used for sale to population and other sectors. It is assumed in the model that investments can be official only. The stock of open product $Q_X^X(t)$ increases due to production and decreases due to shipment to agents Y, Z, L and for investments $I_X(t)$. It is considered that charges on investments from the own product coincide with incomes of them. There is a fix share of own product stock that goes on foreign market $X_X^{XO}(t)$.

$$\frac{dQ_X^X}{dt} = (1 - q_X) Y_X - (a_X^{XL} + a_X^{XY} + a_X^{XZ} + a_X^{XO}) Q_X^X - c_X^X I_X, \quad (2)$$

where

$$I_X = \frac{b_K^X W^X}{p_X^X c_X^X + p_Y^X c_Y^X + p_Z^X c_Z^X},$$

$$p_X^X = \min \{ p_X^Y, p_X^Z \},$$

$$X_X^{XO} = a_X^{XO} Q_X^X.$$

The stock of intermediate product Y of agent X grows by purchase of formal product Y of agent Y on price $p_Y^X(t)$ and of informal product W of agent Y on price $p_W^X(t)$ and decreases by use it in production and investment. The stock of intermediate product Z of agent X grows by purchase of formal product Z on price $p_Z^X(t)$ and of informal product U of agent Z on price $p_U^X(t)$ and decreases by use it in production and investment.

$$\frac{dQ_Y^X}{dt} = \frac{b_Y^{XY} W^X}{p_Y^X} + \frac{b_W^{XY} W^X}{p_W^X} - a_Y^X Q_Y^X - c_Y^X I_X,$$

$$\frac{dQ_Z^X}{dt} = \frac{b_Z^{XZ} W^X}{p_Z^X} + \frac{b_U^{XZ} W^X}{p_U^X} - a_Z^X Q_Z^X - c_Z^X I_X. \quad (3)$$

The stock of labor in sector X grows by purchase from regional households L of formal labor L on official wage $s_L^X(t)$ and informal labor B on unofficial wage $s_B^X(t)$ and decreases by demand of labor of agent X.

$$\frac{dQ_L^X}{dt} = \frac{b_L^{XL} W^X}{s_L^X} + \frac{b_B^{XL} B^X}{s_B^X} - a_L^X Q_L^X. \quad (4)$$

The stock of capital in sector X grows by investment $b_K^X W^X(t)$, and decreases by decreasing of capital of agent X with tempo μ_K^X and by use of capital in production of sector X.

$$\frac{dQ_K^X}{dt} = b_K^X W^X - \mu_K^X Q_K^X - a_K^X Q_K^X. \quad (5)$$

A value of credit $C^{BX}(t)$ from regional bank system B to agent X is limited by residual value of productive assets, which is assumed to be proportional to the capital stock.

$$C^{BX} = \sigma^X Q_K^X, \quad \sigma^X > 0. \quad (6)$$

Debt $Z^X(t)$ of agent X to the bank system B grows by new loans $C^{BX}(t)$ and by current interest rate $r(t)$ on debt, and decreases by repayments $H^{XB}(t)$.

$$\frac{dZ^X}{dt} = C^{BX} + rZ^X - H^{XB}, \quad H^{XB} = b_H^{XB} W^X. \quad (7)$$

The stock of open ("white") money $W^X(t)$ of economic agent X grows when agent X takes bank credits, sales the goods on foreign market, sales the goods on domestic markets, takes transfers from regional consolidated budget $T^{GX}(t)$ and receipts of the "washed" money from a shadow turnover $b_B^X B^X(t)$. It decreases due to payments of wages to households L , due to consumption of intermediate products of adjacent sectors Y and Z , due to payments on credits and transfers of taxes to the consolidated budget.

$$\begin{aligned} \frac{dW^X}{dt} = & wp_X^O X_X^{XO} + C^{BX} + (p_X^L a_X^{XL} + p_X^Y a_X^{XY} + p_X^Z a_X^{XZ}) Q_X^X - \\ & - (b_Y^{XY} + b_Z^{XZ} + b_W^{XY} + b_U^{XZ} + b_L^{XL} + b_H^{XB}) W^X - \\ & - T^{XG} + T^{GX} + b_B^X B^X. \end{aligned} \quad (8)$$

Here $w(t)$ is Ruble /USD exchange rate, $T^{XG}(t)$ is transfer payments of taxes to the consolidated regional budget, $T^{GX}(t)$ is transfer from the budget to sector.

Transfers of agent X to the consolidated budget $T^{XG}(t)$ develop from the profit tax $T_1^{XG}(t)$, the added value tax $T_2^{XG}(t)$, the excises $T_3^{XG}(t)$, the uniform social tax $T_4^{XG}(t)$, the customs duties on export $T_5^{XG}(t)$.

$$\begin{aligned} T^{XG} &= T_1^{XG} + T_2^{XG} + T_3^{XG} + T_4^{XG} + T_5^{XG}, \\ T_5^{XG} &= n_5 wp_X^O X_X^{XO}, \\ T_4^{XG} &= n_4 b_L^{XL} W^X, \\ T_3^{XG} &= n_3^X \left[wp_X^O X_X^{XO} + (p_X^L a_X^{XL} + p_X^Y a_X^{XY} + p_X^Z a_X^{XZ}) Q_X^X \right] \\ T_2^{XG} &= n_2 \left\{ wp_X^O X_X^{XO} + (p_X^L a_X^{XL} + p_X^Y a_X^{XY} + p_X^Z a_X^{XZ}) Q_X^X - \right. \\ &\quad \left. - (b_Y^{XY} + b_Z^{XZ} + b_H^{XB}) W^X - T_3^{XG} - T_4^{XG} - T_5^{XG} \right\}, \\ T_1^{XG} &= n_1 \left\{ wp_X^O X_X^{XO} + (p_X^L a_X^{XL} + p_X^Y a_X^{XY}) Q_X^X - \right. \\ &\quad \left. - (b_Y^{XY} + b_Z^{XZ} + b_H^{XB} + b_L^{XL}) W^X - T_2^{XG} - T_3^{XG} - T_4^{XG} - T_5^{XG} \right\} \end{aligned} \quad (9)$$

The informal share in total output q_X defines a gain of stocks both opened $Q_X^X(t)$ and shadow $Q_V^X(t)$ products. The stock of shadow product V in sector X decreases due to deliveries to households and allied sectors.

$$\frac{dQ_V^X}{dt} = q_X Y_X - (a_V^{XL} + a_V^{XY} + a_V^{XZ}) Q_B^X. \quad (10)$$

The stock of shadow money of informal share of production grows by sale of informal product as final product to households L and, as intermediate product, to adjacent sectors Y and Z , a part b_B^X of shadow money stock has time to wash, a part b_B^{XG} gets as penal sanctions in a profitable part of the regional government consolidated budget, and a part b_B^{XL} goes to the households as shadow incomes.

$$\begin{aligned} \frac{dB^X}{dt} = & (p_V^L a_V^{XL} + p_V^Y a_V^{XY} + p_V^Z a_V^{XZ}) Q_V^X - \\ & - (b_B^{XL} + b_B^X + b_B^{XG}) B^X. \end{aligned} \quad (11)$$

The equations for sectors Y and Z can be write by analogy to equations (1) - (11) for sector X .

Change of stock for final product X intended to agent L (households), $Q_X^L(t)$ defines change of a consumer price index p_X^L on product X .

$$\frac{dQ_X^L}{dt} = a_X^{XL} Q_X^X - \frac{b_X^{LX} W^L}{p_X^L}, \quad (12)$$

$$\frac{dp_X^L}{dt} = \alpha_X^L \left(\frac{b_X^{LX} W^L}{p_X^L} - a_X^{XL} Q_X^X \right). \quad (13)$$

Changes of stocks of all other products in all markets and change of the corresponding prices are similarly defined. Thus we assume, that the product of the same sectors acts in other quality in the different markets and has the other price, that it is, as a matter of fact, other product.

It is considered that increase of the open and shadow wages can occur both at shortage of the labour, and due to increase of consumer prices on production of the sector. For example, the open wage at sector X is determined by equation

$$\begin{aligned} \frac{ds_L^X}{dt} = & \left[\alpha_L^X \left(\frac{b_L^{XL} W^X}{s_L^X} - a_L^{LX} Q_L^{LX} \right) \right. \\ & \left. + \frac{\beta_L^X s_L^X}{p_X^L} \left(\frac{b_X^{LX} W^L}{p_X^L} - a_X^{XL} Q_X^X \right) \right]_+. \end{aligned} \quad (14)$$

Here the following designation is used: $A+ = A$, if $A > 0$ and $A+ = 0$, if $A \leq 0$. It is supposed that the gain of wages at surplus of a manpower does not exceed a gain of the price, $\beta_L^X \leq \alpha_X^L$.

The stock of money on the regional consolidated budget accounts increases by tax revenues and decreases by transfers to production sectors and households.

$$\frac{dW^G}{dt} = b_V^{XG} B^X + b_W^{YG} B^Y + b_U^{ZG} B^Z + T^{XG} + T^{YG} + T^{ZG} + T^{LG} - (b_X^{GX} + b_Y^{GY} + b_Z^{GZ} + b_L^{GL}) W^G. \quad (15)$$

The regional bank system is not closed, and the branches of the Russian banks from other regions play greater role in investment decisions. Therefore it was assumed that a part gold and exchange currency reserves of the Russian Federation provide reservation for actives of the region.

The description of the model terminates here because of restrictions on volume.

III. IDENTIFICATION PROBLEM

In this work a method for macroeconomic model parameter estimation based on parallel processing by distributed computing and GRID-technologies is proposed. For an application of this method some measure of similarity between two time-series is required. Complicated economic models with a lot of parameters can be accurately estimated by the proposed method.

A macroeconomic model usually contains a lot of unspecified parameters. Some variables of the macroeconomic model have initial values that are sometimes unknown and should be considered as parameters as well. In most cases mentioned parameters can't be defined on the basis of economic statistics. Moreover, even if all necessary statistics is available the quality of the data isn't always good. That's why only confidence intervals for the unknown parameters can be computed from the data.

For estimation of the unknown parameters time-series for some macro-indexes (calculated by the model) and statistical time-series for these macro-indexes should be compared by means of some measure of similarity. The unknown parameters can be determined implicitly as those parameters, which provide minimum value of the used measure of similarity. Parallel processing on a cluster of workstations or on a supercomputer enables to perform exhaustive search of the parameters within their confidence intervals (determined either from economic sense or from the available statistical data) and estimate their values for reasonable time.

Estimated values of the parameters give new knowledge about the macroeconomic system under investigation and this knowledge can force a researcher to modify the model. Thus macroeconomic model parameter estimation based on parallel processing becomes a powerful tool for mathematical modelling of economic systems.

It was mentioned that calculated and statistical time-series for some macro-indexes should be compared on the basis of some measure of similarity. As used here, two time-series are considered to be similar if they are close as functions of time (in other words, if there is a strong, possibly nonlinear dependence between two time-series). Convolution of the Theil's index and specially designed wavelet based measure of similarity was used as a characteristic of closeness between two time-series.

With linear trend appropriate rescaling with respect to ordinate axe two completely different time-series can become

quite similar and the value of Euclidian distance between these time-series can decreases significantly. Such effect frequently occurs in many real situations.

The Theil index of inequality measures discrepancy of time series X_t and Y_t and the more close it to zero, the more close compared numbers. For convenience of calculations, instead of Theil index we shall use an affinity index.

$$U(X, Y) = 1 - \frac{\sqrt{\sum_{t=t_0}^T (X_t - Y_t)^2}}{\sqrt{\sum_{t=t_0}^T X_t^2 + \sum_{t=t_0}^T Y_t^2}}. \quad (16)$$

Decomposition of model on separate blocks enables define independent parameters for reasonable time due to parallel calculations for model parameter estimation on the fixed intervals of their changes.

For choice for optimal point it is possible to use some convolution of affinity indexes for time series of macroeconomic indexes. For example, if for all macro parameters the adjustment of estimated by model and statistical data has about equal importance, it is possible to maximize compound value of all indexes.

Recording formal

$$K(\bar{a}) \rightarrow \max_{\bar{a} \in A}, \quad (17)$$

where

$$A = \{ \bar{a} \in R^N : a_i^- \leq a_i \leq a_i^+, 1 \leq i \leq N \}, \quad (18)$$

$$K(\bar{a}) = \prod_{j=1}^m U_j(\bar{a}). \quad (19)$$

Here m is a total number of macroeconomic indexes that are compared with statistical ones, $j = 1, \dots, m$. It is necessary to touch only those variants of values of parameters, at which affinity indexes are above some fix positive values, for example, $U_j > 0.85$ ($j = 1, \dots, m$).

Now full search on all unknown parameters, set on an interval, even on the most powerful supercomputers is impossible, as search on a grid from 10 points on each interval on 80 parameters gives 10^{80} variants, that approximately is equal to number of nucleons in the Metagalaxy. Means it is necessary to use the directed search, to reduce number of independent parameters due to additional assumptions by mathematical methods set out in the next section.

IV. OPTIMAL PROBLEMS WITH CONTROL-STATE CONSTRAINTS

Here we consider the optimal control problems from applications in economics. Especially those problems are considered which include control-state variable constraints [7-8]. An optimal control problem can be solved completely or at least qualitatively with the help of the maximum principle

(MP). Difficulties in the numeric solving the optimal control problems with control-state constraints (CSC) might derive from insufficient analysis of model problems and inadequate progress in the MP theory in the case. The MP itself could be a base of the numeric methods. Availability of state variable and CSC of a complex nature leads to a complication of the formulation and properties of the MP.

New objects appear: measure and functional Lagrange multipliers. It becomes necessary to examine the properties of these objects and analyze the interconnections of the various parts of the MP. Otherwise it would be impossible to use the MP. It is known that the MP reduces the initial control problem to the solution of the two-point (multi-point boundary value problems). They include: initial value problem (IVP); the problem of linear (LP) or nonlinear (NP) programming; root finding of transcendental functions. The IVP consists of the two groups of nonlinear differential equations (direct and adjoint).

A number of investigators have tried and rejected shooting methods for certain classes of two-point problems which are particularly sensitive to the initial conditions and are thereby troublesome numerically. On the other hand there are successful examples of shooting methods at the expense of a special parametrisation and continuation (homotopy chains) the solution (R. Bulirsch, H.J. Pesch, J. Stoer, J. Bett, H. Maurer, A. Afanasjev, E. Smoljakov and others).

One of the most effective numerical techniques for the solution of optimal control problems is discretizing the differential equations. This approach combines a nonlinear programming problems with discretization. The resulting problem is characterized by matrices which are large and sparse. The iteration routine requires a good initial guess for the adjoint variables. Such treatment of the problem has been successfully utilized for applications.

This paper extends the continuation methods for ill-posed problems. Our approach utilized the system dynamics and adjoint differential equations defined by the maximum principle. The resulting boundary value problem is characterized by Jacobi or Hesse matrices which are small (dimension) and ill-conditioned.

Method of introduction the parameter helps to overcome the difficulties associated with adjoint initial values.

Among the topics covered are:

- Investigation of the degeneracy phenomenon of the maximum principle for optimal control problems with state constraints and CSC.
- Stiff ordinary differential equations.
- Ill-posed linear and nonlinear programming problems.
- Continuation method in ill-posed boundary valued problems.
- Reentry body problem.
- Optimal control problems in economics.
- Banking activity problems.
- Determination of geometry for optimal trajectory with inequality constraints.

- Parallel calculations.

V. SOME RESULTS OF THE MODEL IDENTIFICATION

Numerical experiments with model were spent to find the efficient variant qualitatively truly reflecting processes, occurring in economy of the Nizhni Novgorod Region. Numerical experiments have shown working effectiveness of full model and its separate parts (Figure 1). It means that the model can be used in the further work. External parameters of this variant can be taken for a basis for more exact identification of parameters of model in the future, and to use the variant as base of numerical qualitative scenario calculations with model.

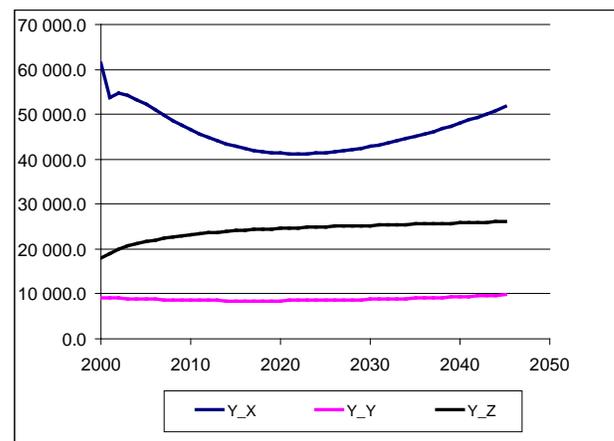


Figure 1 Outputs in production sectors of Nizhni Novgorod Region, where Y_X – output of sector X, Y_Y – output of sector Y, Y_Z – output of sector Z

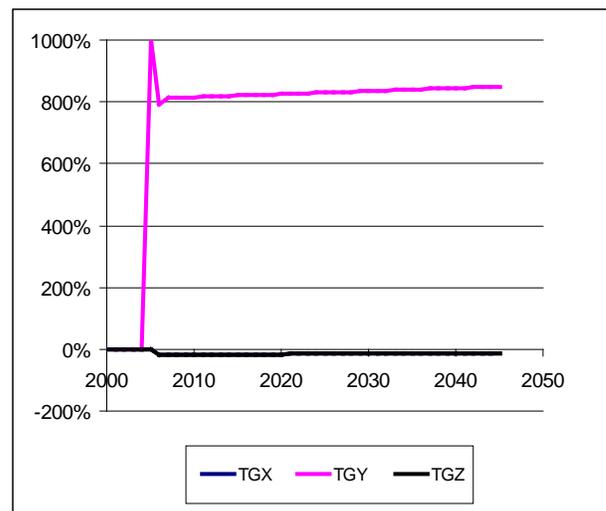


Figure 2 Variation of transfers from regional consolidated budget to production sectors of Nizhni Novgorod Region, where TGX – transfer to sector X, TGY – transfer to sector Y, TGZ – transfer to sector Z

Changes in scenario calculation in comparison with the base variant can be represented by a variation of change of the macroindexes expressed in percentage. If $F(t)$ is a value of

some macroindex at moment of time t in the base variant, and $S(t)$ is a value of the same macroindex in the current scenario, variation $V(t)$ for macroindex changes is defined as $V(t) = 100\% \cdot (S(t)/F(t) - 1)$.

Here results of scenario 1 are presented. In this scenario transfers from consolidated budget to manufacturing sector Y grow up from 2008 they came to stimulate innovation in this sector. But all the cost structure of the sector (including in this case, science, scientific services and education) is not changed. Namely, let budget transfers in the sector Y will increase from 2% to 22% of the money stock of the consolidated budget of the Nizhni Novgorod region.

Real transfers to manufactory sector Y in scenario 1 grow up slightly more than 8 times then in base scenario.

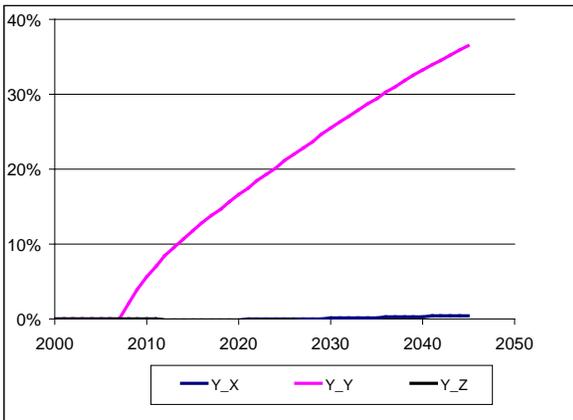


Figure 3 Variations for outputs of regional production sectors (Y_X – output of sector X, Y_Y - output of sector Y, Y_Z - output of sector Z)

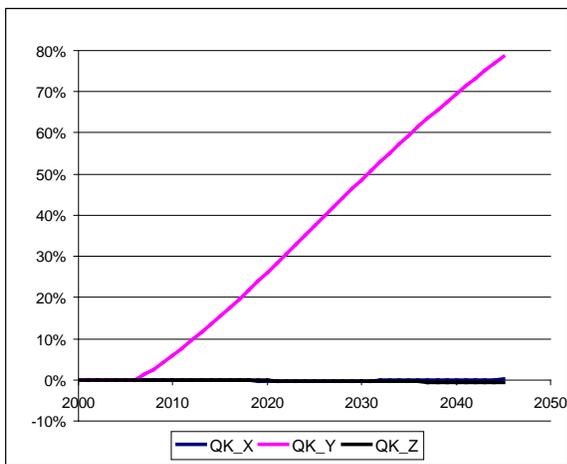


Figure 4 Variations for capitals of regional production sectors (QK_X – capital of sector X, QK_Y - capital of sector Y, QK_Z - capital of sector Z)

Figure 4 show the increase of capital in production sector Y.

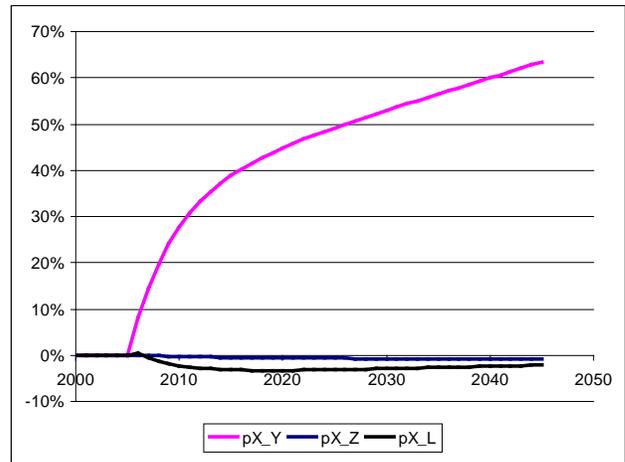


Figure 5 Variations for price indexes on product X on intermediate and consumption markets (pX_Y – price index on X for sector Y, pX_Z – price index on X for sector Z, pX_L – price index on X for households L)

In Figure 5 one can see that the price indexes vary on a miscellaneous in the different domestic markets. Price index of intermediate product for sector Y are increased.

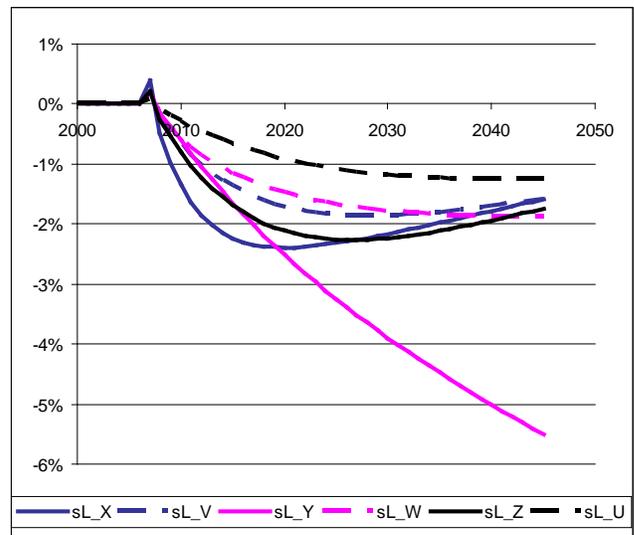


Figure 6 Variations for wage indexes at production sectors (sL_X – wage index for formal labor force at sector X, sL_V – wage index for informal labor force at sector X, sL_Y – wage index for formal labor force at sector Y, sL_W – wage index for informal labor force at sector Y, sL_Z – wage index for formal labor force at sector Z, sL_U – wage index for informal labor force at sector Z)

The change of consumer price indexes conducts to decrease in all wage indexes on all channels of realization: both on official and on shadow ones (Figure 6). From Figure 6 one can see that the wage indexes for shadow channels decrease not so much as they do for official channels, and differently by sector: the more big decrease is for export sector X (official wage for official product X and shadow wage for informal product V), more for innovative sector Y (official wage for

official product Y and shadow wage for informal product W), and less for sector Z (official wage for official product Z and shadow wage for informal product U).

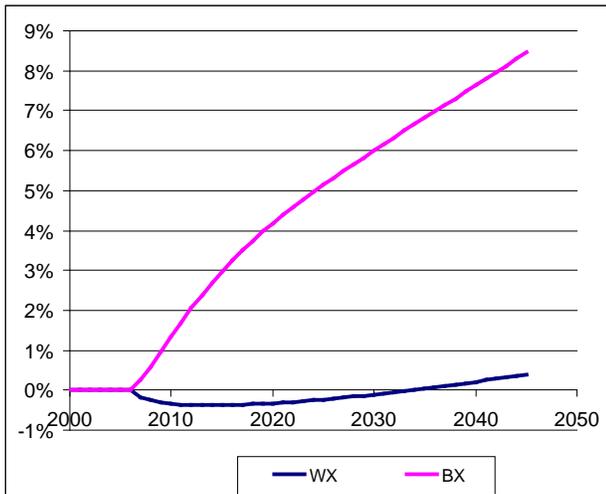


Figure 7 Variations for money of regional production sector X (WX – white money of sector X, BX – black money of sector X)

As you can see from Figure 7 the white money stock in sector X is increased and its black money stock is decreased. The same is true for other sectors.

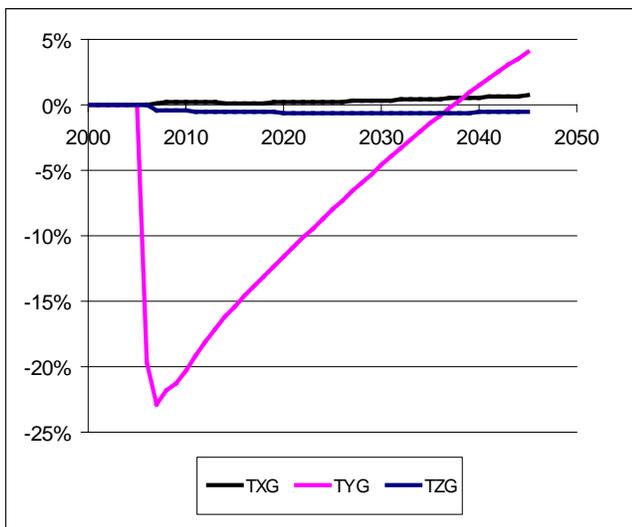


Figure 8 Variations for regional government budget income from production sectors (TXG – regional government budget income from sector X, TYG – regional government budget income from sector Y, TZG – regional government budget income from sector Z)

Let now consider the optimistic scenario, when as a result of support innovation processes of technology transfer in 2010 there was an increase returns on all factors of production, namely, all the degrees in the production function, increased by 5%. Then the gross output at constant 2000 prices increased in all sectors (Figure 9), with the output X industry doubled in

4 times, and for sector Y and Z in 15 times. At the same time, despite the growth in labor productivity, employment increases slightly in the sectors of the economy (Figure 10).

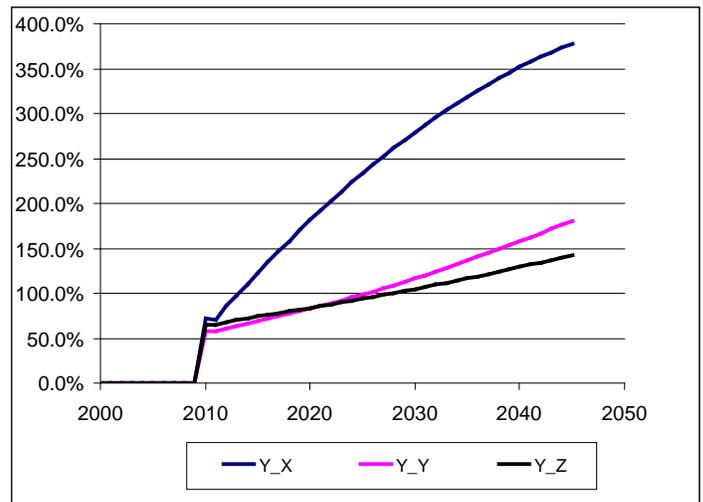


Figure 9 Variations of outputs for optimistic scenario in production sectors of Nizhni Novgorod Region, where Y_X – output of sector X, Y_Y – output of sector Y, Y_Z – output of sector Z

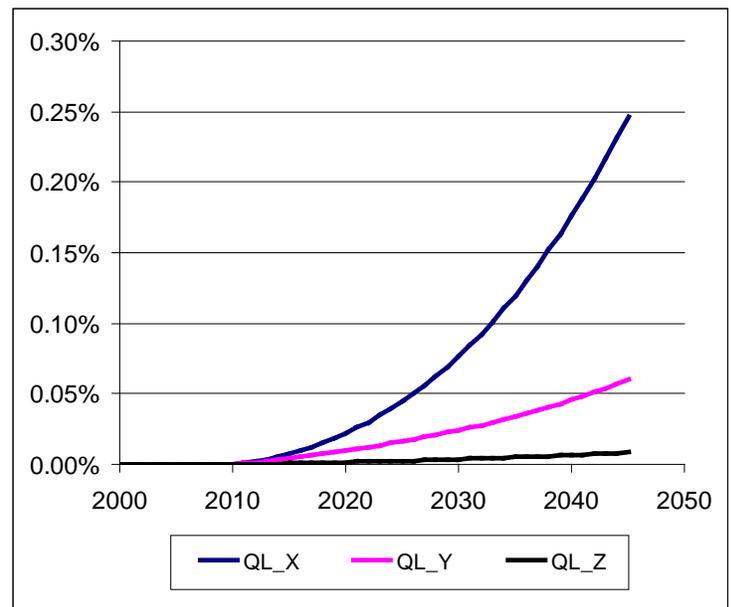


Figure 10 Variations of employments for optimistic scenario in production sectors of Nizhni Novgorod Region, where QL_X – employment in sector X, QL_Y – employment in sector Y, QL_Z – employment in sector Z

All wages are growing, with the exception of legal wage rate sector X of mining and infrastructure industries (Figure 11). Investments of all sectors are grow up (Figure 12).

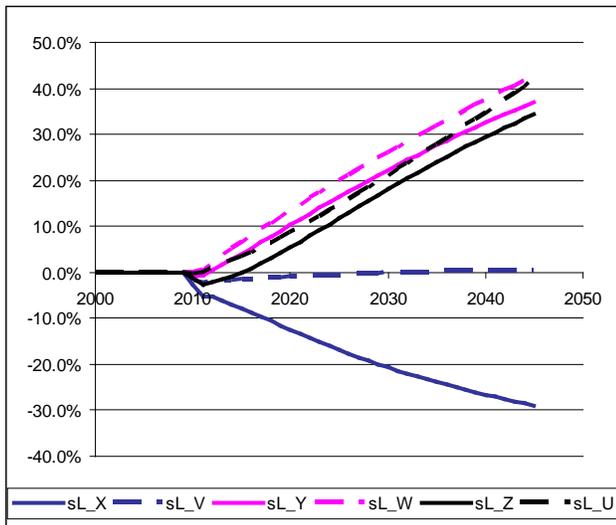


Figure 11 Variations of wages in all production sectors in Nizhni Novgorod Region for optimistic scenario (formal – solid lines, informal – dashed lines)

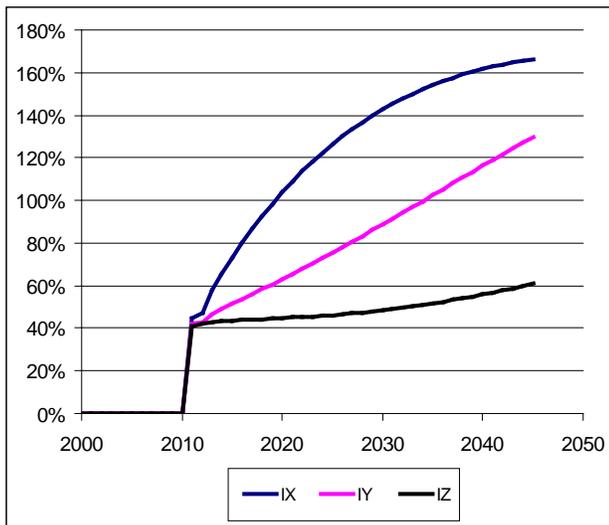


Figure 12 Variations of investments in all production sectors in Nizhni Novgorod Region for optimistic scenario

REFERENCES

[1] A.Ya. Dubovitskii, A.A. Milyutin, "Necessary conditions for a weak extremum in optimal control problems with mixed constraints of the inequality type," *USSR Computational Mathematics and Mathematical Physics*, Vol. 8, Iss. 4, 1968, pp. 24–98.

[2] V. V. Dikusar, "Optimal control problems with mixed constraints," *Differential Equations*, Vol.32, No. 11, 1996, pp. 1462-1468.

[3] V. V. Dikusar, N. N. Olenev, M. Wojtowicz, "Optimal Control Problems with Constraints," in *Abs. IV Int. Conf. on Optimization Methods and Applications "Optimization and applications" (OPTIMA-2013)*. Montenegro, Sep. 22-28, Moscow, 2013, pp. 44-45.

[4] N. Olenev, and N. Mollaverdi, "A Normative Dynamic Model of Regional Economy," *International Journal of Industrial Engineering & Production Research*, June 2011, Vol. 22 N. 2, pp. 99-105. Available: <http://ijiepr.iust.ac.ir/>

[5] V. P. Gergel, N. N. Olenev, V. V. Ryabov, K. A. Barkalov, S. V. Sidorov, "Global optimization in the identification of the multisector model of the Nizhni Novgorod Region economy"

[6] K. Barkalov, N. Olenev, "Parallel global optimization for the problem of a regional economy model identification," in *Proc. Int. Conf. "Numerical Computations: Theory and Algorithms,"* 17–23 June 2013, Falerna, Italy, 2013, p. 45.

[7] A. P. Afanasyev, V. V. Dikusar, A. A. Milyutin, S. A. Chukanov Necessary condition in optimum control, Moscow, Nauka, 1990.

[8] V. V. Dikusar, A. A. Milyutin, Qualitative and numerical methods in maximum principle. Moscow, Nauka, 1989.

Vasily V. Dikusar professor, Dr. Sc., chief researcher of the Department "Problem of Optimization" of Dorodnicyn Computing Center of the Russian Academy of Sciences (RAS). He is Professor of numerical mathematics (1990), Senior Researcher of theory of systems, control and system analysis (1983), associated professor (1993), Doctor of physics and mathematics (1982), PhD of physics and mathematics (1971). Primary Scientific results: 1). Reentry body problem with control-state variables constraints (nonregular case). 2). Optimal control synthesis problem (without Riccati equation). 3). Ill posed linear and quadratic programming problems (the methods of estimation of solution). 4). Continuation methods for solving boundary value problems. 5). Continuous analogs of Newton and gradient methods for ill posed problems. 6). Ordinary differential equation with small (large) parameters (stiff equations, long terminal time, singular perturbed equations). 7). Regularization of degenerated maximum principle. 8). Method of matrix regularization for solving linear algebraic system. 9). An initial assesment of optimal trajectory for inequality constraints. 10). Dubovitsky-Milutin problems with control-state and state variable constraints. 11). As a sample it is proposed consider the flight dynamics problems, environmental problems, inverse electrocardiology problems, financial mathematics, economics, forecasting the prices problems, electroheating problems.

Nicholas N. Olenev is senior researcher of the Department for Mathematical Modeling of Economic Systems at Dorodnicyn Computing Center of the Russian Academy of Sciences (RAS), PhD (1993), docent (2005). He is docent of Moscow institute of physics and Technology (2003), and docent of Peoples Friendship University of Russia (2003).

Marek Wojtowicz is professor of Kazimierz Pulaski University of Technology and Humanities in Radom, Poland.

The decision making process in the system of product design and planning based on Kansei Engineering

Kai-Shuan Shen

Abstract—Based on Kansei Engineering, this sequential study proposes strategies and decision-makings to determine product features and design specification. Then, we mainly probe into users' needs inside a vehicle and propose draft resolution to car interior design, space utilization and function variation for Crossover B-car developers' references. The ideas of product design were verified through experts. Then, the AHP is applied as the strategies to determine product features in this study. In this study, the proposed system could make the process of product design and planning more practical and efficient. Finally, the results of the study could be referred to the designers, managers and researchers for the development of car interior.

Keywords—decision making, design practice, design process(es), production strategy, Kansei Engineering, Analytic Hierarchy Process (AHP)

I. INTRODUCTION

The activity of product design involves with the interaction among human experience, creative thinking and domain-specific knowledge. Hence, a product designer has to put his effort and creativity into a design depending on the type of design problem. In addition, a system can be used to solve problem. For example, the system proposed by Rouse (1986) is important to consider support for the designer's creativity [1]. In this study, the system can be viewed as a decision-making tool for facilitating the process of practical product design and development. In addition, designers' creativity can be inspired in the system. Based on this previous study "The design evaluation process in product design and planning based on consumer appeal", design evaluation plays a critical role in the process of product design and planning because design evaluation is the foundation of strategies and decision-makings in the process of product design and planning. Furthermore, design specifications, principle solutions, and design selection from various initial design alternatives should be determined with full considerations through the assistance of design evaluations during the design conceptualization [2]. In this study, the system performs the decision-making activities according to experienced user and expert evaluation with a clear

procedure in the process of product design and planning. In this system, a strategic way to decide which ideas and design characteristics as criteria have the most advantages can be applied according to Kansei Engineering. Then, after critical criteria were extracted, designers or manager could figure out possible alternatives and chose the best alternative with the assistance of AHP.

The prototype of the proposed system has been developed for the integration of design and product planning. In this study based on Kansei Engineering, as a preference-based design methodology, customers' feedbacks and experts' opinions also are integrated in to this system to assure critical original product ideas to be approved. Hence, this system can assist designers in making a persuasive choice.

Then, marketing strategies can be put forward with the assistance of AHP for product design from integral consideration such as market. The achievements in this study, which have been validated by costumers, experienced users, and experts, are promising and contribute to the efficiency, credibility, practicability and creativity of product design and planning,

II. DESIGN TARGET & DESIGN CASE

In the current market, consumers have more and more needs to B Car and put more and more emphasis on Crossover functions. Many experts believe that crossover style has led the fashion trend of international automobile industries and will get popular around the world. Mercedes-Benz has always been well-known for its luxurious and pragmatic design. However, after Benz has successfully developed A-Class and C-Class, it created a new segment B-Class, between the two classes. The orientation of B-Class, as the concept of "sports tour", caters for consumers' need of "specious room", "outstanding comfort", and "eminent pragmatic" and conquers worldwide young owners with traditional design and driving fun. Benz B-class can make passengers satisfied not merely because of comfortable seats. Furthermore, it integrates elegant design and pragmatic style to create the ideal interior surrounding. The B-class provides extensive interior space even if the body of it is smaller than many other larger cars (Fig 1).



Fig. 1 Benz B170 1.7 exterior and interior design

III. REVIEW OF LITERATURE

A. The process of product design and planning

Product design integrated into the activities of product developing through agent-based system has become a new trend and necessity for product designers and developers. Sun et al (2010) attempted to integrate product design and manufacturing planning through an agent-based concurrent engineering system and presented its fundamental framework and functions [3]. In addition, Chen (2001) proposed a concurrent design evaluation system (CONDENSE) developed to help product designers in evaluating possible design solutions and design alternatives during the early stage of design.

During the process of product design and planning, the determination of design specification is critical and necessary. The design specification is derived from the brief or initial market specification, which is usually a statement of the general objectives and requirements of the product to be developed [4]. The design specification, also called solution principle, performs the best combination of physical effects and preliminary embodiment features to meet the design requirements under the design constraints. For the example of designing a CROSSOVER B-CAR interior, “comfortable position of armrests” can be one of the design specifications to fulfil the design requirement “conforming to ergonomics” under the design constraint “comfortable and convenient”.

Problem encountered in developing design specifications originates from uncertainty in the determination of the design specifications. Then, design specifications must be determined so that the product design process can proceed to the conceptual design stage [5] [6]. In addition, the conceptual design is generally considered to be the highest level of design, involving the most senior product designers working in consultation with management, marketing, and production staff [4].

B. The strategies and decision-makings of product design

The newness of new product development results from of the chosen design strategy and is represented by different new product developments [7]. This argument implies that design strategy plays a critical role in the process of product development. Furthermore, a principal aim of design strategy is to offer variety to consumers using different product features, styles, qualities, sizes, and so on [8]. In this study, design strategy was performed through picking out appealing product features and deciding styles through expert meeting. Therefore, product design strategy in fact reflects a policymaker’s motivation. Furthermore, Baxter (1995) determined that such

motivation can be transformed into a goal [9]. In this study, in order to reach the goal styles of Crossover B-Car interior, designers’ motivations were materialized with the assistances objective statistical analysis and subjective expert decisions and strategies.

Design strategies can be viewed as the policies to achieve a goal. Compared to design strategy, decision-makings of design are applied to execute these policies. Furthermore, effective decision-making can make the process of product design and planning more efficient and can prevent from making mistakes. In this study, decision-making mainly relied on the information based on the results of user and expert evaluations. In addition, the decision-making of design could be more objective due to the assistance of statistical analysis.

C. Kansei Engineering

Kansei Engineering is a design method based on semantic evaluation. The term “Kansei Engineering” comes from director of Japan Mazda Car Company in 1986, whose address it in a conference of world automobile technology, managers’ conference of U.S.A automobile industrial and class speech in Michigan University. He stressed on that car have to contribute to creation of culture for extending the theory of car culture. Japanese scholar Nagamachi brought out a new consumer-oriented technology of product development-Kansei Engineering. He defined Kansei Engineering as the technology of transforming consumers’ feeling, demands, and impressions to product into the elements of design and production function [10]. More specifically, Kansei engineering is used to grasp vague demands of the consumer, and develop the car based on the user’s word. This was based upon the analysis data showing a relationship between human impressions and interior design [11].

There are four main aspects which Kansei Engineering explores: (1) how to grasp the consumer’s feeling (Kansei) about the product in terms of ergonomic and psychological estimation, (2) how to identify the design characteristics of the product from the consumer’s Kansei, (3) how to build Kansei Engineering as an ergonomic technology, and (4) how to adjust product design to the current societal change or people’s preferences trend [12].

“Kansei” is similar to a semiotic system, which is designed to determine human’s affection and preferences. Hence, from the aspect of Kansei Engineering which can leads a user-centered design study, what kind of feeling a product can brings to users is a very import issue to designers. For example, Nagamachi (2008) used the method of Kansei Engineering to find a hierarchy of the values in a customer’s life [13]. In addition, subjective evaluations were carried out by semantic differential methods, and then analysed by using multivariate analyses [11].

Kansei Engineering is utilized in the automotive, electrical appliance, construction, clothing and other industries [12]. For example, Nissan, Mazda and Mitsubishi were eager to implement Kansei Engineering and they began to produce many kinds of newly designed passenger cars [12]. In addition, Kansei Engineering can also be used for studying car interior style and design detail such as how the design of a meter or a

steering wheel can affect humans' feeling while using them. Jindo (1997) attempted to explore styling or design specification of car interiors by Kansei engineering, especially regarding the speedometer and steering wheel of a passenger car [11].

D.AHP (Analytic Hierarchy Process)

Saaty proposed Analytic Hierarchy Process (AHP) in 1971 in order to solve the decision-making and multi-aspect problem which cannot be solely settled through the measurement of quantities [14]. AHP can make complicated question systematized and hierarchically decompose different aspects. In addition, after sequences of ideas are sought out through the judgment of quantification, they can be synthetically evaluated for providing policymaker with adequate information of selecting appropriate proposals. A policymaker can simultaneously diminish the risk of wrong decision-making by sharing risk for individual element of each factor. Hence, as far as a policymaker is concerned, hierarchical structure of is helpful to the understanding to things. In addition, while a policymaker faces the situation of selecting appropriate proposals, it is necessary to proceed to the evaluation of every alternative one according to certain criteria [14].

Thomas L. Saaty, the author of the well-known AHP method, has recently been gaining popularity. AHP analyses and incorporates comparisons of importance between the elements of the system (criteria and alternatives) as perceived by the decision maker. In this paper a new approach, which integrated Kansei Engineering into AHP, was applied for selecting design alternatives for Crossover B-Car interior. Comparing to traditional multi-criteria decision analysis, this approach can help decision makers to sieve out critical criteria and their corresponding factors and to select the best alternatives with a more professional and strict procedure.

E. Quantification Type I

This study also adopted the Quantification Type I method as a tool to analyse the importance of the appeal factor of Crossover B-Car interior. Quantification I method was used to analysed the relationships between the subjective evaluation scores and design elements [11]. This technique is also generally used in Japan to examine the relationship between quantitative data (the scores in this work) and qualitative data (the design categories of the samples evaluated). By using the multiple linear regression methods, Hayashi's Quantification Theory Type I (Hayashi, 1950) can statistically predict the relationship between a response value and categorical values. Moreover, in product design, Hayashi's Quantification Theory Type I can also be used to evaluate the weights of the factors form users' preferences [15] [16][17].

The statistical method can be widely applied to research analysis especially for Kansei Engineering. Nagamachi (2008) introduced that Quantification Theory Type I is an excellent technique which is feasible to construct the relationships between design elements and Kansei images. This study also adopted the Quantification Type I method as an analysis tool of the appeal of a CROSSOVER B-CAR interior. In addition, upper-level and lower-level items were measured and quantified

using the importance-level of the original evaluation. Using this method, we were able to analyse the importance of appeal. In addition, the importance of upper-level and lower-level items to the original evaluation were measured and quantified. In particular, the partial correlation coefficients indicate the extent to which each design element contributes to an explanation of the evaluation adjective concerned (Jindo, 1997). In this study, the technique of Quantification Theory Type I was transferred to a type of mathematic formula and was executed through Excel Macro for statistical analysis.

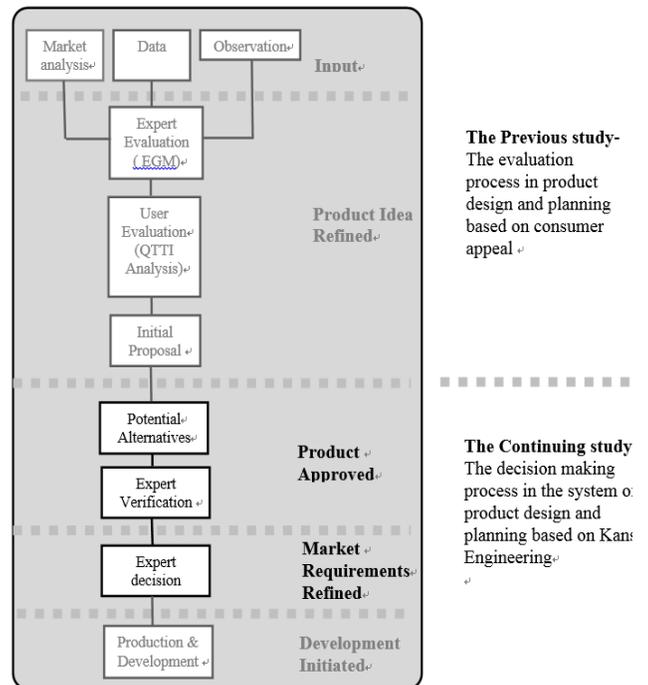


Fig. 2 The system based on Kansei Engineering for the process of product design and planning

IV. RESEARCH OBJECTIVES

This study aims to explore the process of product design and planning based on Kansei Engineering-based AHP. In particular, the application of strategies and the determination of decision-makings are two critical issues to be probed in this study. Therefore, the selections of design criteria and the most proper alternative through careful evaluation in the Kansei Engineering-based system are explored in this study. In addition, expert and consumer opinions could be integrated through the assistance of the system to adopt design decision and strategies. More specifically, design specification was mainly determined according to the results of design evaluation. Then, designers or managers could decide the best alternatives among potential ones.

V.METHODS & RESEARCH PROCEDURES

In this paper, Kansei Engineering was used for extracting critical criteria and factors for the Analytic Hierarchy Process (AHP). In addition, AHP is applied to decide the best alternative for the design of Crossover B-Car Taiwan. Decision

of the appropriate alternatives can be viewed as a complex multi-criteria decision-making problem that requires an extensive evaluation process of the possible alternatives and factors as diverse as design issues. The decision-making process includes the identification of three alternatives of Crossover B-Car interior and 23 criteria grouped into eight critical factors. Five professional editors of auto magazines, six experienced consumers acted as decision makers (DMs). The comparisons between the elements were identified and analyzed using the AHP multi-criteria decision method. The strengths and weaknesses of AHP as a multi-criteria decision analysis tool are also described in the paper. The main findings of this research have proved that Kansei Engineering-based AHP is a useful and efficient tool to help managers to make their decision process traceable and reliable.

This study can be divided into 2 main stages according to general process of product planning (Fig 2). The study reveals the following two stages according to the previous study with the first two stages.

A. Product Approved

Potential alternatives-

In the stage of “product approved”, potential alternatives could be proposed by designers according to the critical appeal factors and design characteristics of Crossover B-Car interior, which were refined through the evaluation process of the previous study. In addition, expert meeting could approve these alternatives by leaving practical ones for the next stage.

In addition, attractive design specifications were determined according to the results of the evaluation conducted through the previous study and were further sifted by experts. Hence, design characteristics with higher “category scores” and appeal factors with higher “partial correlation coefficient” were listed as the priorities for design specification. Furthermore, designers would propose 6 alternatives according to the 8 appeal factors. Each alternative would integrate 3 critical appeal factors with higher “partial correlation coefficient” from the 8 ones and several corresponding design characteristics with higher “category scores”. Hence, designers could apply their creativity to figure out car features for each practical alternative.

Expert verification-

Experts’ decisions can be performed through the assistance of AHP to determine the weights on desired factors in each alternative of Crossover B-Car Interior.

A. Market Requirements Refined

In the stage of “market requirements refined” in the process of product design and development, after specialists’ verification for the potential alternatives, they had to hold a meeting to decide the best alternative from these ones by considering market requirements.

VI. PRACTICES & RESULTS

A. Product Approved

Potential alternatives-

According to the results of my previous study “The evaluation process in product design and planning based on consumer appeal”, the 8 critical appeal factors have being already formed according to EGM. Then, 6 alternatives were disclosed from designer brainstorming and each of them has its own distinctive styles. Furthermore, designers had to select the 3 appeal factors from the 8 ones to present the unique features of each alternative. In addition, the design specifications corresponding to the three factors were mainly determined by the weights of importance based on the result QTTI, including partial correlation coefficients and category scores. Hence, design items or categories with higher scores or partial correlation coefficients were first considered to be left in the list of design specifications. Table I, II, III, IV, V, VI, VII, VIII shows the order of items by partial correlation coefficients corresponding to their categories with the highest scores in the 8 factors. Then, 6 alternatives based on the different consumers’ needs and their critical reasons were proposed. 3 alternatives are listed for examples as the following:

The first one, “Warm and Joyful Family”, is designed for family members by the following aspects.

- Economical and Useful- artificial dark leather and water-proof weave cloth for easy clearance, simple mechanism for durable and stable needs, simple and readable instrument cluster for easy reading, big module and simple mechanics for the purpose of durable and stable needs
- Customized for Consumer needs- pure air for family members’ health
- Flexible and Innovative Space Usage- the variation of door opening and seat folding for family members, utilization of invisible space such as dashboard surrounding and air-conditioning outlets for the needs of various ingenious collection and storage, hooks on the back of seats, the door used as the slope for the convenience of disable passenger getting in/out of the car
- Others- multimedia for family enjoyments, an integrated module of knobs and buttons for easy control, a comfortable steering wheel and gear stick , GPS, audio/video digital entertainment system, referable colors, such as dark, gray or brown.
- Innovative and Extraordinary- other equipment such as power supply and water storage
- Flexible and Innovative Space Usage- various and flexible interior room setting for adapting different activities
- Others- vivid and brilliant color for presenting sporty style, a digital GPS (Global Positioning System) providing the information of a height above sea level and instant road situation, water-proof or dirty-proof texture such as artificial leather or clothing, sporty colors such as orange, yellow, and green, sufficient power supply, the equipment of water storage.

Order ^o	Items ^o	Partial ^o Correlation ^o Coefficients ^o	The Category With The Highest Score In Each Item ^o	Category Scores ^o
1 ^o	Sporty and recreational ^o	0.728 ^o	Vivid and brilliant color ^o	0.131 ^o
2 ^o	Exquisite and quality ^o	0.565 ^o	Metal elements and strips ^o	0.091 ^o
3 ^o	Tasty ^o	0.541 ^o	Leather interior ^o	0.050 ^o

Table. I The order of items by partial correlation coefficients corresponding to their categories with the highest scores in “Fashionably-designed”

Order ^o	Items ^o	Partial ^o Correlation ^o Coefficients ^o	The Category With The Highest Score In Each Item ^o	Category Scores ^o
1 ^o	The space conforming user needs ^o	0.647 ^o	Family space ^o	0.095 ^o
2 ^o	Sufficient illumination ^o	0.638 ^o	Light attached to the rear view mirror ^o	0.106 ^o
3 ^o	Conforming to user habits ^o	0.602 ^o	Adjustable exits of air conditionals ^o	0.120 ^o
4 ^o	Conforming to ergonomics ^o	0.558 ^o	Real seats capable of leaning and adjusting ^o	0.242 ^o

Table. II The order of items by partial correlation coefficients corresponding to their categories with the highest scores in “Comfortable and convenient”

Order ^o	Items ^o	Partial ^o Correlation ^o Coefficients ^o	The Category With The Highest Score In Each Item ^o	Category Scores ^o
1 ^o	Safe and stable ^o	0.781 ^o	Tranquility inside the car ^o	0.103 ^o
2 ^o	Worth more than its cost ^o	0.652 ^o	For daily use ^o	0.174 ^o
3 ^o	Endurable ^o	0.137 ^o	Leather ^o	0.011 ^o

Table. III The order of items by partial correlation coefficients corresponding to their categories with the highest scores in “Economical and useful”

Order ^o	Items ^o	Partial ^o Correlation ^o Coefficients ^o	The Category With The Highest Score In Each Item ^o	Category Scores ^o
1 ^o	Customized for the needs of the specific population ^o	0.816 ^o	Design for enthusiasts of sport or recreation ^o	0.230 ^o
2 ^o	Customized for individual preferences ^o	0.519 ^o	Optional look style of seats ^o	0.158 ^o

Table.IV The order of items by partial correlation coefficients corresponding to their categories with the highest scores in “Customized for consumer needs”

Order ^o	Items ^o	Partial ^o Correlation ^o Coefficients ^o	The Category With The Highest Score In Each Item ^o	Category Scores ^o
1 ^o	Digitization ^o	0.800 ^o	The application of digital technology, such as digital GPS ^o	0.194 ^o
2 ^o	Technical metallic sense ^o	0.595 ^o	Blue LED ^o	0.077 ^o
3 ^o	3C ^o	0.320 ^o	Peripheral equipment supporting 3C products ^o	0.138 ^o

Table.V The order of items by partial correlation coefficients corresponding to their categories with the highest scores in “Technical”

Order ^o	Items ^o	Partial ^o Correlation ^o Coefficients ^o	The Category With The Highest Score In Each Item ^o	Category Scores ^o
1 ^o	Outstanding sense of touch ^o	0.608 ^o	The knob with oil pressure ^o	0.071 ^o
2 ^o	Outstanding sense of sight ^o	0.416 ^o	Well-designed form, such as the top form of a gear stick ^o	0.102 ^o
3 ^o	Outstanding sense of smell ^o	0.416 ^o	The scent of boutique ^o	0.067 ^o
4 ^o	Outstanding sense of hearing ^o	0.385 ^o	The tranquility inside car interior ^o	0.052 ^o

Table.VI The order of items by partial correlation coefficients corresponding to their categories with the highest scores in “Extraordinarily Experienced”

Order ^o	Items ^o	Partial ^o Correlation ^o Coefficients ^o	The Category With The Highest Score In Each Item ^o	Category Scores ^o
1 ^o	Extraordinary ^o	0.788 ^o	Changeable interior style ^o	0.196 ^o
2 ^o	Innovative ^o	0.492 ^o	Innovative color match ^o	0.095 ^o

Table.VII The order of items by partial correlation coefficients corresponding to their categories with the highest scores in “Innovative and Extraordinary”

Order ^o	Items ^o	Partial ^o Correlation ^o Coefficients ^o	The Category With The Highest Score In Each Item ^o	Category Scores ^o
1 ^o	Integrated and various ingenious collection and storage ^o	0.729 ^o	Integrated storage space ^o	0.115 ^o
2 ^o	Various and flexible interior room setting ^o	0.692 ^o	Front seat variation ^o	0.102 ^o

Table.VIII The order of items by partial correlation coefficients corresponding to their categories with the highest scores in “Flexible and Innovative Space Usage”

The third one, “Technical and Pleasurable E Generation” style, mainly appeals to the application of technological multi-function to achieve the goal of integrating humans’ behavior and thinking with it and focuses on the following three aspects. Then, the sense of technical metallic, fashion, modern and future can be performed through metal elements, quality soft plastic, lively patterns, and grains and dots.

- Technological- a digital system, supply for connecting with popular and fashionable PDA (Person Digital Assistant) or minicomputer for building a multimedia entertainment center, technological metal texture for presenting cool, tough, young and dynamic motion sense , digital technological system of application and multimedia system , a digitized multi-functional instrument cluster integrating road seating, the application of mobile communication

	1 ^o	1 ^o	1 ^o	2 ^o	2 ^o	2 ^o	5 ^o	5 ^o	1.94 ^o	0.20 ^o
Economical and useful ^o	1 ^o	1 ^o	1 ^o	2 ^o	2 ^o	2 ^o	5 ^o	5 ^o	1.94 ^o	0.20 ^o
Flexible and Innovative Space Usage ^o	1 ^o	1 ^o	1 ^o	2 ^o	2 ^o	2 ^o	5 ^o	5 ^o	1.94 ^o	0.20 ^o
Customized for consumer needs ^o	1 ^o	1 ^o	1 ^o	2 ^o	2 ^o	2 ^o	5 ^o	5 ^o	1.94 ^o	0.20 ^o
Innovative and Extraordinary ^o	1/2 ^o	1/2 ^o	1/2 ^o	1 ^o	1 ^o	1 ^o	3 ^o	3 ^o	1.01 ^o	0.11 ^o
Comfortable and convenient ^o	1/2 ^o	1/2 ^o	1/2 ^o	1 ^o	1 ^o	1 ^o	3 ^o	3 ^o	1.01 ^o	0.11 ^o
Technical ^o	1/2 ^o	1/2 ^o	1/2 ^o	1 ^o	1 ^o	1 ^o	3 ^o	3 ^o	1.01 ^o	0.11 ^o
Extraordinarily Experienced ^o	1/5 ^o	1/5 ^o	1/5 ^o	1/3 ^o	1/3 ^o	1/3 ^o	1 ^o	1 ^o	0.36 ^o	0.04 ^o
Fashionably-designed ^o	1/5 ^o	1/5 ^o	1/5 ^o	1/3 ^o	1/3 ^o	1/3 ^o	1 ^o	1 ^o	0.36 ^o	0.04 ^o

Table.IX the radar figure of “warm and joyful family”

- Customized for Consumer needs- touch panel used for facilitating users ‘operating and shrinking users’ controlling interface to accommodate personal digital devices, such as a laptop or mobile phone, cool color or two-color series

- Flexible and Innovative Space Usage- collection and storage for personal digital equipment,

- Others- illumination used for lighting and making cool atmosphere such as LED and blue light, integrated controllers for easy control

Expert Verification-

Three chief editors of auto publication and two experienced

designers are invited for evaluating the six alternatives based on their specialized fields. Experts had to decide the weights of factors for each alternative by AHP. After getting the weights, radar charts (Fig 3) which showed unique features with the advantages of each alternative were made. Furthermore, the best three alternatives plus their corresponding factors with weights

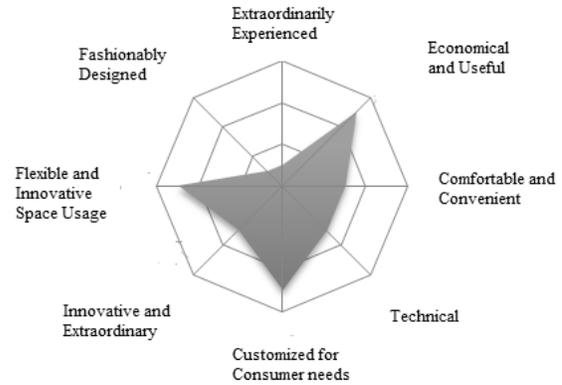


Fig. 4 The radar figure of “warm and joyful family”

which were determined by experts would further proceed to “market requirement refined” in this study. As a result, “recreational and Lohas Explorer”, “technical and Pleasurable E Generation”, and “warm and joyful family” were selected as the final alternatives as following:

First, “Warm and Joyful Family” was recommended because many consumers with family have the need of not only of large interior capacity but also small car exterior body. Hence, although more passengers can be accommodated for family member, it would not be at the cost of the good flexibility of car body for driving or parking. Hence, “warm and Joyful Family”, is designed for family members. The weights of the 8 critical types, based on the result of AHP (Table IX), are showed in the radar chart (Fig 4).

- Economical and Useful (20%)- Design specifications were mainly determined according to the analytical results of QTTI. Determined design specifications included artificial dark leather and water-proof weave cloth for easy clearance, simple mechanism for durable and stable needs, simple and readable instrument cluster for easy reading, big module and simple mechanics for the purpose of durable and stable needs
- Customized for Consumer needs (20%)- Design specifications were mainly determined according to the analytical results of QTTI. Determined design specifications included pure air for family members’ health
- Flexible and Innovative Space Usage (20%)- Design specifications were mainly determined according to the analytical results of QTTI. Determined design specifications included the variation of door opening and seat folding for family members, utilization of invisible space such as dashboard surrounding and air-conditioning outlets for the

	1 ^o	1 ^o	1 ^o	2 ^o	2 ^o	5 ^o	5 ^o	5 ^o	2.17 ^o	0.22 ^o
Comfortable and convenient ^o	1 ^o	1 ^o	1 ^o	2 ^o	2 ^o	5 ^o	5 ^o	5 ^o	2.17 ^o	0.22 ^o
Flexible and Innovative Space Usage ^o	1 ^o	1 ^o	1 ^o	2 ^o	2 ^o	5 ^o	5 ^o	5 ^o	2.17 ^o	0.22 ^o
Innovative and Extraordinary ^o	1 ^o	1 ^o	1 ^o	2 ^o	2 ^o	5 ^o	5 ^o	5 ^o	2.17 ^o	0.22 ^o
Economical and useful ^o	1/2 ^o	1/2 ^o	1/2 ^o	1 ^o	1 ^o	3 ^o	3 ^o	3 ^o	1.16 ^o	0.12 ^o
Customized for consumer needs ^o	1/2 ^o	1/2 ^o	1/2 ^o	1 ^o	1 ^o	3 ^o	3 ^o	3 ^o	1.16 ^o	0.12 ^o
Technical ^o	1/5 ^o	1/5 ^o	1/5 ^o	1/3 ^o	1/3 ^o	1 ^o	1 ^o	1 ^o	0.42 ^o	0.04 ^o
Extraordinarily Experienced ^o	1/5 ^o	1/5 ^o	1/5 ^o	1/3 ^o	1/3 ^o	1 ^o	1 ^o	1 ^o	0.42 ^o	0.04 ^o
Fashionably-designed ^o	1/5 ^o	1/5 ^o	1/5 ^o	1/3 ^o	1/3 ^o	1 ^o	1 ^o	1 ^o	0.42 ^o	0.04 ^o

Table. X The weights of 8 critical factors in “recreational and Lohas explorer” by AHP

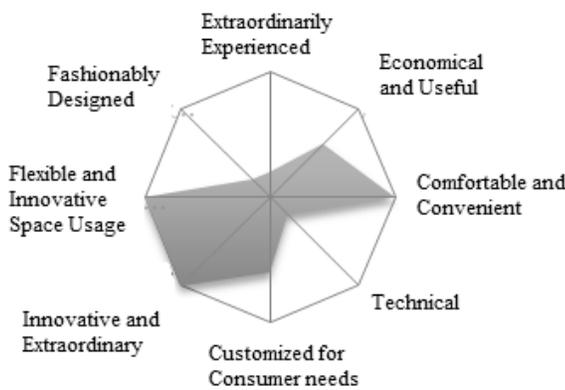


Fig. 5 The radar figure of “recreational and Lohas explorer”

needs of various ingenious collection and storage, hooks on the back of seats, the door used as the slope for the convenience of disable passenger getting in/out of the car

- Others (40%)- Design specifications were mainly determined according to the analytical results of QTTI. Determined design specifications included multimedia for family enjoyments, an integrated module of knobs and buttons for easy control, a comfortable steering wheel and gear stick, GPS, audio/video digital entertainment system, referable colors, such as dark, gray or brown.

Secondarily, “Recreational and Lohas Explorer” was recommended because many consumers with open mind desire to do outdoor activities so they need large as well as dirty and

scratch proof interior space with enough power supply and manageable storage. Hence, “Recreational and Lohas Explorer” style, mainly appeals to the styles which satisfies people who are engaged in outdoor recreational activities considers three aspects from consumers’ needs, including the three following. The weights of the 8 critical types are showed in the radar chart (Fig 5) and (Table X).

- Comfortable and Convenient (22%)- Design specifications were mainly determined according to the analytical results of QTTI. Determined design specifications included enough illumination convenient for outdoor activities, a convenient and ergonomic gear, artificial leather or water-proof weave cloth, comfortable and controllable steering wheel and gear stick, simple and readable instrument cluster.

- Innovative and Extraordinary (22%)- Design specifications were mainly determined according to the analytical results of QTTI. Determined design specifications included equipment such as power supply and water storage.

- Flexible and Innovative Space Usage (22%)- Design specifications were determined according to the analytical results of QTTI. Determined design specifications included various and flexible interior room setting for adapting different activities.

- Others (34%)- Design specifications were mainly determined according to the analytical results of QTTI. Determined design specifications included vivid and brilliant color for presenting sporty style, a digital GPS (Global Positioning System) providing the information of a height above sea level and instant road situation, water-proof or dirty-proof texture such as artificial leather or clothing, sporty color such as orange, yellow, and green, sufficient power supply, the equipment of water storage.

Thirdly, “technical and pleasurable E Generation” was recommended because E generation welcome new electronic product such as PDA, tablet PC, and mobile phone. Hence, car interior with the capability of supporting the new products and mobile communication is a potentially orientation for development. Hence, “Technical and Pleasurable E Generation” style, mainly appeals to the application of technological multi-functions to achieve the goal of integrating humans’ behavior and thinking with it and focuses on the following three aspects. Then, the sense of technical metallic, fashion, modern and future can be performed through metal elements, quality soft plastic, and lively patterns, grains and dots. The weights of the 8 critical types can be showed in the radar chart (Fig 6) and table (Table XI).

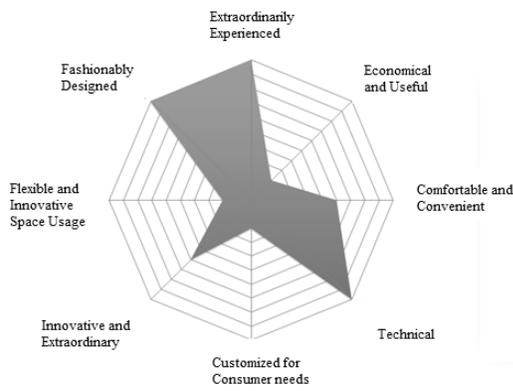


Fig. 6 the radar figure of “Technical and Pleasurable E Generation”

	Extraordinarily Experienced ^a	Economical and useful ^a	Comfortable and convenient ^a	Innovative and Extraordinary ^a	Fashionably-designed ^a	Technical ^a	Customized for consumer needs ^a	Flexible and Innovative Space Usage ^a
Warm and Joyful Family ^a	0.13 ^a	0.00 ^a	0.01 ^a	0.06 ^a	0.06 ^a	0.01 ^a	0.01 ^a	0.04 ^a
Recreational and Lohas Explorer ^a	0.17 ^a	0.00 ^a	0.01 ^a	0.00 ^a	0.00 ^a	0.03 ^a	0.05 ^a	0.05 ^a
Technical and Pleasurable E Generation ^a	0.11 ^a	0.01 ^a	0.00 ^a	0.01 ^a	0.02 ^a	0.02 ^a	0.03 ^a	0.01 ^a

Table. XV The scores of the 3 alternatives by calculating the weights of the 8 factors of AHP

- Technological (22%)- Design specifications were mainly determined according to the analytical results of QTTI. Determined design specifications included a digital system, supply for connecting with popular and fashionable PDA (Person Digital Assistant) or minicomputer for building a multimedia entertainment center, technological metal texture for presenting cool, tough, young and dynamic motion sense, digital technological system of application and multimedia system, a digitized multi-functional instrument cluster integrating road seating, the application of mobile communication

- Customized for Consumer needs (22%)-Design specifications were mainly determined according to the analytical results of QTTI. Determined design specifications included touch panel used for facilitating users ‘operating and shrinking users’ controlling interface to accommodate personal digital devices, such as a laptop or mobile phone, cool color or

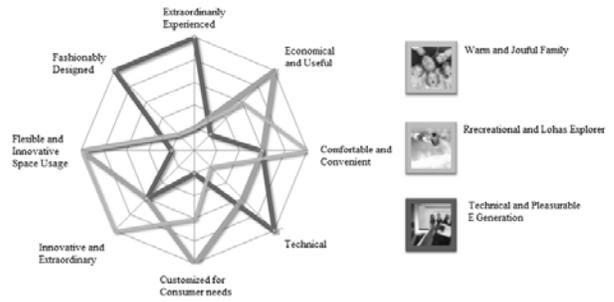


Fig. 7 The radar figure of the three target population

two-color series.

- Flexible and Innovative Space Usage (22%)- Design specifications were mainly determined according to the analytical results of QTTI. Determined design specifications included collection and storage for personal digital equipment.

- Others (34%)- Design specifications were mainly determined according to the analytical results of QTTI. Determined design specifications included illumination used for lighting and making cool atmosphere such as LED and blue light, integrated controllers for easy control.

In addition, the radar charts showing the 3 alternatives (Fig 7) reveals common and different points among them based on the variable weights of 8 critical types. Then, according to the market analysis and experts’ opinions, the strategies for car size adopt the principle of longer wheelbase and exterior width, as well as shorter exterior length to create larger interior space and smaller exterior volume. Hence, a comfortable interior space and nimble exterior volume are created for the characteristics of CROSSOVER B-CAR.

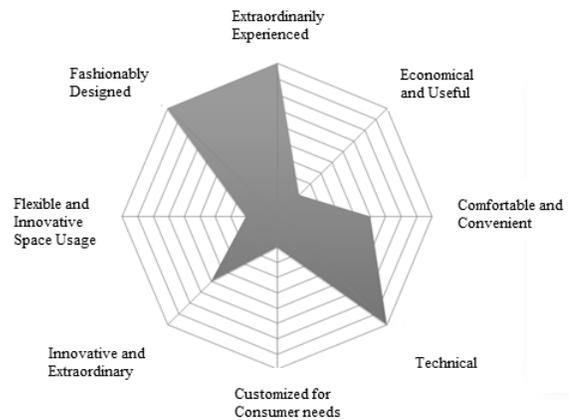


Fig. 6 The radar figure of “Technical and Pleasurable E Generation”

REFERENCES

- [1] Rouse, W. B., 1986, A note on the nature of creativity in engineering: implications for supporting system design. *Information Processing and Management*, 22, 279± 285.
- [2] Chen, C.H., Occena, L. G., & Fok, S. C. (2001), CONDENSE: a concurrent design evaluation system for product design, *International Journal of Production Research*, 39(3), 413-433
- [3] Sun, Y., Zang, Y.F., & Nee, A.Y.C. (2010), A distributed multi-agent environment for product design and manufacturing planning, *International Journal of Production Research*, 39(4), 625-645
- [4] Walsh, V., Roy, R., Bruch, M., Potter, S., 1992, *Winning by Design: Technology, Product Design and International Competitiveness* (Oxford: Blackwells).
- [5] Ma, C.H., 1982, *Industrial Design: The Creativity and Implementation of Product Development* (Taiwan: Long San).
- [6] Pahl, G., & Beitz, W., 1996, *Engineering Design: A Systematic Approach* (London: Springer).
- [7] Cooper, R.G., *Industrial firms' new product strategies*. *J. Business Res.*, 1985, 13(2), 107–122.
- [8] Kotler, P., *Marketing Management: Analysis, Planning, Implementation, and Control*, 1991 (Prentice-Hall:Englewood Cliffs, NJ).
- [9] Baxter, M., *Product Design: A Practical Guide to Systematic Methods of New Product Development*, 1995 (Chapman & Hall: London).
- [10] Nagamachi, M. (2002) "Kansei engineering as a powerful consumer-oriented technology for product development", *Applied Ergonomics*, 33, pp. 289-294
- [11] Jindo, T., & Hirasago, K.(1997), Application studies to car interior of Kansei engineering , *Industrial Ergonomic* 19, 105-114
- [12] Nagamachi, M. (1995), "Kansei Engineering: A new ergonomic consumer-oriented technology for product development", *International Journal of Industrial Ergonomics*, 15, pp. 3-11.
- [13] Nagamachi, M. (2008) "Perspectives and the new trend of Kansai/affective engineering", *The TQM Journal*, Vol. 20 Iss:4, pp.290-298
- [14] Saaty, T. L. (1990), "How to make a decision: The analytic hierarchy process ", *European Journal of Operational Research*, Vol. 48 Iss:1, pp. 9-26
- [15] Hayashi, C. (1950) On the Quantification of Qualitative Data from the Mathematico-Statistical Point of view, *Annals of the Institute of Statistical Mathematics*, Vol. 2.
- [16] Iwabuchi, C. et. al, (2001) *Data Management and Analysis by Yourself*, Japan: Humura publishing, pp180-185.
- [17] Sugiyama, K. et. al, (1996) *The basic for Survey and Analysis by Excel*, Japan: Kaibundo publishing, pp.51-62.

	Extraordinarily Experienced ^o	Customized for consumer needs ^o	Technical ^o	Fashionably-designed ^o	Innovative and Extraordinary ^o	Comfortable and convenient ^o	Economical and useful ^o	Flexible and Innovative Space Usage ^o
Flexible and Innovative Space Usage ^o	0.03 ^o	0.06 ^o	0.06 ^o	0.07 ^o	0.13 ^o	0.21 ^o	0.22 ^o	0.22 ^o
Economical and useful ^o	0.15 ^o	0.24 ^o	0.24 ^o	0.24 ^o	0.26 ^o	0.21 ^o	0.22 ^o	0.22 ^o
Comfortable and convenient ^o	1.89 ^o	0.15 ^o	0.15 ^o	0.24 ^o	0.21 ^o	0.39 ^o	0.21 ^o	0.11 ^o
Innovative and Extraordinary ^o	0.18 ^o	0.18 ^o	0.18 ^o	0.18 ^o	0.21 ^o	0.07 ^o	0.11 ^o	0.11 ^o
Fashionably-designed ^o	0.09 ^o	0.06 ^o	0.06 ^o	0.06 ^o	0.07 ^o	0.04 ^o	0.07 ^o	0.11 ^o
Technical ^o	0.09 ^o	0.06 ^o	0.06 ^o	0.06 ^o	0.07 ^o	0.05 ^o	0.06 ^o	0.06 ^o
Customized for consumer needs ^o	0.09 ^o	0.06 ^o	0.06 ^o	0.06 ^o	0.07 ^o	0.04 ^o	0.06 ^o	0.06 ^o
Extraordinarily Experienced ^o	0.03 ^o	0.02 ^o	0.02 ^o	0.02 ^o	0.03 ^o	0.02 ^o	0.04 ^o	0.04 ^o

Consistency Index = (8.28-8)/(8-1)=0.04

Table. XIII The scores of 8 critical factors by calculating the weights of AHP

	Warm and Joyful Family ^o	Recreational and Lohas Explorer ^o	Technical and Pleasurable E Generation ^o
Warm and Joyful Family ^o	0.20 ^o	0.20 ^o	0.20 ^o
Recreational and Lohas Explorer ^o	0.12 ^o	0.22 ^o	0.12 ^o
Technical and Pleasurable E Generation ^o	0.04 ^o	0.12 ^o	0.04 ^o

Table. XIV The weights of 3 alternatives by calculating the weights of the 8 factors by AHP

Longitudinal dispersion coefficient as sensitivity parameter in water quality simulation model

Yveta Velísková, Marek Sokáč

Abstract— Paper deals with sensitivity analysis of the outputs from a numerical model MIKE 11 (one dimensional model) related to values of the longitudinal dispersion coefficients, used as model input data. These coefficients are one of the most important characteristics, which impact the pollution dispersion modelling in rivers. Determination of their correct values was a subject of previous research studies, now the results and experience from these studies are used in this contribution. A large set of numerical simulations were performed with various values of the longitudinal dispersion coefficient ($D= 5 - 50 \text{ m}^2.\text{s}^{-1}$). Simulations were performed at the Hron River in the part from Slovenská Ľupča to the river estuary. To separate the flow velocity contribution on the total pollutant dispersion process, two series of numerical simulations were performed: with the dispersion coefficients values depending on the flow velocity and dispersion coefficient values, which are independent on the river flow velocity. All mentioned simulations were performed at two hydrologic conditions: annual discharge (Q_a) along the modelled river part and with the nearby minimum flow situation (Q_{355}). Results, achieved till today, confirm relatively high model sensitivity on input values of dispersion coefficients. The influence rate is discussed in the contribution and documented by figures.

Keywords—dispersion, numerical simulation, sensitivity analysis, surface water.

I. INTRODUCTION

The present legislation evaluating quality of water bodies (WF) in Slovakia is based on implementation of the Water Framework Directive (2000/60/ES). Concerning the Directive it is required eco-morphological monitoring of WF, which is based on evaluation of the rate of anthropogenic impact. It does not refer only to river bed, but also the state of environment nearby to stream is taking into consideration.

Dispersion coefficients are the crucial input parameters of transport processes models [2], [5], [6]. Aim of this contribution is description of fact how the values of longitudinal dispersion coefficients, used as a one-

dimensional model input data, affect the results of water quality numerical simulation.

II. THEORETICAL BASIS

In this chapter we are going to discuss the basic theoretical terms and relationships related to dispersion in open streams only to the extent important in relation to the scope of the article.

One-dimensional advection-diffusion equation is the simplest mathematical formulation of dispersion:

$$\frac{\partial AC}{\partial t} + \frac{\partial QC}{\partial x} - \frac{\partial}{\partial x} \left(AD \frac{\partial C}{\partial x} \right) = -AKC + C_s \cdot q \quad (1)$$

where C is a concentration of relevant substance (mg.l^{-1}), D is a dispersion coefficient ($\text{m}^2.\text{s}^{-1}$), A is a flow area (m^2), Q is a discharge in a stream ($\text{m}^3.\text{s}^{-1}$), K is a coefficient expressing the effect of chemical and biological processes on dissolved substance (s^{-1}), C_s is a concentration of pollutant source, q is a discharge of source, x is a length (m) and t is time (s).

This equation includes two basic transport mechanisms:

1. advective (or convective) transport caused by fluid flow
2. dispersion transport caused by concentration gradient

Advection- diffusion equation is based on the assumptions, that

- a substance under consideration is homogeneously distributed over the cross section and an ideal mixing (immediate homogenisation in cross section) is taken into account even for resource/abstraction of substance,
- a substance is conservative (is not subject to chemical and biological processes) or its interaction with environment can be described using the first order differential equation:

$$\frac{dC}{dt} = K \cdot C \quad (2)$$

- Fick's law of diffusion is applied, i.e. dispersion transport is proportional to the concentration gradient.

The first assumption, namely ideal mixing and homogenous distribution of substances in cross section, is a result of one-dimensional description of model area using the MIKE 11 model [3]. Such description is applicable where one dimension predominates over other (for example the river

This work was supported in part by the APVV, grant nr. 0274-10, Centre of excellence ITMS 26240120004 and 7th FP project „Goldfish”, FP7-ICT-2009-6, Grant agreement no: 269985.

Yveta Velísková is with the Institute of Hydrology Slovak Academy of Science, Bratislava, phone +421 2 49268 280, e-mail: veliskova@uh.savba.sk

Marek Sokáč is with the Department of Sanitary and Environmental Engineering, Faculty of Civil Engineering, Slovak University of Technology, Bratislava, e-mail: marek.sokac@stuba.sk

where the length is a determining dimension and the phase of cross-sectional mixing can be neglected). However, this assumption cannot be applied to water reservoirs in any case where distribution of hydraulic parameters over the depth and width of flow cross section plays a significant role (this situation requires at least two or three-dimensional simulation).

However, one-dimensional models for simulation mixing in streams need the values of longitudinal dispersion coefficient. As it follows from references [1-8], this coefficient derivation is achieved by several ways: from the own experience or that from the references, over the qualified estimates, up the special calculations application. As dispersion coefficient value is determined by the turbulence intensity in the given stream section, its magnitude depends upon its main hydraulic characteristics: form and magnitude of its cross section profile, its flow velocity and its longitudinal slope.

III. MODEL MIKE

MIKE 11, developed by the Danish Hydraulic Institute, is a set of modules for simulation of flow, water quality and sediment transport in rivers, channels, river mouths, irrigation systems and other surface water resources.

It is a dynamic one-dimensional model tool for detailed proposal and management of simple and complex river and channel systems. Simulation results can be used in engineering, water management, water quality management and planning applications.

MIKE 11 consists of several modules for: hydrodynamics (basic module), hydrology, cohesive sediment transport, water quality and non-cohesive sediment transport. In this study only the basic hydrodynamic and advection-dispersion transport modules are used for simulation. Generally, a module structure of the model offers great flexibility: each of modules can be processed separately; data transfer among modules is automated; coupling of physical processes is easier; quick and simple application and development of new modules is possible.

The type of database technology used in the model provides efficient data storage and retrieval. Data organization and handling is unified within the whole model system. MIKE 11 is operated through the interactive control menu system.

Hydrodynamic module is a basic computational module required to run other computational modules. The module simulates unsteady flow in streams through the finite difference method by using implicit computational scheme. The computational scheme is applicable to vertically homogenous stream in conditions of subcritical and supercritical flows. The hydrodynamic module requires the most data from all modules that are used. The data are stored in several datasets and in some cases it is possible to import data from other formats (e.g. text files). The data can be categorized into two large groups:

1. geometric data (describing stream channel geometry, dimensions and topology) and

2. hydraulic parameters (hydraulic roughness of stream channel, etc.)

Advection-dispersion module is based on the one-dimensional equation of mass conservation of dissolved and suspended solids (e.g. salts or cohesive sediments). This module requires the outputs of hydrodynamic module in time and space (discharges, water levels, hydraulic radius and flow areas). Advection-dispersion equation (1) is solved numerically using implicit scheme with the finite differences. Suddenly changed concentrations can be simulated in this way. For advection-dispersion module, it is important to specify the substances considered in simulation and the value of dispersion coefficient. This value can be specified according to the equation:

$$D = a \cdot v^b \quad (3)$$

where a is a dispersion factor and v is a flow velocity (m/s). The resulting value of dispersion coefficient can be defined by setting the minimum and maximum values.

IV. SIMULATION OF POLLUTION TRANSPORT IN HRON RIVER

Simulations for the study of the effect of dispersion coefficient and pollutant concentrations along the river were carried out using the numerical model of the Hron River in the section from Slovenská Ľupča (river km 183.84) up to the river mouth (river km 0.0). Numerical simulations were performed using the above-mentioned MIKE 11 model. For all alternatives of simulation we took into consideration a uniform load of receiving body (the same concentration of pollutant). The ammonia pollution was taken into account in simulation and it was specified as a concentration of ammonia nitrogen $N-NH_4$ in the model. The model way of getting the pollution into the river is similar to accidental pollution of river, i.e. discharge (release) of larger wastewater volumes with relatively high concentrations of ammonia nitrogen (800 mg/l) during two hours. It is clear that these inputs to the numerical model represent extreme emergency situation. However, our objective is to analyse the sensitivity of simulation results to data used as model inputs regardless of the impacts of accident on water quality in a stream. To clarify this issue it is important to note that the decrease in concentrations (see charts in Fig.4) is caused not only by the dispersion but also by mixing the water in the Hron River with its tributaries (dilution) and by biological oxidation processes that are included in the MIKE 11 model (water quality module – WQ). The objective of this article is to determine the sensitivity of numerical models to input values of the longitudinal dispersion coefficient while maintaining other processes having an effect on changes in quality of water in stream (physical, chemical and biological processes of water self-purification).

The value of dispersion coefficient D was selected through the dispersion factor a and exponent b according to (3). The values in the range from 5 to 50 were used for dispersion

factor a . Two series of numerical simulations were performed in order to eliminate the effect of flow velocity on pollution dispersion from the results of simulation. One series was done using the exponent value in (3) $b = 0$, and thus the value of dispersion coefficient did not depend on flow velocity. The value of exponent in the second series of trials was $b = 1$ and the dispersion coefficient was a function of flow velocity according to (3).

All the mentioned simulations were carried out in two alternatives – hydrological conditions: average annual discharge Q_a in particular cross sections along the whole length of river, and the discharge close to the minimum value - Q_{355} .

V. OUTCOMES AND DISCUSSION

We have obtained a number of curves showing the distribution of concentration along the whole length of the monitored section of river. Then the curves were analyzed. The analysis was aimed at the change in distribution of concentrations at different values of the dispersion coefficient while eliminating the effect of flow velocity ($b=0$). One of such cases is shown in Figure 1 (for Q_a and Q_{355}). It is apparent that the time-concentration curves in considered cross section differ in dispersion coefficient value from each other. The model responds to the changed value of dispersion characteristics. The values of maximum concentrations also differ from each other regarding the value D and definitely the discharge condition applied to particular model situation.

Other two figures - figures 2 and 3 – show the comparison of simulation results proving the effect of dispersion coefficient, whether it is the function of flow velocity in a given river section or not. According to (3), when b equals 0, than the value of dispersion coefficient D taken into account in simulation is not affected by the flow velocity. On the contrary, if $b \neq 0$, than the value of dispersion coefficient becomes the function of flow velocity and thus it is influenced by this parameter. The results of the simulations show (see figure) that such method of setting the values has an effect on distribution of pollutant concentration in stream – maximum concentration values and concentration distribution curve shape are changed. These differences are less significant for Q_a ($Q_a > Q_{355}$) than for Q_{355} . That is due to the fact that flow velocity in stream at Q_a is higher and close to the value of 1 m/s compared to Q_{355} where the velocities are considerably lower. However, it is important to take into account the effect of dispersion factor a of (3) on the value of dispersion coefficient D .

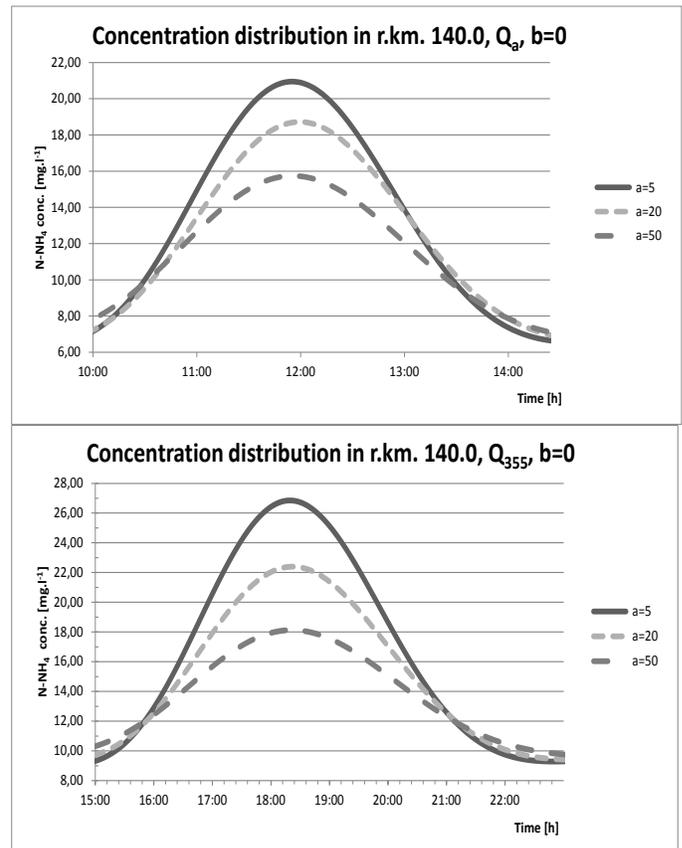


Fig. 1. Concentration curves at river kilometer 140.0 and Q_a and Q_{355} ($b=0$)

Another output of the series of simulations was the distribution of maximum concentrations along the simulated section of the Hron River. This graphical output has a great significance in the assessment of the change in water quality due to potential accidental pollution.

The Figure 4 indicates the distribution of concentrations for both simulated discharges (Q_a and Q_{355}) as well as for different values of dispersion coefficient D (5; 20; 50). As shown in the figure, the effect of selected value for dispersion coefficient is evident. Despite the fact that this case is a theoretical simulation of accidental pollutant discharge into a stream and the concentrations indicated in the charts are also only theoretical (illustrative), it is important to be aware of the differences in maximum concentrations at particular river kilometer for different dispersion coefficient values. Such differences could be of crucial importance to river biota in a real situation.

VI. CONCLUSION

The article deals with the sensitivity of MIKE numerical model outputs to used values of dispersion coefficients. These

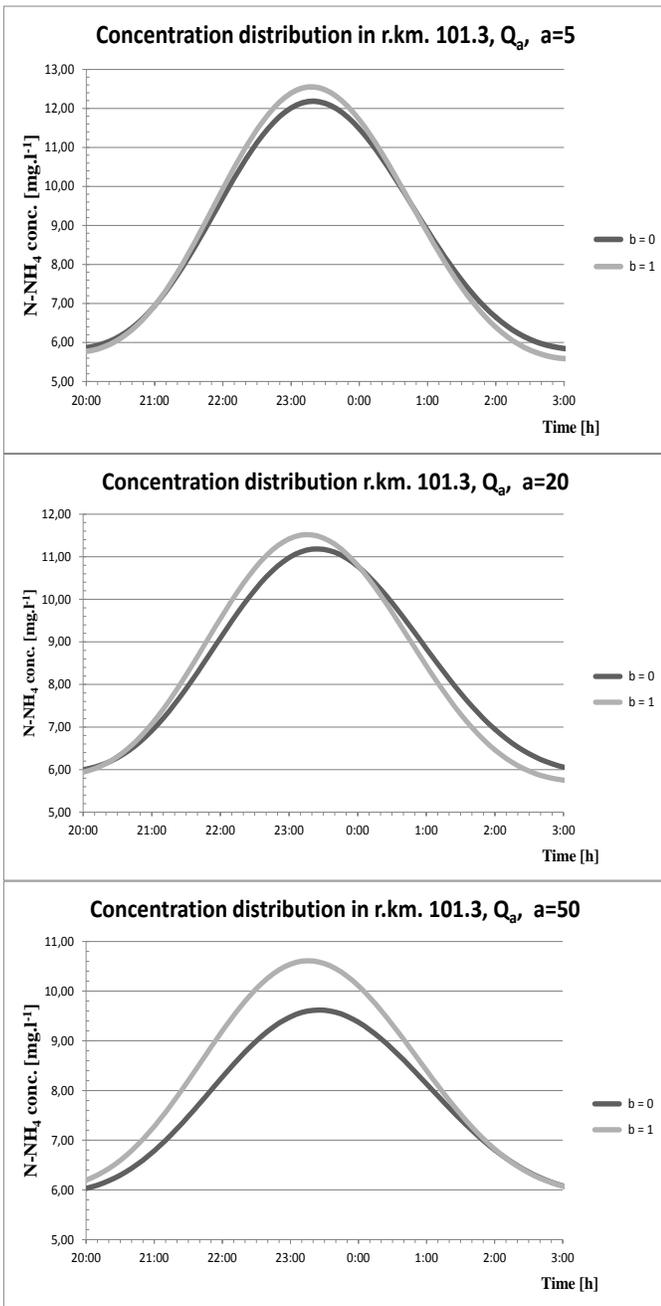


Fig. 2 Concentration curve at river kilometer 101.3 for different values of coefficient a (equation 3) (considering the effect of flow velocity $b=1$ or without the effect of flow velocity $b=0$) - discharge Q_a

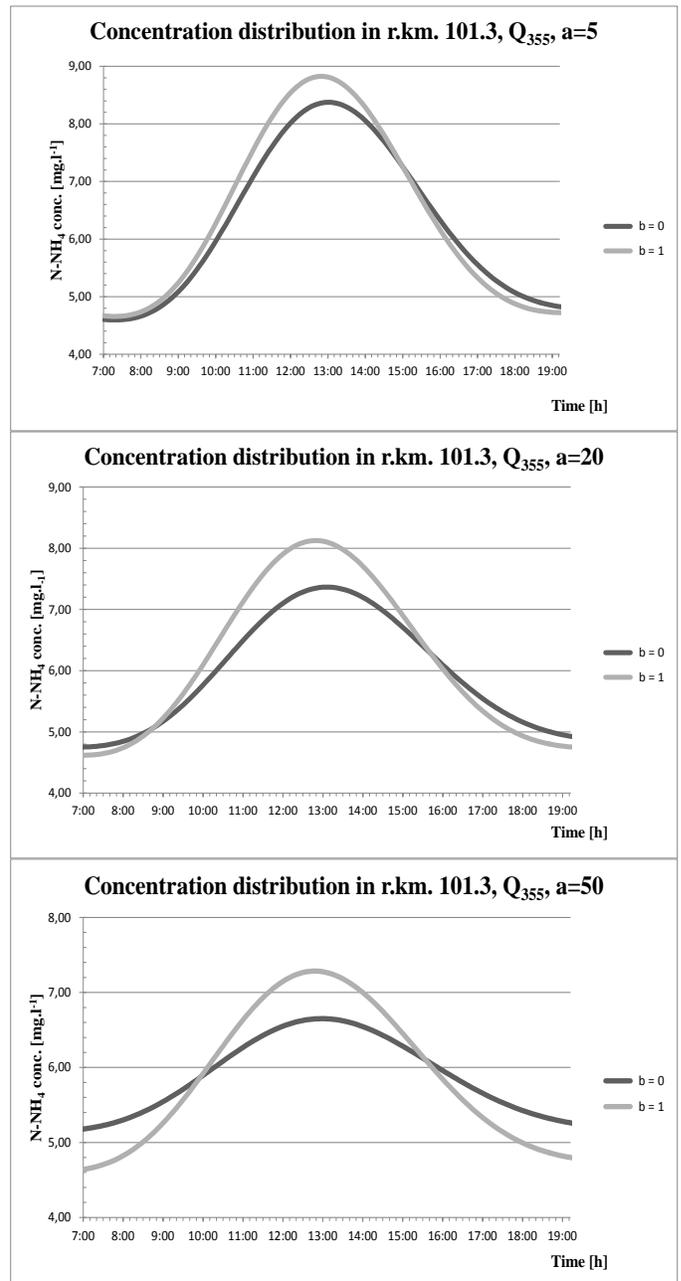


Fig. 3 Concentration curve at river kilometer 101.3 for different values of coefficient a (equation 3) (considering the effect of flow velocity $b=1$ or without the effect of flow velocity $b=0$) – discharge Q_{355}

coefficients are one of the main characteristics affecting the dispersion phenomenon in a stream.

Figure 1 clearly shows how this phenomenon affects pollution transport – when the value of dispersion coefficient is low, the dispersion phenomenon is minimal and concentrations of transported substance are higher than in case when the dispersion is applied in a larger extent.

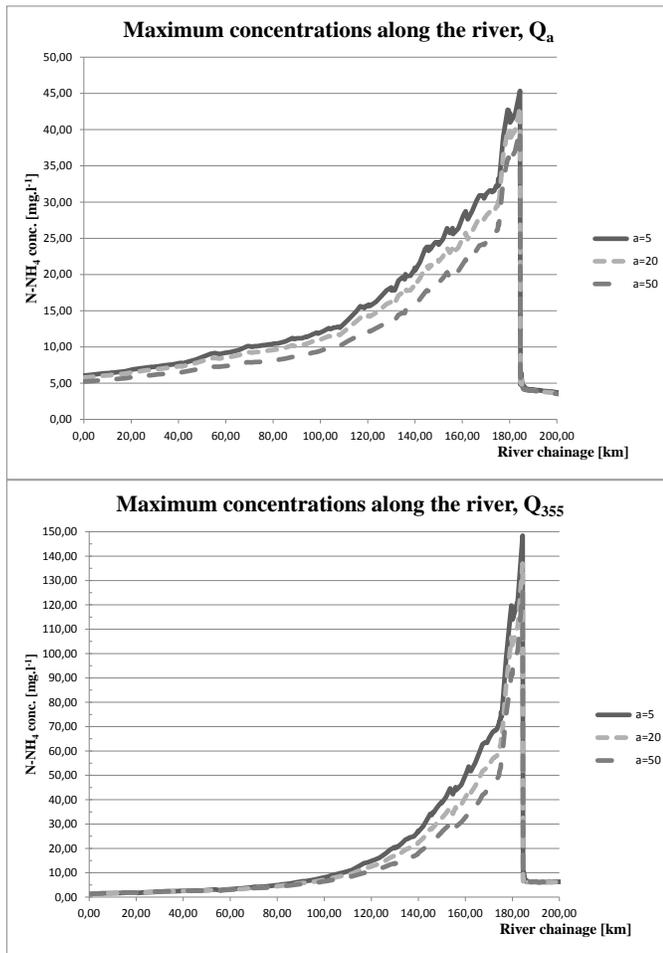


Fig. 4 Maximum concentration curve along the simulated section of the Hron River

As it was stated previously, if the real situation in stream is assessed, the differences in the values of maximum concentrations at particular river kilometers for different values of dispersion could be of crucial importance to river biota. Therefore, it is important to select the value of dispersion coefficients in simulation model very precisely which is also confirmed by the results of this study.

We expected more explicit result when we firstly intended to find out the sensitivity of water quality numerical model to used values of dispersion coefficients. As the outcomes of simulations indicate, it is necessary to continue with the analysis because there are still many factors entering the assessment and affecting the explicitness of effects.

ACKNOWLEDGMENT

This paper was prepared with the support of the Scientific Grant Agency APVV within the implementation of the project no. APVV-0274-10 „Kvantifikácia vplyvu vstupných údajov a parametrov modelového prostriedku na presnosť výstupov simulačných modelov disperzie v povrchových tokoch“. (Quantification of input data and model parameters influences on correctness of outputs of dispersion models for surface water), as well as by the 7th FP project “Detection of

Watercourse Contamination in Developing Countries using Sensor Networks”, Acronym: Goldfish., FP7-ICT-2009-6, Grant agreement no: 269985. It is also the result of the project implementation ITMS 26240120004 Centre of excellence for integrated flood protection of land supported by the Research & Development Operational Programme funded by the ERDF.

REFERENCES

- [1] Kosorin, K.: Dispersion coefficient for natural cross- sections of surface watercourses, *J. Hydrol. Hydromech.*, 43, 1995, 1-2, 93-101 (in Slovak).
- [2] Říha, J. et al.: Water quality and its modeling. NOEL, 2000, Brno, 269 p. (in Czech)
- [3] Sokáč, M.: The influence of discontinuous pollution sources on the receiving water body (in Slovak). Slovak University of Technology, Bratislava, Faculty of Civil Engineering, Edition of scientific works, sheet 87, ISBN 978-80-227-3328-1, 104 p.
- [4] Swamee, P.K.; Pathak, S.K.; Sohrab, M.: Empirical relations for longitudinal dispersion in streams. *Journal of Environmental Engineering* 126 (11), 2000, pp. 1056-1062.
- [5] Velísková, Y., Dulovičová, R., Sokáč, M. (2009a): Determination of the Size of the Longitudinal Dispersion of the Flow with a Free Water Surface in a Prismatic Channel. Part 1: Straight Stretch of Route. (in Slovak) *Acta Hydrologica Slovaca*, 1, 35–43. (in Slovak).
- [6] Velísková, Y., Dulovičová, R., Sokáč, M. (2009b): Determination of the Size of the Longitudinal Dispersion of the Flow with a Free Water Surface in a Prismatic Channel. Part 2: Straight Curved line. (in Slovak) *Acta Hydrologica Slovaca*, 2, 328–335. (in Slovak).
- [7] Velísková, Y., Sokáč, M., Dulovičová, R. (2009c): Determination of longitudinal dispersion coefficient in sewer networks. WMHE 2009. Vol. I. : Eleventh International Symposium on Water Management and Hydraulic Engineering. Ohrid, Macedonia, 1.-5.9.2009., Skopje : University Ss. Cyril and Methodius, 2009., ISBN 978-9989-2469-6-8., S. 493-498.
- [8] Velísková, Y., Sokáč, M., 2010: Sensitivity of Water Quality Numerical Model for Used Dispersion Coefficient Values. *Acta Hydrologica Slovaca*, 2, 210–218. (in Slovak)

The Tactical Model based on a Multi-Depot Vehicle Routing Problem

P. Stodola, and J. Mazal

Abstract—The Multi-Depot Vehicle Routing Problem is a famous problem formulated more than 50 years ago. Since that time, a lot of exact, heuristic and metaheuristic methods have been proposed in order to find a feasible solution for this NP-hard problem. The first part of this paper presents the original algorithm of the authors based on the Ant Colony Optimization theory. This part introduces pivotal principles of the algorithm, along with conducted experiments and acquired results on benchmark instances in comparison with rival state of the art methods. The primary part of the article deals with the tactical model based on our problem solution: optimal supply distribution. The model has become a part of our tactical information system which serves as a tool for commanders to support them in their decision making process. The model is introduced in terms of problem formulation, implementation, and application in practical situations in the domain of the military.

Keywords—Ant colony optimization, multi-depot vehicle routing problem, tactical modeling

I. INTRODUCTION

THE Multi-Depot Vehicle Routing Problem (MDVRP) is a well-known problem with many real applications in the areas of transportation, distribution and logistics [1]. In many businesses (e.g. parcel delivery, appliance repair), it is vital to find the optimal solution to this problem as it saves resources for a company, reduces its expenses, shortens time needed to distribute services, and thus makes the company more competitive.

The MDVRP problem consists in computing optimal routes for a fleet of vehicles to drop off goods or services at multiple destinations (customers); each customer should be served only once. The vehicles might start from multiple depots, each located in a different place. The important characteristic is the limited capacity of each vehicle which cannot be exceeded. After visiting the selected customers, each vehicle returns to its depot and might start a new journey to other (so far unvisited) customers with a new load.

MDVRP is an NP-hard problem as it is a generalization of the travelling salesman problem [2], therefore polynomial-time algorithms are unlikely to exist [3]. In this article, we present our original solution approach based on the Ant Colony Optimization (ACO) theory as a new approach to this topical issue. In fact, there have already been some attempts to use this

theory for this problem, but nevertheless, the results of these solutions are not of the quality as when using other contemporary methods (see Table 2). We managed to develop and design the fundamental details and parameters of this approach so that the results are comparable to other state of the art algorithms.

The primary part of this article comprises a tactical model based on our problem solution which has a practical application in the specific domain of the military. It is a model of optimal supply distribution on the battlefield.

This tactical model has been implemented into an actual tactical information system designed to support commanders in their decision making process [4]. A key goal of the model is to provide a tool to support commanders in their decision making as this system include both fundamental and advanced models of military tactics.

II. LITERATURE REVIEW

The solution methods for VRPs can be categorized as exact, heuristic, and metaheuristic. A broad overview of various methods is offered e.g. in [5]. For examples of exact methods, see e.g. [6], or [7], to name a few. Similar to the exact methods, many of heuristics have been developed, see e.g. [8], or [9].

Very popular metaheuristic methods have emerged in the last few years. These can be classified as state space search or evolutionary algorithms. For instance, simulated annealing [10] or genetic algorithms [11], [12] belong to the main evolutionary principles.

The remainder of this section focuses on the ACO methods. The potential of the ACO algorithm has been discovered very soon since it was published [13]. It was successfully applied for various problems [14], [15], [16].

Recently, there have been publications using the ACO theory for MDVRP problems [17], [18], [19]. The solution published in [17] is compared with our algorithm as it uses the standard Cordeau's test instances for evaluation.

III. ANT COLONY OPTIMIZATION ALGORITHM

ACO algorithm is a probabilistic technique for developing good solutions of computational problems. The principle is adopted from the natural world where ants explore their environment to find food; the idea is based on the behavior of ants seeking a path between their colony and a source of food.

P. Stodola is with University of Defence, Brno, Czech Republic (phone: +420 973 442 474; e-mail: petr.stodola@unob.cz).

J. Mazal is with University of Defence, Brno, Czech Republic (e-mail: jan.mazal@unob.cz).

A. Principle of the Algorithm

Fig. 1 presents the ACO algorithm we proposed for MDVRP. The solution found by the algorithm is improved in successive generations (iterations). In point 1, the termination condition is tested, points 2 to 15 cover an individual generation. Each depot employs a colony with the specific number of ants.

```

1. while not terminated
2.   for each ant in a colony
3.     set all nodes as unvisited
4.     while number of unvisited nodes > 0 do
5.       select a depot
6.       compute ant's probability of going to
         unvisited nodes
7.       select a node according to the
         probability
8.       if ant.load + node.load > ant.capacity
         then
9.         return to the depot
10.      else
11.        visit the selected node
12.      return to the depot
13.    save the best solution if found
14.    evaporate pheromone trails
15.    update pheromone trails
16. return the best solution
    
```

Fig. 1 ACO algorithm in pseudo code

In each generation, all ants in all colonies move between individual customers (referred to as nodes in Fig. 1). At first, the state of all nodes is set as unvisited in point 3. The algorithm continues until all nodes are visited (just once). In point 5, the depot (colony) is selected according to the given method; points 6 to 11 apply only to the ant from the selected colony. The ant's probability is computed in point 6; it determines the chance of the ant to go to every remaining unvisited nodes. In point 7, a node to be visited is chosen according to this probability.

Point 8 checks whether the ant can visit the selected node (i.e. whether its current load allows taking the load in the node and thus not exceeding ant's maximum capacity). If not, the ant returns to its colony (emptying its load) and the algorithm continues in point 5. If yes, the ant visits the selected node (delivering node's load and marking it as visited).

In point 12, after visiting all nodes, each ant returns to its colony. Then, if the best solution found in the generation is better than the best solution found in previous ones, it is saved (see point 13). Point 14 ensures evaporating the pheromone trails and in point 15, pheromone trails are updated according to the given method. Then, the next generation begins until the termination condition is met. The best solution found is returned at the end of the algorithm in point 16.

B. Parameters of the Algorithm

The ACO algorithm requires setting a number of parameters influencing the problem solution. Some parameters are adopted from related problems; others are new (see below). The list of all parameters is in Fig. 3.

A crucial parameter (proposed by authors) influences *how depots are selected* (see point 5 in Fig. 1). We propose five possibilities as follows:

- *Random selection*: depot (i.e. its vehicle) is selected randomly.
- *Selection of an idle depot*: depot with the shortest distance travelled so far is selected (i.e. vehicles take turns according to their distance they travelled at the moment of selection).
- *Selection of an idle depot (probability model)*: selection probabilities for all depots are computed based on the distance travelled so far (i.e. depots with shorter routes are more likely to be selected).
- *Selection of a depot with the greatest potential*: depot with the greatest potential is selected. The potential is computed as the sum of all pheromone trails which lead to unvisited customers (at the time of selection) – see formula (1).
- *Selection of a depot with the greatest potential (probability model)*: selection probabilities for all depots are computed based on the sum of pheromone trails to unvisited customers (i.e. depots with the bigger sum are more likely to be selected).

$$\varepsilon^k = \sum_{j \in S_u} \tau_{ij}^k \text{ for all } j \in S_u, \quad (1)$$

where ε^k is a potential for the colony (depot) k ,
 τ_{ij}^k is strength of a pheromone trail from the colony k between nodes i and j ,
 i is an index for the node with the current position of the ant from colony k ,
 S is a set of all nodes,
 S_u is a set of so far unvisited nodes ($S_u \subset S$).

In point 6 in Fig. 1, after a depot (colony) is chosen according to the methods mentioned above, the probabilities of choosing ant's path to the one of so far unvisited nodes are computed according to formula (2).

$$p_{ij}^k = \frac{\tau_{ij}^{\alpha} \cdot \eta_{ij}^{\beta} \cdot \mu_{ij}^{\gamma} \cdot \kappa_{ij}^{\delta}}{\sum_{l \in S_u} \tau_{il}^{\alpha} \cdot \eta_{il}^{\beta} \cdot \mu_{il}^{\gamma} \cdot \kappa_{il}^{\delta}} \text{ for all } j \in S_u, \quad (2)$$

where p_{ij}^k is a probability for an ant from the colony k in a node i to visit a node j ,
 τ_{ij}^k is strength of a pheromone trail from the colony k between nodes i and j ,
 η_{ij} is a multiplicative inverse of the distance between nodes i and j ,

μ_{ij}^k is a so-called savings measure [8],
 κ_{ij}^k is a measure for including the influence of ant's current load [20],
 $\alpha, \beta, \gamma, \delta$ are coefficients controlling the influence of $\tau_{ij}^k, \eta_{ij}, \mu_{ij}^k, \kappa_{ij}^k$ – see formula (2),
 S is a set of all nodes,
 S_u is a set of so far unvisited nodes ($S_u \subset S$).

Number of ants in colonies n_a (see point 2 in Figure 1) is a parameter determining the number of different solutions to be created and evaluated within a generation. Pheromone trails are then updated according to these solutions (based on the given method mentioned above).

Pheromone evaporation coefficient ρ determines the speed of evaporating pheromone trails at the end of each generation (point 14 in Figure 1) – see Formula (3).

$$\tau_{ij}^k = (1 - \rho) \cdot \tau_{ij}^k \text{ for all } i, j \in V, \quad (3)$$

where τ_{ij}^k is strength of a pheromone trail from the colony k between nodes i and j ,
 ρ is the pheromone evaporation coefficient.

C. Experiments and Results

As benchmark problems, we chose Cordeau's MDVRP instances taken from [21], namely p01, p02, p03, p04, p05, p06, p07, p08, p09, p10, p11, p12, p15, p18, and p21 (instances p13, p14, p16, p17, p19, and p20 were not included in the experiments as they incorporate the constraint on the maximum length of a single route, which the algorithm does not support).

Table 1 presents the results. We conducted 100 tests on each instance and registered the best solution found, the mean and standard deviation. The last column shows the difference between our results and the best solutions known so far which were received from [21]. The best known solutions were achieved by various algorithms during the history of benchmark instances.

Table 1 Results for MDVRP benchmark problems

Inst.	NoC	NoD	BKS	OBS	Mean	Stdev	Error
p01	50	4	576.87	576.87	583.15	6.50	0.00%
p02	50	4	473.53	475.86	482.86	3.44	0.49%
p03	75	5	641.19	644.46	650.04	4.12	0.51%
p04	100	2	1001.59	1018.49	1035.39	5.69	1.69%
p05	100	2	750.03	755.71	763.09	3.68	0.76%
p06	100	3	876.50	885.84	899.51	4.89	1.07%
p07	100	4	885.80	895.53	912.48	5.62	1.10%
p08	249	2	4420.95	4445.51	4572.23	66.75	0.56%
p09	249	3	3900.22	3990.19	4145.33	96.89	2.31%
p10	249	4	3663.02	3751.50	3864.92	50.21	2.42%

p11	249	5	3554.18	3657.16	3760.60	38.94	2.90%
p12	80	2	1318.95	1318.95	1320.48	1.90	0.00%
p15	160	4	2505.42	2510.11	2576.27	18.46	0.19%
p18	240	6	3702.85	3741.80	3812.25	37.22	1.05%
p21	360	9	5474.84	5631.12	5788.19	46.64	2.85%

NoC – number of customers, NoD – number of depots
 BKS – best known solution, OBS – our best solution

Table 2 compares results obtained via our algorithm with other results published. Algorithms called GA1 [12], GA2 [22], and GA3 [11] are based on genetic algorithm principles. GJ stands for Gillett and Johnson's algorithm [23]; CGW stands for Chao, Golden and Wasil's algorithm [24]. FIND (Fast improvement, INTensification, and Diversification) is a tabu search based algorithm [9], and finally ACO is another version of an algorithm based on the ACO theory [17]. Best solution values in Table 2 are indicated by bold numbers.

Table 2 Best solutions values obtained by various algorithms

Inst.	Our	GA1	GA2	GA3	GJ	CGW	FIND	ACO
p01	576.9	591.7	622.2	598.5	593.2	576.9	576.9	620.5
p02	475.9	483.1	480.0	478.7	486.2	474.6	473.5	-
p03	644.5	694.5	706.9	699.2	652.4	641.2	641.2	-
p04	1018.5	1062.4	1024.8	1011.4	1066.7	1012.0	1003.9	1585.9
p05	755.7	754.8	785.2	-	778.9	756.5	750.3	-
p06	885.8	976.0	908.9	882.5	912.2	879.1	876.5	-
p07	895.5	976.5	918.1	-	939.5	893.8	892.6	1257.9
p08	4445.5	4812.5	4690.2	-	4832.0	4511.6	4485.1	-
p09	3990.2	4284.6	4240.1	-	4219.7	3950.9	3937.8	9633.2
p10	3751.5	4291.5	3984.8	-	3822.0	3727.1	3669.4	-
p11	3657.2	4092.7	3880.7	-	3754.1	3670.2	3649.0	-
p12	1319.0	1421.9	1319.0	-	-	1327.3	1319.0	-
p15	2510.1	3059.2	2579.3	-	-	2610.3	2551.5	-
p18	3741.8	5462.9	3903.9	-	-	3877.4	3781.0	-
p21	5631.1	6872.1	5926.5	-	-	5791.5	5656.5	-

We can see that our algorithm managed to find better solutions in all cases when compared with the genetic principle based algorithms (GA1, GA2, GA3) and also in case of Gillett and Johnson's algorithm (GJ). The results are also better in 7 cases (and in 1 case the same) in comparison with the CGW algorithm and in 4 cases (and in 2 cases the same) in comparison with the algorithm FIND.

The last column of Table 2 shows the results for another version of the algorithm based on the ant colony optimization theory. As we can see, the results for this algorithm do not compare well with any other algorithm presented; in case of the

instance p09, the error is more than 140% compared to the best known solution.

IV. OPTIMAL SUPPLY DISTRIBUTION MODEL

The ACO algorithm has been integrated into our tactical information system designed to support command decision-making.

This subsystem seeks distribution patterns to provide supplies to friendly elements operating in the area of interest as efficiently as possible. Efficiency is based on the nature of the task at hand; the objective might be to minimize the sum of distances travelled by all vehicles, or minimize the time of the whole operation, or minimize the total fuel consumed by all vehicles.

The system provides a user friendly interface enabling to add, edit and delete nodes (depots and customers). Fig. 2 shows the main dialog for this model. As an example, 4 depots (labelled A to D) and 18 customers were included.

When all nodes are added (including their maximum load/capacity in kilograms), the MDVRP algorithm is executed. Values of algorithm's parameters are set and used to select the best options for the task at hand (see Fig. 3). Note the parameter

called number of cores used; this parameter represents the number of cores of a multi-core processor used for the execution since the ACO algorithm can be parallelized.

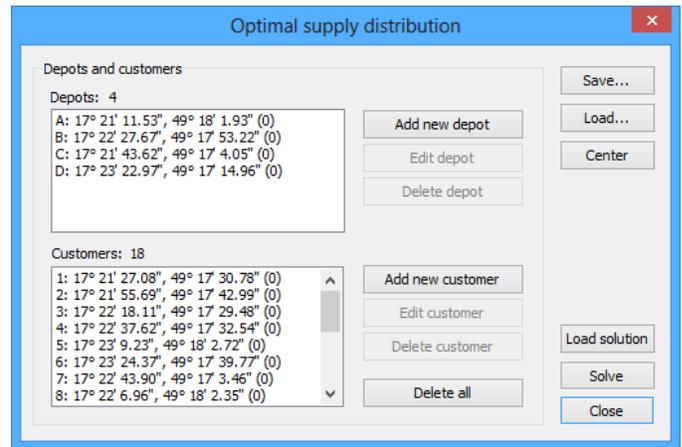


Fig. 2 Dialog for the optimal supply distribution model

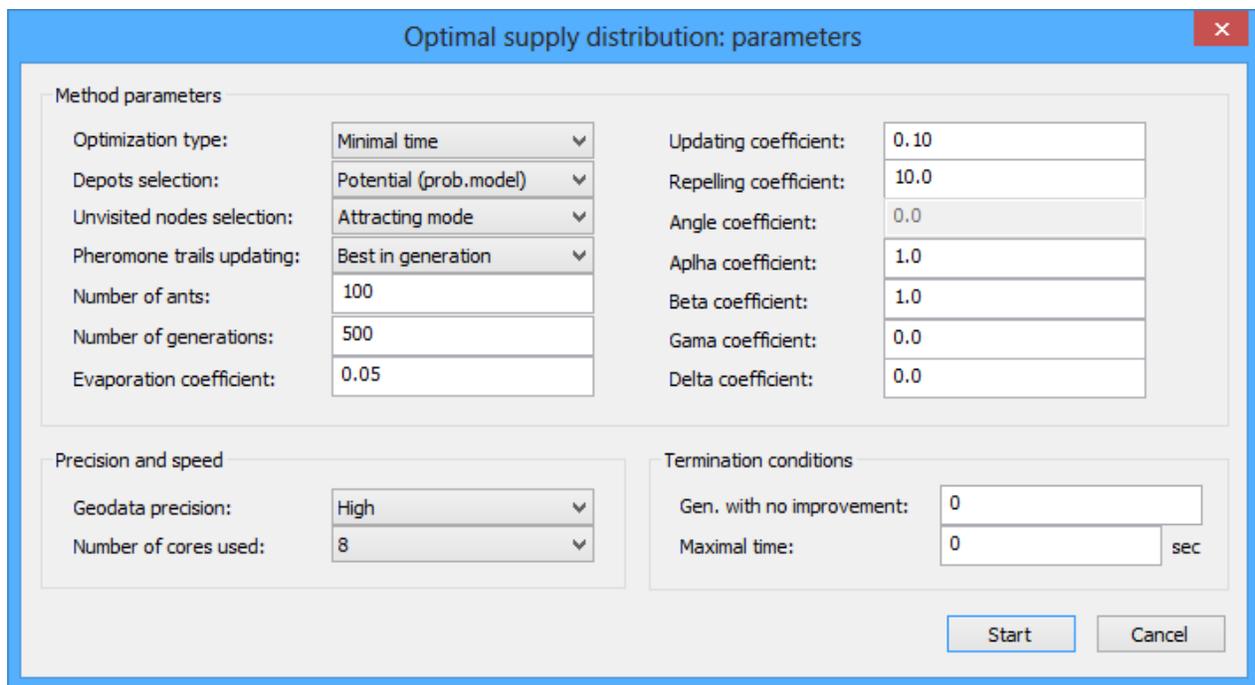


Fig. 3 Parameters for the ACO algorithm

Final routes for all vehicles are displayed both textually and graphically – see Fig. 4. Depots are shown as blue hexagons, customers as blue circles, and the red lines present the optimal

routes for individual vehicles. Although the example is rather simple, the same system can be used for tasks with many depots and hundreds of customers.

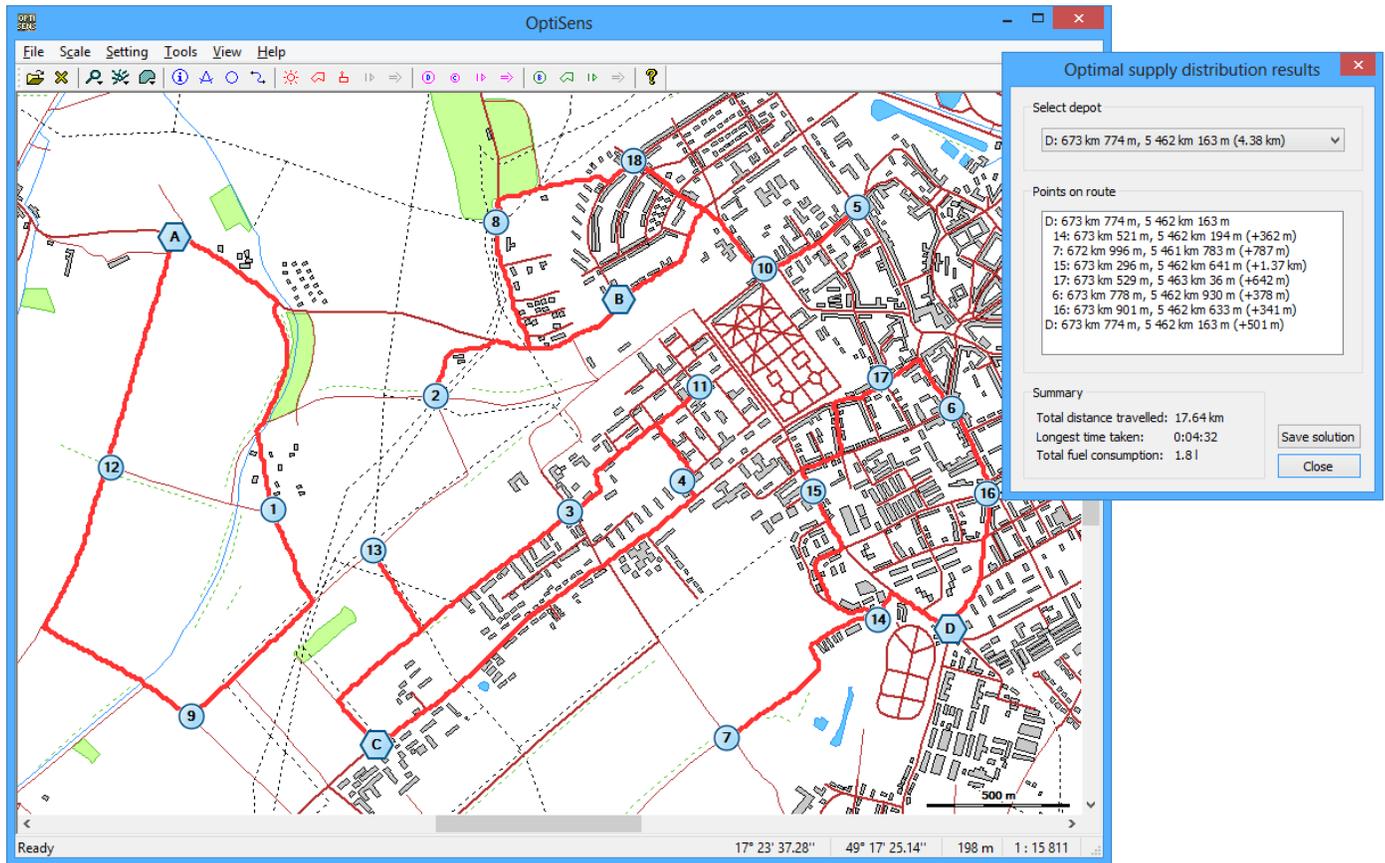


Fig. 4 Solution to the example situation

V. CONCLUSION

The paper presents the approach proposed by authors to the capacitated MDVRP problems based on the ant colony optimization theory. We have developed some new parameters and options not published yet (e.g. methods of selecting depots, method of updating pheromone trails according to the best solution found in a generation), thus contributing to the ACO theory. The new parameters we designed and verified participate on the very good results which the algorithm was able to achieve.

The strengths of the proposed algorithm are as follows:

- Fast convergence close to the optimal solution.
- High quality of solutions (comparable to the state of the art methods).
- Universal applicability (to metric, non-metric, and asymmetric problems).
- Possibility of distributed parallel processing.
- Application of the algorithm without any modification to solve classic VRP or capacitated VRP problems.

The proposed algorithm is of considerable significance in practical application in the domain of the military. It has been implemented into our tactical information system designed to support commanders' decision-making in order to provide the interface to solve the tactical task.

There are also a lot of ways of improving the current version of the algorithm and the system in the future.

Some future perspectives are as follows:

- Distribution of some computation to a GPU processor.
- Distribution of processing not only to the cores of a multi-core processor but also among more computers (to the GRID networks for instance).
- Development of other methods than empirical approach how to find the best parameter setting for various tasks.
- Extension of the algorithm for solving other problems (for instance MDVRP with Time Windows or with Pick-up and Delivering).

REFERENCES

- [1] G. B. Dantzig, and J. H. Ramser, "The Truck Dispatching Problem," in *Management Science*, vol 6, no 1, pp. 80-91, 1959.
- [2] C. Contardo, and R. Martinelli, "A new exact algorithm for the multi-depot vehicle routing problem under capacity and route length constraints," in *Discrete Optimization*, vol. 12, pp. 129-146, 2014.
- [3] M. R. Garey, and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York: W. H. Freeman & Co., 1990.
- [4] P. Stodola, and J. Mazal, "Planning Algorithm and its Modifications for Tactical Decision Support Systems," in *International Journal of Mathematics and Computers in Simulation*, vol. 6, no 1, pp. 99-106, 2012.

- [5] G. Laporte, "The Vehicle Routing Problem: An Overview of Exact and Approximate Algorithms," in *European Journal of Operational Research*, vol. 59, no 3, pp. 345-358, 1992.
- [6] G. Laporte, H. Mercure, and Y. Nobert, "An Exact Algorithm for the Asymmetrical Capacitated Vehicle Routing Problem," in *Networks*, vol. 16, no 1, pp. 33-46, 1986.
- [7] J. Gromicho, J. J. van Hoorn, A. L. Kok, and J. M. J. Schutten, "Restricted Dynamic Programming: A flexible framework for solving realistic VRPs," in *Computers and Operations Research*, vol. 39, no 5, pp. 902-909, 2012.
- [8] G. Clarke, and J. W. Wright, "Scheduling of Vehicles from a Central Depot to a Number of Delivery Points," in *Operations Research*, vol. 12, no 4, pp. 568-581, 1964.
- [9] J. Renaud, G. Laporte, and F. F. Boctor, "A Tabu Search Heuristic for the Multi-Depot Vehicle Routing Problem," in *Computers & Operations Research*, vol. 23, no 3, pp. 229-235, 1996.
- [10] Z. J. Czech, and P. Czarnas, "Parallel Simulated Annealing for the Vehicle Routing Problem with Time Windows," in *10th Euromicro Workshop on Parallel, Distributed and Network-based Processing*, Canary Islands, pp. 376-383, 2002.
- [11] P. Surekha, and S. Sumathi, "Solution To Multi-Depot Vehicle Routing Problem Using Genetic Algorithms," in *World Applied Programming*, vol. 1, no 3, pp. 118-131, 2011.
- [12] R. Thangiah, and S. Salhi, "Genetic Clustering: An Adaptive Heuristic for the Multidepot Vehicle Routing Problem," in *Applied Artificial Intelligence*, vol. 15, no 4, pp. 361-383, 2001.
- [13] M. Dorigo, "Optimization, Learning and Natural Algorithms," Ph.D. dissertation, Milan: Politecnico di Milano, 1992.
- [14] M. Dorigo, and L. C. Gambardella, "Ant Colonies for the Traveling Salesman Problem," in *Biosystems*, vol. 43, no 2, pp. 73-81, 1997.
- [15] A. R. M. de Silva, and G. L. Ramalho, "Ant System for the Set Covering Problem," in *IEEE International Conference on Systems, Man, and Cybernetics*, Tucson, vol. 5, pp. 3129-3133, 2001.
- [16] J. E. Bell, and P. R. McMullen, "Ant Colony Optimization Techniques for the Vehicle Routing Problem," in *Advanced Engineering Informatics*, vol. 18, no 1, pp. 41-48, 2004.
- [17] T. C. M. Caldeira, *Optimization of the Multi-Depot Vehicle Routing Problem: an Application to Logistics and Transport of Biomass for Electricity Production*. Lisbon: Technical University of Lisbon, 2009.
- [18] Jianhua Ma, and Jie Yuan, "Ant Colony Algorithm for Multiple-Depot Vehicle Routing Problem with Shortest Finish Time," in *Communications in Computer and Information Science*, vol. 113, pp. 114-123, 2010.
- [19] K. S. V. Narasimha, E. Kivelevitch, and M. Kumar, "Ant Colony Optimization Technique to Solve Min-Max MultiDepot Vehicle Routing Problem," in *American Control Conference (ACC)*, Montreal, pp. 3980-3985, 2012.
- [20] B. Bullnheimer, R. F. Hartl, and C. Strauss, "Applying the Ant System to the Vehicle Routing Problem," in *Meta-Heuristics: Advances and Trends in Local Search Paradigms for Optimization*, Springer US, pp. 285-296, 1999.
- [21] NEO Web, *Networking and Emerging Optimization*. Malaga: University of Malaga, 2015, <http://neo.lcc.uma.es/vrp/vrp-instances/multiple-depot-vrp-instances/> [Accessed on 30 January 2015].
- [22] B. Ombuki, F. and Hanshar, "An Effective Genetic Algorithm for the Multi-Depot Vehicle Routing Problem," Technical Report No. CS-04-10, St. Catharines, Canada: Brock University, 2004.
- [23] B. E. Gillett, and J. G. Johnson, "Multi-Terminal Vehicle-Dispatch Algorithm," in *Omega*, vol. 4, no 6, pp. 711-718, 1976.
- [24] M. I. Chao, B. L. Golden, and E. Wasil, "A New Heuristic for the Multi-Depot Vehicle Routing Problem that Improves upon Best-known Solutions," in *American Journal of Mathematical and Management Sciences*, vol. 13, no 3-4, pp. 371-406, 1993.

New Results on stability of hybrid stochastic systems

Manlika Rajchakit
 Department of Statistics
 Maejo University
 Chiang Mai, 50290 Thailand
 Email: manlika@mju.ac.th

Abstract—This paper is concerned with robust mean square stability of uncertain stochastic switched discrete time-delay systems. The system to be considered is subject to interval time-varying delays, which allows the delay to be a fast time-varying function and the lower bound is not restricted to zero. Based on the discrete Lyapunov functional, a switching rule for the robust mean square stability for the uncertain stochastic discrete time-delay system is designed via linear matrix inequalities. Finally, some examples are exploited to illustrate the effectiveness of the proposed schemes.

I. INTRODUCTION

Stochastic modelling has come to play an important role in many branches of science and industry. An area of particular interest has been the automatic control of stochastic systems, with consequent emphasis being placed on the analysis of stability in stochastic models. One of the most useful stochastic models which appear frequently in applications is the stochastic differential delay equations. In practice, we need estimate the parameters of systems. If the parameters are estimated using point estimations, the systems are described precisely and hence the study of the systems become relatively easier. On the other hand, if the parameters are estimated using confidence intervals, the systems become stochastic interval equations and the study of such systems are much more complicated.

Switched systems constitute an important class of hybrid systems. Such systems can be described by a family of continuous-time subsystems (or discrete-time subsystems) and a rule that orchestrates the switching between them. It is well known that a wide class of physical systems in power systems, chemical process control systems, navigation systems, auto-mobile speed change system, and so forth may be appropriately described by the switched model [1-4]. In the study of switched systems, most works have been centralized on the problem of stability. In the last two decades, there has been increasing interest in the stability analysis for such switched systems; see, for example, [5-7] and the references cited therein. Two important methods are used to construct the switching law for the stability analysis of the switched systems. One is the state-driven switching strategy [8-10]; the other is the time-driven switching strategy [11-13]. A switched system is a hybrid dynamical system consisting of a finite number of subsystems and a logical rule that manages

switching between these subsystems (see, e.g., [13-15] and the references therein).

The main approach for stability analysis relies on the use of Lyapunov-Krasovskii functional and linear matrix inequality (LMI) approach for constructing a common Lyapunov function [3-10]. Although many important results have been obtained for switched linear continuous-time systems, there are few results concerning the stability of switched linear discrete systems with time-varying delays. In [7-15], a class of switching signals has been identified for the considered switched discrete-time delay systems to be stable under the average dwell time scheme.

This paper studies robust mean square stability problem for uncertain stochastic switched linear discrete-time delay with interval time-varying delays. Specifically, our goal is to develop a constructive way to design switching rule to robustly mean square stable the uncertain stochastic linear discrete-time delay systems. By using improved Lyapunov-Krasovskii functional combined with LMIs technique, we propose new criteria for the robust mean square stability of the uncertain stochastic linear discrete-time delay system. Compared to the existing results, our result has its own advantages. First, the time delay is assumed to be a time-varying function belonging to a given interval, which means that the lower and upper bounds for the time-varying delay are available, the delay function is bounded but not restricted to zero. Second, the approach allows us to design the switching rule for robust mean square stability in terms of LMIs. Finally, some examples are exploited to illustrate the effectiveness of the proposed schemes.

The paper is organized as follows: Section II presents definitions and some well-known technical propositions needed for the proof of the main results. Switching rule for the robust mean square stability is presented in Section III. Numerical examples are provided to illustrate the theoretical results in Section IV, and the conclusions are drawn in Section V.

II. PRELIMINARIES

The following notations will be used throughout this paper. R^+ denotes the set of all real non-negative numbers; R^n denotes the n -dimensional space with the scalar product of two vectors $\langle x, y \rangle$ or $x^T y$; $R^{n \times r}$ denotes the space of all matrices of $(n \times r)$ - dimension. N^+ denotes the set of all

non-negative integers; A^T denotes the transpose of A ; a matrix A is symmetric if $A = A^T$.

Matrix A is semi-positive definite ($A \geq 0$) if $\langle Ax, x \rangle \geq 0$, for all $x \in R^n$; A is positive definite ($A > 0$) if $\langle Ax, x \rangle > 0$ for all $x \neq 0$; $A \geq B$ means $A - B \geq 0$. $\lambda(A)$ denotes the set of all eigenvalues of A ; $\lambda_{\min}(A) = \min\{Re\lambda : \lambda \in \lambda(A)\}$.

Consider a uncertain stochastic discrete systems with interval time-varying delay of the form

$$\begin{aligned} x(k+1) &= (A_\gamma + \Delta A_\gamma(k))x(k) + (B_\gamma + \Delta B_\gamma(k))x(k-d(k)) \\ &\quad + \sigma_\gamma(x(k), x(k-d(k)), k)\omega(k), \\ k \in N^+, \quad x(k) &= v_k, \quad k = -d_2, -d_2 + 1, \dots, 0, \end{aligned} \tag{1}$$

where $x(k) \in R^n$ is the state, $\gamma(\cdot) : R^n \rightarrow \mathcal{N} := \{1, 2, \dots, N\}$ is the switching rule, which is a function depending on the state at each time and will be designed. A switching function is a rule which determines a switching sequence for a given switching system. Moreover, $\gamma(x(k)) = i$ implies that the system realization is chosen as the i^{th} system, $i = 1, 2, \dots, N$. It is seen that the system (1) can be viewed as an autonomous switched system in which the effective subsystem changes when the state $x(k)$ hits predefined boundaries. $A_i, B_i, i = 1, 2, \dots, N$ are given constant matrices and the time-varying uncertain matrices $\Delta A_i(k)$ and $\Delta B_i(k)$ are defined by: $\Delta A_i(k) = E_{ia}F_{ia}(k)H_{ia}$, $\Delta B_i(k) = E_{ib}F_{ib}(k)H_{ib}$, where $E_{ia}, E_{ib}, H_{ia}, H_{ib}$ are known constant real matrices with appropriate dimensions. $F_{ia}(k), F_{ib}(k)$ are unknown uncertain matrices satisfying

$$F_{ia}^T(k)F_{ia}(k) \leq I, \quad F_{ib}^T(k)F_{ib}(k) \leq I, \quad k = 0, 1, 2, \dots, \tag{2}$$

where I is the identity matrix of appropriate dimension, $\omega(k)$ is a scalar Wiener process (Brownian Motion) on $(\Omega, \mathcal{F}, \mathcal{P})$ with

$$E[\omega(k)] = 0, \quad E[\omega^2(k)] = 1, \quad E[\omega(i)\omega(j)] = 0 (i \neq j), \tag{3}$$

and $\sigma_i : R^n \times R^n \times R \rightarrow R^n, i = 1, 2, \dots, N$ is the continuous function, and is assumed to satisfy that

$$\begin{aligned} \sigma_i^T(x(k), x(k-d(k)), k)\sigma_i(x(k), x(k-d(k)), k) &\leq \\ \rho_{i1}x^T(k)x(k) + \rho_{i2}x^T(k-d(k))x(k-d(k)), &\tag{4} \\ x(k), x(k-d(k)) \in R^n, & \end{aligned}$$

where $\rho_{i1} > 0$ and $\rho_{i2} > 0, i = 1, 2, \dots, N$ are known constant scalars. The time-varying function $d(k) : N^+ \rightarrow N^+$ satisfies the following condition:

$$0 < d_1 \leq d(k) \leq d_2, \quad \forall k \in N^+$$

Remark 2.1. It is worth noting that the time delay is a time-varying function belonging to a given interval, in which the lower bound of delay is not restricted to zero.

Definition 2.1. The uncertain stochastic switched system (1) is robustly stable if there exists a switching function $\gamma(\cdot)$ such that the zero solution of the uncertain stochastic switched

system is robustly stable.

Definition 2.2. The system of matrices $\{J_i\}, i = 1, 2, \dots, N$, is said to be strictly complete if for every $x \in R^n \setminus \{0\}$ there is $i \in \{1, 2, \dots, N\}$ such that $x^T J_i x < 0$.

It is easy to see that the system $\{J_i\}$ is strictly complete if and only if

$$\bigcup_{i=1}^N \alpha_i = R^n \setminus \{0\},$$

where

$$\alpha_i = \{x \in R^n : x^T J_i x < 0\}, i = 1, 2, \dots, N.$$

Definition 2.3. The discrete-time system (1) is robustly stable in the mean square if there exists a positive definite scalar function $V(k, x(k)) : R^n \times R^n \rightarrow R$ such that

$$E[\Delta V(k, x(k))] = E[V(k+1, x(k+1)) - V(k, x(k))] < 0, \text{ along any trajectory of solution of the system (1).}$$

Proposition 2.1. [16] *The system $\{J_i\}, i = 1, 2, \dots, N$, is strictly complete if there exist $\delta_i \geq 0, i = 1, 2, \dots, N, \sum_{i=1}^N \delta_i > 0$ such that*

$$\sum_{i=1}^N \delta_i J_i < 0.$$

If $N = 2$ then the above condition is also necessary for the strict completeness.

Proposition 2.2. (Cauchy inequality) *For any symmetric positive definite matrix $N \in M^{n \times n}$ and $a, b \in R^n$ we have*

$$\pm a^T b \leq a^T N a + b^T N^{-1} b.$$

Proposition 2.3. [31] *Let E, H and F be any constant matrices of appropriate dimensions and $F^T F \leq I$. For any $\epsilon > 0$, we have*

$$EFH + H^T F^T E^T \leq \epsilon E E^T + \epsilon^{-1} H^T H.$$

III. MAIN RESULTS

Let us set

$$W_i = \begin{bmatrix} W_{i11} & W_{i12} & W_{i13} \\ * & W_{i22} & W_{i23} \\ * & * & W_{i33} \end{bmatrix},$$

where

$$\begin{aligned} W_{i11} &= Q - P, \\ W_{i12} &= S_1 - S_1 A_i, \\ W_{i13} &= -S_1 B_i, \\ W_{i22} &= P + S_1 + S_1^T + H_{ia}^T H_{ia} + S_1 E_{ib} E_{ib}^T S_1^T, \\ W_{i23} &= -S_1 B_i, \\ W_{i33} &= -Q + 2H_{ib}^T H_{ib} + 2\rho_{i2} I, \end{aligned}$$

$$\begin{aligned}
 J_i &= (d_2 - d_1)Q - S_1 A_i - A_i^T S_1^T + 2S_1 E_{ia} E_{ia}^T S_1^T \\
 &\quad + S_1 E_{ib} E_{ib}^T S_1^T + H_{ia}^T H_{ia} + 2\rho_{i1} I, \\
 \alpha_i &= \{x \in R^n : x^T J_i x < 0\}, \quad i = 1, 2, \dots, N, \\
 \bar{\alpha}_1 &= \alpha_1, \quad \bar{\alpha}_i = \alpha_i \setminus \bigcup_{j=1}^{i-1} \bar{\alpha}_j, \quad i = 2, 3, \dots, N.
 \end{aligned} \tag{5}$$

The main result of this paper is summarized in the following theorem.

Theorem 1. *The uncertain stochastic switched system (1) is robustly stable in the mean square if there exist symmetric positive definite matrices $P > 0, Q > 0$ and matrix S_1 satisfying the following conditions*

(i) $\exists \delta_i \geq 0, i = 1, 2, \dots, N, \sum_{i=1}^N \delta_i > 0 : \sum_{i=1}^N \delta_i J_i < 0.$

(ii) $W_i < 0, \quad i = 1, 2, \dots, N.$

The switching rule is chosen as $\gamma(x(k)) = i$, whenever $x(k) \in \bar{\alpha}_i$.

Proof. Consider the following Lyapunov-Krasovskii functional for any i th system (1)

$$V(k) = V_1(k) + V_2(k) + V_3(k),$$

where

$$\begin{aligned}
 V_1(k) &= x^T(k) P x(k), \quad V_2(k) = \sum_{i=k-d(k)}^{k-1} x^T(i) Q x(i), \\
 V_3(k) &= \sum_{j=-d_1+1}^{-d_2+1} \sum_{l=k+j+1}^{k-1} x^T(l) Q x(l),
 \end{aligned}$$

We can verify that

$$\lambda_1 \|x(k)\|^2 \leq V(k). \tag{6}$$

Let us set $\xi(k) = [x(k) \ x(k+1) \ x(k-d(k)) \ \omega(k)]^T$, and

$$H = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & P & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad G = \begin{pmatrix} P & 0 & 0 & 0 \\ I & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{pmatrix}.$$

Then, the difference of $V_1(k)$ along the solution of the system (1) and taking the mathematical expectation, we obtained

$$\begin{aligned}
 E[\Delta V_1(k)] &= E[x^T(k+1) P x(k+1) - x^T(k) P x(k)] \\
 &= E[\xi^T(k) H \xi(k) - 2\xi^T(k) G^T \begin{pmatrix} 0.5x(k) \\ 0 \\ 0 \\ 0 \end{pmatrix}].
 \end{aligned} \tag{7}$$

because of

$$\xi^T(k) H \xi(k) = x(k+1) P x(k+1),$$

$$2\xi^T(k) G^T \begin{pmatrix} 0.5x(k) \\ 0 \\ 0 \\ 0 \end{pmatrix} = x^T(k) P x(k).$$

Using the expression of system (1)

$$\begin{aligned}
 0 &= -S_1 x(k+1) + S_1(A_i + E_{ia} F_{ia}(k) H_{ia}) x(k) \\
 &\quad + S_1(B_i + E_{ib} F_{ib}(k) H_{ib}) x(k-d(k)) + S_1 \sigma_i \omega(k), \\
 0 &= -\sigma_i^T x(k+1) + \sigma_i^T(A_i + E_{ia} F_{ia}(k) H_{ia}) x(k) \\
 &\quad + \sigma_i^T(B_i + E_{ib} F_{ib}(k) H_{ib}) x(k-d(k)) + \sigma_i^T \sigma_i \omega(k),
 \end{aligned}$$

we have

$$E[-2\xi^T(k) G^T \begin{pmatrix} 0.5x(k) \\ [-S_1 x(k+1) + S_1(A_i + E_{ia} F_{ia}(k) H_{ia}) x(k) \\ + S_1(B_i + E_{ib} F_{ib}(k) H_{ib}) x(k-d(k)) + S_1 \sigma_i \omega(k)] \\ 0 \\ [-\sigma_i^T x(k+1) + \sigma_i^T(A_i + E_{ia} F_{ia}(k) H_{ia}) x(k) \\ + \sigma_i^T(B_i + E_{ib} F_{ib}(k) H_{ib}) x(k-d(k)) + \sigma_i^T \sigma_i \omega(k)] \end{pmatrix}]$$

Therefore, from (7) it follows that

$$\begin{aligned}
 E[\Delta V_1(k)] &= E[x^T(k) [-P - S_1 A_i - S_1 E_{ia} F_{ia}(k) H_{ia} \\
 &\quad - A_i^T S_1^T - H_{ia}^T F_{ia}^T(k) E_{ia} S_1^T] x(k) \\
 &\quad + 2x^T(k) [S_1 - S_1 A_i - S_1 E_{ia} F_{ia}(k) H_{ia}] x(k+1) \\
 &\quad + 2x^T(k) [-S_1 B_i - S_1 E_{ib} F_{ib}(k) H_{ib}] x(k-d(k)) \\
 &\quad + 2x^T(k) [-S_1 \sigma_i - \sigma_i^T A_i - \sigma_i^T E_{ia} F_{ia}(k) H_{ia}] \omega(k) \\
 &\quad + x(k+1) [S_1 + S_1^T] x(k+1) \\
 &\quad + 2x(k+1) [-S_1 B_i - S_1 (E_{ib} F_{ib}(k) H_{ib})] x(k-d(k)) \\
 &\quad + 2x(k+1) [\sigma_i^T - S_1 \sigma_i] \omega(k) \\
 &\quad + x^T(k-d(k)) [-\sigma_i^T B_i - \sigma_i^T E_{ib} F_{ib}(k) H_{ib}] \omega(k) \\
 &\quad + \omega^T(k) [-2\sigma_i^T \sigma_i] \omega(k),
 \end{aligned}$$

By assumption (3), we have

$$\begin{aligned}
 E[\Delta V_1(k)] &= E[x^T(k) [-P - S_1 A_i - S_1 E_{ia} F_{ia}(k) H_{ia} \\
 &\quad - A_i^T S_1^T - H_{ia}^T F_{ia}^T(k) E_{ia} S_1^T] x(k) \\
 &\quad + 2x^T(k) [S_1 - S_1 A_i - S_1 E_{ia} F_{ia}(k) H_{ia}] x(k+1) \\
 &\quad + 2x^T(k) [-S_1 B_i - S_1 E_{ib} F_{ib}(k) H_{ib}] x(k-d(k)) \\
 &\quad + x(k+1) [S_1 + S_1^T] x(k+1) \\
 &\quad + 2x(k+1) [-S_1 B_i - S_1 E_{ib} F_{ib}(k) H_{ib}] x(k-d(k)) \\
 &\quad - 2\sigma_i^T \sigma_i],
 \end{aligned}$$

Applying Proposition 2.2, Proposition 2.3, condition (2) and assumption (4), the following estimations hold

$$\begin{aligned}
 -S_1 E_{ia} F_{ia}(k) H_{ia} - H_{ia}^T F_{ia}^T(k) E_{ia}^T S_1^T &\leq S_1 E_{ia} E_{ia}^T S_1^T + H_{ia}^T H_{ia}, \\
 -2x^T(k) S_1 E_{ia} F_{ia}(k) H_{ia} x(k+1) &\leq
 \end{aligned}$$

$$\begin{aligned}
 & x^T(k)S_1E_{ia}E_{ia}^TS_1^Tx(k) + x(k+1)^TH_{ia}^TH_{ia}x(k+1), \\
 & -2x^T(k)S_1E_{ib}F_{ib}(k)H_{ib}x(k-d(k)) \leq \\
 & x^T(k)S_1E_{ib}E_{ib}^TS_1^Tx(k) + x(k-d(k))^TH_{ib}^TH_{ib}x(k-d(k)), \\
 & -2x^T(k+1)S_1E_{ib}F_{ib}(k)H_{ib}x(k-d(k)) \leq \\
 & x^T(k+1)S_1E_{ib}E_{ib}^TS_1^Tx(k+1) + x(k-d(k))^TH_{ib}^TH_{ib}x(k-d(k)), \\
 & -\sigma_i^T(x(k), x(k-d(k)), k)\sigma_i(x(k), x(k-d(k)), k) \leq \\
 & \rho_{i1}x^T(k)x(k) + \rho_{i2}x^T(k-d(k))x(k-d(k)).
 \end{aligned}$$

Therefore, we have

$$\begin{aligned}
 E[\Delta V_1(k)] &= E[x^T(k)[-P - S_1A_i - A_i^TS_1^T \\
 & + 2S_1E_{ia}E_{ia}^TS_1^T \\
 & + S_1E_{ib}E_{ib}^TS_1^T + S_2E_{ia}E_{ia}^TS_2^T \\
 & + H_{ia}^TH_{ia} + 2\rho_{i1}I]x(k) \\
 & + 2x^T(k)[S_1 - S_1A_i]x(k+1) \\
 & + 2x^T(k)[-S_1B_i - S_2A_i]x(k-d(k)) \quad (8) \\
 & + x(k+1)[S_1 + S_1^T + H_{ia}^TH_{ia} \\
 & + S_1E_{ib}E_{ib}^TS_1^T]x(k+1) \\
 & + 2x(k+1)[S_2 - S_1B_i]x(k-d(k)) \\
 & + x^T(k-d(k))[2H_{ib}^TH_{ib} \\
 & + 2\rho_{i2}I]x(k-d(k))],
 \end{aligned}$$

The difference of $V_2(k)$ is given by

$$\begin{aligned}
 E[\Delta V_2(k)] &= E\left[\sum_{i=k+1-d(k+1)}^k x^T(i)Qx(i) \right. \\
 & - \sum_{i=k-d(k)}^{k-1} x^T(i)Qx(i)] \\
 &= E\left[\sum_{i=k+1-d(k+1)}^{k-d_1} x^T(i)Qx(i) \right. \\
 & + x^T(k)Qx(k) - x^T(k-d(k))Qx(k-d(k)) \\
 & + \sum_{i=k+1-d_1}^{k-1} x^T(i)Qx(i) \\
 & - \left. \sum_{i=k+1-d(k)}^{k-1} x^T(i)Qx(i)\right]. \quad (9)
 \end{aligned}$$

Since $d(k) \geq d_1$ we have

$$\sum_{i=k+1-d_1}^{k-1} x^T(i)Qx(i) - \sum_{i=k+1-d(k)}^{k-1} x^T(i)Qx(i) \leq 0,$$

and hence from (9) we have

$$\begin{aligned}
 E[\Delta V_2(k)] &\leq E\left[\sum_{i=k+1-d(k+1)}^{k-d_1} x^T(i)Qx(i) \right. \\
 & + \left. x^T(k)Qx(k) - x^T(k-d(k))Qx(k-d(k))\right]. \quad (10)
 \end{aligned}$$

The difference of $V_3(k)$ is given by

$$\begin{aligned}
 E[\Delta V_3(k)] &= E\left[\sum_{j=-d_2+2}^{-d_1+1} \sum_{l=k+j}^k x^T(l)Qx(l) \right. \\
 & - \sum_{j=-d_2+2}^{-d_1+1} \sum_{l=k+j+1}^{k-1} x^T(l)Qx(l)] \\
 &= E\left[\sum_{j=-d_2+2}^{-d_1+1} \left[\sum_{l=k+j}^{k-1} x^T(l)Qx(l) + x^T(k)Q(\xi)x(k) \right. \right. \\
 & - \left. \sum_{l=k+j}^{k-1} x^T(l)Qx(l) \right. \\
 & \left. - x^T(k+j-1)Qx(k+j-1)\right] \\
 &= E\left[\sum_{j=-d_2+2}^{-d_1+1} [x^T(k)Qx(k) \right. \\
 & - \left. x^T(k+j-1)Qx(k+j-1)]\right] \\
 &= E[(d_2 - d_1)x^T(k)Qx(k) \\
 & - \sum_{j=k+1-d_2}^{k-d_1} x^T(j)Qx(j)]. \quad (11)
 \end{aligned}$$

Since $d(k) \leq d_2$, and

$$\sum_{i=k+1-d(k+1)}^{k-d_1} x^T(i)Qx(i) - \sum_{i=k+1-d_2}^{k-d_1} x^T(i)Qx(i) \leq 0,$$

we obtain from (10) and (11) that

$$\begin{aligned}
 E[\Delta V_2(k) + \Delta V_3(k)] &\leq E[(d_2 - d_1 + 1)x^T(k)Qx(k) \\
 & - x^T(k-d(k))Qx(k-d(k))]. \quad (12)
 \end{aligned}$$

Therefore, combining the inequalities (8), (12) gives

$$E[\Delta V(k)] \leq E[x^T(k)J_i x(k) + \psi^T(k)W_i \psi(k)], \quad (13)$$

where

$$\psi(k) = [x(k) \ x(k+1) \ x(k-d(k))]^T,$$

$$W_i = \begin{bmatrix} W_{i11} & W_{i12} & W_{i13} \\ * & W_{i22} & W_{i23} \\ * & * & W_{i33} \end{bmatrix},$$

$$W_{i11} = Q - P,$$

$$W_{i12} = S_1 - S_1A_i,$$

$$W_{i13} = -S_1B_i,$$

$$W_{i22} = P + S_1 + S_1^T + H_{ia}^TH_{ia} + S_1E_{ib}E_{ib}^TS_1^T,$$

$$W_{i23} = -S_1B_i,$$

$$W_{i33} = -Q + 2H_{ib}^TH_{ib} + 2\rho_{i2}I,$$

$$J_i = (d_2 - d_1)Q - S_1A_i - A_i^TS_1^T + 2S_1E_{ia}E_{ia}^TS_1^T$$

$$+ S_1E_{ib}E_{ib}^TS_1^T + H_{ia}^TH_{ia} + 2\rho_{i1}I.$$

Therefore, we finally obtain from (13) and the condition (ii) that

$$E[\Delta V(k)] < E[x^T(k)J_i x(k)],$$

$$\forall i = 1, 2, \dots, N, \quad k = 0, 1, 2, \dots$$

We now apply the condition (i) and Proposition 2.1., the system J_i is strictly complete, and the sets α_i and $\bar{\alpha}_i$ by (5) are well defined such that

$$\bigcup_{i=1}^N \alpha_i = R^n \setminus \{0\},$$

$$\bigcup_{i=1}^N \bar{\alpha}_i = R^n \setminus \{0\}, \quad \bar{\alpha}_i \cap \bar{\alpha}_j = \emptyset, i \neq j.$$

Therefore, for any $x(k) \in R^n, k = 1, 2, \dots$, there exists $i \in \{1, 2, \dots, N\}$ such that $x(k) \in \bar{\alpha}_i$. By choosing switching rule as $\gamma(x(k)) = i$ whenever $x(k) \in \bar{\alpha}_i$, from the condition (13) we have

$$E[\Delta V(k)] \leq E[x^T(k)J_i x(k)] < 0, \quad k = 1, 2, \dots,$$

which, combining the condition (6), Definition 2.3 and the Lyapunov stability theorem [16], concludes the proof of the theorem in the mean square.

Remark 1. Note that the result proposed in [2–10, 13–15] for switching systems to be asymptotically stable under an arbitrary switching rule. The asymptotic stability for switching linear discrete time-delay systems studied in [4–9] was limited to constant delays. In [7–12], a class of switching signals has been identified for the considered switched discrete-time delay systems to be stable under the averaged well time scheme.

IV. NUMERICAL EXAMPLES

Example 1. (Stability) Consider the uncertain switched discrete-time system (1), where the delay function $d(k)$ is given by

$$d(k) = 1 + 4\sin^2 \frac{k\pi}{2}, \quad k = 0, 1, 2, \dots$$

and

$$(A_1, B_1) = \left(\begin{bmatrix} -0.1 & 0.01 \\ 0.02 & -0.2 \end{bmatrix}, \begin{bmatrix} -0.7 & 0.01 \\ 0.02 & 0.3 \end{bmatrix} \right),$$

$$(A_2, B_2) = \left(\begin{bmatrix} -0.2 & 0.02 \\ 0.03 & -0.3 \end{bmatrix}, \begin{bmatrix} -0.5 & 0.02 \\ 0.04 & 0.12 \end{bmatrix} \right),$$

$$(H_{1a}, H_{1b}) = \left(\begin{bmatrix} 0.1 & 0 \\ 0 & 0.2 \end{bmatrix}, \begin{bmatrix} 0.2 & 0 \\ 0 & 0.3 \end{bmatrix} \right),$$

$$(H_{2a}, H_{2b}) = \left(\begin{bmatrix} 0.4 & 0 \\ 0 & 0.5 \end{bmatrix}, \begin{bmatrix} 0.1 & 0 \\ 0 & 0.2 \end{bmatrix} \right),$$

$$(E_{1a}, E_{1b}) = \left(\begin{bmatrix} 5.3 & 0 \\ 0 & 3.4 \end{bmatrix}, \begin{bmatrix} 3.2 & 0 \\ 0 & 5.5 \end{bmatrix} \right),$$

$$(E_{2a}, E_{2b}) = \left(\begin{bmatrix} 3.5 & 0 \\ 0 & 3.3 \end{bmatrix}, \begin{bmatrix} 2.2 & 0 \\ 0 & 4.3 \end{bmatrix} \right),$$

$$(F_{1a}, F_{1b}) = \left(\begin{bmatrix} 0.1 & 0 \\ 0 & 0.2 \end{bmatrix}, \begin{bmatrix} 0.2 & 0 \\ 0 & 0.3 \end{bmatrix} \right),$$

$$(F_{2a}, F_{2b}) = \left(\begin{bmatrix} 0.2 & 0 \\ 0 & 0.5 \end{bmatrix}, \begin{bmatrix} 0.1 & 0 \\ 0 & 0.2 \end{bmatrix} \right).$$

By LMI toolbox of Matlab, we find that the conditions (i), (ii) of Theorem 1 are satisfied with $d_1 = 1, d_2 = 5, \delta_1 = 1, \delta_2 = 1$ and

$$P = \begin{bmatrix} 1.1329 & -0.0010 \\ -0.0010 & 1.7289 \end{bmatrix}, Q = \begin{bmatrix} 0.0506 & -0.0011 \\ -0.0011 & 0.4454 \end{bmatrix},$$

$$S_1 = \begin{bmatrix} -0.0169 & 0.0002 \\ 0 & -0.0798 \end{bmatrix}.$$

In this case, we have

$$(J_1(S_1, Q), J_2(S_1, Q)) =$$

$$\left(\begin{bmatrix} -0.2170 & -0.0026 \\ -0.0026 & -1.8633 \end{bmatrix}, \begin{bmatrix} -0.3591 & -0.0016 \\ -0.0016 & -2.0531 \end{bmatrix} \right).$$

Moreover, the sum

$$\delta_1 J_1(R, Q) + \delta_2 J_2(R, Q) = \begin{bmatrix} -0.5761 & -0.0042 \\ -0.0042 & -3.9164 \end{bmatrix}$$

is negative definite; i.e. the first entry in the first row and the first column $-0.5761 < 0$ is negative and the determinant of the matrix is positive. The sets α_1 and α_2 are given as

$$\alpha_1 = \{(x_1, x_2) : -0.2170x_1^2 - 0.0052x_1x_2 - 1.8633x_2^2 < 0\},$$

$$\alpha_2 = \{(x_1, x_2) : 0.3591x_1^2 + 0.0032x_1x_2 + 2.0531x_2^2 > 0\}.$$

Obviously, the union of these sets is equal to $R^2 \setminus \{0\}$. The switching regions are defined as

$$\bar{\alpha}_1 = \{(x_1, x_2) : -0.2170x_1^2 - 0.0052x_1x_2 - 1.8633x_2^2 < 0\},$$

$$\bar{\alpha}_2 = \alpha_2 \setminus \bar{\alpha}_1.$$

By Theorem 1 the uncertain system is robustly stable and the switching rule is chosen as $\gamma(x(k)) = i$ whenever $x(k) \in \bar{\alpha}_i$.

V. CONCLUSION

This paper has proposed a switching design for the robust stability of uncertain stochastic switched discrete time-delay systems with interval time-varying delays. Based on the discrete Lyapunov functional, a switching rule for the robust stability for the uncertain stochastic switched discrete time-delay system is designed via linear matrix inequalities. Numerical examples are provided to illustrate the theoretical results.

ACKNOWLEDGEMENT

This work was supported by the Office of Agricultural Research and Extension Maejo University Chiangmai Thailand, the Thailand Research Fund Grant, the Higher Education Commission and Faculty of Science, Maejo University, Thailand (TRG5780203).

REFERENCES

- [1] M. de la Sen, Global Stability of Polytopic Linear Time-Varying Dynamic Systems under Time-Varying Point Delays and Impulsive Controls, *Mathematical Problems in Engineering*, vol. 2010, Article ID 693958, 33 pages, 2010. doi:10.1155/2010/693958
- [2] K. Ratchagit, V.N. Phat, Stability and stabilization of switched linear discrete-time systems with interval time-varying delay, *Nonlinear Anal. Hybrid Syst.*, Vol. 5, 2011, 605–612. DOI: 10.1016/j.nahs.2011.05.006
- [3] VN. Phat, Y. Kongtham, and K. Ratchagit, LMI approach to exponential stability of linear systems with interval time-varying delays, *Linear Algebra Appl.*, Vol. 436, 2012, 243-251. doi: 10.1016/j.laa.2011.07.016
- [4] P. Niamsup, M. Rajchakit, G. Rajchakit, Guaranteed cost control for switched recurrent neural networks with interval time-varying delay, *JOURNAL OF INEQUALITIES AND APPLICATIONS*, 2013. DOI: 10.1186/1029-242X-2013-292
- [5] D. Liberzon, A.S. Morse, Basic problems in stability and design of switched systems, *IEEE Control Syst. Mag.*, **19**(1999), 57-70.
- [6] A.V. Savkin and R.J. Evans, *Hybrid Dynamical Systems: Controller and Sensor Switching Problems*, Springer, New York, 2001.
- [7] Z. Sun and S.S. Ge, *Switched Linear Systems: Control and Design*, Springer, London, 2005.
- [8] F. Gao, S. Zhong and X. Gao, Delay-dependent stability of a type of linear switching systems with discrete and distributed time delays, *Appl. Math. Computation*, **196**(2008), 24-39.
- [9] C.H. Lien, K.W. Yu, Y.J. Chung, Y.F. Lin, L.Y. Chung and J.D. Chen, Exponential stability analysis for uncertain switched neutral systems with interval-time-varying state delay, *Nonlinear Analysis: Hybrid systems*, **3**(2009),334–342.
- [10] G. Xie, L. Wang, Quadratic stability and stabilization of discrete-time switched systems with state delay, In: *Proc. of the IEEE Conference on Decision and Control, Atlantics*, December 2004, 3235-3240.
- [11] S. Boyd, L.E. Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.
- [12] D.H. Ji, J.H. Park, W.J. Yoo and S.C. Won, Robust memory state feedback model predictive control for discrete-time uncertain state delayed systems, *Appl. Math. Computation*, **215**(2009), 2035-2044.
- [13] G.S. Zhai, B. Hu, K. Yasuda, and A. Michel, Qualitative analysis of discrete- time switched systems. In: *Proc. of the American Control Conference*, 2002, 1880-1885.
- [14] W.A. Zhang, Li Yu, Stability analysis for discrete-time switched time-delay systems, *Automatica*, **45**(2009), 2265-2271.
- [15] F. Uhlig, A recurring theorem about pairs of quadratic forms and extensions, *Linear Algebra Appl.*, **25**(1979), 219-237.
- [16] R.P. Agarwal, *Difference Equations and Inequalities*, Second Edition, Marcel Dekker, New York, 2000.

Vision-based Navigation and System Identification of Unmanned Underwater Survey Vehicle

Seda Karadeniz Kartal, M. Kemal Leblebicioğlu

Abstract—In this study, a nonlinear mathematical model for an unmanned underwater survey vehicle is obtained. The Navigation problem is solved using the inertial navigation system and vision-based measurements. These are integrated to obtain vehicle's navigation data more accurately. In addition, the magnetic compass, depth sensor and Pitot tube are used in order to support vehicle's attitude, velocity and depth information. Performance of the resultant navigation system can be analyzed by creating suitable system state, measurement and noise models. The navigational data of the vehicle can be obtained improved using the extended Kalman filter. The mathematical model of the vehicle includes some unknown parameters such as added mass and drag parameters. They are obtained based on a system identification study of the vehicle using this estimated navigation data. All of this study is performed in Matlab/Simulink environment.

Keywords — inertial navigation, integration navigation system, mathematical modelling, unmanned underwater vehicle, vision-based navigation, system identification.

I. INTRODUCTION

UNMANNED underwater vehicles are used both in civilian and military areas frequently. They are most important tools to observe underwater. The unmanned underwater survey vehicle (SAGA) used in this study, is a remotely operated underwater survey vehicle specifically developed for the purpose of investigation of underwater and equipped with a camera and a two dimensional sonar (see Figure1) [1]. It is very easy to obtain navigational data and high resolution video, underwater observation, using this vehicle.

This work is supported from TÜBİTAK (Turkish Scientific and Technical Research Institute) 1001 project which's code is 111E267.

S. Karadeniz Kartal is currently working toward a Ph.D. degree at Middle East Technical University, Ankara, Turkey. Also, she is a research assistant in the Department of Electrical and Electronics Engineering of METU (e-mail: kseda@metu.edu.tr). Üniversiteler Mah. Dumlupınar Blv. No: 1, 06800 Çankaya Ankara/TURKEY Phone: +90 312 210 23 02 Fax: +90 312 210 23 04.

M. Kemal Leblebicioğlu a full professor in the Department of Electrical and Electronics Engineering of METU since 1999, Ankara, Turkey. (e-mail: kleb@metu.edu.tr). Üniversiteler Mah. Dumlupınar Blv. No: 1, 06800 Çankaya Ankara/TURKEY Phone: +90 312 210 23 02 Fax: +90 312 210 23 04.



Fig. 1 SAGA

The determination of the vehicle position is important in almost all the applications of these kind of vehicles. There are different navigation systems for determination of vehicle position. In this study, the inertial navigation system (INS) and vision-based data collection system are integrated to determine the position of the vehicle. INS measures the angular rate and acceleration. This measurement data is used to obtain vehicle position, velocity and attitude by integration. There are some errors involved with this procedure and they quickly increases with time in such a way that they are almost useless after a very short time if no correction is performed. On the other hand, the data collection rate is very fast (above 50 Hz). In the vision-based measurement system, the camera is located above in the pool of dimensions $5\text{ m} \times 5\text{ m}$. Its position is known (The distance of the camera to the pool surface is measured). While the vehicle moves in the pool, the camera records the video of the moving vehicle continuously. The position of the vehicle at a given instant is obtained using the known position of the camera and the distance between the vehicle and camera (the details are given in Section IV). The accuracy of this system is higher than the INS only case (since the vision based measurements acts like GPS data), but the data collection rate from the vision based measurement system is relatively small than the INS. These two navigation systems are integrated to obtain a navigation system which is more accurate. In addition, the depth data, the attitude data, and the velocity of the vehicle, are measured, respectively, with the depth sensor, the magnetic compass and Pitot tube.

In general mathematical models of unmanned underwater vehicles are nonlinear but linear models can be preferred

sometimes [2]. It is focused on a nonlinear mathematical model in this study (since it is more accurate). Usually, mathematical models of unmanned underwater vehicles (UUV) have some unknown parameters such as added mass and damping coefficients. The added mass parameters and damping coefficients can be obtained from some hydrodynamic software programs such as WAMIT, VERES and SEAWAY and, for example, SOLIDWORKS. If the structure of the vehicle has some symmetry with respect to pitch and yaw planes, added mass parameters can be obtained approximately using the strip theory [2]. For more accuracy, values of these parameters can be improved by a system identification study based on the navigational data from pool experiment. In this study, the initial approximate values of these parameters are used in the original mathematical model (especially, in the design of the autopilots). Then, these parameters are improved with a system identification study using the data of coming from the navigation system.

II. MATHEMATICAL MODEL

The mathematical model of an unmanned underwater vehicle is obtained as shown in the equations below [2].

$$M(\dot{v}) + C(v)v + D(v)v + g(\eta) = \tau = u \quad (1)$$

$$\dot{\eta} = J(\eta)v \quad (2)$$

M: The mass of the vehicle,
 C: Centrifugal and Coriolis forces matrix,
 D: Damping matrix,
 g: Gravitational and Buoyancy forces matrix,
 τ: Input vector,
 v: The linear and angular velocity vector of the vehicle,
 η: The position and attitude vector of the vehicle,
 J: Transformation matrix.

$$\vec{\eta} = [\eta_1^T, \eta_2^T]^T \quad \eta_1^T = [x, y, z]^T \quad \eta_2^T = [\theta, \psi]^T \quad (3)$$

$$\vec{v} = [v_1^T, v_2^T]^T \quad \vec{v}_1^T = [u, v, w]^T \quad \vec{v}_2^T = [p, q, r]^T \quad (4)$$

$$\vec{\tau} = u \quad \vec{\tau}_1^T = [X, Y, Z]^T \quad \vec{\tau}_2^T \equiv [K, M, N]^T \quad (5)$$

$$= [\tau_1^T, \tau_2^T]^T$$

u is the column matrix consists of moments and forces produced from thrusters. The vehicle has three thrusters. Two of them are located horizontally, at right and left sides. The last one is located in vertically. The motion in the x -axis (surge motion) and the rotation around the z -axis (yaw angle) are accomplished by horizontal thrusters. The motion in the z -axis (heave motion) and the rotation around the y -axis (pitch angle) are managed from the vertical thruster. In this study, the 3D motion is realized by a suitable combination of right, left and vertical thrusters.

III. INERTIAL NAVIGATION SYSTEM

The inertial measurement unit (IMU) is the main part of the inertial navigation system. The IMU is constructed by accelerometers and gyroscopes in three orthogonal axes.

Accelerometers measure the force vector, f_{ib}^b . Gyroscopes measure the angular rate vector, w_{ib}^b . The position, velocity and attitude information of the vehicle are obtained from these measured angular rate and force (acceleration) data by integration. The position of the vehicle is obtained by twice integration of measured force. And the attitude of the vehicle is obtained by integration of measured angular rate. These data have error since integration error [3].

A. Kinematic

In navigation studies, the linear and angular motions are measured in different coordinate systems. The kinematic units such as position, velocity, acceleration and angular rates are usually expressed in three different frames [3].

- Object frame, α ;
- Reference frame, β ;
- Resolving frame, γ .

The notation $x_{\beta\alpha}^\gamma$ is means that the vector x is determined in the frame α with respect to frame β , and the result is expressed in the resolving frame γ .

B. IMU Modelling and Inertial Navigation Process

The output information of the accelerometer and gyroscope are modelled by the following equations [3].

$$\tilde{f}_{ib}^b = b_a + (I_3 + M_a)f_{ib}^b + w_a \quad (6)$$

$$\tilde{w}_{ib}^b = b_g + (I_3 + M_g)w_{ib}^b + G_g f_{ib}^b + w_g \quad (7)$$

\tilde{f}_{ib}^b : The output data of accelerometer; force (acceleration) information,

\tilde{w}_{ib}^b : The output data of gyroscope; angular rate information,

f_{ib}^b : The actual force (acceleration),

w_{ib}^b : The actual angular rate,

b_a and b_g : The bias errors of accelerometer and gyroscope,

M_a and M_g : The scale-factor and cross-coupling errors of accelerometer and gyroscope,

w_a and w_g : The random noise in accelerometer and gyroscope measurements,

G_g : The further error source of gyroscope, gyro dependent bias,

I_3 : 3×3 Unit matrix.

The inertial navigation process integrates the output of the IMU to produce position, velocity and attitude information. The navigation equations are constructed in four stages. These stages are attitude update, specific-force frame transformation, velocity update and position update. The reference frame is chosen as the Earth-Centered Earth-Fixed frame (ECEF).

C. Attitude Update

The attitude update of Earth-frame navigation equations is obtained as shown below, using the angular rate, w_{ib}^b information [3].

$$C_b^e(+)\approx C_b^e(-)(I_3 + \Omega_{ib}^b \tau_i) - \Omega_{ie}^e C_b^e(-)\tau_i \quad (8)$$

The matrix C_b^e , is the attitude matrix of the vehicle and Ω_{ib}^b is the measured angular rate of the IMU. It is the angular rate of the body coordinate frame related to the inertial coordinate frame and it is a skew-symmetric matrix. The matrix Ω_{ie}^e is the earth angular rate with respect to the inertial coordinate frame and it is a skew-symmetric matrix. The constant τ_i is the integration time range.

D. Specific-Force Frame Transformation

The vehicle force information of the body-fixed coordinate frame related to inertial coordinate frame f_{ib}^b is updated and resolved in the earth coordinate frame. This force is updated using the updating attitude vector, C_b^e [3].

$$f_{ib}^e(t) \cong \frac{1}{2}(C_b^e(+)+C_b^e(-))f_{ib}^b \quad (9)$$

E. Velocity Update

The vehicle's velocity related to the earth frame is updated as follows, using the earth centered force and gravity [3].

$$v_{eb}^e(+)\approx v_{eb}^e(-)+(f_{ib}^e+g_b^e(r_{eb}^e(-))-2\Omega_{ie}^e v_{eb}^e(-))\tau_i \quad (10)$$

F. Position Update

The vehicle's position related to the earth frame is updated as follow, using the updating velocity data [3].

$$r_{eb}^e(+)=r_{eb}^e(-)+(v_{eb}^e(-)+v_{eb}^e(+))\frac{\tau_i}{2} \quad (11)$$

IV. VISION-BASED NAVIGATION SYSTEMS

The vision-based navigation system as seen in figure 2 consists of the camera which is located above the pool (5m × 5m). This camera is positioned at a known height from the pool. The angle (a), or the distance d , can be obtained using the captured vision. In particular, the distance d is proportional to the number of pixels between the center of the image and the vehicle's center in the captured image. The depth data is taken from depth sensor in the vehicle. While the vehicle moves in the pool, the camera whose position is known records visions of the vehicle. The vertical distance between camera and vehicle, h is obtained from known vehicle's depth and height of the camera.

V. INTEGRATION NAVIGATION SYSTEM

The vision-based measurement system and INS are integrated to produce the navigation solution which is more accurate. The position and attitude data of the vehicle originally comes from IMU. Also, the attitude data of the vehicle is obtained from the magnetic compass mounted on the vehicle. The depth data of the vehicle is obtained from the depth sensor mounted on the vehicle. The velocity data of the vehicle is taken from Pitot tube. The vision-based measurement system acts as a kind of GPS data to improve the position information produced by IMU [4]. The Loosely

Coupled integration architecture is chosen as the basic navigation fusion algorithm. In this study, the data collection rate of the camera is 1Hz, INS is 50 Hz, magnetic compass is 10 Hz and the other sensors are 1 Hz. The synchronization of the different data collection rates is achieved by holding the lower rates at their previous values until new data comes in. In this way, the system works as if all the data is collected at the highest data collection rate.

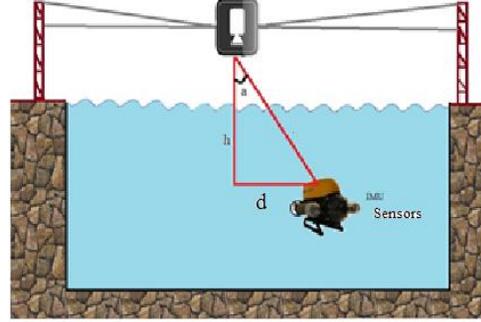


Fig. 2 the schematic illstrution of integrated navigation system

VI. STATE SELECTION and SYSTEM MODEL

In this study, the error states are estimated. The extended Kalman filter is used as the estimation algorithm [5].

Error states are chosen as the attitude, velocity and position of the vehicle related to earth frame and the bias errors of accelerometer and gyroscope. These are shown as follows.

$$x_{eb}^e = [\partial\varphi_{eb}^e \quad \partial v_{eb}^e \quad \partial r_{eb}^e \quad b_a \quad b_g]^T \quad (12)$$

$\partial\varphi_{eb}^e$: The attitude error vector of the vehicle,

∂v_{eb}^e : The velocity error vector of the vehicle,

∂r_{eb}^e : The position error vector of the vehicle.

The attitude variation of the vehicle with respect to time is

$$\partial\dot{\varphi}_{eb}^e = \hat{C}_b^e b_g - \Omega_{ie}^e \partial\varphi_{eb}^e. \quad (13)$$

The gravity error changes with respect to the position error. In this application, the gravity variation related to position variation is small and neglected. So, the variation of the velocity error is shown below.

$$\partial\dot{v}_{eb}^e = \hat{C}_b^e b_a - (\hat{C}_b^e \hat{f}_{ib}^b) \wedge \partial\varphi_{eb}^e - 2\Omega_{ie}^e \partial v_{eb}^e \quad (14)$$

The variation of the position error is then as follows.

$$\partial\dot{r}_{eb}^e = \partial v_{eb}^e \quad (15)$$

The state space representation of the system is:

$$\dot{x} = Fx + Qw \quad (16)$$

$$\begin{bmatrix} \delta\varphi_{eb}^e \\ \delta v_{eb}^e \\ \delta r_{eb}^e \\ \dot{b}_a \\ \dot{b}_g \end{bmatrix} = \begin{bmatrix} -\Omega_{ie}^e & 0_3 & 0_3 & 0_3 & \hat{C}_b^e \\ -\hat{C}_b^e \hat{f}_{ib}^b \wedge & -2\Omega_{ie}^e & 0_3 & \hat{C}_b^e & 0_3 \\ 0_3 & I_3 & 0_3 & 0_3 & 0_3 \\ 0_3 & 0_3 & 0_3 & 0_3 & 0_3 \\ 0_3 & 0_3 & 0_3 & 0_3 & 0_3 \end{bmatrix} \begin{bmatrix} \delta\varphi_{eb}^e \\ \delta v_{eb}^e \\ \delta r_{eb}^e \\ b_a \\ b_g \end{bmatrix} \quad (17)$$

$$\Omega_{ie}^e = \begin{bmatrix} 0 & -w_{ie}^e & 0 \\ w_{ie}^e & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (18)$$

The covariance matrix of the system noise is given by [3].

$$Q = \begin{pmatrix} n_{rg}^2 I_3 & 0_3 & 0_3 & 0_3 & 0_3 \\ 0_3 & n_{ra}^2 I_3 & 0_3 & 0_3 & 0_3 \\ 0_3 & 0_3 & 0_3 & 0_3 & 0_3 \\ 0_3 & 0_3 & 0_3 & n_{bad}^2 I_3 & 0_3 \\ 0_3 & 0_3 & 0_3 & 0_3 & n_{bgd}^2 I_3 \end{pmatrix} \tau_s \quad (19)$$

The values of n_{rg}^2 , n_{ra}^2 , n_{bad}^2 ve n_{bgd}^2 are the power density of random noise of gyroscope, accelerometer, the power density of bias error of the accelerometer and the gyroscope, respectively. The value of σ_{ra} is the standard deviation of the measurement force noise in the accelerometer and the value of σ_{rg} is the standard deviation of the measurement angular rate noise in the gyroscope. Also, τ_i is the time range of the integration. The values of σ_{bad} , σ_{bgd} are bias deviation of the accelerometer and gyroscope and the time related to these deviations are τ_{bad} and τ_{bgd} .

$$\begin{aligned} n_{ra}^2 &= \sigma_{ra}^2 \tau_i \\ n_{rg}^2 &= \sigma_{rg}^2 \tau_i \\ n_{bad}^2 &= \sigma_{bad}^2 \tau_{bad} \\ n_{bgd}^2 &= \sigma_{bgd}^2 \tau_{bgd} \end{aligned} \quad (20)$$

The data from the IMU comes in discrete time. Hence, the equivalent representation of the Equation 16 in discrete time is used.

$$x_{k+1} = \Phi_k x_k + w_k \quad (21)$$

The state transition matrix, Φ_k is derived from the system dynamics matrix F and time range $\Delta t = t_{k+1} - t_k$ [6]. This

matrix is assumed to be constant during the sampling interval. The second and higher order terms in the expansion below are neglected.

$$\Phi_k = e^{F\Delta t} = I + F\Delta t + \frac{(F\Delta t)^2}{2!} + \dots \quad (22)$$

VII. MEASUREMENT MODEL

The measurement model is composed of the attitude, velocity and position errors. The attitude data comes from IMU and magnetic compass. The velocity data comes from IMU and Pitot tube. Also, the position data is comes from IMU and the vision-based measurement system. In addition, the depth information is provided from depth sensor [7]. The IMU data rate is taken as the reference data rate (i.e., the highest data collection rate among measurements). The error model of IMU is considered to be the static error model (i.e., a time invariant representation is assumed) [8]. The measurement noise is assumed as the zero mean Gaussian noise. The measurement matrix for the extended Kalman filter is given by [3]:

$$z = m - m_{ref} \quad (23)$$

The variation of the measurement vector related to the earth frame is obtained as follow.

$$\partial z_k = \begin{bmatrix} \partial\varphi_{eb}^e \\ \delta v_{eb}^e \\ \delta x_{eb}^e \end{bmatrix} = \begin{bmatrix} \varphi_{m.c} - \varphi_{imu} \\ v_{ptt} - v_{imu} \\ x_{imu} - x_{kamera} - x_{depth} \end{bmatrix} \quad (24)$$

H_k , measurement matrix is obtained as:

$$H_k = \begin{bmatrix} I_3 & 0_3 & 0_3 & 0_3 & 0_3 \\ 0_3 & I_3 & 0_3 & 0_3 & 0_3 \\ 0_3 & 0_3 & I_3 & 0_3 & 0_3 \end{bmatrix} \quad (25)$$

The measurement covariance matrix is chosen as:

$$R = \begin{bmatrix} R_{11} & 0_3 & 0_3 \\ 0_3 & R_{22} & 0_3 \\ 0_3 & 0_3 & R_{33} \end{bmatrix} \quad (26)$$

VIII. SIMULATION RESULTS

The error states are estimated using the extended Kalman filter algorithm for the 3D motion. Results have been obtained in a Matlab/Simulink environment.

The actual velocity, attitude and position data with respect to time is shown in figures 3, 4 and 5, respectively. The vehicle moves in 3D with surge (u) and heave (w) speeds and yaw (φ) and pitch (θ) angles. The estimated position error shown in figure 6 is obtained from the integrated navigation system. As seen in figure 6, estimated depth error is smaller

than the other position errors (x and y), since an additional sensor (i.e., depth sensor) is used to support the measurements in the z direction. Figures 7 and 8, show the estimated attitude error and estimated velocity error, for the integrated navigation system, respectively. As observed from the simulation results, the estimated error states are small in the integrated navigation system. Thus, the more accurate position of the vehicle is obtained by this navigation system.

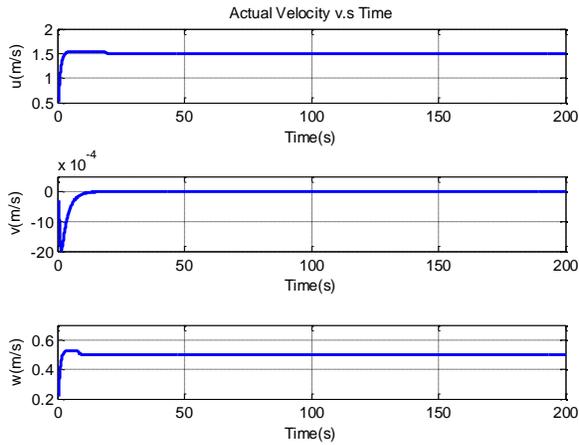


Fig. 3 the actual velocity of the vehicle

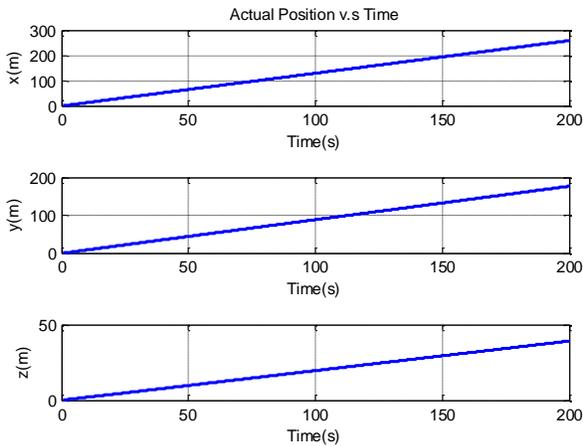


Fig. 4 the actual position of the vehicle

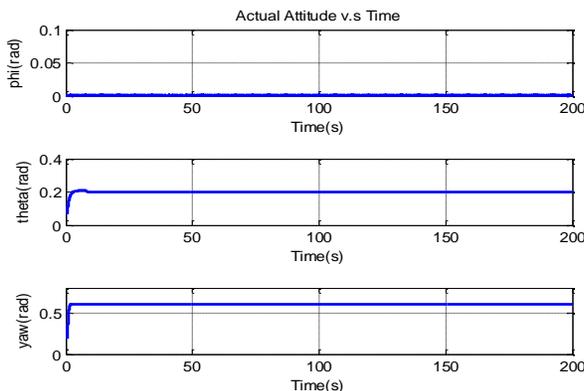


Fig. 5 the actual attitude of the vehicle

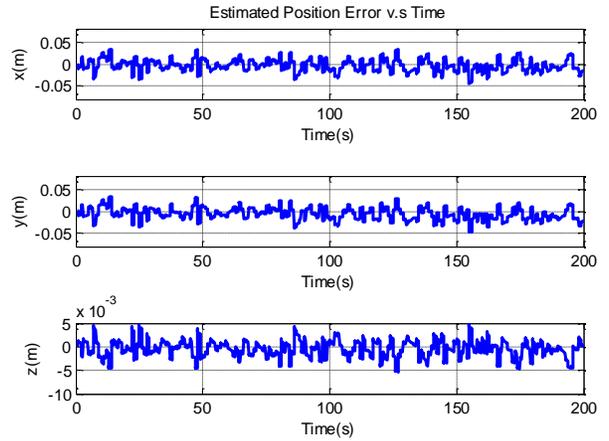


Fig. 6 the estimated position errors

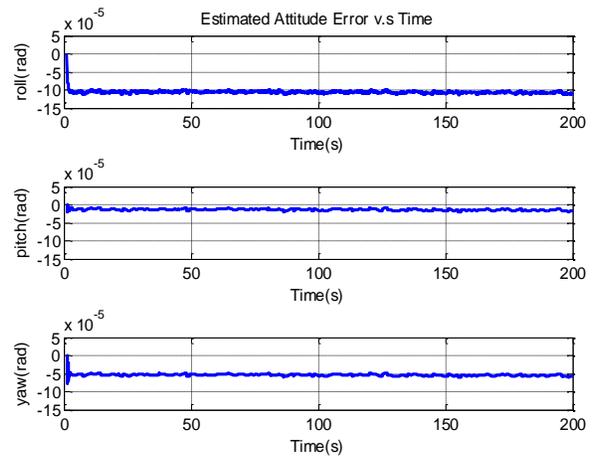


Fig. 7 the estimated attitude errors

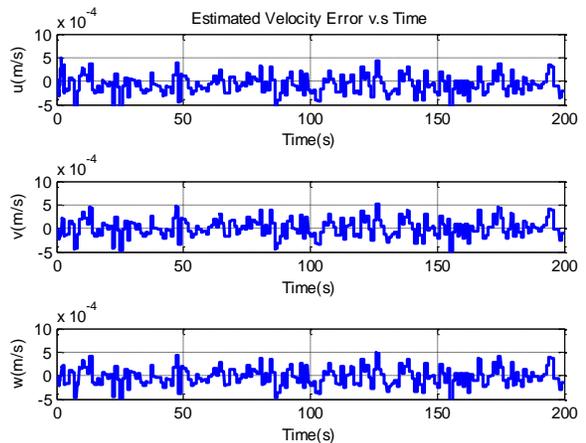


Fig. 8 the estimated velocity errors

IX. SYSTEM IDENTIFICATION

The mathematical model of SAGA used in this study involves some unknown parameters such as added mass and damping coefficients. These parameters are originally obtained approximately. Their more accurate values are to be obtained by a system identification study. An uncoupled motion model is preferred in order to concentrate on certain parameters more accurately in the measurements. These motions and parameters are shown in the following table.

Motion	Parameter
Surge	$X_{\dot{u}}$ and $X_{\ddot{u}}$
Sway	$Y_{\dot{v}}$ and $Y_{\ddot{v}}$
Heave	$Z_{\dot{w}}$ and $Z_{\ddot{w}}$
Roll	$K_{\dot{p}}$ and $K_{\ddot{p}}$
Pitch	$M_{\dot{q}}$ and $M_{\ddot{q}}$
Yaw	$N_{\dot{r}}$ and $N_{\ddot{r}}$

Table.1 the decoupled motion of the vehicle and related parameters

Since the vehicle cannot perform sway and roll motion, the parameters of $Y_v, Y_{\dot{v}}, K_p$ ve $K_{\dot{p}}$ cannot be identified accurately. Other parameters are obtained by solving an optimization problem whose cost function, (J) is determined as

$$\text{minimize } J = \sum_{i=1}^n [d_{\text{measured}} - d_{\text{model}}]^2 \quad (27)$$

$d_{\text{model}} \approx$ added mass and damping parameters

This optimization problem is solved using “fminsearch” algorithm of Matlab optimization toolbox. d_{model} in this optimization problem is generated from the simulation model by using the (recorded) inputs obtained from the physical system in experimentation [4].

A. Surge Motion Test

The right and left thrusters are activated with equal force. The vehicle moves in the x-direction with surge speed (u). This surge motion equation is as follows.

$$\sum X = (m - X_{\dot{u}})\dot{u} - X_u u \quad (28)$$

where,

$\sum X$: The total force with direction x direction,
 m : The mass of the vehicle,
 u : The surge speed of the vehicle.

The parameters related to this surge motion, $X_{\dot{u}}$ ve X_u as shown in table 2, are obtained from the solution of the associated optimization problem.

B. Yaw Motion Test

The right and left thrusters are activated with different forces. The vehicle rotates about the z-axis with angular rate (r) and yaw angle (φ) in the xy plane (i.e., yaw plane). This motion equation is as follows.

$$\sum N = (I_z - N_{\dot{r}})\dot{r} - N_r r \quad (29)$$

where,

$\sum N$: The total moment in the yaw plane,
 I_z : The inertia tensor of the vehicle in the z-axes,
 r : The angular rate component in the z-axes.

The parameters related to yaw plane motion, $N_{\dot{r}}$ ve N_r as shown in table 2, are obtained from the solution of the associated optimization problem.

C. Pitch Motion Test

The vertical thruster is activated. The vehicle moves in the xz plane (i.e., pitch plane) with heave speed (w) and pitch angle (θ). The pitch motion equation is as follows.

$$\begin{bmatrix} \sum Z \\ \sum M \end{bmatrix} = \begin{bmatrix} (m - Z_{\dot{w}})\dot{w} - Z_w w \\ (I_y - M_{\dot{q}})\dot{q} - M_q q + (z_G - z_B)W \sin\theta \end{bmatrix} \quad (30)$$

where,

$\sum Z$: The total force in the xz plane,
 $\sum N$: The total moment in the xz plane,
 z_G, z_B : The gravity and buoyancy forces component along the z axis,
 I_y : The inertia tensor of the vehicle along the y axis,
 w, q : The heave speed in the z direction and the angular rate component in the y -axis.

The parameters are related to pitch motion $Z_w, M_q, Z_{\dot{w}}$ ve $M_{\dot{q}}$ as shown in table 2, are obtained from the solution of the associated optimization problem.

Motion	Damping Parameters	Added Mass Parameters
Surge	$X_{\dot{u}} = -11.7439$	$X_u = -1.6300$
Yaw	$N_{\dot{r}} = -16.846$	$N_r = -0.0144$
Pitch	$Z_{\dot{w}} = -18.4213$	$Z_w = -1.1366$
Pitch	$M_{\dot{q}} = -0.6004$	$M_q = -0.1547$

Table. 2 added mass and damping parameters as the result of optimization

X. CONCLUSION

The nonlinear mathematical model of SAGA is obtained from a combination of a system identification and a navigation study. The navigation problem is solved by integrating the IMU data with all the remaining sensor information, and in particular, the vision based measurement system. As usual, all the measurement data is noisy. The fusion is performed using an Extended Kalman Filter algorithm. It is seen that the estimated state error is much smaller in the integrated navigation system. The system identification of SAGA is achieved by using the estimated data from the integrated navigation system. The unknown parameters, added mass and damping coefficients are obtained more accurately as the result of the system identification study.

In the future, the real physical world pool experiment will be performed for SAGA. During the system identification study, classical optimization algorithms in MATLAB (fminsearch algorithm) as well as evolutionary optimization techniques such as differential evolution, particle swarm optimization and genetic algorithms will be used and compared.

THANKS

This study is supported from TÜBİTAK (Turkish Scientific and Technical Research Institute) 1001 project which's code is 111E267. We offer our thanks by the way.

REFERENCES

- [1] <http://www.desistek.com.tr>.
- [2] T.I. Fossen, *Guidance and Control of Ocean Vehicles*, John Wiley & Sons Inc., 1994.
- [3] Paul D. Groves, *Principles of GNSS, Inertial, and Multi-Sensor Integrated Navigation Systems*, Boston:Artech House, (2008).
- [4] Hsin-Hung Chen, "Vision Based Tracking with Projective Mapping for Parameter Identification of Remotely Operated Vehicles", *OCEANS*, Page; 983 - 994, 2008.
- [5] John L. Crassidis and John L. Junkins, *Optimal Estimation of Dynamic Systems*, 2004.
- [6] R.G. Brown and Y.C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering*, John Wiley and Sons Inc., 1997.
- [7] Andreas Løberg Carlsen, *Navigational Assistance for Mini-ROV*, M.Sc.Thesis in Norwegian University of Science and Technology, July 2010.
- [8] Naser El-Sheimy, Haiying Hou, and Xiaoji Niu, "Analysis and Modelling of Inertial Sensors Using Allan Variance", *IEEE Transactions on Instrumentation and measurement*, Cilt.57, No.1 Sayfa. 140-149, January 2008.

Seda Karadeniz Kartal received her Bachelor's degree (2005) in electrical and electronics engineering from Firat University. She is currently working toward a Ph.D. degree at Middle East Technical University, Ankara, Turkey.

She is a research assistant at Middle East Technical University. Her research interests include modelling, control, guidance, navigation and system identification of unmanned underwater vehicle.

Kemal Leblebicioğlu received his Ph.D. from the Department of Mathematics of the Middle East Technical University (METU), in 1988.

He has been a full professor in the Department of Electrical and Electronics Engineering of METU since 1999. He has a background in optimization, optimal control theory, computer vision, intelligent systems, flight control, walking robots, and unmanned vehicles.

Dr. Leblebicioğlu has been a member of numerous international program committees and has been (co-) author of more than 40 journal papers, book chapters, and numerous conference papers. He is on several IFAC technical committees. He was the editor of the journal ELEKTRİK published by

TÜBİTAK (Turkish Scientific and Technical Research Council) from 1996 to 2009. He is on the editorial board of several Turkish technical journals. He organized an international workshop on computer vision and intelligent systems and two national conferences on control. He conducted several R&D projects as project leader and researcher and gave consultancy to BİLTEN (TÜBİTAK's Research Center on Electronics and Information Technologies), 3rd Supply and Maintenance Center of Turkish Air Forces (Ankara), TÜBİTAK-İltaren, and to ASELSAN on the topic of "real time automatic visual tracking of targets." Now he is studying unmanned vehicles, in particular, unmanned air and underwater vehicles. He completed a project (supported by TÜBİTAK) on the construction of a special unmanned under-sea vehicle, called ULISAR, which is a mixture of an ROV and a fully automated system. His current project, also supported by TÜBİTAK, is on the coordinated guidance of multiple unmanned air vehicles. Presently, he is also working as an advisor for the Bahçeşehir University Foundation, BUEV, about matters related to education, developing educational systems to help students and teachers, intelligent testing, and student evaluation systems.

The Video Game As Practice For Developing Virtual Reality Sports Jumping Skills in Children 5 Years. Case Study of Innovative Practices in Educational Institutions of Bogotá, Colombia

J. LÓPEZ, L.COY, J.CAVIATIVA, Y.GUZMAN, A.GUTIERREZ

Abstract — The world of video games is a field of constant development. The possibility of applying video games to learning and developing of bodily skills has permitted deploying this experience; the object of this quasi-experimental research consisted in analyzing the effects of using video games where physical activity and sports practice are exerted, promoting development on locomotor skills like jumping in children 5 years of age, belonging to state educational institutions in Engativá in the city of Bogotá (Colombia) for which 30 children were selected: 15 (control) and 15 (intervention). These children were subjected to the application of tests during continuous weekly sessions during three moments: warm up, interaction and application with the running virtual game, and stretching of the principal muscle groups of lower limbs. Before the applications, anthropometric data (weight, height, and BMI) were collected from the study population. The study found that the motivation induced by the sports practice, upon including the use of virtual reality video games with repetition exercises with bodily extension and elongation movements, influences the locomotor jump pattern in children 5 years of age. These permit broadening the capacities of sports practices as part of the bodily and sports formation process in state educational institutions.

Keywords: Virtual reality video games, locomotor patterns, sports skills, innovative strategies

I. INTRODUCTION

Use of virtual reality games has extended to diverse human activities in populations of all ages. Besides the technological advances, among the reasons why the use of these games has increased are the family dynamics where parents spend their days in working activities and children spend their leisure time on video games, the internet, and television, given that the time to share with the family in outdoor activities has been reduced. This has contributed to the growth of the digital culture throughout the world. Information and Communication Technologies (ICTs), as well as virtual reality games have become strategic allies in learning processes, given that interactivity increases motivation, a process underlies learning. In this same sense, said games have been used as cognitive and physical

rehabilitation methods [1], in neurological pathologies, like strokes and Alzheimer's disease [2].

Therein, the use of video games, by using and developing national standards as pedagogic strategy when employed with child population, may provide an exhaustive framework of interest and investigation and – in relation to quality standards – may be related to the integration of new technologies that, according to the Ministry of Information Technologies and Telecommunications (MINTIC, 2014), must be appropriated from the inclusion of educational strategies that support the teaching practice as of the institutional management and evaluation of their impact upon child populations.

Another important aspect to keep in mind is how mental learning and motor learning are generated in human beings. Thus, mental learning is defined as the acquisition and improvement de knowledge and intellectual capacities and skills [3]; psychomotor learning is the process through which the child relates, is aware of and adapts to his/her surroundings, including aspects like expressive and comprehensive language, visual-motor coordination, gross motor skills, balance, and the social-affective aspect, which is related to self-esteem. Through the manipulation of objects and domain of space by walking, the child acquires sensory-motor experiences that will allow him or her to construct concepts, which will translate into ideas and will develop their thinking and capacity to reason [4,5]. This association is evidenced during the children's early stages when their cognitive processes are produced through their motor activity, by manipulating and moving objects around them, to create significance as a result of the construction of learning through their context, bringing the children to stimulate their motor processes and motor skills. This inter-relation must be reinforced through the combination of the teaching (mental learning) with motor activities like the use of virtual reality games (motor learning).

Motor learning and, thus, significant learning can be influenced by virtual reality games through their influence on the pre-motor cortex. This cortex has direct connections with

the primary motor area and contributes with 30% of the axons that make up the corticospinal and corticobulbar tracts. The pre-motor cortex is constituted by two components: lateral and medial. The lateral component facilitates development of conditional tasks with visual cues and the medial component participates in the selection and initiation of movements through more internal than external signals, motivation [5]; virtual reality games can influence directly on motor learning processes by providing visual cues and impacting directly on motivation.

This contextualization seeks to apply, through a quasi-experimental design, the development of sports jumping skills in children 5 years of age from the influence of the practice of virtual reality video games. This technology is seen as an innovative learning strategy within educational institutions, promoting, in the area of sports, development of skills appertaining to the human bodily formation that must be developed from early ages, through the stimulus of mental skills concerning processes like attention and memory, generating effects on the performance of coordinated locomotor patterns that must be developed between 5 and 7 years of age and which are fundamental to start the practice of a sports discipline.

Besides the study's principal objective, it sought to highlight that the use of this technology is useful as an innovative pedagogical strategy in the physical education formation field in state institutions as a support tool for the teacher and a complement in conducting outdoor activities. Given the need for surveillance and control by government entities on the prevention of disease related to lack of physical activity or sedentary lifestyles, this type of technology is suggested as a way of staying active at home, propitiating family union and healthy lifestyles.

II. CONCEPTUAL FRAMEWORK

It is important to introduce the basic element of study, the evolution of video games and their transition onto the world of sports education. With this concept, it may be stated that it is an ubiquitous tool in the activities of professors interested in innovating as part of the formation process; concretely, it could be defined as a document containing the organized records of all the data and knowledge that refer to physical education and video games and which serve as foundation for the diagnosis of a useful educational application in this area.

Games have been broadly addressed by different academic disciplines like math, history, philosophy, sociology, motor praxeology, etc. [6]. The importance of games for human beings and for culture is more than a universal consensus; in the end, it fulfills the ludic and socialization needs of all animals, including humans, according to that indicated by Caillois 1958 cited by Bortoleto 2006 [6]. This is why ludic activities in general and the game motor constitute a fundamental content for the educational and/or recreational

environment, with this being largely the responsibility of physical education teacher.

Virtual reality games have also been used to maintain active lifestyles [7], [8], [9]. Some authors establish the relationship between the practice of physical activity and motor development [10]. Due to these findings, it may be inferred that a relationship exists between using virtual reality games and the maturity of basic locomotor patterns.

In spite of the benefits derived from using virtual reality games, some studies have observed negative effects from their practice, mainly for the visual system [11] and the musculoskeletal system [12].

Studies recommend increasing research on attention processes and their relationship with the maturity of locomotor patterns implied in the use of these types of video games. Due to this, the objective of this study was to identify the influence of the practice of virtual reality games on attention processes related with the jump motor pattern in children 5 years of age.

A. *Virtual Reality Video Games*

A videogame is an interactive information technology program destined for entertainment and which can operate in diverse devices: computers, game consoles, mobile phones, etc. It integrates audio and video and permits enjoying experiences that, in many cases, would be difficult to experience in reality.

The characteristics of video games include: the quality of graphics (at the beginning in two dimensions, and currently in 3D), game control must be easy to use and intuitive, and sound (from the speaker to surround sound) [13].

Diverse types of video games exist, among them we could name adventure games (intelligence tests or puzzle solving to advance), arcade (skills activities), sports, strategy (coordinate actions), role playing (players manage a personality and it evolves during the game according to the user's decisions), and simulation (some type of action is simulated like, for example piloting an airplane) [13].

Virtual reality games enter an exclusive range of tools in which users can venture creatively to where the limit of their imagination permits. Therein lays, quite possibly, the biggest attraction, given that imagination and creativity have an opportunity to be executed in an unlimited and artificial "world". The origin of these games is the Department of Defense of the United States, where they were created as material for an aviation class during the 1970s for flight simulations by practicing and not risking lives [14].

B. Locomotor Patterns

Children progressively develop skills in movements, from the first involuntary reflex movements to highly complex abilities. The early childhood period (2 to 7 years of age) is critical for the development of elemental motor patterns. Children who do not mature motor patterns during this period frequently exhibit difficulties in carrying out more complex motor skills like sports movements.

Locomotor patterns are those movements that allow children to explore space; these include: walking, running, high and long jumping, hopping, galloping, and climbing. These fundamental movements are observable behavioral patterns that can be divided into three stages [15]:

Initial Stage: This stage represents the first guided goal the child tries to execute; its movements are characterized by an inappropriate sequence, restricted use of the body, and poor rhythmic coordination. There is almost no spatial and temporal integration; the two-year-old child's movements of locomotion, manipulation, and balance are typical of this stage.

Elemental Stage: There is greater control and better rhythmic coordination of the fundamental movements. Temporal and spatial elements are more coordinated; however, the pattern is still exaggerated or restricted. Children between 3 and 4 years of age present a great variety of movements during the Elemental Stage.

Mature Stage: It is characterized by mechanical efficiency, coordination, and controlled performance. Children between 5 and 7 years of age can and should be in this Mature Stage.

C. Relationship between Video games and Physical education

Currently, the standards of creating and designing video games from Information and Communication Technologies serve to design learning materials, given that their construction employs cognitive modification elements.

Under this design conception of teaching – learning situations emerge: 1. Speed as greater experience to process information upon designing arcade video games; 2. connectivity, from synchronously and asynchronously operability, which is why it is supported on the capacity to solve problems; 3. constant action in which children and adolescents rarely need manuals to learn the operation of information technology elements; thereby, learning them intuitively [16]. 4. Guidance in solving problems, guided from the approach to the capacity to design in which constant revision of the action exists; 5. Immediate reward in which adolescents request usefulness of contextualized knowledge; 6. Importance of fantasy, which is a primordial element of

current video games; 7. Positive vision of technology; without the existence of fears associated to it, which permits its use.

With respect to sports video games reference model, these games reproduce known sports, like: football, basketball, or golf, which are available in 2D and 3D. These games require coordination and strategy, particularly if the player has to manage a team [17] [6]; according to this, age considerations exist, given that activities are associated to the level or age group to which the games are destined. Likewise, the language level used must be adequate for the age group.

The video game's action time must be optimal to complete the challenges, thus, ensuring [17] that students have the necessary time to end the levels of the match and benefit from the educational characteristics.

D. Pedagogic Considerations of Using Video Games

A pedagogic structure based on the structure of video game use within physical education and sports education that has been traditionally used by teachers, without other support than gymnastic artifacts or open spaces centered on rules can be used to emulate conditions of static locomotion. Because of this, which when implementing an evaluation from measuring with learning curves with the arcade game, it is considered necessary to have a simple curve adequate for the age of application – permitting players (boys, girls) to make mistakes and start again.

Every activity designed for use in the educational context must consider using characteristics in said design that promote the pedagogical needs of the population in which these will be employed. These activities must be constructed under ludic concepts, facilitating conceptualization and manipulation by the students to whom they are aimed [6]. Experience in the implementation of video game standards and their use permits enhancement and sufficient maturity on the development of the individual's own processes.

According to the aforementioned, another pedagogic consideration is extracted from the video game's own content, which should illustrate the subject taught, [17] although the content may not be strictly related with the study plan. As it regularly occurs, it may contribute a clear and simplified representation of any of the concepts taught.

Clear objectives. According to this, professors should make sure that the objectives of the game are clearly defined for students to know exactly what is asked of them. Frustrating situations may arise if the instructions are not precise and students could feel blocked because they do not know how to advance in the game.

Clear progress. Professors should check if the player's progress is shown in markers or progress bars and will help students to have a positive attitude with respect to their

provision and will show them that their actions influence their progress, which should motivate players to become responsible for their learning activities.

E. Follow up and Evaluation of Using Video Games and Physical Activity

A standard in designing said activities and themes on virtual reality has opted for working from the evaluation and follow-up systems that permit an adequate follow-up system of the student. For the thematic interest, the child's progress can be measured from the performance and evaluations prior to its use. Said conditions and aspects can be analyzed. This was not included in most video games and simulators; however, some are compatible [17] [6] with sharable content object reference model (SCORM) and can be integrated onto a learning management system (LMS), which permits following the process of students and identifying the points that need more attention.

III. METHODOLOGY

This study was developed through a quasi-experimental design with pre-test, post-test, and a control group; the study was approved by the ethics committee at an educational institution in Bogotá, Colombia. The intervention group was exposed to the use of virtual reality video games from the Nintendo Wii console, during two sessions per week of virtual reality games for one month. Each session was carried out in three parts or moments: a first part for warm up; then, 15 minutes of interaction with the game, and – finally – stretching of the main muscle groups of the lower limbs. The control group was not exposed to any type of interaction with the video game.

La study population fulfilled the following inclusion criteria: children 5 years of age who belonged to educational institutions located in the city of Bogotá, who had never practiced virtual reality sports games, and who had all their physical and mental capacities, whose locomotor jump pattern was at the Initial Stage, and whose parents or legal guardians had signed the informed consent. Informed consent was also obtained from the children.

After signing the paperwork to participate in the study, the study population was evaluated pre- and post-intervention, which contained personal and anthropometric data, as well as data on the stages of their fundamental patterns.

Upon ending the intervention and evaluation processes, the data were used comparatively among the groups and among the evaluations, thus, identifying if any influence existed of the video games on the locomotor jump pattern in 5-year-old children.

IV. RESULTS AND DISCUSSION

The total study population was 30 children of which 15 belonged to the control group and 15 belonged to the intervention group. The anthropometric characteristics of the study population and of the intervention population are shown in Table 1. From these data (weight and height), the body mass index (BMI) was obtained.

Table 1. Anthropometric characteristics of the study population

	Weight	Height	BMI
Control	19.8 ± 2.39	108.8 ± 7.14	16.7 ± 1.90
Intervention	18.0 ± 2.86	107.8 ± 6.41	15.5 ± 1.88

From the table, it may be noted that the BMI in the population is normal, with the presence of low weight, overweight, or obesity conditions.

Figure 1 describes changes in the locomotor jump pattern stage, in the control and intervention groups, that is, children who were in the Elemental Stage and moved on to the Initial Stage or who were in the Initial Stage and moved on to the Mature Stage.

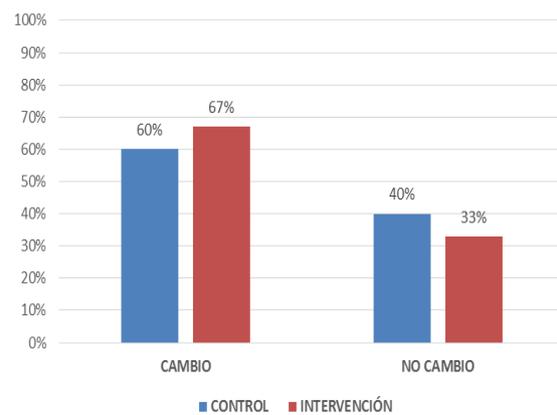


Figure 1. Percentage of change during the stage of the locomotor jump pattern in the control group and intervention group

The start of these types of ludic-motor situations adapted and/or created based on motor skills –locomotor jump pattern – whose percentage was increased from the use of virtual reality video games, making use of the ludic and motivational potential interpreting the action that can be generated when using an arcade video game with educational characteristics for the area of primary physical education.

Development of this measurement permits defining if the increased incursion into the new Information and Communication Technologies from improved strategies of using video games may differ on the motivation of the sports practice.

Although references of studies on the effect of virtual reality games in developing locomotor patterns have been limited, this study agrees with investigations that have found positive

effects of these types of games in learning motor activities in patients with neurological damage, such as: cerebrovascular disease, spinal cord injury, multiple sclerosis, Parkinson's disease, and brain damage due to trauma [18]. Said effect can be attributed to three key elements: repetition, sensory feedback, and the motivation these games provide due to the interaction and immersion in the video game.

Controversy exists about the quality of motor activities learnt. Prior studies consider that the movements are similar or equivalent, this is why they are recommended to learn motor tasks [19]; however, due to the difference in perception less precision has been observed in the movements [20], [21], [22].

Another point of debate is the applicability of motor activities in reality; nevertheless, some investigations have found that virtual reality games provide motor learning in the three dimensions of space, corresponding to the movements made in the real world [16]. The relationship of motor learning with intellectual learning is such that virtual reality systems are used for training in non-medical practices, like aviation, nuclear and industrial systems, and in medical practices like training for endoscopic surgery, general surgery, vascular surgery, orthopedics, and neuro-rehabilitation.

As a Project to include technology in the development of locomotor skills, using virtual reality video games under professional guidance, from teacher support and advice, it permits diminishing displacement to adequate scenarios for sports practices and using space in state educational institutions. Thus, body awareness [6], refers to open spaces as pedagogical elements in which bodily management is part of the expression represented by the game itself as element of formation and not as the sole formation element of the individual who interacts.

Finally, state institutions, from the need to incorporate these types of innovative pedagogical strategies onto infant populations, must permit incorporating technologies in the generation of cognitive processes as in the development of physical skills in a forefront educational system; distancing boys, girls, and later future adolescents from useless activities and waste of time. Said development and design of motivational techniques under the conception of greater physical activity will make useful times of dedication to sports education, development of cognitive skills and, hence, physical skills, contributing to a holistic formation in adolescents in future societies.

The execution of the influence of the use of video games as strategy permits exercising diverse psychomotor coordination skills, while delving into the knowledge of the regulations and strategies of sports. This, applied to the case study for Santa Fé in the city of Bogotá, Colombia will provide the education system a redefinition of its use improvement component in the system of sports education strategies.

V. CONCLUSIONS

This research starts a path where investigations may be conducted to include broader study samples, with more objective tests to evaluate the locomotor pattern, increased amount of video game sessions, and statistical tests to determine the effects of virtual reality games on the development of motor patterns on children. This research found a slight increase in the number of children who moved on to a higher stage in the jump pattern in the group conducting two weekly sessions of virtual reality games for one month (intervention group) with respect to the control group.

Boys and girls five years of age may benefit from using virtual reality video games upon the development of skills and mental agility processes from attention and memory on problem solving processes upon identifying execution errors, verification and repetition as accomplishment process. Said causality is associated to the motor development of extension, flexion, and elongation of lower limbs in boys and girls measured by sports-type arcade video games in Colombian state educational institutions; likewise, they will benefit from their right to recreation and physical activity, which ensures them independent attention and proprioceptive promotion of motivation to sports activities, among them long jump, discus throw, running, and basic gymnastics recognized as having limited access for this population.

The vision created upon using virtual reality video games as of the development of motor patterns through sports applications in education will allow rapid development and engagement to the current educational strategy in basic primary school, and the diagnoses applied to it by professionals in adolescents sports therapy. The theme of sports education has many parts that are quite difficult to address in a single investigation; however, these types of proposals and implementation are necessary from the innovative incorporation of new technologies onto education, which can contribute to the knowledge of this transcendental theme for our country on issues of education and sports health.

Using video games in educational institutions is beginning to gain force in Colombia from the guiding policies by the Ministry of New Technologies (MINTIC); however, given that this research was conducted on state educational entities, the differences from the educational environment and from health prevention permit its implementation in curricular systems to be done gradually and, currently, few have adopted it through their own initiative because this emerges from the skills of the primary physical education teacher, his/her academic competency, and particular openness to using video games as support tools for the teaching tasks.

The link between an adequate low-cost pedagogical strategy, like incorporating video games in the classroom, and integrating innovative non-conventional strategies aimed at

learning attention and memorization processes, make using virtual reality applications into versatile tools and of easy access to state educational institutions, which by developing inclusion processes in curricular plans in the area of sports sciences permit exerting control and improving the development of basic locomotor skills on adolescents population, specifically children between 5 and 7 years of age that enable their future involvement in sports practices.

Given the increasing amount of virtual reality video games in existence or which can be developed, this research focuses an analysis sample of a sports video game on a study group, permitting the analysis of the capacity that can be generated by its use in physical and sports education. Designing arcade sports-type video games broadens the possibility of restructuring objectives on students' performance and greater physical activity, but it is mainly a task of state institutions of primary basic education and professors from primary physical education to contribute to greater performance of locomotor skills by the population of application from the incorporation of pedagogical strategies that include their adequate use and under particular objectives.

VI. BIBLIOGRAPHY

- [1]. GOMEZ MORA, M. Aplicación de realidad virtual en la rehabilitación cognitiva. *In: Revista Vínculos. Ciencia, Tecnología y Sociedad.* January-June, 2013. Vol 10, no 1. 130-135.
- [2] FRACER A., MICHELLE A, WALTER B., Estudios en Tecnología de la Salud e Informática. *In: Revisión anual de la ciberterapia y telemedicina,* 2012, Vol. 154, p. 229-234.
- [3] MEINEL K., SCHNABEL., G. Teoría del movimiento. [On line] 2004. Available in internet: http://books.google.es/books?id=_pCVYGDg4EEC&dq=learning+mot+or&hl=es&source=gbs_navlinks_s [citado en 18 Noviembre de 2012].
- [4] TORRES, C. La actividad lúdica y su incidencia en el desarrollo psicomotriz de los niños y las niñas de 0 a 6 años del primer año de educación básica de la escuela " Mercedes de González de Moscoso" del barrio Bellavista, parroquia Manú, Cantón Saraguro, Provincia de Loja, periodo 2009-2010. [On line] 2011. Available in internet: <http://dspace.unl.edu.ec:8080/xmlui/bitstream/handle/123456789/3377/TORRES%20TORRES%20CECILIA%20DEL%20CARMEN.pdf?sequence=1> [cited 18 November 2012].
- [5] PURVES, D. Neurociencia. Editorial Médica Panamericana. 2007.
- [6] AVILÉS V. 2004. Taller de experiencia en ONG Raíces. Una experiencia de pedagogía teatral con niños, niñas y adolescentes víctimas de explotación sexual comercial. Oficina Internacional del Trabajo.
- [7] MADDISON R., FOLEY L., Efectos de los videojuegos activos en la composición corporal: un ensayo controlado aleatorio. *In: American Journal of Clinical Nutrition,* June, 2012, vol. 19, p. 156-163.
- [8] ZURANO CONCHES L., Investigadores estudian la eficacia en la población infantil de los videojuegos activos como facilitadores del ejercicio físico. *In: Universitat Politècnica de Valencia,* February, 2011, p. 2
- [9] GRAVES LE, RIDGERS ND, WILLIAMS K, El costo fisiológico del uso del Nindendo Wii en adolescentes, adultos jóvenes y adultos mayores. *In: Journal of Physical activity and Health,* May, 2012, vol. 7, no 3, p. 393-401.
- [8] BAENA A, GRANERO A, RUIZ PJ. Procedimientos e instrumentos para la medición y evaluación del desarrollo motor en el sistema educativo. *In: Journal Sport and Health Research,* June, 2009, Vol. 2, no 2, p. 63-76.
- [11] MURCIA P. L., Incidencia del uso de los videojuegos en alteraciones visuales, ergonómicas en niños de 9 a 14 años [on line]. *In: Ciencia y Tecnología para la salud visual y ocular.* 2004. Available in internet [cited 18 November 2013].
- [10] BELTRAN V. J., VALENCIA A., MOLINA J. P., Los videojuegos activos y la salud de los jóvenes: revisión de la investigación. *In: Revista internacional de medicina y ciencias de la actividad física y el deporte,* March, 2011, Vol. 11, no 41, p. 203-219.
- [13] GALEON. Historia de los Videojuegos, [on line] 2012. Available in internet: www.historia-video-games.galeon.com [cited 25 June 2012].
- [14] PAEZ A., Virtual reality, [on line] Monografias.com, 2007. Available in internet: <http://www.monografias.com/trabajos53/realidad-virtual/realidad-virtual.shtml> [cited 25 June 2012].
- [15] Mc CLENAGHAN Y GALLAHUE. Movimientos fundamentales su desarrollo y su rehabilitación. Editorial Panamericana. 1985.
- [16] CANO DE LA CUERDA R, MUÑOZ-HELLÍN E, ALGUACIL-DIEGO IM, MOLINA-RUEDA F. Telerrehabilitación y neurología. *In: Rev Neurol.* 2010; 51:49—56.
- [17] FELICYA PATRIC, Videojuegos en el Aula. Manual para profesores, ¿Cómo se usan los videojuegos en el aula? European Schoolnet. Hot studio, May, 2009, p. 16-18.
- [18] PEÑASCO-MARTÍN B, REYES-GUZMÁN A, GIL-AGUDO A, BERNAL-SAHÚN A, PÉREZ-AGUILAR B, PEÑA-GONZÁLEZ AI. Aplicación de la realidad virtual en los aspectos motores de la neurorrehabilitación. *In: Rev Neurol* 2010; 51: 481-8.
- [19] HOLDEN, MK. Virtual environments for motor rehabilitation: review. *In: Cyberpsychol Behav* 2005; 8: 187-211.
- [20] VIAU A, FELDMAN AG, MCFADYEN BJ, LEVIN MF. Reaching in reality and virtual reality: a comparison of movement kinematics in healthy subjects and in adults with hemiparesis. *In: J Neuroeng Rehabil* 2004; 1: 11.
- [21] SUBRAMANIAN S, KNAUT LA, BEAUDOIN C, MCFADYEN BJ, FELDMAN AG, LEVIN MF. Virtual reality environments for post-stroke arm rehabilitation. *In: J Neuroeng Rehabil* 2007; 4: 20.
- [22] KNAUT LA, SUBRAMANIAN SK, MCFADYEN BJ, BOURBONNAIS D, LEVIN MF. Kinematics of pointing movements made in a virtual versus a physical 3-dimensional environment in healthy and stroke subjects. *In: Arch Phys Med Rehabil* 2009; 90: 793-802.

First Author: Juan Diego López Vargas, Telecommunications Engineer Santo Tomas University of Colombia, specialist, and doctoral magister in Integration of Information Tencologias Livelihood Organizations at the University of Valencia Spain Politecnca. Consultant and Director of Special Projects and Research at Thinking Colombia University graduate Eccí Colombia.

Second Author: Lena Coy, Physiotherapist Manuela Beltran University of Colombia, Occupational Health specialist Manuela beltran candidate University MA in education at National University of Colombia. Teacher researcher Manuela Beltran University Physiotherapy program.

Third author: Janeth Caviativa, BA in Biology from Colombia University of Education, Education Specialist University of Education of Colombia, Master in Education District of Colombia University, a Doctor of Education candidate Distrial University of Colombia. Teaching Systems Engineering Research Manuela Beltrán University Program.

Four Author Yohan Guzman Fourth Degree in Biology University of Education of Colombia, Specialist and Master of Education candidate Universdiad Pedagogica of Colombia. Teaching Basic Science Research Program Universdada Manuela Beltrán.

Fifth Author. Adriana Gutierrez Physiotherapist National University of Colombia and a Master in Neuroscience National University of Colombia. ECCI University Teaching Biomedical Engineering Program

On Selection of Efficient Fuzzy Models Incorporated with Multi-Objective Reactive Power Control

Ragab A. El-Sehiemy

Department of Electrical Engineering University of Kafrelsheikh, Kafrelsheikh, Egypt

Email of the author: elsehiemy@eng.kfs.edu.eg

Abstract- This paper proposes an efficient multi-objective procedure for selection of the best suitable linear fuzzy models that improve the performance of the reactive power problem at different operating conditions. The proposed linear fuzzy models are triangular and trapezoidal are investigated for the control and dependent variables. Added to that, the selection procedure is carried out incorporated with several control strategies which accomplished with the variant operating conditions. The choice of the best membership models for each variable of the control and dependent variables are applied for each operating condition followed by the sufficient control actions that able to eliminate the emergency effects. One or more objective functions are considered for each operating condition according to the specified control strategy to enhance the system performance. The proposed MFLP is applied to the West Delta region system as a part of the Egyptian Unified network.

Keywords: reactive power control, multi-objective, emergency, emergency conditions, linear fuzzy modeling.

I. Introduction

The ability of fuzzy logic to represent the sorts of qualitative statements, employed by humans, has found favor among many engineers and it is effective in solving multi objective problems. The choice of shape depends on the individual application. Different fuzzy models have been presented in [1] to solve the fuzzy-based optimal power dispatch problem. Abou El Ela et al. [2] solved the optimal active power dispatch problem using MFLP technique involving preventive action constraints. In [3], a multi-objective fuzzy-based incorporated artificial bee algorithm to solve economical/environmental problem and discrete OPF problem respectively is presented. A dynamic fuzzy interactive approach is developed for distributed generation expansion planning in [4]. In [5] and [6] the multi-machine system is considered whereas in [6] fuzzy logic is interacted with differential evaluation algorithm for robust power system stabilizer with minimum rules. For automated distribution system, an optimal switch placement problem is formulated as multi-objective fuzzy model then solved via modified shuffled frog leaping algorithm [7]

In [8], the optimal reactive power dispatch (ORPD)-based comparison studies between two linear fuzzy models for control and dependent system constraints has been discussed. These linear fuzzy models namely triangular and trapezoidal fuzzy models were used to solve the ORPD problem. It is concluded from that paper, the triangular shapes of fuzzy models are the best for minimizing the

power losses at normal operating condition for both control and dependent variables. Also, the use of triangular membership shape aims at minimizing the deviation of control and dependent variables from the permissible operational settings. An enhancing of reactive power management considering both technical and economic issues is proposed in [9]-[11].

One of the major operating tasks of a power system is to maintain the bus load voltages within their limits for high quality consumer services. The electric power loads are not constant but vary from time to time. Any change in the power demand causes lower or higher voltages [12]. The loss minimization is one of the important objectives in operating the transmission networks [13]. Appropriate provision for reactive power is essential for power systems in order to ensure secure and reliable operation of power systems. Reactive power is strongly related to bus voltages throughout a power network, and hence reactive power services have a significant effect on system security. Insufficient reactive power supply can result in voltage collapse, which has been one of the reasons for some recent major blackouts [14]. Wu et al [15] described an optimal power flow (OPF) based approach for assessing the minimal reactive power support for generators in deregulated power systems. He et al. [16] presented a method to optimize reactive power flow with respects to multiple objectives while maintaining voltage security. The management of reactive power reserves in order to improve static voltage stability by using a modified particle swarm optimization algorithm was presented in [17].

In this paper, the ORPD problem is solved using MFLP technique to determine the efficient fine fuzzy tuning model incorporated with the optimal settings of control variables with of power system variables. The specific objectives are to minimize the real power losses, maximize the reactive power reserve while satisfying limits on all the variables.

Rest of this paper is organized as follows: formulation of the fuzzy based ORPD problem is introduced in Section 2. The linear fuzzy modeling for the objectives and constraints of the ORPD problem is presented in Section 3. The proposed procedure for maximal reactive power reserve is described in Section 4. Application results of case studies are presented in Section 5. The outcome of the current work is concluded in the last section.

II. FUZZY BASED ORPD PROBLEM

The optimal reactive power dispatch problem is formulated as a fuzzy constrained optimization problem to

minimize the real power losses. In this paper, the sensitivity parameters are used to represent the objectives and dependent variables in terms of the control variables [11].

The problem can be presented as follows:

$$\min \Delta F = \begin{bmatrix} \partial \tilde{F} / \partial \tilde{v}_g & \partial \tilde{F} / \partial \tilde{Q}_S & \partial \tilde{F} / \partial \tilde{t}_{ij} \end{bmatrix} \begin{bmatrix} \tilde{v}_g \\ \tilde{Q}_S \\ \tilde{t}_{ij} \end{bmatrix} \quad (1)$$

subjected to:

$$\Delta v_{g_i}^{\min} \leq \Delta \tilde{v}_{g_i} \leq \Delta v_{g_i}^{\max} \quad i \in Ng - \text{slack bus} \quad (2)$$

$$\Delta Q_{S_j}^{\min} \leq \Delta \tilde{Q}_{S_j} \leq \Delta Q_{S_j}^{\max} \quad j \in N_s \quad (3)$$

$$\Delta t_{ij}^{\min} \leq \Delta \tilde{t}_{ij} \leq \Delta t_{ij}^{\max} \quad i, j \in N_t \quad (4)$$

$$\Delta v_{l_i}^{\min} \leq \Delta \tilde{v}_{l_i} \leq \Delta v_{l_i}^{\max} \quad i \in N_b - Ng \quad (5)$$

$$\Delta Q_{G_j}^{\min} \leq \Delta \tilde{Q}_{G_j} \leq \Delta Q_{G_j}^{\max} \quad j \in Ng - \text{slack bus} \quad (6)$$

$$\Delta Q_{f_k}^{\min} \leq \Delta \tilde{Q}_{f_k} \leq \Delta Q_{f_k}^{\max} \quad k \in N_l \quad (7)$$

Where (\tilde{F}) is the fuzzy real losses in the transmission network; (\tilde{v}_g, \tilde{v}_l) is the fuzzy bus voltage of generator and load buses respectively; (\tilde{Q}_S) is the fuzzy reactive output from the switchable bus; (\tilde{t}_{ij}) is the fuzzy tap point of the transformer tap changer; (\tilde{Q}_G) is the fuzzy reactive output from generators and (\tilde{Q}_f) is the fuzzy reactive flow through lines. Ng is the number of generators; Ns is the number of switchable buses; Nt is the number of transformer tap changer; Nb is the number of buses; Nl is the number of transmission lines. The symbols (min, max and Δ) refer to minimum, maximum and change of any variable value, respectively.

The dependent variables (y) are represented in terms of control variables (x), as referred in equations (5)-(7), as:

$$y = C_{yx} \cdot x \quad (8)$$

where, C_{yx} is the sensitivity parameters of the dependent variables in terms of the control variables

Reactive power reserve of the generators is the ability of the generators to support bus voltages under increased load or disturbance condition. The reactive power reserve of any generator can be represented as:

$$Q_{G_{i,res}} = Q_{G_{i,max}} - Q_{G_i} \quad , i = 1, 2, \dots, Ng \quad (9)$$

Where, $Q_{G_{i,res}}$ is the reactive power reserve of generator i; $Q_{G_{i,max}}$ is the maximum limit of reactive power output of generator i which is the maximum limit of reactive power that the machine can supply; Q_{G_i} is the reactive power output of generator i at a certain operating condition.

The reactive power reserve of switchable devices can be represented as:

$$Q_{S_{j,res}} = Q_{S_{j,max}} - Q_{S_j} \quad , i = 1, 2, \dots, N_s \quad (10)$$

Where, $Q_{S_{j,res}}$ is the reactive power reserve of a switchable VAR source at bus j; $Q_{S_{j,max}}$ is the maximum limit of reactive power output of a switchable reactive power source at bus j; Q_{S_j} is the reactive power output of a switchable reactive power source at bus j at a certain operating condition.

Additional objective function, F_a , is to minimize the control variables' adjustments to alleviate the load voltage and VAR violations with an overall minimal adjustment of the control variables as:

$$\min F_a = \sum_{g=1}^{Ng} C_{vg} |\Delta v_g| + \sum_{i=1}^{Ns} C_{si} |\Delta Q_{Si}| + \sum_1^{Nt} C_{t_{ij}} |\Delta t_{ij}| \quad (11)$$

The values of C_{vg} , C_{si} , and C_{ti} are chosen to reflect the relative positions of control variables. These factors are set equal 1.0, which means equal priority to all of the control variables.

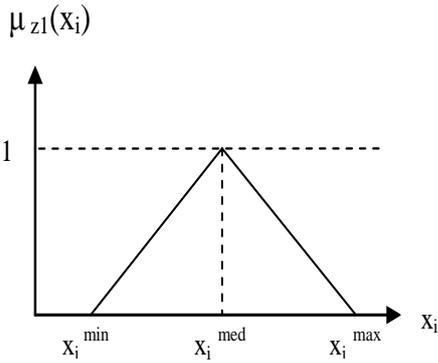
III. FUZZY ORPD MODELING

There are various types of membership functions which are commonly used in fuzzy set theory to solve the optimal active power dispatch in power systems. One of the best membership functions to represent the control and dependent variables in power systems was the triangular shape [9] and [10]. In this paper, an effort is employed for identify the best fuzzy membership model for reactive power control problem.

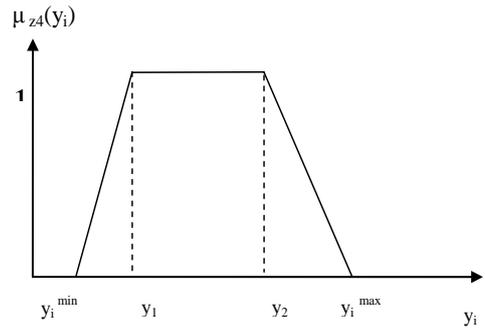
A) Fuzzy modeling of constraints

The triangle fuzzy modeling for the control variables (x) is shown in Figure 1.a. These control variables are the voltage at generators buses, reactive power output at switchable buses and transformer tap changer. It is seen that a membership function equal to 1 is assigned to x_i^{med} . x_i^{min} and x_i^{max} are the minimum and maximum limits of a control variable (x_i), respectively. x_i^{med} is a point between the minimum and maximum limits of the control variables. With the best tuning of the control variables especially the generators voltage to enforce it towards desired values to enhance voltage security and it is less than the maximum limit of each one.

Similarly, a triangle fuzzy modeling for the dependent variables (y_j) is shown in Figure 1.b. It is seen that a membership function equal to 1 is assigned to y_j^{med} . Each dependent variable is represented by two linear constraints for the upper and lower limits. y_j^{min} and y_j^{max} are the minimum and maximum limits of each dependent variable (y_j), respectively. y_j^{med} is a point between the minimum and maximum limits of each dependent variable and it is less than the maximum limit of each one.

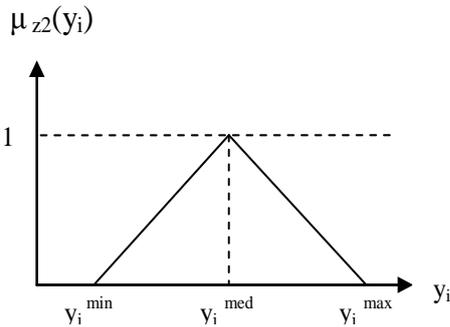


a. Control variables (x_i)



b. Dependent variable

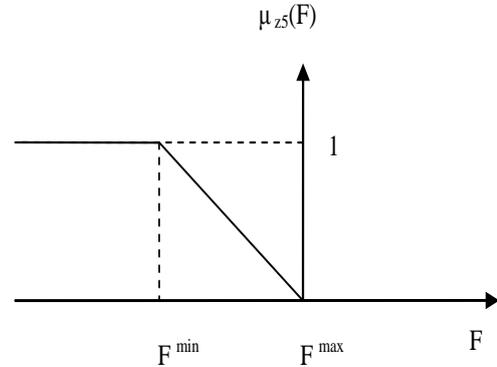
Figure 2 Trapezoidal membership model



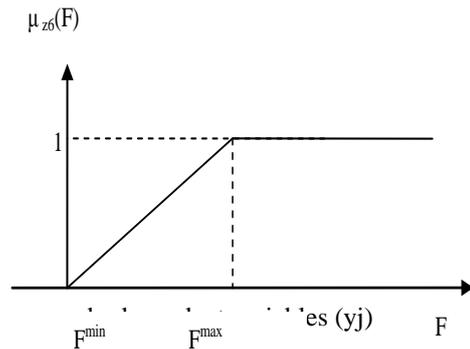
b. Dependent variables

Figure 1 Triangular membership model

Similarly, the trapezoidal fuzzy modeling for the control variables (x) is shown in Figure 2.a. It is seen that a membership function equal to 1 is assigned to the interval $[x_1 \ x_2]$. x_i^{\min} and x_i^{\max} are the minimum and maximum limits of a control variable (x_i), respectively. x_1 and x_2 are the two arbitrary points between the minimum and maximum limits of the control variables, with best tuning of the control variables especially the generators voltage to enforce it towards desired values to enhance voltage security. In a similar manner, the trapezoidal fuzzy modeling for the dependent variables (y_j) is shown in Figure 2.b. y_1 and y_2 are two arbitrary points between the minimum and maximum limits of each dependent variable.

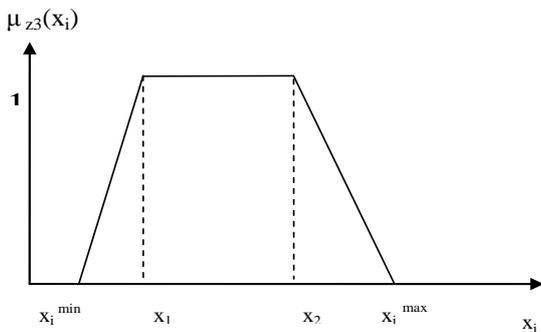


a) minimum model



b) maximum model

Figure 3 Fuzzy membership modeling of objective functions



a. control variables (x_i)

The fuzzy modeling of the minimization model (minimizing the incremental of real power losses) and maximization model (maximizing the reactive power reserve) are shown in Figure 3.a and 3.b, respectively. F^{\min} and F^{\max} are the minimum and maximum limits of different objective functions

IV. PROPOSED CONTROL STRATEGIES

The choice of the on-line control actions is dependent on the nature of the system operating condition. The following sections show various control actions (preventive/ corrective control actions) based on the operating condition. The proposed control actions allow the operator to ramp the constraints that suitable to the operating condition. Also, the availability of different options for the same operating condition helps the power system operator to choose the most suitable corrective actions according the operation requirements. For example, in the predicted emergency, preparing certain reserve from different reactive power resources gives the operators the multiple options based on the system performances as:

- The highest reduction in power losses
- The lowest level of voltage deviation
- The maximum of reactive power reserve level
- The highest loadability level for transmission systems
- In the power market environment, the highest reduction in reactive power sources.
- The lowest reactive power costs

The control settings are optimally calculated with satisfying the operating constraints in order to achieving the needed single or multiple objective functions. The considered operating conditions are:

- Case 1: normal operating condition
- Case 2: predicted emergency condition
- Case 3: combines both normal and predicted emergency condition:

V. Multi-objective Fuzzy Linear Programming Technique

For managing the reactive power control problem, the MFLP technique is performed by maximizing the minimum of the satisfaction parameters as:

$$\text{maximize } \lambda, \tag{12}$$

where,

$$\lambda = \min \{ \mu_{z1}, \mu_{z2}, \dots, \mu_{zi} \} \tag{13}$$

where, μ_{zi} is the membership functions of the constraints for control and dependent variables as well as the objectives constraints of real power losses and reactive power reserves, within range of [0-1] for all constraints.

VI. APPLICATIONS

A) Test Systems

The test system is that of the West Delta region [11], which is as a part of the Unified Egyptian Network and consists of 52-bus and 8 generation buses. These buses are connected by 108 lines. Shunt compensation limits at buses 18, 20 and 42 have been assumed between 0 p.u and 1 p.u (the base voltage is 66 kV, while the base MVA is 100). On Load Tap Changer (OLTC) limits between buses 4-25 and 11-28 have been assumed between 0.9 and 1.1 p.u.

B) Procedure for Fuzzy Membership Modeling

Table 1 shows the twelve fuzzy models with multiple objective functions are considered in order to choose the best fuzzy membership model which is suitable for each of control and dependent variables at variant operating conditions. The Newton Raphson load flow method is applied to obtain the initial operating conditions for both tested system. The operating voltage range for all load buses is $1 \pm 5\%$. Consequently, the minimum and maximum operational voltages at load buses are located in the range 0.95 to 1.05.

Table 1 Fuzzy membership models of control & dependent variables

Variables	Δv_g	ΔQ_s	Δt_{ij}	Δv_l	ΔQ_g	Δq_f
Model 1.	1	1	1	1	1	1
Model 2.	1	1	1	1	1	2
Model 3.	1	1	1	1	2	2
Model 4.	1	1	1	2	2	2
Model 5.	1	1	2	2	2	2
Model 6.	1	2	2	2	2	2
Model 7.	2	2	2	2	2	2
Model 8.	2	2	2	2	2	1
Model 9.	2	2	2	2	1	1
Model 10.	2	2	2	1	1	1
Model 11.	2	2	1	1	1	1
Model 12.	2	1	1	1	1	1

- 1) refers to triangular membership function,
- 2) refers to trapezoidal membership function

C) Results & Discussions

Table 2- Table 4 show the results of proposed ORPD using the twelve fuzzy modeling at different operating conditions (Cases 1-3), respectively. For case 1, the maximum reduction of the real power losses is obtained (13.52%) at models 1 and 11. While, the lowest real power losses (3.99%) is occurred at model 8. The ORPD using fuzzy modeling (models 1-12) has more real power losses reduction. While, the real power loss reduction is dependent on the selecting of the intermediate points of membership functions for the control and dependent variables. Figure 5 shows the voltage profile at load buses 18, 20 and 21. These buses have a voltage levels below the minimum voltage limits (0.95 pu). All fuzzy models remove the violation in these voltages and preserve them within the permissible limits.

Considering, Case 2, the maximum decreasing in the voltage deviation is obtained (0.0224) at model 6. The lowest reduction in voltage deviation (0.0252) is occurred at model 12. The maximum reactive power reserve is obtained (3.00), when the function F_3 is considered as an objective function, at models 2, 4, 6, 9 and 12. While, the minimum reactive power reserve in F_3 (2.9984) is occurred at model 7. With respect to Case 3, the proposed ORPD results for different fuzzy modeling for predicted emergency operation condition are reported. The maximum reduction of the real power losses is obtained (1.6684%) at model 9. While, the real power losses isn't decreased at some cases. The maximum decreasing in the voltage deviation is obtained (0.0224), when the function F_2 is

considered as an objective function, at model 6. While, the minimum decreasing in F_2 (0.0259) is occurred at case 8. The maximum reactive power reserve is obtained (3.0), when the function F_3 is considered as an objective function, at models 1, 2, 4 and 12. The voltage deviations are minimized, and the reactive power reserve is maximized for all different fuzzy modeling cases.

Table 2 comparison between different fuzzy modeling for first operating condition (case 1)

Variables	Case 1		
	Ploss F1	VD F2	Qres F3
Initial condition	0.1808	0.0395	2.9994
Model 1.	0.1564	0.0411	1.9036
Model 2.	0.1581	0.0559	1.5417
Model 3.	0.1658	0.0292	1.8012
Model 4.	0.1671	0.0194	1.6535
Model 5.	0.1656	0.0187	1.6652
Model 6.	0.1655	0.0197	1.604
Model 7.	0.1658	0.0207	1.6041
Model 8.	0.1736	0.0242	2.9998
Model 9.	0.1695	0.0195	3
Model 10.	0.1568	0.0403	1.9202
Model 11.	0.1565	0.0409	1.9055
Model 12.	0.1568	0.0403	1.9252

Table 3 comparison between different fuzzy modeling for second operating condition (case 2)

Variables	Case 2		
	Ploss F1	VD F2	Qres F3
Initial condition	0.1808	0.0395	2.9994
Model 1.	0.1835	0.024	2.9998
Model 2.	0.1874	0.0243	3
Model 3.	0.1813	0.0232	2.9998
Model 4.	0.1829	0.0236	3
Model 5.	0.181	0.0233	2.9998
Model 6.	0.1833	0.0224	4
Model 7.	0.1811	0.0242	2.9994
Model 8.	0.181	0.0233	2.9998
Model 9.	0.1799	0.0251	3
Model 10.	0.1809	0.0249	2.9998
Model 11.	0.1813	0.0249	2.9996
Model 12.	0.1814	0.0252	3

Table 4 comparison between different fuzzy modeling for third operating condition (case 3)

Variables	Case 3		
	Ploss F1	VD F2	Qres F3
Initial condition	0.1808	0.0395	2.9994
Model 1.	0.1828	0.0238	3
Model 2.	0.1817	0.0232	3
Model 3.	0.1803	0.0236	2.9998
Model 4.	0.1815	0.0237	3
Model 5.	0.1806	0.0237	2.9996
Model 6.	0.1839	0.0224	2.9996
Model 7.	0.1848	0.0225	2.9998
Model 8.	0.1781	0.0259	2.9996
Model 9.	0.1778	0.0252	2.9994
Model 10.	0.1787	0.0248	2.9996
Model 11.	0.1838	0.0235	2.9996
Model 12.	0.1805	0.0236	3

Table 5 summarizes the best fuzzy models for each operating conditions. The following general remarks give the best fuzzy modeling for each of the control and dependent variables as:

- 1) For normal operating condition, triangular fuzzy modeling for all of control and dependent variables leads to the lowest levels of power loss reduction. While, the lowest levels of power losses is achieved by model 8 at other operating condition. The reduction of power losses is achieved when the fuzzy modeling of transmission reactive power flow is triangular model.
- 2) The second objective function has the lowest levels at models 2, 9, and 7 at normal, predicted emergency and normal & predicted emergency conditions, respectively. This objective is affected greatly by the fuzzy modeling of reactive power flow and reactive power at generation buses.
- 3) For the third objective function, the fuzzy modeling of control variables should triangular while in other operating conditions, the control variables are modeled using triangular membership functions. In the other hand, trapezoidal membership model is preferred for dependent variable in all operating conditions expect at normal operating condition.

Table 5 Summary of Best models for variant operating conditions

Case	F1	F2	F3
Case 1: Normal Operating	Model 1	Model 2	Model 5
Case 2: Predicted Emergency	Model 8	Model 9	Model 11
Case 3: Normal & Predicted emergency	Model 8	Model 7	Model 9

VII. CONCLUSION

This paper presents an efficient MFLP procedure for the management of reactive power using different fuzzy models. The proposed technique in order to minimize the real power losses with enhancing the voltage security at all buses to overcome any emergency that may occur in power system. The MFLP technique has been successfully applied to achieve multi objective functions, which are required to obtain the optimal reactive power reserve for power systems. The optimal control actions are maximized the reactive power reserves to avoid any emergency condition and to restore the system to the normal state. With the use of the MFLP technique, the best tuning of power system variables is obtained by achieving the proposed objectives. Therefore, the proposed procedure allows the system operator to solve the emergency condition problem with minimum increase of power losses. For normal operating condition, triangular fuzzy modeling for all of control and dependent variables leads to the lowest levels of power loss reduction. While, the lowest levels of power losses is achieved by model 8 at other operating condition. The reduction of power losses is achieved when the fuzzy modeling of transmission reactive power flow is triangular model. The second objective function has the lowest levels at models 2, 9, and 7 at normal, predicted emergency and normal & predicted emergency conditions, respectively. This objective is affected greatly by the fuzzy modeling of reactive power flow and reactive power at generation buses. For the third objective function, the fuzzy modeling of control variables should triangular while in other operating conditions, the control variables are modeled using triangular membership functions. In the other hand, trapezoidal membership model is preferred for dependent variable in all operating conditions expect at normal operating condition.

REFERENCES

- [1] A. A. Abou EL-Ela, M. Bishr, S. Allam, and R. El-Sehiemy, "Optimal power dispatch using different fuzzy constraints power systems," *International Energy Journal*, Volume 8, Issue 3, September 2007.
- [2] A. A. Abou EL-Ela, M. Bishr, S. Allam, and R. El-Sehiemy, "Optimal Preventive Control Actions Using Multi-Objective Fuzzy Linear Programming Technique", *Electric Power System Research Journal*, Vol. 74, Issue 1, April (2005), pp. 147-155.
- [3] A. Khorsandi, S. H. Hosseinian, A. Ghazanfari, "Modified artificial bee colony algorithm based on fuzzy multi-objective technique for optimal power flow problem," *Electr. Power Syst. Res.*, 2013, 95, pp. 206–213.
- [4] M.E. Jahromi, M. Ehsan, A. F. Meyabadi, "A dynamic fuzzy interactive approach for DG expansion planning," *Int. J. Electr. Power Energy Syst.*, 2012, 43, (1), pp. 1094–1105.
- [5] J. Talaq, "Optimal power system stabilizers for multi machine systems," *Int. J. Electr. Power Energy Syst.*, 2012, 43, (1), pp. 793–803.
- [6] A. Khodabakhshian, R.Hemmati, "Multi-machine power system stabilizer design by using cultural algorithms," *Int. J. Electr. Power Energy Syst.*, 2013, 44, (1), pp. 571–580.
- [7] V. S. Vakula, K.R. Sudha, "Design of differential evolution algorithm-based robust fuzzy logic power system stabilizer using minimum rule base," *IET. Gener. Transm. Distrib.*, 2012, 6, (2), pp. 121–132.
- [8] I. G. Sardou, M. Banejad, R.A., Hooshmand, A.Dastfan, "Modified shuffled frog leaping algorithm for optimal switch placement in distribution automation system using a multi-objective fuzzy approach," *IET. Gener. Transm. Distrib.*, 2012, 6, (6), pp. 493–502
- [9] A.A. Abou, R. El-Sehiemy, and A. M. Shaheen, "Multi-Objective Fuzzy Based Procedure for Optimal Reactive Power Dispatch Problem", *Proceedings of the 14th International Middle East Power Systems Conference (MEPCON'10)*, Cairo University, Egypt, December 19-21, (2010), Paper ID 312. pp. 941-946.
- [10] R. El Sehiemy, A. A. El- Ela, A. Shaheen, "Multi-objective fuzzy-based procedure for enhancing reactive power management," *IET Gener., Transm. & Distrib.*, 7, 12 (2013): 1453-1460.
- [11] R. El Sehiemy, A. A. El-Ela, A. Shaheen, "A Multi-Objective Fuzzy-based Procedure for Reactive Power-based Preventive Emergency Strategy," *International Journal of Engineering Research in Africa*, Germany, 13 (2015): 91-102
- [12] K.R.C. Mamundur, R.D. Chenoweth, Optimal control of reactive power flow for improvements in voltage profiles and for real power loss minimization, *IEEE Trans. Power Appar. Systems.* 100 (1981), pp. 3185–3194.
- [13] O. Gavasheli and L. A. Tuan, 'Optimal Placement of Reactive Power Supports for Transmission Loss Minimization: The Case of Georgian Regional Power Grid', *Large Engineering Systems Conference on Power Engineering Montreal, Quebec, Canada, 2007*, 10-12 October, pp. 125–130
- [14] US-Canada Power System Outage Task Force, *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations*, Issued April (2004).
- [15] H. Wu, C.W.Yu, N. Xu, X. J. Lin, "An OPF based approach for assessing the minimal reactive power support for generators in deregulated power systems", *International Journal of Electrical Power & Energy Systems* 30 (1), (2008), pp. 23–30.
- [16] R. He, G. A. Taylor, Y. H. Song "Multi-objective optimal reactive power flow including voltage security and demand profile classification", *International Journal of Electrical Power & Energy Systems* 30 (5), (2008), pp. 327–36.
- [17] L.D. Arya, L.S. Titare and D.P. Kothari "Improved particle swarm optimization applied to reactive power reserve maximization", *Electrical Power and Energy Systems* 32 (2010) pp. 368–374

Application of the orthogonal invariants of three-dimensional operators in some hydrodynamic problems and Hubble expansion law

Ilia R. Lomidze

Abstract – On the base of complete set of orthogonal invariants of operators which is built in our previously series of articles, the system of nonlinear differential equations (NDE) for orthogonal invariants of ideal gas (liquid) hydrodynamic velocity’s Jacobi matrix is obtained when the flow is barochronic. It is shown that only two regimes of barochronic flow are possible – a potential and/or a solenoidal one. Exact solutions of the NDE system are obtained; there are found polynomial relations between the Jacobi matrix’s invariants and it is proved that these relations are integrals of motion. Using the obtained results the 3-dimensional hydrodynamic Euler equations are solved and hydrodynamic velocity’s and medium density’s time and space dependence are found. It is shown that the hydrodynamic velocity of potential barochronic flow depends on radius-vector (for arbitrary choosing origin of coordinates frame) satisfies the nonrelativistic Hubble law. This result seems interesting taking into consideration that barochronic flow naturally describes long-scale evolution of the Universe. The sufficient and necessary conditions are find for the solution of hydrodynamic Euler equations of solenoidal barochronic flow having form of the primitive wave or of the double wave.

Keywords – Barochronic flow, exact solutions of hydrodynamic Euler equations, Hubble expansion law.

if

I. INTRODUCTION

In the series of articles [1-3] the classification problem of the operators in n -dimensional Euclidean space \mathbb{E}^n (and of their matrices) for orthogonal transformations by a complete set of their orthogonal polynomial invariants has been solved (in these studies the corresponding theorems for the unitary transformations group in n -dimensional unitary vector space \mathbb{U}^n have been proved too).

The method developed can be successfully used in various physics problems. The results obtained in this way usually are more correct and detailed then obtained by other authors with methods differ from our ones. Our method allows to find some new solutions [4] that were not found by methods used previously [5-7].

In present paper, the orthogonal invariants of three-dimensional matrix are used to solve three-dimensional nonlinear equation in partial derivatives, describing some hydro- and aero-dynamic problems. The similar method has been used in [5-7] where the differential equations (DE)

system was obtained for algebraic invariants of Jacoby matrix of hydrodynamic velocity field. In [5-7] authors investigate types of symmetries of general solutions of the DE system for some hydrodynamic models. But the set of algebraic invariants used in [5-7] is not complete. In [4] we have shown that the application of complete polynomial basis of orthogonal invariants gets all (smooth) solutions of corresponding physics problems in covariant form. The number of arbitrary parameters in obtained solutions can not be reduced in general.

Below we use the following main results of studies [1-3]:

- I. For arbitrary linear operator \mathbf{T} in real three-dimensional Euclidean space \mathbb{E}^3 there exists orthonormal canonical basis (CB) determined uniquely up to the simultaneous reflection of all coordinate axis, such that the corresponding matrix $T=[T_{ik}]$ (generally non-symmetric) of this operator gets the form

$$T = \begin{bmatrix} s_1 & \omega_3 & -\omega_2 \\ -\omega_3 & s_2 & \omega_1 \\ \omega_2 & -\omega_1 & s_3 \end{bmatrix}; \quad \begin{matrix} (s_1 \leq s_2 \leq s_3, \\ T_{12} \equiv \omega_3 \geq 0, \\ T_{13} \equiv -\omega_2 \geq 0) \end{matrix} \quad (1)$$

if $s_1 = s_2 < s_3$ (or if $s_1 < s_2 = s_3$) then applying the corresponding rotation $s_1 = s_2 < s_3$ in the coordinate plane $x_1 O x_2$ (or in the plane $x_2 O x_3$) that does not change the significance $T_{12} \equiv \omega_3$ ($T_{23} \equiv \omega_1$), the entry T_{13} always may be reduced to 0; if $s_1 = s_2 = s_3$ then the entries T_{13} and T_{23} may be reduced to 0.

- II. The complete set of algebraic (polynomial) orthogonal invariants of given operator \mathbf{T} has been built, which is reciprocally connected with the entries of the matrix T of this operator in the CB. In three-dimensional real Euclidean space this set in general case consists from the following six polynomial invariants ($\text{tr} M = \sum M_{kk}$):

$$\text{tr} S^\lambda, \quad (\lambda = \overline{1,3}) \quad \text{tr} A^2, \quad \text{tr}(SA^2), \quad \text{tr}(SAS^2A^2), \quad (2)$$

where

$$S_{ik} = S_{ki} = (T_{ik} + T_{ki})/2, \quad A_{ik} = -A_{ki} = (T_{ik} - T_{ki})/2.$$

It is easy to show (see [1] and [4]), that in CB one has

$$\begin{aligned} \text{tr} S^\lambda &= s_1^\lambda + s_2^\lambda + s_3^\lambda, & \lambda &= \overline{1,3}; \\ \text{tr} A^2 &= -2(\omega_1^2 + \omega_2^2 + \omega_3^2) \equiv -2\bar{\omega}^2; \\ \text{tr}(SA^2) &= s_1\omega_1^2 + s_2\omega_2^2 + s_3\omega_3^2 - \bar{\omega}^2 \text{tr} S; & (3) \\ \text{tr}(SAS^2A^2) &= \omega_1\omega_2\omega_3 \det \left[s_k^{-l-1} \right]_{k,l=\overline{1,3}} = \\ &= \omega_1\omega_2\omega_3 (s_2 - s_1)(s_3 - s_1)(s_3 - s_2). \end{aligned}$$

I. R. Lomidze is with the Georgian Technical University, Tbilisi, 0175 Georgia (phone: +995-555-484-928; e-mail: lomiltu@gmail.com).

Hence two operators are orthogonally similar iff they have coincident set of invariants (2). The method developed we use to the problem of barochronic flow of ideal gas.

II. EQUATIONS FOR BAROCHRONIC FLOW OF IDEAL GAS

Definition. The flow of continued medium is called *barochronic* if its pressure P and the density ρ depend on time only but not on the coordinates in Euclidean space \mathbb{E}^3 [5].

It is clear that three dimensional barochronic flow physically can be fulfilled only in infinite uniform space. So, the investigation of such regime is rather interesting from the cosmological point of view because it gives us the possibility to separate effects caused by gravitation (curvature of space) from the purely kinematical effects in plane infinite uniform (homogeneous) space.

It is also evident that while the gradient of pressure and density are zero the flow can happen only by inertia, and the nontrivial dependence of hydrodynamic velocity on coordinates and time is caused by the initial field of velocity, that is considered to be continuous and sufficiently smooth.

For barochronic flow of ideal gas the continuity equation [8] and the Euler equations take the form of non-linear differential equations

$$\partial_i \rho + \rho \partial_j u_j = 0, \quad \partial_i u_k + u_j \partial_j u_k = 0, \quad (k = \overline{1,3}) \quad (4)$$

for the density $\rho = \rho(t)$ and for hydrodynamic velocity vector $\vec{u} = (u_1, u_2, u_3)$; $\partial_t \equiv \partial / \partial t$, $\partial_j \equiv \partial / \partial x_j$ ($j = \overline{1,3}$). As usually, the summarization is meant by repeated indices.

Splitting the Jacoby matrix $J(t, \vec{x}) = [\partial_i u_k]_{k,i=\overline{1,3}}$ on the symmetry and skew-symmetry parts

$$J(t, \vec{x}) = S + A, \quad S = \frac{1}{2} [\partial_i u_k + \partial_k u_i]_{k,i=\overline{1,3}}, \quad A = \frac{1}{2} [\partial_i u_k - \partial_k u_i]_{k,i=\overline{1,3}},$$

we have shown in [4] that the following statements are fulfilled:

Proposition 1. The dependence of Jacoby matrix on time is determined by the equation

$$J(t, \vec{x}) = J(0, \vec{x})(E_3 + tJ(0, \vec{x}))^{-1}, \quad (5)$$

where E_3 denotes unit 3×3 matrix, and the matrix $(E_3 + tJ(0, \vec{x}))$ is not singular.

Theorem 1. All algebraic invariants of Jacoby matrix $J(t, \vec{x}) = [\partial_k u_l]$ are time dependent only and the next correlations are valid:

$$\partial_t^m \text{tr}[J(t, \vec{x})]^l = (-1)^m l(l+1) \dots (l+m-1) \text{tr}[J(t, \vec{x})]^{l+m}, \quad (6)$$

$$\begin{aligned} \partial_t \text{tr}(S^\lambda A^\zeta S^\mu A^\eta) = & \\ & -(\lambda + \eta) \text{tr}(S^{\lambda+1} A^\zeta S^\mu A^\eta) - (\mu + \zeta) \text{tr}(S^\lambda A^\zeta S^{\mu+1} A^\eta) \\ & - \lambda \text{tr}(S^{\lambda-1} A^{\zeta+2} S^\mu A^\eta) - \mu \text{tr}(S^\lambda A^\zeta S^{\mu-1} A^{\eta+2}) \\ & - \zeta \text{tr}(S^\lambda A^{\zeta-1} S A^\mu A^\eta) - \eta \text{tr}(S^\lambda A^\zeta S^\mu A^{\eta-1} S A). \end{aligned} \quad (7)$$

$$(l, m, \lambda, \mu, \zeta, \eta \in \mathbf{N})$$

Theorem 2. The elements of polynomial basis of invariants (2) satisfy the closed system of ordinary differential equations

$$\begin{aligned} (S_1 \equiv \text{tr} S = \text{tr}(S+A) = \text{div} \vec{u}(t, \vec{x}), \quad \gamma \equiv -\text{tr} A^2 = (\text{rot} \vec{u})^2 / 2, \quad f' \equiv \partial_t f): \\ (\text{tr} S)' = -\text{tr} S^2 - \text{tr} A^2; \\ (\text{tr} A^2)' = -4 \text{tr}(S A^2); \\ (\text{tr} S^2)' = -2 \text{tr} S^3 - 2 \text{tr}(S A^2); \\ \text{tr}(S A^2)' = \text{tr}(S^2 A^2) - 4 S_1 \text{tr}(S A^2) + (S_1^2 - \text{tr} S^2) \text{tr} A^2 - (\text{tr} A^2)^2 / 2; \\ \text{tr}(S^2 A^2)' = \\ -2 S_1 \text{tr}(S^2 A^2) - (2 \text{tr} S^2 + \text{tr} A^2) \text{tr}(S A^2) \\ - \text{tr} S^3 \text{tr} A^2 + S_1 \text{tr} S^2 \text{tr} A^2 + S_1 (\text{tr} A^2)^2 / 5; \\ \text{tr}(S A S^2 A^2)' = -3 S_1 \text{tr}(S A S^2 A^2). \end{aligned} \quad (8)$$

The seventh equation in (8) can be integrated directly:

$$\text{tr}(S A S^2 A^2) = C_7 \exp \left[-3 \int S_1(t) dt \right],$$

whyle the first six equations after corresponding simplifications give us the equations:

$$S_1''' + 4 S_1 S_1'' + 3(S_1')^2 + 6 S_1^2 S_1' + S_1^4 = 0, \quad (10)$$

$$\begin{aligned} \gamma''' + 6 S_1 \gamma'' + 6 \gamma'(S_1' + 2 S_1^2) \\ + 2 \gamma(S_1'' + 3(S_1')^2 + 4 S_1^3) - 4 S_1 \gamma^2 / 5 = 0. \end{aligned} \quad (11)$$

III. SOLUTION OF THE EQUATIONS

It is possible to find the general solutions of (10)-(11) by using the dependence (5). Indeed, taking into account that $\det(E_3 + tJ(0, \vec{x})) = 1 + c_1 t + c_2 t^2 + c_3 t^3 = q(t) \neq 0$ (c_1, c_2, c_3 being the coefficients of characteristic polynomial of matrix $J(0, \vec{x})$), we can find time dependence of invariants $S_1(t)$ and $\gamma(t)$:

$$S_1(t) = \text{tr} S(t, \vec{x}) = q^{-1}(t)(c_1 + 2c_2 t + 3c_3 t^2) = q'(t) / q(t), \quad (12_1)$$

$$\gamma(t) = -\text{tr} A^2(t, \vec{x}) = q^{-2}(t)(b_0 + b_1 t + b_2 t^2). \quad (12_2)$$

The formulas (12_{1,2}) give general solutions of equations (10)-(11) and contain six constants, presented in terms of space derivatives of the initial (smooth) field of hydrodynamic velocity. But the constants $c_1, c_2, c_3, b_0, b_1, b_2$ can't be chosen arbitrary, being constrained by the equations (10)-(11).

Putting (12₁) in (11), after simplifications, we obtain the equation

$$(q^2 \gamma)''' - 4(q^2 \gamma)^2 q^{-3} q' / 5 = 0.$$

Using here the formula (12₂), we obtain that the two possibilities exist only

$$\gamma(t) = 0 \Leftrightarrow b_0 = b_1 = b_2 = 0 \Leftrightarrow \text{rot} \vec{u}(t, \vec{x}) = 0 \quad (13_1)$$

and/or

$$q'(t) = 0 \Leftrightarrow c_1 = c_2 = c_3 = 0 \Leftrightarrow \text{div} \vec{u}(t, \vec{x}) = 0. \quad (13_2)$$

Thus we have proved the following important

Theorem 3. The smooth vector field describing the hydrodynamic velocity of barochronic flow of ideal gas is either potential either solenoid.

Let us discuss each of these two possibilities separately.

A. If the barochronic flow is potential i.e. if

$$\text{rot} \vec{u}(t, \vec{x}) = 0, \quad A_{lk} = (u_{l,k} - u_{k,l}) / 2 = 0,$$

$$u_{l,k} = u_{k,l} = S_{lk}; \quad \vec{u}(t, \vec{x}) = \text{grad} \varphi(t, \vec{x}),$$

then the hydrodynamic velocity has the form (t_0 is an arbitrary constant):

$$\vec{u}(t, \vec{x}) = \vec{x}(t+t_0)^{-1}, \quad (14)$$

and the time dependence of density is described by formulas

$$\rho(t) = \rho_0 |1+t/t_0|^{-3}, \quad \text{if } |t_0| = -3\rho_0/\rho_0' > 0,$$

$$\rho(t) = \rho_1 t^{-3}, \quad \rho_1 > 0, \quad \text{if } t_0 = 0.$$

B. In the case of solenoid barochronic flow from the formulas (12_{1,2}) and (13₂) we get

$$c_1 = c_2 = c_3 = 0, \quad q(t) = 1, \quad \gamma(t) = b_0 + b_1 t + b_2 t^2.$$

Hence, the matrix $J(0, \vec{x})$ has the rank $r \leq 2$. As it is shown in [4], there is fulfilled the following

Theorem 4. *If the barochronic flow of ideal gas is solenoid, then all three coefficients of characteristic polynomial of Jacoby matrix are equal to zero for arbitrary moment of time and there are possible the following cases ($r = \text{rank}[u_{i,j}(t, \vec{x})]_{i,j=1,3}$):*

- 1°. $r=0 \Leftrightarrow \text{tr}S^2 = -\text{tr}A^2 = 0 \Leftrightarrow S = A = 0$;
- 2°. $r=1 \Leftrightarrow \text{tr}S^2 = -\text{tr}A^2 > 0, \text{tr}S^3 = \text{tr}(SA^2) = 0$ (solutions are simple waves) (then it is evident that $\text{tr}(S^2A^2) = -(\text{tr}S^2)^2/2 < 0, \text{tr}(SAS^2A^2) = 0$);
- 3°. $r=2$ in all other cases.

The Theorem 4 involves the next corollaries:

Corollary 1. If barochronic flow of ideal gas is solenoid, then the Jacoby matrix $J(t, \vec{x}) = [u_{j,k}(t, \vec{x})]$ is time dependent only and is completely determined up to the real orthogonal transformation by three independent invariants of Jacoby matrix $J(0, \vec{x})$, which do not depend on the space coordinates:

$$\text{tr}S^2 = -\text{tr}A^2 = b_0, \quad 3\text{tr}(SA^2) = -\text{tr}S^3 = 3b_1/4, \quad \text{tr}(SAS^2A^2);$$

besides this, for the considering flow the invariant $\text{tr}(S^2A^2) = (b_2 - b_0^2)/2$ is constrained by the invariants of basis (2) with the inequalities $0 \leq b_1^2 \leq (2b_0/3)^3$. So, we have:

$$\begin{aligned} (4\text{tr}(SAS^2A^2))^2 &= [\text{tr}S^2(2\text{tr}(S^2A^2) - \text{tr}A^2\text{tr}S^2) - (2\text{tr}(SA^2))^2]^2 \\ &\quad - 2(2\text{tr}(S^2A^2) - \text{tr}A^2\text{tr}S^2)^3; \end{aligned}$$

$$0 \leq 6(\text{tr}S^3)^2 = 54(\text{tr}(SA^2))^2 \leq (\text{tr}S^2)^3 = -(\text{tr}A^2)^3 \geq 0. \quad (\text{tr}S = 0)$$

Corollary 2. Solution of the simple-wave-type exists iff $\text{tr}S^3 = \text{tr}(SA^2) = 0$. The solutions of the simple-/or double-wave-type can't be merged if the flow remains barochronic, because the corresponding criteria can't be fulfilled simultaneously and do not change in time.

Corollary 3. The components of hydrodynamic velocity $u_j(t, \vec{x}), (j=1,3)$, of solenoid barochronic flow in the canonical basis of the matrix $J_{jk}(0, \vec{0})$ are described by formulas

$$u_j(t, \vec{x}) = (J_{jk}(0, \vec{0}) - tJ_{jk}^2(0, \vec{0}))(x_k - u_k^0 t) + u_j^0, \quad (j=1,3) \quad (16)$$

$$\vec{u}(t, \vec{x}) = (\mathbf{J} - t\mathbf{J}^2)(\vec{x} - \vec{u}^0 t) + \vec{u}^0, \quad (16')$$

where $u_j^0 = u_j(0, \vec{0}), \langle \vec{e}_j | \mathbf{J} | \vec{e}_k \rangle = J_{jk}(0, \vec{0}) = u_{j,k}(0, \vec{0}), (j, k=1,3)$; the initial moment of time and the origin of coordinates are chosen arbitrary according to the condition of barochronicity.

Thus, the invariants (2) of Jacoby matrix allow to determine the regime of barochronic flow:

- a) The flow is potential, if $\text{tr}A^2 = (\text{rot}\vec{u})^2 = 0$; then the set of the invariants (2) contains only one independent invariant that is $S_1 = \text{tr}S$, and its initial value $S_1(0) = 3t_0^{-1}$ determines the initial values and the time dependence of all other invariants, as well as the time dependence of the hydrodynamic velocity $\vec{u}(t, \vec{x})$ and the gas density $\rho(t)$;
- b) The flow is solenoid, if $\text{tr}S = \text{div}\vec{u} = 0$; then there are at most three independent values among invariants (2) (and in the set of their initial values), which completely determine the Jacoby matrix $[u_{j,k}(0, \vec{x})]$ in its (orthonormal) CB [1-4] and together with constants $u_j^0 = u_j(0, \vec{0}) (j=1,3)$ (the initial values of velocity), allow to find the hydrodynamic velocity $\vec{u}(t, \vec{x})$. The six real constants $b_0, b_1, b_2, u_1^0, u_2^0, u_3^0$ are usually independents i.e. in general case their number can't be reduced.

IV. HUBBLE EXPANSION LAW

As the three-dimensional barochronic flow physically can be fulfilled only in the infinite homogeneous space that is considered in present paper to be three-dimensional Euclidean space \mathbb{E}^3 , we get the statement:

There exists the nonstationary (potential) solution of Euler equation for hydrodynamic velocity of ideal gas, that satisfies (formally) the well known Hubble law in its non relativistic form

$$\vec{u}(t, \vec{r}) = \vec{r} |t+t_0|^{-1} \equiv H \vec{r}. \quad (17)$$

It have to admit that "Hubble constant" H in the equation (17) is $H = |t+t_0|^{-1}$, and therefore the rate of expansion of such "barochronic Universe" has the character of uniform (potential) flow with constant expansion velocity \vec{u}_0 :

$$\vec{r}(t) = \vec{r}_0 + \vec{u}_0 t. \quad (18)$$

Thereby, it looks like rather interesting to investigate the same problem in non Euclidean curved space, taking into account relativistic effects. Such investigation is in progress.

REFERENCES

- [1] I. R. Lomidze, *Criteria of Unitary and Orthogonal Equivalence of Operators*. Bull. Acad. Sci. Georgia, **141**, №3, 1991, p. 481-483.
- [2] I. Lomidze, *On Some Generalisations of the Vandermonde Matrix and Their Relations with the Euler Beta-function*. Georgian Math. J., **v. 1** (1994), №4, p.405-417.
- [3] I. Lomidze, *Criteria of Unitary Equivalence of Hermitian Operators with Degenerate Spectrum*. Georgian Math. J., **v.3**, №2, 1996, p.141-152. <http://www.jeomj.rmi.acnet.ge/GMJ/>, **v.3**, №2, 1996, p.141-152.
- [4] I. Lomidze, J. I. Javakhishvili, *Application of Orthogonal Invariants of Operators in Solving Some Physical Problems*. JIRN Communications, P5-2007-31, Dubna, 2007.

- [5] L.V. Ovsyannikov *Isobaric gas motions*, Differential Equations, 1994, v.30, № 10, p.1656-1662.
- [6] L.V. Ovsyannikov, A. P. Chupakhin, *Regular partially invariant submodels of gas dynamics equations*, J. Nonlinear Math. Phys. v. 2, № 3/4, 1995, p.236-246.
- [7] L.V. Ovsyannikov, A. P. Chupakhin, *Regular partially invariant submodels of equations of gas dynamics* Journal of Applied Math. and Mechanics. v. 60, Issue 6, 1996, p.969–978.
- [8] L. D. Landau, E. M. Lifshitz, *Fluid Mechanics* (book), Second Edition, Vol. 6 (Course of Theor. Phys. Series), Oxford, Elsevier, 1987, ch. 1.

Features gas explosion in a cylindrical tube with a hole on the side

Iurii H. Polandov, Vitaliy A. Babankov and Sergei A. Dobrikov

Abstract – We performed numerical modeling of gas explosions in a cylindrical tube with a diameter of 200 mm and a length of 1500 mm, with closed ends and has 5 holes in the side. During explosion was opened only one of them, with passage sections of the openings diameters varied from 20 mm to 70 mm. Ignition is always carried out with the same end. It was found that when you open the 2nd hole (counting from the ignition device) and hole diameter from 50 mm to 61 mm inside the tube develop intense pressure oscillations with an amplitude of up to 15 kPa with a frequency close to the natural frequency of the internal volume (200 Hz).

Modeling technique used of Large Particle Method (LPM) describes these fluctuations, including their excitation, although the system of equations is not explicitly visible in this mechanism.

Keywords – large particles method, gas explosion, tube, hole, vibrations, singing flame Higgins.

I. INTRODUCTION

A method of self-excitation of pressure oscillations in the tube [1] is known as "singing" flame of Higgins, occurring during the combustion of a gaseous fuel in a long tube. In this case, the tube is vertical, the burner is located inside the tube in the bottom quarter, the position of the flame in the tube is steady. It is known a device Rijke, the main element of which is also a vertical tube. This device, named a simple generator

The authors express their gratitude to the Ministry of Education and Science of the Russian Federation for the financial support of research carried out in the Scientific - Educational Center "Fluid and Gas Mechanics. Physics of Combustion" at the State University - Education - Science - Production Complex on 2014- 2016 (theme №20, "Investigation of the mechanics of gas explosion in domestic premises and rationale of measures to reduce the hazard)."

Iurii H. Polandov is a supervisor of Scientific and Educational Center "Fluid and Gas Mechanics. Physics of Combustion" at the State University – Education – Science - Production Complex, 302020, Orel, Naugorskoe shosse, 29, Russia (corresponding author to provide phone: 8-910-304-37-89, fax 8-4862-41-32-95, e-mail: polandov@yandex.ru).

Vitalii A. Babankov is a research associate of Scientific and Educational Center "Fluid and Gas Mechanics. Physics of Combustion" at the State University – Education – Science - Production Complex, 302020, Orel, Naugorskoe shosse, 29, Russia, (phone: 8-4862-41-32-95, fax 8-4862-41-32-95, e-mail xenosv@mail.ru).

Sergei A. Dobrikov is trainee researcher Scientific and Educational Center "Fluid and Gas Mechanics. Physics of Combustion" at the State University – Education – Science - Production Complex, 302020, Orel, Naugorskoe shosse, 29, Russia, (phone: 8-4862-41-32-95, fax 8-4862 -41-32-95, e-mail xenosv@mail.ru).

of self-oscillation, draws energy not from the flame, but from electric spiral. There is a Helmholtz resonator [3], in which the oscillation frequency can be calculated with confidence. According to studies, described in [4], the cause of excitation of such oscillations is the existence of zones of varying viscous friction along the tube. This statement gives reason to believe that the mechanism of excitation of oscillations has thermoacoustic character.

In this article we consider the process of flame front propagation in a cylindrical tube with a gas explosion. Of course, in this case, there is a lot to do with the processes described, so quite naturally raised the question of the possibility of excitation of vibrations and explosions. However, there are serious differences in the conditions of processes that do not give a clear answer to this question. Firstly, the position of the flame front does not remain in the same place as in the above cases, but quickly moved along the tube. And secondly, the gas flows from the side surface of the hole, but not from the end of it.

Since the processes of gas-dynamic and thermal processes are described in terms of computational fluid dynamics (CFD), then the appropriate formulation of mathematical models can answer the question of the possible existence of oscillations of the gas in the tube when a gas explosion. Furthermore, in order to determine the conditions under which the oscillation can occur.

In this regard, it is unclear why the authors of [5], which used the FLACS in the CFD tool to describe a gas explosion in a cell measuring 4.6 meters x 4.6 meters x 3 meters, it was not possible to simulate the acoustic vibrations, arising in the process, although these fluctuations were mentioned in the article and it was stated on their impact on the explosion pressure. Note that the physical experiments on gas explosion in a referred chamber were conducted by other authors [6].

II. WORKING TOOLS

In drawing up the mathematical model of the process we have made the following assumptions concerning the simulated environment:

1. The initial mixture of propane-air is a homogeneous and stoichiometric;
2. The difference between the thermodynamic characteristics of the original mixture and the combustion products is negligible;

3. The gases in the physical process is inviscid and are ideal;
4. The combustion reaction occurs at the boundary of the original mixture and the combustion products.

Given the assumptions the problem is reduced to modeling the dynamics of the gas with uniform properties by using one of the methods for the unsteady multidimensional problems of fluid mechanics (CFD). The choice of a particular method is limited by arbitrary geometry of the computational domain, as well as the possibility of taking into account the availability of features in the simulated currents. As a basic system of equations to describe the dynamics of the medium was used known system of Euler equations in divergence form, closable equation of state.

Given the assumptions the problem is reduced to modeling the dynamics of the gas with uniform properties by using one of the methods for the unsteady multidimensional problems of fluid mechanics (CFD). The choice of a particular method is limited by arbitrary geometry of the computational domain, as well as the possibility of taking into account the availability of features in the simulated currents. As a basic system of equations to describe the dynamics of the medium was used known system of Euler equations in divergence form, closable equation of state.

On the domain of integration is superimposed Euler (fixed) grid of rectangular cells with sides Δx , Δy и Δz . Numerical solution of the system is carried out by large particles method, LPM, [10], which is based on the idea of Harlow «n particles" in the cell, allowing "splitting" of physical processes. But in the LPM solids replaced by a single liquid, fill the entire volume of the cell. This explains the name of the method. Method of large particles as well as other modern methods such as Godunov method [5], FLACS [6] et al., Allow us to study the gasdynamic flow without a priori information about the structure of the solution. The calculation consists of repetitive time steps. In turn, each such step includes three steps:

1. "Euler" stage, when neglected all effects associated with the movement of the fluid (mass flow through the faces of the cells is not);
2. "Lagrangian" stage, where the calculated mass flow through the cell boundaries;
3. The final stage, which determines the final flow parameters on the basis of conservation laws for each cell and the entire system as a whole.

The system includes equations that describe the process of heat and mass transfer with the environment and the process of propagation the flame. Cooling processes on the chamber walls are estimated on the basis of physical experiments carried out according to pressure drop in the explosion in a closed chamber. To calculate the flow through the open border to border pressure cell is assumed equal to the average between the pressure in the chamber and atmospheric. For the simulation of flame propagation introduced an additional

parameter "mass fraction of combustion products in the cell." Inside the cell the combustion products and the starting mixture divided by the flame front, which moves relative to the moving direction of the original gas mixture with the velocity of the laminar combustion. Take into account the dependence of the rate of flame propagation on the temperature of the initial mixture.

Ribs cells taken equal $\Delta x = \Delta y = \Delta z = 0,01$ m, the time step $\Delta t = 5 \cdot 10^{-7}$ c with that by a wide margin meets the criterion of stability Courant - Friedrichs - Lewy. Stock of taken in view of the fact that the introduction of the mechanism of flame spread has a negative effect on the stability of the account.

The calculated form the boundaries of repeated design a real camera, which is a cylinder $d = 200$ mm and $L = 1500$ mm, with muffled ends and is equipped with five holes, equally spaced along the length of the cylinder (Fig. 1). Starting positions of hole are closed, except one variable from №1 to №5. During the experiment, one hole is sealed with a piece of paper 0.1 mm thick. The diameter of the open hole varied from 20 to 70 mm. Chamber was filled with a stoichiometric mixture of propane-air gas. Ignition of the gas was produced always at one and the same place at the left end of the tube. Pressure was measured at two points located on both ends of the tube.

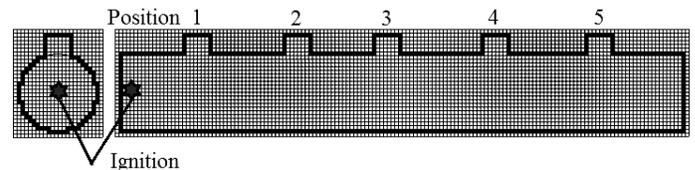


Fig. 1 - Estimated area

III. THE ADEQUACY OF THE MODEL

The adequacy of the numerical model is confirmed by comparing the results of calculations and data of physical experiments.

For this purpose, given the published data [7] for the physical gas explosion in the chamber of Fig. 1 at the position of the hole in the position 3 and 40 mm in diameter, and they are compared with calculations which are obtained for the same physical conditions of the experiment (Fig. 2). It can be seen that the physical and numerical experiments on approximately the same pressure stroke, despite the rather complicated dynamics of the process (Fig. 2a). Moreover, in both cases (Fig. 2B and 2C) were observed the pressure oscillations with amplitude about 1 kPa and a frequency of about 200 Hz, which coincides in time with the flame hit the hole. We can also note that the testimony of the first and second pressure sensors are out of phase, that is, we are dealing with a standing sound wave.

We see the adequacy of the model.

IV. 2. THE RESULTS OF THE EXPERIMENT

The results of the experiment are shown in Fig. 3. Data are shown in their absolute values. In the graph on the

ordinate postponed the value of excess pressure in the explosion, and on the horizontal axis - the number of position the open hole.

It is seen that these numerical experiments confirmed the known strong influence on the size of the hole of the explosion pressure: in this case, by increasing the hole size from 20 mm

Fig. 6 shows the pattern of the flame front in the experiment. The initial section of the flame front takes a certain shape of a tulip. Due to expansion of the area of the combustion front at this stage (up to 0.05) dramatically increases the pressure in the chamber. Then, reaching the edge of the side wall of the chamber forms the shape of an octopus. After entering the combustion products through the open hole (0.05) and

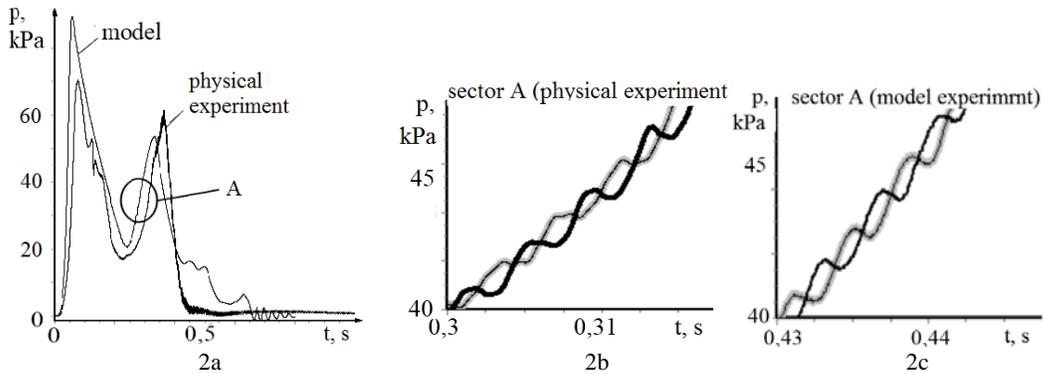


Fig. 2 - Numerical and physical explosion at the end of the hole №3 with a diameter 40 mm

to 70 mm the pressure decreases from 5 to 30 times. They also indicate that the hole effectiveness as means the protection with explosions depends from the distance between the opening and the source of ignition. Fig. 4 shows the result of translation to experimental data where the ordinate postponed the ratio of the pressure of the explosion at each position of the hole to the pressure of the explosion at the near position (№1). The graph shows the anomaly that occurs when the diameters of the hole over at least 55 mm and 61 mm, and only in the position №2.

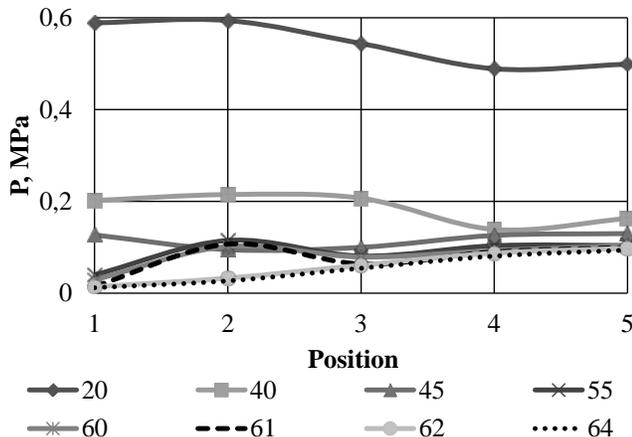


Fig. 3 - Dependence of the pressure of the explosion on the size and position of the hole

We have noticed that in these cases the camera develop intense pressure fluctuations (Fig. 5), the frequency of which varies from 150 to 210 Hz. The oscillation amplitude in this case reaches a value of 15 kPa. The maximum amplitude was at a frequency of 200 Hz.

shortening the "tentacles" pressure begins to fall sharply. At this time, begin to develop pressure fluctuations, reaching a maximum when the cross-section of the hole is fully occupied flame. After 0.1 seconds the front area increases again, which explains the increase in pressure. At the time of 0.15 with flames completely detached from the hole and vibrations begin to fade. At 0.35 with burning practically stops.

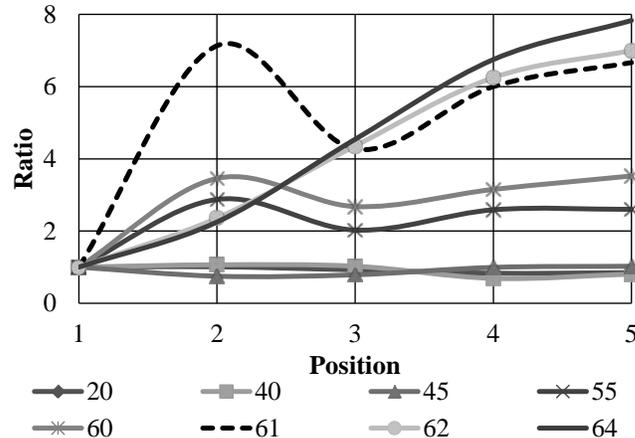


Fig. 4 - dependence of pressure of the explosion on the size and position of the hole relative units

We draw attention to two points:

- Firstly, the vibrations begin when the flame enters the hole;
- Secondly, the area of the flame front varies synchronously with the pressure variations, moreover the value of amplitude of the flame in relative units is not less than the pressure.

This suggests that the pressure fluctuations is associated with fluctuations of the area of the flame front. Or is vice versa.

In addition, we also note that the hole № 2 located at a distance from sources of ignition for about a quarter of the

length of the chamber (a little more) "singing" flame Higgins. In this case the tube becomes similar to a device that produces an effect Higgins. However, in our case, according to the scale of oscillations, we are talking about the "very loud" flame.

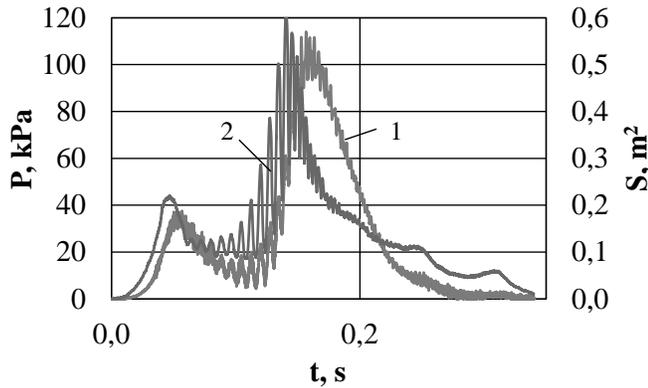


Fig. 5 - Dynamic pressure chamber. (Hole 2, $d = 60$ mm);
1 - the pressure in the chamber; 2 - the area of the combustion front

It was found that the range of parameters for which there are vibrations in the flame spread along the length of the tube, has clear boundaries. For example, if we consider the effect of the diameter of the holes, as shown in Fig. 7, with values less than the diameter of the hole 61 mm there are fluctuations, and with a diameter of 62 mm and more - there are no fluctuations.

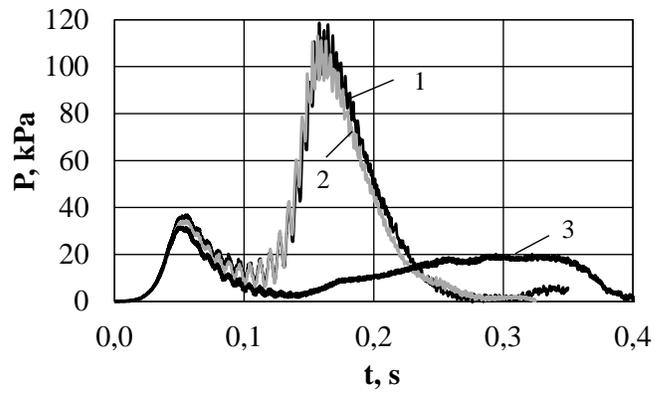


Fig. 7. The border existence of a stable "very loud" flame between the values of the hole with diameter 61 mm and 62 mm

Fluctuations occur at the coincidence of three factors:

- Flame front is in the first half of the length of the chamber;
- Tube is located near the quarter length of the chamber;
- Tube have a certain size.

The fact that we are dealing with the effect of close to "singing" flame Higgins agrees fact that in other cases, such intense vibrations no.

Of course, this approval should be checked using a physical experiment.

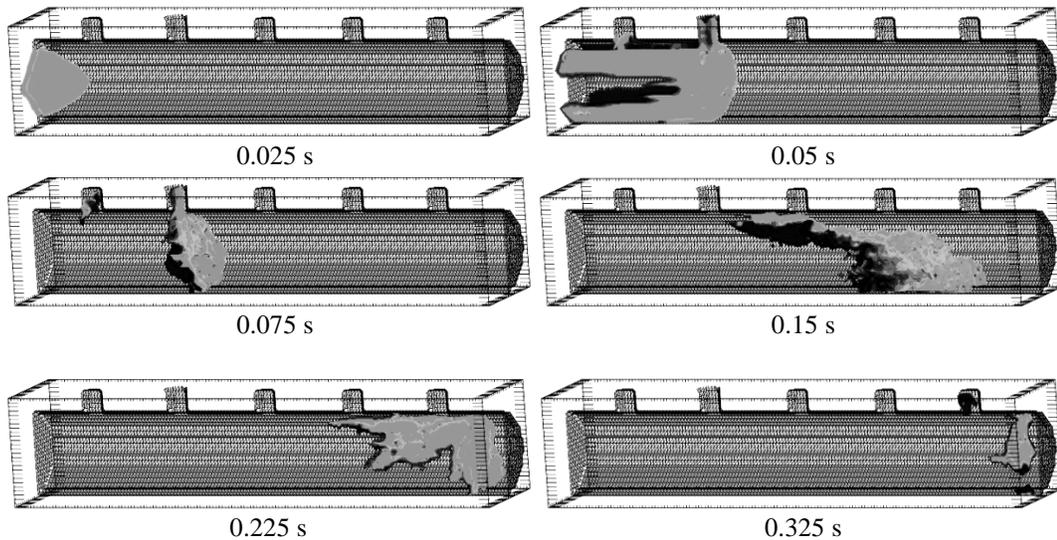


Fig. 6. The calculated position of the flame front at different times (see Fig. 5)

We observe oscillations, based on resonance effect, when the process in the hole performs the role of periodic external forces. The gas in the tube acts as an oscillation circuit with the natural frequencies. At vibrating combustion the explosion pressure increases three times as compared with pressure at ordinary combustion.

V. CONCLUSION

- Tube with closed ends and an opening on the side in the case of the explosion of gas can be "tuned" to vibrating combustion.
- Vibration combustion increases the explosion pressure in the tube.
- Approval needs to be tested on a physical experiment.

REFERENCES

- [1] Higgins B. On the sound produced by a current of hydrogen gas passing through a tube // *Journal Natural Philosophy, Chemistry and the Arts*. 1802. Vol.1. P.129.
- [2] Rijke P.L. Notiz uber eine neue Art, die in einer am beiden Enden offenen Rohre enthaltene Luft in Schwingungen zu versetzen // *Pogg. Ann. Phys. und Chem.* 1859. Vol. 107. S. 339–345.
- [3] Hermann von Helmholtz. On the sensations of tone as a physiological basis for the theory of music. Alexander John Ellis. — Longmans, Green, 1885. — 576 c.
- [4] Rauschenbach B. V. *Vibrazionnoe gorenie*. M.: Fizmatgiz, 1961. 500 s.
- [5] Godunov, S. K. (1959), A Difference Scheme for Numerical Solution of Discontinuous Solution of Hydrodynamic Equations, *Math. Sbornik*, 47, 271–306, translated US Joint Publ. Res. Service, JPRS 7226, 1969.
- [6] Helene H. Pedersen, Prankul Middha. Modelling of Vented Gas Explosions in the CFD tool FLACS. A publication of AIDIC. The Italian Association of Chemical Engineering Online at: www.aidic.it/cet.
- [7] Polandov Iu. H., Barg M.A., Babankov V.A. On effectiveness of overflow explosive valve// *Proceedings of the 6th European Congress on Computational Methods in Applied Sciences and Engineering*, September 10–14, 2012, Vienna / Austria, 2012.
- [8] C. Regis Bauwens, Jeff Chaffee and Sergey Dorofeev. Venting is used to reduce the consequences of explosions. *FM Global, Research Division, Norwood, MA, USA, 3rd ICHS, September 16-18, 2009.*

Iurii H. Polandov is a supervisor of Scientific and Educational Center “Fluid and Gas Mechanics. Physics of Combustion” at the State University – Education – Science – Production Complex, 302020, Orel, Naugorskoe shosse, 29, Russia (corresponding author to provide phone: 8-910-304-37-89, fax 8-4862-41-32-95, e-mail: polandov@yandex.ru).

He studied at the Moscow Aviation Institute (University) and graduated in 1966, there has defended candidate and doctoral dissertations. Author is a permanent member GRACM (Greek Association of Computational Mechanics), the Congress participants and Eccomas GRACM, was the scientific director of 3 the Dissertation on gas explosions, is a member of the editorial board of the journal "Safety", on the topic of gas explosions has published 31 articles in the Russian press.

The Authors express their gratitude to the Ministry of education and science of Russia on the financial support of the state task Scientific and educational center: theme №20, "Investigation of the mechanics of gas explosion in domestic premises and rationale of measures to reduce the hazard".

Double Check of Optimization Results using Neural Network and Statistical Methods

Natalja Fjodorova, Marjana Novič

Abstract—The goal of optimization is to select the best elements (objects) with regard to some criteria from some set of available alternatives. The application of optimization methods in industry is of great importance nowadays. It contributes to the increasing of quality of product and the productivity as well as reduction of energy consumption, waste and operational costs. In the study the traditional designs of experiment (DOEs) based on statistical method (response surface design) was complimented with neural network (NN) mapping method which enables to get 2D image of studied technological process (as a 2 dimensional map of properties of product) and select multiple optima. The implementation of both methods supports the double check of optimization results and expands options for selection of multiple optima. The final solution can be taken on the basis of compromise decision. Implementation of neural network mapping technique together with parametric estimation models were demonstrated for improvement of technological process of pigment dying of high performance fibers. Proposed method is simple in use and not time consuming. It can be recommended for the use in different industries for improvement of existing (ongoing) process as well as at the stage of development of new product.

Keywords—design of experiment, feed forward bottle neck neural network, neural network mapping, optimization, surface response design.

I. INTRODUCTION

The reason for the popularity of experimental design strategies and optimization methods is the competitive environment of today's marketplace in many manufacturing and service industries.

The goal of design of experiment (DOE) is to find desired factor settings so that a process average or a quality characteristic of key product properties are close to the target (on aim) and the variability is as small as possible. In chemical engineering optimization can be used to improve the production, economic and environmental performance or other criteria and simultaneously meet the specification requirements. Different algorithms can be used to solve the optimization problem. The type of relationship between input parameters and output response (linear or non-linear)

This work was supported by the Slovenian Ministry of Higher Education, Science and Technology (grant P1-017).

Natalja Fjodorova- Laboratory of Chemometrics, National Institute of Chemistry, Hajdrihova 19, SI- 1000 Ljubljana, Slovenia; (corresponding author: phone: +386 41 488 292; fax: +386 01/476 03 00; e-mail: natalja.fjodorova@ki.si)

Marjana Novič - Laboratory of Chemometrics, National Institute of Chemistry, Hajdrihova 19, SI- 1000 Ljubljana, Slovenia; (e-mail: marjana.novic@ki.si).

determines the choice of applied technique. A few examples of different approaches for optimization of different processes are represented in papers [1-4].

Some of the tools for optimization of non-linear processes using regression methods are described in the book [5]. Besides the regression methods (especially for non-linear processes) the neural network methods can be applied for solving optimization problems. Many papers [6-11] discuss the application of NNs for optimization with combination of other methods like genetic algorithm (GA).

The statistical regression method, particularly, response surface design (RSD) [12] has been applied in the study because it is widely used in industry and science. A relatively new method, namely, feed forward bottle neck neural network (FFBN NN) mapping technique for optimization was applied in this work. The basic goal of the optimization method using a neural network (NN) is to replace the model equations by an equivalent NN using mapping technique that allows one to identify multiple optima easily. A 2D map of output parameters (responses) overlapped with locations corresponding to the combinations of input parameters (setting points) enables visualization of optimal setting parameters of technological processes in the 2D map. Implementation of the FFBN NN mapping technique enables improvement of the quality of industrial products as well as findings multiple optimal solutions in the development of the new products. The application of FFBN neural network mapping technique for pigment dying of aramid and arimid fibers was published in the paper [13]. In this study we considered FFBN NN method versus RSD and compare obtained results for aramid fibers.

In our study, first, the FFBN NN was applied and several optimums were determined. Second, the RSD was performed and optimal setting parameters were found. Finally, the optimal setting parameters obtained using the FFBN NN method were checked in the regression model of RSD. The desirability of optimum parameter settings in both methods was compared and correlation between them was demonstrated. Implementation of both methods supports double check of process which is very important for reliability of settings.

II. MATERIALS AND PROCESS

A. Aramid fibers

Poly-amide benzimidazole (PABI) fibers (in Russian literature known under the trade name SVM) [14] were used in the present study. These fibers relate to the group of aramid fibers based on aromatic para-aromatic polyamide with

heterochains. The PABI fibers have extremely high modulus and strength, are heat resistant at the temperatures of 200-250°C and can be used at high operating temperature. Therefore, they are widely used in production of protective clothing (i.e. bulletproof vests) [15].

PABI fibers are generally undyeable by using classical methods. In the study we used continuous pigment dyeing process at the stage of formation of PABI fibers combined with thermo-spinning.

B. The pigment dyeing process

Pigment dyeing bath used in the study contains the following components: X1-pigment- phthalocyanine blue (highly thermostable); X2- binder latex on the bases of butadiene and vinylidene chloride in ratio (30:70); X3- anti-migration agent- manutex RS on the bases of sodium alginate; X4- dispersing agent- prevocell Wof. The process was performed at different temperatures X5. For the references see the patent [16]. The concentrations of components in dyeing bath used in the study and temperatures are represented in Table I.

TABLE I. CODED AND UNCODED VALUES OF INDEPENDENT INPUT VARIABLES AT 5 LEVELS (-2, -1, 0, 1, 2) FOR PIGMENT DYEING PROCESS OF ARAMID FIBER

Variables (factors)	Coded levels				
	-2	-1	0	+1	+2
X1- concentration of binder, %	10	15	20	25	30
X2- concentration of pigment %,	0,25	0,50	0,75	1,00	1,25
X3 - concentration of anti-migration agent, %	0,025	0,050	0,075	0,100	0,125
X4- concentration of dispersing agent, %	0,02	0,04	0,06	0,08	0,10
X5 - temperature, °C	300	350	400	450	500

The operation scheme of module for the continuous pigment dyeing of PABI fibers combined with thermo-spinning is represented in Fig. 1.

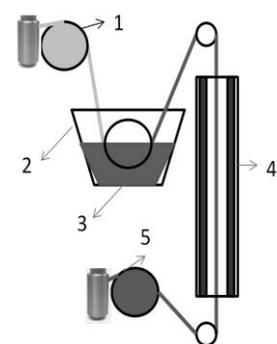


Fig. 1. The operation scheme of pigment dyeing module.

The distinctive feature of our proposed method is that first the fiber pass through the dyeing bath with pigment composition, thereafter the impregnated fibers go through the heat chamber with infrared radiation (heating) at the stage of thermo-spinning. The thermo-fixation of dye pigment composition here takes place at temperature of 350-500°C.

Note:1- let off roll with original arimid fiber; 2-pigment dye bath; 3- pigment dyeing suspension; 4- heat chamber with infrared radiation; 5-take up roller device for colored fiber.

The following three response variables (represented the quality of dyed fibers) were considered: y_1 - color strength, y_2 - tensile strength, y_3 - elongation to break.

III. METHODS

A. The feed-forward bottleneck neural network

The FFBN neural network was applied in the study (so called auto associative neural network). For the references about this technique and its application see articles [17-22]. Multidimensional data sets are difficult to interpret and visualize. The FFBN neural network was used for compression and visualization of the data in 2D maps.

The input vector in the FFBN neural network can be represented as a vector of $x_i = \{x_{i1} x_{i2} x_{i3} \dots x_{im}\}$, where “m” corresponds to the number of factors (m=5 in our model). In the FFBN each i-th object is projected onto a two dimensional map with coordinate h_{i1}/h_{i2} . In our model “i” corresponds to a number of run (from 1 to 32 in our case). See Fig. 2.

The FFBN neural network is formed by means of mapping and de-mapping the hidden layer. The signals in the two hidden nodes are taken as two coordinates for each input object, enabling a 2D projection of experimental objects onto a 2D map. In other words, the two neurons in the hidden layer produce, for each input object x_i , a corresponding pair of coordinates ($H = \{h_1, h_2\}$). Thus, in our study we obtained the 2D map with distribution of 32 experimental settings (like was determined in the plan of experiment).

For each of the 32 experimental settings the corresponding value of Y (Y_1, Y_2, Y_3) was determined in the course of the experiment. The projection of Y onto H_1/H_2 coordinate gave the contour plots of response Y (Y_1, Y_2, Y_3). Overlapping the projection of 32 experimental objects (obtained from the FFBN neural network 2D map) with responses contour plots at the same coordinates (H_1/H_2) enables visualization and determining of optimal settings corresponding to the Y optimal values.

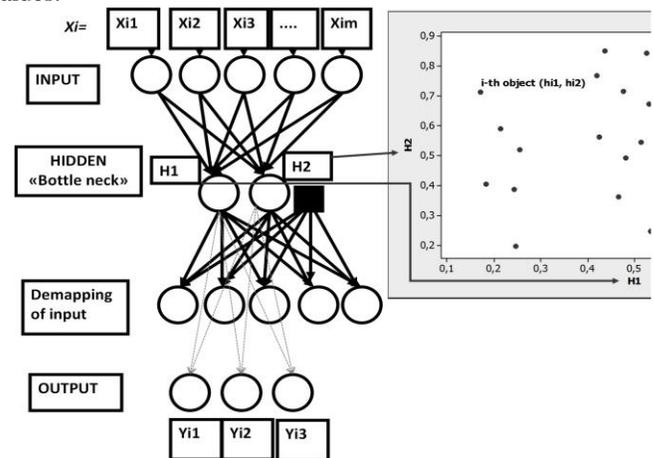


Fig. 2. The FFBN neural network mapping of i-th object.

B. The response surface methodology

Response surface methods (RSM) are used to examine the relationship between response variables (y_n) and a set of quantitative experimental factors (x_m). A general form of this

type of response function can be represented as equation:

$$y = f(x_1, x_2, \dots, x_m),$$

where y is the response and x_1, x_2, \dots, x_m are quantitative levels of factors of interest. Function f here defines the response surface.

Response Surface Methodology (RSM) is the general term for collection of statistical techniques that are useful for analysing problems influenced by several variables where the objective is to understand curvature for the purpose of optimizing the response or tolerancing the X_s . Among the techniques are: central composite designs, method of steepest ascent, evolutionary operation, simplex, and numerous others.

The goal of the study was to determine the factor levels for which the response variables (y_1 - y_3) are optimal (maximal in our case) or to find factors setting that simultaneously optimize several responses (so called generalized response Y_{n_gen}).

The generalized response Y_{n_gen} as well as individual responses y_1, y_2, y_3 can be determined for five factors x_1 - x_5 using the equations 1-4 in the course of RSD.

$$y_1 = f(x_1, x_2, \dots, x_5) + \alpha \tag{1}$$

$$y_2 = f(x_1, x_2, \dots, x_5) + \beta \tag{2}$$

$$y_3 = f(x_1, x_2, \dots, x_5) + \gamma \tag{3}$$

$$Y_{n_gen} = f(x_1, x_2, \dots, x_5) + \varepsilon \tag{4}$$

where x_1, x_2, \dots, x_5 are quantitative levels of considered factors.

The central composite design (CCD) [5, 12, 23] based on a quadratic model was used in the study.

Experimental data were analyzed using the response surface regression procedure using Minitab 15 software and fitted to a second-order polynomial model.

Minitab has a Response Optimizer that provides with an optimal solution for the input variable combinations and an optimization plot. This command in our study was based on the results of previously performed RSD.

The MINITAB's Response Optimizer identifies the combination of input variable settings that jointly optimize a set of responses. We examined Y_1 - Y_3 as well as generalized response Y_{n_gen} . Joint optimization satisfies the requirements for all the responses in the set. The desirability (D) is a measure of how well you have satisfied the goals for considered responses.

The opportunity exists to check the desirability of any settings. Therefore, in the study we calculated the desirability of optimal settings obtained using FFBN neural network method. This was done to see how data obtained in both methods (RSD and FFBN) are correlated with each other.

IV. RESULTS AND DISCUSSION

A. Plan of experimental design

Five independent variables (which affect the quality of pigment dyeing) namely concentration binder latex (x_1 , (%)), concentration of pigment (x_2 , (%)), concentration of anti-migration agent (x_3 , (%)), concentration of dispersing agent (x_4 , (%)) and temperature (x_5 , (°C)) of thermo fixation were chosen.

Each of the 5 independent variables were explored at 5 levels: -2; -1; 0; +1 and +2. The coded and uncoded values are given in Table 1. The design matrix with 32 runs (number of experiments) for Central Composite Design (CCD) was composed and represented in Table II. As the dependent variables we explored the following responses: y_1 - color strength; y_2 - tensile strength and y_3 -% elongation to break for PABI fiber.

TABLE II. THE EXPERIMENTAL PLAN OF CCD WITH FIVE INDEPENDENT VARIABLES (X1-X5) IN CODED UNITS AND VALUES OF RESPONSES Y1-Y3 IN THE EXPERIMENT WITH 32 RUNS

No run	X1-pigment	X2-binder	X3-anti-migration agent	X4-dispersing agent	X5-temperature	Y1,color strength	Y2,tensile strength	Y3,elongation to break
1	+1	+1	+1	+1	+1	4.3	102.3	3.4
2	-1	+1	+1	+1	-1	3.3	90.2	3.9
3	+1	-1	+1	+1	-1	4.3	91.6	3.8
4	-1	-1	+1	+1	+1	4.6	100.5	3.5
5	+1	+1	-1	+1	-1	3.9	92.3	3.8
6	-1	+1	-1	+1	+1	4.2	101.2	3.6
7	+1	-1	-1	+1	+1	4.7	102.7	3.5
8	-1	-1	-1	+1	-1	4.3	90.1	3.9
9	+1	+1	+1	-1	-1	3.6	91.7	3.8
10	-1	+1	+1	-1	+1	3.4	101.1	3.5
11	+1	-1	+1	-1	+1	3.8	102.5	3.4
12	-1	-1	+1	-1	-1	3.5	91.1	3.9
13	+1	+1	-1	-1	+1	3.9	102.6	3.4
14	-1	+1	-1	-1	-1	3.4	90.6	3.8
15	+1	-1	-1	-1	-1	4.9	102.3	3.5
16	-1	-1	-1	-1	+1	4.7	101.9	3.4
17	-2	0	0	0	0	3.2	98.6	4.1
18	2	0	0	0	0	3.3	99.1	4
19	0	-2	0	0	0	5	97.9	4
20	0	2	0	0	0	3	98.3	4.1
21	0	0	-2	0	0	3.5	98.5	4
22	0	0	2	0	0	3.9	98.7	4.1
23	0	0	0	-2	0	3.5	98.6	4
24	0	0	0	2	0	3.9	98.7	4.1
25	0	0	0	0	-2	3.5	90.4	3.8
26	0	0	0	0	2	4.6	105.5	3
27	0	0	0	0	0	4.7	98.8	4.1
28	0	0	0	0	0	4.8	98.7	4.1
29	0	0	0	0	0	4.7	98.7	4
30	0	0	0	0	0	4.8	98.8	4
31	0	0	0	0	0	4.8	98.7	4.1
32	0	0	0	0	0	4.7	98.8	4

B. Analysis of FFBN neural network maps

The architecture of FFBN neural network applied in our work is shown in Fig. 3.

The neural networks use vectors for treatment of information. The input data in Fig. 3 (see left upper corner) is represented as 5 vectors. Each vector (X_1 - X_5) represents an individual parameter (X_1 - X_5) at different levels (-2, -1, 0, +1, +2) as determined by 32 experimental settings.

For the 5 factors (X_1 - X_5) with independent variables (5 input parameters) a special architecture of error back-propagation neural network (5, 2, 5) was used, in which the data are fed into the 5-nodes input layer and then transferred through the 2- nodes hidden layer (so called bottleneck) to the 5-nodes output layer. The two hidden nodes of the hidden layer (bottle neck) produce two coordinates ($H=\{h_1, h_2\}$) for each input object X_i like was explained in section Methods.

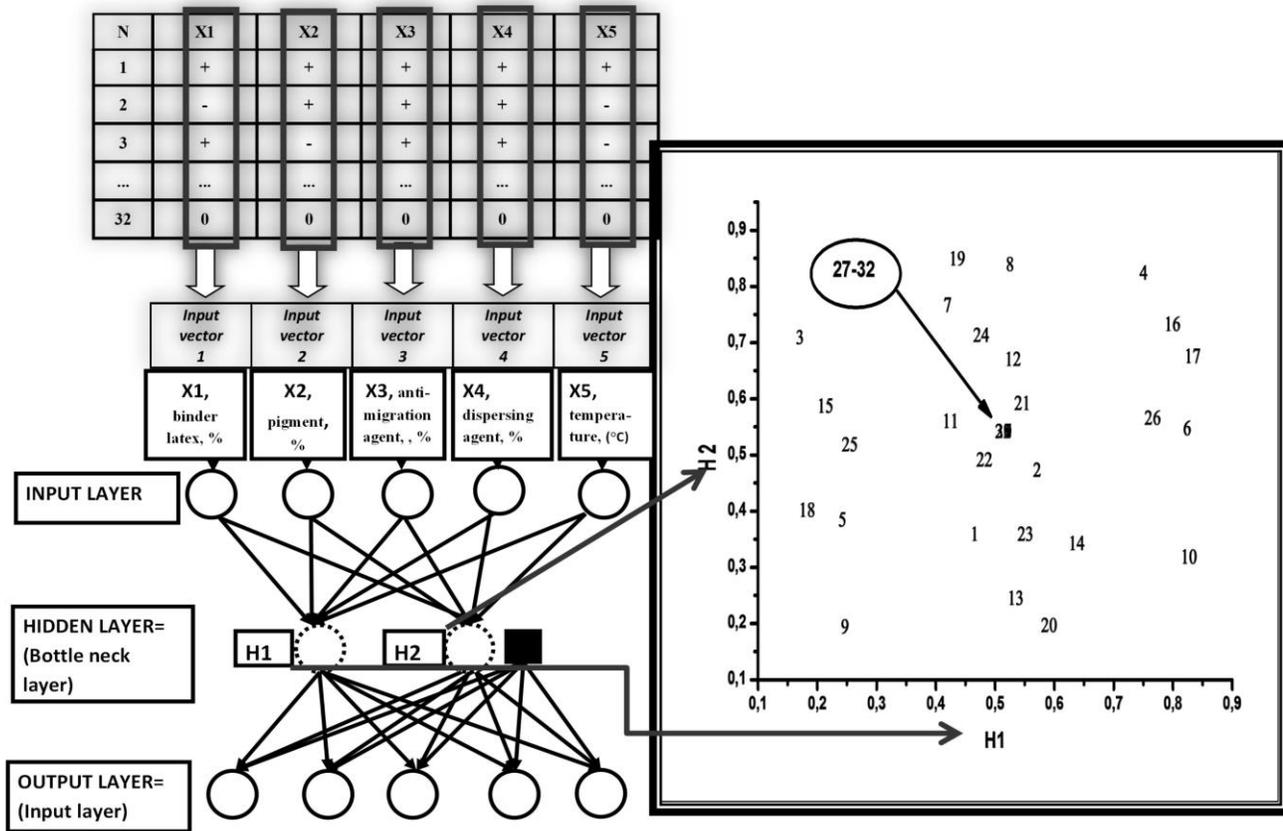


Fig. 3. The architecture of FFBN neural network applied in the study and projection of 32 objects (corresponding to the 32 experimental settings) from two hidden layers H1/H2 into 2D map.

The projection of 32 objects into the H_1/H_2 plot is shown on the right side in Fig. 3. This way the five-dimensional representation space was transformed into a 2D space (H_1/H_2) as the 32 varied data-points (32 experimental settings in design of experiment).

It should be highlighted that the distribution of the 32 setting points in the 2D map is independent from values of responses Y_1 - Y_3 , which is the intrinsic property of the considered neural network. Values of responses Y_1 - Y_3 for 32 experimental settings were measured in the course of the experiment. Then the projection of Y values onto H_1/H_2 coordinates was made and the contour plots of Y were obtained overlapped with a 2D map with the 32 setting points. See Fig. 4, where (a)-corresponds to contour plot of generalized response Y_{n_gen} , (b) relates to individual response Y_{n1} - color strength, (c)- to individual response Y_{n2} - tensile strength and (d)- to individual response Y_{n3} - elongation to break. Overlapping the projection of 32 objects with responses contour plots enables the determination of multiple optima.

A. Determination of optimums using FFBN NN mapping technique

2D projection of setting points as well as responses related to the quality of studied product enables easy determination of

several optima and understanding of the dynamic of the studied process. Take a look at Fig. 4. In the contour plots the dark grey area corresponds to maximum and the more light grey relates to the minimal values. The following optimums were investigated in the study: the central point (setting at zero level 0,0,0,0,0) corresponding to setting points 27-32 and setting points № 19, 16, 15, 7 and 4 marked with circles in Fig. 4 (a). The main goal of optimization in our study was to find the best color strength keeping in mind physical properties that should stay at the level that meets specification requirements for PABI fibers.

Fig. 4 (b, c, d) demonstrates that setting points 16, 15, 7 and 4 correspond to the best color strength while the elongation reduced sacrificing for increase of tensile strength.

Thus, the point № 19 belongs to the highest value of response (most dark grey area). The combination 19 gave the best color strength without significant reduction of mechanical properties of fibers (neither of elongation nor tensile strength). The point №19 corresponds to the following levels of parameters: $X_1=0$, $X_2=-2$, $X_3=0$, $X_4=0$, $X_5=0$. Therefore, the optimum was set for factors X_1 and X_3 - X_6 at zero level (0) and for factor X_2 at minimal level (-2) that corresponds to the minimal concentration of pigment.

A few optima exist. A final decision should be based upon a compromise, taking into account expert opinion based on understanding the nature of the studied polymers, their mechanical properties and the mechanism of action between binder, pigment and fiber during the coloring process.

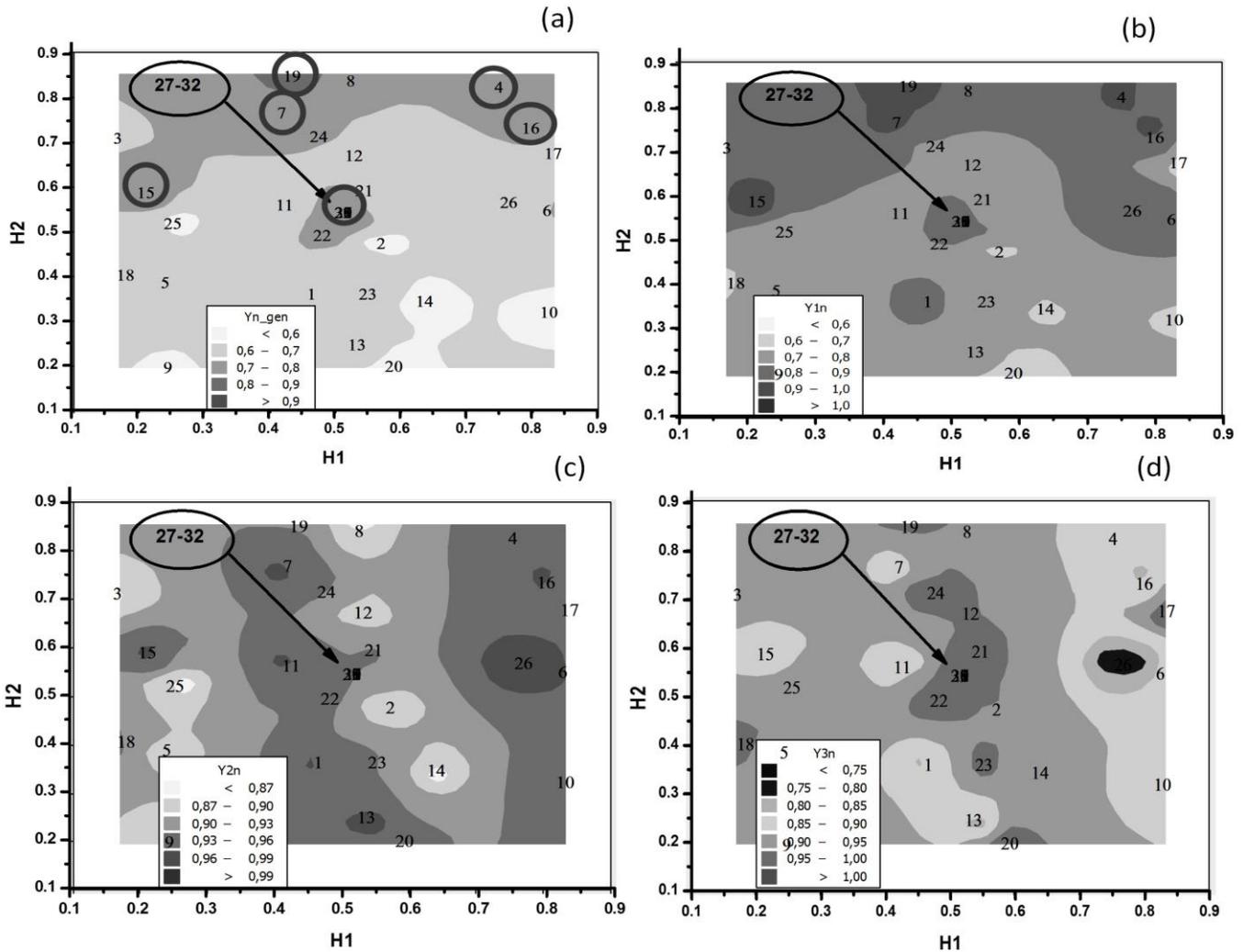


Fig. 4. The map of 32 factor settings (experimental condition) overlapped with contour plots of (a)-generalized response Y_{n_gen} and individual responses: (b)- Y_{1n} - color strength, (c)- Y_{2n} - tensile strength and (d)- Y_{3n} - elongation to break.

B. Analysis of response surface design (RSD)

The central composite design (CCD) with 5 factors, 32 total runs, 1 block, 16 cube runs and 6 total center points (replicates) was performed using the Minitab 15 software program. The simple and combined effects of five input variables (x_1 - x_5) on the quality of dyed fiber (y_1 - y_3) were determined. Minitab calculates regression coefficients for each Y (y_1, y_2, y_3) and p-values. The p-values were used as a tool to check the significance of each coefficient.

Reduced model equations (5-7) were obtained:

$$y_1 = 4,648 - 0,733x_2 - 1,091x_1^2 \quad (5)$$

$$y_2 = 98,965 + 1,858x_1 + 8,758x_5 \quad (6)$$

$$y_3 = 4,097 - 0,358x_5 - 0,836x_5^2 \quad (7)$$

The obtained results show that for color strength y_1 of PABI fiber, only the concentration of pigment x_2 and square of concentration of binder latex x_1^2 have significant influence; for tensile strength y_2 the most significant parameters are concentration binder latex (x_1) and temperature x_5 . Elongation to break y_3 appeared to be significantly dependent only on temperature x_5 .

C. Determination of optima using RSD

Determination of optima using RSD was performed using the response optimizer option in Minitab 15. We are dealing with a predictive model here. The set of values for the independent variables (X_1 - X_5) that correspond to the technological conditions of the product (for 32 setting points) are fed into the model while the response variables related to the product quality were determined in the course of experiment. The aim is to find a set of values for independent variables for which the predictive model yields the desired response. The program performs a search for response target in the independent variable space. For each selected set of independent values, the model prediction is evaluated and compared with the desired response.

In the first part of our study the RSD was used to find factor regions (parameters settings- (X_1 - X_5)) that produce the best combinations of each of individual responses Y_1 - Y_3 . The optimization plot layout is represented in Fig. 5a and shows how the factors X_1 - X_5 affect the predicted responses Y_1 - Y_3 . Minitab calculates optimal settings for the input variables along with desirability values to indicate how well those settings achieve the response targets. In Fig. 5a the composite desirability (0.87420) is fairly close to 1, which indicates the settings appear to achieve favorable results for all responses as

a whole. The most important response is color strength y_1 of PABI fiber – more important than tensile strength of PABI fiber y_2 and elongation to break of PABI fiber y_3 . The highest individual desirability equal to 1 was obtained for color strength y_1 , the desirability for tensile strength y_2 was equal to 0.70212 and desirability for elongation to break y_3 was equal to 0.83182.

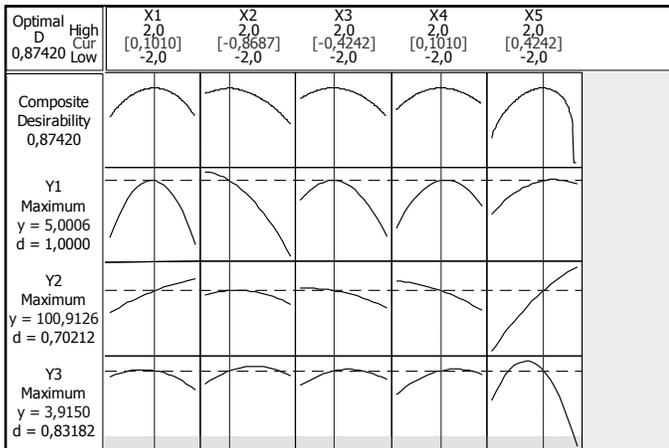


Fig. 5a. The optimization plot layout for factors X1-X5 (coded unites) for responses Y1-Y3.

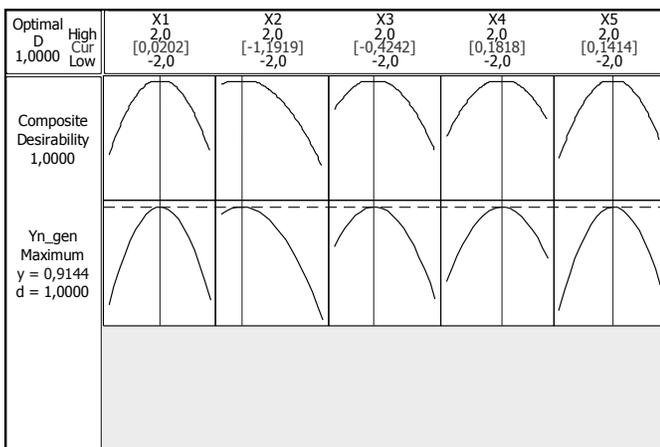


Fig. 5b. The optimization plot layout for factors X1-X5 (coded unites) for generalized response Y_{n_gen} .

In the second part of our study, we examined the generalized value of response Y. At first we divided each of Y (y_1 - y_3) on its maximal value and got y_{n1} - y_{n3} in coded units in the range 0-1. The goal of optimization was to reach the maximum for each of Y, therefore to get generalized value of Y (Y_{n_gen}) we multiply $y_{n1} * y_{n2} * y_{n3} = Y_{n_gen}$

Fig. 5b represents the optimization plot layout for factors x_1 - x_5 (coded unites) for generalized response Y_{n_gen} (coded unites). The highest desirability equal to 1 was obtained in this case.

The optimization plot is interactive; we can adjust input variable settings on the plot to search for more desirable solutions. The possibility exists to explore the desirability of settings obtained using the neural network method. We used this property to find desirability of results obtained in neural network method which is described below.

D. Comparison of optimal results in both methods (FFBN NN and RSD)

The optimization plot enables the changing of settings. We entered the optimal setting points obtained using the FFBN NN mapping method into the Minitab response optimizer. We considered the following points: **19**, **16**, **15**, **7**, **4** which correspond to the following combination of factors in coded units: **19** (0,-2, 0, 0, 0); **16** (-1,-1,-1,-1,+1), **15**(+1,-1,-1,-1,-1), **7**(+1,-1,-1,+1,+1), **4**(-1,-1,+1,+1,+1).

The desirability of different optimum settings is illustrated in Fig. 6.

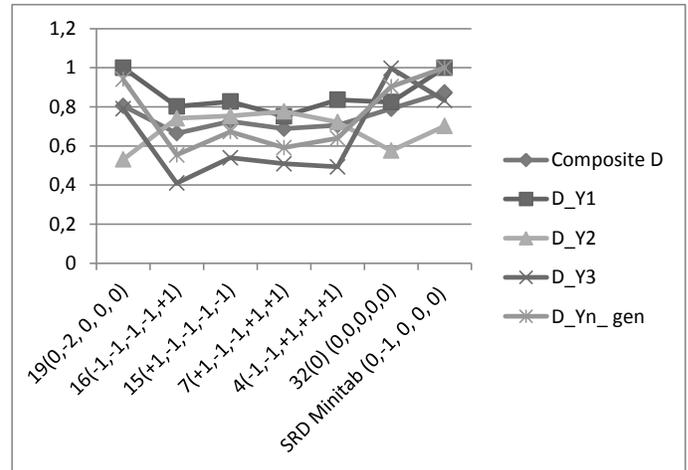


Fig. 6. Composite desirability (D) and desirability of different setting points related to studied optimums.

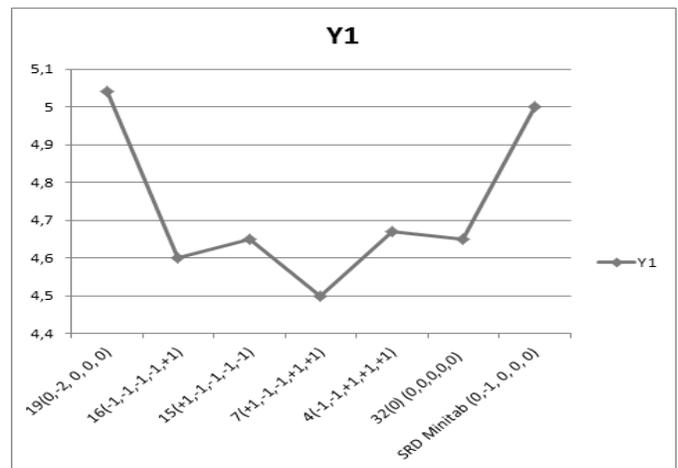


Fig. 7. The value of color strength Y1 at different optimal settings for the following responses: Y1, Y2, Y3, Y_{n_gen}

The highest desirability belongs to point **19** (0,-2, 0, 0, 0) in the case of NN and to RSD Minitab setting (0,-1, 0, 0, 0).

The result obtained using the neural network model shows slightly better color strength (5.04) than in the RSD model (5.00) reaching the highest possible desirability 1.000 in case of point **19** (0,-2,0,0,0) as well as for the result obtained using the RSD method (see Fig. 7).

V. CONCLUSION

The goal of our study was to determine optimum operation conditions: factor levels of X1-X5 that produce the maximum responses Y1-Y3: color strength, tensile strength and elongation to break) in the process of pigment dyeing of high performance PABI fibers.

The feed forward bottle neck neural network (FFBN NN) provided 2D map of whole technological process with all optimums while response surface methodology (RSM) was based on second order polynomial regression.

We demonstrated the FFBN NN network method for finding optima in technological processes in comparison with the traditional RSD method. The visible projection of response in 2D map (in the FFBN method) enables to determine more numbers of optimal solutions than RSD method.

Both methods demonstrated closed results. Therefore, their integration provides double check and finding a more reliable solution.

The colour strength obtained using the FFBN NN method appeared to be slightly better (5,04) than by using RSD (5,00).

The FFBN NN algorithm was developed at the laboratory of chemometrics at the National Institute of Chemistry Ljubljana. This algorithm is easy to use, non-time consuming and provides the ability to obtain visualization of process parameters in a 2D map. Therefore, it can be recommended for finding optimum parameters in technological processes in different industry processes as well as in the Six Sigma (improvement phase).

VI. REFERENCES

- [1] M. Pishvae, M. Rabhani, S. Torabi, "A robust optimization approach to closed-loop supply chain network design under uncertainty", *Appl. Math. Modell.*, vol. 35, issue 2, pp. 637-649, 2011.
- [2] M. Hamdy, A. Hasan, K. Siren, "Applying a multi-objective optimization approach for Design of low-emission cost-effective dwellings", *Build. Environ.*, vol. 46, issue 1, pp. 109-123, 2011.
- [3] A. Wu, J. Zhang, H. Chung, "Decoupled optimal design for power electronic circuits with adaptive migration in coevolutionary environment", *Appl. Soft Comput. J.*, vol. 11, issue 1, pp. 23-31, 2011.
- [4] J. Wang, Z. Zhai, Y. Jing, et al., "Particle swarm optimization for redundant building cooling heating and power system", *Appl. Energy*, vol. 87, issue 12, pp. 3668-3679, 2010.
- [5] L. Eriksson, E. Johansson, N. Kettaneh-Wold et al., *Design of Experiments: Principles and Applications*. 3-d ed. Umee, Sweden: UMETRICS AB, 2008.
- [6] B. Ozcelik, T. Erzurumlu, "Comparison of the warpage optimization in the plastic injection molding using ANOVA, neural network model and genetic algorithm", *J. Mater. Process. Technol.*, vol.171, issue 3, pp. 437-445, 2006.
- [7] J. Zheng, Q. Wang, P. Zhao et al., "Optimization of high-pressure die-casting process parameters using artificial neural network", *International journal, advanced manufacturing technology*, vol. 44, №7-8, pp. 667-674, 2009.
- [8] Y. Park, S. Rhee, "Process modeling and parameter optimization using neural network and genetic algorithms for aluminum laser welding automation", *The Int. J. Adv. Manuf. Technol.*, vol. 37, issue 9-10, pp. 1014-1021, 2008.
- [9] S. Changyu, W. Lixia, L. Qian, "Optimization of injection molding process parameters using combination of artificial neural network and genetic algorithm method", *J. Mater. Process. Technol.*, vol. 183 (2), pp. 412-418, 2007.
- [10] S. Sette, L. Boullart et al., "Optimizing the Fiber-to-Yarn Production Process with a Combined Neural Network/Genetic Algorithm Approach", *Text. Res. J.*, vol. 67 (2), pp. 84-92, 1997.
- [11] S. I. Faridah, A. B. Nordin. "Multi-output Hybrid GA-NN with Adaptive Mechanism". *Proceedings of the 2013 International Conference on Applied Mathematics and Computational Methods*, pages 232-237, 2013
- [12] R. Myer, D. Montgomery, *Response surface methodology*, New York: John Wiley, 2002.
- [13] N. Fjodorova, M. Novič, T. Diankova, "Optimization of pigment dyeing process of high performance fibers using feed-forward bottleneck neural networks mapping technique", *Anal Chim Acta*, vol.705, no.1-2, pp.148-154, 2011.
- [14] K. Perepelkin. "Chemical Fibers with Specific Properties for Industrial Application and Personnel Protection", *J. Ind. Text.*, vol. 31 (2), pp. 87-102, 2001.
- [15] M. Jassal, S.Ghosh, "Aramid fibres – An overview", *Indian J. Fibre Text. Res.*, vol. 27, pp. 290-306, 2002.
- [16] T. Diankova et al., Patent of Russian Federation № 2158793, *Bulletins of Inventions*, №31, MKI D, 2000.
- [17] M. Kramer, "Nonlinear principal component analysis using autoassociative neural networks", *AIChE J.*, vol. 37(2), pp. 233-243, 1991.
- [18] M. Novic, N. Groselj, "Bottle-neck type of neural network as a mapping device towards food specifications", *Anal. chim. Acta*, vol. 649, pp. 68-74, 2009.
- [19] M. Daszykowski, B. Walczak, D. Massart, "A journey into low-dimensional spaces with Autoassociative Neural Networks", *Talanta*, vol. 59 (6), pp. 1095-1105, 2003.
- [20] M. Daszykowski, B. Walczak, D. Massart, "Projection methods in chemistry", *Chemom. Intell. Lab. Syst.*, vol. 65, pp. 97-112, 2003.
- [21] D. Livingstone, G. Hesketh, D. Clayworth, "Novel methods for the display of multivariate data using neural networks", *J. Mol. Graphics*, vol. 9, pp. 115-118, 1991.
- [22] R. Kocjan, J. Zupan, "Application of a feed-forward artificial neural network as a mapping device", *Chem. Inf. Comput. Sci.*, vol. 37 (6), pp. 985-989, 1997.
- [23] G. Box, W. Hunter and J. Hunter, *Statistics for Experiments*, New York: John Wiley and Sons, Inc., pp.653, 1978.

On the use of conditional expectation estimators

Sergio Ortobelli, Tommaso Lando

Abstract— This paper discusses two different methods to estimate the conditional expectation. The kernel non-parametric regression method allows to estimate the regression function, which is a realization of the conditional expectation $E(Y|X)$. A recent alternative approach consists in estimating the conditional expectation (intended as a random variable), based on an appropriate approximation of the σ -algebra generated by X . In this paper, we propose a new procedure to estimate the distribution of the conditional expectation based on the kernel method, so that it is possible to compare the two approaches by verifying which one better estimates the true distribution of $E(Y|X)$. In particular, if we assume that the two-dimensional variable (X, Y) is normally distributed, then the true distribution of $E(Y|X)$ can be computed quite easily, and the comparison can be performed in terms of goodness-of-fit tests.

Keywords—Conditional Expectation, Kernel, Non Parametric, Regression.

I. INTRODUCTION

Within a bivariate probabilistic framework, this paper discusses different methods to estimate the conditional expected value. On the one hand, several well known methods are aimed at estimating the regression function $g(x) = E(Y|X = x)$, which represents a realization of the random variable $E(Y|X)$. In particular, the kernel non-parametric regression (see [1] and [2]) allows to estimate $E(Y|X = x)$ as a locally weighted average, based on the choice of an appropriate kernel function: the method yields consistent estimators, provided that the kernel functions and the random variable Y satisfy some conditions, described in Section II. On the other hand, an alternative methodology was recently introduced by [3] for estimating the random variable $E(Y|X)$: this method has been proved to be consistent without requiring any regularity assumption. In this paper we stress the

This paper has been supported by the Italian funds ex MURST 60% 2014, 2015 and MIUR PRIN MISURA Project, 2013–2015, and ITALY project (Italian Talented Young researchers). The research was also supported through the Czech Science Foundation (GACR) under project 13-13142S and through SP2013/3, an SGS research project of VSB-TU Ostrava, and furthermore by the European Regional Development Fund in the IT4Innovations Centre of Excellence, including the access to the supercomputing capacity, and the European Social Fund in the framework of CZ.1.07/2.3.00/20.0296 (to S.O.) and CZ.1.07/2.3.00/30.0016 (to T.L.).

S.O. Author is with University of Bergamo, via dei Caniana, 2, Bergamo, Italy; and VŠB -TU Ostrava, Sokolská třída 33, Ostrava, Czech republic; e-mail: sergio.ortobelli@unibg.it.
T.L. Author is with University of Bergamo, via dei Caniana, 2, Bergamo, Italy; and VŠB -TU Ostrava, Sokolská třída 33, Ostrava, Czech republic; e-mail: tommaso.lando@unibg.it.

difference between the two methods, that are actually aimed at different estimates (i.e. the mathematical function $g(x)$ vs. the random variable $E(Y|X)$) and therefore are not comparable. In order to compare these two different methodologies, we propose a method to estimate the distribution of $E(Y|X)$ based on the kernel non-parametric formula proposed by [1] and [2]. Then, if we know the real distribution of $E(Y|X)$ (which, for instance, can be easily computed in case of normality), then we can perform a simulation analysis, drawing a bivariate random sample from (X, Y) , and finally investigate which estimated distribution better fits to the true one.

The paper is organized as follows: in Section II we present the different methodologies and their properties; in Section III we examine a method to compare the two estimators, with assumption of normality; in Section IV we briefly illustrate the financial interpretation and possible application of the conditional expected value.

II. METHODS

In this section we describe two different procedures to evaluate the conditional expected value between two random variables. Let $X: \Omega \rightarrow \mathbb{R}$ and $Y: \Omega \rightarrow \mathbb{R}$ be integrable random variables in the probability space $(\Omega, \mathfrak{F}, P)$. Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be a random sample of independent observations from the bi-dimensional variable (X, Y) . The first procedure is aimed at estimating the conditional expectation of Y given $X = x$, which is a mathematical function of X ; the second method yields an unbiased and consistent estimator of the random variable $E(Y|X)$.

The kernel non-parametric regression

It is well known that, if we know the form of the function $g(x) = E(Y|X = x)$ (e.g. polynomial, exponential, etc.), then we can estimate the unknown parameters of $g(x)$ with several methods (e.g. least squares). In particular, if we do not know the general form of $g(x)$, except that it is a continuous and smooth function, then we can approximate it with a non-parametric method, as proposed by [1] and [2]. Thus, $g(x)$ can be estimated by:

$$\hat{g}_n(x) = \frac{\sum_{i=1}^n y_i K\left(\frac{x-x_i}{h(n)}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{h(n)}\right)}, \quad (1)$$

where $K(x)$ is a density function such that i) $K(x) < C < \infty$; ii) $\lim_{x \rightarrow \pm\infty} |xK(x)| = 0$; iii) $h(n) \rightarrow 0$ when $n \rightarrow \infty$. The function $K(x)$ is denoted by *kernel*, observe that kernel functions are generally used for estimating probability densities non-parametrically (see [4]). It was proved in [1] that if Y is quadratically integrable then $\hat{g}_n(x)$ is a consistent estimator for $g(x)$. In particular, observe that, if we denote by

$f(x, y)$ the joint density of (X, Y) , the denominator of (1) converges to the marginal density of $X \int f(x, y)dy$, while the numerator converges to the function $\int yf(x, y)dy = \int_{-\infty}^{\infty} \int_{\{X=x\}} yP(dx, dy)$ (note that, if X is continuous, the function $\int_{\{X=x\}} yP(dx, dy) / \int_{-\infty}^{\infty} \int_{\{X=x\}} P(dx, dy)$ has to be intended as a regular conditional probability).

The OLP method

We now describe an alternative non-parametric approach [3] for approximating the conditional expectation, the method is denoted by “*OLP*”, which is an acronym of the authors’ names. Define by \mathfrak{F}_X the σ -algebra generated by X (that is, $\mathfrak{F}_X = \sigma(X) = X^{-1}(\mathcal{B}) = \{X^{-1}(B) : B \in \mathcal{B}\}$, where \mathcal{B} is the Borel σ -algebra on \mathbb{R}). Observe that the regression function is just a “pointwise” realization of the random variable $E(Y|\mathfrak{F}_X)$, which can equivalently be denoted by $E(Y|X)$. The following methodology is aimed at estimating $E(Y|X)$ rather than $g(x)$.

\mathfrak{F}_X can be approximated by a σ -algebra generated by a suitable partition of Ω . In particular, for any $k \in \mathbb{N}$, we consider the partition $\{A_j\}_{j=1}^{b^k} = \{A_1, \dots, A_{b^k}\}$ of Ω in b^k subsets, where b is an integer number greater than 1 and:

- $A_1 = \left\{ \omega : X(\omega) \leq F_X^{-1} \left(\frac{1}{b^k} \right) \right\}$,
- $A_h = \left\{ \omega : F_X^{-1} \left(\frac{h-1}{b^k} \right) < X(\omega) \leq F_X^{-1} \left(\frac{h}{b^k} \right) \right\}$, for $h = 2, \dots, b^k - 1$
- $A_{b^k} = \Omega - \cup_{j=1}^{b^k-1} A_j = \left\{ \omega : X(\omega) > F_X^{-1} \left(\frac{b^k-1}{b^k} \right) \right\}$.

Starting with the trivial sigma algebra $\mathfrak{F}_0 = \{\emptyset, \Omega\}$, we can obtain a sequence of sigma algebras generated by these partitions, for different values of k ($k=1, \dots, m, \dots$). For instance, $\mathfrak{F}_1 = \sigma\{\emptyset, \Omega, A_1, \dots, A_b\}$ is the sigma algebra generated by $A_1 = \{\omega : X(\omega) \leq F_X^{-1}(1/b)\}$, $A_s = \{\omega : F_X^{-1}(\frac{s-1}{b}) < X(\omega) \leq F_X^{-1}(\frac{s}{b})\}$, $s=1, \dots, b-1$ and $A_b = \{\omega : X(\omega) > F_X^{-1}((b-1)/b)\}$. Generally:

$$\mathfrak{F}_k = \sigma \left(\{A_j\}_{j=1}^{b^k} \right), k \in \mathbb{N}. \tag{2}$$

Hence, it is possible to estimate the random variable $E(Y|\mathfrak{F}_X)$ by

$$E(Y|\mathfrak{F}_k)(\omega) = \sum_{j=1}^{b^k} \frac{1_{A_j}(\omega)}{P(A_j)} \int_{A_j} Y dP = \sum_{j=1}^{b^k} E(Y|A_j) 1_{A_j}(\omega), \tag{3}$$

where $1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$. Indeed, by definition of the conditional expectation, we can easily verify that $E(Y|\mathfrak{F}_k)$ is the unique \mathfrak{F}_k -measurable function such that, for any set $A \in \mathfrak{F}_k$, (that can be seen as a union of disjoint sets, in particular $A = \cup_{A_j \in \mathfrak{F}_k} A_j$) we obtain the equality

$$\int_A E(Y|\mathfrak{F}_k) dP = \int_A Y(\omega) dP(\omega) \tag{4}$$

It is proved in [3] that $E(Y|\mathfrak{F}_k)$ is a consistent estimator of the random variable $E(Y|X)$, that is, $\lim_{k \rightarrow \infty} E(Y|\mathfrak{F}_k) = E(Y|X)$ a.s.

The method, as defined by (3), requires only that Y is an integrable random variable. From a practical point of view, given n i.i.d. observations of Y , if we know the probability p_i corresponding to the i -th outcome y_i , we obtain:

$$E(Y|A_j) = \sum_{y_i \in A_j} y_i p_i / P(A_j). \tag{5}$$

Otherwise, we can give uniform weight to each observation, which yields the following consistent estimator of $E(Y|A_j)$:

$$\frac{1}{n_{A_j}} \sum_{y_i \in A_j} y_i, \tag{6}$$

where n_{A_j} is the number of elements of A_j . Therefore, we are always able to estimate $E(Y|\mathfrak{F}_k)$, which in turn is a consistent estimator of the conditional expected value $E(Y|X)$.

III. COMPARISON IN CASE OF NORMALITY

If we assume that X and Y are jointly normally distributed, i.e. $(X, Y) \sim N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$, we can obtain the distribution of the random variable $E(Y|X)$ quite easily. Indeed, we know that

$$g(x) = E(Y|X = x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X), \tag{7}$$

therefore, as $X \sim N(\mu_X, \sigma_X)$, we obtain that

$$E(Y|X) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (X - \mu_X) \sim N(\mu_Y, |\rho| \sigma_Y). \tag{8}$$

Of course, if we simulate data from (X, Y) and approximate $E(Y|X)$ with the estimator $E(Y|\mathfrak{F}_k)$ defined in (3), we can finally compare the true and the theoretical (estimated) distribution by performing a goodness-of-fit test. Differently, the kernel non-parametric regression method does not allow to estimate $E(Y|X)$, but only yields a consistent estimator of $g(x)$. However, assume that the random variable X' is independent from X and moreover $X =_d X'$ (that is, $X' \sim N(\mu_X, \sigma_X)$ and $\rho(X, X') = 0$): in this case we can estimate $E(Y|X')$ with

$$g_n(X') = \frac{\sum_{i=1}^n y_i K \left(\frac{X' - x_i}{h(n)} \right)}{\sum_{i=1}^n K \left(\frac{X' - x_i}{h(n)} \right)}, \tag{9}$$

and thereby we can also estimate the distribution of $E(Y|X)$, because $E(Y|X') =_d E(Y|X)$. Obviously, the estimate depends on the choice of the kernel function K . It is proved that $E(Y|\mathfrak{F}_k)$ converges almost surely to $E(Y|X)$ ($E(Y|\mathfrak{F}_k) \rightarrow_{a.s.} E(Y|X)$). Moreover, note that also $g_n(X')$ satisfies a weaker convergence property (convergence in distribution). Indeed, we have that

$$g_n(X') \rightarrow_{a.s.} E(Y|X') =_d E(Y|X), \tag{10}$$

thus we obtain that $g_n(X') \rightarrow_d E(Y|X)$.

Finally, it is possible to compare the two methods by verifying which one better estimates the distribution of $E(Y|X)$, future studies will be focused on this issue.

IV. CONDITIONAL EXPECTATION AND FINANCIAL APPLICATIONS

The conditional expectation of a random variable given another can be especially useful for financial applications. In particular, we can use conditional expectation estimators

either for ordering the investors choices as suggested by [3] or to evaluate and exercise those arbitrage opportunities when applies in the market.

We recall the classical definitions of first and second-orders stochastic dominance.

First order stochastic dominance: X *FSD* Y if and only if $F_X(t) \leq F_Y(t), \forall t \in \mathbb{R}$

or, equivalently X *FSD* Y if and only if $E(g(X)) \geq E(g(Y))$ for any increasing function g .

Second order stochastic dominance (increasing concave order): X *SSD* Y if and only if $\int_{-\infty}^t F_X(u)du \leq \int_{-\infty}^t F_Y(u)du, \forall t \in \mathbb{R}$ or, equivalently X *SSD* Y if and only if $E(g(X)) \geq E(g(Y))$ for any increasing and concave function g . Obviously X *FSD* Y implies also X *SSD* Y .

If we assume that X and Y are, for instance, two different gambles or investments, the financial interpretation of stochastic orders follows straightforward. Indeed, in this case X *FSD* Y means that X is stochastically “larger” than Y , while X *SSD* Y indicates a larger expectation of gain and generally an inferior “risk”. Those investors who prefer X to Y , provided that X *SSD* Y , are generally defined non-satiable risk averse investors. The following property characterizes the second order stochastic dominance in terms of conditional expectation.

Super-martingale property. X *SSD* Y if and only if there exist two random variables X', Y' defined on the same probability space that have the same distribution of X and Y such that:

$$E(Y'|X') \leq X' \text{ a. s.}$$

The proof of this property arises from the analysis proposed by Strassen [5] and is a well-known result of ordering theory (see also [6]-[7] and the references therein).

Thus, using the empirical evaluation of the conditional expected value, we can attempt to order the investors choices as suggested by [3]. On the other hand, using the fundamental theorem of arbitrage, we know that there exist no arbitrage opportunities in the market if there exists a risk neutral martingale measure under which the discounted price process results a martingale. So, when we assume that the filtration $\{\mathfrak{F}_s\}$ is the one generated by the price process $\{X_t\}_{t>0}$ (assumed to be a Markov process) then we get that $E(X_t|\mathfrak{F}_s) = E(X_t|X_s)$. Therefore, this property the conditional expected value estimator and the fundamental theorem of arbitrage can be used to estimate the risk neutral measure and the presence of arbitrage opportunities in the market.

V. CONCLUSION

In this paper, we deal with two methodologies for estimating the conditional expectation, studying their properties and analyzing their differences. We also propose a procedure to estimate the distribution of $E(Y|X)$ based on the kernel method. We observe that the OLP method proposed by [3] yields a consistent estimator of the random variable $E(Y|X)$, while the generalized kernel method, proposed in eq.

(9) yields a consistent estimator of the distribution function of $E(Y|X)$. In future work it will be possible to compare these two methodologies by verifying which one better estimates the distribution of $E(Y|X)$, based on simulation analysis and goodness-of-fit tests. Moreover, we recall that these estimators may have several financial applications such as ordering investors’ opportunities or identifying arbitrage opportunities.

REFERENCES

- [1] E. A. Nadaraya, “On estimating regression,” *Theory of Probability and its Applications*, vol. 9, no. 1, pp. 141-142, 1964.
- [2] G. S. Watson, “Smooth regression analysis,” *Sankhya, Series A*, vol. 26, no. 4, pp. 359-372, 1964.
- [3] S. Ortobelli, F. Petronio, T. Lando, “A portfolio return definition coherent with the investors preferences,” under revision in *IMA-Journal of Management Mathematics*.
- [4] V. A. Epanechnikov, “Non-parametric estimation of a multivariate probability density,” *Theory of Probability and its Applications*, vol. 14, no. 1, pp. 153-158, 1965.
- [5] V. Strassen, “The existence of probability measures with given marginals,” *Ann. Math. Statist.*, vol. 36, no. 2, pp. 423-439, 1965.
- [6] A. Müller, D. Stoyan, *Comparison methods for stochastic models and risks*, New York: John Wiley & Sons, 2002.
- [7] M. Shaked, G. Shanthikumar, *Stochastic orders and their applications*. New York: Academic Press Inc. Harcourt Brace & Company, 1994.

LQR control of a quadrotor helicopter

Demet Canpolat Tosun¹, Yasemin Işık², Hakan Korul³

Abstract— This paper focuses on a quadrotor model, named as Qball-X4 developed by Quanser. The quadrotor simulation model includes both linear and nonlinear X, Y, and Z position, roll/pitch and yaw dynamics. The Linear Quadratic Regulator (LQR) control technique is used to control the height, X and Y position, yaw and roll/pitch angle. The results of position control are obtained through simulations to reach desired attitudes. Various simulation parameters have been tested to demonstrate the validity of the proposed control system design and the effectiveness of the reconfigurable controller design in LQR control. Simulation results are presented for the position controls along X, Y, and Z axis, roll/pitch and yaw angles of the Qball-X4.

Keywords— Quadrotor, Qball-X4, LQR control, axis control, angle control, Matlab/Simulink

I. INTRODUCTION

Unmanned Aerial Vehicles (UAVs) has been the research subject of several recent applications. As an example of unmanned aerial vehicle systems, quadrotors are taken into account with the simple mechanical structure, being affordable and easy to fly.

In this study, the quadrotor named as Qball-X4 which is developed by Quanser is used. The Qball-X4 is a test platform suitable for a wide variety of UAV research applications. The Qball-X4 is propelled by four motors fitted with 10-inch propellers. The quadrotor is covered within a protective carbon fiber cage. The Qball-X4 ensures safe operation as well as opens the possibilities for a variety of novel applications with this proprietary design.

The Qball-X4 has onboard avionics data acquisition card (DAQ), named HiQ, and the embedded Gumstix computer to measure onboard sensors and drive the motors. Many research applications are enabled through the HiQ which has a high-resolution inertial measurement unit (IMU) and avionics input/output (I/O) card. Besides, the Qball-X4 comes with real-time control software, QuaRC. By means of the QuaRC, developers and researchers can rapidly develop and test

controllers through a Matlab/Simulink interface.

QuaRC is a rapid-prototyping and production system for real-time control that is so tightly integrated with Simulink that it is virtually transparent. QuaRC consists of a number of components that make this seamless integration possible [2]:

- **QuaRC Code Generation:** QuaRC extends the code generation capabilities of Simulink Coder by adding a new set of targets, such as a Windows target and QNX x86 target. These targets appear in the system target file browser of Simulink Coder. These targets change the source code generated by Simulink Coder to suit the particular target platform. QuaRC automatically compiles the C source code generated from the model, links with the appropriate libraries for the target platform and downloads the code to the target.
- **QuaRC External Mode Communications:** QuaRC provides an "external mode" communications module that allows the Simulink diagram to communicate with real-time code generated from the model.
- **QuaRC Target Management:** Generated code is managed on the target by an application called the QuaRC Target Manager. It is the QuaRC Target Manager that allows generated code to be seamlessly downloaded and run on the target from Simulink.

QuaRC's open-architecture structure allows user to develop powerful controls. QuaRC can target the Gumstix embedded computer. The Gumstix computer automatically generates codes and executes controllers on-board the vehicle. With this structure, users can observe sensor measurements and tune parameters in realtime from a host computer while the controller is performing on the Gumstix [1].

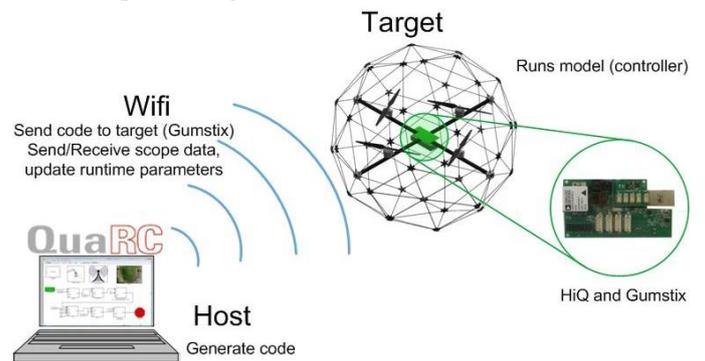


Fig. 1 Communication hierarchy [1]

The interface between the Qball-X4 and Matlab/Simulink is the QuaRC. The developed controller models in Simulink are

¹MSc. Student, Faculty of Aeronautics and Astronautics, Anadolu University 26470 Eskisehir / Turkey (e-mail: demetcanpolat@anadolu.edu.tr)

²Asst. Prof., Faculty of Aeronautics and Astronautics, Anadolu University 26470 Eskisehir/Turkey (e-mail: yaisik@anadolu.edu.tr)

³Asst. Prof., Faculty of Aeronautics and Astronautics, Anadolu University, Turkey (e-mail: hkorul@anadolu.edu.tr)

downloaded and compiled into executables on the Gumstix by the QuaRC. The configuration of the system is as shown in Fig. 1.

The required hardware and software for Qball-X4 are as follows [1]:

- Qball-X4: Qball-X4 as shown in Fig. 2,
- HiQ: QuaRC aerial vehicle data acquisition card (DAQ),
- Gumstix: The QuaRC target computer. An embedded, Linux-based system with QuaRC runtime software installed,
- Batteries: Two 3-cell, 2500 mAh Lithium-Polymer batteries,
- Real-Time Control Software: The QuaRC-Simulink configuration.

In the literature, several research studies performed in both simulations and experiments with the Qball-X4.

Some of these are as follows:

Sadeghzadeh, Mehta, Chamseddine, and Zhang proposed a Gain-Scheduled PID controller for fault-tolerant control of the Qball-X4 system in the presence of actuator faults [3].

Abdolhosseini, Zhang and Rabbath developed an efficient Model Predictive Control (eMPC) strategy and tested it on the unmanned quadrotor helicopter testbed Qball-X4 to address the main drawback of standard MPC with high computational requirement [4].

Hafez, Iskandarani, Givigi, Yousefi and Beaulieu proposed a control strategy for tactic switching, going from line abreast formation to dynamic encirclement. Their results show that applying the MPC strategy solves the problem of tactic switching for a team of UAVs (Qball-X4) in simulation [5].

Abdolhosseini, Zhang, and Rabbath have tried to design an autopilot control system for the purpose of three-dimensional trajectory tracking of the Qball-X4. Besides, they successfully implemented a constrained MPC framework on the Qball-X4 to demonstrate effectiveness and performance of the designed autopilot in addition to the simulation results [6].

Chamseddine, Zhang, Rabbath, Fulford and Apkarian worked on actuator fault-tolerant control (FTC) for Qball-X4. Their strategy is based on Model Reference Adaptive Control (MRAC). Three different MRAC techniques which are the MIT rule MRAC, the Conventional MRAC (C-MRAC) and the Modified MRAC (M-MRAC) have been implemented and compared with a Linear Quadratic Regulator (LQR) controller [7].

In this study, the LQR control technique has been used to control the three-dimensional motion of the Qball-X4

II. THE QBALL-X4 MODEL

In this section, the dynamic model of the Qball-X4 is described. Both nonlinear and linearized models are described to develop controllers.

The axes of the Qball-X4 are denoted (x, y, z) as shown in Fig. 2. The angles of the rotation about x, y, and z are roll/pitch, and yaw, respectively. The global workspace axes are denoted (X, Y, Z) and are defined with the same orientation as the Qball-X4 sitting upright on the ground.

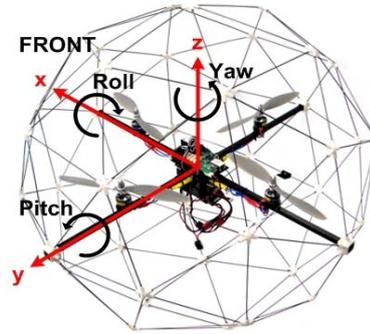


Fig. 2 Qball-X4 axes and sign convention [1]

The Qball-X4 uses brushless motors. They are mounted to the frame along the X and Y axes and to the four speed controllers which are also mounted to the frame. The motors and propellers are configured so that the front and back motors spin clockwise and the left and right motors spin counter-clockwise [1].

The relationship between the thrust (F_i) generated by i th motor and the i th PWM input (u_i) is [1]:

$$F_i = K \frac{w}{s + w} u_i \quad (1)$$

where w is the actuator bandwidth and K is a positive gain.

The calculated and verified parameters through experimental studies by Quanser are stated in Table I.

A state variable, \mathbf{v} , is defined to represent the actuator dynamics as follows:

$$\mathbf{v} = \frac{w}{s + w} \mathbf{u} \quad (2)$$

A. Height Model

The vertical motion of the Qball-X4 results from all thrusts generated by the four propellers. Therefore, the height dynamics can be written as [1]:

$$M\ddot{Z} = 4F \cos(r) \cos(p) - Mg \quad (3)$$

where F is the thrust generated by each propeller M is the mass of the quadrotor, Z is the height and r and p are the roll and pitch angles, respectively. With the assumption that the roll and pitch angles are close to zero, Eq. (3) is linearized and written in the following state space form as follows [1]:

$$\begin{bmatrix} \dot{z} \\ \ddot{z} \\ \dot{\phi} \\ \dot{s} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{4K}{M} & 0 \\ 0 & 0 & -w & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} z \\ \dot{z} \\ \phi \\ s \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} u + \begin{bmatrix} 0 \\ -g \\ 0 \\ 0 \end{bmatrix} \quad (4)$$

B. X-Y Position Model

The motion along the X and Y axes are coupled to roll and pitch motions, respectively. The motions are caused by changing roll/pitch angles. With the assumption that the yaw angle is zero, the dynamics of motion along the X and Y axes can be written as [1]:

$$M\ddot{X} = 4F \sin(p) \quad (5)$$

$$M\ddot{Y} = -4F \sin(r) \quad (6)$$

By assuming the roll and pitch angles are close to zero, linearized equations gives the following state-space models [1]:

$$\begin{bmatrix} \dot{X} \\ \ddot{X} \\ \dot{\phi} \\ \dot{s} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{4K}{M} p & 0 \\ 0 & 0 & -w & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} X \\ \dot{X} \\ \phi \\ s \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ w \\ 0 \end{bmatrix} u \quad (7)$$

$$\begin{bmatrix} \dot{Y} \\ \ddot{Y} \\ \dot{\phi} \\ \dot{s} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{4K}{M} r & 0 \\ 0 & 0 & -w & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} Y \\ \dot{Y} \\ \phi \\ s \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ w \\ 0 \end{bmatrix} u \quad (8)$$

C. Roll/Pitch Model

The roll/pitch motion is modelled as shown in Fig. 3 with the assumption that the rotations about the x and y axes are decoupled.

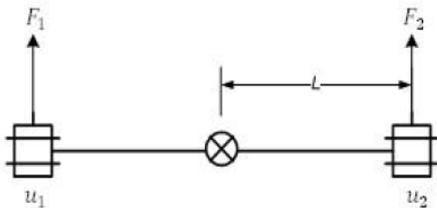


Fig. 3 The roll/pitch axis model [1]

As shown in Fig. 3, two propellers causes the motion in each axis. The difference in the generated thrusts produces the rotation around the center of gravity. The roll/pitch angle, θ , can be formulated using the following dynamics [1]:

$$J\ddot{\theta} = \Delta F L \quad (9)$$

where L is the distance between the propeller and the center of gravity, and

$$J = J_{roll} = J_{pitch} \quad (10)$$

are the rotational inertia of the device in roll and pitch axes.

The difference between the forces generated by the motors are represented as follows [1]:

$$\Delta F = F_1 - F_2 \quad (11)$$

The following state space representation can be derived from the dynamics of the motion and the actuator dynamics [1]:

$$\begin{bmatrix} \dot{\theta} \\ \ddot{\theta} \\ \dot{v} \\ \dot{s} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{KL}{J} & 0 \\ 0 & 0 & -w & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \theta \\ \dot{\theta} \\ v \\ s \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ w \\ 0 \end{bmatrix} \Delta F \quad (12)$$

A fourth state denoted as $\dot{s} = \theta$ can be defined to facilitate the use of integrator in the feedback structure and the augmented system dynamics can be rewritten as follows [1]:

$$\begin{bmatrix} \dot{\theta} \\ \ddot{\theta} \\ \dot{\phi} \\ \dot{s} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{KL}{J} & 0 \\ 0 & 0 & -w & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \theta \\ \dot{\theta} \\ \phi \\ s \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ w \\ 0 \end{bmatrix} \Delta F \quad (13)$$

D. Yaw Model

Yaw motion is caused by the difference between torques exerted by the two clockwise and the two counter-clockwise rotating propellers.

The relation between the torque, τ , generated by each propeller and the PWM input (u) is [1]:

$$\tau = K_y u \quad (14)$$

where K_y is a positive gain. Yaw motion is modeled by the following equation [1]:

$$J_y \ddot{\psi} = \Delta \tau \quad (15)$$

In this equation, J_y is the rotational inertia about the z axis, and the ψ is the yaw angle.

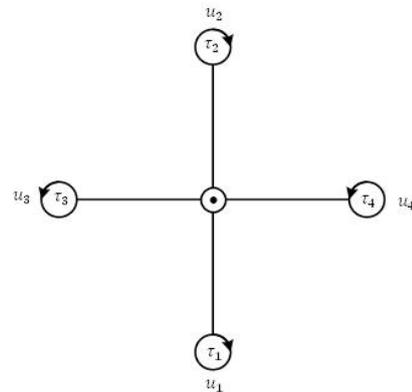


Fig. 4 The yaw axis model with propeller direction of rotation [1]

The resultant torque of the motors, $\Delta\tau$, can be calculated from

$$\Delta\tau = \tau_1 + \tau_2 - \tau_3 - \tau_4 \tag{16}$$

The yaw dynamics can be written in state-space form as follows [1]:

$$\begin{bmatrix} \dot{\psi} \\ \ddot{\psi} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \psi \\ \dot{\psi} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{K_y}{J_y} \end{bmatrix} \Delta\tau \tag{17}$$

Table I System parameters [1]

Parameter	Value
K	120 N
w	15 rad/s
I_{roll}	0.03 kg.m ²
I_{pitch}	0.03 kg.m ²
M	1.4 kg
K_y	4 N.m
J_y	0.03 kg.m ²
L	0.2m

III. LQR CONTROL

Linear quadratic regulator (LQR) is one of the most commonly used optimal control techniques for linear systems. This control method takes into account a cost function which depends on the states of the dynamical system and control input to make the optimal control decisions.

A system can be expressed in state space form as

$$\dot{x} = Ax + Bu \tag{18}$$

$$y = Cx \tag{19}$$

and suppose we want to design state feedback control

$$u = -Kx \tag{20}$$

to stabilize the system.

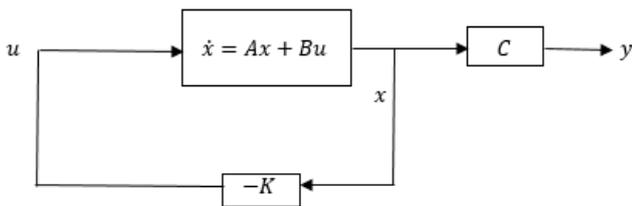


Fig. 5 LQR controller diagram

The closed-loop system using this control becomes

$$\dot{x} = (A - BK)x \tag{21}$$

The design of K is a tradeoff between the transient response and the control effort. The optimal control approach to this tradeoff is to define a cost function and search for the control, $u = -Kx$, that minimizes this cost function.

$$J = \int_0^{\infty} (x^T Q x + u^T R u) dt \tag{22}$$

where Q is an $n \times n$ positive definite matrix and R is an $n \times n$ positive definite matrix, both are symmetric.

The LQR gain vector K is given by

$$K = R^{-1} B^T P \tag{23}$$

where, P is a positive definite symmetric constant matrix obtained from the solution of matrix algebraic reccatti equation

$$A^T P + PA - PBR^{-1}B^T P + Q = 0 \tag{24}$$

The objective in optimal design is to select the K that minimizes the cost function as stated above. The cost function also known as performance index J can be interpreted as an energy function, so that making it small keeps small the total energy of the closed-loop system [8].

As seen from cost function, both the state $x(t)$ and control input $u(t)$ have weights on the total energy of the system. Therefore, if J is small, $x(t)$ and $u(t)$ can not be too large and as a control objective, if we minimize the cost function, the cost function will be an infinite integral $x(t)$. This means that $x(t)$ goes zero as t goes to infinity and guarantees the stability of the closed-loop system.

A. Height Control

For the height control model of the Qball-X4, the state matrices, A and B , obtained from the state-space form of the height model and the gain matrix K is calculated from the Q and R matrices which are chosen suitable for the system. Eventually, the height control model of the Qball-X4 is constructed through Matlab/Simulink as shown in Fig. 6.

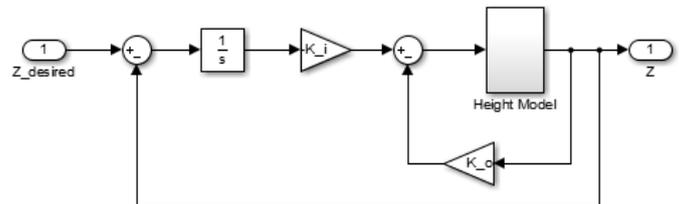


Fig. 6 Simulink model for the height control

B. X-Y Position Control

The Simulink model for X and Y position control is constructed by obtaining state matrices and the suitable weight matrices. The X and Y position control models are as shown in Fig. 7 and Fig. 8, relatively.

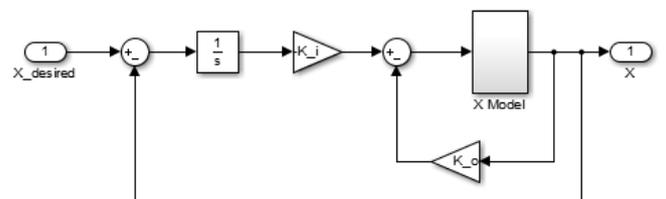


Fig. 7 Simulink model for the X position control

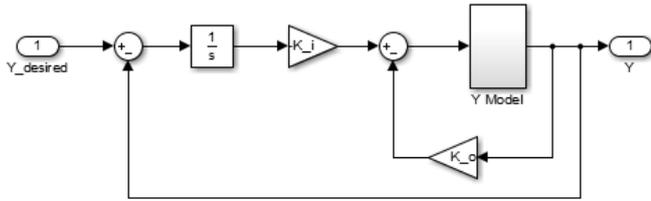


Fig. 8 Simulink model for the Y position control

C. Roll/Pitch Angle Control

In like manner, the state matrices are obtained from the state space form of the roll/pitch model and the weight matrices (Q and R) are assigned and the gain matrix K is calculated to construct the control model.

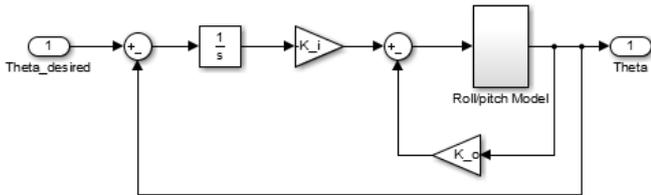


Fig. 9 Simulink model for the roll/pitch angle control

D. Yaw Angle Control

The yaw angle control of the Qball-X4 is constructed as shown in Fig. 10 by means of Simulink.

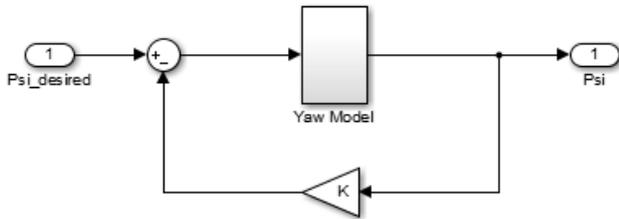


Fig. 10 Simulink model for the yaw angle control

IV. SIMULATION RESULTS

The simulation results obtained from the models shown in the previous section are shown as follows.

If we examine the X and Y positions control to reach a desired value (2m), the results shown in Fig. 12 and Fig. 13 which met our design criteria with no overshoot and a response time of approximately 6 seconds are obtained.

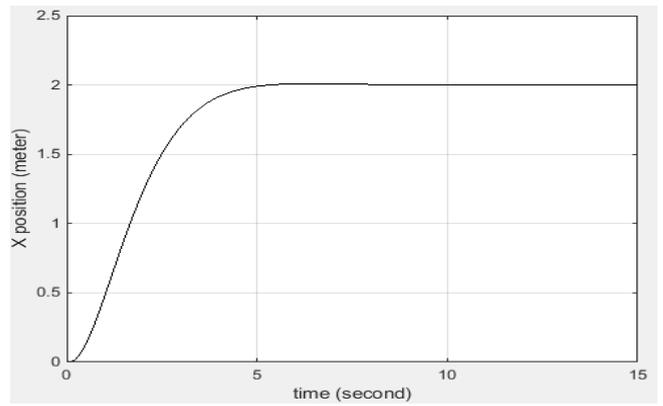


Fig. 6 The X position response

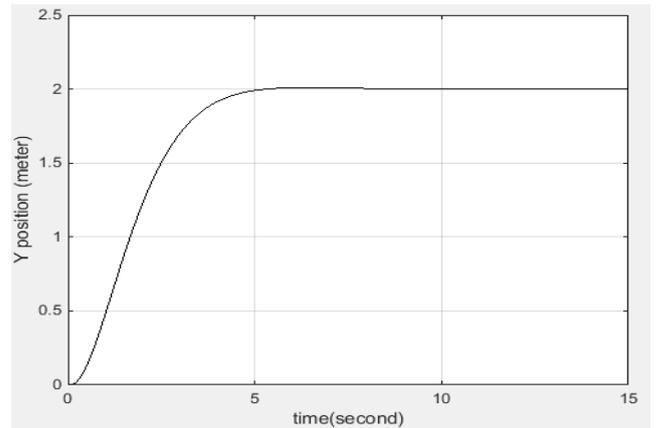


Fig. 7 The Y position response

The vertical motion control of the device is performed via the Simulink model in Fig. 6 and the simulation results are shown in the Fig. 14 which reaches the desired height (2m) in approximately 6 seconds with no overshoot.

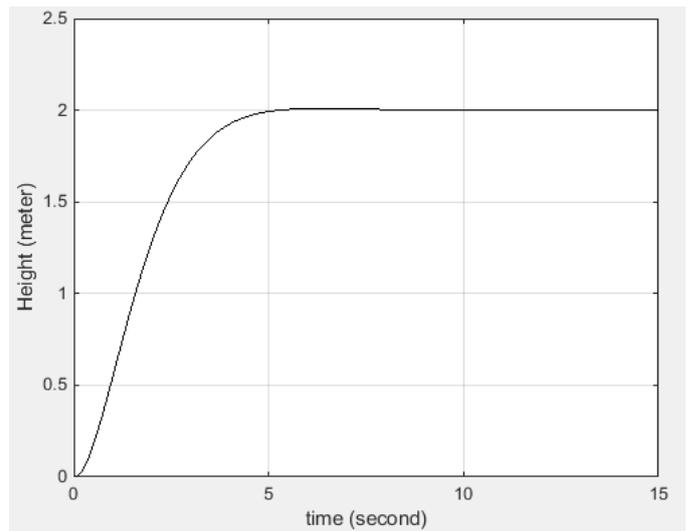


Fig. 8 The height response

The simulation results for the roll/pitch control models are shown in Fig. 15 which includes a small overshoot (%1.2) and reaches the desired value of the roll/pitch angle in approximately 0.4 seconds.

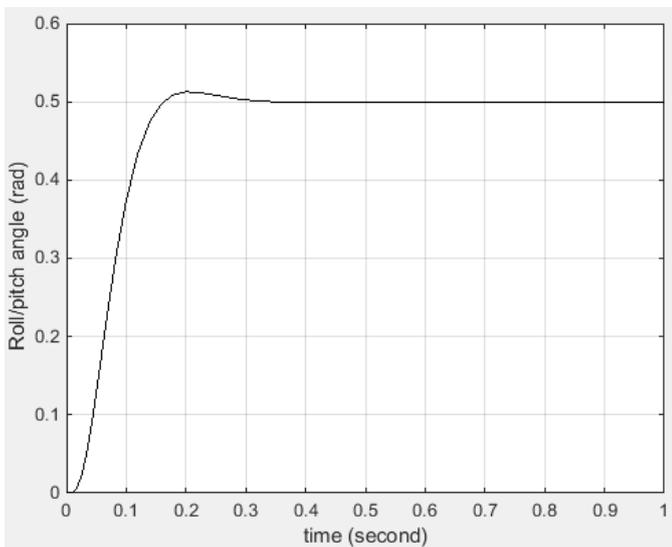


Fig. 9 The roll/pitch angle response

The simulation results which belong to the yaw angle control of the device are shown in Fig. 16. The control objective is to keep the yaw angle 0.5 radian. The Fig. 16 shows that the model reaches the desired yaw angle in 5 seconds with no overshoot.

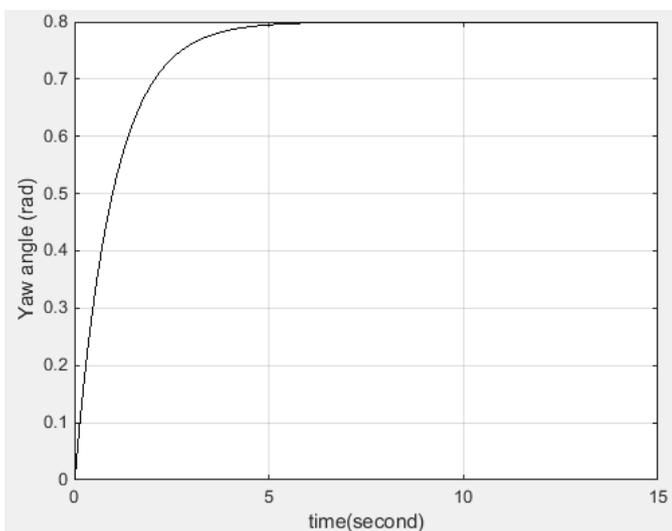


Fig. 10 The yaw angle response

The linear stability of the system is assured in simulation environment with the control gains which are designed with the weighting matrices, Q and R .

V. RESULTS AND FUTURE WORKS

In this study, the position controls along X, Y and Z axis, roll/pitch and yaw angle controls are performed in the Matlab/Simulink for the Qball-X4 quadrotor model. The LQR controllers are designed for each model. The suggested controllers are tested in simulation environment. The simulation results show that the performance specifications are met through choosing suitable weight matrices for each controller. Because of the LQR technique deals with balance between low control effort and faster response, the matrices are chosen to meet this two performance criteria.

As a conclusion, to meet the control objective, the following directions should be assured:

- Getting system dynamics as closely as possible the real system
- Calculating the control gains with choosing appropriate weighting matrices.

The future work is to test the proposed controllers experimentally on the Qball-X4 testbed with the positional data obtained from the external camera system (Optitrack camera system). Thus, the real-time performance of the proposed LQR controller would be examined. Then, performing the research applications suitable for the Qball-X4, including:

- Path planning,
- Obstacle avoidance,
- Sensor fusion,
- Fault-tolerant control, and more

will be the key subjects of the next study.

REFERENCES

- [1] Quanser Inc., "Qball-X4 user manual," Canada, 2010.
- [2] Quanser Inc., "QUARC user's guide". Available: http://www.quarcservice.com/ReleaseNotes/files/quarc_user_guide.html
- [3] I. Sadeghzadeh, A. Mehta, A. Chamseddine, and Y. Zhang, "Active fault tolerant control of a quadrotor uav based on gainscheduled pid control," Electrical & Computer Engineering (CCECE), vol. ., pp. 1-4, May 2012.
- [4] M. Abdolhosseini, Y. M. Zhang, C. A. Rabbath, "An efficient model predictive control scheme for an unmanned quadrotor helicopter," Journal of Intelligent and Robotic Systems, v.70 n.1-4, p.27-38, April 2013.
- [5] A. T. Hafez, M. Iskandarani, S. N. Givigi, S. Yousefi, A. Beaulieu, "UAVs in formation and dynamic encirclement via model predictive control," 19th IFAC World Congress, Cape Town, South Africa, August 24-29, 2014
- [6] M. Abdolhosseini, Y. M. Zhang, C. A. Rabbath, "Trajectory tracking with model predictive control for an unmanned quad-rotor helicopter: theory and flight test results," ICIRA'12, Proceedings of the 5th international conference on Intelligent Robotics and Applications, Part I, LNAI 7506, pp. 411-420, 2012.
- [7] J. A. Chamseddine, Y.M. Zhang, C.A. Rabbath, C. Fulford, J. Apkarian, "Model reference adaptive fault tolerant control of a quadrotor UAV," in: AIAA Infotech@Aerospace, St. Louis, Missouri, USA, 29-31 March 2011.
- [8] F.L. Lewis, "Linear quadratic regulator (LQR) state feedback design," [PDF document]. Retrieved from Lecture Notes Online Web site: <http://www.uta.edu/utari/acs/Lectures/lqr.pdf>.

Molecular dynamics simulations for lithographic production of carbon nanotube structures from graphene

D Fülep, I Zsoldos and I László

Abstract — In the present work we shall study the topological and energetical conditions for the growing of perfect nanotubes and their Y-junctions. For this purpose Density Functional based Tight-Binding (DFTB) Molecular Dynamics (MD) simulations were performed for producing carbon nanotubes and their Y-junction from graphene nanoribbons.

Keywords — molecular dynamics simulation, Density Functional Tight Binding method, graphene, carbon nanotube

I. INTRODUCTION

Although the exceptional electric properties of carbon nanotubes has already been proved in several publications [1], until now only very few electric devices were presented or realized [2-4]. This fact can be explained by the lack of well controlled reliable technology for nanotube or nanotube network construction. Nanotube construction from nanoribbons is a promising possibility. Nanoribbons can be produced with the help of nanolithography [5] and various chemical compounds [6]. Various ribbon structures as the L [7], the T [8] and the Z [9] structures were suggested for various nano-electric building blocks. Experimental and simulational methods are used for the study of nano electric networks [10] and functional units [11].

For the time being the accuracy of nanoribbon cutting from graphene is about few nm, and only one order of magnitude is missing to the atomic accuracy. There are studies for the instabilities at nanoribbon edges and nanotubes are obtained in molecular dynamics simulations from two nanoribbons [12]. It was demonstrated in molecular dynamics simulations that graphene patterns with atomic accuracy can develop in a self organizing way to the predetermined fullerenes or nanotubes [13-15].

The support of the grant “Developments of materials for the automotive industry” with identification number TÁMOP-4.2.2.A-11/1/KONVY-2012-0029 is gratefully acknowledged.

Dávid Fülep: Faculty of Technology Sciences, Széchenyi István University, H-9126 Győr, Hungary (fulep@sze.hu)

Ibolya Zsoldos: Faculty of Technology Sciences, Széchenyi István University, H-9126 Győr, Hungary (zsoldos@sze.hu)

István László: Department of Theoretical Physics, Institute of Physics, Budapest University of Technology and Economics, H-1521 Budapest, Hungary (laszlo@eik.bme.hu)

E-mail: fulep@sze.hu, zsoldos@sze.hu, laszlo@eik.bme.hu

The importance of nanotube production from two nanoribbons comes from the fact, that in this way open ended carbon nanotubes are developed but the one pattern nanotubes are always closed at one end [13]. Nanotube growing from nanoribbons is not a trivial task although the idea has already been published [12, 16].

In the present work we shall study the topological and energetical conditions for the growing of perfect nanotubes and their Y-junctions. For this purpose Density Functional based Tight-Binding (DFTB) Molecular Dynamics (MD) simulations were performed for producing carbon nanotubes and their Y-junction from graphene nanoribbons. The constant temperature simulations were controlled with the help of Nosé-Hoover thermostat. In our systematic study we obtained critical curvature energies and determined topological conditions for nanotube productions from one over the other put two parallel graphene nanoribbons.

II. THE METHOD

The interatomic interaction was calculated with the help of Density Functional Tight Binding method [17]. The nanoribbons were cut out from a graphene sheet of interatomic distance $r=1.42 \text{ \AA}$. After putting the two nanoribbons one over the other with parallel position, the nanotube formation was simulated in a molecular dynamics calculation with constant environmental temperature [18-19]. The time step was $\Delta t = 0.7\text{fs}$ and the Verlet algorithm [20] gave the velocity. The initial atomic displacements during the time step of $\Delta t = 0.7\text{fs}$ were sorted randomly and they gave the initial velocities by appropriate scaling. In this scaling we supposed an initial kinetic temperature T_{init} . This initial temperature was chosen from the range of $T_{\text{init}}=1000\text{K}$ and 1100 K . We have found that the final structure was depending more strongly on the direction of the initial velocities than the actual value of T_{init} . That is by scaling of the initial temperature in the above mentioned range the final structure was not strongly changing. As the formation of new bonds decreased the potential energy and increased the kinetic energy we had to keep the temperature constant. In a constant energy calculation the kinetic energy obtained by forming new bonds destroyed other bonds of the structure. We used Nosé-Hoover thermostat [18-19, 21-22] for the constant temperature simulation. It is evident that in the Nosé-Hoover thermostat there is an oscillation of the temperature but it cannot destroy the

structure formation. In the following the temperature of the calculation will mean the temperature of the thermostat. If the constant temperature were realized with the help of random scaling of the kinetic energy we could not distinguish the temperature of the environment and the structure. This is why we can speak about the T_{init} temperature and the temperature of the Nosé-Hoover thermostat (the environment temperature).

III. RESULTS

We were studying armchair and zigzag nanotubes. The initial structure contained two congruence graphene nanoribbons put one over the other at a distance of 3.4 \AA (Figure 1). We calculated the interatomic forces between the carbon atoms and we were waiting new bond formations at the edges of the ribbons. We wanted to obtain the predefined nanotube in a self organizing process. According to our simulations the formation conditions were depending on the type of the nanotube.

In the cases of straight nanotubes we found topological and energetic conditions for the perfect growing of the nanoribbons.

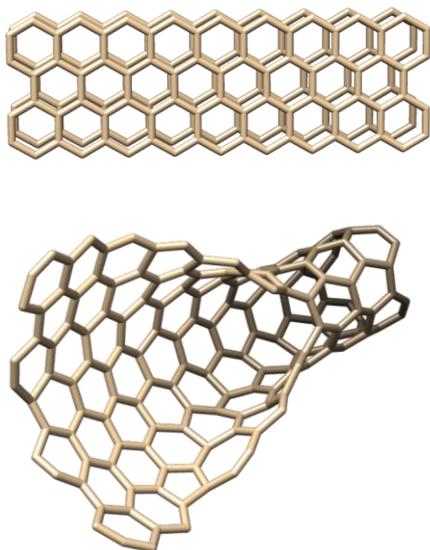


Figure 1. Simulation of armchair nanotubes. The initial (upper) and the final (lower) structures. The simulation parameters are the followings: 1000K simulation temperature, 22.13 \AA of length and 7.10 \AA of width for the parallel nanoribbons.

The basic problem of armchair nanotube formation can be seen on Figure 1. The simulation process of two congruence and parallel nanoribbons was performed at 1000 K temperature. We observed the initial growing together at both side of the ribbons but the process stopped at the established structure of the figure. At one side there is a tendency to form a graphene sheet. According to our computations there is a critical curvature energy over which the heat energy of the environment cannot produce the energy sufficient for overtake energy barrier of the bond formation. By increasing the temperature the structure could overtake this barrier but it

could destroy the other bonds as well. The correct formation of nanotube can happen only if the corresponding curvature energy is less than a critical curvature energy of 0.18 eV . This critical curvature energy corresponds to the nanotube (5,5) of radius 3.3 \AA which is obtained from the ribbons of widths 9.23 \AA as we can see in Figure 2.

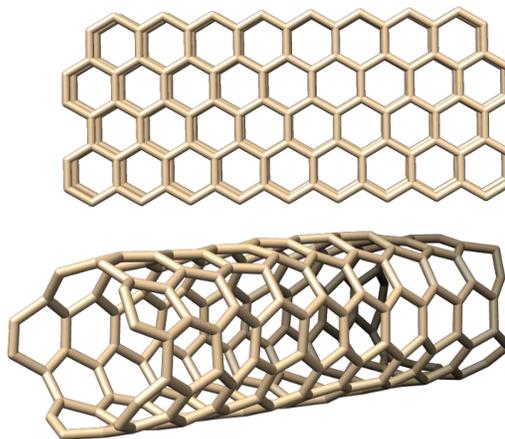


Figure 2. Simulation of armchair nanotubes. The initial (upper) and the final (lower) structures. The simulation parameters are the followings: 1000K simulation temperature, 22.13 \AA of length and 9.23 \AA of width for the parallel nanoribbons.

In the case of zigzag nanotubes the critical curvature energy is less, the critical ribbon width is greater than the same value at the armchair nanotubes. In Figure 3 the width of the two parallel ribbons of the initial model is 9.23 \AA . We can see the structures after the simulation times of 1.7 ps and 2.8 ps . There is a tendency of constructing a flat structure here, as well.



Figure 3. Simulation of zigzag nanotubes. The initial (top) and the structures after a simulation time of 1.7 ps (central) and 2.8 ps (bottom). The simulation parameters are the followings: 1000 K simulation temperature, 85.91 \AA of length and 13.01 \AA of width for the parallel nanoribbons.

The critical ribbon width of zigzag nanotubes is larger than it was in the case of the armchair nanotubes: The critical width of 15.99 Å corresponds to the nanotube (14,0) and the critical curvature energy of 0.1 eV. This case is shown in Figure 4: at 0.4 ps we observed an initial nanotube formation and at 2.8 ps we obtained a perfect zigzag nanotube.

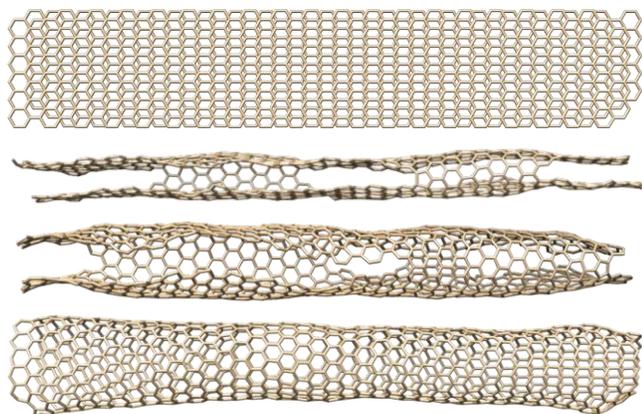


Figure 4. Simulation of zigzag nanotubes. The initial (top) and the developed structures at 0.45ps, 0.77ps and 2.8ps. The simulation parameters are the followings: 1050 K simulation temperature, 85.91 Å of length and 15.99 Å of width for the parallel nanoribbons.

Building from parallel graphene nanoribbons can give chances for controlled reliable technology in the case of more complicated carbon nanostructures, according to molecular dynamics simulations. The most important unit of the nanoelectronic networks, the carbon nanotube Y-junction can grow from parallel ribbons by self-assembled way. The example of an armchair Y-junction is shown in the followings.

In Figure 5 several initial models of the simulation can be seen:

- In Figure 5.a. the width of the ribbons is less than the critical width.
- In Figure 5.b. the width of the ribbons equals to the critical width.
- In Figure 5.c. the width of the ribbons is larger than the critical width.

The correct Y-junction structure cannot grow up from nanoribbons having width less than the critical width. During the simulation flat graphene parts appear in the structure even if some parts start to grow in tube form, as it can be seen in Figure 6.

The correct Y-junction structure can grow up from nanoribbons having width same or larger than the critical width, as it can be seen in Figure 7-8.

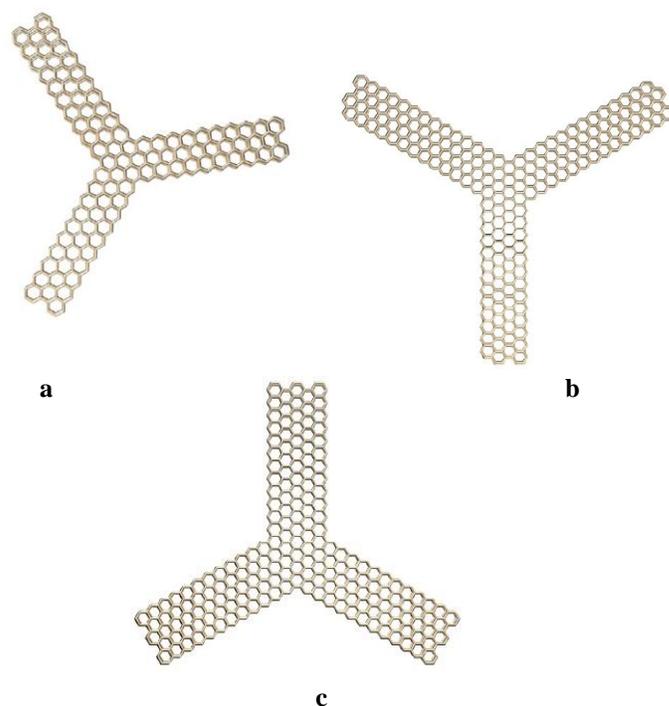


Figure 5. Initial graphene nanoribbon models for the molecular dynamics simulation.

- a:** The width of the ribbons is less than the critical width.
- b:** The width of the ribbons equals to the critical width.
- c:** The width of the ribbons is larger than the critical width.



Figure 6. Results of the molecular dynamics simulation starting from the model of Figure 5.a.



Figure 7. Results of the molecular dynamics simulation starting from the model of Figure 5.b.



Figure 8. Results of the molecular dynamics simulation starting from the model of Figure 5.c.

IV. CONCLUSION

From our molecular dynamics simulations we obtained the following conditions for straight nanotube formation from two parallel nanoribbons put one over the other:

- For armchair nanotubes the critical ribbon width is 9.23 Å corresponding to the critical curvature energy of 0.18eV.
- For zigzag nanotubes we obtained the critical ribbon width of 15.99 Å and the corresponding critical curvature energy of 0.1eV.

In the case of more complicated carbon nanostructures there is possibility for the self-assembly growing from graphene nanoribbons, which was shown on the example of an armchair carbon nanotube Y-junction.

REFERENCES

- [1] Avouris P 2002 Molecular electronics with carbon nanotubes *Accounts Chem. Res.* 35 1026-1034
- [2] Tans S J, Verschueren A R M, Dekker C 1998 Room-temperature transistor based on a single carbon nanotube *Nature* 393 49-52
- [3] Yao Z, Postma H W C, Balents L, Dekker C 1999 Carbon nanotube intramolecular junctions *Nature* 402 273-276
- [4] Keren K, Berman R S, Buchstab E, Sivan U, Braun E 2003 DNA-templated carbon nanotube field-effect transistor *Science* 302 1380-1382
- [5] Tapasztó L, Dobrik G, Lambin P, Biro L P 2008 Tailoring the atomic structure of graphene nanoribbons by scanning tunnelling microscope lithography *Nature Nanotechnology* 3 397-401
- [6] Nemes-Incze P, Magda G, Kamarás K, Biró L P 2010 Crystallographically selective nanopatterning of graphene on SiO₂ *Nano Research* 3 110-116

- [7] Xie Y E, Chen Y P, Sun L Z, Zhang K W, Zhong J 2009 The effect of corner form on electron transport of L-shaped graphene nanoribbons *Physica B-Condensed Matter* 404 1771-1775
- [8] Kong X L, Xiong Y J 2010 Resonance Transport of Graphene Nanoribbon T-Shaped Junctions *Chinese Phys. Letters* 27 047202
- [9] Wang Z F, Li Q X, Shi Q W, Wang X P, Hou J G, Zheng H X, Yao Y, Chen J 2008 Ballistic rectification in a Z-shaped graphene nanoribbon junction *Applied Physics Letter* 92 133119
- [10] Areshkin D A, White C T 2007 Building Blocks for Integrated Graphene Circuits *Nanoletters* 7 3253-3259
- [11] Areshkin D A, Nikolic B K 2009 I-V curve signatures of nonequilibrium-driven band gap collapse in magnetically ordered zigzag graphene nanoribbon two-terminal devices *Phys. Review B* 79 205430
- [12] Han S S, Lee K S, Lee H M 2004 Nucleation mechanism of carbon nanotube *Chemical Physics Letters* 383 321-325
- [13] László I, Zsoldos I 2012 Graphene-based molecular dynamics nanolithography of fullerenes, nanotubes and other carbon structures *Europhysics Letters* 99 63001
- [14] László I, Zsoldos I 2012 Molecular dynamics simulation of carbon nanostructures: The C₆₀ buckminsterfullerene *Phys. Status Solidi B* 249 2616-2619
- [15] László I, Zsoldos I 2014 Molecular dynamics simulation of carbon nanostructures: The D_{5h} C₇₀ fullerene *Physica E* 56 427-430
- [16] He L, Lu J Q, Jiang H 2009 Controlled Carbon-Nanotube Junctions Self-Assembled from Graphene Nanoribbons *Small* 5 2802-2806
- [17] Allen M P, Tildesley D J 1996 *Computer Simulation of Liquids* (Clarendon Press, Oxford)
- [18] Frenkel D, Smit B 1996 *Understanding Molecular Simulation* (Academic Press, San Diego)
- [19] Porezag D, Frauenheim T, Köhler T, Seifert G and Kaschner R 1995 Construction of tight-binding-like potentials on the basis of density-functional theory: Application to Carbon *Phys. Rev. B* 51 12947-12957
- [20] Verlet L, 1967 Computer experiments on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules *Phys. Rev.* 159 98-103
- [21] Nosé S, 1984 A molecular dynamics method for simulation in the canonical ensemble *Mol. Phys.* 52 255-268
- [22] Hoover W G 1985 Canonical dynamics: Equilibrium phase-space distributions *Phys. Rev. A* 31 1695-1697

Swarm Optimization-Based Personalization of Interactive Systems

Alexander Nikov, Stefka Stoeva, and Tricia Rambharose

Abstract—An approach for personalization of interactive systems based on particle swarm optimization (SOPA) is proposed. It restructures a hierarchical navigational menu to give the shortest path from one application functional interaction point (AFIP) to another. This allows adaptation to user preferences and minimizes the total selection time of menu options. SOPA is using a Particle Swarm Optimization (PSO) model to process the user interaction frequencies between one AFIP to another. Thus the discovering of usage patterns and the analysis of navigation behavior of interactive system user is supported. By SOPA the initial interaction tree structure of an interface is transformed into an optimized hierarchical menu structure. A case study simulating user interaction with a recruitment website illustrates SOPA. A comparison with Backpropagation algorithm shows lower training error for SOPA. Menu selection speed is improved, as the total number of clicks for user to get from one point of interaction to another is minimized. In this way, the personalized menu structure minimizes spanning time for frequently selected menu options and enhances the user efficiency of the interface.

Keywords—Computational intelligence, interactive systems, personalization, simulation.

I. INTRODUCTION

PERSONALIZATION is a key factor to be considered when designing user interfaces. Personalized interfaces help make user experience satisfying, efficient, minimize frustration and have a major impact on a firm's profitability [5], [19], [20], [22]. Interfaces can be personalized dynamically or customized. Customized personalization involves direct requests by the user [23]. There is a growing body of research however which indicates that dynamic, automatic or adaptive personalization allows users to achieve their goals faster, reduce navigational overhead, and increase satisfaction [4]. Many aspects of an interface can be personalized dynamically to improve usability [12]. Menu structures are a ubiquitous and mainstream channel of interface navigation. A myriad of experiments with various software mechanisms have been performed by researchers in order to create effective menu structures for all classes of

devices and systems.

Interactions with large menu structures pose a problem with menu design and structure. To increase the menu-item capacity the Hotbox widget [13] was developed which combines several GUI techniques. Menus are also static so even though many rows are used to present menu items, some menu options may still have a deep navigation path which has to be repeatedly traversed. This method, like many others [26], requires the user to learn a new way of interaction or navigation. This is both inconvenient and undesirable for users.

A split menu [21] splits a menu into two sections. Adaptation to user behavior is done by moving high-frequency menu selections to the top split and less frequent selections downward. Even though this menu structure was not static, it does not consider grouping similar submenus options together and does make the menu structure more efficient rather considers a menu as having only two sections.

The technique jumping menus [3] is most similar to the work done in current paper. This menu design approach addressed issues faced with navigation using cascading pull-down menus. The tactic used here is automatic horizontal 'jumping' of the cursor to the right into open sub-menu levels when a mouse click is detected inside a parent item. However, vertical cursor steering also impacts menu selection time. Deep menu structures will become monotonous and tedious to users especially if the user has to continuously 'jump to' the same submenus to reach their desired tasks. Another model which addresses similar issues with cascading pull-down menus is the use of a force field algorithm [2].

The disadvantage of most of the research previously carried out is that the techniques attempted to improve user interaction with menu interfaces by use of additional hardware or software or required taxing users' memory and defying taught habits to adjust to drastic changes in the GUI. Furthermore, menu structure techniques for all classes of devices and systems are very diverse and most do not apply to the specific application of the approach used in this paper. Even the techniques previously researched which are similar to the concept of this paper have the disadvantage of not adapting to user preferences.

In this paper, emphasis is placed on cascading pull-down menus. A Swarm Optimization-based Personalization Approach (SOPA) is proposed to minimize spanning time for

A. Nikov is with The University of the West Indies, St. Augustine, Trinidad and Tobago (phone: 868-6622002 ext. 84127; e-mail: alexander.nikov@sta.uiw.edu).

S. Stoeva is with the Bulgarian National Library, BG-1504 Sofia, Bulgaria (e-mail: greenapple_s2000@yahoo.com).

T. Rambharose is with Medullan Trinidad, Eastern Main Road, Red Hill, D'Abadie, Trinidad and Tobago (e-mail: tricia.rambharose@gmail.com).

frequently selected menu options. SOPA makes use of user logs, a web usage mining (WUM) technique, to calculate the frequencies of transition from one end point: application functional interaction point (AFIP) to another. WUM techniques have been used to discover web usage patterns and performs analysis of navigational behavior of web user [1].

One of the latest advancements in computational intelligence modelling [16], [18] the particle swarm optimization (PSO) algorithm is used to transform the initial interaction tree structure of an interface to an optimized hierarchical menu structure. Menu selection speed is improved, as the total number of clicks for users to get from one terminal point of interaction to another is minimized. The next part of this paper outlines detailed steps of the SOPA approach. This is followed with a simulation case study to which SOPA is applied.

II. SWARM OPTIMIZATION-BASED PERSONALIZATION APPROACH (SOPA)

In the following the main steps of SOPA approach are explained (cf. Fig. 1).

At **step 1** AFIPs of initial interaction structure (tree) are determined. The initial interaction structure of the interactive system is presented as a tree. In Fig. 2 $IP_1, IP_2, \dots, IP_{13}$ are the interaction points of the interaction structure presented as a tree hierarchy. Dialogue functional interaction points (DFIP) are connected with functions responsible for manipulation of dialogue objects, e.g. windows, menus. They correspond to nonterminal nodes of interaction tree (cf. IP_1, IP_2, IP_3, IP_7). Application functional interaction points (AFIP) are connected with functions changing the properties of application objects, e.g. files, text documents, paragraphs, rows, fonts, printing, saving, etc. They correspond to terminal nodes of interaction tree (cf. $IP_4, IP_5, IP_6, IP_8, IP_9, IP_{10}, IP_{11}, IP_{12}, IP_{13}$).

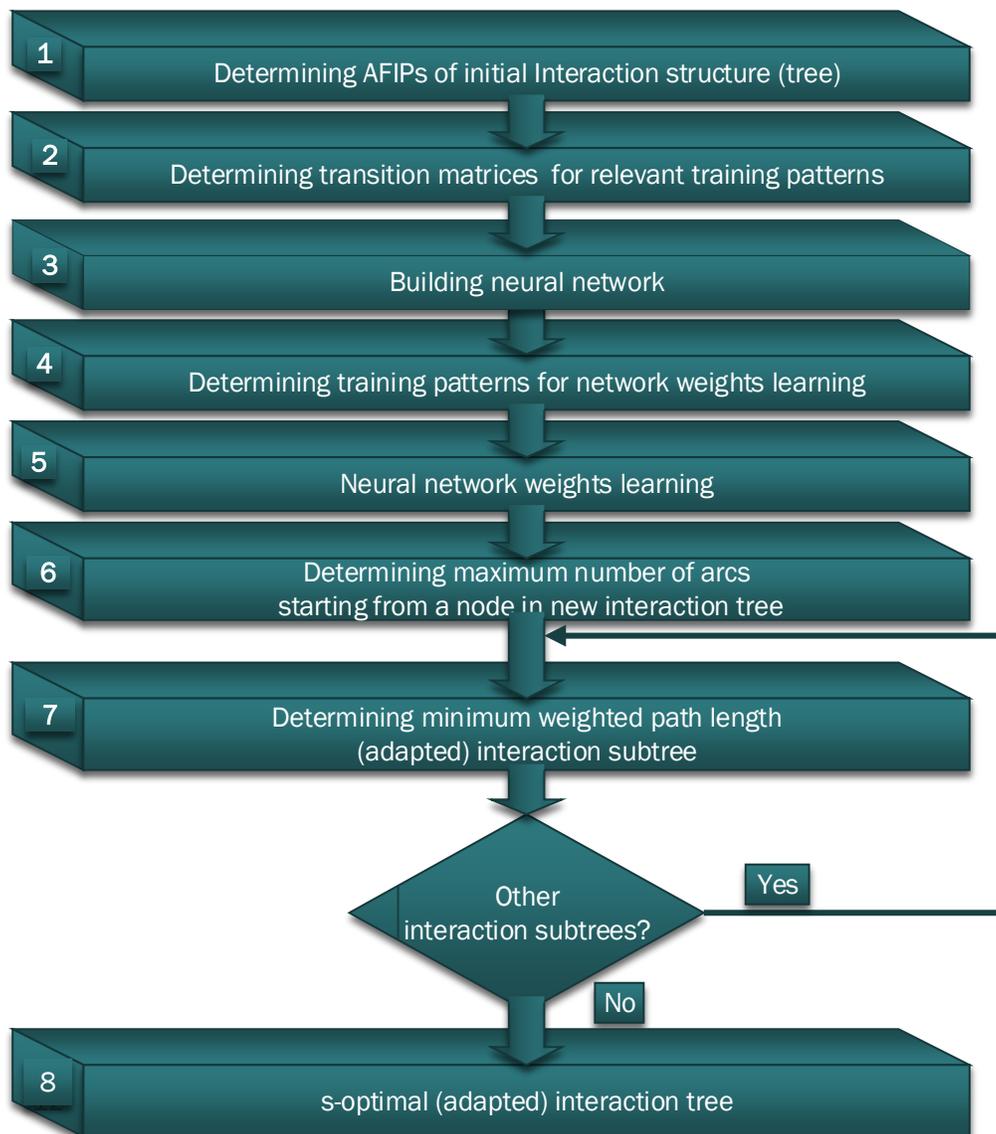


Fig. 1 SOPA sequence of steps

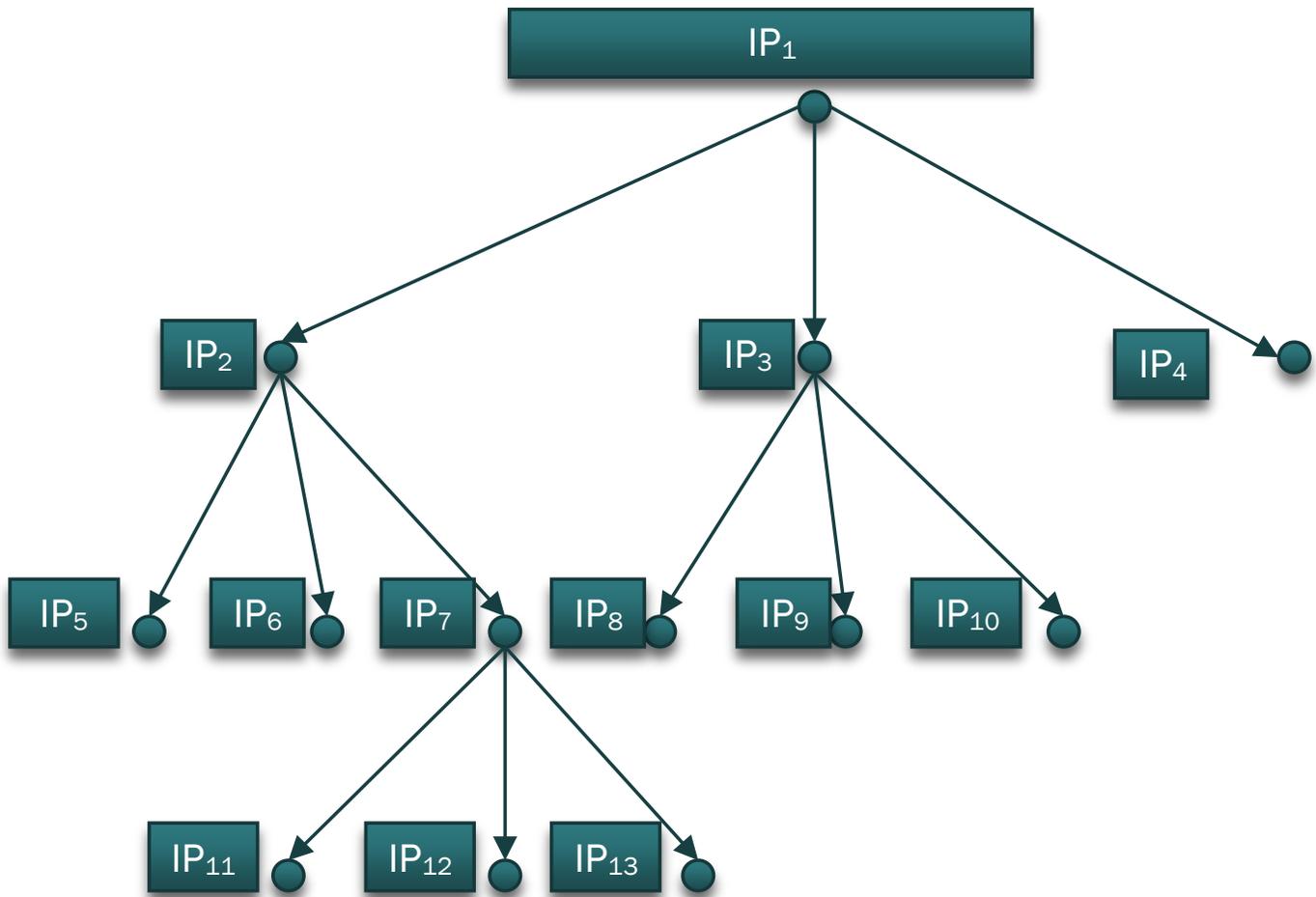


Fig. 2 Interaction structure example

For collection of training patterns with an interactive system at **step 2** are taken the transitions between all N AFIPs. For this purpose are used logs of particular work sessions from the beginning to completing the selected tasks. To each pattern corresponds a transition matrix, which cells consist of the frequencies of transitions between the relevant AFIPs. To the pattern with number m , $m=1, \dots, M$, corresponds the transition matrix U^m , where M is the number of all patterns. The cells of accumulation transition matrix \mathbf{K} contain the sum of frequencies of relevant patterns.

On the basis of accumulation transition matrix \mathbf{K} at **step 3** a two-layered neural network (NN) structure is determined [10], [15]. The input neurons $N^{(0)}, N^{(0)} \leq N$, of network layer 0 are relevant to AFIPs, which row sum transition frequencies are $\sum_{j=1}^N f_{ij} > 0, i=1, \dots, N$. The neurons $N^{(1)}, N^{(1)} \leq N$, of network layer 1 are relevant to AFIPs, which column sum transition frequencies are

$\sum_{i=1}^N f_{ij} > 0, j=1, \dots, N$. At this step the initial weights of network are determined using the maximal and minimal frequency respective element of accumulation transition matrix \mathbf{K} .

The following explanation of constructing and using this network can be given: The weights at first layer $w_{ij}^{(1)}$ based on transition frequencies between AFIPs can be used for forecasting (prediction) of the next user action. The neural network weights $w_{1j}^{(2)}$ at the second layer present the importance of AFIPs used for constructing the optimal interaction tree according to modified Huffman's algorithm [9]. This tree ensures the minimal total number of clicks for user to get from one AFIP to another AFIP

At **step 4** the matrix \mathbf{T} of the training patterns for network weights learning is determined. It includes the input values of the neural network for each training pattern and the target values.

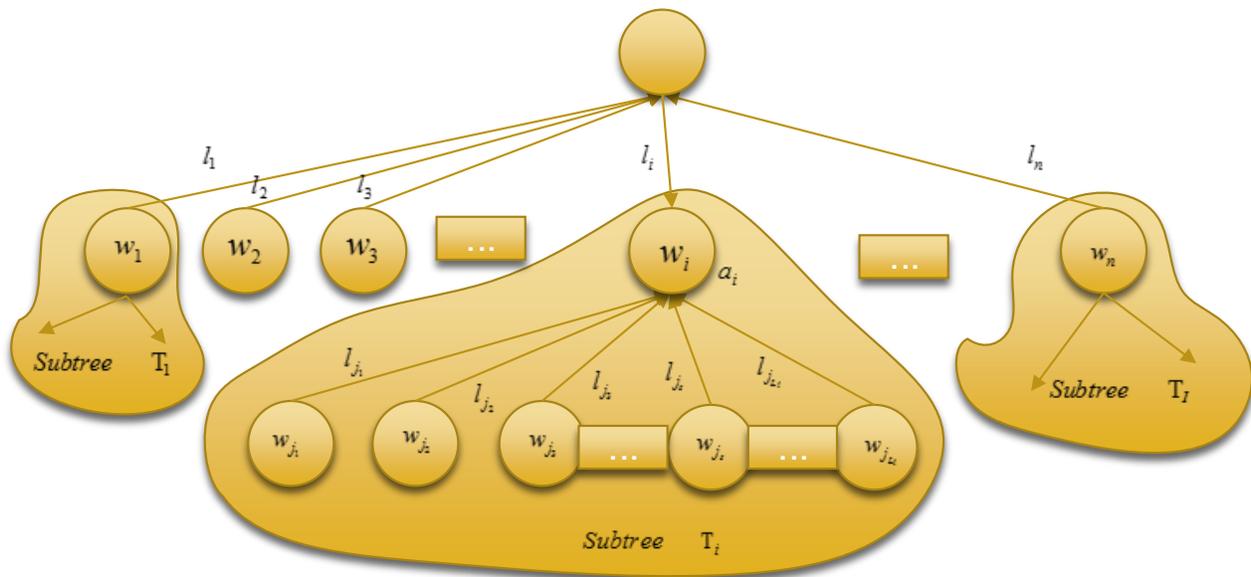


Fig. 3 Example of an interaction tree split into subtrees

For network-weights learning at **step 5** the computational intelligence technique Particle Swarm Optimization (PSO) algorithm [6], [11] is used. Here each input to the NN is viewed as a particle in multidimensional space. The particles are the input values for each row of the \mathbf{T} matrix. They use information about their own best solution and the global best solution to minimize the error between the NN outputs and the targets. NN trained by PSO was found to have good accuracy in searching for the global best result [25]. The average of final weights after training all patterns are the new set of input weights for training the patterns again, if necessary. Training stops when a calculated average error after training one pass over all patterns satisfy a stated error goal

At **step 6** the maximum number of arcs r starting from a node in adapted interaction tree are determined. According to functional and ergonomic requirements and constraints [7] the maximal number of arcs r starting from a node in the interaction tree is determined. For this purpose the Miller's number 7 ± 2 [14] is used. The interaction tree is now split into subtrees. With menu interfaces there are groups of menu options which must be together under a particular heading. These groups form the subtrees in the menu structure. For an optimized menu tree it is desirable that the options in these groups remain fixed, however the ordering of items within the groups can be optimized according to user preferences.

At **step 7** s-optimal interaction tree is constructed. The sequential application of Huffman algorithm [9] to each subtree, using the derived weights after NN training, generates the s-optimal interaction tree (cf. Fig. 3). Here the path length vector $\lambda = (l_1, l_2, \dots, l_L)$ is s-optimal for the weight vector $W = (w_1, w_2, \dots, w_L)$ iff the path length vector $\lambda_i = (l_{j_1}, l_{j_2}, \dots, l_{j_{L_i}})$ is minimal according to

Huffman for the weight vector $W_i = (w_{j_1}, w_{j_2}, \dots, w_{j_{L_i}})$ of each subtree T_i of the tree \mathbf{T} considering semantic implications of grouping.

During **step 8** an interaction tree is generated, which is s-optimal. This new tree structure should arrange the menu items in such a way that the path for a user to get from one option to another is the minimum path, therefore requiring minimum number of clicks. Depending on size of data collected from user interaction periodically the user will be offered a modification/personalization of interaction structure [16].

III. SIMULATION CASE STUDY

SOPA was applied to a simulated real life example of a prototype of a job recruitment website. It was used to illustrate the feasibility of SOPA. A segment of the initial interaction tree is given in Fig. 4 with the weights for each AFIP after training.

PSO-based SOPA was compared with the most popular neural networks approach the backpropagation (BP) algorithm. In this case study the adaptive learning rate backpropagation algorithm [24] was used. Both approaches were trained using very similar parameters for comparability. All networks for both approaches were trained for 2000 iterations maximum, or till a specified error goal of 0.01 was reached. The initial layer weights for training all patterns for the first time were the values calculated at Step 3. A text file was randomly generated to simulate an input log file of $M=4$ patterns of $N=32$ AFIPS each. Generated output included an error graph, a text file stating relevant training results, Huffman optimized subtrees structures and the s-optimal tree structure. The training with PSO was significantly longer than with BP but however resulted in a much smaller training error (cf. Fig. 5).

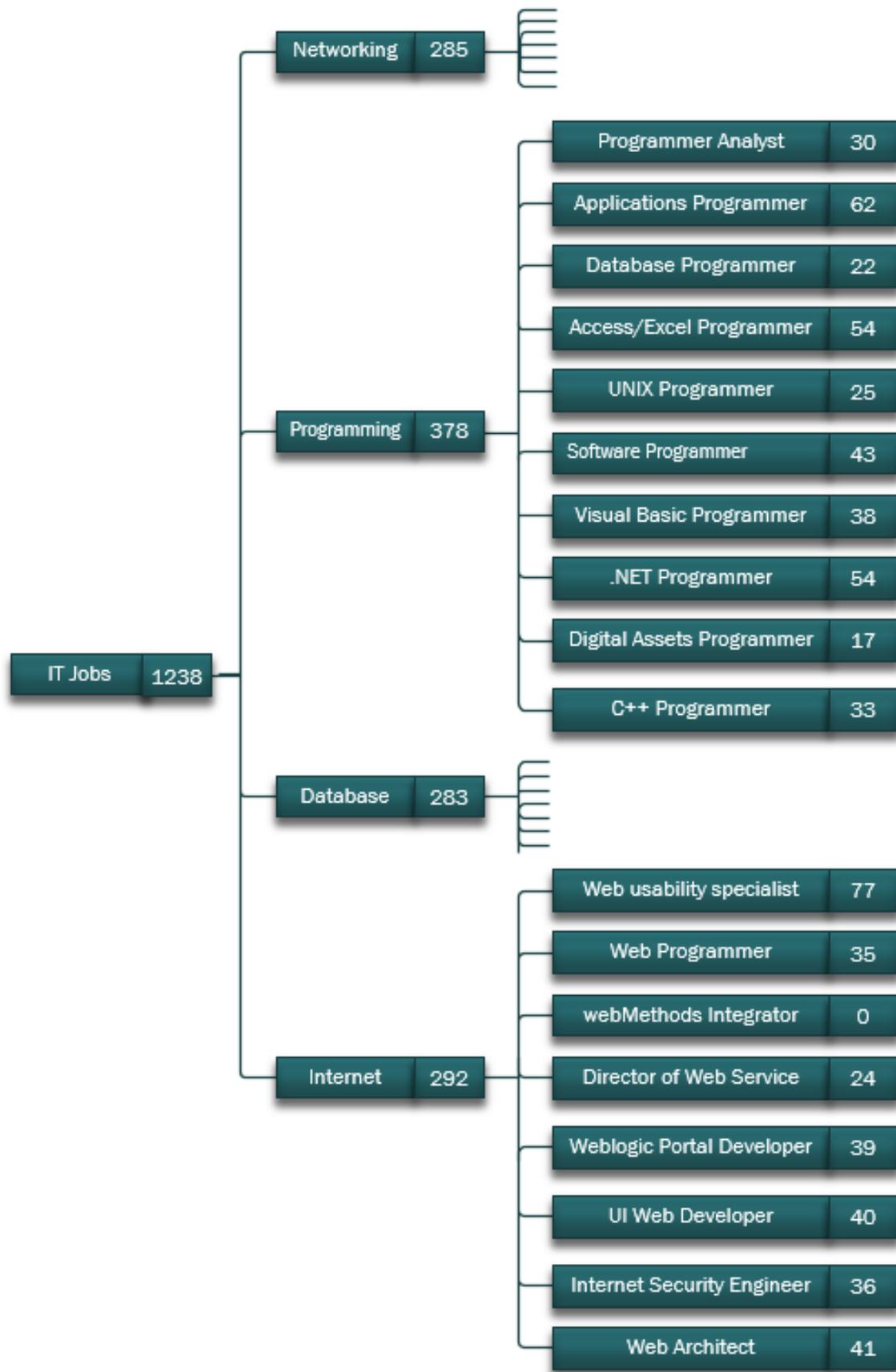


Fig. 4 Segment of initial interaction tree showing weights after training with PSO

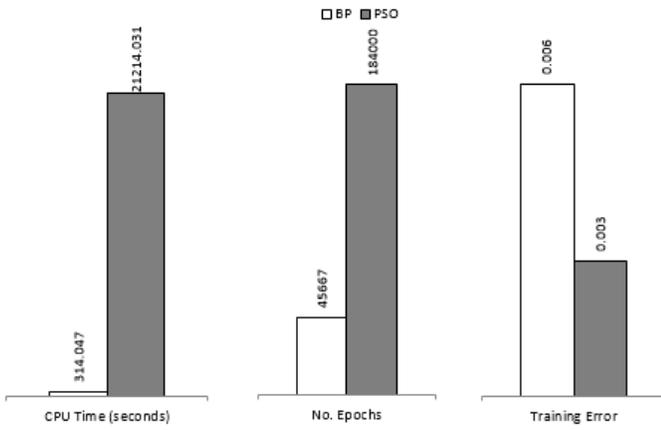


Fig. 5 BP and PSO training results

r was set equal to 5 conforms according to Miller’s rule. The user is no longer presented with a longer list of options than they can mentally process at first glance. The highest valued weights are first presented to the user and the lower value weighted menu options are shown at a deeper submenu. On Fig. 6 is presented a segment of the resulting interaction structure assuming optimal (minimal path length) very fast access to all AFIPs

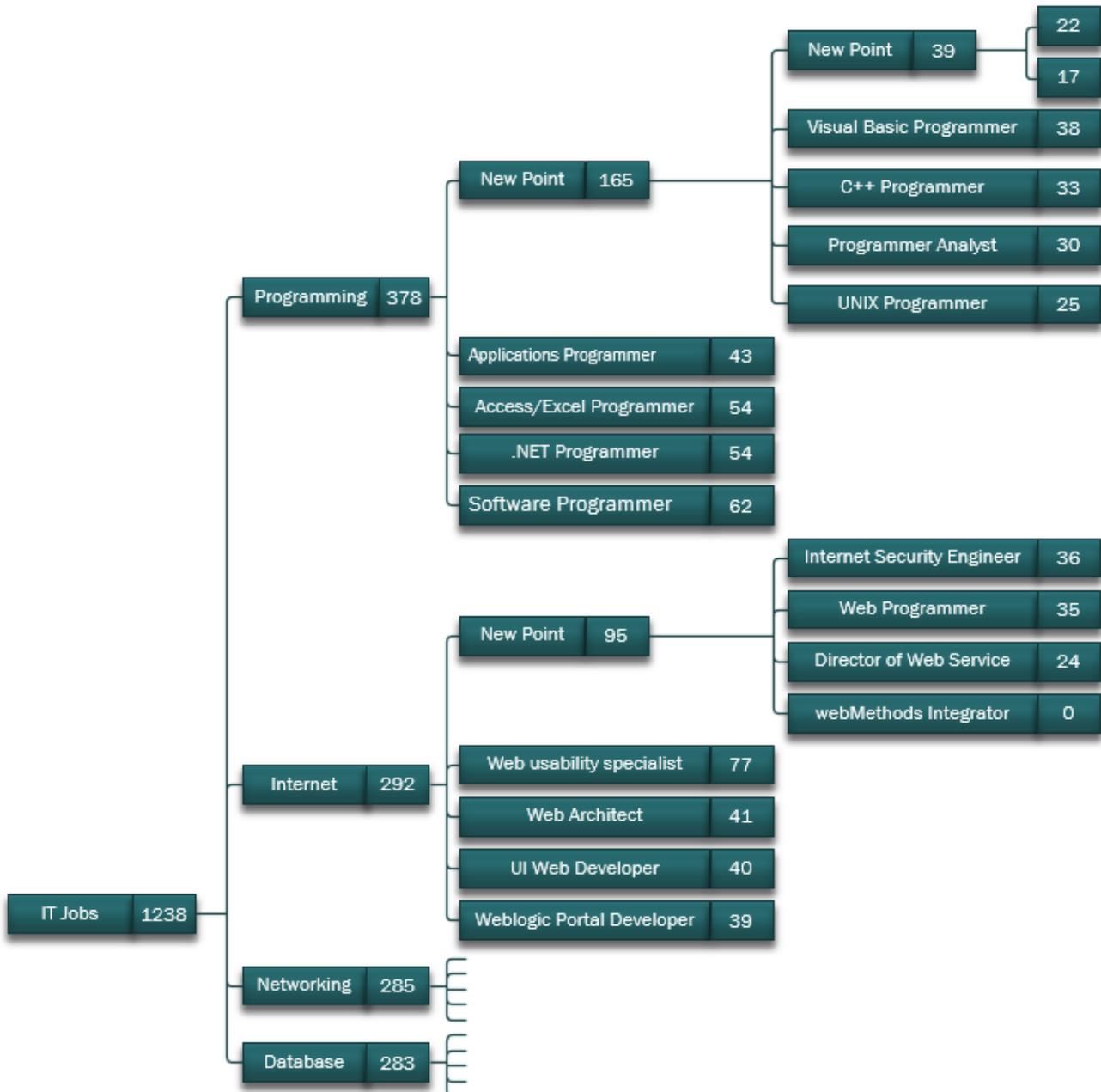


Fig. 6 Segment of s-optimal interaction tree using PSO weights

IV. CONCLUSIONS

Personalized navigation in user interfaces poses a problem across the board for all devices and systems. Menu structures are the most ubiquitous navigation tool. The simplicity to this navigation technique however should not be taken for granted by haphazardly grouping menu options into too deep nor too narrow a hierarchical structure. The current structure of cascading pull-down menus does not cater for the frequency of user transitions between menu selections. These static menu structures make navigation monotonous, error prone and time consuming especially when users have to navigate to deep submenus.

The SOPA approach provides adaptable navigation in cascading pull-down menu structures. This new technique takes the initial interaction tree and transition frequencies as input and creates an optimized interaction menu structure with minimized total number of clicks to terminal menu items. This novel approach can dramatically decrease menu selection time for both expert and novice users of an interface.

Further research is planned with real data for testing the algorithm, for measuring user task performance and for collecting subjective user feedback by a questionnaire.

REFERENCES

- [1] B. Abdurrahman, "Web usage mining for analyzing unique behavior of web user," in *Proc. International Conference on Electrical Engineering and Informatics*, 2007, pp. 356-359.
- [2] D. Ahlstrom, "Modeling and improving selection in cascading pull-down menus using Fitts' law, the Steering Law and Force Fields," in *Proc. CHI 2005, ACM Press*, 2005, pp. 61-70.
- [3] D. Ahlstrom, R. Alexandrowicz, and M. Hitz, "Improving menu interaction: a comparison of standard, force enhanced and jumping menus," in *Proc. CHI 2006, ACM Press* 2006, pp. 1067-1076.
- [4] P. Brusilovsky, "Adaptive Navigation Support," in P. Brusilovsky, A. Kobsa and W. Nejdl (Eds.). *The Adaptive Web, Springer*, 2007, pp. 263-290.
- [5] G. Chakraborty, V. Lala, and D. Warren, "An empirical investigation of antecedents of B2B Websites' effectiveness," *Journal of Interactive Marketing*, 16(4), pp. 51-72, 2002.
- [6] R. Eberhart, and J. Kennedy, "A new optimizer using particles swarm theory," in *Proc. Sixth International Symposium on Micro Machine and Human Science*, 1995, pp. 39-43.
- [7] *Ergonomic requirements for office work with visual display terminals (VDTs). Guidance on usability*, ISO 9241-11 standard, 1998.
- [8] T. Halverson, and A. J. Hornof, "A Computational Model of "Active Vision" for Visual Search in Human-Computer Interaction," *Human-Computer Interaction*, Vol. 26, pp. 285-314, 2011.
- [9] D. A. Huffman, "A method for the construction of minimum redundancy codes," in *Proc. IRE 40(9)*, 1952, pp. 1098-1101.
- [10] B. Irie, and S. Miyake, "Capability of three-layered perceptrons," in *Proc. of IEEE Int. Conf. On Neural Networks*, 1988, pp. 641-648.
- [11] J. Kennedy, and R. Eberhart, "Particle Swarm Optimization," in *Proc. IEEE Int'l. Conf. on Neural Networks*, 1995, pp. 1942-1948.
- [12] P. T. Kortum, and A. Bangor, "Usability Ratings for Everyday Products Measured With the System Usability Scale," *Intl. Journal of Human-Computer Interaction*, 29(2), pp. 67-76, 2013.
- [13] G. Kurtenbach, G. W. Fitzmaurice, R. N. Owen, and T. Baudel, "The Hotfix: efficient access to a large number of menu-items," in *Proc. CHI 1999, ACM Press*, 1999.
- [14] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *The Psychological Review*, Vol. 8, pp. 81-97, 1956.
- [15] D. Nguyen, and B. Widrow, "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights," in *Proc. IJCNN*, 1990, pp. 21-26.
- [16] A. Nikov, H. -G. Lindner, and T. Georgiev, "A control structure for adaptive interactive systems," in H. -J. Bullinger and J. Ziegler (Eds.). *Human-Computer Interaction: Ergonomics and User Interfaces*, Lawrence Erlbaum Associates Inc. Publishers, London, Mahwah, New Jersey, pp. 351-356, 1999.

- [17] T. Rambharose, and A. Nikov, "Computational intelligence-based personalization of interactive web systems," *WSEAS Transactions on Information Science and Applications*, 7(4), pp. 484-497, 2010.
- [18] T. Rambharose, and A. Nikov, "Personalization of web-based systems based on computational intelligence modelling," in *Proc. 4th WSEAS International Conference on Computer Engineering and Applications CEA'10*, 2010, pp. 170-175.
- [19] S. Ramnarayan, "Perceived effectiveness of personalization," *Journal of Business & Economics Research*, 3(9), pp. 41-49, 2005.
- [20] P. Saariluomaand, and J. P. P. Jokinen, "Emotional Dimensions of User Experience: A User Psychological Analysis," *Intl. Journal of Human-Computer Interaction*, 30(4), pp. 303-320, 2014.
- [21] A. Sears, and B. Shneiderman, "Split menus: effectively using selection frequency to organize menus," *ACM Transactions on Computer-Human Interaction (TOCHI)*, 1(1), pp. 27-51, 1994.
- [22] K. Y. Tam, and S. Y. Ho, "Web personalization: is it effective?" *IT Professional*, 5(5), pp. 53-57, 2003.
- [23] D. S. Weld, "Automatically Personalizing User Interfaces," in *Proc. IJCAI'03, Morgan Kaufmann*. 2003, pp. 1613-1619.
- [24] P. J. Werbos, "Beyond regression: New tools for predictions and analysis in the behavioral science," Ph.D. Thesis, Harvard University, 1974.
- [25] C. Zhang, H. Shao, and Y. Li, "Particle swarm optimization for evolving artificial neural network," in *Proc. of IEEE Int. Conf. on System, Man, and Cybernetics*, 2000, pp. 2487-2490.
- [26] G. Zhang, G. Shen, J. Staiger, A. Troy, and S. Jiayang, *FcAWN: Concept analysis as a formal method for automated web-menu design*, Shaker Verlag, 2004.



Alexander Nikov received his Ph.D. in 1985 in Biomedical Engineering at Technical University of Sofia, Bulgaria, and the Dr. habil. in 2005 in Human Factors Engineering and Human-Computer Interaction at Brunswick University of Technology, Germany. He is an Associate Professor in Computer Science at the University of the West Indies, Trinidad and Tobago and head of User Experience Living Lab member of European Network of Living Labs.

He has developed and modified methods of multivariate statistics, fuzzy logic, neural networks and swarm optimization and applied them to user experience design, human-computer interaction, human factors, industrial design, ecology and medicine. The results his research were published in 156 books, journals and conference proceedings. His current research includes computational-intelligence-based machine learning and its application to user experience design (UXD): emotional, personalized and workplace UXD. (cf. <http://www2.sta.uwi.edu/~anikov>)

PMSG wind system control for time-variable wind speed by imposing the DC Link current

Ciprian Sorandaru, Sorin Musuroi, Gheza-Mihai Erdodi and Doru-Ionut Petrescu

Abstract—This paper presents a method for controlling a wind system, - wind turbine (WT) + permanent magnet synchronous generator (PMSG) - so as to reach an optimal energetic operation at a time-variable wind speed. Wind speed and momentary mechanical angular speed of the PMSG impose the generator load value in the energetically optimal region. By energy balance measurements made with speed and power measurements, the generator load is determined so that the system has been brought into the energetic optimal region. It analyzes the maximum power operation in a WT by changing the load to the generator, while the wind speed significantly varies over time. The coordinates of the maximum power point (MPP) changes over time and they are determined by the values of the wind speed and mechanical inertia. Not always wind system can be lead in a timely manner in the MPP. The speed variation of the wind speed and the inertia value are two fundamental elements on which the MPP operation depends. By prescribing the amount of DC link current, I_{cc} , the main circuit of the converter can achieve a simple and useful system tuning WT + PMSG. Operation control method in the optimal energy region of the WT is based on the knowledge of the I_{cc} current value, which is determined by wind speed and momentary mechanical angular speed, MAS.

Keywords—Permanent Magnet Synchronous Generator, mathematical model of Wind Turbine, maximum power point, wind system.

I. INTRODUCTION

IN the literature [1-21] various mathematical models of wind turbines (MM-WT) offered by building companies and/or obtained under laboratory conditions are presented, far different from those in real conditions operation [7, 12, 19]. For this reason the final result, especially the obtained electrical energy has a value less than the maximum possible at the maximum power point (MPP) operating at optimal mechanical angular speed (MAS). In most works is treated the operation of the wind turbine (WT) at MPP. [3, 5, 11, 21]. In some cases, [7, 9, 15, 11, 21], there are used mathematical

C. Sorandaru is with Politehnica University of Timisoara, Department of Electrical Engineering, Timisoara, Bd. V. Parvan 2, Romania (corresponding author to provide phone: +40-256-403466; fax: +40-256-403452; e-mail: ciprian.sorandaru@upt.ro).

S. Musuroi is with Politehnica University of Timisoara, Department of Electrical Engineering, Timisoara, Bd. V. Parvan 2, Romania (e-mail: sorin.musuroi@upt.ro).

G.M. Erdodi is with Politehnica University of Timisoara, Department of Mechanical Engineering, Timisoara, Bd. M. Viteazu 1, Romania (e-mail: geza.erdodi@erlendieselservice.ro).

D.I. Petrescu is with Politehnica University of Timisoara, Department of Mechanical Engineering, Timisoara, Bd. M. Viteazu 1, Romania (e-mail: petrescu.doru@yahoo.com).

models which are only partially valid, because of the continuous varying weather conditions. The laboratory conditions where they have obtained the turbine characteristics are different from those in real operation [11, 15, 17].

Recent works [1, 2, 3, 4] use control algorithms based on the measurement of wind speed and prescribing optimal speed of the mechanical angular speed in the MPP region. The estimation of the optimal MAS on the basis of the wind speed is a complex problem solved by mathematical calculations and with specialized simulation software [2, 3, 5].

Method of bringing the wind system operating point in the MPP region, by appropriately modifying the electric generator load requires the measurement of the wind speed and is quite powerful, [17,19,21], in certain circumstances. It can analyze these variations in time by knowing the wind speed and given the values of the moments of inertia.

There are geographical areas where the wind speed changes its value in less time [8, 9, 17]. In Romania, the wind speed varies in time and therefore the method can be applied in certain areas only after a prior study.

The method is based on the dependency of the power of WT on MAS, that means the function $P_{WT}(w)$ has, at a certain speed, a maximum value for MAS, ω_{OPTIM} (Fig. 1).

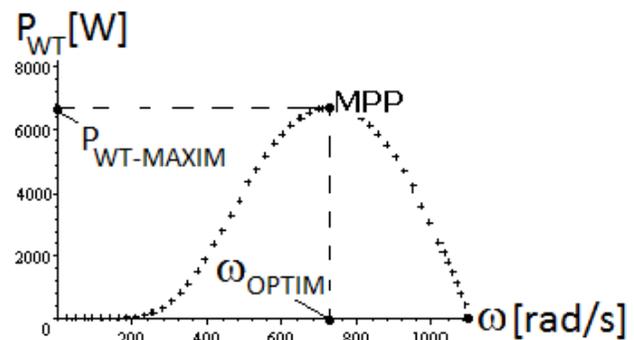


Fig.1.Power characteristic of the WT

For wind speed which does not change his value over time, the operation in the MPP region can be performed quite simply. For wind speeds which significantly vary over time, the problem becomes complex and sometimes unsolvable (if the wind quickly changes the speed).

Analysis of the MPP operation is done by simulation using specific mathematical models for WT and PMSG

By changing the PMSG load, the system try to reach the MPP region and the transient phenomena can be visualized by solving the movement equation WT+PMSG system.

II. THE MATHEMATICAL MODEL OF THE WIND TURBINE

We will use a classical turbine model [14], which allows the estimation of the reference angular speed w_{ref} . The mathematical model of the WT allows also the calculation of the optimal speed, so as the captured energy will be a maximum one.

The power given by the WT can be calculated using the following equation:

$$P_{WT} = \rho \pi R_p^2 C_p(\lambda) V^3 \quad (1)$$

where: r - is the air density, R_p - the pales radius, $C_p(l)$ - power conversion coefficient, $l = R\omega/V$, V - the wind speed, ω - mechanical angular speed (MAS).

The power conversion coefficient, $C_p(l)$, could be calculated as follows:

$$C_p(\lambda) = c_1 \left(\frac{c_2}{\lambda} - c_3 \right) e^{-\frac{c_4}{\lambda}}, \quad (2)$$

$$\frac{1}{\lambda} = \frac{1}{\lambda} - 0.0035, \quad (3)$$

$c_1 - c_4$ are data-book constants.

$$\frac{1}{\lambda} = \frac{1}{\lambda} - 0.0035 = \frac{V}{R\omega} - 0.0035 = \frac{V}{1.5\omega} - 0.0035$$

By replacing, we can obtain the the power conversion coefficient as follows:

$$C_p(\lambda) = c_1 \left(\frac{c_2}{\lambda} - c_3 \right) e^{-\frac{c_4}{\lambda}} = c_1 \left(c_2 \left(\frac{V}{1.5\omega} - 0.0035 \right) - c_3 \right) e^{-c_4 \left(\frac{V}{1.5\omega} - 0.0035 \right)} \quad (4)$$

And the power given by the wind turbine can be calculated as follows:

$$P_{WT}(\omega, V) = \rho \pi R^2 C_p(\lambda) V^3 = 1.225\pi 1.5^2 c_1 \left(c_2 \left(\frac{V}{1.5\omega} - 0.0035 \right) - c_3 \right) e^{-c_4 \left(\frac{V}{1.5\omega} - 0.0035 \right)} V^3$$

or

$$P_{WT}(\omega, V) = \rho \pi R^2 C_p(\lambda) V^3 = k_1 \left(k_2 \left(\frac{V}{\omega} - 0.0525 \right) - c_3 \right) e^{-k_3 \left(\frac{V}{\omega} - 0.0525 \right)} V^3 \quad (6)$$

Where $k_1 = 1.225\pi 1.5^2$, $k_2 = c_2/1.5$, $k_3 = c_4/1.5$.

For the wind turbine WT, the producer gives the experimental power characteristics, $P_{WT}(\omega, V)$, or torque characteristics $T_{WT}(\omega, V)$, the last ones being known as mechanical experimental characteristics.

$$T_{WT}(\omega, V) = \frac{P_{WT}(\omega, V)}{\omega} = k_1 \left(k_2 \left(\frac{V}{\omega} - 0.0525 \right) - c_3 \right) e^{-k_3 \left(\frac{V}{\omega} - 0.0525 \right)} V^3 / \omega. \quad (7)$$

The maximum value of the function $P_{WT}(\omega, V)$ is achieved for a reference MAS ω_{ref} , as follows:

$$\frac{dP_{TV}}{d\omega} = \frac{d}{d\omega} \left(k_1 \left(k_2 \left(\frac{V}{\omega} - 0.0525 \right) - c_3 \right) e^{-k_3 \left(\frac{V}{\omega} - 0.0525 \right)} V^3 \right) = 0 \quad (8)$$

and it yields

$$\omega_{ref} = \omega_{OPTIM} = 400 \cdot k_3 \frac{k_2}{400 \cdot k_2 + 21 \cdot k_3 k_2 + 400 \cdot k_3 c_3} \cdot V = k_4 \cdot V \quad (9)$$

This result proves the direct link between reference speed and wind speed.

By replacing this result, it yields:

$$P_{WT-MAX}(V) = k_p \cdot V^3 \quad (10)$$

This result proves a cubic dependency of the WT power on the wind speed.

If the wind speed has large variations, this result must be reanalyzed.

The mathematical model of the PMSG

To analyze the behavior of the system WT-PMSG for the the time-varying wind speeds, it uses orthogonal mathematical model for permanent magnet synchronous generator (PMSG) given by the following equations [5]:

$$\begin{cases} -U\sqrt{3} \sin \theta = R_1 I_d - \omega L_q I_q \\ U\sqrt{3} \cos \theta = R_1 I_q + \omega L_d I_d + \omega \Psi_{PM} \\ T_{PMSG} = p_1 (L_d - L_q) I_d I_q + I_q \Psi_{PM} \end{cases} \quad (11)$$

where: U - stator voltage

I_d, I_q - d-axis and q-axis stator currents

θ - load angle

R_1 - phase resistance of the generator;

L_d - synchronous reactance after d axis;

L_q - synchronous reactance after q axis;

Ψ_{PM} - flux permanent magnet;

T_{PMSG} - PMSG electromagnetic torque

III. OPERATING CONTROL IN THE MPP REGION

The study of operation in the MPP region will be performed by simulation using the following mathematical models.

The mathematical model for the WT (MM-WT)

For the wind turbine, the producer provides the (5) experimental power characteristics [14], $P_{WT}(w, V)$

$$P_{WT}(\omega, V) = 1191.5 \cdot (V/\omega - 0.02) \cdot e^{-98.06 \cdot (V/\omega)} \cdot V^3 \quad (12)$$

The reference MAS, ω_{ref}

The maximum value of the function $P_{WT}(w, V)$ is obtained for the reference MAS, ω_{ref} , by differentiation:

$$\frac{dP_{WT}(\omega, V)}{d\omega} = \frac{d}{d\omega} \left(1191.5 \cdot (V/\omega - 0.02) \cdot e^{-98.06 \cdot (V/\omega)} \cdot V^3 \right) = 0 \quad (13)$$

$$\omega_{ref} = 31.115 \cdot V \quad (14)$$

For this value of MAS, the maximum power is obtained:

$$1191.5 \cdot (V/\omega - 0.02) \cdot e^{-98.06 \cdot (V/\omega)} \cdot V^3 = 0.61884 \cdot V^3 \quad (15)$$

$$P_{WT-MAX} = 0.61884 \cdot V^3 \quad (16)$$

The mathematical model for the PMSG (MM-PMSG)

From the nominal values of the PMSG [1], for the nominal power: $P_N = 5$ [kW], it yields $R_1 = 1.6$ [W], $L_d = 0.07$ [H], $L_q = 0.08$ [H], $\Psi_{PM} = 1.3$ [Wb].

From the equations of the PMSG, it obtains

$$\begin{cases} -R I_d = 1.6 I_d - \omega \cdot 0.08 \cdot I_q \\ -R I_q = 1.6 I_q + \omega \cdot 0.07 \cdot I_d + \omega \Psi_{PM} \\ T_{PMSG} = -0.01 \cdot I_d I_q + I_q \Psi_{PM} \\ \Psi_{PM} = 1.3 \\ P = (I_d^2 + I_q^2) \end{cases} \quad (17)$$

$$P_{PMSG} = 4225R\omega^2 \frac{4\omega^2 + 625R^2 + 2000R + 1600}{(1250R^2 + 4000R + 3200 + 7\omega^2)^2} \quad (18)$$

$$T_{PMSG} = 845\omega(5R + 8) \cdot \frac{4\omega^2 + 625R^2 + 2000R + 1600}{(1250R^2 + 4000R + 3200 + 7\omega^2)^2} \quad (19)$$

$$U_{CC} = 500 \text{ [V]} \quad (20)$$

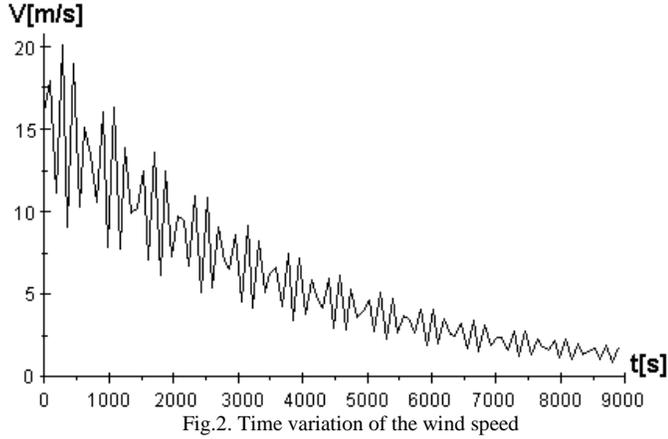
$$P_{PMSG} = U_{CC} \cdot I_{CC} \quad (21)$$

$$I_{CC} = P_{CC}/U_{CC} \quad (22)$$

3.1. Case study for time-variable wind speed

For a sinusoidal time-variable wind speed, as presented in Fig. 2, with $T = 35$ [s]:

$$V(t) = (16 - 6 \cdot (\sin 0.17943t))e^{-t/3600} \quad (23)$$



The wind speed is continuously monitored and the equivalent wind speed and the optimum DC link current are calculated at discrete time intervals $\Delta t = T$.

The value of the DC link current I_{CC} is also continuously monitored, depending on the error:

$$\Delta I = I_{CC} - I_{CC-OPTIM} \quad (24)$$

the load resistance R is consequently modified.

The control of the wind system is realized based on the two measurements, presented above:

1. Wind speed
2. Current I_{CC}

Using [1], for the time interval $\Delta t = [a, a+T]$ we can define an equivalent wind speed, as follows:

$$V_{ECH} = \sqrt{\frac{1}{T} \int_a^{a+T} \left((16 - 6 \cdot (\sin 0.17943t))e^{-t/3600} \right)^2 dt} \quad (25)$$

With a period of 35 [s], optimal MAS is calculated starting from $t=40$ [s] (i.e. 40, 75, 110 ... [s]), using the dependency:

$$\omega_{OPTIM} = 31.817 \cdot V_{ECH} \quad (26)$$

The following results are obtained:

- For the interval $\Delta t = 5+40$ [s], $V_{ECH} = 17.187$ [m/s] and $\omega_{OPTIM} = 546.84$ [rad/s]
- For the interval $\Delta t = 40+75$ [s], $V_{ECH} = 17.021$ [m/s] and $\omega_{OPTIM} = 541.56$ [rad/s]
- For the interval $\Delta t = 75+110$ [s], $V_{ECH} = 16.856$ [m/s] and $\omega_{OPTIM} = 536.31$ [rad/s]

3.1.1. The control system by imposing the current I_{CC}

The power acquired by the PMSG is found in the intermediate circuit power and, from this equation, the I_{CC} current is obtained. (21) and (23) - (25)

$$I_{CC} = \frac{P_{PMSG}}{U_{CC}} = \left(4225R\omega^2 \frac{4\omega^2 + 625R^2 + 2000R + 1600}{(1250R^2 + 4000R + 3200 + 7\omega^2)^2} \right) / 500 \quad (27)$$

Time evolution of the process

The simulations are based on the mechanical equation:

$$J \frac{d\omega}{dt} = T_{WT} - T_{PMSG} \quad (28)$$

where J is equivalent inertia moment, T_{PMSG} is the torque of PMSG, T_{WT} is the torque of WT. By imposing the conduction angle of the converter between PMSG and the network, different values for load resistance and thus for the current I_{CC} are obtained.

The system is lead in the optimal energy region by imposing a DC link current, as results from energy balance, presented below:

To obtain the optimum MAS, ω_{OPTIM} , the PMSG load must be adjusted based on:

- kinetic energy variations of the moving parts
- optimum MAS to be reached at the moment $t=45$ [s]

From the mechanical equation, it yields:

$$J \frac{d\omega}{dt} \omega = \omega \cdot T_{WT} - \omega \cdot T_{PMSG} \quad (29)$$

$$J \cdot (\omega_k^2 - \omega_{k-1}^2) / 2 = \int_{t_{k-1}}^{t_k} P_{WT} \cdot dt - \int_{t_{k-1}}^{t_k} P_{PMSG} \cdot dt \quad (30)$$

The energy to be captured by the PMSG, during $\Delta t = t_k - t_{k-1}$ time interval is:

$$W_{PMSG} = \int_{t_{k-1}}^{t_k} P_{PMSG} \cdot dt = \int_{t_{k-1}}^{t_k} P_{WT} \cdot dt - J \cdot (\omega_k^2 - \omega_{k-1}^2) / 2 = E(\Delta t) - J \cdot (\omega_k^2 - \omega_{k-1}^2) / 2 \quad (31)$$

Where $E(\Delta t)$ is the value of energy to be captured during Δt time interval. It has two components:

1. $\int_{t_{k-1}}^{t_k} P_{WT} \cdot dt$ - energy captured by the wind turbine
2. $J \cdot (\omega_k^2 - \omega_{k-1}^2) / 2$ - rotational kinetic energy

The control process has two steps:

Step 1: bringing the system in the optimum energetic region

Step 2: keeping the system in the optimum energetic region

Step 1: bringing the system in the optimum energetic region Could be done in two ways:

- a. By loading the generator at maximum power if the initial MAS is greater than the optimum value
- b. By no-load operation if the initial MAS is less than the optimum value

a. PMSG loading at maximum admissible power:

Starting from an initial speed $\omega(0)=555$ m/s, from (a1) we can obtain min and max values for WT power:

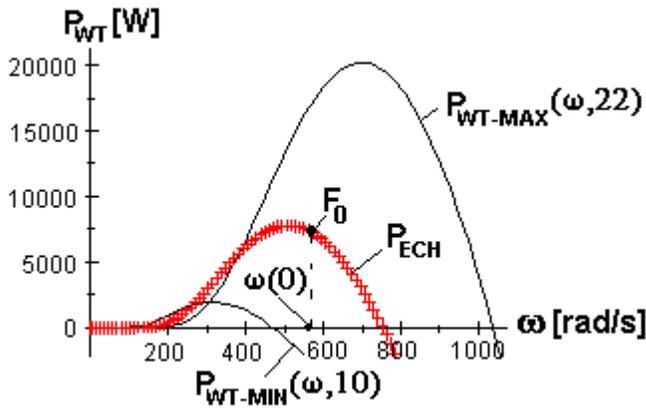


Fig. 3. The power characteristics

It is necessary to bring the WT at optimal speed and only after connect the generator to the grid.

b. Generator operates at no-load

- Measurement of wind speed and calculation of ω_{OPTIM} ;
- Measurement of MAS and comparison with ω_{OPTIM} ;
- When $\omega = \omega_{OPTIM}$, the generator is connected to the grid.

After calculations, the following values are obtained:

$$\begin{aligned} \omega_{OPTIM-40} &= 546.84 \text{ [rad/s]} \\ P_{PMSG-40} &= 7075.9 \text{ [W]} \\ R &= 453.85 \text{ [W]} \end{aligned}$$

Step 2: keeping the system in the optimum energetic region
Load at $t=75$ [s]

The energy captured by the PMSG, W_G , in the interval $Dt = 40+75$ [s] can be estimated by measuring the electrical energy during this interval or, by simulations, from the mechanical equation and using the PMSG power.

During this interval, the variation of the kinetic energy is:

$$W_{KINETIK-REAL} = J \cdot (\omega^2(75) - \omega^2(40))/2 = -6778.9 \text{ [J]} \quad (32)$$

The electric energy captured by the PMSG, during the same interval, is:

$$W_G(35) = 2.4710 \times 10^5 \text{ [J]} \quad (33)$$

The wind energy captured by the wind turbine is:

$$E(35) = 2.4028 \times 10^5 \text{ [J]} \quad (34)$$

It can prove the conservation of energy, with a very small error ($\approx 10^{-2}$ %).

Remark 1: Practically, based on the variations of kinetic energy and energy captured by the PMSG, the wind energy can be obtained.

To reach optimum MAS

$$\omega_{OPTIM-75} = 541.56 \text{ [rad/s]} \quad (35)$$

it would be necessary a load for the generator calculated from energy equation:

Required kinetic energy:

$$W_{KINETIK-REQ} = J \cdot (\omega_{OPTIM-75}^2 - \omega^2(40))/2 = -1.1494 \times 10^5 \text{ [J]} \quad (36)$$

Wind energy captured in this time interval:

$$E(35) = W_G(35) + W_{KINETIK-REAL} = 2.4032 \times 10^5 \text{ [J]} \quad (37)$$

The required energy for the PMSG is:

$$W_{PMSG-REQ}(35) = E(35) - W_{KINETIK-REQ} = 3.5522 \times 10^5 \text{ [J]} \quad (38)$$

By estimation a medium power during this interval,

$$P_{PMSG-MED} = W_{PMSG-REQ}(35)/35 = 10149 \text{ [W]} \quad (39)$$

Using power equation and with $w = 544.2$, the required load to reach the optimal region is:

$$R_{PMSG-REQ-75} = 311.64 \text{ [\Omega]} \quad (40)$$

In these conditions, the power to be prescribed to the PMSG ($P_{PMSG-P-75}$) is:

$$P_{PMSG-P-75} = 10001 \text{ [W]} \quad (41)$$

Remark 2: The captured wind energy is about two times greater than the variations of kinetic energy. So, for $t=75$ [s] we have obtained the following values: (35), (40), (41).

The process can be represented as in Fig. 4

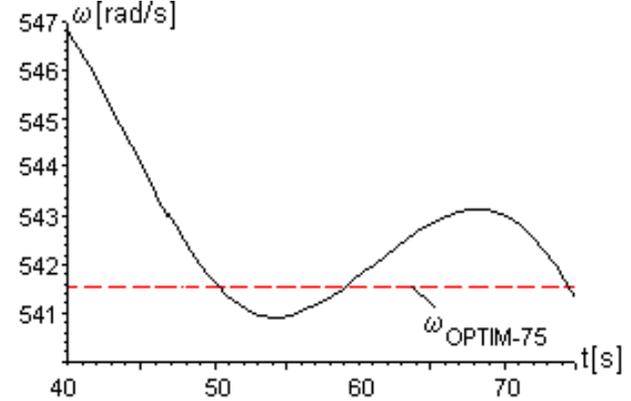


Fig.4. Time variation of MAS for $R=311.64$ [Ω]

Remark 3: It can observe that at $t=50$ [s] the system reach $\omega_{OPTIM-50} = 541.56$ [rad/s] and based on this remark we can prescribe the new value for the PMSG load and it isn't necessary to wait until $t=75$ [s].

In the same way, for $t=110$ [s], the results are:

$$\omega_{OPTIM-110} = 536.31 \text{ [rad/s]}$$

$$P_{PMSG-P-75} = 12650 \text{ [W]} \quad (42)$$

$$R_{PMSG-REQ-75} = 238.61 \text{ [\Omega]} \quad (43)$$

The process is represented in Fig. 5

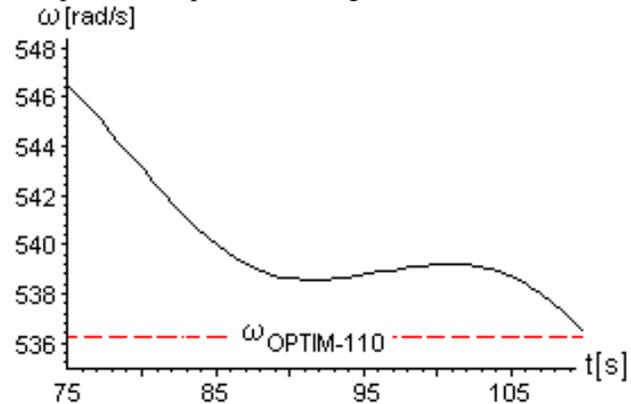


Fig. 5. Time variation of MAS for $R=238.61$ [Ω]

Remark 4: The control of the PMSG load has a dead-time of 35 [s], because the optimal load can be done only after processing the data from interval $\Delta t = 75+110$ [s]. The time variation of MAS with (REAL) and without (IDEAL) considering the dead-time is presented in Fig. 6.

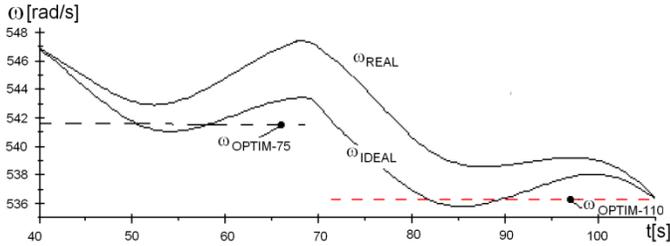


Fig.6. Time variation of MAS – ideal and real

The control algorithm

By measuring the wind speed, the optimal MAS can be calculated. Comparing the optimal MAS with the current MAS, the required power for the PMSG and, consequently the optimum DC link current are obtained.

The algorithm is presented below:

1. measure of wind speed and calculation of $\omega_{OPTIM-tk}$
2. measure MAS of PMSG and calculation the real kinetic energy
3. estimation of the captured wind energy
4. estimation of the kinetic energy, necessary to lead the system at $\omega_{OPTIM-tk}$
5. estimation of the energy from the PMSG to lead the system to MAS
6. calculation of medium PMSG power, corresponding to the energy estimated at 5.
7. calculation of the PMSG load from the power estimated at 6.
8. calculation of the PMSG power.

The value of the optimum DC link current is achieved by an appropriate control of the switches of the power electronic converter (Fig.7.)

The wind speed is measured using an anemometer. The optimum DC link current is calculated and, after that, the converter is controlled with the output value of the regulator R.

3.2. The relationship between the wind speed and the DC link current

The relationship is presented below:

$$U_{cc} \cdot I_{cc} = k_{cc} \cdot V^3 \tag{44}$$

$$I_{cc} = k_1 \cdot V^3 \tag{45}$$

Where $k_1 = WT+PMSG$ constant and $V =$ wind speed.

The constant k_1 is obtained from $I_{cc-OPTIM}$ for the values obtained at $t=75[s]$.

The DC link current is obtained from the wind speed, using the relationship:

$$I_{cc} = 4.0562 \times 10^{-3} \cdot V_{ECH}^3 \tag{46}$$

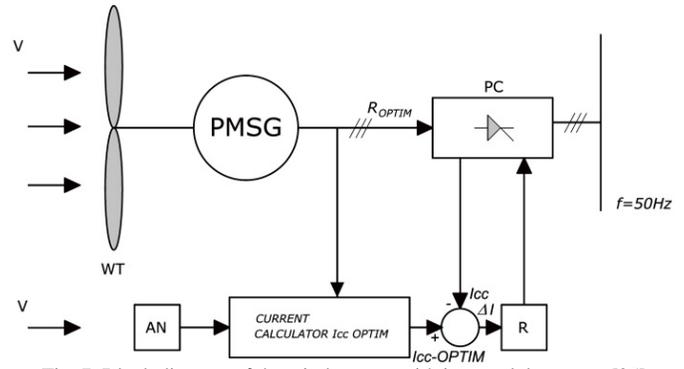


Fig. 7. Block diagram of the wind system with imposed dc current [25]

IV. CONCLUSIONS

The simulations presented in this paper have described the time evolution of the significant variables of process: current, speed, power, imposing the PMSG load. The best results are obtained by imposing the optimal value of load current, $I_{cc-OPTIM}$. By knowing the optimal value of the load current, the PMSG load can be adjusted so that the PMSG operates at the maximum energy. The speed variation of wind speed in time and the inertia value are two fundamental elements upon which the MPP operation. By prescribing the optimal DC link current, I_{cc} , from intermediate circuit of the converter, a simple and useful adjustment WT PMSG system can be achieved. Operation control method in the optimal energy of WT is based on the knowing of the I_{cc} value, which is determined by wind speed and momentary mechanical angular speed, MAS. By analyzing several cases were able to establish basic parameters leading to an optimal operation. By measuring the wind speed, the MAS, and calculation of the optimal load current, the operation in the energetically optimal region can be performed. The control algorithm based on energy balance measurements made by MAS and electrical energy, has been validated by simulations.

REFERENCES

- [1]. Babescu M, Borlea I, Jigoria Oprea D., "Fundamental aspects concerning Wind Power System Operation Part.2, Case Study" Medina Tunisia 2012 IEEE MELECON, 2012,25-28 March 978-1-4673-0783-3
- [2]. Babescu,M, Borlea I, Jigoria Oprea D, "Fundamental aspects concerning Wind Power System Operation Part.1, Matematical Models" Medina Tunisia 2012 IEEE MELECON, 2012 ,25-28 March,978-1-4673-0783-3
- [3]. M. Babescu, O.Gana, L.Clotea"Fundamental Problems related to the Control of Wind Energy Conversion Systems-Maximum Power Extraction and Smoothing the Power Fluctuations delivers to the Grid"OPTIM-13th International Conference, Brasov, Romania
- [4]. Babescu M, Borza I,Gana O., Lacatusu F., "Comportarea sistemului electroenergetic eolian la variatii rapide ale vitezei vântului" Producerea , transportul si utilizarea energiei, pp 11-24,Editura RISOPRINT Cluj-Napoca, 2010, ISSN 2066-4125.
- [5]. Babescu M, Boraci,R, Chioreanu C, Koch C, Gana O "On Functioning of the Electric Wind System at its Maximum Power" ICCO-CONTI 2010, Timisoara, Romania, May 27-29, 2010.
- [6]. Bej,A-Turbine de vânt-ISBN 973-625-098-9,Editura POLITEHNICA Timisoara,2003
- [7]. Barakati S.M, M.Kazerani, and J.D.Aplevich, "Maximum Power Tracking Control for a Wind Turbine System Including a Matrix Converter ", IEEE Trans. Energy Convers., vol. 24, no. 3, pp.705-713, September 2009

- [8]. Chen Z., Spooner E. – “Grid power with variable speed turbines”, IEEE Trans. Power Electron., vol. 16, no. 2, pp. 148-154, Jun. 2001
- [9]. El Aimani S., Francois B., Minne F., Robyns B. – “Comparativw analysis of control structures for variable speed wind turbine”, in Proc. CESA, Lille, France, Jul. 9-11, 2003,
- [10]. Gavris M.L. – “Dual Input DC-DC Converters for Renewable Energy Processing” – Teza de doctorat, feb. 2013, Univ. “POLITEHNICA TIMISOARA”
- [11]. Gertmar – Windturbines. Berlin, Germany: Springer-Verlag, 2000-
- [12]. Jeong H G, Seung R H, Lee K B – An Improved Maximum Power Point Tracking Method for Wind Power Systems – Energies 2012, 5, 1339-1354; doi:10.3390 /en5051339 energies ISSN 1996-1073, www.mdpi.com/journal/energies
- [13]. [13] Jiao S., Hunter G., Ramsden V., Patterson D. – “Control system design for a 20 KW wind turbine generator with a boost converter and battery bank load”, in Proc. IEEE PESC, Vancouver, BC, Canada, Jun. 2001, pp. 2203-2206
- [14]. Kim K.H., Van T.L., Lee D.C., Song S.H., Kim E.H. - "Maximum output Power Tracking Control in Variable-Speed Wind Turbine System Considering Rotor Inertisl Power", in IEEE transaction on industrial electronics, vol.60, no.8, august 2013, pp.3207-3217
- [15]. Koutroulis E, Kalaitzakis K – Design of a Maximum Power Tracking System for Wind-Energy-Conversion Applications-486 IEEE Transactions on industrial electronics, Vol. 53, No. 2, April 2006
- [16]. Luca D., Nichita C., Diop A. P., Dakyo B., Ceanga E. – “Load torque estimators for wind turbines simulators”, in Proc. EPE Conf., Graz, Austria, Sep. 2001
- [17]. Nishikata S, Tatsuta F - A New Interconnecting Method for Wind Turbine/Generators in a Wind Farm and Basic Performances of the Integrated System - IEEE Transactions on Industrial Electronics, vol 57, Nr.2, p.468-476, ISSN 0278-0046, feb.2010.
- [18]. Örs M – Maximum Power Point Tracking for Small Scale Wind Turbine With Self-Excited Induction Generator-CEAI, Vol.11, No.2, pp. 30-34, 2009, Technical University of Cluj-Napoca, Romania
- [19]. K K, Tiwari Dr. A.N – Maximum Power Point Tracking Of Wind Energy Conversion System With Synchronus Generator-International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 5, July - 2012 ISSN: 2278-0181 MMEEC Gorakhpur-273010
- [20]. Petrilă D.P. – Energy Conversion and Storage Control for Small Wind Turbine Systems – Teza de doctorat, feb 2013, Univ. “POLITEHNICA TIMISOARA”
- [21]. Petru T. – “Modeling wind turbines for power system studies”, Ph. D. dissertation, Chalmers, Goteborg, Sweden, Jun. 2003
- [22]. Quaschnig V. – Understanding Renewable Energy Systems, ISBN 1-84407-128-6, London Carl Hanser Verlag GmbH & Co KG, 2005.
- [23]. V.D. Müller, O.Gana, L.S. Bocîi, M. Popa, "The Leading Of The Eolian Power Systems In Order To Maximise The Power And To Flatten The Fluctuations Of The Generated Power." La Gestión De Los Sistemas De Energía Eólica Para Maximizar La Potencia Y Para Aplanar Las Fluctuaciones De La Energía Generada (Recibido E115.01 de 2012. Aceptado E123.09. De 2012) Facultad De Ingenierias-Universidad Antioquia-Columbia, Spain
- [24]. D. Vatau, F.D. Surianu, “Monitoring of the Power Quality on the Wholesale Power Market in Romania”, Proceedings of the 9th WSEAS International Conference on Electric Power Systems, High Voltages, Electric Machines, Genova, Italy, October 17-19, 2009, pp.59-64
- [25]. G.M. Erdodi, D.I. Petrescu, C. Sorandaru, S. Musuroi, “The determination of the maximum energetic zones for a wind system, operating at variable wind speeds”, ICSTCC Sinaia, 2014

Dual Approach to Complex Ecological System Analysis and Modeling

Migdat Hodzic (*), Mirsad Hadzikadic (**), Ted Carmichael (**), Suvad Selman (*)

(*) International University of Sarajevo, Sarajevo, Bosnia and Herzegovina

(**) University of North Caroline at Charlotte

mhodzic@ius.edu.ba (Contact Author), mirsad@uncc.edu, tedsaid@gmail.com, sselman@ius.edu.ba

Abstract - In this paper we present new, dual approach to analysis and simulation of a complex nonlinear ecological system of preys and predators, using classic nonlinear dynamic Lotka-Volterra mathematical model (LVM) in parallel with an Agent Based model (ABM), using model attributes description of the system. We propose to implement this dual approach using "mathematical" approach together with an "agent based" approach using appropriate modeling environments, such as Matlab and NetLogo. As the system models become more complex we aim at using both LVM and AMB to reinforce each other and check each other findings. This way the validity of the model and its usefulness would be greatly increased, and some long standing ecological paradoxes may be explained and qualified.

Keywords: *Mathematical Modeling; Agent Based Modeling; Lotka-Volterra Equation; Predator and Prey; Complexity and Stability; Structure*

1. INTRODUCTION

In analysis and simulation of complex ecological systems, a researcher often starts with a nonlinear Lotka Volterra model (LVM) of predator prey dynamic system [1]. The problem with this approach is that the LVM is very simplified model and apart from a detailed stability analysis [1], there are no real life complex ecological dynamic system models which are flexible and useful enough. Some of the reasons are (i) Lack of any general model build up methodology, (ii) Lack of any structural analysis of complex dynamical ecological models, and (iii) Very few results explaining some well know ecological paradoxes. We aim to address some of these important issues. In this paper, examples of various Single Prey Single Predator (SPSP) as well as Multiple Prey Multiple Predator (MPMP) models are introduced in a gradual way, from simple to more complex ones. Our goal is to gain insight into (i) Predator-prey population, (ii) Structural properties of the models, (iii) Understanding of stability in multispecies communities, and (iv) Improve usability, robustness and adaptivity of LVM ecological models. With this approach we aim to go towards analytical description of the key classic problems in ecology, such as (i) Paradox of the Plankton, (ii) Paradox of the Enrichment, (iii) Oksanen's description and trophic level numbers, and other general Complex Systems paradigms such as (iv) Adaptivity and (v) Emergence. We also compare LVM analytical stability results with simulated ABM results. We propose to take advantage of flexibility that ABM offers, and in doing so acquire key feedback to reinforce and improve nonlinear mathematics of the LVM as well. This way we can build very complex but usable predator-prey ecological models which are also mathematically tractable.

2. NONLINEAR MODEL LINEARIZATION

As a starting point, we can assume the most general non linear ecological model described as:

$$S: dX/dt = f[X(t),t] \quad (1)$$

Any well-behaved non linear system can be linearized around equilibrium points X^* of the function $f[X(t)]$. This approach works well close to equilibrium points. The other advantage is that there are well known theoretical stability results for linear complex systems [1,4,6,7].

Unfortunately, linearization may be very restrictive and limited in its usefulness, hence analysis of real nonlinear ecological predator-prey systems will produce more realistic results. But, nonlinear problems are not easy to deal with. We propose here a step-by-step build-up of nonlinear models which will allow us to better understand effects of nonlinearities and interconnections in multi species environments.

3. GENERAL ECOLOGICAL NONLINEAR MODEL

General ecological nonlinear model in the context of our interest in this paper is described by [1]:

$$S: dX/dt = A(t,X) X \quad (2)$$

where X is vector of (for example aquatic) species. The model in (2) is obviously a nonlinear one, but has an appearance of a linear system. The vector X may be as simple as a two dimensional vector (one pray, one predator), or it could consist of 10s and 100s of species arranged in some logical conglomerate of prey and predator species, all collected into the species vector X . Matrix $A(t,X)$ is a "community" matrix with its elements as nonlinear time-dependent functions $a_{ij}=a_{ij}(t,X)$, where "ij" indicates position in the matrix, i for the rows, j for the columns. In the case of X of dimension 2, matrix A is 2 by 2, and its elements are a_{11} , a_{12} , a_{21} , and a_{22} , and they describe self and cross interactions among the two species.

One of our goals is to find a practical way how to model elements of community matrix for a specific ecological system of some aquatic species (small and big fish).

4. SINGLE PREY SINGLE PREDATOR MODELS

Next level of simplification of the ecological model is embodied in the well known nonlinear Lotka-Volterra Model (LVM), which is just a special case of the model (2). Consult [1] for more details. For our purposes in this paper, we will illustrate LVM at first using second order model, with Single Prey Single Predator (SPSP) model. Following that, more complex models will be also given.

A. LVM Solution

Let us assume $X = [X_1, X_2]^T$, X_1 is prey species, X_2 is predator species. The classic LVM [1] is:

$$\begin{aligned} dX_1/dt &= X_1 (A_1 + A_{12} X_2) = A_1 X_1 + A_{12} X_2 X_1 \\ dX_2/dt &= X_2 (A_2 + A_{21} X_1) = A_2 X_2 + A_{21} X_1 X_2 \end{aligned} \quad (3)$$

which can also be written in a compact form as:

$$dX_i/dt = X_i (A_i + A_{ij} X_j) \quad (4)$$

where $i=1,2$ and j is different than i , with $j=1,2$. Here A_1 is the growth rate of the prey. Note that with $A_{12} = 0$ the prey population X_1 continues to increase exponentially, which is equivalent to the absence of any predator X_2 . With $A_{12} < 0$, predator X_2 will control prey population from growing exponentially. For the predator population, growth is dependent on $A_2 < 0$, the rate of predator removal from the system (either by death or migration), and A_{21} , the positive growth rate for predators. The solution to Equations 3 and 4 is periodic, with the predator population always following the prey. Fig. 1 gives an example from a typical SPSP LVM. We assumed constant values of positive A_1 and A_{21} , and negative growth rates A_{12} and A_2 . The other SPS models can be defined, such as positive A_2 and negative A_{21} for the predator, depending on the predator model. The key is to keep the basic model stable.

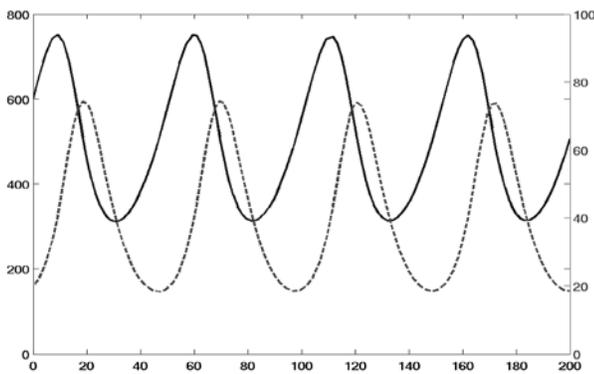


Figure 1. SPSP LVM Population Levels
(Prey Solid, Predator Dashed)

In terms of the general nonlinear model given in (2), and with no time dependency, the community matrix A is:

$$A(t,X) = A = \quad (5)$$

$a_{11}(t,X)$	$a_{12}(t,X)$
$a_{21}(t,X)$	$a_{22}(t,X)$

with:

$$\begin{aligned} a_{11} &= A_1 \\ a_{12} &= A_{12} X_2 \\ a_{21} &= A_{21} X_1 \\ a_{22} &= A_2 \end{aligned} \quad (6)$$

General LVM stability results are given in [1].

B. ABM Solution

The original LVM Equations 3 and 4 are very simple ecological model. They assume, for example, unlimited food available to the prey, and so the prey (and predator) growth rates are limited by corresponding “growth”

coefficients. In these equations, the growth coefficient is A_1 for the prey and A_{21} for the predators. As a comparison, in ABM model, the growth rate for both populations can be determined by how successful they are at finding food. This can be modeled as a stochastic process which averages out to a stable rate across each population, hence corresponding to large extent to LVM approach, in the limit. Other effects can be incorporated as well, per modeling flexibility of ABM approach. Fig. 2 gives a typical agent based snapshot of simulation control window. Various model attributes are easily defined. For example, the predators are not consumed, but they disappear from the simulation at a constant rate by reaching the end of their programmed lifetime. This is represented by negative A_2 . Their population increases linearly based on the prey consumption. This is proportional to the number of both populations, and thus represented by $A_{21} X_1 X_2$.

In the ABM, when the food is increased initially, both growth values, A_1 and A_{21} , temporarily go out of equilibrium and they both increase initially. In the steady state, the prey growth rate remains constant, because their population growth is offset by increased predation, due to an increase in the predator population. The predator population, however, stays elevated, and so increased competition means that their growth rate returns to the original value, for initial food availability. Note that the coefficient A_2 , the rate of predator removal (death or migration) from the system model, is determined by the predator attribute *age* and a limited lifetime for each individual. The prey also has an attribute for age, but in practice, very few fish die of old age. This is particularly true at higher levels of resources, because their average age drops as a consequence of fish being born faster while their population remains stable. It is this last fact that seems to cause the system to eventually become unstable, at very high levels of resources. We will compare this with general stability results in [1], in our future paper.

There is a limit to how quickly fish can be consumed after being spawned. As the limit is reached, endogenous spatial heterogeneities appear in time with increased volatility in both populations. Per Fig. 2, ABM gives lots of flexibility to model the system, but essentially gives no analytical insight and the solution such as the case with LVM. That is the essence of our dual approach here, i.e.

- (i) Use ABM for its flexibility, and
- (ii) LVM for its mathematical elegance

This way we can learn about using ABM to improve or change LVM. One obvious idea is to make the LVM model in Equations 3 and 4 times varying to some extent (Section D). The next model illustrates adding a “crowding” term in LVM which corresponds to species dynamic when disconnected from the other specie(s).

C. Crowding Effect In LVM

The extended LVM with crowding effect is as follows:

$$dX_i/dt = X_i (A_i + \sum A_{ij} X_j) \quad (7)$$

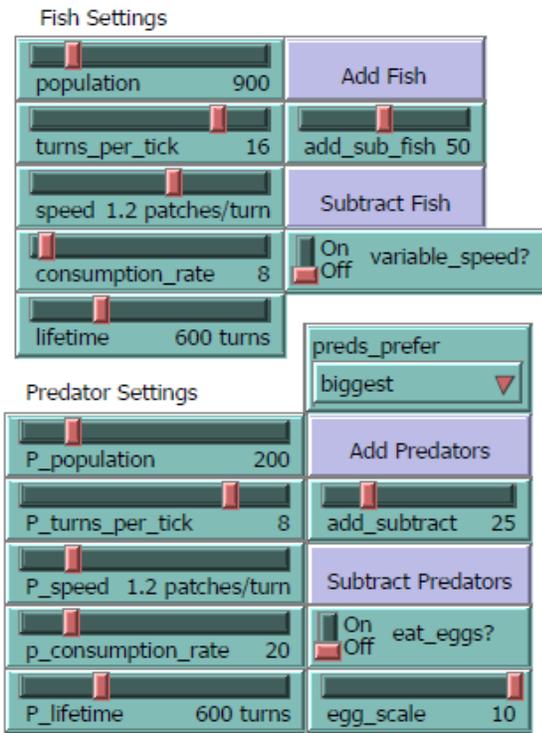


Figure 2. SPSP ABM simulation control window

where $i = 1,2$ and sum \sum is over $j = 1,2$. This would be equivalent to prey self multiplication without predator in dX_i/dt . In this case community matrix (Equation 5) elements are:

$$\begin{aligned} a_{11} &= A_1 + A_{11}X_1 \\ a_{12} &= A_{12} X_1 \\ a_{21} &= A_{21} X_2 \\ a_{22} &= A_2 + A_{22}X_2 \end{aligned} \quad (8)$$

In this model, A_{12} and A_{21} are negative, with newly introduced A_{11} and A_{22} positive. Another option is to go back to our original ABM in Equations 3 and 4 and simulation of Section B. In that case the last two equations in (8) change to:

$$\begin{aligned} a_{21} &= A_{21} + A_{22} X_2 \\ a_{22} &= A_2 \end{aligned} \quad (9)$$

with A_{21} and A_{22} positive and A_2 negative. The key is not to allow either model to let the predator grow out of control (become unstable). When using ABM simulation, the "crowding" effects can be implemented according to options in Fig. 2, or by adding new options and additional model attributes.

Next step is to accommodate time varying community matrix. Again, the ABM model of Fig. 2 can accommodate this by simple addition of proper agent model attribute which translates easily in to LVM equation for $A=A(t)$. We can also add dependency on populations themselves, i.e. $A=A(t,X)$. That is discussed next.

D. Time Varying LVM Community Matrix

The time varying LVM in general is:

$$dX_i/dt = X_i [A_i(t,X) + \sum A_{ij}(t,X) X_j] \quad (10)$$

where $i=1,2$ and sum \sum is over $j=1,2$. The Equation 10 is an extension of Equation 7, where we added time varying and population dependencies in the model. This can be presented in the compact form as:

$$dX/dt = A(t,X) X \quad (11)$$

with:

$$A = A(t,X) = \quad (12)$$

$a_{11}(t,X)$	$a_{12}(t,X)$
$a_{21}(t,X)$	$a_{22}(t,X)$

and for example:

$$a_{11}(t,X) = A_1(t,X) + A_{11}(t,X) X_1 \quad (13)$$

similarly for the rest of the coefficients in (8). Note that community matrix elements are functions of both X_1 and X_2 . This will give us lots of freedom in modeling dynamic of two interconnected species. The modeling should be done in individual steps (coefficient by coefficient) so we have full understanding of making even the simplest change. Both ABM and LVM approaches to compare and simulate accordingly, follow.

Example 1. Coefficients only functions of time, not of X , i.e. in (13), we have:

$$a_{ij}(t,X) = a_{ij}(t), i,j=1,2 \quad (14)$$

Example 2. Coefficients only functions of X , not of time, i.e. from (8,13), we have:

$$a_{ij}(t,X) = a_{ij}(X), i,j=1,2 \quad (15)$$

Example 3. Coefficients functions of local populations X_1 or X_2 only, but not of time, i.e.

$$a_{ij}(t,X) = a_{ij}(X_j), i,j=1,2 \quad (16)$$

where we assumed local dependencies only, for example $a_{11}(X_1)$ is function of X_1 and not of X_2 , etc. Obviously we can have more complicated case such as:

Example 4. Coefficients only functions of X_1 and/or X_2 but not of time, i.e.

$$a_{ij}(t,X) = a_{ij}(X_j), i \neq j, a_{ii}(t,X) = a_{ii}(X_j, X_j), i,j=1,2 \quad (17)$$

where we left the "crowding" coefficients functions of only their corresponding specie population.

Finally, we introduce time and have the following time varying version of Example 4.

Example 5. Coefficients functions of time as well as of X_1 and/or X_2 , i.e.

$$a_{ij}(t,X) = a_{ij}(t, X_j), i \neq j, a_{ii}(t,X) = a_{ii}(t, X_j, X_j), i,j=1,2 \quad (18)$$

As we develop complex LVM and ABM, our approach here is to follow the above formulas in implementing LVM and ABM to model corresponding features into both models. This way we will be able to precisely interpret

every step of the two models. For example, in Example 5 earlier, we would agree on what does $A_{12}(t, X_2)$ mean in terms of t and X_2 , similarly for other coefficients. That is then modeled in ABM via appropriate attributes of Fig. 2. As indicated bellow, Reference [1] has an extensive analysis of stability of LVM, which would be corroborated with carefully designed experiments in ABM simulations.

5. ABM SPSP SIMULATION

In this Section we summarize and discuss various details of ABM SPSP simulation, based on setup of Fig. 2, which can address and attempt to simulate various models and Examples of Section 4.

Simulation [9] was run for 1000 time step chunks which generate ‘counts’ for all the variables at each step, i.e. (i) New prey or predators introduced, (ii) Food consumed by prey or prey consumed by predators, (iii) Predator deaths (due to end of lifetime), and (iv) Population sizes. Without any changes during the 1000 time steps, the simulation (under certain settings) is stable, and each variable is averaged across the recorded time. Each food level had plenty of time to stabilize. The data from 1001 thru 2000 were used only for 0.20 food (20% chance of food growth per patch, per step); then changed to 0.30, ignored the next 1000 steps and used 3001 thru 4000 for 0.30, etc., for 0.40.

In the simulation the predators are assumed of the same size, but the prey grows larger as they eat, starting at 0. Hence, the measurement below of “biomass” for the prey as indicated in Table 1.

TABLE 1. ABM SIMULATION SUMMARY

Food Rate	Fish Average	Fish St. Deviation	Predator Population	Fish Consumption
0.20	1154.73	33.56	158.97	4562
0.30	1140.39	39.66	329.09	6841
0.40	1140.45	36.27	462.10	9120

Food Rate	Predator Consumpt.	New Fish	New Predators	Predator Death
0.20	61.41	74.19	2.124	2.117
0.30	125.88	144.2	4.392	4.391
0.40	177.87	225.7	6.158	6.179

We can use the simpler fish model where, like the predators, each fish can be exactly the same size and would not change. The effects of fish size can be easily removed from the model. So, for example, to calculate A_{12} we don’t really need the size of the fish, only the population size. Various options such as “big fish are easier to catch” can be also implemented in ABM. This can be programmed into the model, for example under the option labeled “Predator Preferences” which can be added to Fig. 2.

In the ABM simulation, we get the following behaviors:

- a) Completely stable
- b) Oscillating-but-stable, and
- c) Oscillating-but-unstable.

In the stable settings neither the fish nor the predator population changes. Since $dX_1/dt = 0$, we can (presumably) say that in terms of LVM, Equations 3 and 4:

$$A_1X_1 = A_{12}X_2X_1 \tag{19}$$

which corresponds to prey growth rate equal to predator consumption rate of the prey. Similarly $dX_2/dt = 0$, hence:

$$A_2X_2 = A_{21}X_1X_2 \tag{20}$$

which simply means that the predator growth rate equals the predator death rate. Per Equation (19), A_1 , the growth rate of prey, irrespective of the number of prey, is equal to the consumption rate of predators times the number of predators. The death rate for the predators, in terms of the number of prey, irrespective of the number of predators, is equal to the consumption rate of preys times the number of preys, as in (20) above. In addition to that, in Table 1, we see that the last two columns, new predators and predator deaths (A_2X_2), are approximately equal, corresponding to the stable state, irrespective of the food rate.

Calculating the above during out-of-equilibrium periods (transients) is trickier, such as right after prey food is increased from 0.20 to 0.30. It becomes trickier when other nonlinear effects emerge. For example, increasing the food produces spatial inhomogeneities, i.e. there are areas where food or prey becomes scarce for a time, and the consumption rates vary across the simulation space. This emerges from the ABM simulation itself. However, these can be modeled by a random “jump” to predator and prey movement. Hence each turn, fish in the simulation jumps to new, random spatial coordinates. This particularly affects the predator numbers, as they consume prey faster only while the system is out of equilibrium. Once the system stabilizes (steady state) with a higher predator population, then each particular predator consumes roughly the same amount of fish as before the change.

This is the key dynamic of the ABM: the predators are more in number but have harder time finding prey, and the prey (more food) are more in number, making it easier for the predators to find prey. These two dynamics balance out, so that the equilibrium consumption rate per predator is the same. These numbers would grow in a kind of an S-curve, right after the food for the prey is increased, so that the number of predators would start to grow, then grow faster, then grow slower, and then stabilize at a new, higher level. The way to measure this would be to divide the total predator consumption rate by the total number of new predators. Per the assumptions of LVM, the fish growth is proportional to the amount of fish-food consumed, and the predator growth is proportional to the amount of fish (prey) consumed. See Table 2 bellow.

TABLE 2. PREDATOR CONSUMPTION / NEW PREDATORS

Food Rate	Predator Consumption (Fish Biomass)	New Predators	Quotient
0.20	61.41	2.124	28.91
0.30	125.88	4.392	28.66
0.40	177.87	6.158	28.88

Here, A_{21} , the predator growth rate, is a constant, even when the food available to the fish increases. So each predator eats same amount before and after the prey food is

increased. The prey growth rate, A_1 , increases with more food, but the prey population size does not, because prey is consumed faster by the predators. As expected, these are all about the same size. The predator consumption rate is not the number of *fish* consumed, but the total *biomass*, which controls predator numbers. Hence the predators eat more, but there are more predators, so one eats about the same, after the new equilibrium. The consumption growth is stable initially, Fig. 3,4. In Fig. 5 we see an S-curve which flattens out, when the new equilibrium is reached.

6. MULTIPLE PREY MULTIPLE PREDATOR MODEL

We now extend the SPSP LVM (Section 3) to a general nonlinear model to include multiple species (MPMP):

$$dX_i/dt = X_i [A_i(t,X) + \sum A_{ij}(t,X) X_j] \tag{20}$$

where $i = 1, 2, \dots, n$, and sum \sum is over all $j = 1, 2, \dots, n$.

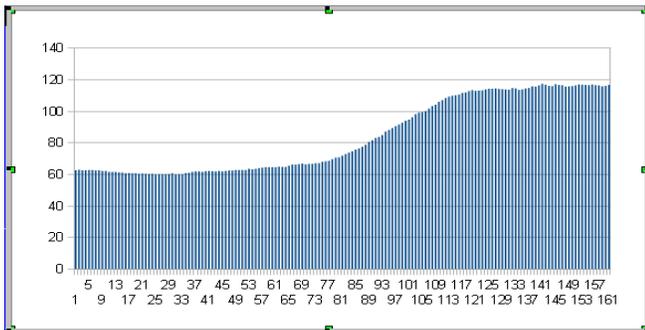


Figure 3. Total Predator Consumption Rates

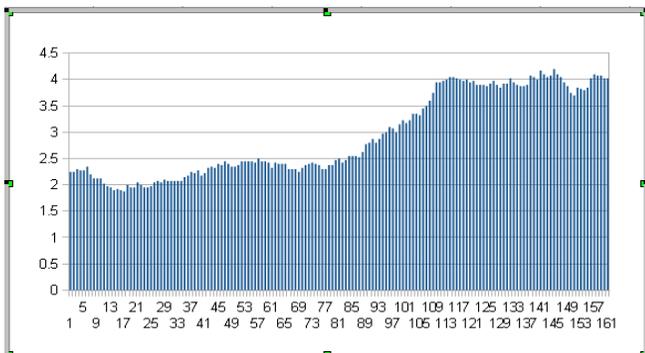


Figure 4. Total New Predator Growth

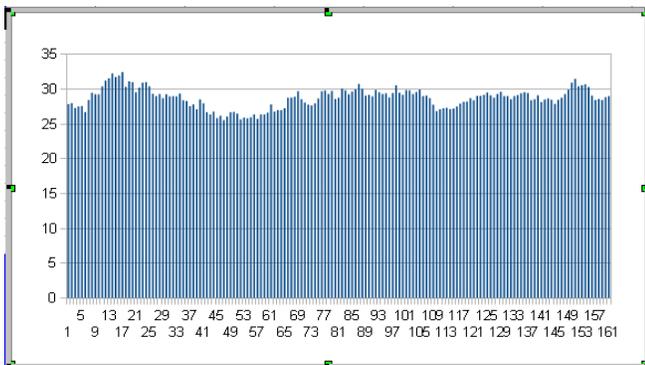


Figure 5. Predator Consumption / New Predator Growth

We can model 2 preys 1 predator, 4 preys 2 predators, 10 preys 3 predators, etc., hence building up complexity of the

LVM's. Here are some specific examples, where we continued from Example 5 (Section 4D) and increased the number of species. This may be influenced by a specific multispecies situation, such as an aquatic fish environment with a variety of preys and predators involved.

Example 6. Two preys one predator, coefficients time and functions of X_1, X_2 , and X_3 or the total vector X , i.e.:

$$\begin{aligned} a_{ii}(t,X) &= A_i(t,X) + A_{ii}(t,X_i) X_i \\ a_{ij}(t,X_j) &= A_{ij}(t,X_j) X_j \end{aligned} \tag{21}$$

where $i,j=1,2,3, i \neq j$, and X_3 is a predator. In compact form, community matrix $A(t,X)$ is now represented as 3x3 array:

$$A(t,X) = \begin{bmatrix} a_{11}(t,X) & a_{12}(t,X_2) & a_{13}(t,X_3) \\ a_{21}(t,X_1) & a_{22}(t,X) & a_{23}(t,X_3) \\ a_{31}(t,X_1) & a_{32}(t,X_2) & a_{33}(t,X) \end{bmatrix} \tag{22}$$

Example 7. Four preys (species 1,2,4,5) and two predators (3,6), for simplicity, and coefficients functions of time as well as of X_i , or the total vector X :

$$\begin{aligned} a_{ii}(t,X) &= A_i(t,X) + A_{ii}(t,X_i) X_i, i=1,2,3,4,5,6 \\ a_{ij}(t,X_j) &= A_{ij}(t,X_j) X_j, i,j=1,2,3,4,5,6; i \neq j \\ a_{ij}(t,X_4) &= 0, i=1,2,3; j=4,5,6 \\ a_{ij}(t,X_4) &= 0, j=1,2,3; i=4,5,6 \end{aligned} \tag{23}$$

The community matrix $A(t,X)$ is now 6x6 array:

$$A(t,X) = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 & 0 & 0 \\ a_{31} & a_{32} & a_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & a_{44} & a_{45} & a_{46} \\ 0 & 0 & 0 & a_{54} & a_{55} & a_{56} \\ 0 & 0 & 0 & a_{64} & a_{65} & a_{66} \end{bmatrix} \tag{24}$$

and it consists of two decoupled predator prey systems. Any of the zero coefficients a_{ij} indicates lack of influence of j -th specie to i -th specie. Assuming that predators can prey on all of the species, but not on each other, we have:

Example 8. Community matrix $A(t,X)$ is still 6x6, with less 0 elements (“*” are also 0 for this Example):

$$A(t,X) = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 & 0 & a_{16} \\ a_{21} & a_{22} & a_{23} & 0 & 0 & a_{26} \\ a_{31} & a_{32} & a_{33} & 0 & 0 & * \\ 0 & 0 & a_{43} & a_{44} & a_{45} & a_{46} \\ 0 & 0 & a_{53} & a_{54} & a_{55} & a_{56} \\ 0 & 0 & * & a_{64} & a_{65} & a_{66} \end{bmatrix} \tag{26}$$

If predators prey on each other, then we have Example 9:

Example 9. Community matrix $A(t,X)$ is still 6x6, with even less 0 elements, “*” are a_{36} and a_{63} respectively.

Example 10. Here we have an overlapping model, where two almost decoupled specie communities share a common four (**boldfaced**) elements:

$$A(t,X) = \tag{27}$$

a_{11}	a_{12}	a_{13}	0	0	0
a_{21}	a_{22}	a_{23}	0	0	0
a_{31}	a_{32}	a_{33}	a_{34}	0	0
0	0	a_{43}	a_{44}	a_{45}	a_{46}
0	0	0	a_{54}	a_{55}	a_{56}
0	0	0	a_{64}	a_{65}	a_{66}

This model can be handled by an approach in [5] where the model is “expanded” to decouple it effectively. Finally we add environmental effects [1] into LVM by:

$$S: \frac{dX}{dt} = A(t,X) X + B(t,X) \tag{28}$$

where $B(t,X)$ models external effects of the environment (food, space, temperature). Let us look at Example 6, and add environmental vector $B(t,X)$. We obtain community matrix $A(t,X)$ in (22), with species vector $X=[X_1, X_2, X_3]^T$ and the corresponding environmental vector is:

$$B(t,X) = [B_1(t,X), B_2(t,X), B_3(t,X)]^T \tag{29}$$

Or even simpler case, where each environmental component depends only on individual specie, i.e.

$$B(t,X) = [B_1(t,X_1), B_2(t,X_2), B_3(t,X_3)]^T \tag{30}$$

As it was discussed in Section 4B and 5, ABM introduced food supply into the model and the above environment vector is the right place to introduce the food supply, as a **control** input into the LVM (future work research subject).

As the community matrices become larger and more complex, we note that there are certain structural properties in the way "0" elements are placed. This is calling for approaches described in [2,4,5,6] which take advantage of special structures to simplify calculations and expose key structural properties of the models. There are elements of "overlapping" components in community matrices, which can be "expanded and contracted" [5] in building effective controls in multispecies communities. As the number of species grow, smart shuffling of the position of species in the vector X may produce hierarchical structure of community matrix $A(t,X)$ [4], producing much simpler controls and simpler stability analysis, as the overall community matrix is split into subsystems (agents) hierarchically interconnected.

7. STABILITY AND COMPLEXITY

There are some key existing mathematical results related to LVM which can be used and which can accommodate multi-species modeling and stability in particular [1]. They give regions of stability estimates and point to specific reasons for instability and balance between stability and complexity. These regions can be tested using both LVM and ABM approaches which will add a measure of confidence and practicality to the stability results. As several ecology researchers (not mathematicians) pointed out in literature, there seems to be a balance in competing multi-species environments between numbers of inter connections among the species versus interconnection strengths. Our (obvious) mathematical conjecture is:

If we denote by N number of interconnections for a given species (in a multi species environment) and by I their intensity, then:

$$N \text{ times } I = \text{Constant} \tag{31}$$

where equality sign is just a measure of closeness of two sides of the expression. We could rephrase this intuitive notion and add stochastic measure by using Expected Value $E(\)$ as:

$$E(N \text{ times } I) = \text{Constant} \tag{32}$$

where intensity I may be represented by some norm. In this context the LVM would need to be expressed in a stochastic form by adding certain stochastic processes either in random parameters in the community matrix elements, or as an additive colored or white noise process to the model itself. We will consider this in future work.

8. CONCLUSION

In this paper we set the scene for a robust and effective, dual model based approach (LVM, AMB) to build simple-to-complex predator-prey ecological models and examples of Single Prey Single Predator (SPSP) as well as Multiple Prey Multiple Predator (MPMP) models. This approach aims to produce practical results which can be used in real life ecological problems, and to better understand classic notions in multi-species models, as (i) Paradox of the Plankton, (ii) Paradox of the Enrichment, (iii) Oksanen's description and tropic level numbers, and other general Complex Systems paradigms such as (iv) Adaptivity and (v) Emergence. Our dual approach relies on methodology of step-by-step model build-up and reinforcement using two very different approaches, i.e. “mathematical” LVM and “ad-hoc” ABM. Proposed approach adds to the overall rigorosity of the obtained results and their validation and interpretations, by meticulously checking and comparing results of ABM and LVM as more and more complex models are built. In the research which follows, we will (i) Explore specific examples from this paper using appropriate computing environments such as Matlab, Mathematica, NetLogo, and (ii) Compare theoretical LVM stability results [1] with ABM modeling.

9. REFERENCES

- [1] D. D. Siljak, Large-Scale Dynamic Systems, Stability Structure, North Holland, New York, 1978.
- [2] M. I. Hodzic, “Some Extensions to Classic Lotka - Volterra Modeling For Predator Prey Applications”, SEJSC, Vol. 3/1, 2014.
- [3] A. I. Zecevic and D. D. Siljak, Control of Complex Systems, Structural Constraints and Uncertainty, Springer, Berlin, 2010.
- [4] M. I. Hodzic, Estimation of Large Sparse Systems, Chapter 3, Stochastic Large-Scale Engineering Systems, edited by Spyros G. Tzafestas, and Keigo Watanabe, Marcel Dekker, N. York, 1992.
- [5] M. I. Hodzic, R. Krtolica and D. D. Siljak, A Stochastic Inclusion Principle, Chapter 30, Differential Equations, Stability and Control, Edited by Saber Elaydi, Marcel Dekker, New York, 1991.
- [6] M. I. Hodzic and D. D. Siljak, “Decentralized Estimation and Control With Overlapping Information Sets”, IEEE Transactions on Automatic Control, January, 1986.
- [7] D. D. Siljak, Decentralized Control of Complex Systems, Academic Press, San Diego, 1991.
- [8] M. Hadzikadic, T. Carmichael and C. Curtin, “Complex Adaptive Systems and Game Theory: An Unlikely Union”, Wiley, 2010.
- [9] M. Hadzikadic and T. Carmichael, UNCC, USA, Private Communications, 2010-2013.

Performance estimate for a Proton Exchange Membrane Fuel Cell: Sensitivity Analysis aimed to Optimization

Enrico Testa, Paolo Maggiore, Lorenzo Pace and Matteo D. L. Dalla Vedova

Abstract— A fuel cell based system performance is basically determined by the amount of current density the stack is able to produce, given a well-defined quantity of reactants flowing through it. Starting from a Proton Exchange Membrane fuel cell (PEMFC) distributed parameters model, considering all the aspects influencing the cell behavior, a Multidisciplinary Design Optimization (MDO) process based on a surrogate model is presented. A Monte Carlo Simulation approach is chosen to perform a sensitivity analysis to estimate the effects of key parameters on performances. This analysis allows the definition of a ranking Pareto plot to operate design variables reduction, decreasing the problem complexity and increasing the orthogonality of the input design matrix. The main purpose is to find a suitable and validated method able to reduce the time expense for a complete simulation, so to originate useful input to a multi-disciplinary design optimization.

Keywords—MDO, PEM fuel cell, Surrogate model.

I. INTRODUCTION

IN a quite complex and always growing context, Proton Exchange Membrane Fuel Cell (PEMFC) technology is gaining increasing interest in the automotive and aerospace industry in the last decade, mainly due to its environmental sustainability. Many industrial and academic studies are carried out in this field, specifically implementing PEMFC Computational Fluid Dynamics (CFD) models and experimental characterization. Some CFD models gained interest and diffusion during the last decade thanks to the improvements in computer science technologies and the consequent reduction of computational time, enabling even more robust and complex models. The use of numerical modelling allows a great flexibility in the design and analysis of fuel cells [1].

E. Testa is with the Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino, Corso Duca degli Abruzzi, 24 - 10129 - Torino, ITALY. (e-mail: enrico.testa@polito.it).

P. Maggiore is with the Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino, Corso Duca degli Abruzzi, 24 - 10129 - Torino, ITALY. (e-mail: paolo.maggiore@polito.it).

L. Pace is with the Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino, Corso Duca degli Abruzzi, 24 - 10129 - Torino, ITALY. (e-mail: lorenzo.pace@polito.it).

M. D. L. Dalla Vedova is with the Department of Mechanical and Aerospace Engineering (DIMEAS), Politecnico di Torino, Corso Duca degli Abruzzi, 24 - 10129 - Torino, ITALY. (corresponding author to provide phone: +390110906850; e-mail: matteo.dallavedova@polito.it).

An extensive bibliographical review has been carried out so to retrieve the most affordable solution, both on the side of PEMFC simulation [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24] and on the side of MDO [25, 26, 27, 28], so to understand the main needs of the two main involved topics. The fuel cell model used here is referred to [29] and validated with empirical values given in literature. The model here used considers all the most important physical aspects involved, and it is based on previous CFD PEM fuel cell models available in literature. The model is implemented in *CD-adapco Star-CCM+* software environment, with an extensive use of user-defined functions. The main characteristics of this model are a 3D simulation domain comprising both fluid and solid regions, a steady-state solution, the adoption of a multi-component gas and non-isothermal conditions. The flow considered is single-phase. The basic equations for the computation of the fluid flow, the diffusivity of the reactants and the ionic conductivity of the membrane are the same that can be found in [22]. The electrochemical model uses the standard electrochemical laws implemented in [18]. The main difference consists in using an arcsine function instead of a logarithmic one for the electrochemical activation losses. The presence of liquid water and its effects on the cell performance (occlusion of catalyst reaction sites and flooding phenomena) are considered despite the single-phase presence. This approach was based on Dawes et al. [18]. The liquid water presence and quantity is calculated from the value of relative and absolute humidity, and the electrochemical and fluid-dynamics performances are scaled (degraded) based on liquid water calculated in each cell of the computational domain. The level of liquid water presence is quantified with the saturation (s , dimensionless) value [18]. The phase-change of vapor into liquid water is considered (imposing gas sinks) and modelled as in the ANSYS FLUENT fuel cell modules manual [30]. The geometrical domain simulated comprised only a single channel of the cell to limit the computational cost. The presence of the other fuel cell channels can be also simulated varying the starting value of the saturation variable. In this model, the simulation comprises not only the membrane electrode assembly (membrane, catalyst layers and gas diffusion layers), but also the fluid channels, the solid bipolar plates (affecting the heat transfer) and the cooling water flowing on the opposite side of the plates.

Together with the indirect liquid water presence simulation, the other main aspect differentiating this model from others is the reduction in the porosity of the gas diffusion layers given by the clamping pressure of the stack, as discussed in [13]

II. METHODOLOGY

A design space evaluation was performed considering the performance of the fuel cell from a fluid dynamics and electrochemical point of view. A set of parameters was selected and then split into two main sets.

The **boundary conditions values** (first set of design variables), also defined "**uncontrollable input noises**" or "**noise factors**" [31, 32], are reported below:

Cathode exchange current density, i_{0c} : the exchange current density is an important electrochemical parameter related to the kinetics of the chemical reactions. This variable depends upon many physical and electro-chemical factors, as the noble metal particles used, their shape and distribution over the catalytic surfaces and the micro-structure of the supporting surfaces. In the model, it is defined for both the cathode and the anode sides. This variable is usually measured in A/cm^2 . The higher its value, the faster the chemical reactions. A quicker chemical reaction has the direct effect of lowering the detrimental voltage losses, since it implies a lower amount of energy absorbed by the reaction itself (in the form of a voltage loss), improving the power output. The cathode exchange current density for a PEM fuel cell is usually in the range of $0.01 - 5 A/cm^2$, while the anodic reaction exhibits usually an exchange current density of about $1000 - 3000 A/cm^2$ [3]. From an electrochemical point of view, the cathode exchange current density produces the well-identifiable initial voltage drop at very low current densities. The initial voltage drop translates the whole fuel cell polarization curve into lower voltage values, i.e. it reduces the overall cell efficiency. Therefore, this control factor is expected to have a strong influence on the cell performance.

Anode exchange current density, i_{0a} .

Condensation rate, r_{cond} . The condensation rate is a gain factor (measured in $1/s$) directly related to the kinetics of water vapour condensation into liquid form. This value is usually defined in the range of $100 - 200 1/s$ by commercial software, e.g. ANSYS Fluent [30] for the simulation of generic multiphase flows contemplating a transition from vapour to liquid form. Fluent PEMFC model sets this value to $100/s$.

Saturation coefficient, sat_{rate} : this parameter is another gain factor used in the definition of the saturation variable (s), implemented for simulating major or minor quantities of liquid water presence inside the porous media. It is defined as the ratio of volume occupied by liquid water divided by void volume available within the dry porous structure of the Gas Diffusion Layer (GDL). When simulating a portion of fuel cell, it allows the user to take into account the presence of the whole cell, i.e. the liquid water produced by the part of the cell that is not really considered in the simulation can be modelled by assuming a suitable value of the sat_{rate} .

The effects of the presence of liquid water are here considered. On the other hand, the **tuning parameters**, also defined "**control factors**", are summarized below:

Anode inlet gas temperature, T_a : the temperature of the gas mixture entering the cell at the hydrogen side.

Anode inlet relative humidity, Rh_a : the relative humidity of the gas mixture entering the cell at the hydrogen side. In case of PEMFCs, the polymer membrane requires high level of humidity to operate properly as electrolytic element of the cell. Despite the cell produces liquid water as a chemical by-product at the cathode side, it is often not sufficient to guarantee a proper membrane operation. For this reason, it is often required to inject hydrogen at a high level of humidification for medium-large fuel cell stack.

Cathode inlet gas temperature, T_c : the temperature of the gas mixture (H_2 and H_2O) entering the cell at the oxygen side.

Cathode inlet relative humidity, Rh_c : the relative humidity of the gas mixture (O_2 , N_2 and H_2O) entering the cell at the oxygen side.

Compression (of the GDL), $compr$: the effects of the torque applied to clamp the stack. The clamping force, required to prevent reactants leakage and a good contact between the electric conductive parts, has the counteracting effect of reducing the porosity of the gas diffusion layers, directly reducing the void volume available to the reactants. In the model, the gas permeability and diffusivity are reduced as function of the dry porosity of the GDL influenced by the stack clamping pressure. The reduction in electrical contact resistance between catalyst, GDL and electrodes given in case of stronger clamping forces are not considered in this model.

Geometrical parameters (gas channel width, gas channel length, etc.) are defined as "**controllable inputs**" since their uncertainty level can be controlled during the manufacturing process [5, 31]. They are not involved in the presented sensitivity analysis, since this study is done for a fixed fuel cell geometry. The design space evaluation is often performed through the use of a Design of Experiments (DoE) technique.

The advantage of using a DoE consists in a maximum amount of knowledge gained with a minimum expense of numerical trials. Due to the fact that analysis processes are often time consuming, an efficient exploration of the entire design space requires a systematic samples distribution. The objective is to get many representative details of the correlation between system response and design parameters, while at the same time minimizing the number of design evaluations [27, 31, 32, 33, 34, 35]. Several strategies can be used to generate appropriate samples [12, 33], namely Monte Carlo Simulation (MCS), Latin Hypercube Sampling (LHS),

Optimal Latin Hypercube Sampling (OLHS) and Factorial Designs.

According to literature references [35] and thanks to the low complexity of this model, the MCS approach is used here. For the purposes of this work, a Sample Random Sampling (SRS) technique is used [26, 28, 35], consisting in generating random values according to a certain distribution.

This technique has been preferred by the authors because it gives an absolutely random distribution. As a result, this technique generates random sample points instead of dividing the distribution into N intervals of equal probability. The major drawback of this approach is that the design points may be clustered in some regions of the design space whereas other parts could be almost unexplored. To avoid this situation, (i.e., to achieve an almost even distribution of the design points), a significant number of simulations is required. As a consequence, this method should be used just in case of fast running analyses, while other algorithms should be considered for a complete covering of the design space. A uniform Probability Density Function (PDF), instead of a common Normal one, is adopted to model the random behavior, obtaining in this way a matrix containing the generated values of input parameters. The choice of a uniform PDF is motivated by the fact that a sensitivity analysis is performed evaluating all the values the parameters could assume, without having values with different likelihood, in a range included between upper and lower bounds. The PDF is also allowed thanks to the absence of geometrical parameters considered. Moreover, selected DoE techniques have to ensure the orthogonality of the generated matrix of design variables (i.e. its transpose is equal to its inverse) to ensure a good fit of meta-models [36]. An advantage of using orthogonal design variables as a basis for fitting data is that the inputs can be decoupled in the analysis of variance [25].

Orthogonality implies the estimates of the effects are uncorrelated, where any pair of independent variables is linearly independent. The most familiar measure of dependence between two quantities is the **Pearson product-moment correlation coefficient** ($\rho_{X,Y}$) also called Pearson's correlation [37, 38]. It is obtained by dividing the covariance of the two variables by the product of their standard deviations.

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad (1)$$

The larger the correlation, the less independent the parameters and the less orthogonal the design matrix is. If the design matrix is not orthogonal, a coupling exists in the matrix, so that the interaction effects of independent variables are not distinguishable.

As a second step, a sensitivity analysis was done to evaluate which parameters have the major weight on the system response. This was done thanks to the *iSight* software, providing some useful visual tools. After performing the DoE, a 2nd order polynomial was chosen as approximated function. The function fits the responses on a discrete set of samples after calculating the function coefficients thanks to the least square method. The unknown model function can be approximated by a 2nd order Taylor series. Moreover, it is possible to use one-dimensional cuts through the response surface to quantify the influence of the parameters separately. The influence of the design parameters is displayed in a classic Pareto plot, where positive effects on the responses are marked in blue, whereas negative effects are colored in red.

The last presented representation is more direct than other graphic results, giving the designer a useful tool to better understand which design parameters could be neglected because of their poor effect on global performances [19].

According to common techniques of robust design [31], the two sets of parameters are kept separated and two different sensitivity analyses are done to evaluate the influence of each set on the outputs separately.

Furthermore, an overall analysis considering all of the parameters at the same time would require a definitely higher number of trials, determining an unacceptable amount of time spent in simulating, as the noise and control factors could sensibly influence each other. Therefore, for the purposes of the present work, two different analyses gave a satisfactory result at a feasible computational cost. The reference output monitored is the current density: at a fixed user defined operating voltage, the higher its value, the higher the power output available from the fuel cell.

Considering the number of trials to be simulated and the splitting of the sensitivity analysis into two different ones, an amount of 100 simulations has been chosen as a compromise between computational cost and sufficiently reliable preliminary results. The sensitivity analysis tool provides different graphs and post-processing features. During a preliminary analysis, the most meaningful charts are the scatter plots and the Pareto plots, presented later in the text.

Considering the operating conditions at which the cell is investigated, the authors decided to start with the analysis of possible flooding (at low voltage and high current density), opting for 0.2 V. Only one single point of the polarization curve is analyzed, being anyway one of the most representative ones, where the cell is particularly sensible to change in performances. Also a validation of the approach based on the simulation of a single point could better test the goodness of the methodology.

III. RESULTS AND DISCUSSIONS

After the run completion, the first step of the design space evaluation consisted in determining the level of orthogonality of the input variables. Two different correlation matrices are presented for both noise and control factors (Table 1 and 2, respectively), obtained for the present case study.

As it can be seen, both noise and control factors present a very low correlation. These results are important because, as stated before, this is a useful preliminary step to get a good fitting response of the meta-model, avoiding to get confounding behaviors. If the correlation factor is not close to a zero value, there is some level of confounding of the independent variables and this would affect the ability to estimate the source of variability in the system response and the associated model coefficients [39]. The ability of the surrogate model to approximate the reality in a better way is given by the lack of void spaces in the design space. If void spaces are present, the surrogate model would consider regions not covered by data, making the model error excessive.

	i_{0a}	i_{0c}	r_{cond}	sat_{rate}
i_{0a}	1	-0.127	-0.114	0.0725
i_{0c}	-0.127	1	0.124	-0.029
r_{cond}	-0.114	0.124	1	-0.066
sat_{rate}	0.0725	-0.029	-0.066	1

Tab. 1 Noise factors correlation matrix, showing the mutual influence of each variable on the others. 1 indicates a perfect match; 0 indicates a complete non-correlation.

	T_a	Rh_a	T_c	Rh_c	$Compr$
T_a	1	-0.082	0.238	0.102	0.0064
Rh_a	-0.082	1	0.026	-0.091	-0.062
T_c	0.238	0.026	1	0.12	0.11
Rh_c	0.102	-0.091	0.12	1	0.0352
$Compr$	0.0064	-0.062	0.11	0.0352	1

Tab. 2 Control factors correlation matrix, showing the mutual influence of each variable on the others. 1 indicates a perfect match; 0 indicates a complete non-correlation.

Figure 1 shows the scatter plots originating from the present analysis on control factors. Examining such plots, the most interesting result is the direct proportionality between the current density and anode relative humidity. The Pareto plot that is obtained (shown in Figure 3, top part) clearly shows the relative weight of each of the selected control factors on the objective variable. The anode temperature and the anode relative humidity can be identified as the two main parameters affecting the current density. On the other hand, no clear correlation can be made for relative humidity and temperature at the cathode, neither for compression rate. The same procedure is adopted for the noise factors. Scatter plots for noise factors are provided in Figure 2. It is quite clear that the most leading parameter is the cathode exchange current density. Moreover, the anode exchange current has more effects than condensation and saturation, and it cannot be neglected. Such behavior is reflected by the Pareto plot shown in the bottom part of Figure 3. Saturation and condensation rate are instead not directly contributing.

As a matter of fact, at the cathode side the membrane is continuously humidified thanks to the water produced by the electrochemical reaction. On the other hand, the anode usually experiences difficulty in keeping the right membrane humidification, since the presence of liquid water is strictly connected to its transport through the membrane itself and the hydrogen inlet humidification.

Despite the membrane is thin, a good amount of membrane humidity must be guaranteed at its two sides, being humidification at only one side not enough. Therefore, the strong importance of humidity at the anode becomes clear.

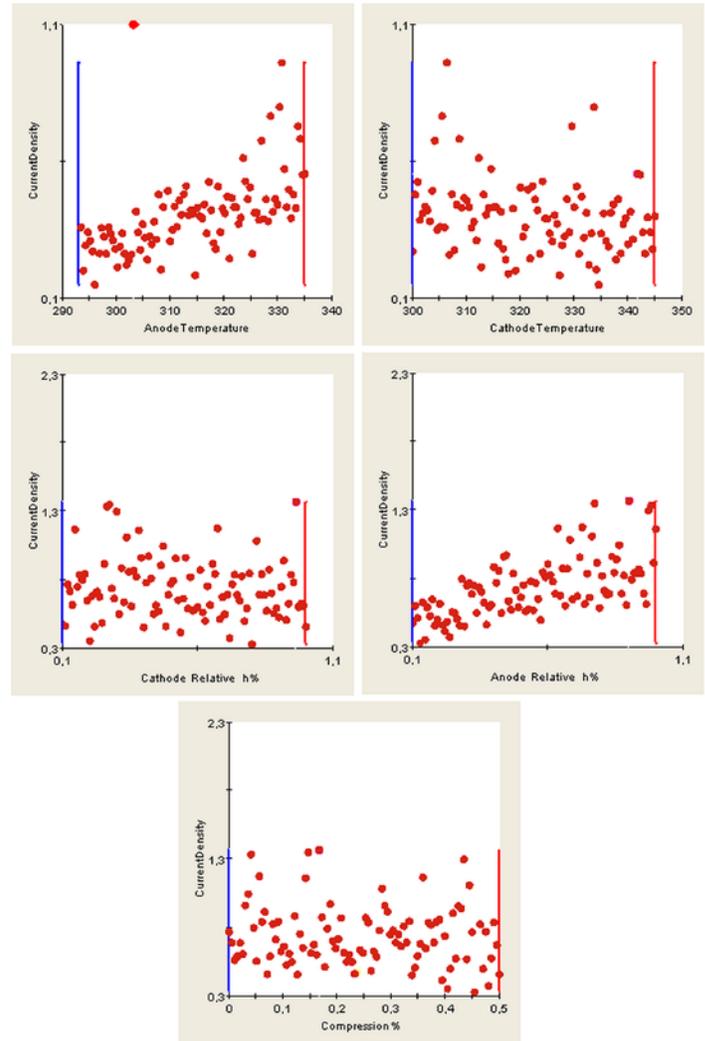


Fig. 1 Scatter plots obtained for the sensitivity analysis of the control factors. Each red point represents a simulation point of the DoE. For each plot, its corresponding control factor is given in the x-axis, between its lower and upper limits. The objective function is given on the y-axis.

Being the membrane humidification directly proportional to the electric conductivity of the membrane, a higher value of humidification means a higher electric current.

This is the reason why it is extremely important to monitor and correctly set the right value of anodic temperature and humidity. The reason for the importance of the anode temperature could be justified considering the operating point here simulated, equal to 0.2 V in output. At this low voltage value, the current production is high, meaning a high liquid water production at the cathode side. The back-diffusion of water through the membrane, given by the gradient of concentration at the two sides, is enhanced and can counterbalance the electro-osmotic drag. A lower relative humidity at the anode (i.e. a high inlet temperature) helps in removing the excess water, which could imply water flooding. This behavior, on the other hand, is opposed at low or medium current densities, where very high anode relative humidity is always required to prevent the membrane drying.

The cathode exchange current density shows a great influence on the polarization curve, being perfectly in-line with the physical explanation already given.

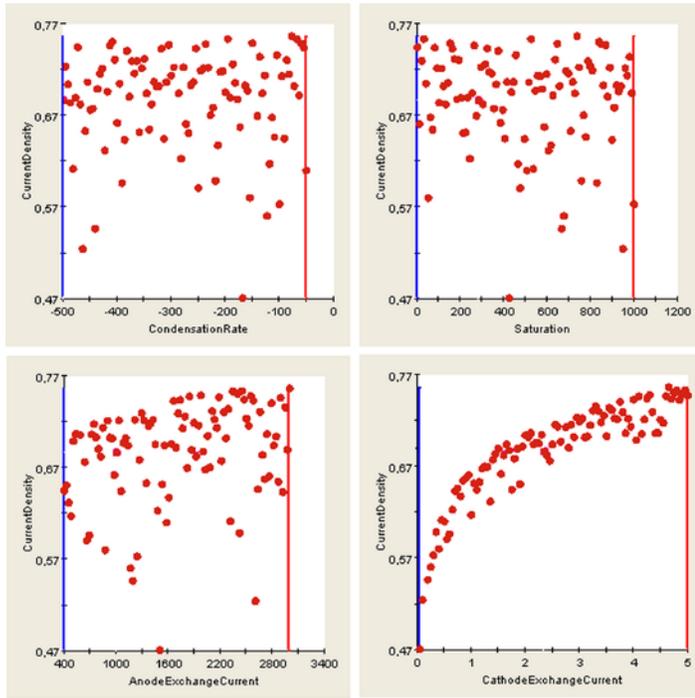


Fig. 2 Scatter plots obtained for the sensitivity analysis of the noise factors. Each red point represents a simulation point of the DoE. For each plot, its corresponding noise factor is given in the x-axis, between its lower and upper limits. The objective function is given on the y-axis.

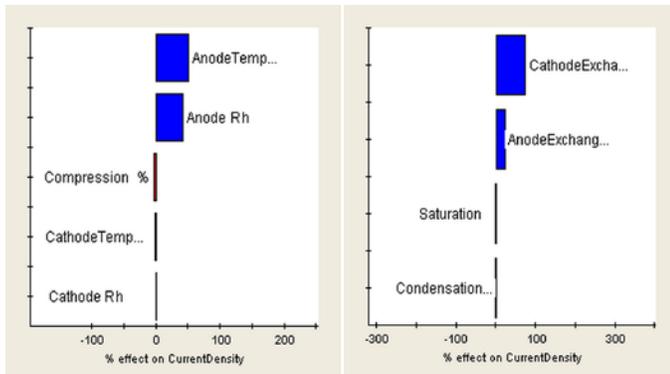


Fig. 3 Pareto plot of the control factors (right), with the noise factors kept to a constant value; Pareto plot of the noise factors (left), with the control factors are kept to a constant value. On the x-axis the percentage importance of each control factor on the objective function is shown. Blue colour for positive effects on the objective function and red for negative effects. It is noticeable how the main control factors are temperature and relative humidity at the anode, while the main noise factor is exchange current density both at anode and cathode.

IV. CONCLUSIONS

The presented work individuates the key factors affecting the behavior of a PEM FC model. The provided sensitivity analysis is a pivotal input to any MDO process that could be applied to such models, with the aim to reduce computational effort without affecting significantly the representativeness, thus leading to an increase in the efficiency of the model itself.

Further activities in this sense will be to realize a surrogate model based on the output of this work and to obtain a MDO able to provide and validate the best solution in terms of maximum current density produced at a given voltage.

REFERENCES

- [1] D. M. Bernardi & M. W. Verbrugge (1991). *Mathematical model of a gas diffusion electrode bonded to a polymer electrolyte*. *AIChE journal*, 37(8), 1151-1163.
- [2] S. Dutta, S. Shimpalee & J. W. Van Zee (2000). *Three-dimensional numerical simulation of straight channel PEM fuel cells*. *Journal of Applied Electrochemistry*, 30(2), 135-146.
- [3] [2] S. Dutta, S. Shimpalee & J. W. Van Zee (2001). *Numerical prediction of mass-exchange between cathode and anode channels in PEM fuel cell*. *Int J Heat and Mass Transf* 44:2029-2042.
- [4] T. E. Springer, T. A. Zawodzinski, S. Gottesfled (1991). *Polymer electrolyte fuel cell model*. *J Electrochem Soc* 138:2334-2342.
- [5] K. W. Lum, J. J. McGuirk (2005). *Three-dimensional model of a complete polymer electrolyte membrane fuel-cell model formulation, validation and parametric studies*. *J Power Sources* 143:103-124. doi: 10.1016/j.jpowsour.2004.11.032
- [6] B. R. Sivertsen, N. Djalali N (2005). *CFD-based modelling of proton exchange membrane fuel cells*. *J Power Sources* 141:65-78. doi: 10.1016/j.jpowsour.2004.08.054
- [7] A. Z. Weber & J. Newman (2004). *Modeling transport in polymer-electrolyte fuel cells*. *Chemical Reviews*, 104(10), 4679-4726.
- [8] K. Z. Yao, K. Karan, K. B. McAuley, P. Oosthuizen, B. Peppley & T. Xie(2004). *A review of mathematical models for hydrogen and direct methanol polymer electrolyte membrane fuel cells*. *Fuel Cells*, 4(1-2), 3-29.
- [9] C. Y. Wang (2004). *Fundamental models for fuel cell engineering*. *Chemical Reviews*, 104(10), 4727-4766.
- [10] D. Cheddie & N. Munroe (2005). *Review and comparison of approaches to proton exchange membrane fuel cell modeling*. *Journal of Power Sources*, 147(1), 72-84.
- [11] A. Z. Weber et al. *A Critical Review of Modeling Transport Phenomena in Polymer-Electrolyte Fuel Cells*. *Journal of The Electrochemical Society* 161.12 (2014): F1254-F1299.
- [12] N. P. Siegel, M. W. Ellis, D. J. Nelson, M. R. Von Spakovsky (2003). *A two-dimensional computational model of PEMFC with liquid water transport*. *J Power Sources* 128:173-184.
- [13] M.A.R.S. Al-Baghdadi (2009). *A CFD study of hygro-thermal stresses distribution in PEM fuel cell during regular cell operation*. *J. Renew Energy* 34:674-682.
- [14] D.S. Falcão, P.J. Gomes, V.B. Oliveira, C. Pinho C, A.M.F.R. Pinto (2011). *1D and 3D numerical simulations in PEM fuel cells*. *Int J Hydrogen Energy* 36:12486-12498. doi: 10.1016/j.ijhydene.2011.06.133
- [15] Z. H. Wang, C.Y. Wang & K.S. Chen (2001). *Two-phase flow and transport in the air cathode of proton exchange membrane fuel cells*. *Journal of Power Sources*, 94(1), 40-50.
- [16] C.Y. Wang & P. Cheng (1997). *Multiphase flow and heat transfer in porous media*. *Advances in heat transfer*, 30, 93-196.
- [17] S. Um, C.Y. Wang, & K.S. Chen (2000). *Computational fluid dynamics modeling of proton exchange membrane fuel cells*. *Journal of the Electrochemical society*, 147(12), 4485-4493.
- [18] J.E. Dawes, N.S. Hanspal, O.A. Family, A. Turan (2009) *Three-dimensional CFD modelling of PEM fuel cells: an investigation into the effects of water flooding*. *J Chem Engin Sci* 64:2781-2794. doi: 10.1016/j.ces.2009.01.060.

- [19] C.H. Min, Y.L. He, X.L. Liu, B.H. Yin, W. Jiang, W.Q. Tao (2006) *Parameter sensitivity examination and discussion of PEM fuel cell simulation model validation. Part II: results of sensitivity analysis and validation of the model*. J Power Sources 160:374-385. doi: 10.1016/j.jpowsour.2006.01.080
- [20] G.H. Guvelioglu H.G. Stenger (2005). *Main and interaction effects of PEM fuel cell design parameters*. J Power Sources 2: 424-433. doi: 10.1016/j.jpowsour.2005.06.009
- [21] M. Secanell, R. Songprakor, N. Djilali, A. Suleman (2010). *Optimization of a proton exchange membrane fuel cell membrane electrode assembly*. Struct Multidisc Optim 40:563-583. doi: 10.1007/s00158-009-0387-z
- [22] N. Pourmahmoud, S. Rezaadeh, I. Mirzaee, V. Heidarpoor (2011). *Three-dimensional numerical analysis of proton exchange membrane fuel cell*. J Mech Sci And Tech 25:2665-2673. doi: 10.1007/s12206-011-0743-y
- [23] N. Ahmadi, S. Rezaadeh, I. Mirzaee, N. Pourmahmoud (2012). *Three-dimensional computational fluid dynamic analysis of the conventional PEM fuel cell and investigation of prominent gas diffusion layers effect*. J Mech Sci And Tech 26:2247-2257. doi: 10.1007/s12206-012-0606-1
- [24] C. Kim, H. Sun (2012). *Topology optimization of gas flow channel in an automotive fuel cell*. Int J Automot Tech 13:783-789. doi: 10.1007/s12239-012-0078-4
- [25] M. Secanell, J. Wishart, P. Dobson (2011). *Computational design and optimization of fuel cells and fuel cell systems: a review*. Journal of Power Sources, 196(8), 3690-3704.
- [26] B. Mukhtar, S. Javaid Zaidi, M. Naim Faqir (2010). *Multi-objective function optimization for PEM fuel cell system*. ECS Trans 26:77-88. doi: 10.1149/1.3428978
- [27] R.Y. Rubinstein, D.P. Kroese (2007). *Simulation and the Monte Carlo method* (2nd ed.). John Wiley & Sons, New York.
- [28] K. Deb (2001). *Multi-objective optimization using evolutionary algorithms*. John Wiley & Sons, New York.
- [29] O. Vigna Suria, E. Testa, P. Peraudo, P. Maggiore (2011). *A PEM fuel cell distributed parameters model aiming at studying the production of liquid water within the cell during its normal operation: model description, implementation and validation*. In SAE World Congress 2011, MI, USA.
- [30] ANSYS (2011). *ANSYS FLUENT fuel cell modules manual*. ANSYS, Inc, USA.
- [31] D3.4.3 (2009) *Optimisation and robust design capabilities – achievements, needs and orientations*. CRESCENDO FP7-234344 © Copyright CRESCENDO Consortium.
- [32] A.V. Bernstein, A.P. Kuleshov, A.P. (2009). *Construction of orthogonal non-linear manifolds in the problem of dimension reduction*. In: Proceedings of 7th International School – Seminar on Multivariate statistical analysis and econometrics.
- [33] J.F.M. Barthelemy, R.T. Haftka (1993). *Approximation concepts for optimum structural design – A review*. Struct Optim 5:129-144. doi: 10.1007/BF01743349
- [34] N.A.C. Cressie (1993). *Statistics for spatial data*, revised edition. John Wiley & Sons, New York.
- [35] M. Papadrakakis, V. Papadopoulos (1996). *Robust and efficient methods for stochastic finite element analysis using Monte Carlo simulation*. Computer Methods in Appl Mech and Engin 134:325-340.
- [36] P.N. Peraudo, C. Abbondanza, P. Maggiore (2012). *A multi-objective design optimization approach for the preliminary design of high-speed low pressure turbine disks for green engine architectures*. 14th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, September 17-19, Indianapolis, IN. AIAA 2012-5606. doi: 10.2514/6.2012-5606
- [37] B. Tabachnick, G. Fidell, S. Linda (2007). *Using multivariate statistics* (5th ed.). Pearson International Edition, Boston.
- [38] S.M. Stigler (1989). *Francis Galton's account of the invention of correlation*. Stat Sci 4:73-79.
- [39] D.P. Francis, A.J. Coats, D. Gibson (1999). *How high can a correlation coefficient be? Effects of limited reproducibility of common cardiological measures*. Int J Cardiol 69:185-199.

Enrico Testa received his B.Sc. and M.Sc. degrees from Turin Polytechnic in Aerospace Engineering in 2007 and 2010, respectively. Then he received the Ph.D. degree in Aerospace Engineering in the same university in 2014. His major research areas dealt with fuel cell systems modeling for aerospace applications. At present, he is working in the automotive powertrain industry.

Paolo Maggiore is a professor at the Mechanical and Aerospace Engineering Department of Politecnico di Torino, that joined in 1992, where he teaches aerospace general systems engineering. Currently his students are involved in projects ranging from hydrogen fuel cell powered airplanes and UAVs, and health monitoring of flight controls, to multi-disciplinary design optimization of aerospace systems design

Lorenzo Pace graduated in Aerospace Engineering at Politecnico di Torino in 2008. Since 2008 to 2011, he worked as an assistant researcher, following studies about system experimental testing and modelization in the aerospace field, with a focus to energy saving techniques. Since 2012 to 2014 he completed a PhD in Aerospace Engineering at Politecnico di Torino, with the contribution of Thales Alenia Space, focused on the application of Model Based System Engineering to verification in the space industry.

Matteo D. L. Dalla Vedova received the M.Sc. and the Ph.D. from the Politecnico di Torino in 2003 and 2007, respectively. He is currently assistant researcher at the Department of Mechanics and Aerospace Engineering. His research activity is mainly focused on the aeronautical systems engineering and, in particular, is dedicated to design, analysis and numerical simulation of on board systems, study of secondary flight control systems and conception of related monitoring strategies, development of prognostic algorithms for aerospace servomechanism and study of innovative primary flight control architectures.

The Mathematical model of reflection of plane waves in a transversely isotropic magneto-thermoelastic medium under rotation

Abo-el-nour N. Abd-alla and Fatimah Alshaikh

Abstract— In this paper, the effect of rotation and magnetic field on plane waves in transversely isotropic thermoelastic medium under the Green-Lindsay theory with two relaxation times of generalized thermoelasticity has been discussed. The governing equations of generalized magneto-thermoelasticity of rotating transversely isotropic medium are solved in plane and a cubic velocity equation is obtained to show the existence of three quasi plane waves in the medium. The reflection of quasi plane waves is considered from a thermally insulated and stress-free surface, where three relations between the reflection coefficients are obtained. The reflection coefficients are computed for Cobalt material and presented graphically.

Keywords— Cobalt Material, Magneto-thermo-elastic coupling; Reflection coefficients; Rotation; Thermal relaxation times; Transversely isotropic materials.

I. INTRODUCTION

Two generalizations to the coupled theory were introduced. The first is due to Lord and Shulman [1] who obtained a wave-type heat equation by postulating a new law of heat conduction to replace the classical Fourier's law. Since the heat equation of this theory is of the wave-type. The second generalization to the coupled theory of elasticity is what is known as the theory of thermoelasticity with two relaxation times is obtained by Green and Lindsay [2] in an explicit version of the constitutive equations. These theories were extended by Sherief [3] and by Dhaliwal and Sherief [4] to include the effects of anisotropy. In this theory a modified law of heat conduction including both the heat flux and its time derivative replaces the conventional Fourier's law. The heat equation associated with this theory is a hyperbolic one and hence automatically eliminates the paradox of infinite speeds of propagation inherent in both the uncoupled and the coupled theories of thermoelasticity. For many problems involving steep heat gradients and when short time effects are sought this theory is indispensable.

Increasing attention is being devoted to the interaction between magnetic fields and strain in a thermoelastic solid due to its many applications in the fields of geophysics, plasma physics and related topics. In the nuclear field, the extremely high temperatures and temperature gradients as well as the

magnetic fields originating inside nuclear reactors influence their design and operations [5].

A comprehensive review of the earlier contributions to the subject can be found in [6], [7], [8] and [9]. Among the authors who considered the generalized magneto-thermoelastic equations are Nayfeh and Nemat-Nasser [10] who studied the propagation of plane waves in a solid under the influence of an electromagnetic field. They have obtained the governing equations in the general case and the solution for some particular cases. Choudhuri [11] extended these results to rotating media.

Abd-alla and his coworkers [12], [13] and [14] studied the phenomenon of the reflection and/or transmission of plane waves from free surface of a magneto-thermo-elastic solid half-space with many assumption. Othman and Song [15] investigated the reflection of magneto-thermo-elastic waves when the medium of reflected wave is homogeneous and isotropic. Singh et al. [16] applied Green-Naghdi's theory of generalized thermoelasticity to study the reflection of P and SV waves from the free surface of a magneto-thermoelastic solid half-space. Singh and Yadav [17] studied the effect of rotation and magnetic field on plane wave in transversely isotropic thermoelastic medium with one relaxation time. Wave propagation phenomenon in solids is important due to its relevance in composite engineering, geology, seismology, seismic exploration, control system and acoustics [18]. In view of the fact that most large bodies, like the earth, the moon, and other planets, have angular velocity and their own magnetic field. The study of wave propagation in a generalized thermoelastic media with additional parameters like rotation, electric, magnetic, anisotropy, porosity, viscosity, microstructure, temperature and other parameters provide vital information about existence of new or modified waves. The reflection phenomenon of plane wave propagation in an anisotropic solid has been studied by many researchers. For example, Keith and Crampin [19] studied the reflection and refraction of seismic waves at a plane interface between two dissimilar anisotropic media. Singh and Khurana [20] investigated the reflection of propagation of plane waves at the free surface of a monoclinic elastic half-space. Sharma [21] discussed the reflection of thermoelastic waves from the stress-free thermally insulated boundary of a transversely isotropic solid half-space. Kumar and Singh [22] illustrated the effects of rotation and imperfection on reflection and transmission of plane waves in anisotropic generalized thermoelastic media. Recently, Abd-alla and his co-workers [23]-[29] investigated many problems concerning the reflection or the reflection and

Acknowledgment:

This work was supported by the Deanship of Scientific Research of Jazan University under the Grant no. 25/G5/1434.

A.N. Abd-alla, and F. Alshaikh, Department of Mathematics, Faculty of Science, Jazan University, Jazan, Saudi Arabia.

Corresponding author: *e-mail address*: aboelnourabdalla@yahoo.com

refraction phenomena in anisotropic piezoelectric or in thermo-piezoelectric materials with the existence of the initial stresses or not.

The aim of this paper is to study the influence of two thermal relaxation times according to the theory of Green-Lindsay, the rotation and the magnetic field on the reflection of plane waves in a transversely isotropic magneto-thermoelastic medium. Reflection of the plane waves from a stress-free thermally insulated surface studied to obtain the reflection coefficients of various reflected waves. The effects of anisotropy, rotation, thermal relaxation times and magnetic fields illustrated graphically on these coefficients.

II. FORMULATION OF THE PROBLEM AND THE BASIC EQUATIONS

We consider the elastic medium is rotating uniformly with an angular velocity $\boldsymbol{\Omega} = \Omega \mathbf{n}$, where \mathbf{n} is a unit vector representing the direction of the axis of rotation. The displacement equation of motion in the rotating frame of reference has two additional terms [11]: Centripetal acceleration, $\boldsymbol{\Omega} \times (\boldsymbol{\Omega} \times \mathbf{u})$ due to time varying motion only and the Corioli's acceleration $2\boldsymbol{\Omega} \times \dot{\mathbf{u}}$, where \mathbf{u} is the dynamic displacement vector. These terms do not appear in non-rotating media.

Therefore, the basic equation for a homogeneous and transversely isotropic magneto-thermoelastic half space in the context of Green-Lindsay theory may be taken in a unified form without body forces, body couples and heat sources can be written as:

(a) Equation of motion with Lorentz force \mathbf{f} is given as

$$\nabla \cdot \boldsymbol{\sigma} + \mathbf{f} = \rho[\ddot{\mathbf{u}} + \{\boldsymbol{\Omega} \wedge (\boldsymbol{\Omega} \wedge \mathbf{u}) + (2\boldsymbol{\Omega} \wedge \dot{\mathbf{u}})] \quad (1)$$

$$\mathbf{f} = \mathbf{J} \times \mathbf{B} \quad (2)$$

(b) Maxwell equations:

$$\text{curl} \mathbf{H} = \mathbf{J}, \text{curl} \mathbf{E} = -\dot{\mathbf{B}}, \text{div} \mathbf{B} = 0, \mathbf{B} = \mu_e \mathbf{H} \quad (3)$$

(c) Generalized heat conduction equation according the (G-L) theory:

$$\mathbf{K} \nabla^2 T = T_0 \boldsymbol{\gamma} \nabla \dot{\mathbf{u}} + \rho C_e (\dot{T} + t_0 \ddot{T}) \quad (4)$$

(d) Strain-displacement relations:

$$\boldsymbol{\varepsilon} = \frac{1}{2} (\nabla \mathbf{u} + \nabla \mathbf{u}^T) \quad (5)$$

(e) Constitutive relation between stress and kinematic, electric and thermal fields:

$$\boldsymbol{\sigma} = \mathbf{c} \boldsymbol{\varepsilon} - \boldsymbol{\gamma} (T + t_1 \dot{T}) \quad (6)$$

(f) Generalized Ohm's law in deformable continua is

$$\mathbf{J} = \sigma_o [\mathbf{E} + (\dot{\mathbf{u}} \times \mathbf{B})], \quad (7)$$

The elastic part $\boldsymbol{\sigma}^{el} = \mathbf{c} \boldsymbol{\varepsilon}$ of the stress tensor $\boldsymbol{\sigma}$ in equation (6), for the transversely isotropic materials, may be expressed as [24]:

$$\mathbf{c} \boldsymbol{\varepsilon} = \boldsymbol{\sigma}^{el} = \lambda \text{tr} \boldsymbol{\varepsilon} \mathbf{I} + 2\mu \boldsymbol{\varepsilon} + \alpha (\mathbf{n}^T \boldsymbol{\varepsilon} \mathbf{I} + \text{tr} \boldsymbol{\varepsilon} \mathbf{N}) + 2(\mu_o - \mu)(\mathbf{N} \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon} \mathbf{N}) + \beta (\mathbf{n}^T \boldsymbol{\varepsilon} \cdot \mathbf{N}) \quad (8)$$

If the direction of the symmetric axis \mathbf{n} is along the y -axis, for $y \leq 0$ only the following material constants do not vanish:

i) the elastic moduli $c_{11}, c_{12}, c_{13}, c_{33}$ and c_{44} ; ii) the thermal elastic coupling moduli γ_1 and γ_3 ; So in Eq. (8) the elastic parameters, using elastic moduli in Voigt notation, become:

$$\begin{cases} \lambda = c_{12}, & 2\mu = (c_{11} - c_{12}), & \alpha = (c_{13} - c_{12}) \\ 2(\mu_o - \mu) = (2c_{44} - c_{11} + c_{12}) \\ \beta = (c_{11} + c_{33} - 2c_{13} - 4c_{44}) \end{cases} \quad (9)$$

The superimposed dot represents time derivative and the symbols appearing in the previous equations are listed in table (1) in the end of this paper. The complete geometry of the problem is shown in Fig. 1

We consider the problem of a magneto-thermoelastic half-space ($y \leq 0$) with the small effect of temperature gradient on the conduction current \mathbf{J} is neglected. We assume that $\mathbf{H} = \mathbf{H}_o + \mathbf{h}$ where $\mathbf{H} = (0, 0, H_o)$. The perturbed magnetic field \mathbf{h} is so small that the product of \mathbf{h}, \mathbf{u} and their derivatives can be neglected while linearizing the field equations. Also, we consider the magnetic field constant intensity $\mathbf{H} = (0, 0, H_o)$ acts parallel to the bounding plane (take as the direction of the z -axis). Thus, all quantities considered will be functions of the time variable t and of the coordinates x and y [10].

Substituting from Eqs. (3) and (7) into Eq. (1), one may obtain

$$c_{11} u_{,xx} + c_{12} v_{,xy} + c_{66} (u_{,yy} + v_{,xy}) - \gamma_1 (T + t_1 \dot{T})_{,x} + \mu_e H_o^2 (u_{,xx} + v_{,xy}) = \rho (\ddot{u} - \Omega^2 u - 2\Omega \dot{v}), \quad (10)$$

$$c_{22} v_{,yy} + c_{12} u_{,xy} + c_{66} (v_{,xx} + u_{,xy}) - \gamma_2 (T + t_1 \dot{T})_{,y} + \mu_o H_o^2 (u_{,xy} + v_{,yy}) = \rho (\ddot{v} - \Omega^2 v + 2\Omega \dot{u}), \quad (11)$$

Furthermore, the heat conduction equation may be written in explicit form as:

$$K_1 T_{,xx} + K_2 T_{,yy} - \rho C_e (\dot{T} + t_0 \ddot{T}) = T_o (\gamma_1 \dot{u}_{,x} + \gamma_2 \dot{v}_{,y}), \quad (12)$$

where we have used $c_{66} = 0.5(c_{11} - c_{12})$ and the electromagnetic body force ($\mathbf{f} = \mathbf{J} \times \mathbf{B}$), $\boldsymbol{\gamma}$ and \mathbf{K} were written by their components. We should note that for G-L theory the thermal relaxation time must satisfy the relation $t_1 \geq t_o > 0$.

The solution of equations (10), (11) and (12) can be chosen for a harmonic wave propagated in the direction in the following form

$$\{u, v, T\} = (A, B, C) \exp[ik(x \sin \theta - y \cos \theta - ct)], \quad (13)$$

where the wave normal lies in the xy -plane, and makes an angle θ with the y -axis. Substituting from Eq. (13) into Eqs. (10), (11) and (12), we can get a system of three homogeneous equations:

$$A[D_1 - \zeta\Omega^*] + B[D_2 + 2\frac{\Omega}{\omega}i\zeta] + \tau C^* \sin \theta = 0, \quad (14)$$

$$A[D_2 + 2\frac{\Omega}{\omega}i\zeta] + B[D_3 + \zeta\Omega^*] - \tau_n \bar{\gamma} C^* \cos \theta = 0, \quad (15)$$

$$i\lambda(\varepsilon\zeta \sin \theta)A - i\lambda(\bar{\gamma}\varepsilon\zeta \cos \theta)B + \frac{1}{\omega}(\zeta - D_5)C^* = 0, \quad (16)$$

where we have used

$$D_1 = (c_{11} + \mu_e H_0^2) \sin^2 \theta + c_{66} \cos^2 \theta,$$

$$D_2 = -(c_{12} + c_{66} + \mu_e H_0^2) \sin \theta \cos \theta,$$

$$D_3 = (c_{22} + \mu_e H_0^2) \cos^2 \theta + c_{66} \sin^2 \theta,$$

$$D_4 = K_1 \sin^2 \theta + K_2 \cos^2 \theta,$$

with the help of the following of dimensionless quantities

$$\zeta = \rho c^2, \quad \omega = Kc, \quad C^* = (\omega\gamma_1 C / k), \quad \Omega^* = 1 + (\Omega^2 / \omega^2),$$

$$\bar{\gamma} = \frac{\gamma_2}{\gamma_1}, \quad \varepsilon = \frac{T_o \gamma_1^2}{\rho C_e}, \quad \lambda_n = nt_o + \frac{i}{\omega},$$

$$\lambda_1 = t_o + \frac{i}{\omega}, \quad D_5 = \frac{D_4}{\tau_1 C_e}, \quad \alpha = \frac{(i/\omega)}{t_o + (i/\omega)} = \frac{(i/\omega)}{\lambda_1}.$$

To get the non-trivial solution of the system of Eqs. (14), (15) and (16), the determination of the factor of matrix must vanish. Therefore,

$$\begin{vmatrix} D_1 - \zeta\Omega^* & D_2 + 2i\zeta(\Omega/\omega) & \lambda_n \sin \theta \\ D_2 - 2i\zeta(\Omega/\omega) & D_3 - \zeta\Omega^* & -\lambda_n \bar{\gamma} \cos \theta \\ i\omega\alpha\varepsilon\zeta \sin \theta & -i\omega\alpha\bar{\gamma}\varepsilon\zeta \cos \theta & \zeta - D_5 \end{vmatrix} = 0 \quad (17)$$

This yields the following cubic equation in ζ :

$$A_0 \zeta^3 + A_1 \zeta^2 + A_2 \zeta + A_3 = 0 \quad (18)$$

where,

$$A_0 = 4(\Omega/\omega)^2 - \Omega^{*2},$$

$$A_1 = \Omega^* (D_1 + D_3) + (\Omega^{*2} - 4(\Omega/\omega)^2) D_5 - i\lambda_n \varepsilon \omega \alpha (\sin^2 \theta + \bar{\gamma}^2 \cos^2 \theta) \Omega^*$$

$$A_2 = D_2^2 - D_1 D_3 + \Omega^* (D_1 D_5 + D_3 D_5) + i\lambda_n \varepsilon \omega \alpha \times (2\bar{\gamma} D_2 \sin \theta \cos \theta + \bar{\gamma}^2 D_1 \cos^2 \theta + D_3 \sin^2 \theta)$$

$$A_3 = (D_1 D_3 - D_2^2) D_5$$

Here, the three roots $\zeta_j = \rho c_j^2$, ($j=1,2,3$) of equation (13)

corresponding to the complex phase velocities c_j , ($j=1,2,3$) of

qP, qSV and qT waves, respectively. If we write $c_j^{-1} = V_j^{-1} - i\omega q_j$, ($j=1,2,3$), then clearly V_j and q_j are the

speeds of propagation and the attenuation coefficients of the qP, qSV and qT waves [17].

III. THE BOUNDARY CONDITIONS

The required boundary conditions at the free surface $y = 0$ are vanishing of the normal stresses, tangential stresses and normal component of the heat flux vector, i.e.,

$$\sigma_{yy} + \sigma_{yy}^* = 0, \quad \sigma_{yx} + \sigma_{yx}^* = 0 \quad \text{and} \quad T_{,y} = 0 \quad \text{on} \quad y = 0, \quad (19)$$

where

$$\sigma_{yy} = c_{12} u_{,x} + c_{22} v_{,y} - \beta_2 (T + t_1 \dot{T}),$$

$$\sigma_{yy}^* = -\mu_e H_o^2 (u_{,x} + v_{,y}),$$

$$\sigma_{yx} = c_{44} (u_{,y} + v_{,x}), \quad \sigma_{yx}^* = 0.$$

IV. REFLECTION FROM FREE SURFACE

In this section, we shall drive the relations between the reflection coefficients, when (qP or qT or qSV) wave becomes incident at a thermoelastic solid half-space ($y \geq 0$) with thermally insulated and stress-free surface $y = 0$. The positive y -axis is taken into the half-space. For the incident wave at free surface, there will be three reflected waves, i.e., reflected qP, reflected qT and reflected qSV. Accordingly, if the wave normal to the incident wave (qP or qT or qSV) makes angle θ_o with the positive direction of y -axis and those of reflected qP, qSV and qT waves make θ_1 , θ_2 and θ_3 with the same direction. The complete geometry showing the incident and reflected waves is depicted in figure (1). The appropriate displacement components and temperature field which must satisfy the boundary conditions (19) at $y=0$ are:

$$u = A_o e^{\eta_o} + A_1 e^{\eta_1} + A_2 e^{\eta_2} + A_3 e^{\eta_3}, \quad (20)$$

$$v = F_o A_o e^{\eta_o} + F_1 A_1 e^{\eta_1} + F_2 A_2 e^{\eta_2} + F_3 A_3 e^{\eta_3}, \quad (21)$$

$$T = G_o A_o e^{\eta_o} + G_1 A_1 e^{\eta_1} + G_2 A_2 e^{\eta_2} + G_3 A_3 e^{\eta_3}, \quad (22)$$

with

$$\eta_o = ik_o (x \sin \theta_o - y \cos \theta_o - c_o t),$$

$$\eta_l = ik_l (x \sin \theta_l + y \cos \theta_l - c_l t) \quad , \quad l = 1,2,3$$

where c_o is the velocity of the incident qP or qT or qSV wave, k_l , ($l=0,1,2,3$) are complex wave numbers, $c_l = \zeta_l / \rho$ ($l = 0,1,2,3$).

Substituting from Eqs. (20-22) into Eq.(10) and Eq. (11), we get in the case of incident wave

$$F_0 = -N_1 / N_2, \quad G_0 = (k_o / \xi \lambda_n) (N_3 / N_2) \quad (23)$$

where

$$N_1 = \gamma_1 \sin \theta_o [(D_{2o} - 2i(\Omega/\omega)\zeta_o] + \gamma_2 \cos \theta_o (D_{1o} - \zeta_o \Omega^*),$$

$$N_2 = \gamma_2 \cos \theta_o [(D_{2o} + 2i(\Omega/\omega)\zeta_o] + \gamma_1 \sin \theta_o (D_{3o} - \zeta_o \Omega^*),$$

$$N_3 = [(4(\Omega/\omega)^2 - \Omega^{*2})\zeta_o^2 + \Omega^* (D_{1o} + D_{3o})\zeta_o + D_{2o}^2 - D_{1o} D_{3o}],$$

Where

$$D_{1o} = (c_{11} + \mu_e H_o^2) \sin^2 \theta_o + c_{66} \cos^2 \theta_o,$$

$$D_{2o} = -(c_{12} + c_{66} + \mu_e H_o^2) \sin \theta_o \cos \theta_o,$$

$$D_{3o} = (c_{22} + \mu_e H_o^2) \cos^2 \theta_o + c_{66} \sin^2 \theta_o.$$

In the case of reflected waves, one may get the following displacement components and temperature field which are suitable to satisfy the boundary conditions (19) at $y=0$.

$$u = A_l e^{\eta_l}, v = F_l A_l e^{\eta_l} \text{ and } T = G_l A_l e^{\eta_l}, l = 1, 2, 3 \quad (24)$$

with

$$F_l = -N_{1l} / N_{2l}, G_l = -(k_l / \omega \lambda_n)(N_{3l} / N_{2l}), \quad (25)$$

where

$$\begin{aligned} N_{1l} &= \gamma_1 \sin \theta_l [(D_{2l} - 2i(\Omega / \omega) \zeta_l] - \gamma_2 \cos \theta_l (D_{1l} - \zeta_l \Omega^*), \\ N_{2l} &= \gamma_2 \cos \theta_l [(D_{2l} + 2i(\Omega / \omega) \zeta_l] - \gamma_1 \sin \theta_l (D_{3l} - \zeta_l \Omega^*), \\ N_{3l} &= [(4(\Omega / \omega)^2 - \Omega^{*2}) \zeta_l^2 + \Omega^* (D_{1l} + D_{3l}) \zeta_l + D_{2l}^2 - \\ &D_{1l} D_{3l}], \end{aligned}$$

where

$$D_{1l} = (c_{11} + \mu_e H_o^2) \sin^2 \theta_l + c_{66} \cos^2 \theta_l,$$

$$D_{2l} = (c_{12} + c_{66} + \mu_e H_o^2) \sin \theta_l \cos \theta_l,$$

$$D_{3l} = (c_{22} + \mu_e H_o^2) \cos^2 \theta_l + c_{66} \sin^2 \theta_l.$$

Using Eqs. (20-22) which corresponding to incident and reflected waves in the boundary conditions (19) yield

$$\begin{aligned} &-(c_{22} - \mu_e H_o^2)(F_o A_o (ik_o \cos \theta_o) - F_1 A_1 (ik_1 \cos \theta_1) - \\ &F_2 A_2 (ik_2 \cos \theta_2) - F_3 A_3 (ik_3 \cos \theta_3)) + (c_{12} - \mu_e H_o^2) \times \\ &(A_o (ik_o \sin \theta_o) + A_1 (ik_1 \sin \theta_1) + A_2 (ik_2 \sin \theta_2) + \\ &A_3 (ik_3 \sin \theta_3)) - \gamma_2 (1 - it_1 \omega)(G_o A_o + G_1 A_1 + G_2 A_2 + G_3 A_3) = 0 \\ &- ik_o (A_o \cos \theta_o - A_o F_o \sin \theta_o) + ik_1 (A_1 \cos \theta_1 + A_1 F_1 \sin \theta_1) + \\ &ik_2 (A_2 \cos \theta_2 + A_2 F_2 \sin \theta_2) + ik_3 (A_3 \cos \theta_3 + A_3 F_3 \sin \theta_3) = 0 \\ &- ik_o (G_o A_o \cos \theta_o) + ik_1 (G_1 A_1 \cos \theta_1) + ik_2 (G_2 A_2 \cos \theta_2) + \\ &ik_3 (G_3 A_3 \cos \theta_3) = 0. \end{aligned} \quad (27)$$

Since the phases of the waves should be the same for each value of x , then we obtain the following relations must be valid $\eta_o = \eta_1 = \eta_2 = \eta_3$,

and so

$$k_o \sin \theta_o = k_1 \sin \theta_1 = k_2 \sin \theta_2 = k_3 \sin \theta_3$$

and

$$k_o c_o = k_1 c_1 = k_2 c_2 = k_3 c_3.$$

Therefore, the Snell's law may be deduced as

$$\frac{\sin \theta_o}{V_o} = \frac{\sin \theta_1}{V_1} = \frac{\sin \theta_2}{V_2} = \frac{\sin \theta_3}{V_3}, \quad (29)$$

Multiplying Eqs. (26-28) by the factor $(i / k_o A_o)$ and using Snell's law (29).

putting

$$\frac{A_i}{A_o} = X_i, \quad (30)$$

one may get the following system of three non-homogenous equations

$$\sum_{j=1}^3 a_{ij} X_j = b_i, \quad i = 1, 2, 3 \quad (31)$$

where

$$a_{1L} = F_L (c_{22} - \mu_e H_o^2) [(V_o / V_L)^2 - \sin^2 \theta_o]^{1/2} + (c_{12} - \mu_e H_o^2) \sin \theta_o + (\omega \lambda_n \gamma_2 / k_o) G_L,$$

$$a_{2L} = [(V_o / V_L)^2 - \sin^2 \theta_o]^{1/2} + F_L \sin \theta_o,$$

$$a_{3L} = G_L [(V_o / V_L)^2 - \sin^2 \theta_o]^{1/2},$$

$$b_1 = (c_{22} - \mu_e H_o^2) F_o \cos \theta_o - (c_{12} - \mu_e H_o^2) \sin \theta_o - (\omega \lambda_n \gamma_2 / k_o) G_o,$$

$$b_2 = \cos \theta_o - F_o \sin \theta_o, \quad b_3 = G_o \cos \theta_o, \quad L = 1, 2, 3.$$

From Eqs. (30). the analytical expressions of reflection coefficients X_1, X_2 and X_3 for incident qP waves may be obtained. While the analytical expressions of reflection coefficients Y_1, Y_2 and Y_3 for incident qSV and the analytical expressions of reflection coefficients Z_1, Z_2 and Z_3 for incident qT are related to reflection coefficients X_j with the relations:

$$Y_1 = F_1^* X_1, \quad Y_2 = F_2^* X_2, \quad Y_3 = F_3^* X_3 \quad (32)$$

$$Z_1 = G_1^* X_1, \quad Z_2 = G_2^* X_2, \quad Z_3 = G_3^* X_3 \quad (33)$$

where

$$F_j^* = F_j / F_o, \quad G_j^* = G_j / G_o, \quad j = 1, 2, 3. \quad (34)$$

V. NUMERICAL RESULTS AND DISCUSSION

The Cobalt material is chosen for numerical evaluations which has the following physical data [21].

$$c_{11} = 3.071 \times 10^{11} \text{ Nm}^{-2}, \quad c_{12} = 1.650 \times 10^{11} \text{ Nm}^{-2}$$

$$c_{22} = 1.27 \times 10^{11} \text{ Nm}^{-2}, \quad C_e = 4.27 \times 10^2 \text{ JKg}^{-2} \text{ deg}^{-1},$$

$$\gamma_1 = 7.040 \times 10^6 \text{ Nm}^{-2} \text{ deg}^{-1}, \quad \rho = 8.836 \times 10^3 \text{ Kgm}^{-3}$$

$$\gamma_2 = 6.9 \times 10^6 \text{ Nm}^{-2} \text{ deg}^{-1}, \quad K_1 = 0.69 \times 10^2 \text{ Wm}^{-2} \text{ deg}^{-1},$$

$$K_2 = 0.69 \times 10^2 \text{ Wm}^{-2} \text{ deg}^{-1}, \quad T_o = 298 \text{ K}.$$

Using Mathcad program, the modulus of the reflection coefficient for different quasi plane waves are calculated numerically versus the angle of incidence of (qP or qSV or qT) waves for different thermal, rotation and magnetic parameters from Eqs. (30), (32) and (33). These variations of the reflection coefficients are presented graphically in Figs. 2a, 2b, 2c, 3a, 3b, 3c, 4a, 4b and 4c. We can summarize the following remarks:

First, the variation of the reflection coefficients ratios X_1, X_2 and X_3 versus the angle of incidence θ_o for incident of qP wave under the theory of (G-L) is studied in the following cases: (i) The first group when the dimensionless rotation parameter is $\Omega' = 4$, the imposed primary magnetic field is $H_o = 10^5 \text{ A/m}$ and the variation of the thermal relaxation time with $t_1 = (2, 4, 6, 8) \times 10^{-12} \text{ s.}$, the results are shown in Figs. (2a, 2b, 2c).

In this group, figure 2a shows that the curve of X_1 (when $t_1 = 2 \cdot 10^{-12}$ s., solid red curve) for each value of θ_o in the interval $(0^\circ, 55^\circ)$ increases monotonically from 1.133 at $\theta_o = 0^\circ$ up to 2.215 at $\theta_o = 55^\circ$ and decreases up to the minimum value 1.161 at $\theta_o = 90^\circ$. Moreover, the values of X_1 decrease with variations in the thermal relaxation time t_1 on the same interval $\theta_o = (0^\circ, 55^\circ)$. While in the interval $\theta_o = (56^\circ, 90^\circ)$, the variations of X_1 have a small increases as the thermal relaxation time t_1 increases.

Figs. 2b and 2c, show the reflection coefficients X_2 and X_3 . In these cases, the qualitative behaviors are approximately similar to the standard normal distribution in probability.

(ii) The second group when $H_o = 10^5$ A/m, $t_1 = 5 \times 10^{-12}$ s. and the variation of the dimensionless rotation $\Omega' = 4.0, 4.3, 4.6, 4.9$, the results are given in Figs. (3a, 3b, 3c).

(iii) The third group when $\Omega' = 4$, $t_1 = 5 \times 10^{-12}$ s. and the variation of $H_o = (1,2,3,4) \times 10^5$ A/m, the results are potted in Figs. (4a, 4b, 4c).

From Figures 5 and 6, it is noted that the parameter of proportionality F_1^* and G_1^* are equal one (described by the solid (red) lines). Which means that the reflection coefficients Y_1 and Z_1 of the qSV and qT waves coincide on the reflection coefficients X_1 of the qP wave. Moreover, Figure 5 presents the parameters of proportionality F_2^* and F_3^* which show the relations between X_2 and X_3 of qP wave with Y_2 and Y_3 of qSV, respectively. It is easy to see that F_2^* and F_3^* change their values, respectively, in (0.698, 1.519) and (0.071, 0.948) when the values of the angle of incident θ_o changes in the interval $(0^\circ, 90^\circ)$. Similarly, Figure 6 shows the parameters of proportionality G_2^* and G_3^* which give, respectively, the relations between X_2 and X_3 of qP wave with Z_2 and Z_3 of qT. It is clear that G_2^* and G_3^* change their values, respectively, between (4.540, 2.145) and (0.033, 0.779) when the values of the angle of incident θ_o changes on the interval $(0^\circ, 90^\circ)$.

CONCLUSIONS

The solution of governing equations of magneto-thermoelastic transversely isotropic medium illustrates the existence of three quasi-plane waves, namely, qP, qSV and qT waves. Reflection of these plane waves is presented at stress free and thermally insulated surface of transversely isotropic magneto-thermoelastic solid half-space in the context of Green-Lindsay

theory of thermoelasticity. The reflection coefficient ratios depend on the angle of incidence, angle of reflection, rotation, thermal relaxation times and elastic parameters. The numerical computations showed that the reflection coefficients of various quasi-plane waves are affected significantly due to the presence of relaxation times, rotation, imposed magnetic field and anisotropy in the medium. The reflection coefficients of qSV and qT waves are proportional with the reflection coefficients of qP wave. However, the reflection coefficients of qT wave are much less than that of qP and QSV waves at each angle of incidence. This study is important to solve problems with factors such as elastic field, thermal field, magnetic field and rotation coexist.

References

- [1] H. Lord, Y. Shulman, A generalized dynamical theory of thermoelasticity, *J. Mech. Phys. Solid*, 15 pp. 299–309 (1967).
- [2] A. Green, K. Lindsay, Thermoelasticity, *J. Elasticity*, 2 pp. 1–7(1972).
- [3] H.H. Sherief, On generalized thermoelasticity, Ph.D. thesis, University of Calgary, Canada, (1980).
- [4] R.S. Dhaliwal, H.H. Sherief, Generalized thermoelasticity for anisotropic media, *Quart. Appl. Math.*, 38 pp. 1-8 (1980).
- [5] J.L. Nowinski, Theory of thermoelasticity with applications, Sijthoff & Noordhoff International Publishers, Alphen Aan Den Rijn (1978).
- [6] G. Paria, Magneto-elasticity and magneto-thermoelasticity, *Adv. Appl. Mech.*, 10 pp. 73–112 (1967).
- [7] W. Nowacki, "Magnetoelasticity," Chapter II, in *Electromagnetic interactions in elastic solids*, edited by H. Parkus (Springer, Vienna), pp.158-183 (1979).
- [8] F.C. Moon, "Magneto-solid mechanics," John Wiley & Sons, New-York (1985).
- [9] Abd-alla A.N., Nonlinear constitutive equations for thermo-electroelastic materials. *Mechanics Research Communications* 24 (3) pp. 335–346 (1999).
- [10] A. Nayfeh, S. Nemat-Nasser, Electromagneto-thermoelastic plane waves in solids with thermal relaxation, *J. Appl. Mech. Series E*, 39 pp. 108–113 (1972).
- [11] S. Choudhuri, Electro magneto-thermoelastic plane waves in rotating media with thermal relaxation, *Int. J. Engrg. Sci.*, 22 pp. 519–530 (1984).
- [12] A.N. Abd-alla, and S.M. Abo-Dahab, The influence of the viscosity and the magnetic field on reflection and transmission of waves at interface between magneto-viscoelastic materials, *Meccanica* 43 (4), pp. 437-448 (2008).
- [13] A.N. Abd-alla, A. A. Yahia and S. M. Abo-Dahab. On Reflection of the Generalized Magneto-thermo-viscoelastic Plane Waves. *Chaos, Solitons and Fractals*, 16 pp.211–231(2003).
- [14] A.N. Abd-alla, Relaxation effects on reflection of generalized magneto-thermo-elastic waves, *Mechanics Research Communications* 27 (5) pp. 591-600 (2000).

[15] M.I.A. Othman, Y.Q. Song, Reflection of magneto-thermoelastic waves from a rotating elastic half-space, *International J. Engng. Sci.*, 46, pp.459-474, (2008).

[16] B. Singh, L. Singh and S. Deswal, Reflection of plane waves from a free surface of a generalized magneto-thermoelastic solid half-space with diffusion, *Journal of Theoretical and Applied Mechanics*, 52 (2) pp. 385-394 (2014).

[17] B. Singh and A.K.Yadav, Reflection of plane waves in a rotating transversely isotropic magneto-thermoelastic solid half space, *Journal of Theoretical and Applied Mechanics*, Sofia, 42(3) pp. 33-60, (2012).

[19] Keith, C.M., S. Crampin. *Seismic Body Waves in Anisotropic Media: Reflection and Refraction at a Plane Interface*. Geophy. J. Royal Astro. Soc., 49 pp. 181–208 (1977).

[20] Singh, S. J., S. Khurana., Reflection of P and SV waves at the free surface of a monoclinic elastic half-space, *Proc. Indian Acad. Sci. (Earth Planet. Sci.)* 111 pp. 401–412 (2002).

[21] J.N. Sharma, V. Kumar and S.P. Sud., Plane harmonic waves in orthorhombic thermoelastic, materials *J. Acoust. Soc. Am.* 107, pp. 293- 305 (2000).

[22] R. Kumar and M. Singh., Effects of rotating and imperfection on reflection and transmission of plane waves in anisotropic generalized thermoelastic media. *J. Sound Vibr.*, 324 pp. 773–797 (2009).

[23] A.N. Abd-alla, I. Giorgio, L. Galantucci, A. Hamdan and D. del Vescovo, Wave reflection at a free interface in an anisotropic pyroelectric medium with nonclassical thermoelasticity, Submitted to *Continuum Mechanics and thermodynamics*, (2014).

[24] A.N. Abd-alla, A.M. Hamdan, I. Giorgio and D. Del Vescovo, The mathematical model of reflection and refraction of longitudinal waves in thermo-piezoelectric materials, *Online: Archive of Applied Mechanics*, 84, pp.1229-1248 (2014).

[25] A.N. Abd-alla, F.A. AlShaikh, and A.Y. Al-Hossain, The reflection phenomena of quasi-vertical transverse waves in piezoelectric medium under initial stresses, *Meccanica* 47 (3) , pp.731-744 (2012).

[26] A.N. Abd-alla, H.A. Eshaq, and H. El Haes, The phenomena of reflection and transmission waves in smart nano materials, *Journal of Computational and Theoretical Nanoscience* 8 (9) pp.1670-1678 (2011).

[27] A.N. Abd-alla, A.Y. Al-Hossain, F.A. Farhoud and M. Ibrahim, The mathematical model of reflection and refraction of plane quasi-vertical transverse waves at interface nano-composite smart material, *Journal of Computational and Theoretical Nanoscience* 8 (7) pp.1193-1202 (2011).

[28] A.N. Abd-alla, F.A. Alshaikh, Reflection and refraction of plane quasi-longitudinal waves at an interface of two piezoelectric media under initial stresses, *Archive of Applied Mechanics* 79 (9) pp. 843-857 (2009).

[29] A.N. Abd-alla, F.A. Alshaikh, The effect of the initial stresses on the reflection and transmission of plane quasi-vertical transverse waves in piezoelectric materials, *World Academy of Science, Engineering and Technology* 38 pp. 660-668 (2009).

Table 1 Nomenclature

ρ	mass density
\mathbf{u}	mechanical displacement
ε	strain tensor
σ	stress tensor
\mathbf{c}	fourth order tensor tiffness

\mathbf{B}	magnetic induction vector
\mathbf{H}	Total magnetic field vector
\mathbf{E}	electric field vector
\mathbf{f}	Electromagnetic body force (Lorentz force)
\mathbf{J}	electric current density vector
\mathbf{H}_o	Perturbed magnetic field vector \mathbf{h}
T	temperature
T_0	reference uniform temperature of the body
\mathbf{K}	heat conduction second order tensor
$\mathbf{\Omega}$	Rotation vector
γ	thermal elastic coupling tensor
C_e	specific heat at constant strain
t_0, t_1	relaxation times
\mathbf{n}	vector of the symmetric axis
\mathbf{N}	second order tensor $\mathbf{n} \otimes \mathbf{n}$
\mathbf{I}	identity tensor of second order
$\lambda, \mu, \mu_0,$	
α, β	elastic constitutive parameters
μ_e	Magnetic permeability of the medium
σ_o	Electric conductivity of the medium
ω	Circular frequency
C	complex phase velocity
θ	Angle of propagation measured from the normal to the half-space

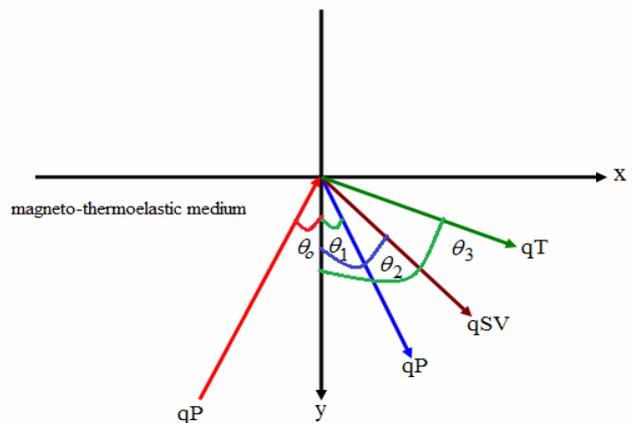


Fig. 1 Geometry of the problem

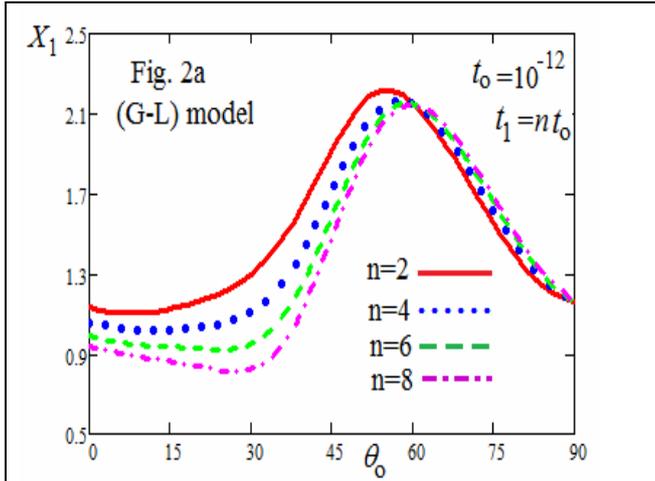


Fig.2a The reflection coefficient X_1 versus θ_o for (G-L) model when $\Omega' = 4$, $H_o = 10^5$ and the variation of the thermal relaxation time with ($n=2,4,6,8$).

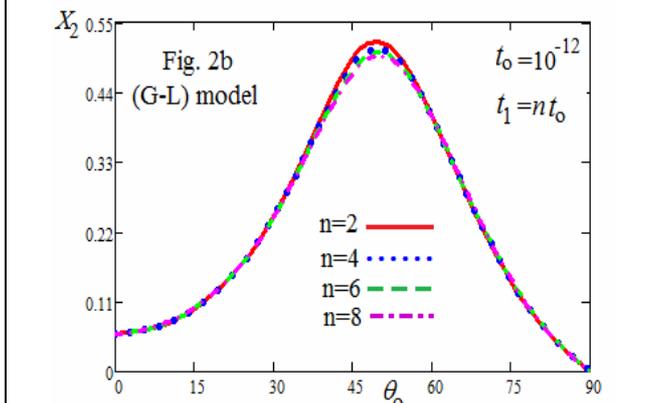


Fig.2b The reflection coefficient X_2 versus θ_o for (G-L) model when $\Omega' = 4$, $H_o = 10^5$ and the variation of the thermal relaxation time with ($n=2,4,6,8$).

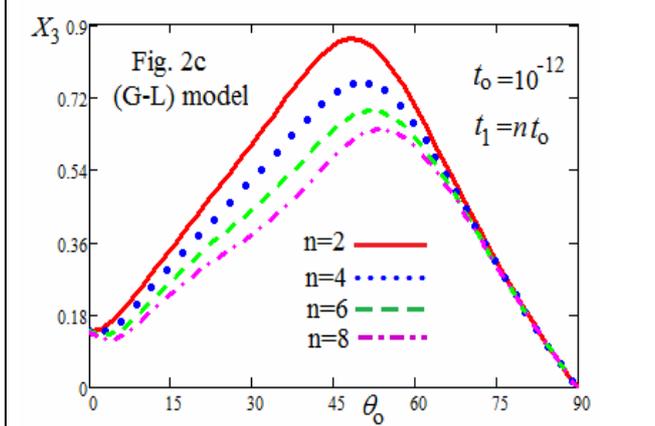


Fig.2c The reflection coefficient X_3 versus θ_o for (G-L) model when $\Omega' = 4$, $H_o = 10^5$ and the variation of the thermal relaxation time with ($n=2,4,6,8$).

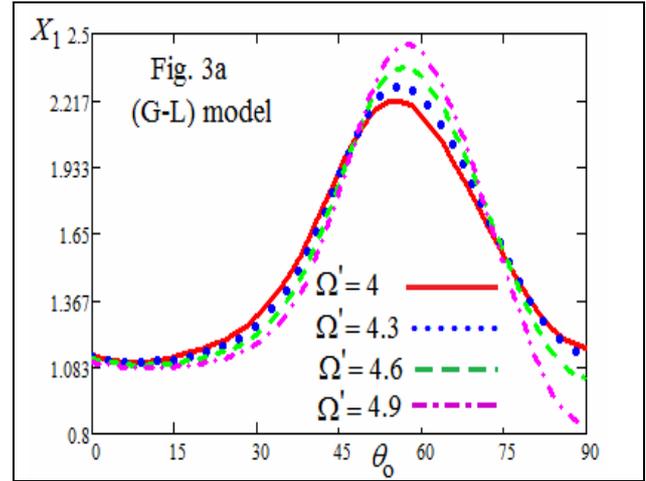


Fig.3a The reflection coefficient X_1 versus θ_o for (G-L) model when $H_o = 10^5$, $t_1 = 5 \times 10^{-12}$ and the variation of the dimensionless rotation $\Omega' = 4, 4.3, 4.6, 4.9$.

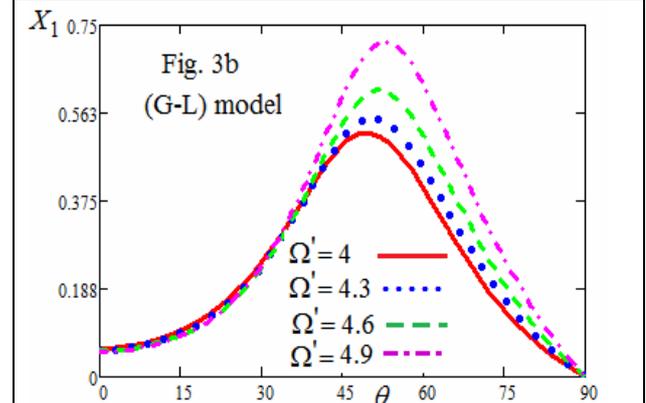


Fig.3b The reflection coefficient X_2 versus θ_o for (G-L) model when $H_o = 10^5$, $t_1 = 5 \times 10^{-12}$ and the variation of the dimensionless rotation $\Omega' = 4, 4.3, 4.6, 4.9$.

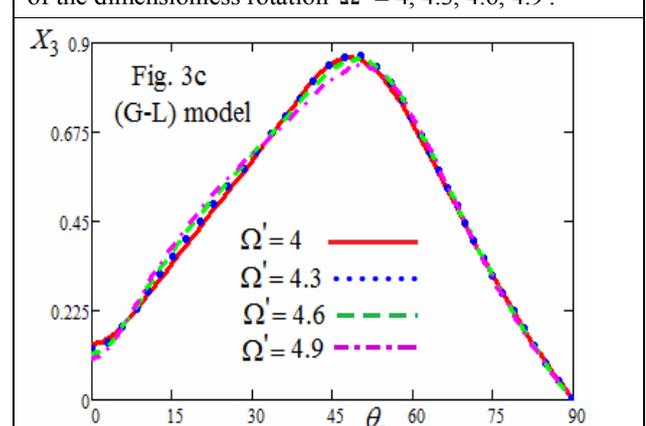


Fig.3c The reflection coefficient X_3 versus θ_o for (G-L) model when $H_o = 10^5$, $t_1 = 5 \times 10^{-12}$ and the variation of the dimensionless rotation $\Omega' = 4, 4.3, 4.6, 4.9$.

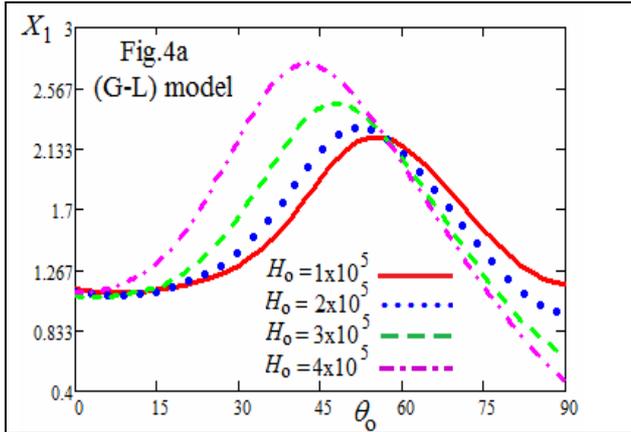


Fig.4a The reflection coefficient X_1 versus θ_o for (G-L) model when $\Omega' = 4$, $t_1 = 5 \times 10^{-12}$ and the variation of the imposed magnetic field $H_o = (1, 2, 3, 4) \times 10^5$.

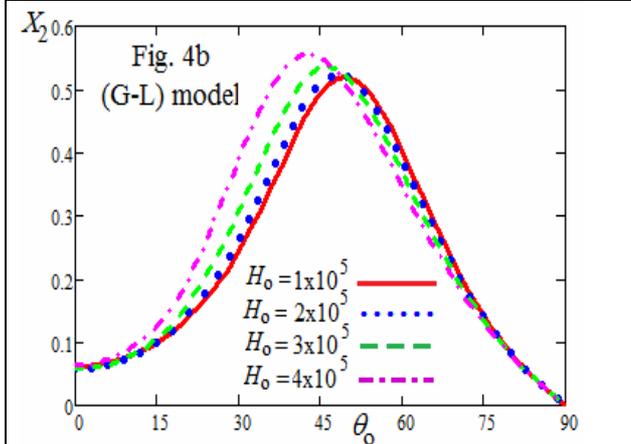


Fig.4b The reflection coefficient X_2 versus θ_o for (G-L) model when $\Omega' = 4$, $t_1 = 5 \times 10^{-12}$ and the variation of the imposed magnetic field $H_o = (1, 2, 3, 4) \times 10^5$.

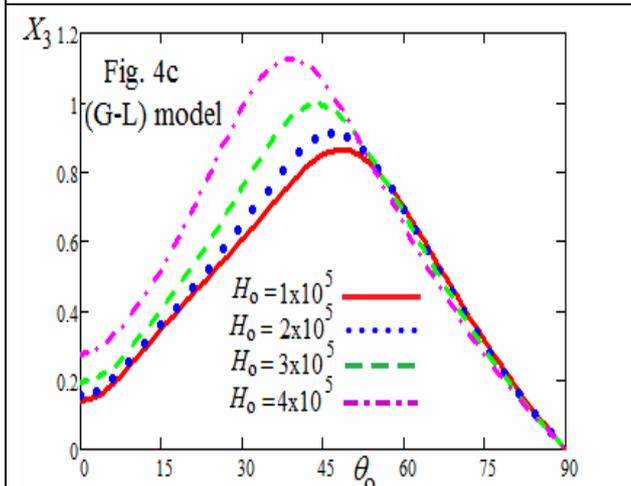


Fig.4c The reflection coefficient X_3 versus θ_o for (G-L) model when $\Omega' = 4$, $t_1 = 5 \times 10^{-12}$ and the variation of the imposed magnetic field $H_o = (1, 2, 3, 4) \times 10^5$.

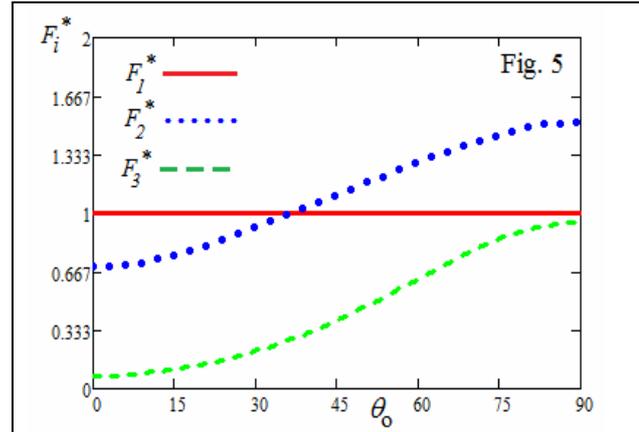


Figure 5 shows the changes of the parameters of proportionality F_1^* , F_2^* and F_3^* as a function of the angle of incident θ_o .

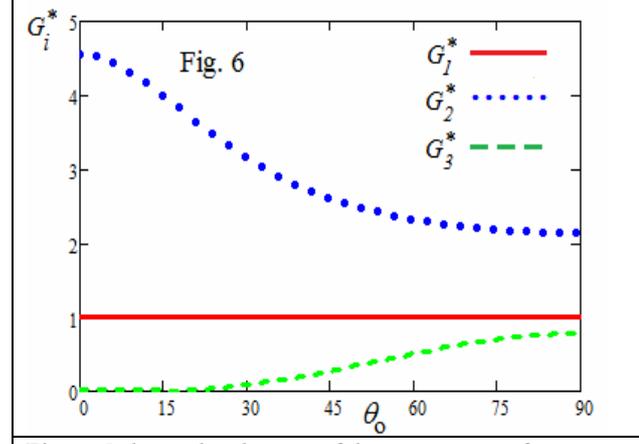


Figure 5 shows the changes of the parameters of proportionality G_1^* , G_2^* and G_3^* as a function of the angle of incident θ_o .

Prediction and evaluation of response to breast cancer chemotherapy by use of multifractal analysis

Jelena Vasiljevic, Jelena Pribic, Ksenija Kanjer, Wojtek Jonakowski, Jelena Sopta, Dragica Nikolic Vukosavljevic and Marko Radulovic

Abstract—Neoadjuvant chemotherapy increases survival of patients with locally advanced breast cancer. Very accurate predictors of chemotherapy response are essential for effective chemotherapeutic management due to the pronounced individual heterogeneity in breast cancer. Predictive molecular determinants for conventional chemotherapy are only emerging and still incorporate a high degree of predictive variability. Taken together, there is a pressing need for improvements of predictive performance. We addressed this issue by exploring the value of tumour histology image analysis as a novel tool for prediction and evaluation of chemotherapy response. Fractal analysis was applied to hematoxylin/eosin stained archival diagnostic breast tumour biopsies derived from 106 patients diagnosed with invasive breast cancer and treated with anthracycline. Based on the results it is concluded that multifractal analysis of breast tumour tissue prior to chemotherapy can predict the pathological complete response, partial pathological response and progressive/stable disease with accuracies ranging from 82% - 91%. Multifractal comparison of tumour sections before and after chemotherapy, also based on 350 representative histology images for each group, has achieved a discrimination accuracy between the groups of 73%. This study indicates for the first time the potential value of multifractal analysis as a simple and cost-effective quantitative clinical tool for prediction and evaluation of chemotherapy response.

Keywords—breast cancer, chemotherapy, histology, prediction

I. INTRODUCTION

Death by breast cancer is mainly caused by a metastatic relapse at distant sites. For that reason, besides local surgery and radiotherapy, patients are also administered postoperative systemic therapy with the aim to reduce the risk of recurrence and death through eradicating distant metastatic deposits. Patients with advanced breast cancer are additionally administered a preoperative (neoadjuvant) chemotherapy. Due to the high heterogeneity of breast cancer, the individual sensitivity to a particular chemotherapy is also highly heterogeneous. By determining the right treatment for each individual patient, predictive markers reduce relapse rates and prolong survival in invasive breast cancer.

Hormonal receptors and c-erbB2 expression present obvious markers for decisions on hormonal and erbB2-targeted therapies [1]. Unfortunately, the choice of

a predictive marker is far less straightforward for conventional chemotherapy

A number of molecular and pathological predictive tools are used in both prediction and evaluation of breast cancer chemotherapy response, including the proliferation marker Ki67 [2], PET/CT imaging [3], MRI imaging [4], histopathological tumour examination, and residual tumor size [5]. Even the most recent candidate predictors of tumour chemotherapy response such as microRNAs, proliferation index, TIMP-1, Lin-28 and gene panels Oncotype DX, MammaPrint and immune-related gene signatures still exhibit predictive variability with consequent uncertain therapeutic guidance [6-8].

In parallel to these molecular predictive methods, digital pathology emerges as a tool to analyse histology images, based on the fact that morphological changes of tumour tissue reflect the sum effect of a very large number of molecular changes that may be difficult or even impossible to fully acquire and interpret by use of available molecular methods. Tumour tissue histology image analysis may thus present a convenient readout of the molecular alterations in breast cancer. Multifractal analysis is one of the digital pathology approaches, known as powerful morphometry tool for quantitative assessment of complex pathological structures [9]. However, its potential use in breast cancer therapy prediction and evaluation has not been investigated. We thus hypothesized that multifractal analysis of tumour histology may prove useful for improvement of the currently poor accuracy of chemotherapy efficacy prediction and evaluation methods. This task was approached by a neoadjuvant therapy model which has been accepted as an ideal *in vivo* assessment of therapy response because the tumour remains *in situ* throughout treatment, thus allowing the exact evaluation of the chemotherapy response [8].

II. METHODS

Criteria for the selection of patients for this retrospective study were as follows: 1) an incisional biopsy of the primary breast cancer confirming invasive carcinoma before commencing the treatment and 2) primary locally advanced breast cancer that was strictly not operable.

Prior to surgery, all patients were treated with standard anthracycline-based chemotherapy (5-

fluorouracil 500 mg/m², doxorubicin 50 mg/m² and cyclophosphamide 500 mg/m² intravenously). Breast tumour response was evaluated after chemotherapy completion by pathohistological examination of the resected surgical material including measurement of the residual tumor size, optical microscopy and immunohistochemical analysis according to recommendations of International Expert Panel [5], as previously described in detail [8].

Tumour tissue sections were cut at 5-µm thickness from the paraffin blocks and stained with haematoxylin/eosin stain as described [8]. Representative tissue sections were selected for each patient by a pathologist and digital microscopic images acquired at x400 magnification using Olympus BX-51 light microscope and a mounted Olympus digital camera. Multifractal analysis of binary digital medical images was performed by use of FracLac software and the obtained parameters subsequently analysed statistically for differences between the three groups of tissues by use of DTREG 10.3.0. A model of Single Tree of the classification type was performed.

Split-sample cross-validation was performed by randomly splitting the original sample into training set and validation sets in ten validation cycles by use of DTREG software.

III. RESULTS AND DISCUSSION

The value of multifractal analysis in prediction and evaluation of the chemotherapy response was approached by a neoadjuvant chemotherapy (CT) model as presented on Fig. 1.

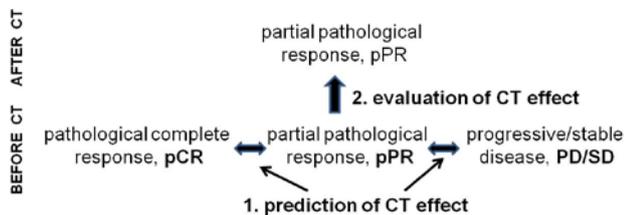


Fig. 1 the conceptual scheme of patient group comparisons by use of multifractal analysis

Due to the fact that this study was retrospective, histology images of tumours before the therapy were divided in three response categories as indicated on Fig. 1 according to their actual response to chemotherapy: pathological complete response (pCR), partial pathological response (pPR) and progressive/stable disease (PD/SD). Significant discrimination between such groups by multifractal analysis indicates its potential in important clinical tasks of chemotherapy prediction and evaluation. Tumour samples of patients with pCR and PD/SD were not available as these patients did not undergo surgery. The pPR group underwent a tumour extraction surgery and was analysed both before and after chemotherapy in order to test the value multifractal analysis in evaluation of the tumour chemotherapy response (Fig. 1). Multifractal analysis was performed on binary black and white digital

images, derived from original colour images. Examples of the typical analysed histology sections are shown on Fig. 2, indicating the absence of obvious visual differences between the three response groups before therapy. Even a detailed pathological microscopic analysis or molecular predictor tools are often unable to provide clues which could serve as the basis for a reliable prediction of chemotherapy response [2].

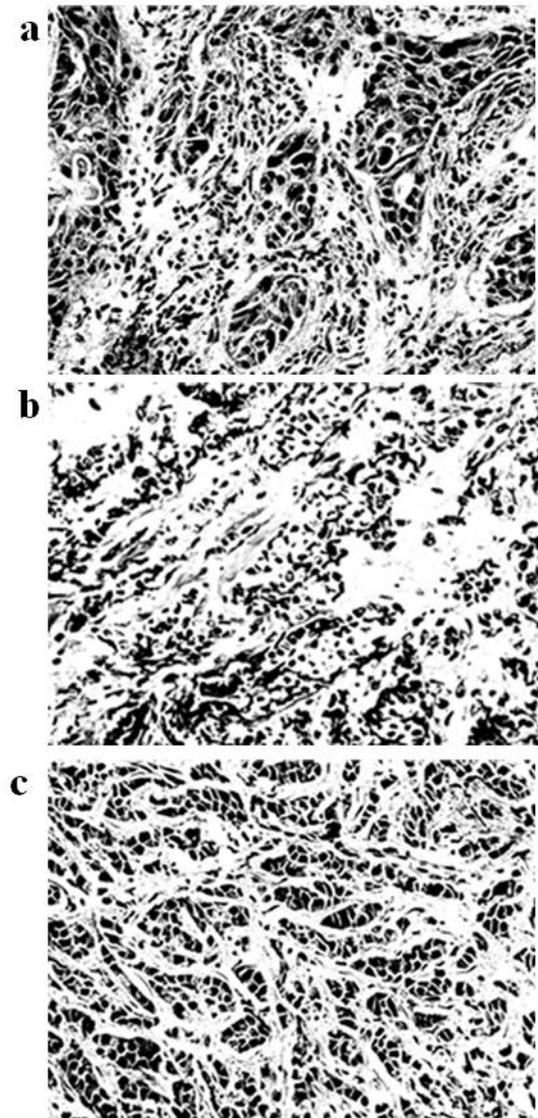


Fig. 2 comparison of pre-therapy breast tumor histological images by multifractal analysis: a) pathological complete response, b) partial pathological response and c) progressive/stable disease groups.

Table 1 indicates that multifractal analysis is able to distinguish between chemotherapy response groups with good accuracy and thus may indeed become useful for the prediction of individual response to chemotherapy. Accuracy represents the percentage of times that the predicted and observed outcomes match [10]. Prediction accuracies of over 82% achieved for the three chemotherapy responder groups (Table 1) are comparable with those previously obtained by Ki67 as the standard

prediction marker [2] and the PET/CT system which even had an advantage of making predictions according to the actual response to the first cycle of chemotherapy [3].

Table 1 Comparison of chemotherapy responder groups by multifractal analysis

	pCR	pPR	PD/SD
Accuracy (%)*	83/75	91/87	82/73

* training/validation groups

A ten-fold internal split-sample cross-validation was performed as the established internal indicator of model's prognostic overoptimism and stability.

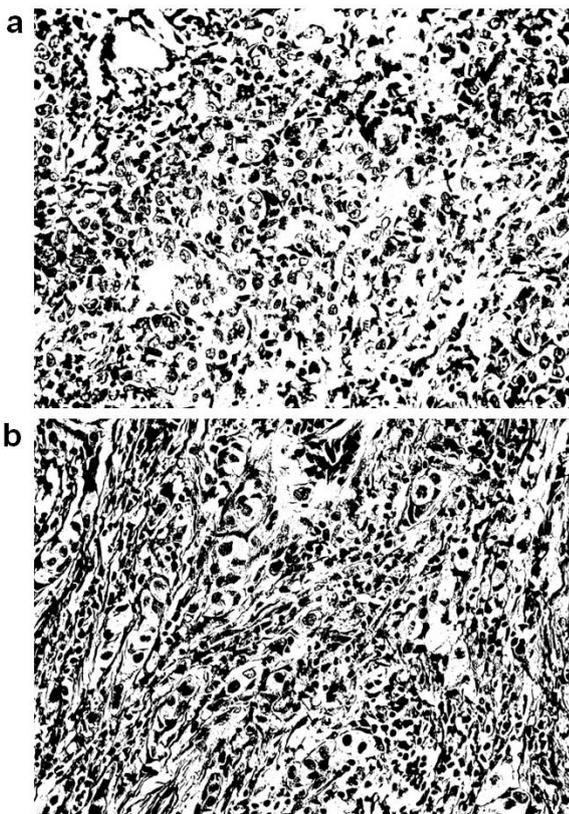


Fig. 3 comparison of the partial response group: a) before and b) after chemotherapy.

Examples of binarized histology images of the same breast tumour from the pPR group are shown on Fig. 3, prior and after chemotherapy (Figs. 3a and 3b, respectively).

It has been established that pathologic complete response (pCR) points to a favourable prognosis [11]. However, within the partial response group the prognosis of disease progression is variable and the extent of chemotherapy response is hard to determine reliably by visual microscopic inspection. We thus examined whether multifractal analysis has the potential to sub-stratify this largest group of chemotherapy responders. Table 2 indicates the potential for use of multifractal analysis in evaluation of chemotherapy response as it differentiates

between tumour histology sections before and after therapy with the accuracy in the training group of 74%

Table 2 Comparison of groups before and after chemotherapy by multifractal analysis

	Training	Validation
Accuracy (%)	74	69

Multifractal analysis delivers a number of parameters, including:

D_{max} - maximum of generalised fractal dimension

$f(\alpha)_{max}$ - maximum of $f(\alpha)$ multifractal spectrum

$\alpha_{f_{max}}$ - α which corresponds to $f(\alpha)_{max}$

$Q_{D_{max}}$ - Q which corresponds to D_{max}

$f(\alpha)_{min}$ - minimum of $f(\alpha)$

$\alpha_{f_{min}}$ - α which corresponds to $f(\alpha)_{min}$

With many available fractal parameters, we set out to identify the one which is most important based on the ability to correctly discriminate between histology images from different patient groups as on Fig. 1. This was achieved by use of variable importance classification by DTREG software. Remarkably, the most important parameters for these two tasks are different, with $f(\alpha)_{max}$ determined as crucial for the prediction and $f(\alpha)_{min}$ for the evaluation of chemotherapy response (Table 3).

Table 3 The most important multifractal parameters for the tasks of chemotherapy prediction and evaluation

	Prediction	Evaluation
Parameter	$f(\alpha)_{max}$	$f(\alpha)_{min}$

Interpretation of this result is based on the fact that multifractal analysis describes structure features from both local and global points of view. The local regularity is described by the Hölder exponent (α) while global regularity is reflected in the multifractal spectrum $f(\alpha)$ which describes the distribution of α [12]. High values of Hölder exponent (α) reflect high local changes of the observed structure around a given point [13]. $f(\alpha)_{max}$ and $f(\alpha)_{min}$ are the parameters of the multifractal spectrum, with $f(\alpha)_{max}$ denoting the fractal dimensions with the maximum probability, while $f(\alpha)_{min}$ denotes the rarest value of α . In view of these facts and the variable importance classification result it can be concluded that prediction and evaluation of chemotherapy response by multifractal analysis are based on distinct global properties of the tumour histology image.

The importance of quantitative imaging for the prediction of chemotherapy sensitivity is based on the fact that improvements in predicting chemotherapy complete response versus non-complete response allow more optimal chemotherapy decisions, thus affecting the quality of life and survival. On the other hand, improvements of the evaluation of chemotherapy effects facilitate clinical testing of new chemotherapeutics and enable more accurate prognosis of survival.

We assume that prognostic value of multifractal analysis derives from its capacity to extract yet unidentified microscopic qualities of a tumour. It may be speculated that among such known clues are the signs of apoptosis, occurrence and distribution of mitotic cells, vascularization, cellularity, tissue growth patterns and probably other unknown qualities [14-16]. A similar set of histological clues may also be responsible for a difference in multifractal scoring of pre- and post-chemotherapy histological images.

IV. CONCLUSION

Improvements of prediction and evaluation of chemotherapy efficacy are of high clinical relevance due to the major impact of chemotherapy on quality of life and survival. We hypothesized that multifractal analysis could provide a valuable addition to existing clinicopathological and molecular prognosticators, based on its powerful discriminating morphometric quantitative description of the irregular structures typical of tumour growth. It is here shown for the first time that multifractal analysis of breast tumour tissue has a potential of aiding both in the prediction and evaluation of response to chemotherapy. The two effects were based on the recognition of distinct global fractal properties of the tumour histological image. Cost-effectiveness of this method derives from rapid analysis of standard clinical material and presents a clear advantage over the methods which are currently used for these clinical tasks.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education and Science, Republic of Serbia, Science and Technological Development grants OI-175059, III 45005, TR 32025 and TR32037.

REFERENCES

- [1] A. Goldhirsch, J. H. Glick, R. D. Gelber, A. S. Coates, B. Thurlimann, H. J. Senn, "Meeting highlights: international expert consensus on the primary therapy of early breast cancer 2005", *Ann Oncol*, 16, 2005, pp. 1569-1583.
- [2] A. Sueta, Y. Yamamoto, M. Hayashi, S. Yamamoto, T. Inao, M. Ibusuki, K. Murakami, H. Iwase, "Clinical significance of pretherapeutic Ki67 as a predictive parameter for response to neoadjuvant chemotherapy in breast cancer: is it equally useful across tumor subtypes?", *Surgery*, 155, 2014, pp. 927-935.
- [3] S. M. Lee, S. K. Bae, T. H. Kim, H. K. Yoon, S. J. Jung, J. S. Park, C. K. Kim, "Value of 18F-FDG PET/CT for early prediction of pathologic response (by residual cancer burden criteria) of locally advanced breast cancer to neoadjuvant chemotherapy", *Clinical Nuclear Medicine*, 39, 2014, pp. 882-886.
- [4] J. H. Chen, S. Bahri, R. S. Mehta, A. Kuzucan, H. J. Yu, P. M. Carpenter, S. A. Feig, M. Lin, D. J. Hsiang, K. T. Lane, J. A. Butler, O. Nalcioglu, M. Y. Su, "Breast cancer: evaluation of response to neoadjuvant chemotherapy with 3.0-T MR imaging", *Radiology*, 261, 2011, pp. 735-743.
- [5] M. Kaufmann, G. N. Hortobagyi, A. Goldhirsch, S. Scholl, A. Makris, P. Valagussa, J. U. Blohmer, W. Eiermann, R. Jackesz, W. Jonat, A. Lebeau, S. Loibl, W. Miller, S. Seeber, V. Semiglazov, R. Smith, R. Souchon, V. Stearns, M. Untch, G. von Minckwitz, "Recommendations from an international expert panel on the use of neoadjuvant (primary) systemic treatment of operable breast cancer: an update", *J Clin Oncol*, 24, 2006, pp. 1940-1949.
- [6] Y. Sota, Y. Naoi, R. Tsunashima, N. Kagara, K. Shimazu, N. Maruyama, A. Shimomura, M. Shimoda, K. Kishi, Y. Baba, S. J. Kim, S. Noguchi, "Construction of novel immune-related signature for prediction of pathological complete response to neoadjuvant chemotherapy in human breast cancer", *Ann Oncol*, 25, 2014, pp. 100-106.
- [7] R. Rouzier, P. Pronzato, E. Chereau, J. Carlson, B. Hunt, W. J. Valentine, "Multigene assays and molecular markers in breast cancer: systematic review of health economic analyses", *Breast Cancer Research and Treatment*, 139, 2013, pp. 621-637.
- [8] K. Kanjer, S. Tatic, Z. Neskovic-Konstantinovic, Z. Abu Rabi, D. Nikolic-Vukosavljevic, "Treatment response to preoperative anthracycline-based chemotherapy in locally advanced breast cancer: the relevance of proliferation and apoptosis rates", *Pathol Oncol Res*, 19, 2013, pp. 577-588.
- [9] P. W. Huang, C. H. Lee, "Automatic classification for pathological prostate images based on fractal analysis", *IEEE Trans Med Imaging*, 28, 2009, pp. 1037-1050.
- [10] A. C. Justice, K. E. Covinsky, J. A. Berlin, "Assessing the generalizability of prognostic information", *Annals of Internal Medicine*, 130, 1999, pp. 515-524.
- [11] W. F. Symmans, F. Peintinger, C. Hatzis, R. Rajan, H. Kuerer, V. Valero, L. Assad, A. Poniecka, B. Hennessy, M. Green, A. U. Buzdar, S. E. Singletary, G. N. Hortobagyi, L. Pusztai, "Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy", *J Clin Oncol*, 25, 2007, pp. 4414-4422.
- [12] C. J. Evertsz, B. B. Mandelbrot, L. Woog, "Variability of the form and of the harmonic measure for small off-off-lattice diffusion-limited aggregates", *Phys Rev A*, 45, 1992, pp. 5798-5804.
- [13] B. Reljin, M. Paskas, I. Reljin, K. Konstanty, "Breast cancer evaluation by fluorescent dot detection using combined mathematical morphology and multifractal techniques", *Diagn Pathol*, 6 Suppl 1, 2011, pp. S21.
- [14] I. Pantic, S. Pantic, G. Basta-Jovanovic, "Gray level co-occurrence matrix texture analysis of germinal center light zone lymphocyte nuclei: physiology viewpoint with focus on apoptosis", *Microsc Microanal*, 18, 2012, pp. 470-475.
- [15] G. A. Losa, C. Castelli, "Nuclear patterns of human breast cancer cells during apoptosis: characterisation by fractal dimension and co-occurrence matrix statistics", *Cell and Tissue Research*, 322, 2005, pp. 257-267.

[16] W. Huang, X. Li, Y. Chen, M. C. Chang, M. J. Oborski, D. I. Malyarenko, M. Muzi, G. H. Jajamovich, A. Fedorov, A. Tudorica, S. N. Gupta, C. M. Laymon, K. I. Marro, H. A. Dyvorne, J. V. Miller, D. P. Barbodiak, T. L. Chenevert, T. E. Yankeelov, J. M. Mountz, P. E. Kinahan, R. Kikinis, B. Taouli,

F. Fennessy, J. Kalpathy-Cramer, "Variations of dynamic contrast-enhanced magnetic resonance imaging in evaluation of breast cancer therapy response: a multicenter data analysis challenge", *Transl Oncol*, 7, 2014, pp. 153-166.

Exponential stability of linear hybrid systems with interval time-varying delays

Grienggrai Rajchakit

Department of Mathematics
Maejo University
Chiangmai, Thailand
Email: kreangkri@mju.ac.th

Abstract—This paper is concerned with exponential stability of switched linear systems with interval time-varying delays. The time delay is any continuous function belonging to a given interval, in which the lower bound of delay is not restricted to zero. By constructing a suitable augmented Lyapunov-Krasovskii functional combined with Leibniz-Newton's formula, a switching rule for the exponential stability of switched linear systems with interval time-varying delays and new delay-dependent sufficient conditions for the exponential stability of the systems are first established in terms of LMIs.

I. INTRODUCTION

Switched time-delay systems have been attracting considerable attention during the recent years [1-10], due to the significance both in theory development and practical applications. However, it is worth noting that only the state time delay is considered, and the time delay in the state derivatives is largely ignored in the existing literature. If each subsystem of a switched system has time delay in the state derivatives, then the switched system is called switched neutral system [10–14]. Switched neutral systems exist widely in engineering and social systems, many physical plants can be modelled as switched neutral systems, such as distributed networks and heat exchanges. For example, in [11–16], a switched neutral type delay equation with nonlinear perturbations was exploited to model the drilling system. Unlike other systems, the neutral has time-delay in both the state and derivative. However, it is well-known that time-delay in the system may be a source of instability or bad system performance. Thus many researchers try to study them to find stability criteria for such system with time-delay to be stable. Most of the known results on this problem are derived assuming only that the time-varying delay $h(t)$ is a continuously differentiable function, satisfying some boundedness condition on its derivative: $\dot{h}(t) \leq \delta < 1$. This paper gives the improved results for the exponential stability of switched linear systems with interval time-varying delay. The time delay is assumed to be a time-varying continuous function belonging to a given interval, but not necessary to be differentiable. Specifically, our goal is to develop a constructive way to design switching rule to the exponential stability of switched linear systems with interval time-varying delay. By constructing argument Lyapunov functional combined with LMI technique, we propose new criteria for the exponential

stability of the switched linear system. The delay-dependent stability conditions are formulated in terms of LMIs.

The paper is organized as follows: Section II presents definitions and some well-known technical propositions needed for the proof of the main results. Delay-dependent exponential stability conditions of the switched linear system are presented in Section III.

II. PRELIMINARIES

The following notations will be used in this paper. R^+ denotes the set of all real non-negative numbers; R^n denotes the n -dimensional space with the scalar product $\langle \cdot, \cdot \rangle$ and the vector norm $\| \cdot \|$; $M^{n \times r}$ denotes the space of all matrices of $(n \times r)$ -dimensions; A^T denotes the transpose of matrix A ; A is symmetric if $A = A^T$; I denotes the identity matrix; $\lambda(A)$ denotes the set of all eigenvalues of A ; $\lambda_{\min/\max}(A) = \min/\max\{\text{Re}\lambda; \lambda \in \lambda(A)\}$; $x_t := \{x(t+s) : s \in [-h, 0]\}$, $\|x_t\| = \sup_{s \in [-h, 0]} \|x(t+s)\|$; $C([0, t], R^n)$ denotes the set of all R^n -valued continuous functions on $[0, t]$; Matrix A is called semi-positive definite ($A \geq 0$) if $\langle Ax, x \rangle \geq 0$, for all $x \in R^n$; A is positive definite ($A > 0$) if $\langle Ax, x \rangle > 0$ for all $x \neq 0$; $A > B$ means $A - B > 0$. $*$ denotes the symmetric term in a matrix.

Consider a linear system with interval time-varying delay of the form

$$\begin{aligned} \dot{x}(t) &= A_{\gamma}x(t) + D_{\gamma}x(t - h(t)), \quad t \in R^+, \\ x(t) &= \phi(t), \quad t \in [-h_2, 0], \end{aligned} \quad (1)$$

where $x(t) \in R^n$ is the state; $\gamma(\cdot) : R^n \rightarrow \mathcal{N} := \{1, 2, \dots, N\}$ is the switching rule, which is a function depending on the state at each time and will be designed. A switching function is a rule which determines a switching sequence for a given switching system. Moreover, $\gamma(x(t)) = i$ implies that the system realization is chosen as the i^{th} system, $i = 1, 2, \dots, N$. It is seen that the system (1) can be viewed as an autonomous switched system in which the effective subsystem changes when the state $x(t)$ hits predefined boundaries. $A_i, D_i \in M^{n \times n}$, $i = 1, 2, \dots, N$ are given constant matrices, and $\phi(t) \in C([-h_2, 0], R^n)$ is the initial function with the norm

$\|\phi\| = \sup_{s \in [-h_2, 0]} \|\phi(s)\|$; The time-varying delay function $h(t)$ satisfies

$$0 \leq h_1 \leq h(t) \leq h_2, \quad t \in R^+.$$

The stability problem for switched system (1) is to construct a switching rule that makes the system exponentially stable.

Remark 2.1. It is worth noting that the time delay is a time-varying function belonging to a given interval, in which the lower bound of delay is not restricted to zero.

Definition 2.1. Given $\alpha > 0$. The switched linear system (1) is α -exponentially stable if there exists a switching rule $\gamma(\cdot)$ such that every solution $x(t, \phi)$ of the system satisfies the following condition

$$\exists N > 0 : \|x(t, \phi)\| \leq N e^{-\alpha t} \|\phi\|, \quad \forall t \in R^+.$$

We end this section with the following technical well-known propositions, which will be used in the proof of the main results.

Definition 2.2. The system of matrices $\{J_i\}, i = 1, 2, \dots, N$, is said to be strictly complete if for every $x \in R^n \setminus \{0\}$ there is $i \in \{1, 2, \dots, N\}$ such that $x^T J_i x < 0$.

It is easy to see that the system $\{J_i\}$ is strictly complete if and only if

$$\bigcup_{i=1}^N \alpha_i = R^n \setminus \{0\},$$

where

$$\alpha_i = \{x \in R^n : x^T J_i x < 0\}, i = 1, 2, \dots, N.$$

We end this section with the following technical well-known propositions, which will be used in the proof of the main results.

Proposition 2.1. [17] *The system $\{J_i\}, i = 1, 2, \dots, N$, is strictly complete if there exist $\delta_i \geq 0, i = 1, 2, \dots, N, \sum_{i=1}^N \delta_i > 0$ such that*

$$\sum_{i=1}^N \delta_i J_i < 0.$$

If $N = 2$ then the above condition is also necessary for the strict completeness.

Proposition 2.2. (Cauchy inequality) *For any symmetric positive definite matrix $N \in M^{n \times n}$ and $a, b \in R^n$ we have*

$$\pm a^T b \leq a^T N a + b^T N^{-1} b.$$

Proposition 2.3. [18] *For any symmetric positive definite matrix $M \in M^{n \times n}$, scalar $\gamma > 0$ and vector function $\omega : [0, \gamma] \rightarrow R^n$ such that the integrations concerned are well defined, the following inequality holds*

$$\left(\int_0^\gamma \omega(s) ds \right)^T M \left(\int_0^\gamma \omega(s) ds \right) \leq$$

$$\gamma \left(\int_0^\gamma \omega^T(s) M \omega(s) ds \right).$$

Proposition 2.4. [19] *Let E, H and F be any constant matrices of appropriate dimensions and $F^T F \leq I$. For any $\epsilon > 0$, we have*

$$EFH + H^T F^T E^T \leq \epsilon E E^T + \epsilon^{-1} H^T H.$$

Proposition 2.5. (Schur complement lemma [20]). *Given constant matrices X, Y, Z with appropriate dimensions satisfying $X = X^T, Y = Y^T > 0$. Then $X + Z^T Y^{-1} Z < 0$ if and only if*

$$\begin{pmatrix} X & Z^T \\ Z & -Y \end{pmatrix} < 0 \quad \text{or} \quad \begin{pmatrix} -Y & Z \\ Z^T & X \end{pmatrix} < 0.$$

III. MAIN RESULTS

Let us set

$$\mathcal{M}_i = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} & M_{15} \\ * & M_{22} & 0 & M_{24} & S_2 \\ * & * & M_{33} & M_{34} & S_3 \\ * & * & * & M_{44} & M_{45} \\ * & * & * & * & M_{55} \end{bmatrix},$$

$$J_i = Q - S_1 A_i - A_i^T S_1^T, \tag{2}$$

$$\alpha_i = \{x \in R^n : x^T J_i x < 0\}, \quad i = 1, 2, \dots, N,$$

$$\bar{\alpha}_1 = \alpha_1, \quad \bar{\alpha}_i = \alpha_i \setminus \bigcup_{j=1}^{i-1} \alpha_j, \quad i = 2, 3, \dots, N,$$

$$\lambda_1 = \lambda_{\min}(P),$$

$$\lambda_2 = \lambda_{\max}(P) + 2h_2 \lambda_{\max}(Q),$$

$$M_{11} = A_i^T P + P A_i + 2\alpha P + Q,$$

$$M_{12} = -S_2 A_i, \quad M_{13} = -S_3 A_i,$$

$$M_{14} = P D_i - S_1 D_i - S_4 A_i, \quad M_{15} = S_1 - S_5 A_i,$$

$$M_{22} = -e^{-2\alpha h_1} Q, \quad M_{24} = -S_2 D_i,$$

$$M_{33} = -e^{-2\alpha h_2} Q, \quad M_{34} = -S_3 D_i,$$

$$M_{44} = -S_4 D_i, \quad M_{45} = S_4 - S_5 D_i,$$

$$M_{55} = S_5 + S_5^T.$$

The main result of this paper is summarized in the following theorem.

Theorem 1. *Given $\alpha > 0$. The zero solution of the switched linear system (1) is α -exponentially stable if there exist symmetric positive definite matrices P, Q , and matrices $S_i, i = 1, 2, \dots, 5$ such that satisfying the following conditions*

$$(i) \exists \delta_i \geq 0, i = 1, 2, \dots, N, \quad \sum_{i=1}^N \delta_i > 0 : \sum_{i=1}^N \delta_i J_i < 0.$$

$$(ii) \mathcal{M}_i < 0, \quad i = 1, 2, \dots, N.$$

Moreover, the solution $x(t, \phi)$ of the system satisfies

$$\|x(t, \phi)\| \leq \sqrt{\frac{\lambda_2}{\lambda_1}} e^{-\alpha t} \|\phi\|, \quad \forall t \in R^+.$$

Proof. We consider the following Lyapunov-Krasovskii functional for the system (1)

$$V(t, x_t) = \sum_{i=1}^3 V_i,$$

where

$$\begin{aligned} V_1 &= x^T(t)Px(t), \\ V_2 &= \int_{t-h_1}^t e^{2\alpha(s-t)} x^T(s)Qx(s) ds, \\ V_3 &= \int_{t-h_2}^t e^{2\alpha(s-t)} x^T(s)Qx(s) ds. \end{aligned}$$

It easy to check that

$$\lambda_1 \|x(t)\|^2 \leq V(t, x_t) \leq \lambda_2 \|x_t\|^2, \quad \forall t \geq 0, \quad (3)$$

Taking the derivative of V_1 along the solution of system (1) we have

$$\begin{aligned} \dot{V}_1 &= 2x^T(t)P\dot{x}(t) \\ &= 2x^T(t)[A_i^T P + A_i P]x(t) + 2x^T(t)PD_i x(t-h(t)); \\ \dot{V}_2 &= x^T(t)Qx(t) - e^{-2\alpha h_1} x^T(t-h_1)Qx(t-h_1) - 2\alpha V_2; \\ \dot{V}_3 &= x^T(t)Qx(t) - e^{-2\alpha h_2} x^T(t-h_2)Qx(t-h_2) - 2\alpha V_3. \end{aligned}$$

Therefore, we have

$$\begin{aligned} \dot{V}(\cdot) + 2\alpha V(\cdot) &\leq 2x^T(t)[A_i^T P + A_i P + 2\alpha P + 2Q]x(t) \\ &\quad + 2x^T(t)PD_i x(t-h(t)) \\ &\quad - e^{-2\alpha h_1} x^T(t-h_1)Qx(t-h_1) \\ &\quad - e^{-2\alpha h_2} x^T(t-h_2)Qx(t-h_2). \end{aligned} \quad (4)$$

By using the following identity relation

$$\dot{x}(t) - A_i x(t) - D_i x(t-h(t)) = 0,$$

we have

$$\begin{aligned} &2x^T(t)S_1 \dot{x}(t) - 2x^T(t)S_1 A_i x(t) \\ &\quad - 2x^T(t)S_1 D_i x(t-h(t)) = 0 \\ &2x^T(t-h_1)S_2 \dot{x}(t) - 2x^T(t-h_1)S_2 A_i x(t) \\ &\quad - 2x^T(t-h_1)S_2 D_i x(t-h(t)) = 0 \\ &2x^T(t-h_2)S_3 \dot{x}(t) - 2x^T(t-h_2)S_3 A_i x(t) \\ &\quad - 2x^T(t-h_2)S_3 D_i x(t-h(t)) = 0 \\ &2x^T(t-h(t))S_4 \dot{x}(t) - 2x^T(t-h(t))S_4 A_i x(t) \\ &\quad - 2x^T(t-h(t))S_4 D_i x(t-h(t)) = 0 \\ &2\dot{x}^T(t)S_5 \dot{x}(t) - 2\dot{x}^T(t)S_5 A_i x(t) \\ &\quad - 2\dot{x}^T(t)S_5 D_i x(t-h(t)) = 0 \end{aligned} \quad (5)$$

Adding all the zero items of (5) into (4), we obtain

$$\begin{aligned} \dot{V}(\cdot) + 2\alpha V(\cdot) &\leq x^T(t)[A_i^T P + PA_i + 2\alpha P - S_1 A_i \\ &\quad - A_i^T S_1^T + 2Q]x(t) \\ &\quad + 2x^T(t)[e^{-2\alpha h_1} R - S_2 A_i]x(t-h_1) \\ &\quad + 2x^T(t)[-S_3 A_i]x(t-h_2) + 2x^T(t)[PD_i \\ &\quad - S_1 D_i - S_4 A_i]x(t-h(t)) \\ &\quad + 2x^T(t)[S_1 - S_5 A_i]\dot{x}(t) \\ &\quad + x^T(t-h_1)[-e^{-2\alpha h_1} Q]x(t-h_1) \\ &\quad + 2x^T(t-h_1)[-S_2 D_i]x(t-h(t)) \\ &\quad + 2x^T(t-h_1)S_2 \dot{x}(t) \\ &\quad + x^T(t-h_2)[-e^{-2\alpha h_2} Q]x(t-h_2) \\ &\quad + x^T(t-h_2)[-S_3 D_i]x(t-h(t)) \\ &\quad + 2x^T(t-h_2)S_3 \dot{x}(t) \\ &\quad + x^T(t-h(t))[-S_4 D_i]x(t-h(t)) \\ &\quad + 2x^T(t-h(t))[S_4 - S_5 D_i]\dot{x}(t) \\ &\quad + \dot{x}^T(t)[S_5 + S_5^T]\dot{x}(t) \\ &= x^T(t)J_i x(t) + \zeta^T(t)\mathcal{M}_i \zeta(t), \end{aligned} \quad (6)$$

where

$$\zeta(t) = [x(t), x(t-h_1), x(t-h_2), x(t-h(t)), \dot{x}(t)],$$

$$J_i = Q - S_1 A_i - A_i^T S_1^T,$$

$$\mathcal{M}_i = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} & M_{15} \\ * & M_{22} & 0 & M_{24} & S_2 \\ * & * & M_{33} & M_{34} & S_3 \\ * & * & * & M_{44} & M_{45} \\ * & * & * & * & M_{55} \end{bmatrix},$$

$$\begin{aligned} M_{11} &= A_i^T P + PA_i + 2\alpha P + Q, \\ M_{12} &= -S_2 A_i, \quad M_{13} = -S_3 A_i, \\ M_{14} &= PD_i - S_1 D_i - S_4 A_i, \quad M_{15} = S_1 - S_5 A_i, \\ M_{22} &= -e^{-2\alpha h_1} Q, \quad M_{24} = -S_2 D_i, \\ M_{33} &= -e^{-2\alpha h_2} Q, \quad M_{34} = -S_3 D_i, \\ M_{44} &= -S_4 D_i, \quad M_{45} = S_4 - S_5 D_i, \\ M_{55} &= S_5 + S_5^T. \end{aligned}$$

Therefore, we finally obtain from (6) and the condition (ii) that

$$\dot{V}(\cdot) + 2\alpha V(\cdot) < x^T(t)J_i x(t), \quad \forall i = 1, 2, \dots, N, \quad t \in R^+.$$

We now apply the condition (i) and Proposition 2.1., the system J_i is strictly complete, and the sets α_i and $\bar{\alpha}_i$ by (2) are well defined such that

$$\bigcup_{i=1}^N \alpha_i = R^n \setminus \{0\},$$

$$\bigcup_{i=1}^N \bar{\alpha}_i = R^n \setminus \{0\}, \quad \bar{\alpha}_i \cap \bar{\alpha}_j = \emptyset, i \neq j.$$

Therefore, for any $x(t) \in R^n$, $t \in R^+$, there exists $i \in \{1, 2, \dots, N\}$ such that $x(t) \in \bar{\alpha}_i$. By choosing switching rule as $\gamma(x(t)) = i$ whenever $\gamma(x(t)) \in \bar{\alpha}_i$, from (6) we have

$$\dot{V}(\cdot) + 2\alpha V(\cdot) \leq x^T(t) J_i x(t) < 0, \quad t \in R^+,$$

and hence

$$\dot{V}(t, x_t) \leq -2\alpha V(t, x_t), \quad \forall t \in R^+. \tag{7}$$

Integrating both sides of (7) from 0 to t , we obtain

$$V(t, x_t) \leq V(\phi) e^{-2\alpha t}, \quad \forall t \in R^+.$$

Furthermore, taking condition (3) into account, we have

$$\lambda_1 \|x(t, \phi)\|^2 \leq V(x_t) \leq V(\phi) e^{-2\alpha t} \leq \lambda_2 e^{-2\alpha t} \|\phi\|^2,$$

then

$$\|x(t, \phi)\| \leq \sqrt{\frac{\lambda_2}{\lambda_1}} e^{-\alpha t} \|\phi\|, \quad t \in R^+,$$

which concludes the proof by the Lyapunov stability theorem [21]. To illustrate the obtained result, let us give the following numerical example.

IV. NUMERICAL EXAMPLE

Example 4.1. Consider the following the switched stochastic systems with interval time-varying delay (2.1), where the delay function $h(t)$ is given by

$$h(t) = 0.2 + 1.5329 \sin^2 t,$$

and

$$A_1 = \begin{pmatrix} -2 & 0.1 \\ 0.2 & -2.5 \end{pmatrix}, A_2 = \begin{pmatrix} -2.5 & 0.3 \\ 0.2 & -2.9 \end{pmatrix},$$

$$D_1 = \begin{pmatrix} -0.3 & 0.2 \\ 0.1 & -0.39 \end{pmatrix}, D_2 = \begin{pmatrix} -0.5 & 0.2 \\ 0.1 & -0.4 \end{pmatrix}.$$

It is worth noting that, the delay function $h(t)$ is non-differentiable and the exponent $\alpha \geq 1$. Therefore, the methods used in [3, 8, 9, 11 – 15, 17 – 23] are not applicable to this system. By LMI toolbox of Matlab, we find that the conditions (i), (ii) of Theorem 3.1 are satisfied with $h_1 = 0.1, h_2 = 1.7329, \delta_1 = 0.5, \delta_2 = 0.3, \alpha = 0.5, \rho_{11} = 0.1, \rho_{12} = 0.2, \rho_{21} = 0.1, \rho_{22} = 0.2$ and

$$P = \begin{pmatrix} 1.2397 & -0.3984 \\ -0.3984 & 1.3112 \end{pmatrix}, Q = \begin{pmatrix} 1.7931 & -0.0079 \\ -0.0079 & 0.2397 \end{pmatrix},$$

$$R = \begin{pmatrix} 2.3297 & -0.1121 \\ -0.1121 & 1.3397 \end{pmatrix}, U = \begin{pmatrix} 1.7394 & -0.0982 \\ -0.0982 & 0.6321 \end{pmatrix},$$

$$S_1 = \begin{pmatrix} -0.6210 & -0.0335 \\ 0.0499 & -0.3576 \end{pmatrix}, S_2 = \begin{pmatrix} -0.3602 & 0.0170 \\ 0.0298 & -0.3550 \end{pmatrix},$$

$$S_3 = \begin{pmatrix} -0.3602 & 0.0170 \\ 0.0298 & -0.3550 \end{pmatrix}, S_4 = \begin{pmatrix} 0.6968 & -0.0401 \\ -0.0525 & 0.7040 \end{pmatrix},$$

$$S_5 = \begin{pmatrix} -1.4043 & 0.0265 \\ -0.0028 & -0.9774 \end{pmatrix}.$$

In this case, we have

$$(J_1, J_2) = \left(\begin{pmatrix} -1.5667 & -0.0031 \\ -0.0031 & -1.9712 \end{pmatrix}, \begin{pmatrix} -1.5511 & 0.0029 \\ 0.0029 & -1.3297 \end{pmatrix} \right).$$

Moreover, the sum

$$\delta_1 J_1(R, Q) + \delta_2 J_2(R, Q) = \begin{bmatrix} -0.3269 & 0 \\ 0 & -0.7239 \end{bmatrix}$$

is negative definite; i.e. the first entry in the first row and the first column $-0.3269 < 0$ is negative and the determinant of the matrix is positive. The sets α_1 and α_2 are given as

$$\alpha_1 = \{(x_1, x_2) : -1.5667x_1^2 - 0.0062x_1x_2 - 1.9712x_2^2 < 0\},$$

$$\alpha_2 = \{(x_1, x_2) : 1.5511x_1^2 - 0.0058x_1x_2 + 1.3297x_2^2 > 0\}.$$

Obviously, the union of these sets is equal to $R^2 \setminus \{0\}$. The switching regions are defined as

$$\bar{\alpha}_1 = \{(x_1, x_2) : -1.5667x_1^2 - 0.0062x_1x_2 - 1.9712x_2^2 < 0\},$$

$$\bar{\alpha}_2 = \alpha_2 \setminus \bar{\alpha}_1.$$

By Theorem 3.1 the switched stochastic systems (2.1) is 0.5–exponentially stable in the mean square and the switching rule is chosen as $\gamma(x(t)) = i$ whenever $x(t) \in \bar{\alpha}_i$. Moreover, the solution $x(t, \phi)$ of the system satisfies

$$E \{\|x(t, \phi)\|\} \leq E \{1.0239 e^{-0.5t} \|\phi\|\}, \quad \forall t \in R^+.$$

V. CONCLUSION

This paper has proposed a switching design for the exponential stability of switched linear systems with interval time-varying delays. Based on the improved Lyapunov-Krasovskii functional, a switching rule for the exponential stability for the system is designed via linear matrix inequalities.

ACKNOWLEDGMENT

This work was supported by the Office of Agricultural Research and Extension Maejo University Chiangmai Thailand, the Thailand Research Fund Grant, the Higher Education Commission and Faculty of Science, Maejo University, Thailand (TRG5780203).

REFERENCES

- [1] M. C. de Oliveira, J. C. Geromel, Liu Hsu, LMI characterization of structural and robust stability: the discrete-time case, *Linear Algebra and its Applications*, **296**(1999), 27-38.
- [2] V.N. Phat and P.T. Nam (2007), Exponential stability and stabilization of uncertain linear time-varying systems using parameter dependent Lyapunov function. *Int. J. of Control*, **80**, 1333-1341.
- [3] V.N. Phat and P. Niamsup, Stability analysis for a class of functional differential equations and applications. *Nonlinear Analysis: Theory, Methods & Applications* **71**(2009), 6265-6275.
- [4] V.N. Phat, T. Bormat and P. Niamsup, Switching design for exponential stability of a class of nonlinear hybrid time-delay systems, *Nonlinear Analysis: Hybrid Systems*, **3**(2009), 1-10.

- [5] Y.J. Sun, Global stabilizability of uncertain systems with time-varying delays via dynamic observer-based output feedback, *Linear Algebra and its Applications*, **353**(2002), 91-105.
- [6] O.M. Kwon and J.H. Park, Delay-range-dependent stabilization of uncertain dynamic systems with interval time-varying delays, *Applied Math. Computation*, **208**(2009), 58-68.
- [7] H. Shao, New delay-dependent stability criteria for systems with interval delay, *Automatica*, **45**(2009), 744-749.
- [8] J. Sun, G.P. Liu, J. Chen and D. Rees, Improved delay-range-dependent stability criteria for linear systems with time-varying delays, *Automatica*, **46**(2010), 466-470.
- [9] W. Zhang, X. Cai and Z. Han, Robust stability criteria for systems with interval time-varying delay and nonlinear perturbations, *J. Comput. Appl. Math.*, 234 (2010), 174-180.
- [10] V.N. Phat, Robust stability and stabilization of uncertain linear hybrid systems with state delays, *IEEE Trans. CAS II*, **52**(2005), 94-98.
- [11] V.N. Phat and P.T. Nam (2007), Exponential stability and stabilization of uncertain linear time-varying systems using parameter dependent Lyapunov function. *Int. J. of Control*, 80, 1333-1341.
- [12] V.N. Phat and P. Niamsup, Stability analysis for a class of functional differential equations and applications. *Nonlinear Analysis: Theory, Methods & Applications* **71**(2009), 6265-6275.
- [13] V.N. Phat and P.T. Nam (2007), Exponential stability and stabilization of uncertain linear time-varying systems using parameter dependent Lyapunov function. *Int. J. of Control*, 80, 1333-1341.
- [14] V.N. Phat and P. Niamsup, Stability analysis for a class of functional differential equations and applications. *Nonlinear Analysis: Theory, Methods & Applications* **71**(2009), 6265-6275.
- [15] V.N. Phat, Y. Khongtham, and K. Ratchagit, LMI approach to exponential stability of linear systems with interval time-varying delays. *Linear Algebra and its Applications* **436**(2012), 243-251.
- [16] K. Ratchagit and V.N. Phat, Stability criterion for discrete-time systems, *J. Ineq. Appl.*, **2010**(2010), 1-6.
- [17] F. Uhlig, A recurring theorem about pairs of quadratic forms and extensions, *Linear Algebra Appl.*, **25**(1979), 219-237.
- [18] K. Gu, An integral inequality in the stability problem of time delay systems, in: *IEEE Control Systems Society and Proceedings of IEEE Conference on Decision and Control*, IEEE Publisher, New York, 2000.
- [19] Y. Wang, L. Xie and C.E. de SOUZA, Robust control of a class of uncertain nonlinear systems. *Syst. Control Lett.*, **19**(1992), 139-149.
- [20] S. Boyd, L.El Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, SIAM, Philadelphia, 1994.
- [21] J.K. Hale and S.M. Verduyn Lunel, *Introduction to Functional Differential Equations*, Springer-Verlag, New York, 1993.

MODELING OF A SMALL UNMANNED AERIAL VEHICLE

AHMED ELSAYED AHMED
Electrical Engineering Dept.,
Shoubra Faculty of Engineering,
Benha University, Qaliuobia,
Egypt (Telephone: +201007124097),
E-mail: eng_medoelbanna@yahoo.com.

HOSSAM ELDIN HUSSEIN AHMED
Communication Engineering Dept., Faculty of
Electronic Engineering, Menoufia University,
Menouf, Egypt (fax: +20483660716),
E-mail: hhossamkh@yahoo.com.

ASHRAF HAFEZ
Electrical Engineering Dept.,
Shoubra Faculty of Engineering,
Benha University, Qaliuobia.
E-mail: ashrafhafez@hotmail.com.

HALA MOHAMED ABD-ELKADER
Electrical Engineering Dept.,
Shoubra Faculty of Engineering,
Benha University, Qaliuobia
E-mail: hala_mansour56@yahoo.com.

A. N. OUDA
Military technical college
E-mail: ahnasroda@yahoo.com.

Abstract— Unmanned aircraft systems (UAS) are playing increasingly prominent roles in defense programs and defense strategies around the world. Technology advancements have enabled the development of it to do many excellent jobs as reconnaissance, surveillance, battle fighters, and communications relays. Simulating a small unmanned aerial vehicle (SUAV) dynamics and analyzing its behavior at the preflight stage is too important and more efficient. The first step in the UAV design is the mathematical modeling of the nonlinear equations of motion. At first of this thesis a survey with a standard method to obtain the full non-linear equations of motion was utilized, and second is the linearization of the equations according to a steady state flight condition (trimming). This modeling technique is applied to an Ultrastick-25e fixed wing UAV to obtain the valued linear longitudinal and lateral models. At the end the behavior of the states of the non-linear UAV and the resulted linear model is checked by applying a doublet signals in the control surfaces to check the matching between them.

Keywords: UAV, equations of motion, modeling, linearization.

I. Introduction.

Dynamic modeling is an important step in the development and the control of a dynamic system as UAV. This model allows the designers to analyze the system; its possibilities and its behavior depending on various conditions. Especially it's so important for aerial robots where the risk of damage is very high as a fall from a few meters; so the platform can be damaged. Thus, the possibility to simulate and tune a controller before implementing it on the aircraft is highly appreciable. This paper introduces a linear mathematical model for a fixed wing UAV by applying the basic steps of modeling.

2nd section is the introduction of the coordinate frames which are used to transfer any rigid body from frame to another, what are the Euler angles (ψ, θ, ϕ), what is Direct Cosine Matrix (DCM). The definition of stability and wind frames. The extracted angles from rotation of object from body frame to the wind frame (angle of attack (AOA), and Sideslip angles (α, β) respectively).

The 3rd section states the basic used parameters of any fixed wing UAV which are used in the aircraft modeling (geometric parameters, inertia, aerodynamic parameters, and control surfaces), these parameters which is obtained from NASA laboratories will be applied to an Ultrastick-25e (Thor) UAV to get the linear mathematical model of this SUAV.

The 4th section introduces a standard algorithm for an aircraft modeling and converting it from a black box into airframe has 12-state nonlinear equations called equations of motion describe the motion of the UAV. Beginning from the kinematics (the motion analysis without the consideration of forces and momentum, and moments), then adding the force and moment terms expressed in the body frame. The forces applied to the aircraft are the gravity forces, aerodynamic forces, propeller forces; these forces are represented in the body frame, and then converted to the stability frame to get the Lift, Drag, and side forces. Then considering the moments and momentum applied to the aircraft. Then catching the nonlinear equations of motion is the end of this section.

At 5th section linearization of the state equations about an equilibrium point (trimming) is applied first how to calculate the trimmed values according to the steady state flight condition, then decomposition of the dynamic equations into two separated groups (longitudinal and lateral dynamics), the linearization technique for every group is applied separately to derive a linear state space model for longitudinal motion ($A_{lon}, B_{lon}, C_{lon}, D_{lon}$), and a linear state

space model for lateral motion (A_{lat} , B_{lat} , C_{lat} , D_{lat}) according to straight and leveling flight trim conditions. Then the linearization of the roll dynamics by analytical technique is utilized. At the last of this section the comparison between two techniques is utilized.

At the last section a comparison between behavior of nonlinear and linear models is done by applying a doublet response to check the matching between them.

II. UAV Coordinate Frames.

In aerospace applications expressing a given vector in terms of a new Cartesian coordinate frame is commonly needed. In this section the descriptions of the various coordinate frames are illustrated starting with Inertial frame. The angles relating the transfer from vehicle frame to the body frame are the yaw (ψ), pitch (θ), and roll (ϕ). These angles describe the attitude of the aircraft, and commonly called Euler angles. The angles relating the rotation between the body to the stability frame vice versa, and Stability frame to the wind frame are called angle of attack (α), and sideslip angle (β) respectively. These coordinates are discussed briefly as follows [1, 2, and 3].

II.a Coordinate Frames.

- **The Inertial Frame (f^i).**

This frame is the earth fixed frame and called NED frame.

- **Vehicle Frame (f^v).**

The axes of the vehicle frame as in FIG. 1.a.

- **Vehicle-1 Frame (f^{v1}).**

The rotation of the vehicle-1 frame extract the heading angle (ψ) right handed is positive as in FIG. 1.b.

- **Vehicle-2 Frame (f^{v2}).**

The rotation of the vehicle-2 frame extract the pitching angle (θ) right handed is positive as in FIG. 1.c.

- **Body Frame (f^b).**

The rotation of the body frame extract the rolling angle (ϕ) or called as bank angle, right handed is positive as in FIG. 1.d.

The transformation matrix from the vehicle frame to the body frame is DCM which is a function of the Euler angles (ψ, θ, ϕ) is.

$$R_v^b(\phi, \theta, \psi) = R_{v2}^b(\phi) \cdot R_{v1}^v(\theta) \cdot R_{v1}^v(\psi) \quad (1)$$

$$= \begin{pmatrix} c\theta c\psi & c\theta s\psi & -s\theta \\ s\phi s\theta c\psi - c\phi s\psi & s\phi s\theta s\psi + c\phi c\psi & s\phi c\theta \\ c\phi s\theta c\psi + s\phi s\psi & c\phi s\theta s\psi - s\phi c\psi & c\phi c\theta \end{pmatrix}$$

Where $c = \cos$, $s = \sin$

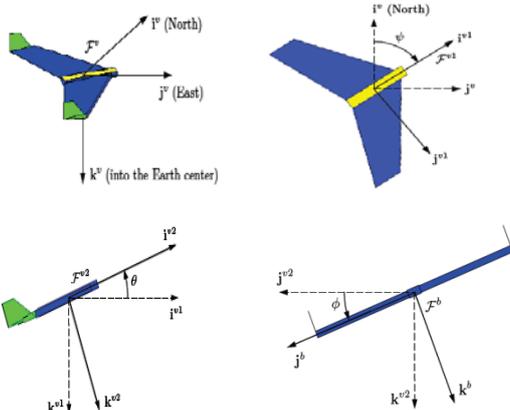


FIG. 1. (a, b, c, and d respectively): The Euler angles and rotational frames illustrations

The rotation sequence $\psi - \theta - \phi$ is commonly used for aircraft and is just one of several Euler angle systems in use [4]. Euler angles representation suffer from a singularity ($\theta = \pm\pi / 2$) also known as the ‘‘gimbal lock’’. In practice, this limitation does not affect the UAV in normal flight mode. [5].

- **Stability Frame (f^s).**

The rotation of stability frame extracts AOA (α) which is the left-handed rotation about the body y^b -axis. The airspeed velocity (V_a) is the velocity of the aircraft relative to the surrounding air.

- **Wind Frame (f^w).**

The angle between the velocity vector and the x^b - z^b plane is called the side-slip angle and is denoted by β . FIG. 2 illustrates the angles extracted from rotation from the body frame to the stability frame to the wind frame (α, β) respectively. The total transformation from the body frame to the wind frame is given by.

$$R_b^w(\alpha, \beta) = R_b^s(\alpha) \cdot R_s^w(\beta) \quad (2)$$

$$= \begin{pmatrix} c\beta c\alpha & -s\beta c\alpha & -s\alpha \\ s\beta & c\beta & 0 \\ c\beta s\alpha & -s\beta s\alpha & c\alpha \end{pmatrix}$$

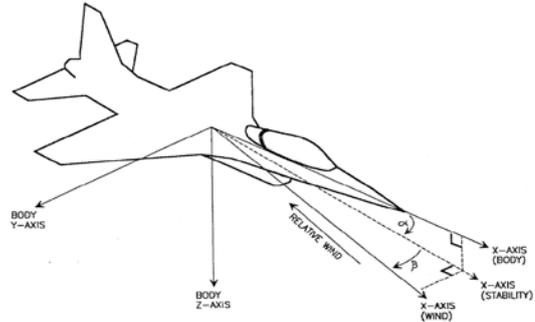


FIG. 2: The rotation angles between the body axis and the wind axis.

II.b. Wind Triangle.

The significant effect of the wind is very important in UAV. The wind triangle briefly illustrates some relations and definitions can be considered in the navigation of the UAV.

The angle between the inertial North (x^i) and the inertial velocity vector projected on the horizontal plane is called the course angle (χ). The crab angle (χ_c) is defined as the difference between the course angle and the heading.

If the wind triangle is projected in the vertical plane another angles can be defined, the flight path angle (γ) is the angle between the horizontal plane and the ground velocity (V_g), so there are two main angles to transform from body frame to flight path frame (χ, γ) [6].

Notes: In the absence of wind,

- The crab angle (χ_c) equal zero.
- The sideslip angle (β) equal zero.
- $V_a = V_g$.

III. Fixed Wing UAV Parameters.

This section presents the basic used parameters of the Ultra Stick-25e (thor). It has a conventional fixed-wing airframe with flap, aileron, rudder, and elevator control surfaces. The maximum deflection of servo actuators equals 25 deg in each direction. The physical details of the aircraft

can be found in table 1, for more details of the University of Minnesota Ultra Stick 25e platform [7, 8].

III.a Geometric Parameters.

The shape of the airfoil determines its aerodynamic properties, and some of its geometrical parameters. Some of aerodynamic parameters are shown in the Figure 2.3.

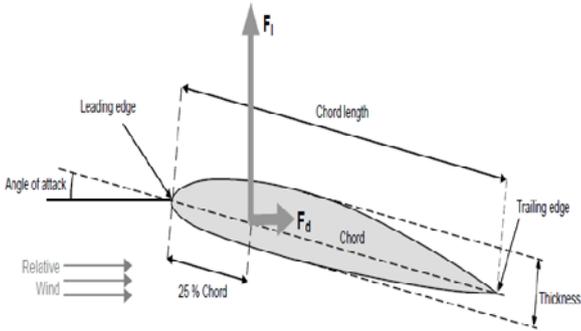


FIG.3: section of airfoil and the applied lift and drag forces

Table 1: Some physical properties of Ultrastick-25e

Property	Symbol	The value
Wing span	b	1.27m
Wing surface area	S	0.3097m ²
Maincord	c	0.25m
Mass	m	1.959kg
Inertia	J _x	0.07151kg.m ²
	J _y	0.08636 kg.m ²
	J _z	0.15364 kg.m ²
	J _{xz}	0.014 kg.m ²

III.b Basic Aerodynamic Parameters.

The dynamics of the UAV is decomposed into longitudinal and lateral dynamics; each of them has some aerodynamic non-dimensional coefficients affect the stability of the aircraft. These coefficients are parameters in the aerodynamic forces and moments equations, and influenced by the airfoil design. A detailed discussion of these coefficients with respect to longitudinal and lateral dynamics existed in [9, 10, and 11].

- **Longitudinal aerodynamic coefficients:** the longitudinal motion acts in the x^b-z^b plane which is called pitch plane and affected by the lift force (f_L), Drag force (f_D), and pitch moment (m). The effectiveness of these forces and moments are measured by lift coefficient (C_L), drag coefficient (C_D), and pitch moment coefficient (C_m). These coefficients influenced by the angle of attack (α), pitch angular rate (q), and elevator deflection (δ_e), but they are nonlinear in the angle of attack; For small α the flow over the wings remain laminar, so no stall conditions will be happened, then we will linearize the equations about this linear zone as in FIG.4.

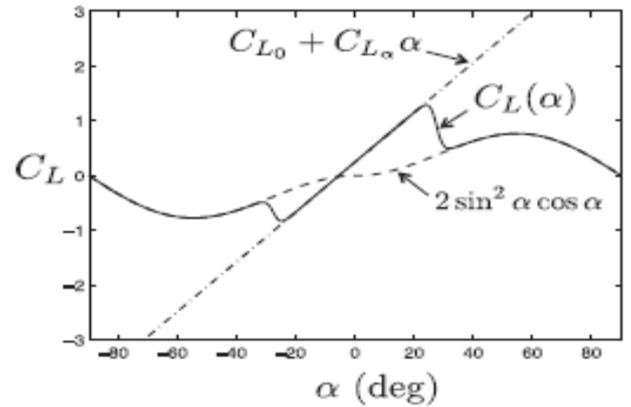


FIG.4: The lift coefficient as a function of α can be approximated by blending a linear function of α (dot-dashed).

- **Lateral aerodynamic coefficients:** the lateral motion which is responsible of the yaw and roll motions. It's affected by the side force (f_y), yaw moment (n), and roll moment (l). The effectiveness of these forces and moments are measured by side force coefficient (C_Y), yaw moment coefficient (C_n), and roll moment coefficient (C_r). These coefficients influenced by sideslip angle (β), yaw angular rate (r), roll angular rate (p), aileron deflection (δ_a), and rudder deflection (δ_r), but they are nonlinear in these parameters.

All of these coefficients should be determined by wind tunnel. Linear approximations for these coefficients and their derivatives are acceptable for modeling purposes and accurate, the linearization is produced by the first-Taylor approximation, and non-dimensionalize of the aerodynamic coefficients of the angular rates [10].

Note:

The effect of Reynolds number and Mach number can be neglected because they are approximately constant in the SUAV dynamics [11].

III.c Fixed Wing UAV Control Surfaces.

As said earlier the designed UAV in this thesis is a standard fixed wing UAV with a standard control surfaces; the elevator, the rudder, the aileron, and the thrust with deflections named δ_e, δ_r, δ_a, and δ_t respectively, the input controls are shown in FIG.5.

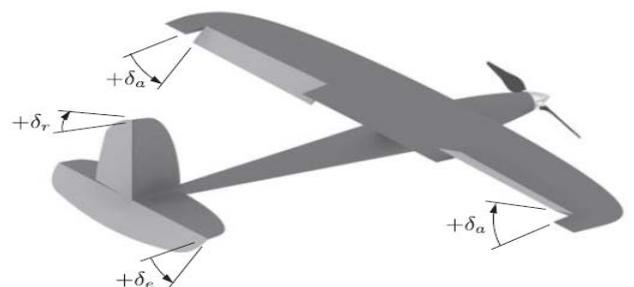


FIG.5: SUAV control surfaces

IV. UAV Flight Dynamics.

The first step of design a controlled UAV is to derive the dynamic model. This section will focus on a standard method for deriving the full nonlinear equations of motion of a fixed wing aircraft.

From beginning, several major assumptions are considered. **First**, the aircraft is rigid. Although aircraft are truly elastic in nature, modeling the flexibility of the UAV will not contribute significantly to the research at hand. **Second**, the earth is an inertial reference frame. **Third**, aircraft mass properties are constant throughout the simulation. **Finally**, the aircraft has a plane of symmetry. The first and third assumptions allow for the treatment of the aircraft as a point mass [12].

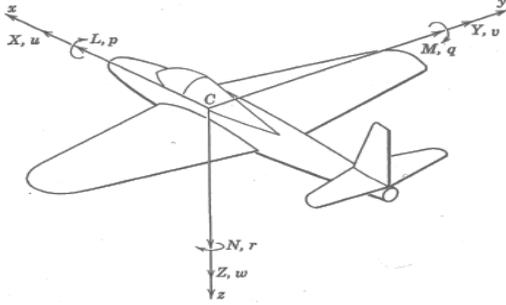


FIG.6: Definitions of UAV body velocities, forces, moments, and angular rates.

IV.a Forces And Moments Applied On The Aircraft.

UAV is subjected to external forces and moments due to gravity, propulsion, and aerodynamics, after applying the newton's second law for translational motion, the applied forces are combined and expressed in the body frame.

$$f_b = (f_x, f_y, f_z)^T. \quad (3)$$

For rotational motion the applied moments are combined and expressed in the body frame. For moments,

$$m_b = (l, m, n)^T. \quad (4)$$

Momentum is defined as the product of inertia matrix j and the angular velocity vector. Due to symmetry of the aircraft about the plane x^b - z^b , the only inertia used in the modeling is j_x , j_y , j_z , and j_{xz} .

$$j_{xy} = j_{yz} = 0.$$

Six degrees of freedom 12-state equations of motion are obtained, but they are not complete, the external forces and moments are not defined yet. The modeling of the forces and moments can be utilized to get finally the nonlinear 12- state equations of motion.

The gravity (f_g), aerodynamic (f_a), and propeller (f_p) forces are the composition of the total forces applied on the body frame (f_x, f_y, f_z). Aerodynamic (m_a), and propeller (m_p) moments is the composition of the total moments applied on the body frame (l, m, n) as shown in FIG. 6, there are no moments produced by the gravity. The above forces will be represented in the stability frame to get F_L , and F_D [13].

IV.b Atmospheric Disturbance.

At the existence of wind the atmospheric disturbances with its two components (steady ambient wind, and wind gusts) can be modeled; the steady ambient wind is

modeled as a constant wind field, the wind gusts is modeled as a turbulence which is generated by passing white noise linear time invariant filter, the Dryden gust model approximations can be considered in the modeling MIL-F-8785C can be used [14].

IV.C Full Nonlinear Equations of Motion.

Finally the following equations of motion are as follows [11].

$$\dot{p}_n = (\cos\theta \cos\psi)u + (\sin\phi \sin\theta \cos\psi - \cos\phi \sin\psi)v + (\cos\phi \sin\theta \cos\psi + \sin\phi \sin\psi)w \quad (5)$$

$$\dot{p}_e = (\cos\theta \sin\psi)u + (\sin\phi \sin\theta \sin\psi + \cos\phi \cos\psi)v + (\cos\phi \sin\theta \sin\psi - \sin\phi \cos\psi)w \quad (6)$$

$$\dot{h} = u \sin\theta - v \sin\phi \cos\theta - w \cos\phi \cos\theta \quad (7)$$

$$\dot{u} = r v - q w - g \sin\theta + \frac{\rho V_a^2 S}{2m} [C_X(\alpha) + C_{X_q}(\alpha) \frac{cq}{2V_a} + C_{X_{\delta e}}(\alpha) \delta_e] + \frac{\rho S_{prop} C_{prop}}{2m} [(k_{moto} \delta_t)^2 - V_a^2] \quad (8)$$

$$\dot{v} = p w - r u + g \cos\theta \sin\phi + \frac{\rho V_a^2 S}{2m} [C_{Y_0} + C_{Y_\beta} \beta + C_{Y_p} \frac{bp}{2V_a} + C_{Y_r} \frac{br}{2V_a} + C_{Y_{\delta a}} \delta_a + C_{Y_{\delta r}} \delta_r] \quad (9)$$

$$\dot{w} = q u - p v + g \cos\theta \cos\phi + \frac{\rho V_a^2 S}{2m} [C_Z(\alpha) + C_{Z_q}(\alpha) \frac{cq}{2V_a} + C_{Z_{\delta e}}(\alpha) \delta_e] \quad (10)$$

$$\dot{\phi} = p + q \sin\phi \tan\theta + r \cos\phi \tan\theta \quad (11)$$

$$\dot{\theta} = q \cos\phi - r \sin\phi \quad (12)$$

$$\dot{\psi} = q \sin\phi \sec\theta + r \cos\phi \sec\theta \quad (13)$$

$$\dot{p} = \Gamma_1 p q - \Gamma_2 q r + \frac{\rho V_a^2 S b}{2} [C_{p_0} + C_{p_\beta} \beta + C_{p_p} \frac{bp}{2V_a} + C_{p_r} \frac{br}{2V_a} + C_{p_{\delta a}} \delta_a + C_{p_{\delta r}} \delta_r] \quad (14)$$

$$\dot{q} = \Gamma_5 p r - \Gamma_6 (p^2 - r^2) + \frac{\rho V_a^2 S c}{2J_y} [C_{m_0} + C_{m_\alpha} + C_{m_q} \frac{cq}{2V_a} + C_{m_{\delta e}} \delta_e] \quad (15)$$

$$\dot{r} = \Gamma_7 p q - \Gamma_1 q r + \frac{\rho V_a^2 S b}{2} [C_{r_0} + C_{r_\beta} \beta + C_{r_p} \frac{bp}{2V_a} + C_{r_r} \frac{br}{2V_a} + C_{r_{\delta a}} \delta_a + C_{r_{\delta r}} \delta_r] \quad (16)$$

Notes:

- The lift and drag terms is nonlinear in (α),
- The propeller thrust is nonlinear in the throttle command.

As we interested in modeling UAV flight under low angle of attack conditions, so simpler linear model can be utilized for

$$C_L(\alpha) = C_{L0} + C_{L\alpha} \alpha$$

$$C_D(\alpha) = C_{D0} + C_{D\alpha} \alpha$$

V. Linearization of Equations of Motion.

Model linearization is based on the small disturbance theory. According to this theory, analysis is done under small perturbations of motion characteristics [15].

Linearization and decoupling of the 12-state equations of motion to produce a reduced linear transfer functions or state space models describing the nonlinear UAV airframe is the most appreciable target to this paper. Low level autopilot control loops for UAV will be designed using This LTI system. The dynamics of aircraft is decoupled into longitudinal (θ, q, h, u, w) and lateral dynamics (ϕ, p, ψ, r, v) . The trimming algorithm at a steady state flight condition will be discussed at the following subsection.

V.a Equilibrium Point and Steady State Flight.

bringing the model under control is done by finding a combination of values of the state and control variables that correspond to a steady-state flight condition then decoupling of the dynamics[2], so the next step is to analyze the dynamics of the aircraft about steady state scenarios or equilibrium points which is actually called trimming technique.

The linearization condition which make $\dot{X} = 0, U = 0$ or constant is supposed. With these conditions the system is called to be at rest (all derivative is equal zero), then examine the behavior of the system near the equilibrium point by slightly perturbing some of the variables. Steady state aircraft flight can be defined as a condition in which all of the motion variables are constant or zero.

- Linear and angular velocities are constants or zero.
- All acceleration components are zero.
- Flat earth.
- Mass of the aircraft is constant.
- Neglecting of the change of atmosphere density due to altitude.

These definitions are available for some aircraft basic scenarios [2].

- Steady wings level flight.
- Steady turning flight.
- Steady wings level climb.
- Climbing turn.

For steady state flight:

$$\dot{p}, \dot{q}, \dot{r}, \dot{u}, \dot{v}, \dot{w} \text{ (or } \dot{V}_a, \dot{\beta}, \dot{\alpha}) = 0$$

$$U = \text{constant}$$

With the following additional constraints according to the flight condition:

- 1- Steady wings level flight:
 $\phi, \dot{\phi}, \dot{\theta}, \dot{\psi} = 0$
- 2- Steady turning flight:
 $\dot{\phi}, \dot{\theta} = 0$, and $\dot{\psi}$ turn rate
- 3- Steady pull-up flight :
 $\phi, \dot{\phi}, \dot{\psi} = 0$, and $\dot{\theta}$ pull – up rate
- 4- Steady climbing turn:
 $\dot{\phi} = 0, \dot{\theta}$ pull – up rate, and $\dot{\psi}$ turn rate

For a fixed wing UAV.

The states are:

$$X = (P_n, P_e, h, u, v, w, \phi, \theta, \psi, p, q, r)^T.$$

The inputs:

$$U = (\delta_e, \delta_t, \delta_a, \delta_r)^T.$$

Our knowledge of the aircraft behavior allows us to specify the required steady state condition so that the trim algorithm converges on an appropriate solution. The generic trim program links to any nonlinear model produces a file containing the steady state values for the states and control inputs for use in the linearization programs.

The aircraft designers must know how to specify the steady state condition; how many of the state and control variables may be chosen independently, and what constraints exist on the remaining variables.

In the process of performing trim calculations for the SUAV the wind will be treated as an unknown disturbance, so the wind speed is zero. The trim calculation algorithm output the trim states and inputs according to the steady state condition. These assumed steady state flights for the Ultrastick-25e as follows.

- I. Steady Straight and Level flight.
- II. Level Climb.
- III. Level Turn.
- IV. Climbing Turn.

So the trimmed outputs and controls are summarized in table 2 which represents a set of trimmed conditions for the Ultrastick-25e model. These values will be used in the next section to evaluate the linear lateral and longitudinal state space models, and then used in the autopilot design according to the desired flight conditions.

Table 2: Trimmed values for Ultrastick-25e (Thor)

	I	II	III	IV
δ_t	0.569	0.721	0.582	0.731
δ_e	-0.0963	-0.102	-0.125	-0.131
δ_r	0.00317	0.00436	-0.00748	-0.00607
δ_a	0.01	0.0138	0.0186	0.0253
V_a	17	17	17	17
β	$3.72 \cdot 10^{-22}$	$3.56 \cdot 10^{-25}$	$-1.51 \cdot 10^{-20}$	$5.8 \cdot 10^{-20}$
α	0.054	0.0529	0.0646	0.0633
h	100	Don't care		
ϕ	-0.00172	-0.00239	0.544	0.547
θ	0.054	0.14	0.0553	0.141
ψ	2.71	2.71	2.71	2.71
p	$5.09 \cdot 10^{-27}$	$5.21 \cdot 10^{-28}$	-0.0193	-0.0492
q	$-7.56 \cdot 10^{-23}$	$1.12 \cdot 10^{-26}$	0.181	0.18
r	$-1.03 \cdot 10^{-24}$	$-8.32 \cdot 10^{-28}$	0.298	0.295
γ	$-9.8 \cdot 10^{-17}$	0.0873	$-1.23 \cdot 10^{-09}$	0.0873

V.b Linear State Space Model.

After obtaining nonlinear 12-state equations of motion and obtaining the trimmed values of different flight conditions, a linearization technique to linearize the equations will be derived to obtain the state space models for longitudinal and lateral dynamics at last.

Calculating the jacobian matrices for LTI equations directly from the nonlinear model are done by assigning the state and control variables from the steady state conditions, and numerically evaluating the partial derivatives in the jacobian matrices. The jacobian matrices may therefore be determined for any steady state flight condition [11]. The linearization program is done to determine A, B, C, D of the state space model.

At last of this subsection the full description of a longitudinal linear state space model with its reduced order modes including the short-period mode, the phugoid mode. Then the description lateral model with its roll mode, the dutch-roll mode, and the spiral-divergence mode.

V.b.1 Longitudinal State Space Model.

The longitudinal state equations are given by:

$$\dot{x}_{lon} \triangleq (u, w, q, \theta, h)^T.$$

And the input (control) vector is defined as:

$$U_{lon} \triangleq (\delta_e, \delta_t)^T.$$

Expressing equations (8), (10), (15), (12), and (7) in terms of x_{lon} and U_{lon} . Assuming that the lateral states are zero (i.e., $\varphi = p = r = \beta = v = 0$) and the wind speed is zero.

V.b.2 Valued Longitudinal Model for Straight and Level Flight.

State space longitudinal model has 5 States ($u, w, \theta, q, \text{ and } -h$), 2 Inputs ($\delta_e, \text{ and } \delta_t$), and 5 Outputs ($V_\alpha, \alpha, \theta, q, \text{ and } h$). The longitudinal linear state space model is SYS_{lon} which has ($A_{lon}, B_{lon}, C_{lon}, D_{lon}$).

$$\begin{aligned}
 - A_{lon} &= \begin{pmatrix} -0.5944 & 0.8008 & -9.791 & -0.8747 & 5.077 * 10^{-5} \\ -0.744 & -7.56 & -0.5294 & 15.72 & -0.000939 \\ 0 & 0 & 0 & 1 & 0 \\ 1.041 & -7.406 & 0 & -15.81 & -7.284 * 10^{-18} \\ -0.05399 & 0.9985 & -17 & 0 & 0 \end{pmatrix} \\
 - B_{lon} &= \begin{pmatrix} 0.4669 & 0 \\ -2.703 & 0 \\ 0 & 0 \\ -133.7 & 0 \\ 0 & 0 \end{pmatrix} \\
 - C_{lon} &= \begin{pmatrix} 0.9985 & 0.05399 & 0 & 0 & 0 \\ -0.003176 & 0.05874 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix} \\
 - D_{lon} &= 0
 \end{aligned}$$

The eigenvalues can be determined by finding the eigenvalues of the matrix A_{lon}

$$|\lambda I - A| = 0$$

Longitudinal Poles:

Eigenvalue	Damping	Frequency
$-0.159 \pm 0.641i$	0.241	0.66 (phugoid)
$-11.7 \pm 10.0i$	0.759	15.4 (short period)

The reduced order modes of the longitudinal dynamics are the approximation of the full linear longitudinal model. At phugoid mode or long period mode is a lightly damped and slow response to the

inputs but the short period mode is fast and acceptable [16].

- The linearized outputs (v, α, q, θ, h) response due to δ_e input are represented in polynomial form eqns. 17-(a, b, c, d, e).

$$\begin{aligned}
 \frac{\bar{v}_\alpha(S)}{\bar{\delta}_e(S)} &= \frac{0.3203 S^5 + 11.53 S^4 + 555 S^3 + 12630 S^2 + 57190 S - 15.45}{S^6 + 29.86 S^5 + 391 S^4 + 1598 S^3 + 632.6 S^2 + 638.1 S + 0.07671} \\
 \frac{\bar{\alpha}(S)}{\bar{\delta}_e(S)} &= \frac{-0.1603 S^5 - 127.4 S^4 - 823.8 S^3 - 330.8 S^2 - 521.1 S - 0.07907}{S^6 + 29.86 S^5 + 391 S^4 + 1598 S^3 + 632.6 S^2 + 638.1 S + 0.07671} \\
 \frac{\bar{q}(S)}{\bar{\delta}_e(S)} &= \frac{-133.7 S^5 - 1859 S^4 - 6751 S^3 - 2245 S^2 - 0.07907 S - 2.806 * 10^{-16}}{S^6 + 29.86 S^5 + 391 S^4 + 1598 S^3 + 632.6 S^2 + 638.1 S + 0.07671} \\
 \frac{\bar{\theta}(S)}{\bar{\delta}_e(S)} &= \frac{-133.7 S^4 - 1859 S^3 - 6751 S^2 - 2245 S - 0.07907}{S^6 + 29.86 S^5 + 391 S^4 + 1598 S^3 + 632.6 S^2 + 638.1 S + 0.07671} \\
 \frac{\bar{h}(S)}{\bar{\delta}_e(S)} &= \frac{2.725 S^4 - 106.8 S^3 - 17590 S^2 - 109100 S - 29300}{S^6 + 29.86 S^5 + 391 S^4 + 1598 S^3 + 632.6 S^2 + 638.1 S + 0.07671}
 \end{aligned}$$

V.b.3 Lateral State Space Model.

Lateral directional equations of motion consist of the side force, rolling moment and yawing moment equations of motion. For the lateral state-space equations, the state is given by

$$\dot{x}_{lat} \triangleq (v, p, r, \phi, \psi)^T,$$

And the input (control) vector is defined as:

$$U_{lat} \triangleq (\delta_a, \delta_r)^T$$

Expressing equations (9), (14), (16), (11), and (13) in terms of x_{lat} and U_{lat} , we get The Jacobians of equations.

V.b.4 Valued Lateral Model for Straight and Level Flight.

The lateral-directional model has five states (v, p, r, ϕ, ψ), two inputs (δ_a, δ_r), and five outputs (β, p, r, ϕ, ψ). The lateral state space model is SYS_{lat} with ($A_{lat}, B_{lat}, C_{lat}, D_{lat}$).

$$\begin{aligned}
 - A_{lat} &= \begin{pmatrix} -0.8726 & 0.8789 & -16.82 & 9.791 & 0 \\ -2.823 & -16.09 & 3.367 & 0 & 0 \\ 0.702 & 0.514 & -2.775 & 0 & 0 \\ 0 & 1 & 0.05406 & -4.088 * 10^{-24} & 0 \\ 0 & 0 & 1.001 & -7.573 * 10^{-23} & 0 \end{pmatrix} \\
 - B_{lat} &= \begin{pmatrix} 0 & 5.302 \\ -156.5 & -5.008 \\ 11.5 & -82.04 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \\
 - C_{lat} &= \begin{pmatrix} 0.05882 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}
 \end{aligned}$$

- $D_{lat} = 0$

The null column in the A_{lat} matrix shows that the state ψ is not coupled back to any other states, and it can be omitted from the state equations when designing the stability augmentation system.

The eigenvalues can be determined by finding the eigenvalues of the matrix A.

$$|\lambda I - A| = 0$$

Lateral-Directional Poles [16]:

Eigenvalue	Damping	Frequency
0	-1	0
-.0138	1.00	0.0138 (spiral)
-1.84 ± 5.28i	0.329	5.59(dutch roll)
-16.1	1.00	16.1 (roll)

In general we found that the roots of the lateral-directional characteristic equation composed of two real roots and a pair of complex roots. These roots will characterize the airplane response [16].

The dutch roll poles are not canceled out of the transfer function p/δ_a complex zeros, thus meaning that Coupling exists between the rolling and yawing motions, the dutch roll mode will involve some rolling motion. These transfer functions validate the decision to use the MIMO state equations for the analysis, so the linearized outputs (β, p, r, ϕ) response due to δ_a input are the eqns. 18-(a, b, c, d)).

$$\frac{\bar{\beta}(S)}{\bar{\delta}_a(S)} = \frac{-19.47 S^2 - 213.8 S - 224.5}{S^4 + 19.74 S^3 + 90.49 S^2 + 502.2 S + 6.89}$$

$$\frac{\bar{p}(S)}{\bar{\delta}_a(S)} = \frac{-156.5 S^3 - 532 S^2 - 4277 S + 123.7}{S^4 + 19.74 S^3 + 90.49 S^2 + 502.2 S + 6.89}$$

$$\frac{\bar{r}(S)}{\bar{\delta}_a(S)} = \frac{11.5 S^3 + 114.8 S^2 - 114.1 S - 2289}{S^4 + 19.74 S^3 + 90.49 S^2 + 502.2 S + 6.89}$$

$$\frac{\bar{\phi}(S)}{\bar{\delta}_a(S)} = \frac{-155.8 S^2 - 525.8 S - 4283}{S^4 + 19.74 S^3 + 90.49 S^2 + 502.2 S + 6.89}$$

And the linearized outputs (β, p, r, ϕ) response due to δ_r input are the eqns. 19-(a, b, c, d)).

$$\frac{\bar{\beta}(S)}{\bar{\delta}_r(S)} = \frac{0.3119 S^3 + 86.8 S^2 + 1302 S - 208.3}{S^4 + 19.74 S^3 + 90.49 S^2 + 502.2 S + 6.89}$$

$$\frac{\bar{p}(S)}{\bar{\delta}_r(S)} = \frac{-5.008 S^3 - 309.5 S^2 - 4304 S + 127.1}{S^4 + 19.74 S^3 + 90.49 S^2 + 502.2 S + 6.89}$$

$$\frac{\bar{r}(S)}{\bar{\delta}_r(S)} = \frac{-82.04 S^3 - 1385 S^2 - 1228 S - 2351}{S^4 + 19.74 S^3 + 90.49 S^2 + 502.2 S + 6.89}$$

$$\frac{\bar{\phi}(S)}{\bar{\delta}_r(S)} = \frac{-9.443 S^2 - 384.4 S - 4370}{S^4 + 19.74 S^3 + 90.49 S^2 + 502.2 S + 6.89}$$

V.c Analytical Linearization of aircraft equations of motion.

In this section the analytical linearization of roll and roll dynamics can be derived to check the matching between state space linearized model and the analytical model.

First: roll or bank angle (ϕ)

The eqn. (11) can be considered to be linearized from this main assumption which is logic for most flight

conditions; the pitch angle (θ) is a small this means that the primarily influence on $\dot{\phi}$ equation is roll rate (p), so

$$\dot{\phi} = p + d_{\phi 1}$$

Second: differentiate the above equation we get:

$$\ddot{\phi} = \dot{p} + \dot{d}_{\phi 1}$$

Third: substitute \dot{p} by eqn. (14) and the equation in the first step we will get the eqn. as follows

$$\ddot{\phi} = -a_{\phi 1} \dot{\phi} + a_{\phi 2} \delta_a + \dot{d}_{\phi 2}$$

Where:

- $a_{\phi 1}, a_{\phi 2}$ are the coefficients of the roll dynamics, they are variables in the aircraft parameters and the trimmed values.
- $d_{\phi 2}$ can be considered as a disturbance on the system.

Fourth: Laplace transfer function is as follows:

$$\phi(s) = \frac{a_{\phi 2}}{s(s + a_{\phi 1})} \left(\delta_a(s) + \frac{1}{a_{\phi 2}} d_{\phi 2}(s) \right)$$

Fifth: the final numerical transfer function of roll (ϕ) for δ_a as input is as follows:

$$\phi(s) \approx \frac{-163.6}{s(s + 16.82)} \delta_a(s)$$

Sixth: roll rate (p) can be approximately considered as the differentiation of the roll angle so

$$p(s) \approx \frac{-163.6}{(s + 16.82)} \delta_a(s)$$

Comparison between the analytical and state space linearization by jacobian matrices is as follows:

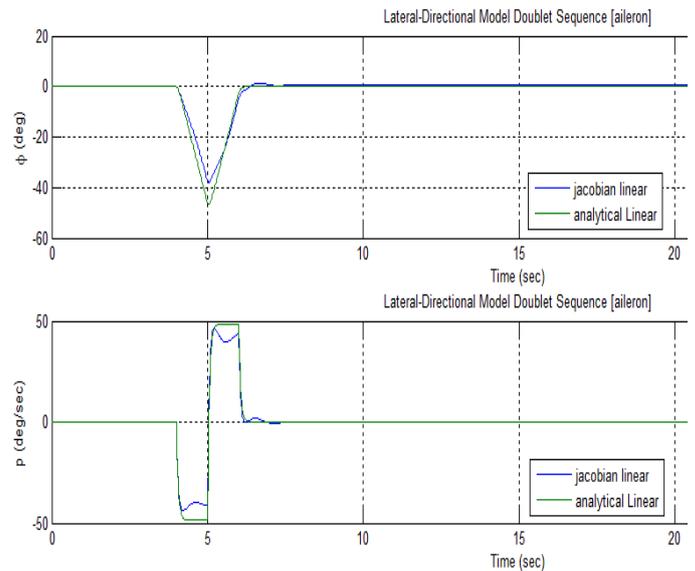


FIG.7: linearized roll and roll rate comparison techniques with applying the doublet signal at the control surface (δ_a).

FIG. 7 illustrates the good matching between the two linearization techniques. The next section illustrates how to validate the linear model.

VI. Validation of Aircraft Model Linearization.

After getting the model some checks of the Ultrastick-25e (thor) longitudinal dynamics responses to (elevator) and lateral dynamics responses to (aileron, rudder) deflections of linear and nonlinear models will be illustrated in the following figures by applying a doublet pulse (a pulse that is symmetric about its reference level (the trim setting) to the control inputs) to see the response of the various outputs.

VI.a Doublet Response of the Linear and Nonlinear Longitudinal Model.

Doublet response of the longitudinal dynamics ($v_a, \alpha, q, \theta, h, a_x, a_z$) of the linear model and nonlinear model is shown in the following figures.

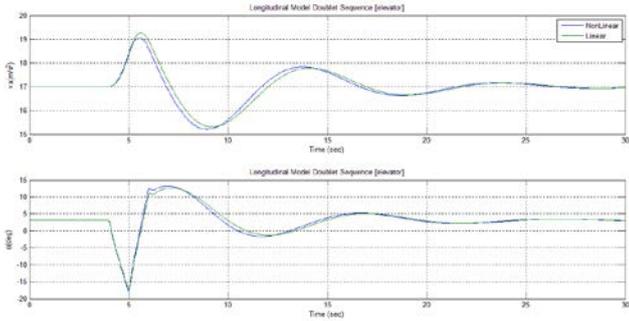


FIG.8: Response of (V_a, θ) of Ultrastick-25e model due to elevator doublet (trim±5 degree).

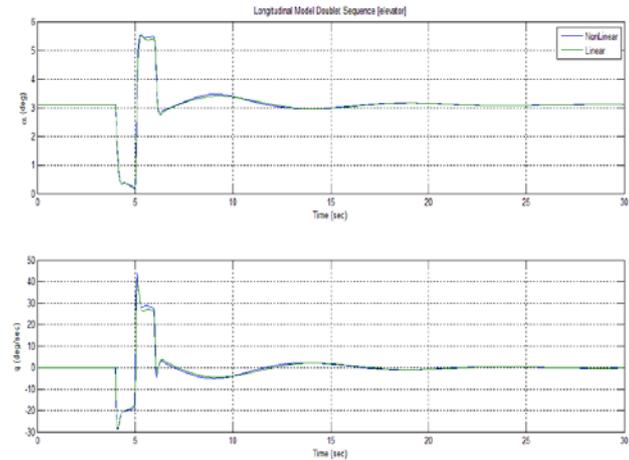


FIG.9: Response of (α, q) of Ultrastick-25e model due to elevator doublet (trim±5 degree).

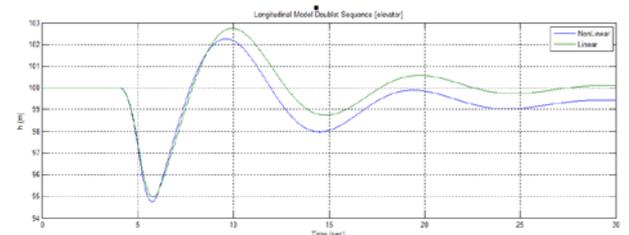


FIG.10: Response of (h) of Ultrastick-25e model due to elevator doublet (trim±5 degree).

VI.b Doublet Response of the Linear and Nonlinear Lateral Model.

Doublet response of the lateral dynamics outputs (β, p, r, ϕ, ψ) response due to doublet δ_a, δ_r of the linear

model and nonlinear model (simulink) are shown in the following figures.

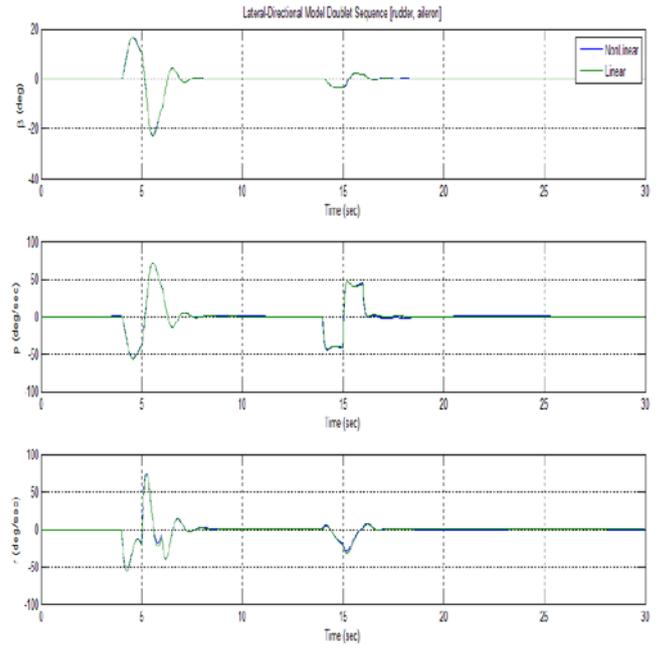


FIG.11: Response of the lateral dynamics (β, p, r) due to 5 degree (aileron, rudder) deflection doublet signal.

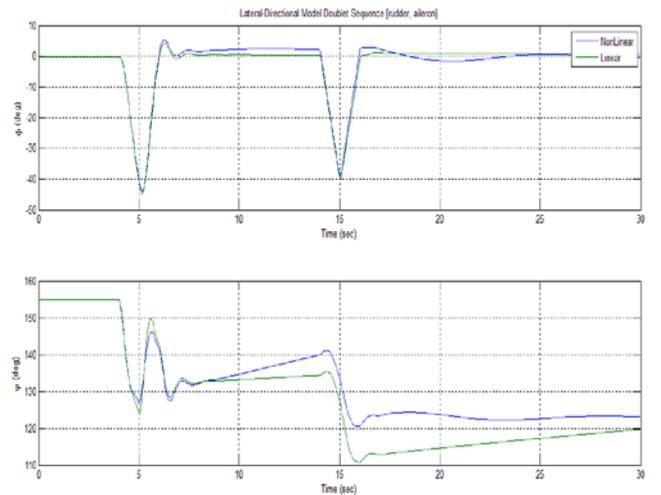


FIG.12: Response of the lateral dynamics (ϕ, ψ) due to 5 degree (aileron, rudder) deflection doublet signal.

VII. Conclusion.

As seen the objective of this thesis is to develop a mathematical model for small unmanned aerial vehicle that can be used for designing an autopilot for it to control the various phases of flights. The nonlinear equations of motion extraction were focused; with these equations the linear longitudinal and lateral models are obtained. With the 6th section the methodology of getting the linear model is acceptable and can approximately describe the behavior of the nonlinear dynamics of UAV, so the resulted one will be used to design the autopilot of the UAV with inner and outer loop.

VIII. References.

[1] Edward AFB CA, “Flying Qualities Phase.Vol2 ”, USAF TEST Pilot School, 1988
 [2] B.L.Stevens and F.L.Lewis, “Aircraft Control and Simulation”, Hoboken, NJ: JohnWiley& Sons, Inc., 2nd ed., 2003.

- [3] R. C. Nelson, "Flight Stability and Automatic Control", Boston, MA:McGraw-Hill, 2nd ed., 1998.
- [4] David A. Coughy, "Introduction To Aircraft Stability And Control", Sibley School Of Mechanical & Aerospace Engineering, Cornell University, Ithaca, New York, 2011
- [5] A. Noth, S. Bouabdallah and R. Siegwart," Dynamic Modeling of Fixed-Wing UAVs", Swiss Federal institute of technology, version 2, 2006.
- [6] D. T. Greenwood, "*Principles of Dynamics*. Englewood Cliffs", NJ: Prentice Hall, 2nd ed., 1988
- [7] Murch, A., Dorobantu, A., and Balas, G., "University of Minnesota UAV Flight Control Research Group," <http://www.uav.aem.umn.edu>, 4 March 2013].
- [8] Murch, A., Paw, Y. C., Pandita, R., Li, Z., and Balas, G., "A Low Cost Small UAV Flight Research Facility," CEAS Conference on Guidance, Navigation, and Control, Munich, Germany, 2011.
- [9] "USAF Stability and Control DATCOM", Flight Control Division, Air Force Flight Dynamics Laboratory, Wright- Patterson Air Force Base, Oh, 1980.
- [10] Xinzhong Chen, Ahsan Kareem," Advances in Modeling of Aerodynamic Forces on Bridge Decks", Journal of Engineering Mechanics, November 2002.
- [11] Randal W.Beard, Timothy W.Mclain, "Small Unmanned Aircraft: Theory and Practice", Princeton university press, 2012.
- [12] Nidal M. Jodeh, "Development Of Autonomous Unmanned Aerial Vehicle Research Platform: Modeling, Simulating, And Flight Testing", Department Of The Air Force Air University, Air Force Institute Of Technology, Wright-Patterson Air Force Base, Ohio, 2006
- [13] Michael V. Cook, "flight dynamics principles: A linear Systems Approach to Aircraft Stability and Control 3rd Edition", Elsevier Ltd, 2013.
- [14] T. R. BEAL. "Digital simulation of atmospheric turbulence for Dryden and von Karman models", Journal of Guidance, Control, and Dynamics, Vol. 16, No. 1 (1993), pp. 132-138.
- [15] Kimon P. Valavanis ,"Advances in Unmanned Aerial Vehicles", University of South Florida, Tampa, Florida, USA,2007.
- [16] David G. Hull,"Fundamentals of Airplane Flight Mechanics", Austin, Texas, 2005.

V_g	Ground speed vector
V_w	Wind speed vector
C_L	Aerodynamic lift coefficient
C_D	Aerodynamic drag coefficient.
C_{m^*}	Aerodynamic pitching moment coefficient
C_{p_x}	Aerodynamic moment coefficient along the x^b -axis
C_{q_x}	Aerodynamic moment coefficient along y^b -axis.
C_{prop}	Aerodynamic coefficient for the propeller.
C_{q_z}	Aerodynamic moment coefficient along the z^b .
C_{X^*}	Aerodynamic force coefficient along x^b
C_{Y^*}	Aerodynamic force coefficient along y^b
C_{Z^*}	Aerodynamic force coefficient along z^b .
δ_a	aileron deflection
δ_e	elevator deflection
δ_r	rudder deflection
δ_t	throttle deflection
f_D	Force due to aerodynamic drag
f_L	Force due to aerodynamic lift
m_b	External moment applied to the airframe
$l, m, \text{ and } n$	the components of m_b in m^b
g	Gravitational acceleration (9.81 m/s ²)
Γ_*	Products of the inertia matrix
h	Altitude
ρ	Density of air.
J	The inertia matrix
$J_x, J_y, J_z, \text{ and } J_{xz}$	Elements of the inertia matrix
k_{motor}	Constant that specifies the efficiency of the motor
S_{prop}	Area of the propeller

Nomenclature.

(x^i, y^i, z^i)	inertial frame axes
(x^v, y^v, z^v)	vehicle frame axes
(x^b, y^b, z^b)	body frame axes
ϕ, θ, ψ	attitude angles, rad
α	Angle of attack.
β	Side slip angle.
χ	Course angle
χ_c	Crab angle
γ	Inertial-referenced flight path angle
u, v, w	inertial velocity components of the airframe projected onto x^b -axis
V_a	Airspeed vector

Gradient-Statistical Algorithm for Calculating Critical Points of Density Probability of Gaussian Mixture

N. N. Aprausheva, V. V. Dikusar, and S. V. Sorokin

Abstract— An algorithm for calculation all critical points of the given probability density of the Gaussian mixture $f(X)$ consists of two parts: 1) modeling a representational sample, 2) discrimination a neighborhood of each critical point of the function $f(X)$, determination of its value and form (maximum, minimum, saddle point). For realization of the second part of the algorithm Gradient, Hessian values of the function $f(X)$ in sample points and cluster-analysis methods are used. Critical points calculation by this algorithm does not need in task of initial conditions.

Keywords— cluster-analysis, critical point, Hessian, sample.

I. INTRODUCTION

FINITE Gaussian mixtures have found wide applications in various fields of science and practice, such as mathematical modeling, pattern recognition, spectroscopy, biology, medicine, chemistry, geology, meteorology, sea fishery, etc. [1-4]. The popularity of finite Gaussian mixtures is stipulated by their identifiability, smoothness, completeness and resolutions [2,5] and requires of solving mathematical problems, such as mode-finding and preliminary estimation of the mode number.

For determination of Gaussian mixture modes an exact analytical expressions do not exist. Only in the simplest case of unimodality of two-component mixture at $\sigma_1 = \sigma_2$, $\pi_1 = \pi_2$, $\rho \leq 2$ the single mode \hat{x} is determined by the formula:

$$\hat{x} = (\mu_1 + \mu_2)2^{-1},$$

σ_1, σ_2 are the variances of the mixture components, π_1, π_2 are their weights, ρ is Mahalanobis distance.

For well-known values of the parameter distributions for calculation of mode and other critical points of probability density various iterative methods are used: Newton, gradient-quadratic, Picard and others [6, 7, 8, 4, 9]. In a

multidimensional space the convergence of the sequence of the iterations, which is generated by this methods, depends on initial conditions, which can lead to a gap of some critical points. For calculation of all critical points of this function the gradient-statistical algorithm (GSA) was developed, which unites statistical elements, gradient method and cluster-analysis. This algorithm does not demand the task of initial conditions, this is its advantage towards other well-known methods.

The probability density of Gaussian mixture in p -dimensional space has the form:

$$f(X) = (2\pi)^{-\frac{p}{2}} \sum_{i=1}^k \pi_i |\Sigma_i|^{-\frac{1}{2}} e^{-\frac{1}{2}(X-\mu_i)\Sigma_i^{-1}(X-\mu_i)'}. \quad (1)$$

$1 \leq p < \infty$, $2 \leq k < \infty$, $X, \mu_i \in R^p$, $\Sigma_i > 0$, $0 \leq \pi_i < 1$, $\sum_{i=1}^k \pi_i = 1$, μ_i is the expectation of i -th component π_i is its weight, Σ_i is its covariance matrix.

Gradient and Hessian of function $f(X)$ are expressed by the formulae [4, 9]:

$$g = \nabla f(X) = \sum_{i=1}^k f_i(X) \Sigma_i^{-1}(\mu_i - X)', \quad (2)$$

$$H = \nabla' \nabla f(X) = \sum_{i=1}^k f_i(X) \Sigma_i^{-1}((\mu_i - X)'(\mu_i - X) - \Sigma_i) \Sigma_i^{-1}. \quad (3)$$

$$f_i(X) = \frac{|\Sigma_i|^{-\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}} \pi_i e^{-\frac{1}{2}(X-\mu_i)\Sigma_i^{-1}(X-\mu_i)'}. \quad (4)$$

II. ALGORITHM DESCRIPTION

From given probability density (1) the p -dimensional representational sample of n independent random vectors is generated,

N. N. Aprausheva, is with Dorodnicyn Computing Centre of RAS, Moscow, Russia. (e-mail: plat@ccas.ru).

V. V. Dikusar, is with Dorodnicyn Computing Centre of RAS, Moscow, Russia. (e-mail: dikussar@yandex.ru).

S. V. Sorokin, is with Dorodnicyn Computing Centre of RAS, Moscow, Russia. (e-mail: author@nrim.go.jp).

$$A = \{X_1, X_2, \dots, X_n\}, \quad 1 \leq p < \infty, \quad (5a)$$

$$X_s = (x_{s1}, x_{s2}, \dots, x_{sp}), \quad s = 1, 2, \dots, n. \quad (5b)$$

Using the formula (2) in each point $X_s \in A$ we calculate the gradient modulus,

$$|\nabla_s| = \left(\sum_{i=1}^p \left(\frac{\partial f(X_s)}{\partial x_i} \right)^2 \right)^{\frac{1}{2}}, \quad s = 1, 2, \dots, n. \quad (6)$$

From the set A we choose subset B , its elements suffice to the conditions $|\nabla f(X_s)| < \varepsilon$, $s \in \{1, 2, \dots, n\}$, ε is a given small positive number,

$$B = \{X_s \in A: |\nabla f(X_s)| < \varepsilon\}, \quad 0 < \varepsilon < 1, \quad s \in \{1, 2, \dots, n\}. \quad (7)$$

From the set B by Sylvester criterion the subset N is selected, which consists of the neighborhoods N_t , $t = 1, 2, \dots, m$, of all modes of the function $f(X)$,

$$N = \{X_s \in B: |H(f(X_s))| < 0\}, \quad s \in \{1, 2, \dots, n\}, \quad (8)$$

$$N = \bigcup_{t=1}^m N_t, \quad (9)$$

$|H(f(X_s))|$ is Hessian determinant of the function $f(X)$ in the point X_s . From the set N the neighborhood of each mode N_t is selected, $t = 1, 2, \dots, m$, by cluster-analysis methods [10, 11, 12] (algorithm FOREL, algorithm k-means and others). The estimation of the mode \hat{X}_t we obtain from the equality:

$$\hat{X}_{0t} = \arg \max_{X_s \in N_t} f(X_s), \quad t = 1, 2, \dots, m, \quad (10)$$

For more exact definition of the mode value \hat{X}_{0t} we may model the sample, which is uniformly distributed in sphere $S_t(\hat{X}_{0t}, r)$ with the center \hat{X}_{0t} and the radius r , we may set

$$\hat{X}_{1t} = \arg \max_{X_s \in S_t} f(X_s), \quad t = 1, 2, \dots, m, \quad (11)$$

Then the mode estimation \hat{X}_t is defined by the equalities:

$$\hat{X}_t = \begin{cases} \hat{X}_{1t}, & \text{if } \hat{X}_{1t} > \hat{X}_{0t}, \\ \hat{X}_{0t}, & \text{if } \hat{X}_{0t} \geq \hat{X}_{1t}. \end{cases} \quad (12)$$

Similarly the neighborhoods of all the local minimums L_t of the function $f(X)$ are discovered. From the set $B \setminus N$ by Sylvester criteria the subset L is selected,

$$L = \{X_s \in B \setminus N: |H(f(X_s))| > 0\}, \quad s \in \{1, 2, \dots, n\}, \quad (13)$$

$$L = \bigcup_{t=1}^l L_t, \quad (14)$$

$$L = \emptyset \quad \text{or} \quad 1 \leq l < m.$$

If in the expression (14) $\neq \emptyset$, then from the set L the neighborhood L_t of each point of local minimum (LM) of the function $f(X)$ is selected by cluster-analysis methods (algorithm KRAB, method k-means and others) [10, 11, 12]. The estimation of the LM point we get by the formula:

$$\tilde{X}_{0t} = \arg \min_{X_s \in L_t} f(X_s), \quad t = 1, 2, \dots, l. \quad (15)$$

Each estimation \tilde{X}_{0t} we can make more exact by additional modeling of uniformly distributed points in the sphere $S_t(\tilde{X}_{0t}, r)$. Then

$$\tilde{X}_{1t} = \arg \min_{X_s \in S_t} f(X_s), \quad t = 1, 2, \dots, l. \quad (16a)$$

end

$$\tilde{X}_t = \begin{cases} \tilde{X}_{1t}, & \text{if } \tilde{X}_{1t} < \tilde{X}_{0t}, \\ \tilde{X}_{0t}, & \text{if } \tilde{X}_{0t} < \tilde{X}_{1t}. \end{cases} \quad (16b)$$

The subset C ,

$$C = B \setminus (N \cup L),$$

does not contain the neighborhoods of the extreme points. It consists of the subset of the neighborhoods of the saddle points S and the subset G of the elements, which obey the following conditions

$$G = \{X_s \in C: f(X_s) < \varepsilon_1 \quad \text{u} \quad ||H(f(X_s))|| < \varepsilon_2\}, \quad (17)$$

$0 < \varepsilon_1 < 1$, $0 < \varepsilon_2 < 1$, ε_1 , ε_2 are small positive numbers, $||H(f(X_s))||$ is modulus of the Hessian determinant in the point $X_s \in C$, $C = S \cup G$. G is a set of points X_s , in which the surface $f(X)$ is nearly flat.

The set U ,

$$U = C \setminus G, \quad U = B \setminus (N \cup L \cup G), \quad (18)$$

is a union of the neighborhoods U_i of all the saddle points (SP) of function $f(X)$,

$$U = \bigcup_{i=1}^c U_i \tag{19}$$

$$U = \bigcup_{i=1}^c U_i \tag{19}$$

$$U = \emptyset \text{ or } U \neq \emptyset. \tag{20}$$

For the determination of the number c in (19) we use the Euler theorem [13]:

$$c = m + l - 1, \tag{21}$$

m is an amount of the modes of the function $f(X)$, l is an amount of the points of their local minima, c is an amount of their saddle points.

The neighbourhood $U_i, i = 1, 2, \dots, c$, of each saddle point is detected by cluster-analysis methods (algorithm k -means, algorithm FOREL and others) with the preliminary research of the structure of the set U , which defined in [12]. The estimation of each saddle point $X_i, i = 1, 2, \dots, v$, is calculated by the formula:

$$X_i = \arg \min_{X_s \in U_i} |\nabla_s|, \quad i = 1, 2, \dots, c, \tag{22}$$

$|\nabla_s|$ is a modulus of the gradient of the function $f(X)$ in the point X_s .

Each saddle point $X_i \in S_i$ is characterized by the following: in its neighborhood there are points $X_s, s \in \{1, 2, \dots, n_i\}$, and points $X_t, t \in \{1, 2, \dots, n_i\}$, for which the inequalities have place:

$$f(X_i) < f(X_s) \text{ and } f(X_i) > f(X_t). \tag{23}$$

$$X_i, X_s, X_t \in U_i \quad s \neq t, \quad s, t \in \{1, 2, \dots, n_i\}.$$

It is possible to make more exact values of the estimations of the critical points of the function $f(X)$ by the well-known algorithms (Newton, Picard, gradient-quadratic) with the starting conditions given by the formulae (11), (16a), (22). For determination of the threshold values in the cluster-analysis algorithms the results of preliminary analysis of the structures of the sets N, L, U are used, which is expanded in [12]

III. THE RESULTS OF THE EXPERIMENTS

The experiment testing of GSA, which has been made on one-dimensional and two-dimensional Gaussian mixtures, gives the positive result.

In table 1 the values of parameters of ten one-dimensional three-component mixtures ($\sigma = 1$) are represented. In table 2

the values of critical points of these mixtures which were got by GSA, are given.

Table 1

Mixture #	Distribution parameters				
	μ_1	μ_2	μ_3	π_1	π_2
1	0	1.41	2	0.33	0.33
2	0	1.8	2	0.5	0.25
3	0	1.5	2	0.45	0.1
4	0	1.41	2.82	0.6	0.2
5	0	1.41	2.82	0.45	0.1
6	0	1.5	3	0.33	0.33
7	0	1.5	3	0.4	0.2
8	0	2.5	5	0.4	0.2
9	0	2	5	0.8	0.1
10	0	2.5	5	0.33	0.33

Table 2

Mixture #	Critical points by GSA				
	Modes			LM	
	\hat{x}_1	\hat{x}_2	\hat{x}_3	\check{x}_1	\check{x}_2
1	1.39				
2	0.90				
3	1.40				
4	0.23				
5	0.23	2.60		1.41	
6	1.50				
7	0.44	2.56		1.50	
8	0.06	2.5	4.94	2.16	2.86
9	0.04	4.96		3.56	
10	0.15	2.5	4.84	1.24	3.76

The investigated two-dimensional mixtures have the following values of parameters: $k = 3, \Sigma_1 = \Sigma_2 = \Sigma_3 = I, I$ is an identity matrix, $\mu_1 = (0, 0), \mu_2 = (0, 2.5), \mu_3 = (2.5, 0), n = 1500$. The weight values π_1, π_2, π_3 are varied. In table 3 their values and estimations of the critical points are represented. They have been got by GSA. In figures 1, 2, 3 the level lines of the surface $f(X)$ and the neighborhoods of the critical points are described. The mode neighborhoods are marked by the red color, the neighborhoods of the saddle points are marked by the green one.

Table 3

Mixture #	Weights		Modes \hat{X}	Saddle points \check{X}
	π_1	π_2		
1	0.75	0.2	(0.009, 0.011)	
2	0.05	0.35	(2.501, 0.013) (0.034, 2.492)	(0.950, 1.262)
3	0.333	0.333	(2.323, -0.001) (0.114, 0.163) (0.012, 2.355)	(1.309, 0.068) (0.060, 1.318)

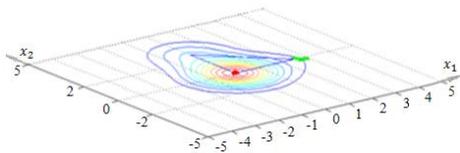


Fig. 1

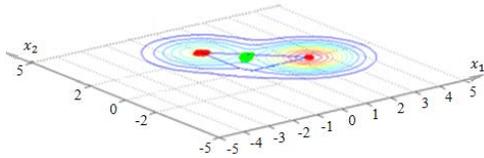


Fig. 2

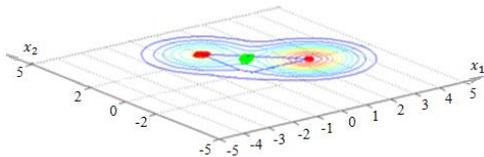


Fig. 3

The gradient-statistical algorithm can be used for the determination of critical points of any smooth function, which has continuous the first and second derivatives, if the closet set K , which contains all its critical points, is known. In this case on the set K the p -dimensional uniformly distributed sample is generated.

REFERENCES

- [1] E. A. Patrick, *Fundamentals of Pattern Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1972; Sov. Radio, Moscow, 1980.
- [2] D. M. Titterington, A. F. M. Smith, and U. E. Makov "Statistical Analysis of Finite Mixture Distributions." in John Wiley, 1985, pp. 53-147.
- [3] T. Ya. Voloshin, I. A. Burdakov, and T. S. Kozenkova, *Statistical Methods for Pattern Recognition Based on the Approximation Approach*. Tikhookean. Okeanol. Inst., Ross. Akad. Nauk, Vladivostok, 1996 [in Russian].
- [4] M. A. Carreira-Perpiñán, *Mode-finding for Mixture of Gaussian Distributions*. Techn. Rep. CS-99-03, Univ. Sheffield, UK, Sheffield, 1999.
- [5] C. A. Robertson, and J. G. Fryer, "Some Descriptive Properties of Normal Mixture," in *Skandinavisk Aktuarietidskrift*, No. 3/4, 1969, pp. 137-149.
- [6] N. N. Aprausheva, and Sorokin S. V. "Mode-Finding for a Mixture of Normal Distributions," in. *Proceedings of VIII All-Russian Conference on Mathematical Methods in Pattern Recognition (Vychisl. Tsentr Ross. Akad. Nauk, Moscow, 1997)*, pp. 4-5 [in Russian].
- [7] B. P. Demidovich, I. A. Maron, *Fundamentals of Computer Mathematics*. Moscow: Nauka, 1970, 664 p. [in Russian].
- [8] N. N. Aprausheva, N. Mollaverdi, and S. V. Sorokin, "Calculation of Stationary Points of Density Probability of the Simplest Gaussian Mixture," in *Tr. ISA RAS. Dinamika neodnorodnykh system*, vyp. 10(2), Moscow: 2006, pp. 113-136 [in Russian].
- [9] M. A. Carreira-Perpiñán, and C. Williams, "On the Number of Modes of a Gaussian Mixture. Inform.," in *Res. Report EDI-INF-RR-0159*. School of Inf., Univ. of Edinburg, 2003, pp. 1-16.
- [10] B. Duran, P. Odell, *Cluster Analysis*. Moscow: Statistika, 1975, 128 p. [in Russian].
- [11] N. G. Zagoruyko, B. N. Elkina, and G. S. Lbov, *Algorithms of Finding Empirical Regularities*. Novosibirsk: Nauka, Sibirskoe otdelenie, 1985, 110 p. [in Russian].
- [12] N. N. Aprausheva, *A New Approach to Detection of Clusters*. Moscow: Computing center RAS, 1993, 65 p. [in Russian].
- [13] V. I. Arnold, *Experimental Mathematics*. Moscow: Fazis, 1999, 63 p. [in Russian].

Hall Current Effect on MHD Free Convection Flow an Inclined Porous Plate with Constant Heat Flux

G.Venkata Ramana Reddy

Abstract— The effect of the Hall current on the magnetohydrodynamic (MHD) natural convection flow from an inclined vertical permeable flat plate with a uniform heat flux is analyzed in the presence of a transverse magnetic field. It is assumed that the induced magnetic field is negligible compared with the imposed magnetic field. The dimensionless momentum equation coupled with energy and mass diffusions are solved by using Nactsheim - Swigert shooting iteration technique. The effects of the various parameters on primary velocity profile, secondary velocity profile, temperature and concentration profile are discussed graphically. The local skin friction coefficient, the local Nusselt number and Sherwood number are shown in tabular form for various values of the parameters.

Keywords— MHD, Heat and mass transfer, Hall Current, inclined Plate, Constant Heat Flux.

INTRODUCTION

Hall current has important contribution in the study of magnetohydrodynamic (MHD) viscous flows. It has many applications in problems of the Hall accelerators as well as in the flight magnetohydrodynamics. The current trend in the application of magnetohydrodynamics is towards a strong magnetic field and a low density of gas. For this reason, the Hall current and ion slip become important. Viscous incompressible fluid flow due to an impulsively started flat plate was examined by Stokes [1]. Rossow [2] examined the flow of a viscous incompressible fluid due to the impulsive motion of an infinite flat plate in the presence of a magnetic field. Sakiadis [3] analyzed analytically and by numerical scheme the boundary layer flow due to a moving flat surface. Laminar compressible boundary layer on a moving flat plate was investigated by Ackroyd [4]. Samuel and Hall [5] obtained the similarity solution for boundary layer flow on a continuous moving porous surface using a series having exponential terms. Sacheti and Bhatt [6], Bhatt and Sacheti [7] investigated Stokes and Rayleigh layers in the presence of a naturally permeable boundary. Hall effects on MHD flows over an accelerated/continuous moving plate are examined by Pop [8], Watanabe and Pop [9], Kiyanjui et al. [10]. Free convection effects on the elasto-viscous fluid flow over an

accelerated plate were examined by Singh et al. [11]. Boundary layer flows in a rotating fluid system are important due to various applications in science and technology. Debnath [12] presented exact solutions of the hydrodynamic and hydromagnetic boundary layer equations in such systems. Takhar and Nath [13], Takhar et al. [14, 15] investigated MHD flow over a stretching surface or moving plate in a rotating fluid. Deka et al. [16] investigated flow over an accelerated plate in a rotating system and Deka [17] examined the Hall effects in such flow in the presence of a magnetic field. Hydromagnetic channel flows in a rotating fluid system are investigated by researchers, e.g. Mandal and Mandal [18], Singh et al. [19], Singh [20], Ghosh [21], Ghosh et al. [22], Seth et al. [23], Guria et al. [24].

The study MHD Free Convection and Mass Transfer Flow of Viscous Incompressible Fluid about an inclined Plate with Hall Current and Constant Heat Flux is investigated.

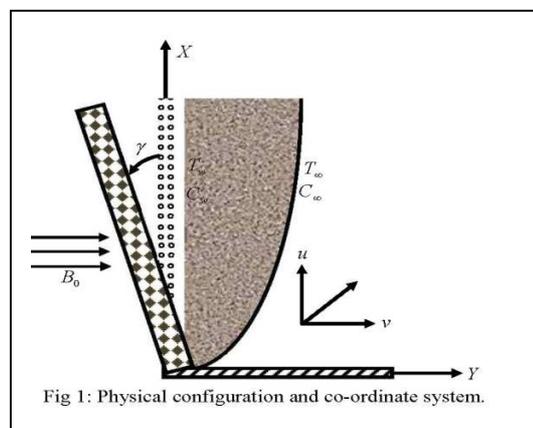


Fig 1: Physical configuration and co-ordinate system.

2. PROBLEM FORMULATION

Consider the steady natural convection boundary layer flow of an electrically conducting and viscous incompressible fluid from a semi-infinite heated permeable vertical inclined flat plate maintained with a uniform surface heat flux in the presence of a transverse magnetic field with the effect of the Hall current. The x axis is along the vertically upward direction, while the y axis is normal to it. The leading edge of the permeable surface is taken to be coincident with the z axis. It is assumed that the uniform heat is supplied from the surface of the plate to the fluid, which is maintained throughout the fluid flow at the uniform rate. The temperature

Dr. G.Venkata Ramana Reddy is working as an Associate Professor in the Department of Mathematics, KL University, Vaddeswaram, Guntur (Dt), Andhra Pradesh, India-522502. Email: gvr1976@kluniversity.in

and concentration at the wall are instantly raised from T_w and C_w to T_∞ and C_∞ respectively. An electrically non-conducting uniform magnetic field of magnitude B_0 is imposed to perpendicular to the flow along the y axis. Let the angle of inclination of the plate is γ and the plate is semi finite. The x component momentum equation reduces to the boundary layer equation if and only if body force is made by gravity, then the body force per unit mass is $F_x = -\rho g_0 \cos \gamma$, where g_0 is the local acceleration due to gravity. Further no body force exists in the direction of y and z , i. e. $\frac{\partial p}{\partial y} = 0, \frac{\partial p}{\partial z} = 0$, and $F_y = 0, F_z = 0$. The x component of pressure gradient at any point in the boundary layer must equal to the pressure gradient in the region outside the boundary layer, in this region $u = 0, v = 0$. Hence x component of pressure gradient become $\frac{\partial p}{\partial x} = -\rho_\infty g \cos \gamma$,

where ρ_∞ is the density of the surrounding fluid at temperature T_∞ . The quantity $\rho - \rho_\infty$ is related to the temperature difference $T - T_\infty$ and concentration (or mass) differences $C - C_\infty$ through the thermal volume expansion coefficient β and volume expansion coefficient β^* by the relation $\frac{\rho - \rho_\infty}{\rho} = -\beta(T - T_\infty) - \beta^*(C - C_\infty)$

$$\therefore F_x - \frac{1}{\rho} \frac{\partial p}{\partial x} = g\beta(T - T_\infty) \cos \gamma + g\beta^*(C - C_\infty) \cos \gamma.$$

We have the generalized ohm's law in the absence of electric field to the case of short circuit problem is of the form

$$\underline{J} + \frac{\omega_e \tau_e}{B_0} \underline{J} \times \underline{B} = \sigma(\underline{E} + \mu_e \underline{q} \times \underline{B}) \quad (1) \text{ where, } \mu_e \text{ is the}$$

magnetic permeability, τ_e is the electron collision time, σ is the time dependent length scale, ω_e is the cyclotron frequency, B_0 is the applied magnetic field.

Since no applied or polarized voltage exist, So the effect of polarization of fluid is negligible, i. e. $\underline{E} \equiv (0,0,0)$. Therefore

$$\text{equation (1) becomes } \underline{J} + \frac{\omega_e \tau_e}{B_0} \underline{J} \times \underline{B} = \sigma \mu_e \underline{q} \times \underline{B} \quad (2)$$

If is assumed that induced magnetic field generated by fluid motion is negligible in comparison to the applied one i. e. $\underline{B} \equiv (0, B_0, 0)$. This assumption is valid because magnetic Reynolds number is very small for liquid metals and partially ionized fluids.

Since the Hall coefficient is $m = \omega_e \tau_e$, so the equation (2) we can write

$$J_z = \frac{\sigma \mu_e B_0}{1 + m^2} (mw + u) \quad (3)$$

$$\text{and } J_x = \frac{\sigma \mu_e B_0}{1 + m^2} (mu - w) \quad (4)$$

The fundamental equations for the steady incompressible MHD flow with the generalized Ohm's law and Maxwell's equations, under the assumptions that the fluid is quasi-neutral, and the ion slip and thermoelectric effects can be neglected. Since the plate is semi-infinite and motion is steady, all physical equations will be the functions of x and y . Thus mathematically the problem reduces to a two dimensional problem given as follows:

$$\frac{\partial u}{\partial x} + \frac{\partial u}{\partial y} = 0 \quad (5)$$

$$u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = \nu \frac{\partial^2 u}{\partial y^2} + g_0 \beta (T - T_\infty) \cos \gamma + g\beta^* (C - C_\infty) \cos \gamma - \frac{\sigma B_0^2}{\rho(1 + m^2)} (u + mw) \quad (6)$$

$$u \frac{\partial w}{\partial x} + v \frac{\partial w}{\partial y} = \nu \frac{\partial^2 w}{\partial y^2} + \frac{\sigma B_0^2}{\rho(1 + m^2)} (mu - w) \quad (7)$$

$$u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} = \frac{k}{\rho c_p} \frac{\partial^2 T}{\partial y^2} \quad (8)$$

$$u \frac{\partial C}{\partial x} + v \frac{\partial C}{\partial y} = D_m \frac{\partial C}{\partial y^2} + \frac{\nu}{c_p} \left[\left(\frac{\partial u}{\partial y} \right)^2 + \left(\frac{\partial w}{\partial y} \right)^2 \right] \quad (9)$$

Subjected to the boundary conditions

$$\left. \begin{aligned} u=0, v=0, w=0, \frac{\partial T}{\partial y} = -\frac{Q}{k}, C=C_w \text{ at } y=0 \\ u \rightarrow 0, w \rightarrow 0, T \rightarrow T_\infty, C \rightarrow C_\infty \text{ as } y \rightarrow \infty \end{aligned} \right\} \quad (10)$$

where u, v, w are the velocity components in the x, y, z direction respectively, ν is the kinematics viscosity, ρ is the density. T, T_w and T_∞ are the temperature of the fluid inside the thermal boundary layer, the plate temperature and the fluid temperature in the free stream, respectively, while C, C_w, C_∞ are the corresponding concentrations. Also, σ is the electric conductivity of the medium, k is the thermal conductivity of the medium, D_m is the coefficient of mass diffusivity, c_p is the specific heat at constant pressure, Q is the constant heat flux per unit area and other symbols have their usual meaning.

In order to solve the above system of equations (6)-(9) with the boundary conditions (10), we adopt the well-defined similarity analysis to attain similarity solutions. For this purpose, the following similarity transformations are now introduced;

$$\eta = y\sqrt{\frac{U_0}{2vx}}, \quad g_0(\eta) = \frac{w}{U_0},$$

$$\theta(\eta) = \frac{k(T-T_\infty)}{Q}\sqrt{\frac{U_0}{2vx}}, \quad \phi(\eta) = \frac{C-C_\infty}{(C_0-C_\infty)},$$

$$\psi = \sqrt{2vxU_0}f(\eta), \quad u = \frac{\partial\psi}{\partial y} = U_0 f'(\eta)$$

and

$$v = -\frac{\partial\psi}{\partial x} = \sqrt{\frac{U_0 v}{2x}} [\eta f'(\eta) - f(\eta)] \quad (11) \text{ Thus,}$$

equations (6)-(10) becomes;

$$f''' + ff'' + G_r \cos \gamma \theta + G_m \cos \gamma \phi - \frac{M}{1+m^2}(f' + mg) = 0 \quad (12)$$

$$g'' + fg' + \frac{M}{1+m^2}(mf' + g) = 0 \quad (13)$$

$$\theta'' + P_r E_c [(f'')^2 + (g')^2] - P_r (f'\theta - f\theta') = 0 \quad (14)$$

$$\phi'' + S_c f\phi' = 0 \quad (15)$$

The corresponding boundary conditions are

$$\left. \begin{aligned} f(\eta) = f_w, f'(\eta) = 0, g(\eta) = 0, \theta'(\eta) = -1, \\ \phi(\eta) = 1 \text{ at } \eta = 0 \\ f'(\eta) \rightarrow \infty, g(\eta) \rightarrow \infty, \theta(\eta) \rightarrow 0, \\ \phi(\eta) \rightarrow 0 \text{ at } \eta \rightarrow \infty \end{aligned} \right\} \quad (16)$$

where $P_r = \frac{k}{\rho C_p \nu}$ is the Prandtl number,

$$E_c = \frac{U_0^3 k}{C_p Q \sqrt{2vxU_0}} \text{ is the Eckert number, } M = \frac{2x\sigma B_0^2}{\rho U_0}$$

is the Magnetic parameter, $S_c = \frac{\nu}{D_m}$ is the Schmidt number,

$$G_r = \frac{g_0 \beta Q \sqrt{2x}}{k U_0 \sqrt{\nu U_0}} \text{ is the local Grashof number,}$$

$$G_m = \frac{2g\beta^*(C_w - C_\infty)x}{U_0^2} \text{ is the local modified Grashof}$$

number, $f_w = v_0 \sqrt{\frac{2x}{U_0 \nu}}$ is the Transpiration parameter.

3. SKIN-FRICTION COEFFICIENTS, NUSSELT NUMBER AND SHERWOOD NUMBER

The quantities of chief physical interest are the skin friction coefficients, the Nusselt number and the Sherwood number. The equation defining the wall skin frictions are

$$\tau_x = \mu \left(\frac{\partial u}{\partial y} \right)_{y=0} \quad \text{and} \quad \tau_z = \mu \left(\frac{\partial w}{\partial y} \right)_{y=0} \quad \text{which are}$$

proportional to $\left(\frac{\partial^2 f}{\partial \eta^2} \right)_{\eta=0}$ and $\left(\frac{\partial g}{\partial \eta} \right)_{\eta=0}$. The Nusselt

number denoted by N_u is proportional to $-\left(\frac{\partial T}{\partial y} \right)_{y=0}$,

hence we have $N_u \propto -\theta'(0)$. The Sherwood number

denoted by S_h is proportional to $-\left(\frac{\partial C}{\partial y} \right)_{y=0}$, hence we

have $S_h \propto -\phi'(0)$. The numerical values of the skin-friction coefficients, the Nusselt number and the Sherwood number are sorted in Tables 1-8.

4. RESULTS AND DISCUSSION

In this study the MHD Free Convection and Mass Transfer Flow of Viscous Incompressible Fluid about an inclined Plate with Hall Current and Constant Heat Flux have been investigated using the Nachtsheim-Swigert shooting iteration technique. To study the physical situation of this problem, we have computed the numerical values of the velocity, temperature, and concentration within the boundary layer and also find the skin friction coefficient, Nusselt number, Sherwood number at the plate. It can be seen that the solutions are affected by the parameters, namely suction parameter f_w , Grashof number G_r , modified Grashof number G_m , magnetic parameter M , Prandtl number P_r , Eckert number E_c , Schmidt number. The values of M and G_r are taken to be large for cooling Newtonian fluid keeping the plate at different angle. The values 0.2, 0.5, 0.73, 2, 3, 4, 5 are considered for P_r . The values 0.1, 0.5, 0.6, 1.0, 2.0, 3.0, 4.0 also considered for S_c . The values of other parameters are however chosen arbitrarily.

Figures (2)-(5), respectively, show the primary velocity, secondary velocity, temperature and concentration profiles for different values of suction parameter f_w . Here $f_w > 0$ corresponds to suction and $f_w < 0$ corresponds to injection at the plate or blowing. From Figure (2-5), it can be seen that the primary velocity, secondary velocity, temperature and concentration profiles decreases with the increase of suction parameter f_w . Figures (6)-(9), respectively, show the the primary velocity, secondary velocity profiles decreased and temperature and concentration profiles increases for different values of M . Figures (10)-(11), respectively, show the cross-flow of primary velocity and secondary velocity, at first increases then decreases with the increase of E_c . Figure (12-13) shows that the temperature profile increase and

concentration profile decreases with the increase of E_c . Figures (14)-(17) show that the primary velocity, secondary velocity profile and concentration profile decreases and temperature profile increases with the increase of S_c . Figures (18)-(19), respectively, shows the cross flow of the primary velocity and secondary velocity with the increase of P_r both of the profile is decrease then increase. Figures (20)-(21) shows that temperature decrease and concentration profile increase with the increase of S_c . Figures (22)-(23), show the cross flow of the primary velocity and secondary velocity with the increase of γ both of the profile is decrease then increase. Figures (24)-(25), show that temperature and concentration profile increases with the increase of γ . Figures (26)-(27), show the cross flow of the primary velocity and secondary velocity with the increase of G_r . Figures (28)-(29), shows that the temperature and concentration profile decreases with the increase of G_r . Figures (30)-(31), show the cross flow of the primary velocity and secondary velocity with the increase of G_m . Figures (32)-(33), shows that the temperature and concentration profile decreases with the increase of G_m .

From figures (34)-(37), show the velocity, secondary velocity, temperature and concentration profile field has a negligible effect for different values of m . Finally the effect of various parameters on the skin friction coefficients (τ_x, τ_w), Nusselt number (N_u) and Sherwood (S_h) are tabulated in Tables 1-8. Table 1 shows that the skin friction coefficient coefficients (τ_x, τ_w) decreases and Nusselt number (N_u) and Sherwood number (S_h) increase with the increase of f_w . Table 2 shows that the skin friction coefficient coefficients τ_x decreases and τ_w increases and Nusselt number (N_u) and Sherwood number (S_h) decreases with the increase of M . Table 3 shows that the skin friction coefficient coefficients (τ_x, τ_w) and Sherwood number (S_h) increases and Nusselt number (N_u) decreases with the increase of E_c . Table 4 shows that the skin friction coefficient coefficients (τ_x, τ_w) and Sherwood number (S_h) decreases and Nusselt number (N_u) increases with the increase of P_r . Table 5 shows that the skin friction coefficient coefficients (τ_x, τ_w) and Nusselt number (N_u) decreases and Sherwood number (S_h) increases with the increase of S_c . Table 6 shows that the skin friction coefficient coefficients (τ_x, τ_w), Nusselt number (N_u) and Sherwood number (S_h) decreases with the increase of γ .

Table 7-8 shows that the skin friction coefficient coefficients (τ_x, τ_w), Nusselt number (N_u) and Sherwood number (S_h) increases with the increase of G_r and G_m .

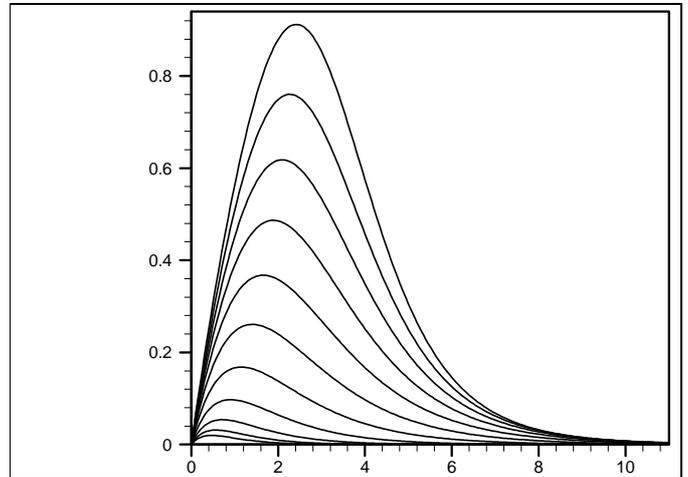


Fig 2: Primary velocity profile for f_w .

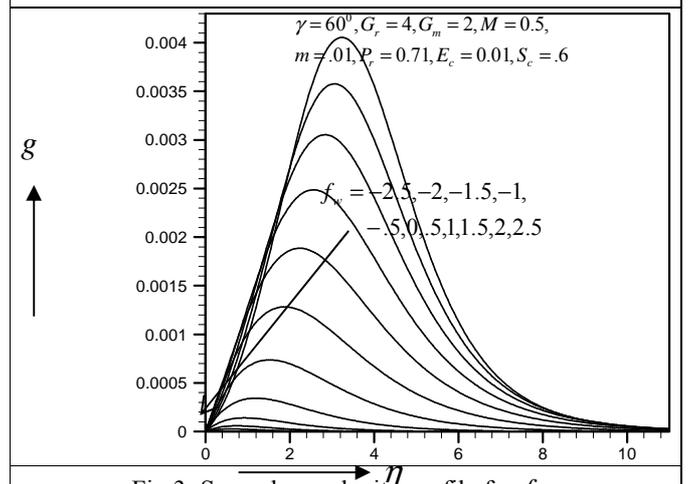


Fig 3: Secondary velocity profile for f_w .

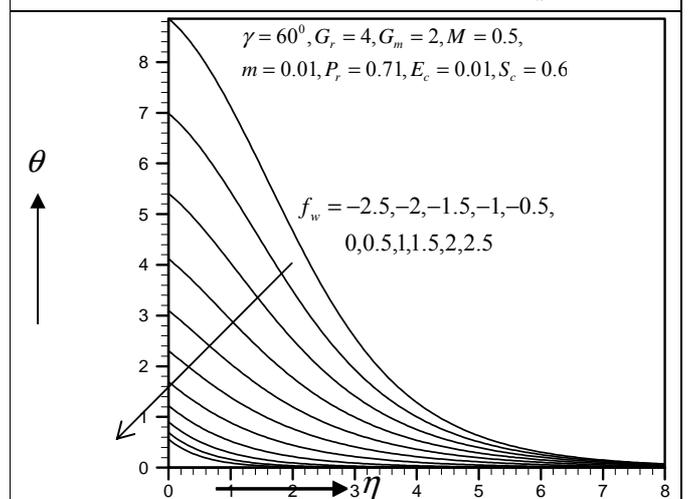


Fig 4: Temperature profile for f_w .

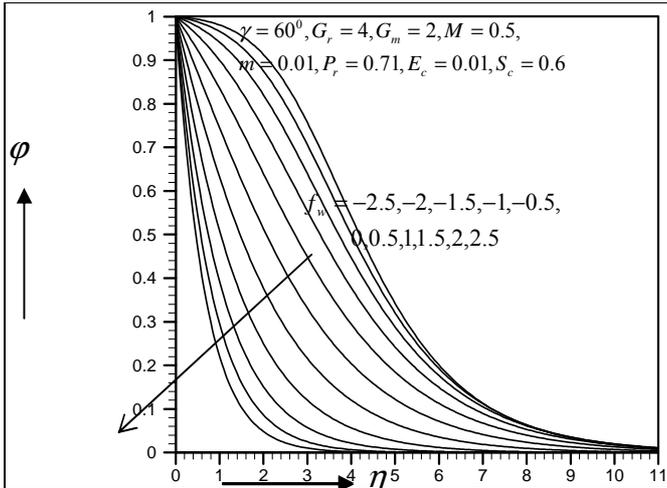


Fig 5: Concentration profile for f_w .

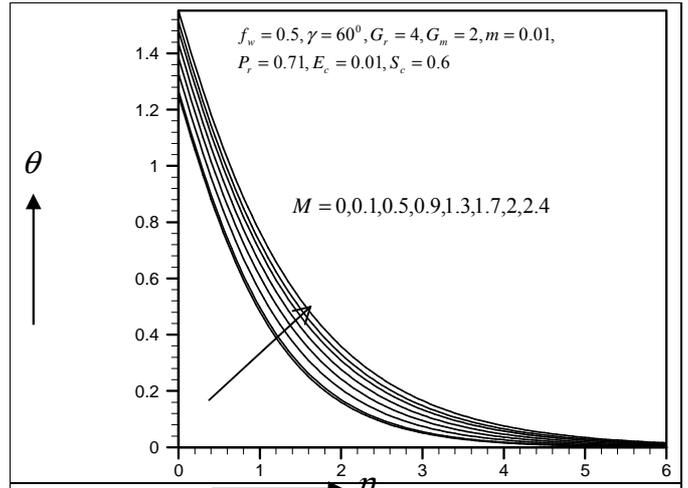


Fig 8: Temperature profile for M .

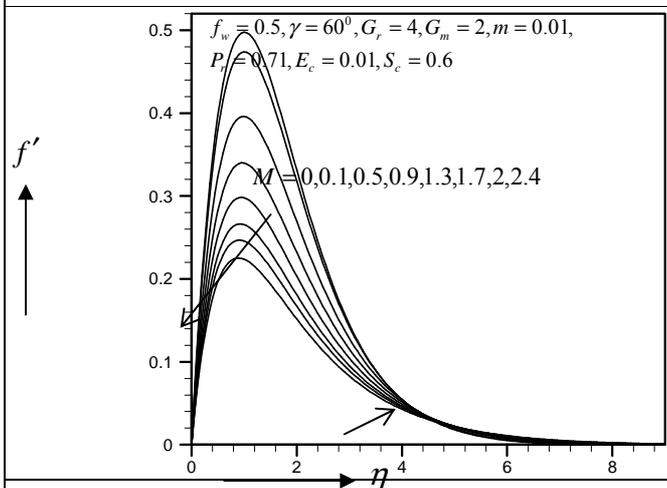


Fig 6: Primary velocity profile for M .

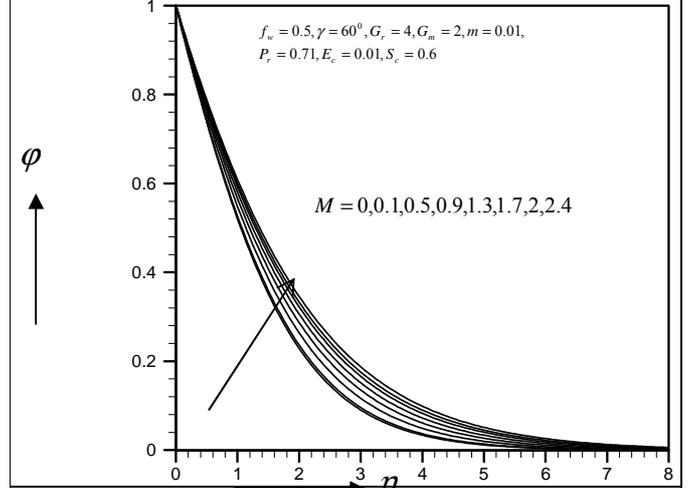


Fig 9: Concentration profile for M .

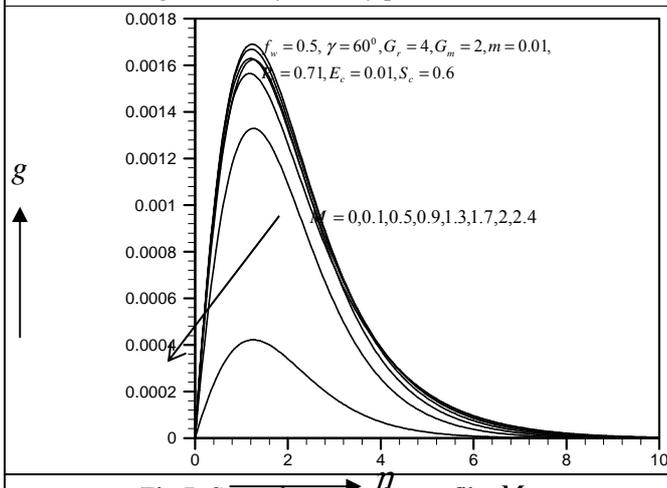


Fig 7: Secondary velocity profile M .

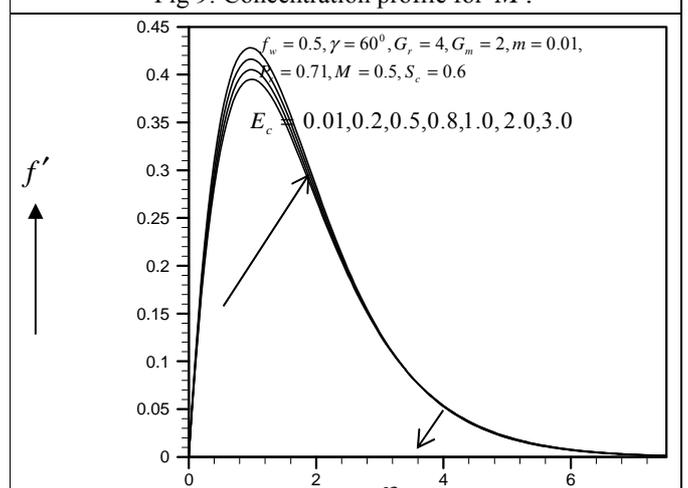


Fig 10: Primary velocity profile for E_c .

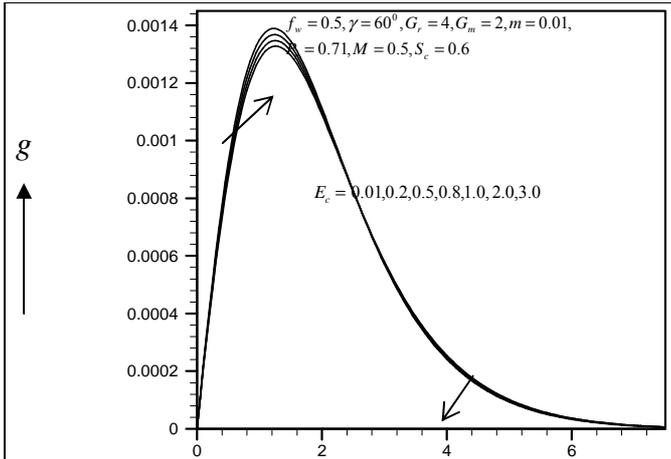


Fig 11: Secondary velocity profile E_c .

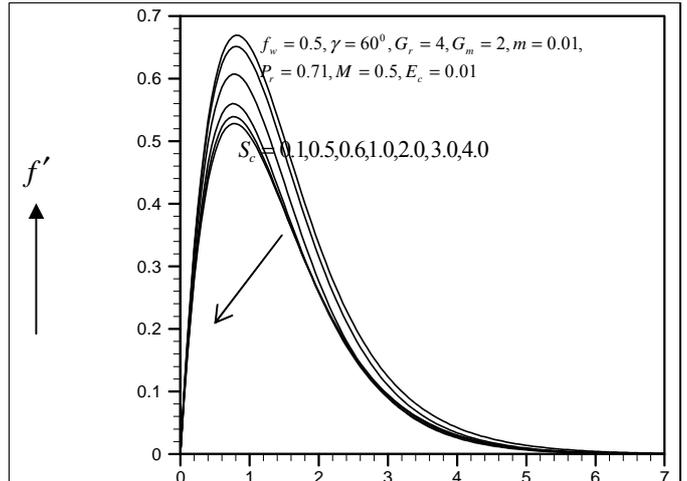


Fig 14: Primary velocity profile for S_c .

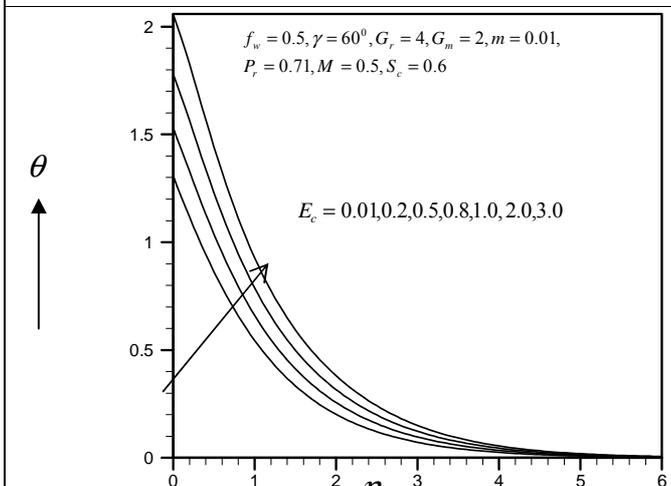


Fig 12: Temperature profile for E_c .

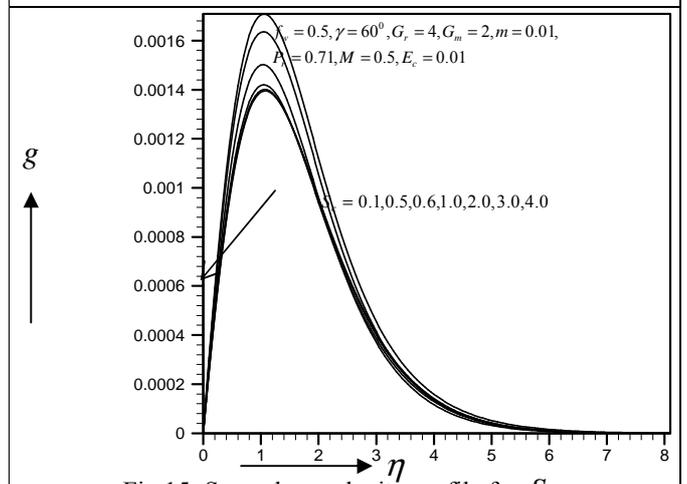


Fig 15: Secondary velocity profile for S_c .

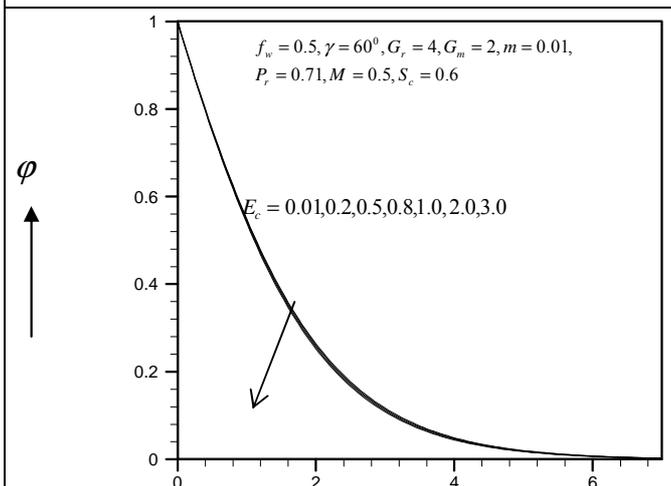


Fig13: Concentration profile for E_c .

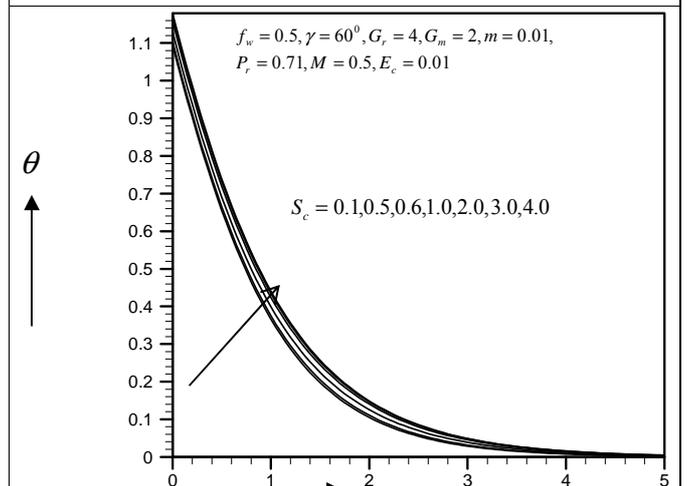


Fig 16: Temperature profile for S_c .

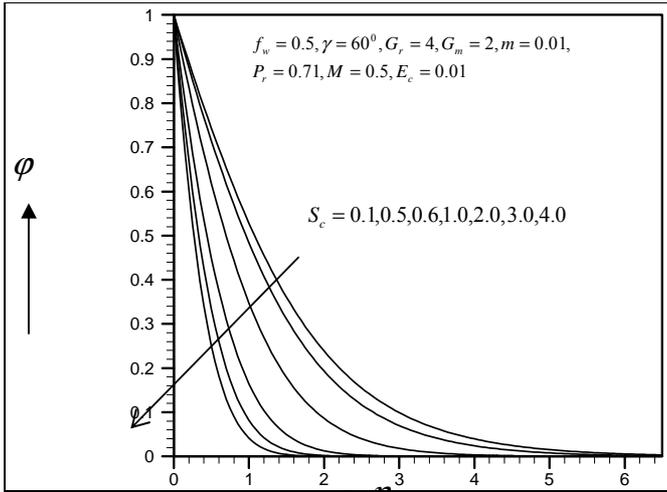


Fig 17: Concentration profile for S_c .

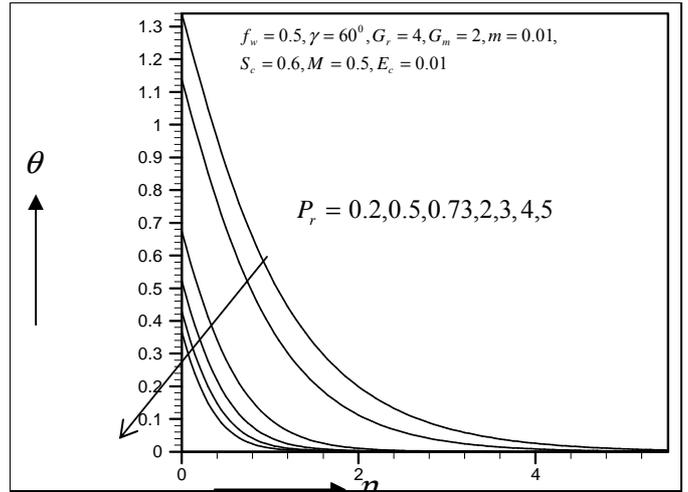


Fig 20: Temperature profile for P_r .

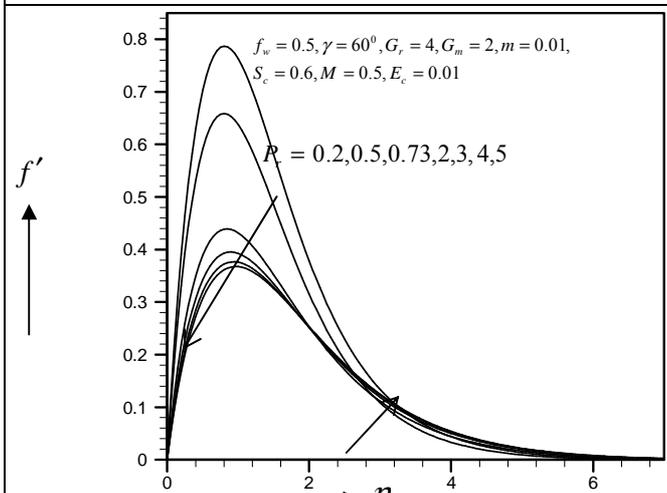


Fig 18: Primary velocity profile P_r .

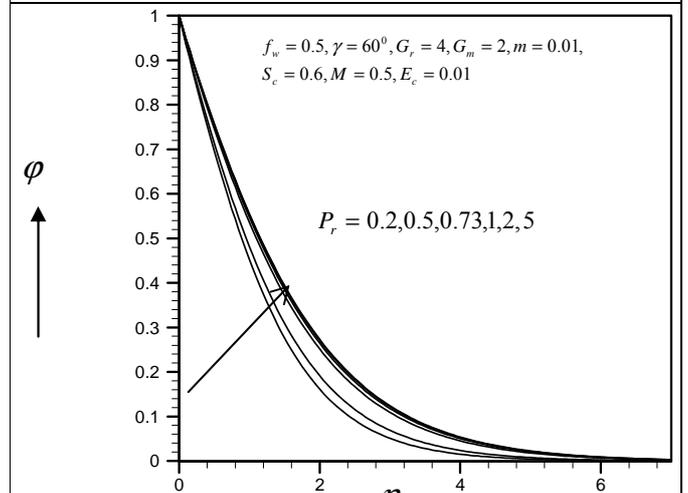


Fig 21: Concentration profile for P_r .

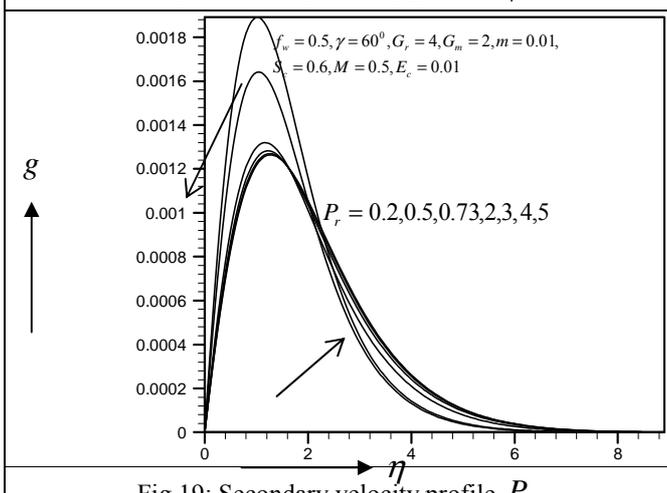


Fig 19: Secondary velocity profile P_r .

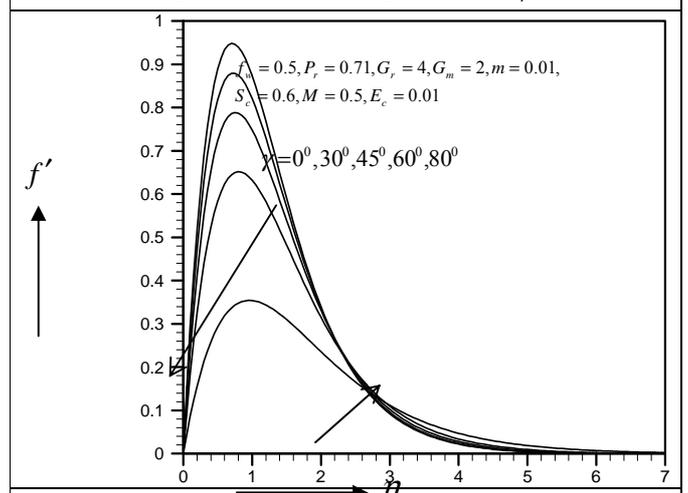


Fig 22: Primary velocity profile for γ .

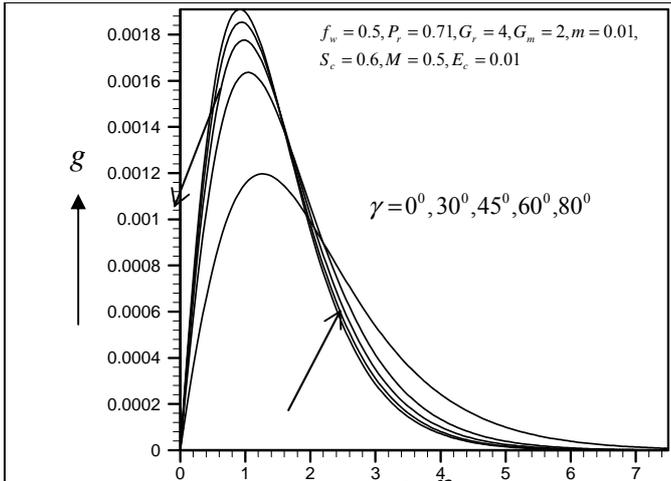


Fig 23: Secondary velocity profile for γ .

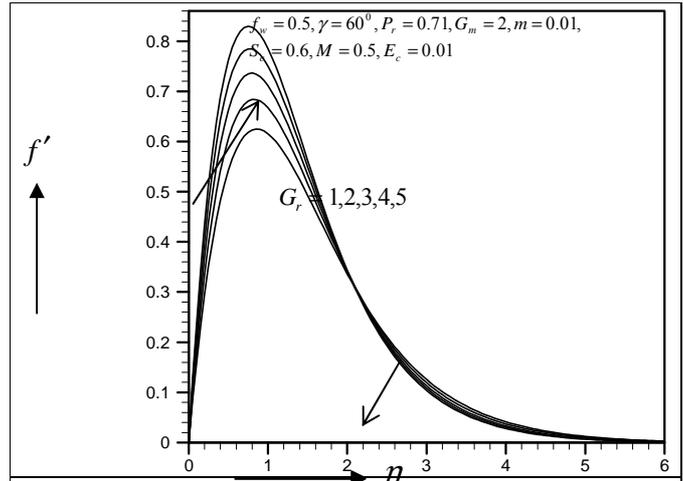


Fig 26: Primary velocity profile for G_r .

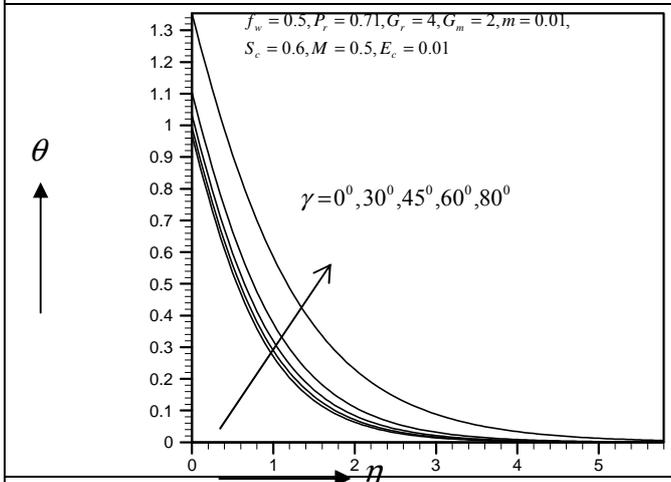


Fig 24: Temperature profile for γ .

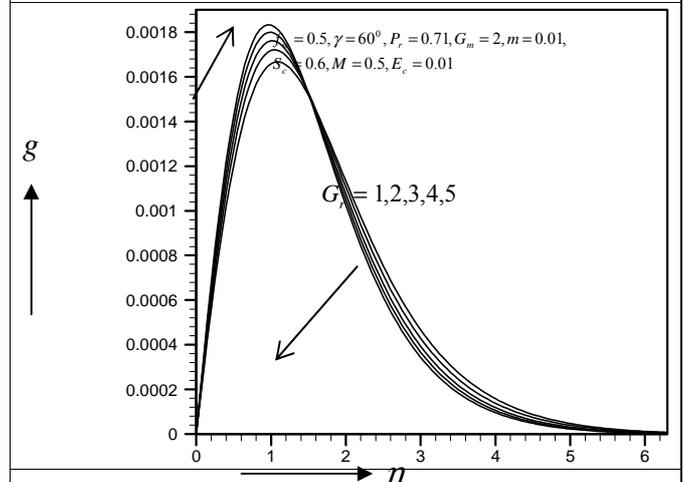


Fig 27: Secondary velocity profile for G_r .

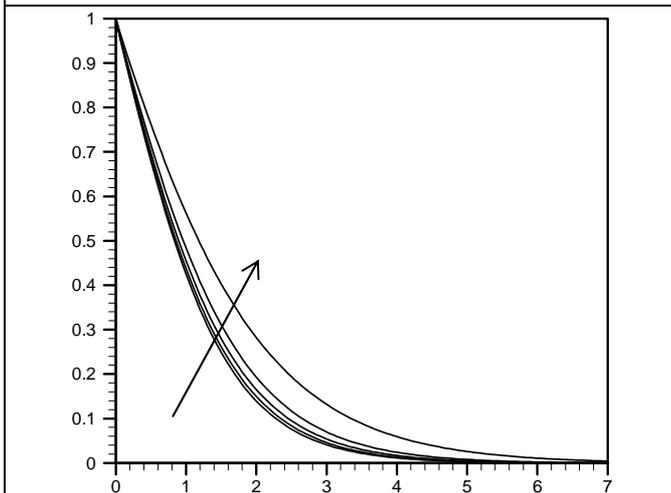


Fig 25: Concentration profile for γ .

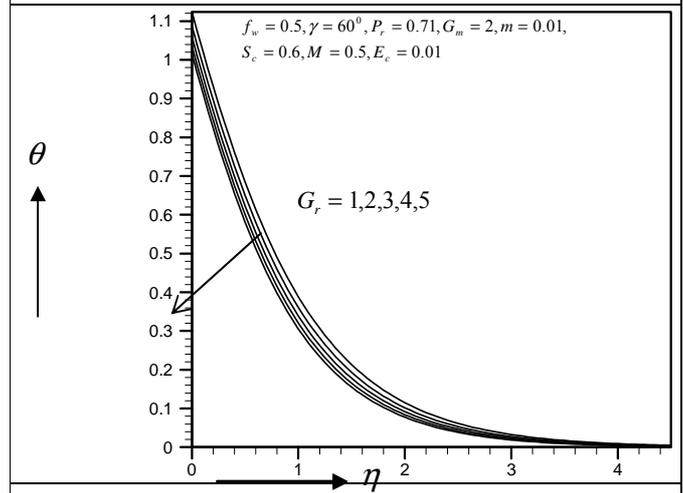


Fig 28: Temperature profile for G_r .

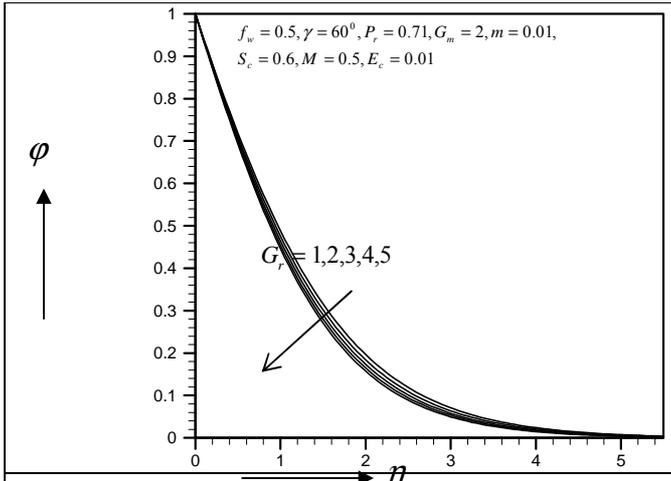


Fig 29: Concentration profile for G_r .

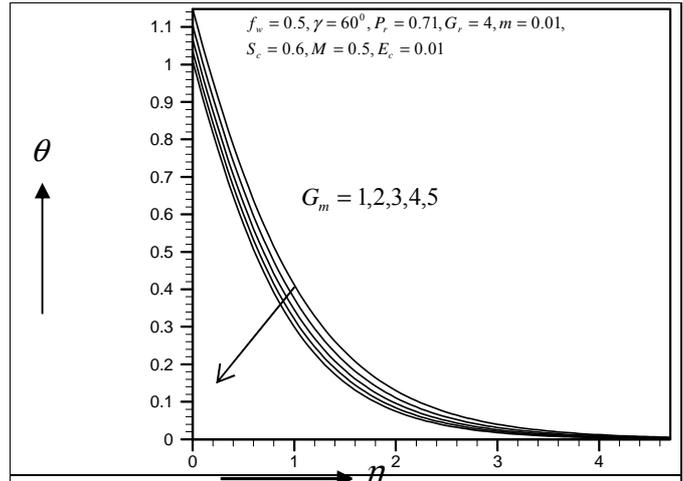


Fig 32: Temperature profile for G_m .

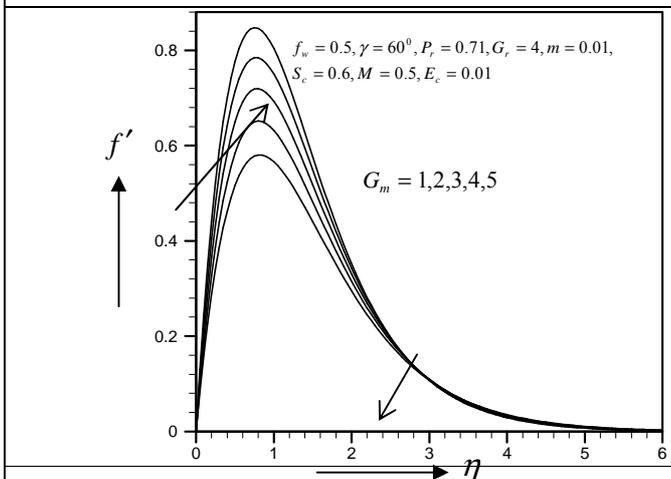


Fig 30: Primary velocity profile for G_m .

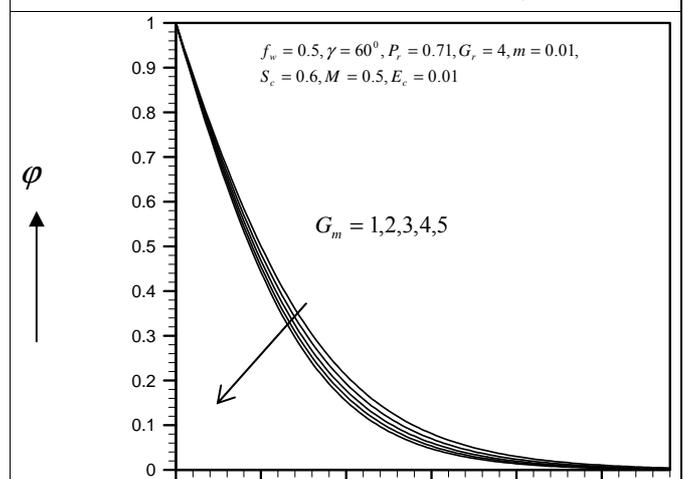


Fig 33: Concentration profile for G_m .

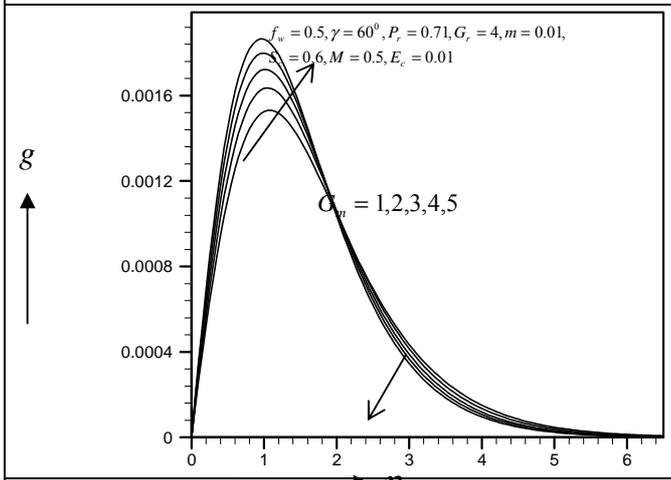


Fig 31: Secondary velocity profile for G_m .

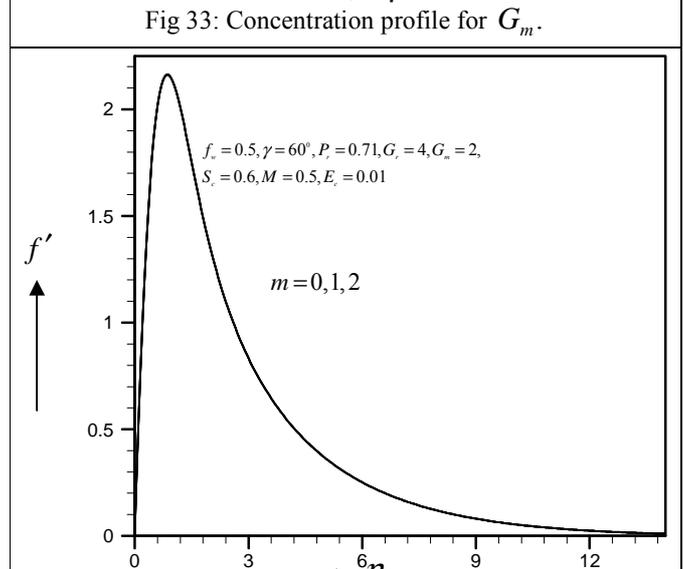


Fig 34: Primary velocity profile for m .

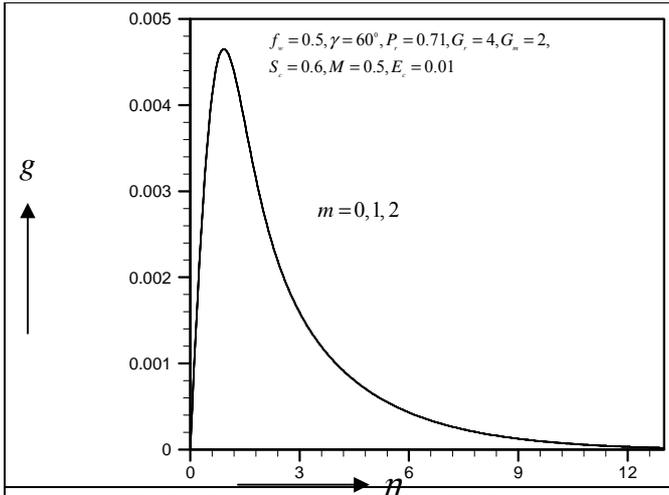


Fig 35: Secondary velocity profile for m .

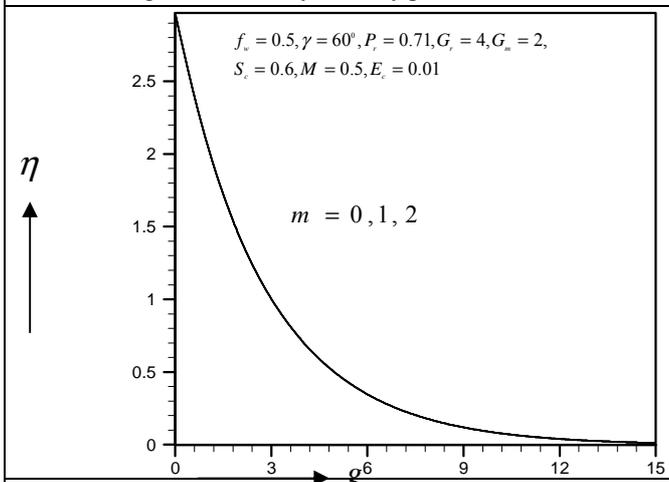


Fig 36: Temperature profile for m .

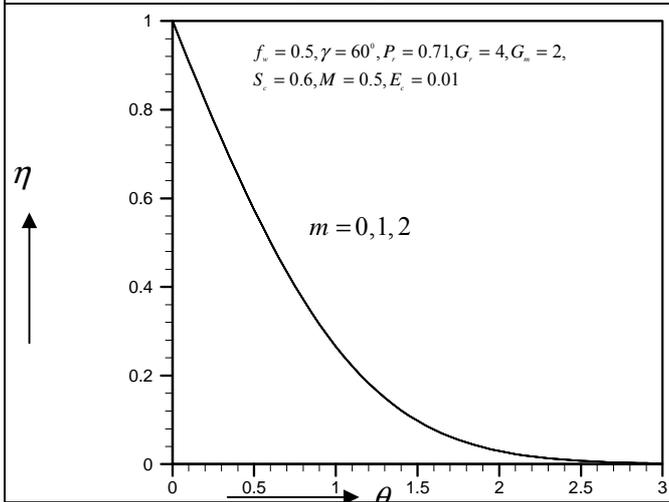


Fig 37: Concentration profile for m .

Table 1. Numerical values of Skin friction coefficient τ_x, τ_w , Nusselt number N_u and Sherwood S_h for different values

of f_w , taking $\gamma = 60^\circ$, $G_r = 4$, $G_m = 2$, $M = 0.5$, $m = 0.01$, $P_r = 0.71$, $E_c = 0.01$, $S_c = 0.6$ as fixed.

f_w	τ_x	τ_w	N_u	S_h
-2.5	0.6668	0.0004	-8.8553	0.0084
-2.0	0.6287	0.0005	-6.9865	0.0178
-1.5	0.5931	0.0007	-5.4077	0.0368
-1.0	0.5585	0.0009	-4.1194	0.0741
-0.5	0.5191	0.0011	-3.1007	0.1427
0.0	0.4644	0.0012	-2.3082	0.2573
0.5	0.3853	0.0010	-1.6914	0.4284
1.0	0.2902	0.0006	-1.2231	0.6558
1.5	0.2076	0.0003	-0.9004	0.9235
2.0	0.1512	0.0001	-0.6940	1.2113
2.5	0.1151	0.0001	-0.5600	1.5061

Table 2. Numerical values of Skin friction coefficient τ_x, τ_w , Nusselt number N_u and Sherwood S_h for different values of M , taking $f_w = 0.5$, $\gamma = 60^\circ$, $G_r = 4$, $G_m = 2$, $m = 0.01$, $P_r = 0.71$, $E_c = 0.01$, $S_c = 0.6$ as fixed.

M	τ_x	τ_w	N_u	S_h
0.0	1.1932	0.0000	-1.2511	0.5666
0.1	1.1505	0.0006	-1.2670	0.5586
0.5	1.0120	0.0021	-1.3286	0.5313
0.9	0.9125	0.0026	-1.3847	0.5096
1.3	0.8378	0.0028	-1.4353	0.4921
1.7	0.7796	0.0028	-1.4811	0.4777
2.0	0.7436	0.0028	-1.5128	0.4684
2.4	0.7029	0.0028	-1.5518	0.4577

Table 3. Numerical values of Skin friction coefficient τ_x, τ_w , Nusselt number N_u and Sherwood S_h for different values of E_c , taking $f_w = 0.5$, $\gamma = 60^\circ$, $G_r = 4$, $G_m = 2$, $m = 0.01$, $P_r = 0.71$, $M = 0.5$, $S_c = 0.6$ as fixed.

Ec	τ_x	τ_w	N_u	S_h
0.1	1.0092	0.0021	-1.3092	0.5309
0.2	1.0150	0.0021	-1.3503	0.5316
0.5	1.0244	0.0021	-1.4168	0.5328
0.8	1.0341	0.0021	-1.4853	0.5339
1.0	1.0407	0.0021	-1.5321	0.5347
2.0	1.0755	0.0022	-1.7809	0.5388
3.0	1.1139	0.0022	-2.0577	0.5432

Table 4. Numerical values of Skin friction coefficient τ_x, τ_w , Nusselt number N_u and Sherwood S_h for different values of Pr , taking

$f_w = 0.5, \gamma = 60^\circ, G_r = 4, G_m = 2, m = 0.01, S_c = 0.6, M = 0.5, E_c = 0.01$ as fixed.

Pr	τ_x	τ_w	N_u	S_h
0.2	3.4971	0.0054	-1.8970	0.7623
0.5	2.4064	0.0035	-1.3401	0.6562
0.7	2.0548	0.0030	-1.1387	0.6180
2.0	1.3704	0.0022	-0.6750	0.5476
3.0	1.1933	0.0021	-0.5234	0.5329
4.0	1.0984	0.0020	-0.4305	0.5261
5.0	1.0409	0.0020	-0.3672	0.5225

Table 5. Numerical values of Skin friction coefficient τ_x, τ_w , Nusselt number N_u and Sherwood S_h for different values of S_c , taking $f_w = 0.5, \gamma = 60^\circ, G_r = 4, G_m = 2, m = 0.01, P_r = 0.71, M = 0.5, E_c = 0.01$ as fixed.

Sc	τ_x	τ_w	N_u	S_h
0.1	2.2727	0.0046	-1.0157	0.2099
0.5	2.0455	0.0031	-1.0939	0.5470
0.6	2.0206	0.0030	-1.1038	0.6158
1.0	1.9525	0.0028	-1.1296	0.8666
2.0	1.8621	0.0026	-1.1585	1.4158
3.0	1.8106	0.0025	-1.1714	1.9240
4.0	1.7751	0.0025	-1.1788	2.4178

Table 6. Numerical values of Skin friction coefficient τ_x, τ_w , Nusselt number N_u and Sherwood S_h for different values of γ , taking $f_w = 0.5, G_r = 4, G_m = 2, m = 0.01, P_r = 0.71, M = 0.5, E_c = 0.01, S_c = 0.6$ as fixed.

γ	τ_x	τ_w	N_u	S_h
0°	3.2825	0.0039	-0.9686	0.6969
30°	2.9778	0.0037	-0.9941	0.6795
45°	2.5835	0.0035	-1.0328	0.6551
60°	2.0206	0.0030	-1.1038	0.6158
80°	0.9409	0.0019	-1.3532552	0.5141570

Table 7. Numerical values of Skin friction coefficient τ_x, τ_w , Nusselt number N_u and Sherwood S_h for different values of G_r , taking $f_w = 0.5, \gamma = 60^\circ, P_r = 0.71, G_m = 2, m = 0.01, S_c = 0.6, M = 0.5, E_c = 0.01$ as fixed

Gr	τ_x	τ_w	N_u	S_h
1.0	1.7983	0.0029	-1.1233	0.6084
2.0	2.0536	0.0031	-1.0878	0.6260
3.0	2.2878	0.0033	-1.0594	0.6411
4.0	2.5072	0.0035	-1.0357	0.6544
5.0	2.7144	0.0036	-1.0155	0.6664

Table 8. Numerical values of Skin friction coefficient τ_x, τ_w , Nusselt number N_u and Sherwood S_h for different values of G_m , taking $f_w = 0.5, \gamma = 60^\circ, P_r = 0.71, G_r = 4, m = 0.01, S_c = 0.6, M = 0.5, E_c = 0.01$ as fixed.

G_m	τ_x	τ_w	N_u	S_h
1.0	1.7760	0.0027	-1.1479	0.5935
2.0	2.0206	0.0030	-1.1038	0.6158
3.0	2.2646	0.0032	-1.0670	0.6361
4.0	2.5072	0.0035	-1.0357	0.6544
5.0	2.7474	0.0037	-1.0088	0.6713

REFERENCES

- [1] G.G. Stokes. On the effect of the internal friction of fluids on the motion of pendulums, *Camb. Phil. Trans.* 9 (1851):8-106.
- [2] V.J. Rossow. On Rayleigh's problem in magnetohydrodynamics, *Phys. Fluids* 3(3) (1960):395-398.
- [3] B.C. Sakiadis. Boundary layer behavior on continuous solid surfaces: II. The boundary layer on a continuous flat surface, *AIChE J.* 7(2) (1961):221-225.
- [4] J.A.D. Ackroyd. On the laminar compressible boundary layer with stationary origin on a moving flat wall, *Proc. Cam. Philos. Soc.* 63 (1967):871-888.
- [5] T.D.M.A. Samuel and I.M. Hall. On the series solution to the laminar boundary layer with stationary origin in a continuous moving porous surface, *Proc. Camb. Phil. Soc.* 73 (1973):223-229.
- [6] N.C. Sacheti and B.S. Bhatt. Stokes and Rayleigh layers in the presence of naturally permeable boundaries, *J. Eng. Mech.* 110 (1984):713-722.
- [7] B.S. Bhatt and N.C. Sacheti. Effect of rotation on Rayleigh layers in presence of naturally permeable boundaries, *J. Eng. Mech.* 113 (1987):1795-1800.

- [8] I. Pop. The effect of Hall currents on hydromagnetic flow near an accelerated plate, *J. Math. Phys. Sci.* 5 (1971):375-379.
- [9] T. Watanabe and I. Pop. Hall effects on magnetohydrodynamic boundary layer flow over a continuous moving flat plate, *Acta Mechanica* 108 (1995):35-47.
- [10] M. Kinyanjui, J.K. Kwanza and S.M. Uppal. Magnetohydrodynamic free convection heat and mass transfer of a heat generating fluid past an impulsively started infinite vertical porous plate with Hall current and radiation absorption, *Energy Conversion and Management* 42(8) (2001):917- 931.
- [11] J. Singh, S.K. Gupta and S. Chandrasekaran. Computational treatment of free convection effect on flow of elastico-viscous fluid past an accelerated plate with constant heat flux, *Applied Mathematics and Computation* 217(2) (2010):685-688.
- [12] L. Debnath. Exact solutions of the unsteady hydrodynamic and hydro-magnetic boundary layer equations in a rotating fluid system, *ZAMM* 55 (1975):431-438.
- [13] H.S. Takhar and G. Nath. Unsteady flow over a stretching surface with a magnetic field in a rotating fluid, *ZAMP* 49 (1998):989-1001.
- [14] H.S. Takhar, A.J. Chamkha and G. Nath. MHD flow over a moving plate in a rotating fluid with magnetic field, Hall currents and free stream velocity. *Int. J. Eng. Sci.* 40(13) (2002):1511-1527.
- [15] H.S. Takhar, A.J. Chamkha and G. Nath. Flow and heat transfer on a stretching surface in a rotating fluid with a magnetic field, *International Journal of Thermal Sciences*, 42(1) (2003):23-31.
- [16] R.K. Deka, A.S. Gupta, H.S. Takhar and V.M. Soundalgekar. Flow past an accelerated horizontal plate in a rotating fluid, *Acta Mechanica* 138 (1999):13-19.
- [17] R.K. Deka. Hall effects on MHD flow past an accelerated plate, *Theoret. Appl. Mech.* 35(4) (2008):333-346.
- [18] G. Mandal and K.K Mandal. Effect of Hall current on MHD couette flow between thick arbitrarily conducting plate in a rotating system, *J. Physical Soc. Japan* 52 (1983):470-477.
- [19] A.K. Singh, N.C. Sacheti and P. Chandran. Transient effects in magneto-hydrodynamic couette flow with rotation: accelerated motion, *Int. J. Engng. Sci.* 32 (1994):133-139
- [20] K.D. Singh. An oscillatory hydromagnetic Couette flow in a rotating system, *ZAMM* 80 (2000):429-432.
- [21] S.K. Ghosh. Effects of Hall current on MHD couette flow in a rotating system with arbitrary magnetic field, *Czech.J. Phys.* 52 (2002):51-63.
- [22] S.K. Ghosh, O.A. Beg, and M. Narahari. Hall effects on MHD flow in a rotating system with heat transfer characteristics, *Meccanica* 44 (2009):741-765.
- [23] G.S. Seth, R. Nandkeolyar, N. Mahto and S.K. Singh. MHD Couette flow in a rotating system in the presence of an inclined magnetic field, *Applied Mathematical Sciences* 3(59) (2009):2919-2932.
- [24] M. Guria, S. Das, R.N. Jana and S.K. Ghosh. Oscillatory Couette flow in the presence of an inclined magnetic field, *Meccanica* 44 (2009):555-564.

An Optimal Manner of distribution of drinking Water Using Heuristic Method

Abdullah Al-Hossain and Said Bourazza.

Abstract—The distribution of the drinking water by a minimum cost experiencing a growing interest in the world. Specially, the minimization of the electric cost absorbed by the pumps. A mathematical formulation is presented in this paper of a real method. We adapted our heuristics to optimize the cost of electric used in three locations in Jazan city, Saudi Arabia. The solutions obtained by our method give exact results in few seconds. which is an advantage for the practical application for real system as shown in the case of Jazan city.

Keywords— Combinatorial optimization, global optimization, Simulated Annealing method, water distribution systems.

I. INTRODUCTION

WE consider the model of combinatorial optimization problem, where the objective function to be minimized gives the cost of the electrical energy consumed in the operating system of the pumps used to provide consumptive water for a community.

We will use one variant of the simulated annealing (SA) as optimization method to solve this problem. The SA was first proposed by Kirkpatrick, Gelatt, and Vecchi [5] is an efficient method for finding the global optimal solution to multidimensional optimization problems.

SA is easy to implement and does not require much computer memory and coding.

SA as computational results show some conflicting results when SA is compared with other algorithms [7]. Ingber and Rosen [4] have proposed a very fast simulated annealing method that is efficient in its strategy and which statistically guarantees to find the global optima. Their results reveal that the method is orders of magnitude more efficient than a GA. On the other hand, Youssef, Sait, and Adiche [10] have performed a comparative study on SA, Tabu Search [3] and evolutionary algorithms by applying them to the same optimization problem. The benchmark problem used is the floor planning of very large scale integrated (VLSI) circuits,

which is a hard multi-criteria optimization problem. This study has shown that Tabu Search and evolutionary algorithms outperform SA. It therefore requires considerable tests of the algorithms to give sound conclusion. However, it is certain that attention is required in the area of choosing an optimal annealing schedule, studying the effect of algorithmic parameters on the performance of SA, and selecting the new solution vector efficiently [7].

There are many variants of simulated annealing algorithm found in relative literature [4], [1], [7]. The main structure is almost preserved and comprises the three following operators: a temperature cooling schedule T , a function neighbor for generating a perturbation and a state transition with an acceptance probability ρ .

A simulated annealing based approach was developed to obtain the least-cost design of a looped water distribution network by Cunha and Sousa [1]. Shieh, Richard, and Peralta [6] presented a simulation/optimization model using a hybrid method combining genetic algorithms and simulated annealing for optimizing an in-situ bioremediation system design.

The paper is organized as follows: In Section 2, we present the mathematical formulation of the problem that will be solved by SA method. At the end of this part we formulate three real problems. The SA algorithm used to solve problems is presented in Section 3. The corresponding numerical results are presented in section 4. Finally, our results are summarized and future work is described in the conclusion.

II. MODEL FORMULATION OF WATER SYSTEM PUMP OPERATIONS

Our work is based on the objective function described by El Mouatasim, Ellaia, and Al-Hossain [2] and Yin Luo Shouqi Yuan, Yue Tang, Jianping Yuan, and Jinfeng Zhang [9]. And taking account the specificity of Jazan city, the problem in three locations is as follows:

A. Swiss Problem:

$$\begin{aligned} \text{Min } Z = & 126.95 \sum_{i=1}^6 \alpha_i + 4186.7 \sum_{i=1}^6 \alpha_i^3 + \\ & + \frac{0.065}{0.3^5} \sum_{i=1}^6 \alpha_i \left(\sum_{k=i}^6 (300 \sum_{m=1}^k (0.67 \alpha_m))^2 \right). \end{aligned}$$

Subject to:

This work was full supported by Jazan University. Many thanks to Scientific Research Deanship, Jazan University and to the director of Jazan General Administration of Drinking Water for helping us and providing the data used in this work.

Abdullah Al-Hossain is associate professor, Department of Mathematics, Faculty of Science - Jazan University, P.B. 2097 Jazan - KSA. (Phone: 00966-73343495; e-mail: aalhossain@jazanu.edu.sa).

Said Bourazza is assistant professor in Department of mathematics, Faculty of science, Jazan University. (e-mail: sbourazza@jazanu.edu.sa).

$$\left\{ \begin{array}{l} 20(\sum_{i=1}^6 40\alpha_i - 200) \geq 300 \\ 20(\sum_{i=1}^6 40\alpha_i - 200) \leq 900 \\ \alpha_i \in \{0, 1\} \text{ for } i \in \{1, 2, 3, 4, 5, 6\} \end{array} \right.$$

B. Mahliya Problem

$$\begin{aligned} \text{Min } Z = & 397.79 \sum_{i=1}^6 \alpha_i + 21195.18 \sum_{i=1}^6 \alpha_i^3 + \\ & + \frac{0.065}{0.2^5} \sum_{i=1}^6 \alpha_i \left(\sum_{k=i}^6 (200 \sum_{m=1}^k (0.67\alpha_m))^2 \right). \end{aligned}$$

Subject to:

$$\left\{ \begin{array}{l} 20(\sum_{i=1}^6 40\alpha_i - 200) \geq 250 \\ 20(\sum_{i=1}^6 40\alpha_i - 200) \leq 700 \\ \alpha_i \in \{0, 1\} \text{ for } i \in \{1, 2, 3, 4, 5, 6\} \end{array} \right.$$

C. Chatee Problem

$$\begin{aligned} \text{Min } Z = & 423.18 \sum_{i=1}^3 \alpha_i + 2093.35 \sum_{i=1}^3 \alpha_i^3 + \\ & + \frac{0.065}{0.2^5} \sum_{i=1}^3 \alpha_i \left(\sum_{k=i}^3 (150 \sum_{m=1}^k (0.67\alpha_m))^2 \right). \end{aligned}$$

Subject to:

$$\left\{ \begin{array}{l} 20(\sum_{i=1}^6 40\alpha_i - 80) \geq 250 \\ 20(\sum_{i=1}^6 40\alpha_i - 80) \leq 900 \\ \alpha_i \in \{0, 1\} \text{ for } i \in \{1, 2, 3\} \end{array} \right.$$

III. SIMULATED ANNEALING

Simulated annealing algorithm begins by generating a single initial solution at random, which is evaluated by the cost function. In our case, a solution is an integer vector consisting of n parameters (number of pumps). Our variant of SA, implemented in this paper, uses the following:

```
Function{recuit} {X0 : is an initial vector chosen randomly}
{ X ← X0, Best X ← X0, T ← 10000,
While{ T > 1 }
{ m ← 0
Repeat
{ Y ← Neighbor (X)
dF ← F(Y) - F(X)
If (dF < 0) Then { X ← Y, best X ← Y }
Else if ( Random < e (-dF/T)) Then X ← Y
m ← m + 1
} Until (m == 100 )
T ← 0.99* T}}
```

For the Neighbor function, we look randomly for the closest solution with respect to the constraints.

IV. EXPERIMENTAL RESULTS

In this section, we propose the mathematical model established in section 2 with some reference to Jazan city conditions in Saudi Arabia. The aim is to minimize the cost of electrical energy consumed in three stations: Swiss district, Chatee District and Mahliya district.

The cost consumed in SAR, where 1 SAR= 3.75 \$.

For solving linear constraints combinatorial optimization problem in this application, we used Mathematica software, and the annealing simulated algorithm for global optimization under linear constraints. We compare Our approach programmed using Mathematica, with the Min function of the software Mathematica.

Note that the experiments performed on a workstation DELL Intel(R) Core™ i3 CPU processor 2.13GHz, 2GB RAM.

The results obtained are presented in this table:

	Mathematica Software		Simulated Annealing	
	The cost	Solution	The cost	Solution
Swiss Problem	393596	$\alpha_1=0.581649,$ $\alpha_2=0.793351,$ $\alpha_3= \alpha_4= \alpha_5=$ $\alpha_6=1$	731930	$\alpha_1=\alpha_2 =1$ $\alpha_3=\alpha_4=$ $=\alpha_5= \alpha_6=1$
Mahliya Problem	334174	$\alpha_1=0.55521,$ $\alpha_2=0.75729,$ $\alpha_3=\alpha_4=\alpha_5=$ $\alpha_6=1$	372132	$\alpha_1 = \alpha_2 =1$ $\alpha_3=0$ $\alpha_4=\alpha_5=\alpha_6=1$
Chatee Problem	16656.8	$\alpha_1=0.446138,$ $\alpha_2=0.928862,$ $\alpha_3=1.$	43572.4	$\alpha_1=\alpha_2=\alpha_3=1$

First, the results are very interesting because they give possible solutions and they are optimal regarding to the constraints. The values given by our solutions are exacts, since α_i takes 0 or 1 (the pump is OFF or ON respectively). However, the value given by the Min function of Mathematica gives a continuous values for α_i in the closed interval [0,1]. And in this last case, the system needs an expert to check if the value is close to zero and decide to switch off the pump or close to 1 and switch ON the pump. Which needs an extra work for the operator to “decode” the solution obtained by the Min function of Mathematica?

Next, we note that in two of the three problems the pumps must necessarily work to satisfy the needs of water users. This gives only 4 hours of rest for the pumps, which might affect their lifespan.

V. CONCLUSION

The numerical computation of the solution, obtained by our variant of the simulated annealing method with appropriate coding for minimizing the cost of the electric pump system, shows that the proposed method gives exact solutions and

works better than the minimization function used by the Mathematica software.

Our work is part of a new policy of the General Administration of distribution of drinking water. Which consist of computerizes the system and builds a geographic information system capable of providing real information in real time. This is why we opted to use the heuristic method for solving these problems.

Our variant of SA can be used as a good method to optimize the management of water distribution by pumping system operations and get results in a short period of time.

REFERENCES

- [1] Cunha, M. C., and Sousa, J., "Water distribution network design optimisation: simulated annealing approach." *J. Water Resour. Plan. Manage.* 125(4), 1999, pp. 215-221.
- [2] El Mouatasim, A., Ellaia, R., Al-Hossain, A., "A continuous approach to combinatorial optimization: Application of water system pumps operations". *Optim. Lett.*, 2012, pp. 177-198.
- [3] Glover, F. (1986) "Future paths for integer programming and links to artificial intelligence". *Comp. and Operations Res.*, (5), 533-549.
- [4] Ingber, L. , Rosen, B., "Genetic algorithms and very fast simulated reannealing: A comparison". *Math. Comput. Modelling* 16, 11, 1992, pp. 87-100
- [5] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M.P., "Optimisation by Simulated Annealing." *Science*, 220, 1983, pp. 671-680.
- [6] Shieh, H. J., Richard C. and Peralta, M., "Optimal In Situ Bioremediation Design by Hybrid Genetic Algorithm-Simulated Annealing." *J. Water Resour. Plan. Manage.* 131(1), 2005, pp. 67-78.
- [7] Suman, B., and Kumar, P., "A survey of simulated annealing as a tool for single and multiobjective optimization." *Journal of Operational Research Society*, 57, 2006, pp. 1143-1160.
- [8] Viessman, W., Hammer, M.J., "Water Supply and Pollution Control". Harper Collins College Publishers, Inc. 5th Ed. 1993.
- [9] Yin Luo, Shouqi Yuan, Yue Tang, Jianping Yuan, and Jinfeng Zhang, "Modeling Optimal Scheduling for Pumping System to Minimize Operation Cost and Enhance Operation Reliability". *Hindawi Publishing Corporation Journal of Applied Mathematics* Volume 2012, Article ID 370502, 19 pages.
- [10] Youssef, H., Sait, S. M., and Adiche, H., "Evolutionary Algorithm, Simulated Annealing and Tabu Search: A comparative study." *Eng. Appl. Artif. Intel.* 14, 2003, pp. 167-181.

Abdullah Yahia Al-Hossain, received his M.Sc (in mathematical statistics) from Iowa State University, USA in 1992 and PhD (in applied statistics) from Lancaster University, UK in 2001. He is currently an Associate Professor, in Mathematics Department, Faculty of Science, Jazan University. He authored and co-authored about 25 conference and journal papers. He served as the chairman the Head of Business Department and the Vice-Dean, Faculty of Business School at KJU. He was also The Dean of Faculty of Engineering & the Acting-Dean for Computer Science Faculty, the Dean, Faculty of science & Acting-Dean of Academic Development Deanship at Jazan University. He is currently the Head of Mathematics Department. and the Dean of Faculty of science at Jazan University.

Certain Integrable Cases in Dynamics of a Multi-Dimensional Rigid Body in a Nonconservative Field

Maxim V. Shamolin

Abstract—This paper is a survey of integrable cases in dynamics of a five-dimensional rigid body under the action of a nonconservative force field. We review both new results and results obtained earlier. Problems examined are described by dynamical systems with so-called variable dissipation with zero mean.

The problem of the search for complete sets of transcendental first integrals of systems with dissipation is quite actual; a large number of works are devoted to it. We introduce a new class of dynamical systems that have a periodic coordinate. Due to the existence of a nontrivial symmetry groups of such systems, we can prove that these systems possess variable dissipation with zero mean, which means that on the average for a period with respect to the periodic coordinate, the dissipation in the system is equal to zero, although in various domains of the phase space, either the energy pumping or dissipation can occur. Based on facts obtained, we analyze dynamical systems that appear in dynamics of a five-dimensional rigid body and obtain a series of new cases of complete integrability of the equations of motion in transcendental functions that can be expressed through a finite combination of elementary functions.

Index Terms—Case of integrability, dynamic part of motion equations, multidimensional rigid body.

I. INTRODUCTION

THIS This paper is a survey of integrable cases in dynamics of a five-dimensional rigid body under the action of a nonconservative force field. We review both new results and results obtained earlier. Problems examined are described by dynamical systems with so-called variable dissipation with zero mean.

We study nonconservative systems for which usual methods of the study of Hamiltonian system is inapplicable. Thus, for such systems, we must directly integrate the main equation of dynamics (see also [1], [2], [3], [4], [5], [6]).

We generalize previously known cases and obtain new cases of the complete integrability in transcendental functions of the equation of dynamics of a five-dimensional rigid body in a nonconservative force field

Of course, in the general case, the construction of a theory of integration of nonconservative systems (even of low dimension) is a quite difficult task. In a number of cases, where the systems considered have additional symmetries, we succeed in finding first integrals through finite combinations of elementary functions [6], [7], [8], [9].

Maxim V. Shamolin is with the Institute of Mechanics, Lomonosov Moscow State University, Moscow, 119192, Russian Federation; e-mail: shamolin@rambler.ru, shamolin@imec.msu.ru (see also <http://shamolin2.imec.msu.ru>).

In basic part we recall general aspects of the dynamics of a free multi-dimensional rigid body: the notion of the tensor of angular velocity of the body, the joint dynamical equations of motion on the direct product $\mathbf{R}^n \times \text{so}(n)$, and the Euler and Rivals formulas in the multi-dimensional case.

We also consider the tensor of inertia of a five-dimensional ($5D$ –) rigid body. In this work, we study one of two possible cases in which there exists two relations between the principal moments of inertia:

(i) there are four equal principal moments of inertia ($I_2 = I_3 = I_4 = I_5$).

Furthermore, we systematize results on the study of equations of motion of a five-dimensional ($5D$ –) rigid body in a nonconservative force field for the case (i). The form of these equations is taken from the dynamics of realistic rigid bodies of lesser dimension that interact with a resisting medium by laws of jet flow when the is influence by a nonconservative tracing force. Under the action of this force, the following two cases are possible. In this case, the velocity of some characteristic point of the body remains constant, which means that the system possesses a nonintegrable servo-constraint (see also [10], [11]).

The results relate to the case where all interaction of the medium with the body part is concentrated on a part of the surface of the body, which has the form of a four-dimensional disk, and the action of the force is concentrated in the direction perpendicular to this disk. These results are systematized and are preserved in the invariant form. Moreover, we introduce an extra dependence of the moment of the nonconservative force on the angular velocity. This dependence can be further extended to cases of the motion in spaces of higher dimension.

Many results of this paper were regularly presented on scientific seminars, including the seminar Actual problems of geometry and mechanics named after Prof. V. V. Trofimov under the supervision of D. V. Georgievskii and M. V. Shamolin.

II. GENERAL DISCOURSE

A. Cases of dynamical symmetry of a five-dimensional body

Let a five-dimensional rigid body Θ of mass m with smooth four-dimensional boundary $\partial\Theta$ be under the influence of a nonconservative force field this can be interpreted as a motion of the body in a resisting medium that fill up five-dimensional domain of Euclidean space \mathbf{E}^5 . We assume that the body

is dynamically symmetric. If the body has two independent principal moments of inertia, then in some coordinate system $Dx_1x_2x_3x_4x_5$ attached to the body, the operator of inertia has the form

$$\text{diag}\{I_1, I_2, I_2, I_2, I_2\}, \tag{1}$$

or the form

$$\text{diag}\{I_1, I_1, I_3, I_3, I_3\}. \tag{2}$$

In the first case, the body is dynamically symmetric in the hyperplane $Dx_2x_3x_4x_5$.

B. Dynamics on $so(5)$ and \mathbf{R}^5

The configuration space of a free, n -dimensional rigid body is the direct product

$$\mathbf{R}^n \times SO(n) \tag{3}$$

of the space \mathbf{R}^n , which define the coordinates of the center of mass of the body, and the rotation group $SO(n)$, which define rotations of the body about its center of mass and has dimension

$$n + \frac{n(n-1)}{2} = \frac{n(n+1)}{2}.$$

Therefore, the dynamical part of equations of motion has the same dimension, whereas the dimension of the phase space is equal to

$$n(n+1).$$

In particular, if Ω is the tensor of angular velocity of a five-dimensional rigid body (it is a second-rank tensor, see [12], [13], [14], [15], [16]), $\Omega \in so(5)$, then the part of dynamical equations of motion corresponding to the Lie algebra $so(5)$ has the following form (see [17], [18]):

$$\dot{\Omega}\Lambda + \Lambda\dot{\Omega} + [\Omega, \Omega\Lambda + \Lambda\Omega] = M, \tag{4}$$

where

$$\begin{aligned} \Lambda &= \text{diag}\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}, \\ \lambda_1 &= \frac{-I_1 + I_2 + I_3 + I_4 + I_5}{2}, \\ \lambda_2 &= \frac{I_1 - I_2 + I_3 + I_4 + I_5}{2}, \\ \lambda_3 &= \frac{I_1 + I_2 - I_3 + I_4 + I_5}{2}, \\ \lambda_4 &= \frac{I_1 + I_2 + I_3 - I_4 + I_5}{2}, \\ \lambda_5 &= \frac{I_1 + I_2 + I_3 + I_4 - I_5}{2}, \end{aligned} \tag{5}$$

$M = M_F$ is the natural projection of the moment of external forces \mathbf{F} acting to the body in \mathbf{R}^5 on the natural coordinates of the Lie algebra $so(5)$, and $[\]$ is the commutator in $so(5)$. The skew-symmetric matrix corresponding to this second-rank tensor $\Omega \in so(5)$ we represent in the form

$$\begin{pmatrix} 0 & -\omega_{10} & \omega_9 & -\omega_7 & \omega_4 \\ \omega_{10} & 0 & -\omega_8 & \omega_6 & -\omega_3 \\ -\omega_9 & \omega_8 & 0 & -\omega_5 & \omega_2 \\ \omega_7 & -\omega_6 & \omega_5 & 0 & -\omega_1 \\ -\omega_4 & \omega_3 & -\omega_2 & \omega_1 & 0 \end{pmatrix}, \tag{6}$$

where $\omega_1, \omega_2, \dots, \omega_{10}$ are the components of the tensor of angular velocity corresponding to the projections on the coordinates of the Lie algebra $so(5)$.

Obviously, the following relations hold:

$$\lambda_i - \lambda_j = I_j - I_i \tag{7}$$

for any $i, j = 1, \dots, 5$.

For the calculation of the moment of an external force acting to the body, we need to construct the mapping

$$\mathbf{R}^5 \times \mathbf{R}^5 \longrightarrow so(5), \tag{8}$$

than maps a pair of vectors

$$(\mathbf{DN}, \mathbf{F}) \in \mathbf{R}^5 \times \mathbf{R}^5 \tag{9}$$

from $\mathbf{R}^5 \times \mathbf{R}^5$ to an element of the Lie algebra $so(5)$, where

$$\begin{aligned} \mathbf{DN} &= \{0, x_{2N}, x_{3N}, x_{4N}, x_{5N}\}, \\ \mathbf{F} &= \{F_1, F_2, F_3, F_4, F_5\}, \end{aligned} \tag{10}$$

and \mathbf{F} is an external force acting to the body. For this end, we construct the following auxiliary matrix

$$\begin{pmatrix} 0 & x_{2N} & x_{3N} & x_{4N} & x_{5N} \\ F_1 & F_2 & F_3 & F_4 & F_5 \end{pmatrix}. \tag{11}$$

Then the right-hand side of system (4) takes the form

$$\begin{aligned} M &= \{x_{4N}F_5 - x_{5N}F_4, x_{5N}F_3 - x_{3N}F_5, \\ &x_{2N}F_5 - x_{5N}F_2, x_{5N}F_1, x_{3N}F_4 - x_{4N}F_3, \\ &x_{4N}F_2 - x_{2N}F_4, -x_{4N}F_1, x_{2N}F_3 - x_{3N}F_2, \\ &x_{3N}F_1, -x_{2N}F_1\}. \end{aligned} \tag{12}$$

Dynamical systems studied in the following, generally speaking, are not conservative; they are dynamical systems with variable dissipation with zero mean (see [12]). We need to examine by direct methods a part of the main system of dynamical equations, namely, the Newton equation, which plays the role of the equation of motion of the center of mass, i.e., the part of the dynamical equations corresponding to the space \mathbf{R}^5 :

$$m\mathbf{w}_C = \mathbf{F}, \tag{13}$$

where \mathbf{w}_C is the acceleration of the center of mass C of the body and m is its mass. Moreover, due to the higher-dimensional Rivals formula (it can be obtained by the operator method) we have the following relations:

$$\mathbf{w}_C = \mathbf{w}_D + \Omega^2\mathbf{DC} + E\mathbf{DC}, \quad \mathbf{w}_D = \dot{\mathbf{v}}_D + \Omega\mathbf{v}_D, \quad E = \dot{\Omega}, \tag{14}$$

where \mathbf{w}_D is the acceleration of the point D , \mathbf{F} is the external force acting on the body (in our case, $\mathbf{F} = \mathbf{S}$), and E is the tensor of angular acceleration (second-rank tensor).

So, the system of equations (4) and (13) of fifteen order on the manifold $\mathbf{R}^5 \times so(5)$ is a closed system of dynamical equations of the motion of a free five-dimensional rigid body under the action of an external force \mathbf{F} . This system have been separated from the kinematic part of the equations of motion on the manifold (3) and can be examined independently.

III. GENERAL PROBLEM ON THE MOTION UNDER A TRACING FORCE

Consider a motion of a homogeneous, dynamically symmetric (case (1)), rigid body with front end face (a four-dimensional disk interacting with a medium that fill the five-dimensional space) in the field of a resistance force \mathbf{S} under the quasi-stationarity conditions.

Let $(v, \alpha, \beta_1, \beta_2, \beta_3)$ be the (generalized) spherical coordinates of the velocity vector of the center of the four-dimensional disk lying on the axis of symmetry of the body,

$$\Omega = \begin{pmatrix} 0 & -\omega_{10} & \omega_9 & -\omega_7 & \omega_4 \\ \omega_{10} & 0 & -\omega_8 & \omega_6 & -\omega_3 \\ -\omega_9 & \omega_8 & 0 & -\omega_5 & \omega_2 \\ \omega_7 & -\omega_6 & \omega_5 & 0 & -\omega_1 \\ -\omega_4 & \omega_3 & -\omega_2 & \omega_1 & 0 \end{pmatrix}$$

be the tensor of angular velocity of the body, $Dx_1x_2x_3x_4x_5$ be the coordinate system attached to the body such that the axis of symmetry CD coincides with the axis Dx_1 (recall that C is the center of mass), and the axes Dx_2, Dx_3, Dx_4, Dx_5 lie in the hyperplane of the disk, and $I_1, I_2, I_3 = I_2, I_4 = I_2, I_5 = I_2, m$ are characteristics of inertia and mass.

We adopt the following expansions in the projections to the axes of the coordinate system $Dx_1x_2x_3x_4x_5$:

$$\mathbf{DC} = \{-\sigma, 0, 0, 0, 0\},$$

$$\mathbf{v}_D = \{v \cos \alpha, v \sin \alpha \cos \beta_1, v \sin \alpha \sin \beta_1 \cos \beta_2,$$

$$v \sin \alpha \sin \beta_1 \sin \beta_2 \cos \beta_3, v \sin \alpha \sin \beta_1 \sin \beta_2 \sin \beta_3\}. \quad (15)$$

In the case (1) we additionally have the expansion for the function of the influence of the medium on the five-dimensional body:

$$\mathbf{S} = \{-S, 0, 0, 0, 0\}, \quad (16)$$

i.e., in this case $\mathbf{F} = \mathbf{S}$.

Then the part of the dynamical equations of motion (including the analytic Chaplygin functions; see below) that describes the motion of the center of mass and corresponds to the space \mathbf{R}^5 , in which tangent forces of the influence of the medium on the four-dimensional disk vanish, takes the form

$$\begin{aligned} & \dot{v} \cos \alpha - \dot{\alpha} v \sin \alpha - \omega_{10} v \sin \alpha \cos \beta_1 + \omega_9 v \sin \alpha \sin \beta_1 \cos \beta_2 - \\ & - \omega_7 v \sin \alpha \sin \beta_1 \sin \beta_2 \cos \beta_3 + \omega_4 v \sin \alpha \sin \beta_1 \sin \beta_2 \sin \beta_3 + \\ & + \sigma(\omega_{10}^2 + \omega_9^2 + \omega_7^2 + \omega_4^2) = -\frac{S}{m}, \quad (17) \end{aligned}$$

$$\begin{aligned} & \dot{v} \sin \alpha \cos \beta_1 + \dot{\alpha} v \cos \alpha \cos \beta_1 - \dot{\beta}_1 v \sin \alpha \sin \beta_1 + \\ & + \omega_{10} v \cos \alpha - \omega_8 v \sin \alpha \sin \beta_1 \cos \beta_2 + \\ & + \omega_6 v \sin \alpha \sin \beta_1 \sin \beta_2 \cos \beta_3 - \\ & - \omega_3 v \sin \alpha \sin \beta_1 \sin \beta_2 \sin \beta_3 - \\ & - \sigma(\omega_9 \omega_8 + \omega_6 \omega_7 + \omega_3 \omega_4) - \sigma \dot{\omega}_{10} = 0, \quad (18) \end{aligned}$$

$$\begin{aligned} & \dot{v} \sin \alpha \sin \beta_1 \cos \beta_2 + \dot{\alpha} v \cos \alpha \sin \beta_1 \cos \beta_2 + \\ & + \dot{\beta}_1 v \sin \alpha \cos \beta_1 \cos \beta_2 - \end{aligned}$$

$$\begin{aligned} & - \dot{\beta}_2 v \sin \alpha \sin \beta_1 \sin \beta_2 - \omega_9 v \cos \alpha + \omega_8 v \sin \alpha \cos \beta_1 - \\ & - \omega_5 v \sin \alpha \sin \beta_1 \sin \beta_2 \cos \beta_3 + \\ & + \omega_2 v \sin \alpha \sin \beta_1 \sin \beta_2 \sin \beta_3 - \\ & - \sigma(\omega_8 \omega_{10} - \omega_5 \omega_7 - \omega_2 \omega_4) + \sigma \dot{\omega}_9 = 0, \quad (19) \end{aligned}$$

$$\begin{aligned} & \dot{v} \sin \alpha \sin \beta_1 \sin \beta_2 \cos \beta_3 + \dot{\alpha} v \cos \alpha \sin \beta_1 \sin \beta_2 \cos \beta_3 + \\ & + \dot{\beta}_1 v \sin \alpha \cos \beta_1 \sin \beta_2 \cos \beta_3 + \\ & + \dot{\beta}_2 v \sin \alpha \sin \beta_1 \cos \beta_2 \cos \beta_3 - \\ & - \dot{\beta}_3 v \sin \alpha \sin \beta_1 \sin \beta_2 \sin \beta_3 + \omega_7 v \cos \alpha - \omega_6 v \sin \alpha \cos \beta_1 + \\ & + \omega_5 v \sin \alpha \sin \beta_1 \cos \beta_2 - \omega_1 v \sin \alpha \sin \beta_1 \sin \beta_2 \sin \beta_3 + \\ & + \sigma(\omega_6 \omega_{10} + \omega_5 \omega_9 - \omega_1 \omega_4) - \sigma \dot{\omega}_7 = 0, \quad (20) \end{aligned}$$

$$\begin{aligned} & \dot{v} \sin \alpha \sin \beta_1 \sin \beta_2 \sin \beta_3 + \dot{\alpha} v \cos \alpha \sin \beta_1 \sin \beta_2 \sin \beta_3 + \\ & + \dot{\beta}_1 v \sin \alpha \cos \beta_1 \sin \beta_2 \sin \beta_3 + \\ & + \dot{\beta}_2 v \sin \alpha \sin \beta_1 \cos \beta_2 \sin \beta_3 + \\ & + \dot{\beta}_3 v \sin \alpha \sin \beta_1 \sin \beta_2 \cos \beta_3 - \omega_4 v \cos \alpha + \omega_3 v \sin \alpha \cos \beta_1 - \\ & - \omega_2 v \sin \alpha \sin \beta_1 \cos \beta_2 + \omega_1 v \sin \alpha \sin \beta_1 \sin \beta_2 \cos \beta_3 - \\ & - \sigma(\omega_3 \omega_{10} + \omega_2 \omega_9 + \omega_1 \omega_7) + \sigma \dot{\omega}_4 = 0, \quad (21) \end{aligned}$$

where

$$S = s(\alpha)v^2, \quad \sigma = CD, \quad v > 0. \quad (22)$$

Further, the auxiliary matrix (11) for the calculation of the moment of the resistance force has the form

$$\begin{pmatrix} 0 & x_{2N} & x_{3N} & x_{4N} & x_{5N} \\ -S & 0 & 0 & 0 & 0 \end{pmatrix}, \quad (23)$$

then the part of the dynamical equations of motion that describes the motion of the body about the center of mass and corresponds to the Lie algebra $\mathfrak{so}(5)$, becomes

$$(\lambda_4 + \lambda_5)\dot{\omega}_1 + (\lambda_4 - \lambda_5)(\omega_4\omega_7 + \omega_3\omega_6 + \omega_2\omega_5) = 0, \quad (24)$$

$$(\lambda_3 + \lambda_5)\dot{\omega}_2 + (\lambda_5 - \lambda_3)(\omega_1\omega_5 - \omega_3\omega_8 - \omega_4\omega_9) = 0, \quad (25)$$

$$(\lambda_2 + \lambda_5)\dot{\omega}_3 + (\lambda_2 - \lambda_5)(\omega_4\omega_{10} - \omega_2\omega_8 - \omega_1\omega_6) = 0, \quad (26)$$

$$\begin{aligned} & (\lambda_1 + \lambda_5)\dot{\omega}_4 + (\lambda_5 - \lambda_1)(\omega_3\omega_{10} + \omega_2\omega_9 + \omega_1\omega_7) = \\ & = -x_{5N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) s(\alpha)v^2, \quad (27) \end{aligned}$$

$$(\lambda_3 + \lambda_4)\dot{\omega}_5 + (\lambda_3 - \lambda_4)(\omega_7\omega_9 + \omega_6\omega_8 + \omega_1\omega_2) = 0, \quad (28)$$

$$(\lambda_2 + \lambda_4)\dot{\omega}_6 + (\lambda_4 - \lambda_2)(\omega_5\omega_8 - \omega_7\omega_{10} - \omega_1\omega_3) = 0, \quad (29)$$

$$\begin{aligned} & (\lambda_1 + \lambda_4)\dot{\omega}_7 + (\lambda_1 - \lambda_4)(\omega_1\omega_4 - \omega_6\omega_{10} - \omega_5\omega_9) = \\ & = x_{4N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) s(\alpha)v^2, \quad (30) \end{aligned}$$

$$(\lambda_2 + \lambda_3)\dot{\omega}_8 + (\lambda_2 - \lambda_3)(\omega_9\omega_{10} + \omega_5\omega_6 + \omega_2\omega_3) = 0, \quad (31)$$

$$\begin{aligned} & (\lambda_1 + \lambda_3)\dot{\omega}_9 + (\lambda_3 - \lambda_1)(\omega_8\omega_{10} - \omega_5\omega_7 - \omega_2\omega_4) = \\ & = -x_{3N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) s(\alpha)v^2, \quad (32) \end{aligned}$$

$$(\lambda_1 + \lambda_2)\dot{\omega}_{10} + (\lambda_1 - \lambda_2)(\omega_8\omega_9 + \omega_6\omega_7 + \omega_3\omega_4) =$$

$$= x_{2N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) s(\alpha)v^2. \quad (33)$$

Thus, the phase space of system (17)–(21), (24)–(33) of fifteenth order is the direct product of the five-dimensional manifold and the Lie algebra $so(5)$:

$$\mathbf{R}^1 \times \mathbf{S}^4 \times so(5). \quad (34)$$

We note that system (17)–(21), (24)–(33), due to the existing dynamical symmetry

$$I_2 = I_3 = I_4 = I_5, \quad (35)$$

possesses cyclic first integrals

$$\omega_1 \equiv \omega_1^0, \omega_2 \equiv \omega_2^0, \omega_3 \equiv \omega_3^0, \omega_5 \equiv \omega_5^0, \omega_6 \equiv \omega_6^0, \omega_8 \equiv \omega_8^0. \quad (36)$$

In the sequel, we consider the dynamics of the system on zero levels:

$$\omega_1^0 = \omega_2^0 = \omega_3^0 = \omega_5^0 = \omega_6^0 = \omega_8^0 = 0. \quad (37)$$

If one considers a more general problem on the motion of a body under a tracing force \mathbf{T} that lies on the straight line $CD = Dx_1$ and provides the fulfillment of the relation

$$v \equiv \text{const} \quad (38)$$

throughout the motion, then instead of F_1 system (17)–(21), (24)–(33) contains

$$T - s(\alpha)v^2, \sigma = DC. \quad (39)$$

Choosing the value T of the tracing force appropriately, one can achieve the equality (38) throughout the motion. Indeed, expressing T due to system (17)–(21), (24)–(33), we obtain for $\cos \alpha \neq 0$ the relation

$$T = T_v(\alpha, \beta_1, \beta_2, \beta_3, \Omega) = m\sigma(\omega_4^2 + \omega_7^2 + \omega_9^2 + \omega_{10}^2) + s(\alpha)v^2 \left[1 - \frac{m\sigma \sin \alpha}{3I_2 \cos \alpha} \Gamma_v \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \right], \quad (40)$$

where

$$\begin{aligned} \Gamma_v \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) = & x_{5N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \sin \beta_1 \sin \beta_2 \sin \beta_3 + \\ & + x_{4N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \sin \beta_1 \sin \beta_2 \cos \beta_3 + \\ & + x_{3N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \sin \beta_1 \cos \beta_2 + \\ & + x_{2N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \cos \beta_1; \end{aligned} \quad (41)$$

here we used conditions (36)–(38).

This procedure can be interpreted in two ways. First, we have transformed the system using the tracing force (control) that provides the consideration of the class (38) of motions interesting for us. Second, we can treat this as an order-reduction procedure. Indeed, system (17)–(21), (24)–(33) generates the following independent system of eighth order:

$$\dot{\alpha}v \cos \alpha \cos \beta_1 - \dot{\beta}_1 v \sin \alpha \sin \beta_1 +$$

$$+ \omega_{10}v \cos \alpha - \sigma \omega_{10} = 0, \quad (42)$$

$$\begin{aligned} & \dot{\alpha}v \cos \alpha \sin \beta_1 \cos \beta_2 + \dot{\beta}_1 v \sin \alpha \cos \beta_1 \cos \beta_2 - \\ & - \dot{\beta}_2 v \sin \alpha \sin \beta_1 \sin \beta_2 - \omega_9 v \cos \alpha + \sigma \omega_9 = 0, \end{aligned} \quad (43)$$

$$\begin{aligned} & \dot{\alpha}v \cos \alpha \sin \beta_1 \sin \beta_2 \cos \beta_3 + \dot{\beta}_1 v \sin \alpha \cos \beta_1 \sin \beta_2 \cos \beta_3 + \\ & + \dot{\beta}_2 v \sin \alpha \sin \beta_1 \cos \beta_2 \cos \beta_3 - \\ & - \dot{\beta}_3 v \sin \alpha \sin \beta_1 \sin \beta_2 \sin \beta_3 + \omega_7 v \cos \alpha - \sigma \omega_7 = 0, \end{aligned} \quad (44)$$

$$\begin{aligned} & \dot{\alpha}v \cos \alpha \sin \beta_1 \sin \beta_2 \sin \beta_3 + \dot{\beta}_1 v \sin \alpha \cos \beta_1 \sin \beta_2 \sin \beta_3 + \\ & + \dot{\beta}_2 v \sin \alpha \sin \beta_1 \cos \beta_2 \sin \beta_3 + \\ & + \dot{\beta}_3 v \sin \alpha \sin \beta_1 \sin \beta_2 \cos \beta_3 - \omega_4 v \cos \alpha + \sigma \omega_4 = 0, \end{aligned} \quad (45)$$

$$3I_2 \dot{\omega}_4 = -x_{5N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) s(\alpha)v^2, \quad (46)$$

$$3I_2 \dot{\omega}_7 = x_{4N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) s(\alpha)v^2, \quad (47)$$

$$3I_2 \dot{\omega}_9 = -x_{3N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) s(\alpha)v^2, \quad (48)$$

$$3I_2 \dot{\omega}_{10} = x_{2N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) s(\alpha)v^2, \quad (49)$$

which, in addition to the permanent parameters specific above, contains the parameter v .

System (42)–(49) is equivalent to the system

$$\begin{aligned} & \dot{\alpha}v \cos \alpha + v \cos \alpha \{ \omega_{10} \cos \beta_1 + \\ & + [(\omega_7 \cos \beta_3 - \omega_4 \sin \beta_3) \sin \beta_2 - \omega_9 \cos \beta_2] \sin \beta_1 \} + \\ & + \sigma \{ -\omega_{10} \cos \beta_1 + [\dot{\omega}_9 \cos \beta_2 - \\ & - (\dot{\omega}_7 \cos \beta_3 - \dot{\omega}_4 \sin \beta_3) \sin \beta_2] \sin \beta_1 \} = 0, \end{aligned} \quad (50)$$

$$\begin{aligned} & \dot{\beta}_1 v \sin \alpha + v \cos \alpha \{ [(\omega_7 \cos \beta_3 - \\ & - \omega_4 \sin \beta_3) \sin \beta_2 - \omega_9 \cos \beta_2] \cos \beta_1 - \omega_{10} \sin \beta_1 \} + \\ & + \sigma \{ [\dot{\omega}_9 \cos \beta_2 - (\dot{\omega}_7 \cos \beta_3 - \\ & - \dot{\omega}_4 \sin \beta_3) \sin \beta_2] \cos \beta_1 + \omega_{10} \sin \beta_1 \} = 0, \end{aligned} \quad (51)$$

$$\begin{aligned} & \dot{\beta}_2 v \sin \alpha \sin \beta_1 + v \cos \alpha \{ [\omega_7 \cos \beta_3 - \\ & - \omega_4 \sin \beta_3] \cos \beta_2 + \omega_9 \sin \beta_2 \} + \\ & + \sigma \{ -[\dot{\omega}_7 \cos \beta_3 - \dot{\omega}_4 \sin \beta_3] \cos \beta_2 - \dot{\omega}_9 \sin \beta_2 \} = 0, \end{aligned} \quad (52)$$

$$\begin{aligned} & \dot{\beta}_3 v \sin \alpha \sin \beta_1 \sin \beta_2 + v \cos \alpha \{ -\omega_4 \cos \beta_3 - \omega_7 \sin \beta_3 \} + \\ & + \sigma \{ \dot{\omega}_4 \cos \beta_3 + \dot{\omega}_7 \sin \beta_3 \} = 0, \end{aligned} \quad (53)$$

$$\dot{\omega}_4 = -\frac{v^2}{3I_2} x_{5N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) s(\alpha), \quad (54)$$

$$\dot{\omega}_7 = \frac{v^2}{3I_2} x_{4N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) s(\alpha), \quad (55)$$

$$\dot{\omega}_9 = -\frac{v^2}{3I_2} x_{3N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) s(\alpha), \quad (56)$$

$$\dot{\omega}_{10} = \frac{v^2}{3I_2} x_{2N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) s(\alpha). \quad (57)$$

Introduce the new quasi-velocities. For this, we transform $\omega_4, \omega_7, \omega_9, \omega_{10}$ by three rotations:

$$\begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{pmatrix} = T_{3,4}(-\beta_1) \circ T_{2,3}(-\beta_2) \circ T_{1,2}(-\beta_3) \begin{pmatrix} \omega_4 \\ \omega_7 \\ \omega_9 \\ \omega_{10} \end{pmatrix}, \quad (58)$$

where

$$T_{3,4}(\beta) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos \beta & -\sin \beta \\ 0 & 0 & \sin \beta & \cos \beta \end{pmatrix},$$

$$T_{2,3}(\beta) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos \beta & -\sin \beta & 0 \\ 0 & \sin \beta & \cos \beta & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$T_{1,2}(\beta) = \begin{pmatrix} \cos \beta & -\sin \beta & 0 & 0 \\ \sin \beta & \cos \beta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Therefore, the following relations hold:

$$\begin{aligned} z_1 &= \omega_4 \cos \beta_3 + \omega_7 \sin \beta_3, \\ z_2 &= (\omega_7 \cos \beta_3 - \omega_4 \sin \beta_3) \cos \beta_2 + \omega_9 \sin \beta_2, \\ z_3 &= [(-\omega_7 \cos \beta_3 + \omega_4 \sin \beta_3) \sin \beta_2 + \\ &+ \omega_9 \cos \beta_2] \cos \beta_1 + \omega_{10} \sin \beta_1, \\ z_4 &= [(\omega_7 \cos \beta_3 - \omega_4 \sin \beta_3) \sin \beta_2 - \\ &- \omega_9 \cos \beta_2] \sin \beta_1 + \omega_{10} \cos \beta_1. \end{aligned} \quad (59)$$

As we see from (50)–(57), we cannot solve the system with respect to $\dot{\alpha}, \dot{\beta}_1, \dot{\beta}_2, \dot{\beta}_3$ on the manifold

$$\begin{aligned} O_1 &= \{(\alpha, \beta_1, \beta_2, \beta_3, \omega_4, \omega_7, \omega_9, \omega_{10}) \in \mathbf{R}^8 : \\ \alpha &= \frac{\pi}{2}k, \beta_1 = \pi l_1, \beta_2 = \pi l_2, k, l_1, l_2 \in \mathbf{Z}\}. \end{aligned} \quad (60)$$

Therefore, on the manifold (60) the uniqueness theorem formally is violated. Moreover, for even k and any l_1, l_2 , an indeterminate form appears due to the degeneration of the spherical coordinates $(v, \alpha, \beta_1, \beta_2, \beta_3)$. For odd k , the uniqueness theorem is obviously violated since the first equation (50) degenerates.

This implies that system (50)–(57) outside (and only outside) the manifold (60) is equivalent to the system

$$\dot{\alpha} = -z_4 + \frac{\sigma v s(\alpha)}{3I_2 \cos \alpha} \Gamma_v \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right), \quad (61)$$

$$\begin{aligned} \dot{z}_4 &= \frac{v^2}{3I_2} s(\alpha) \Gamma_v \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) - \\ &- (z_1^2 + z_2^2 + z_3^2) \frac{\cos \alpha}{\sin \alpha} + \\ &+ \frac{\sigma v s(\alpha)}{3I_2 \sin \alpha} \left\{ -z_3 \Delta_{v,1} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) + \right. \\ &+ z_2 \Delta_{v,2} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) - \\ &\left. - z_1 \Delta_{v,3} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \right\}, \end{aligned} \quad (62)$$

$$\begin{aligned} \dot{z}_3 &= z_3 z_4 \frac{\cos \alpha}{\sin \alpha} + (z_1^2 + z_2^2) \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1} + \\ &+ \frac{\sigma v s(\alpha)}{3I_2 \sin \alpha} \left\{ z_4 \Delta_{v,1} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) - \right. \\ &- z_2 \Delta_{v,2} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \frac{\cos \beta_1}{\sin \beta_1} + \\ &\left. + z_1 \Delta_{v,3} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \frac{\cos \beta_1}{\sin \beta_1} \right\} - \\ &- \frac{v^2}{3I_2} s(\alpha) \Delta_{v,1} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right), \end{aligned} \quad (63)$$

$$\begin{aligned} \dot{z}_2 &= z_2 z_4 \frac{\cos \alpha}{\sin \alpha} - z_2 z_3 \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1} - \\ &- z_1^2 \frac{\cos \alpha}{\sin \alpha} \frac{1}{\sin \beta_1} \frac{\cos \beta_2}{\sin \beta_2} + \\ &+ \frac{\sigma v s(\alpha)}{3I_2 \sin \alpha} \Delta_{v,2} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \left\{ -z_4 + z_3 \frac{\cos \beta_1}{\sin \beta_1} \right\} + \\ &+ \frac{\sigma v s(\alpha)}{3I_2 \sin \alpha} \Delta_{v,3} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \left\{ -z_1 \frac{1}{\sin \beta_1} \frac{\cos \beta_2}{\sin \beta_2} \right\} + \\ &+ \frac{v^2}{3I_2} s(\alpha) \Delta_{v,2} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right), \end{aligned} \quad (64)$$

$$\begin{aligned} \dot{z}_1 &= z_1 z_4 \frac{\cos \alpha}{\sin \alpha} - z_1 z_3 \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1} + \\ &+ z_1 z_2 \frac{\cos \alpha}{\sin \alpha} \frac{1}{\sin \beta_1} \frac{\cos \beta_2}{\sin \beta_2} + \\ &+ \frac{\sigma v s(\alpha)}{3I_2 \sin \alpha} \Delta_{v,3} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \times \\ &\times \left\{ z_4 - z_3 \frac{\cos \beta_1}{\sin \beta_1} + z_2 \frac{1}{\sin \beta_1} \frac{\cos \beta_2}{\sin \beta_2} \right\} - \\ &- \frac{v^2}{3I_2} s(\alpha) \Delta_{v,3} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right), \end{aligned} \quad (65)$$

$$\dot{\beta}_1 = z_3 \frac{\cos \alpha}{\sin \alpha} + \frac{\sigma v s(\alpha)}{3I_2 \sin \alpha} \Delta_{v,1} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right), \quad (66)$$

$$\begin{aligned} \dot{\beta}_2 &= -z_2 \frac{\cos \alpha}{\sin \alpha \sin \beta_1} + \\ &+ \frac{\sigma v s(\alpha)}{3I_2 \sin \alpha \sin \beta_1} \Delta_{v,2} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right), \end{aligned} \quad (67)$$

$$\dot{\beta}_3 = z_1 \frac{\cos \alpha}{\sin \alpha \sin \beta_1 \sin \beta_2} + \frac{\sigma v}{3I_2} \frac{s(\alpha)}{\sin \alpha \sin \beta_1 \sin \beta_2} \Delta_{v,3} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right), \quad (68)$$

where

$$\begin{aligned} \Delta_{v,1} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) &= \\ &= -x_{2N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \sin \beta_1 + \\ &+ x_{3N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \cos \beta_1 \cos \beta_2 + \\ &+ x_{4N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \cos \beta_1 \sin \beta_2 \cos \beta_3 + \\ &+ x_{5N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \cos \beta_1 \sin \beta_2 \sin \beta_3, \\ \Delta_{v,2} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) &= \\ &= -x_{3N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \sin \beta_2 + \\ &+ x_{4N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \cos \beta_2 \cos \beta_3 + \\ &+ x_{5N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \cos \beta_2 \sin \beta_3, \\ \Delta_{v,3} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) &= \\ &= -x_{4N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \sin \beta_3 + \\ &+ x_{5N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \cos \beta_3, \end{aligned} \quad (69)$$

and the function $\Gamma_v(\alpha, \beta_1, \beta_2, \beta_3, \Omega/v)$ can be represented in the form (41).

Here and in the sequel, the dependence on the group of variables $(\alpha, \beta_1, \beta_2, \beta_3, \Omega/v)$ is meant as the composite dependence on $(\alpha, \beta_1, \beta_2, \beta_3, z_1/v, z_2/v, z_3/v, z_4/v)$ due to (59).

The uniqueness theorem for system (50)–(57) on the manifold (60) for odd k violates in the following sense: for odd k through almost all points of the manifold (60), passes a nonsingular phase trajectory of system (50)–(57) intersecting the manifold (60) at right angle and there exists a phase trajectory that at any time instants completely coincides with the point specified. However, physically these trajectories are different since they correspond to different values of the tracing force. Prove this.

As was shown above, to maintain the constraint of the form (38), we must take a value of T for $\cos \alpha \neq 0$ according to (40).

Let

$$\lim_{\alpha \rightarrow \pi/2} \frac{s(\alpha)}{\cos \alpha} \Gamma_v \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) =$$

$$= L \left(\beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right). \quad (70)$$

Note that $|L| < +\infty$ if and only if

$$\lim_{\alpha \rightarrow \pi/2} \left| \frac{\partial}{\partial \alpha} \left(\Gamma_v \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) s(\alpha) \right) \right| < +\infty. \quad (71)$$

For $\alpha = \pi/2$, the required value of the tracing force is define by the equation

$$\begin{aligned} T &= T_v \left(\frac{\pi}{2}, \beta_1, \beta_2, \beta_3, \Omega \right) = \\ &= m\sigma(\omega_4^2 + \omega_7^2 + \omega_9^2 + \omega_{10}^2) - \frac{m\sigma L v^2}{2I_2}. \end{aligned} \quad (72)$$

where $\omega_4, \omega_7, \omega_9, \omega_{10}$ are arbitrary.

On the other hand, maintaining the rotation about some point W by the tracing force, we must choose this force according to the relation

$$T = T_v \left(\frac{\pi}{2}, \beta_1, \beta_2, \beta_3, \Omega \right) = \frac{mv^2}{R_0}, \quad (73)$$

where R_0 is the distance CW .

Relations (72) and (73) define in general, different values of the tracing force T for almost all points of the manifold (60), which proves our assertion.

IV. CASE WHERE THE MOMENT OF A NONCONSERVATIVE FORCE IS INDEPENDENT OF THE ANGULAR VELOCITY

A. Reduced system

Similarly to the choice of Chaplygin analytic functions, we take the dynamical functions s , x_{2N} , x_{3N} , x_{4N} , and x_{5N} in the following form:

$$\begin{aligned} s(\alpha) &= B \cos \alpha, \\ x_{2N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) &= \\ &= x_{2N0}(\alpha, \beta_1, \beta_2, \beta_3) = A \sin \alpha \cos \beta_1, \\ x_{3N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) &= \\ &= x_{3N0}(\alpha, \beta_1, \beta_2, \beta_3) = A \sin \alpha \sin \beta_1 \cos \beta_2, \\ x_{4N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) &= \\ &= x_{4N0}(\alpha, \beta_1, \beta_2, \beta_3) = A \sin \alpha \sin \beta_1 \sin \beta_2 \cos \beta_3, \\ x_{5N} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) &= \\ &= x_{5N0}(\alpha, \beta_1, \beta_2, \beta_3) = \\ &= A \sin \alpha \sin \beta_1 \sin \beta_2 \sin \beta_3, \quad A, B > 0, \quad v \neq 0. \end{aligned} \quad (74)$$

We see that in the system considered, the moment of nonconservative forces is independent of the angular velocity (but depends on the angles $\alpha, \beta_1, \beta_2, \beta_3$).

Herewith, the functions $\Gamma_v(\alpha, \beta_1, \beta_2, \beta_3, \Omega/v)$, $\Delta_{v,s}(\alpha, \beta_1, \beta_2, \beta_3, \Omega/v)$, $s = 1, 2, 3$, in system (61)–(68), take the following form:

$$\Gamma_v \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) = A \sin \alpha,$$

$$\Delta_{v,s} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) \equiv 0, \quad s = 1, 2, 3. \quad (75)$$

Then, due to the nonintegrable constraint (38), outside the manifold (60), the dynamical part of the equations of motion (system (61)–(68)) has the form of the following analytic system:

$$\alpha' = -z_4 + \frac{\sigma ABv}{3I_2} \sin \alpha, \quad (76)$$

$$z_4' = \frac{ABv^2}{3I_2} \sin \alpha \cos \alpha - (z_1^2 + z_2^2 + z_3^2) \frac{\cos \alpha}{\sin \alpha}, \quad (77)$$

$$z_3' = z_3 z_4 \frac{\cos \alpha}{\sin \alpha} + (z_1^2 + z_2^2) \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1}, \quad (78)$$

$$z_2' = z_2 z_4 \frac{\cos \alpha}{\sin \alpha} - z_2 z_3 \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1} - z_1^2 \frac{\cos \alpha}{\sin \alpha} \frac{1}{\sin \beta_1} \frac{\cos \beta_2}{\sin \beta_2}, \quad (79)$$

$$z_1' = z_1 z_4 \frac{\cos \alpha}{\sin \alpha} - z_1 z_3 \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1} + z_1 z_2 \frac{\cos \alpha}{\sin \alpha} \frac{1}{\sin \beta_1} \frac{\cos \beta_2}{\sin \beta_2}, \quad (80)$$

$$\beta_1' = z_3 \frac{\cos \alpha}{\sin \alpha}, \quad (81)$$

$$\beta_2' = -z_2 \frac{\cos \alpha}{\sin \alpha \sin \beta_1}, \quad (82)$$

$$\beta_3' = z_1 \frac{\cos \alpha}{\sin \alpha \sin \beta_1 \sin \beta_2}. \quad (83)$$

Further, introducing the dimensionless variables, parameters, and the differentiation as follows:

$$z_k \mapsto n_0 v z_k, \quad k = 1, 2, 3, 4, \quad n_0^2 = \frac{AB}{3I_2}, \quad (84)$$

$$b = \sigma n_0, \quad \langle \cdot \rangle = n_0 v \langle' \rangle,$$

we reduce system (76)–(83) to the form

$$\alpha' = -z_4 + b \sin \alpha, \quad (85)$$

$$z_4' = \sin \alpha \cos \alpha - (z_1^2 + z_2^2 + z_3^2) \frac{\cos \alpha}{\sin \alpha}, \quad (86)$$

$$z_3' = z_3 z_4 \frac{\cos \alpha}{\sin \alpha} + (z_1^2 + z_2^2) \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1}, \quad (87)$$

$$z_2' = z_2 z_4 \frac{\cos \alpha}{\sin \alpha} - z_2 z_3 \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1} - z_1^2 \frac{\cos \alpha}{\sin \alpha} \frac{1}{\sin \beta_1} \frac{\cos \beta_2}{\sin \beta_2}, \quad (88)$$

$$z_1' = z_1 z_4 \frac{\cos \alpha}{\sin \alpha} - z_1 z_3 \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1} + z_1 z_2 \frac{\cos \alpha}{\sin \alpha} \frac{1}{\sin \beta_1} \frac{\cos \beta_2}{\sin \beta_2}, \quad (89)$$

$$\beta_1' = z_3 \frac{\cos \alpha}{\sin \alpha}, \quad (90)$$

$$\beta_2' = -z_2 \frac{\cos \alpha}{\sin \alpha \sin \beta_1}, \quad (91)$$

$$\beta_3' = z_1 \frac{\cos \alpha}{\sin \alpha \sin \beta_1 \sin \beta_2}. \quad (92)$$

We see that the eighth-order system (85)–(92) (which can be considered as a system on the tangent bundle TS^4 of the four-dimensional sphere S^4 , see below) contains the independent seventh-order system (85)–(91) on its own seven-dimensional manifold.

For the complete integration of system (85)–(92), in general, we need seven independent first integrals. However, after the change of variables

$$\begin{pmatrix} z_4 \\ z_3 \\ z_2 \\ z_1 \end{pmatrix} \rightarrow \begin{pmatrix} w_4 \\ w_3 \\ w_2 \\ w_1 \end{pmatrix},$$

$$w_4 = z_4, \quad w_3 = \sqrt{z_1^2 + z_2^2 + z_3^2}, \quad (93)$$

$$w_2 = \frac{z_2}{z_1}, \quad w_1 = \frac{z_3}{\sqrt{z_1^2 + z_2^2}},$$

system (85)–(92) splits as follows:

$$\alpha' = -w_4 + b \sin \alpha, \quad (94)$$

$$w_4' = \sin \alpha \cos \alpha - w_3^2 \frac{\cos \alpha}{\sin \alpha}, \quad (95)$$

$$w_3' = w_3 w_4 \frac{\cos \alpha}{\sin \alpha}, \quad (96)$$

$$w_2' = d_2(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3) \frac{1 + w_2^2 \cos \beta_2}{w_2 \sin \beta_2}, \quad (97)$$

$$\beta_2' = d_2(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3),$$

$$w_1' = d_1(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3) \frac{1 + w_1^2 \cos \beta_1}{w_1 \sin \beta_1}, \quad (98)$$

$$\beta_1' = d_1(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3),$$

$$\beta_3' = d_3(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3), \quad (99)$$

where

$$d_1(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3) = Z_3(w_4, w_3, w_2, w_1) \frac{\cos \alpha}{\sin \alpha},$$

$$d_2(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3) = -Z_2(w_4, w_3, w_2, w_1) \frac{\cos \alpha}{\sin \alpha \sin \beta_1}, \quad (100)$$

$$d_3(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3) = Z_1(w_4, w_3, w_2, w_1) \frac{\cos \alpha}{\sin \alpha \sin \beta_1 \sin \beta_2},$$

herewith

$$z_k = Z_k(w_4, w_3, w_2, w_1), \quad k = 1, 2, 3, \quad (101)$$

are the functions due to the change of variables (93).

We see that the eighth-order system splits into independent subsystems of lower order: system (94)–(96) has order three and systems (97), (98) (after the change of the independent variable) have order two. Thus, for the complete integration of system (94)–(99) it suffice to specify two independent first integrals of system (94)–(96), one first integral of each system (97), (98), and an additional first integral that attaches Eq. (99).

Note that system (94)–(96) can be considered on the tangent bundle TS^2 of the two-dimensional sphere S^2 .

B. Complete list of invariant relations

System (94)–(96) has the form of a system that appears in the dynamics of a three-dimensional (3D-) rigid body in a field of nonconservative forces.

First, to the third-order system (94)–(96), we put in correspondence the nonautonomous second-order system

$$\begin{aligned} \frac{dw_4}{d\alpha} &= \frac{\sin \alpha \cos \alpha - w_3^2 \cos \alpha / \sin \alpha}{-w_4 + b \sin \alpha}, \\ \frac{dw_3}{d\alpha} &= \frac{w_3 w_4 \cos \alpha / \sin \alpha}{-w_4 + b \sin \alpha}. \end{aligned} \tag{102}$$

Applying the substitution $\tau = \sin \alpha$, we rewrite system (102) in the algebraic form

$$\begin{aligned} \frac{dw_4}{d\tau} &= \frac{\tau - w_3^2/\tau}{-w_4 + b\tau}, \\ \frac{dw_3}{d\tau} &= \frac{w_3 w_4/\tau}{-w_4 + b\tau}. \end{aligned} \tag{103}$$

Later on, introducing the homogeneous variables by the formulas

$$w_3 = u_1\tau, \quad w_4 = u_2\tau, \tag{104}$$

we reduce system (103) to the following form:

$$\begin{aligned} \tau \frac{du_2}{d\tau} + u_2 &= \frac{1 - u_1^2}{-u_2 + b}, \\ \tau \frac{du_1}{d\tau} + u_1 &= \frac{u_1 u_2}{-u_2 + b}, \end{aligned} \tag{105}$$

which is equivalent to the system

$$\begin{aligned} \tau \frac{du_2}{d\tau} &= \frac{1 - u_1^2 + u_2^2 - bu_2}{-u_2 + b}, \\ \tau \frac{du_1}{d\tau} &= \frac{2u_1 u_2 - bu_1}{-u_2 + b}. \end{aligned} \tag{106}$$

To the second-order system (106), we put in correspondence the nonautonomous first-order equation

$$\frac{du_2}{du_1} = \frac{1 - u_1^2 + u_2^2 - bu_2}{2u_1 u_2 - bu_1}, \tag{107}$$

which can be easily reduced to the exact-differential form:

$$d \left(\frac{u_2^2 + u_1^2 - bu_2 + 1}{u_1} \right) = 0. \tag{108}$$

Thus, Eq. (107) has the following first integral:

$$\frac{u_2^2 + u_1^2 - bu_2 + 1}{u_1} = C_1 = \text{const}, \tag{109}$$

which in the previous variables has the form

$$\frac{w_4^2 + w_3^2 - bw_4 \sin \alpha + \sin^2 \alpha}{w_3 \sin \alpha} = C_1 = \text{const}. \tag{110}$$

Remark 1. Consider system (94)–(96) with variable dissipation with zero mean, that becomes conservative for $b = 0$:

$$\begin{aligned} \alpha' &= -w_4, \\ w_4' &= \sin \alpha \cos \alpha - w_3^2 \frac{\cos \alpha}{\sin \alpha}, \\ w_3' &= w_3 w_4 \frac{\cos \alpha}{\sin \alpha}. \end{aligned} \tag{111}$$

It possesses two analytic first integrals of the form

$$w_4^2 + w_3^2 + \sin^2 \alpha = C_1^* = \text{const}, \tag{112}$$

$$w_3 \sin \alpha = C_2^* = \text{const}. \tag{113}$$

Obviously, the ratio of two first integrals (112), (113) is also a first integral of system (111). But for $b \neq 0$, each of the functions

$$w_4^2 + w_3^2 - bw_4 \sin \alpha + \sin^2 \alpha \tag{114}$$

and (113) is not a first integral of system (94)–(96). However, but their ratio is a first integral for any b .

Further, we find the explicit form of the additional first integral of the third-order system (94)–(96). For this, we transform the invariant relation (109) for $u_1 \neq 0$ as follows:

$$\left(u_2 - \frac{b}{2}\right)^2 + \left(u_1 - \frac{C_1}{2}\right)^2 = \frac{b^2 + C_1^2}{4} - 1. \tag{115}$$

We see that the parameters of this invariant relation satisfy the condition

$$b^2 + C_1^2 - 4 \geq 0, \tag{116}$$

and the phase space of system (94)–(96) is stratified into the family of surfaces defined by Eq. (115).

Thus, by relation (109), the first equation of system (106) has the form

$$\tau \frac{du_2}{d\tau} = \frac{2(1 - bu_2 + u_2^2) - C_1 U_1(C_1, u_2)}{-u_2 + b}, \tag{117}$$

where

$$U_1(C_1, u_2) = \frac{1}{2} \{C_1 \pm \sqrt{C_1^2 - 4(u_2^2 - bu_2 + 1)}\}; \tag{118}$$

the integration constant C_1 is defined by condition (116).

Therefore, the quadrature for the search for the additional first integral of system (94)–(96) becomes

$$\begin{aligned} \int \frac{d\tau}{\tau} &= \\ &= \int \frac{(b - u_2) du_2}{2A^0 - C_1 \{C_1 \pm \sqrt{C_1^2 - 4A^0}\}/2}, \end{aligned} \tag{119}$$

$$A^0 = 1 - bu_2 + u_2^2.$$

Obviously, the left-hand side (up to an additive constant) equals

$$\ln |\sin \alpha|. \tag{120}$$

If

$$u_2 - \frac{b}{2} = r_1, \quad b_1^2 = b^2 + C_1^2 - 4, \tag{121}$$

then the right-hand side of Eq. (119) has the form

$$\begin{aligned} &-\frac{1}{4} \int \frac{d(b_1^2 - 4r_1^2)}{(b_1^2 - 4r_1^2) \pm C_1 \sqrt{b_1^2 - 4r_1^2}} - \\ &-b \int \frac{dr_1}{(b_1^2 - 4r_1^2) \pm C_1 \sqrt{b_1^2 - 4r_1^2}} = \\ &= -\frac{1}{2} \ln \left| \frac{\sqrt{b_1^2 - 4r_1^2}}{C_1} \pm 1 \right| \pm \frac{b}{2} I_1, \end{aligned} \tag{122}$$

where

$$I_1 = \int \frac{dr_3}{\sqrt{b_1^2 - r_3^2}(r_3 \pm C_1)}, \quad r_3 = \sqrt{b_1^2 - 4r_1^2}. \quad (123)$$

In the calculation of integral (123), the following three cases are possible.

I. $b > 2$.

$$I_1 = -\frac{1}{2\sqrt{b^2 - 4}} \ln \left| \frac{\sqrt{b^2 - 4} + \sqrt{b_1^2 - r_3^2}}{r_3 \pm C_1} \pm \frac{C_1}{\sqrt{b^2 - 4}} \right| + \frac{1}{2\sqrt{b^2 - 4}} \ln \left| \frac{\sqrt{b^2 - 4} - \sqrt{b_1^2 - r_3^2}}{r_3 \pm C_1} \mp \frac{C_1}{\sqrt{b^2 - 4}} \right| + \text{const.} \quad (124)$$

II. $b < 2$.

$$I_1 = \frac{1}{\sqrt{4 - b^2}} \arcsin \frac{\pm C_1 r_3 + b_1^2}{b_1(r_3 \pm C_1)} + \text{const.} \quad (125)$$

III. $b = 2$.

$$I_1 = \mp \frac{\sqrt{b_1^2 - r_3^2}}{C_1(r_3 \pm C_1)} + \text{const.} \quad (126)$$

Returning to the variable

$$r_1 = \frac{w_4}{\sin \alpha} - \frac{b}{2}, \quad (127)$$

we obtain the final expression for I_1 :

I. $b > 2$.

$$I_1 = -\frac{1}{2\sqrt{b^2 - 4}} \ln \left| \frac{\sqrt{b^2 - 4} \pm 2r_1}{\sqrt{b_1^2 - 4r_1^2} \pm C_1} \pm \frac{C_1}{\sqrt{b^2 - 4}} \right| + \frac{1}{2\sqrt{b^2 - 4}} \ln \left| \frac{\sqrt{b^2 - 4} \mp 2r_1}{\sqrt{b_1^2 - 4r_1^2} \pm C_1} \mp \frac{C_1}{\sqrt{b^2 - 4}} \right| + \text{const.} \quad (128)$$

II. $b < 2$.

$$I_1 = \frac{1}{\sqrt{4 - b^2}} \arcsin \frac{\pm C_1 \sqrt{b_1^2 - 4r_1^2} + b_1^2}{b_1(\sqrt{b_1^2 - 4r_1^2} \pm C_1)} + \text{const.} \quad (129)$$

III. $b = 2$.

$$I_1 = \mp \frac{2r_1}{C_1(\sqrt{b_1^2 - 4r_1^2} \pm C_1)} + \text{const.} \quad (130)$$

Thus, we have found an additional first integral for the third-order system (94)–(96) and we have the complete set of first integrals that are transcendental functions of their phase variables.

Remark 2. We must substitute the left-hand side of the first integral (109) in the expression of this first integral instead C_1 .

Then the additional first integral obtained has the following structure (similar to the transcendental first integral in planar dynamics):

$$\ln |\sin \alpha| + G_2 \left(\sin \alpha, \frac{w_4}{\sin \alpha}, \frac{w_3}{\sin \alpha} \right) = C_2 = \text{const.} \quad (131)$$

Thus, for the integration of the eighth-order system (94)–(99), we have found two independent first integrals. For the complete integration, as was mentioned above, it suffices to find one first integral for each (potentially separated) system

(97), (98), and an additional first integral that attaches Eq. (99).

To find a first integral for each (potentially separated) system (97), (98), we put in correspondence the following nonautonomous first-order equation:

$$\frac{dw_s}{d\beta_s} = \frac{1 + w_s^2 \cos \beta_s}{w_s \sin \beta_s}, \quad s = 1, 2. \quad (132)$$

After integration, this leads to the invariant relation

$$\frac{\sqrt{1 + w_s^2}}{\sin \beta_s} = C_{s+2} = \text{const}, \quad s = 1, 2. \quad (133)$$

Further, for the search for an additional first integral that attaches Eq. (99), to Eqs. (99) and (97) we put in correspondence the following nonautonomous equation:

$$\frac{dw_2}{d\beta_3} = -(1 + w_2^2) \cos \beta_2. \quad (134)$$

Since, by (133),

$$C_4 \cos \beta_2 = \pm \sqrt{C_4^2 - 1 - w_2^2}, \quad (135)$$

we have

$$\frac{dw_2}{d\beta_3} = \mp \frac{1}{C_4} (1 + w_2^2) \sqrt{C_4^2 - 1 - w_2^2}. \quad (136)$$

Integrating the last relation, we arrive at the following quadrature:

$$\mp (\beta_3 + C_5) = \int \frac{C_4 dw_2}{(1 + w_2^2) \sqrt{C_4^2 - 1 - w_2^2}}, \quad C_5 = \text{const.} \quad (137)$$

Integrating this relation we obtain

$$\mp \text{tg}(\beta_3 + C_5) = \frac{C_4 w_2}{\sqrt{C_4^2 - 1 - w_2^2}}, \quad C_5 = \text{const.} \quad (138)$$

Finally, we have the following form of the additional first integral that attaches Eq. (99):

$$\text{arctg} \frac{C_4 w_2}{\sqrt{C_4^2 - 1 - w_2^2}} \pm \beta_3 = C_5, \quad C_5 = \text{const.} \quad (139)$$

Thus, in the case considered, the system of dynamical equations (17)–(21), (24)–(33) under condition (74) has twelve invariant relations: the nonintegrable analytic constraint of the form (38), the cyclic first integrals of the form (36), (37), the first integral of the form (110), the first integral expressed by relations (124)–(131), which is a transcendental function of the phase variables (in the sense of complex analysis) expressed through a finite combination of elementary functions, and, finally, the transcendental first integrals of the form (133) and (139).

Theorem 1. System (17)–(21), (24)–(33) under conditions (38), (74), (37) possesses twelve invariant relations (complete set), five of which transcendental functions from the point of view of complex analysis. Herewith, all relations are expressed through finite combinations of elementary functions.

C. Topological analogies

Consider the following seventh-order system:

$$\begin{aligned} &\ddot{\xi} + b_* \dot{\xi} \cos \xi + \sin \xi \cos \xi - \\ &- [\dot{\eta}_1^2 + \dot{\eta}_2^2 \sin^2 \eta_1 + \dot{\eta}_3^2 \sin^2 \eta_1 \sin^2 \eta_2] \frac{\sin \xi}{\cos \xi} = 0, \\ &\ddot{\eta}_1 + b_* \dot{\eta}_1 \cos \xi + \dot{\xi} \dot{\eta}_1 \frac{1 + \cos^2 \xi}{\cos \xi \sin \xi} - \\ &- (\dot{\eta}_2^2 + \dot{\eta}_3^2 \sin^2 \eta_2) \sin \eta_1 \cos \eta_1 = 0, \\ &\ddot{\eta}_2 + b_* \dot{\eta}_2 \cos \xi + \dot{\xi} \dot{\eta}_2 \frac{1 + \cos^2 \xi}{\cos \xi \sin \xi} + \\ &+ 2\dot{\eta}_1 \dot{\eta}_2 \frac{\cos \eta_1}{\sin \eta_1} - \dot{\eta}_3^2 \sin \eta_2 \cos \eta_2 = 0, \\ &\ddot{\eta}_3 + b_* \dot{\eta}_3 \cos \xi + \dot{\xi} \dot{\eta}_3 \frac{1 + \cos^2 \xi}{\cos \xi \sin \xi} + \\ &+ 2\dot{\eta}_1 \dot{\eta}_3 \frac{\cos \eta_1}{\sin \eta_1} + 2\dot{\eta}_2 \dot{\eta}_3 \frac{\cos \eta_2}{\sin \eta_2} = 0, \quad b_* > 0, \end{aligned} \tag{140}$$

which describes a fixed five-dimensional pendulum in a flow of a running medium for which the moment of forces is independent of the angular velocity, i.e., a mechanical system in a nonconservative field. In general the order of such a system is equal to 8, but the phase variable η_3 is a cyclic variable, which leads to the stratification of the phase space and reduces the order of the system.

The phase space of this system is the tangent bundle

$$T\mathbf{S}^3\{\dot{\xi}, \dot{\eta}_1, \dot{\eta}_2, \dot{\eta}_3, \xi, \eta_1, \eta_2, \eta_3\} \tag{141}$$

of the four-dimensional sphere $\mathbf{S}^4\{\xi, \eta_1, \eta_2, \eta_3\}$. The equation that transforms system (140) to the system on the tangent bundle of the three-dimensional sphere

$$\dot{\eta}_3 \equiv 0, \tag{142}$$

and the equations of great circles

$$\dot{\eta}_1 \equiv 0, \quad \dot{\eta}_2 \equiv 0, \quad \dot{\eta}_3 \equiv 0 \tag{143}$$

define families of integral manifolds.

It is easy to verify that system (140) is equivalent to the dynamical system with variable dissipation with zero mean on the tangent bundle (141) of the four-dimensional sphere. Moreover, the following theorem holds.

Theorem 2. *System (17)–(21), (24)–(33) under conditions (38), (74), (37) is equivalent to the dynamical system (140).*

Indeed it suffices to set $\alpha = \xi$, $\beta_1 = \eta_1$, $\beta_2 = \eta_2$, $\beta_3 = \eta_3$, $b = -b_*$.

V. CASE WHERE THE MOMENT OF A NONCONSERVATIVE FORCE DEPENDS ON THE ANGULAR VELOCITY

A. Introduction of the dependence on the angular velocity

This chapter is devoted to the dynamics of a five-dimensional rigid body in the five-dimensional space. Since the present section is devoted to the study of the motion in the case where the moment of forces depends on the tensor of angular velocity, we introduce this dependence in a more general situation. This also allows us to introduce this dependence for multi-dimensional bodies.

Let $x = (x_{1N}, x_{2N}, x_{3N}, x_{4N}, x_{5N})$ be the coordinates of the point N of application of a nonconservative force (influence of the medium) acting on the four-dimensional disk and $Q = (Q_1, Q_2, Q_3, Q_4, Q_5)$ be the components independent of the tensor of the angular velocity. We consider only linear dependence of the functions $(x_{1N}, x_{2N}, x_{3N}, x_{4N}, x_{5N})$ on the tensor of angular velocity since this introduction itself is not obvious.

We adopt the following dependence:

$$x = Q + R, \tag{144}$$

where $R = (R_1, R_2, R_3, R_4, R_5)$ is a vector-valued function containing the components of the tensor of angular velocity. The dependence of the function R on the components of the tensor of angular velocity is gyroscopic:

$$\begin{aligned} R &= \begin{pmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \end{pmatrix} = \\ &= -\frac{1}{v} \begin{pmatrix} 0 & -\omega_{10} & \omega_9 & -\omega_7 & \omega_4 \\ \omega_{10} & 0 & -\omega_8 & \omega_6 & -\omega_3 \\ -\omega_9 & \omega_8 & 0 & -\omega_5 & \omega_2 \\ \omega_7 & -\omega_6 & \omega_5 & 0 & -\omega_1 \\ -\omega_4 & \omega_3 & -\omega_2 & \omega_1 & 0 \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \end{pmatrix}, \end{aligned} \tag{145}$$

where $(h_1, h_2, h_3, h_4, h_5)$ are some positive parameters.

Since $x_{1N} \equiv 0$, we have

$$\begin{aligned} x_{2N} &= Q_2 - h_1 \frac{\omega_{10}}{v}, \\ x_{3N} &= Q_3 + h_1 \frac{\omega_9}{v}, \\ x_{4N} &= Q_4 - h_1 \frac{\omega_7}{v}, \\ x_{5N} &= Q_5 + h_1 \frac{\omega_4}{v}. \end{aligned} \tag{146}$$

B. Reduced system

Similarly to the choice of the Chaplygin analytic functions

$$\begin{aligned} Q_2 &= A \sin \alpha \cos \beta_1, \quad Q_3 = A \sin \alpha \sin \beta_1 \cos \beta_2, \\ Q_4 &= A \sin \alpha \sin \beta_1 \sin \beta_2 \cos \beta_3, \\ Q_5 &= A \sin \alpha \sin \beta_1 \sin \beta_2 \sin \beta_3, \quad A > 0, \end{aligned} \tag{147}$$

we take the dynamical functions s , x_{2N} , x_{3N} , x_{4N} , and x_{5N} in the following form:

$$\begin{aligned} s(\alpha) &= B \cos \alpha, \quad B > 0, \\ x_{2N} \left(\alpha, \beta_1, \beta_2, \frac{\Omega}{v} \right) &= A \sin \alpha \cos \beta_1 - h \frac{\omega_{10}}{v}, \\ x_{3N} \left(\alpha, \beta_1, \beta_2, \frac{\Omega}{v} \right) &= A \sin \alpha \sin \beta_1 \cos \beta_2 + h \frac{\omega_9}{v}, \\ x_{4N} \left(\alpha, \beta_1, \beta_2, \frac{\Omega}{v} \right) &= A \sin \alpha \sin \beta_1 \sin \beta_2 \cos \beta_3 - h \frac{\omega_7}{v}, \end{aligned} \tag{148}$$

$$x_{5N} \left(\alpha, \beta_1, \beta_2, \frac{\Omega}{v} \right) = A \sin \alpha \sin \beta_1 \sin \beta_2 \sin \beta_3 + h \frac{\omega_4}{v}, \quad h = h_1 > 0, \quad v \neq 0.$$

This shows that in the problem considered, there is an additional damping (but accelerating in certain domains of the phase space) moment of a nonconservative force (i.e., there is a dependence of the moment on the components of the tensor of angular velocity). Moreover, $h_2 = h_3 = h_4 = h_5$ due to the dynamical symmetry of the body.

In this case, the functions $\Gamma_v(\alpha, \beta_1, \beta_2, \beta_3, \Omega/v)$, $\Delta_{v,s}(\alpha, \beta_1, \beta_2, \beta_3, \Omega/v)$, $s = 1, 2, 3$, in system (61)–(68) have the following form:

$$\begin{aligned} \Gamma_v \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) &= A \sin \alpha - \frac{h}{v} z_4, \\ \Delta_{v,1} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) &= \frac{h}{v} z_3, \\ \Delta_{v,2} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) &= -\frac{h}{v} z_2, \\ \Delta_{v,3} \left(\alpha, \beta_1, \beta_2, \beta_3, \frac{\Omega}{v} \right) &= \frac{h}{v} z_1. \end{aligned} \quad (149)$$

Then, due to the nonintegrable constraint (38), outside the manifold (60) the dynamical part of the equations of motion (system (61)–(68)) takes the form of the analytic system

$$\dot{\alpha} = - \left(1 + \frac{\sigma Bh}{3I_2} \right) z_4 + \frac{\sigma ABv}{3I_2} \sin \alpha, \quad (150)$$

$$\dot{z}_4 = \frac{ABv^2}{3I_2} \sin \alpha \cos \alpha -$$

$$- \left(1 + \frac{\sigma Bh}{3I_2} \right) (z_1^2 + z_2^2 + z_3^2) \frac{\cos \alpha}{\sin \alpha} - \frac{Bhv}{3I_2} z_4 \cos \alpha, \quad (151)$$

$$\dot{z}_3 = \left(1 + \frac{\sigma Bh}{3I_2} \right) z_3 z_4 \frac{\cos \alpha}{\sin \alpha} +$$

$$+ \left(1 + \frac{\sigma Bh}{3I_2} \right) (z_1^2 + z_2^2) \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1} - \frac{Bhv}{3I_2} z_3 \cos \alpha, \quad (152)$$

$$\dot{z}_2 = \left(1 + \frac{\sigma Bh}{3I_2} \right) z_2 z_4 \frac{\cos \alpha}{\sin \alpha} -$$

$$- \left(1 + \frac{\sigma Bh}{3I_2} \right) z_2 z_3 \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1} -$$

$$- \left(1 + \frac{\sigma Bh}{3I_2} \right) z_1^2 \frac{\cos \alpha}{\sin \alpha} \frac{1}{\sin \beta_1} \frac{\cos \beta_2}{\sin \beta_2} - \frac{Bhv}{3I_2} z_2 \cos \alpha, \quad (153)$$

$$\dot{z}_1 = \left(1 + \frac{\sigma Bh}{3I_2} \right) z_1 z_4 \frac{\cos \alpha}{\sin \alpha} -$$

$$- \left(1 + \frac{\sigma Bh}{3I_2} \right) z_1 z_3 \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1} +$$

$$+ \left(1 + \frac{\sigma Bh}{3I_2} \right) z_1 z_2 \frac{\cos \alpha}{\sin \alpha} \frac{1}{\sin \beta_1} \frac{\cos \beta_2}{\sin \beta_2} -$$

$$- \frac{Bhv}{3I_2} z_1 \cos \alpha, \quad (154)$$

$$\dot{\beta}_1 = \left(1 + \frac{\sigma Bh}{3I_2} \right) z_3 \frac{\cos \alpha}{\sin \alpha}, \quad (155)$$

$$\dot{\beta}_2 = - \left(1 + \frac{\sigma Bh}{3I_2} \right) z_2 \frac{\cos \alpha}{\sin \alpha \sin \beta_1}, \quad (156)$$

$$\dot{\beta}_3 = \left(1 + \frac{\sigma Bh}{3I_2} \right) z_1 \frac{\cos \alpha}{\sin \alpha \sin \beta_1 \sin \beta_2}. \quad (157)$$

Introducing the dimensionless variables, parameters, and the differentiation as follows:

$$z_k \mapsto n_0 v z_k, \quad k = 1, 2, 3, 4, \quad n_0^2 = \frac{AB}{3I_2}, \quad (158)$$

$$b = \sigma n_0, \quad H_1 = \frac{Bh}{3I_2 n_0}, \quad \langle \cdot \rangle = n_0 v \langle \cdot \rangle',$$

we reduce system (150)–(157) to the form

$$\alpha' = - (1 + bH_1) z_4 + b \sin \alpha, \quad (159)$$

$$z_4' = \sin \alpha \cos \alpha -$$

$$- (1 + bH_1) (z_1^2 + z_2^2 + z_3^2) \frac{\cos \alpha}{\sin \alpha} - H_1 z_4 \cos \alpha, \quad (160)$$

$$z_3' = (1 + bH_1) z_3 z_4 \frac{\cos \alpha}{\sin \alpha} +$$

$$+ (1 + bH_1) (z_1^2 + z_2^2) \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1} - H_1 z_3 \cos \alpha, \quad (161)$$

$$z_2' = (1 + bH_1) z_2 z_4 \frac{\cos \alpha}{\sin \alpha} -$$

$$- (1 + bH_1) z_2 z_3 \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1} -$$

$$- (1 + bH_1) z_1^2 \frac{\cos \alpha}{\sin \alpha} \frac{1}{\sin \beta_1} \frac{\cos \beta_2}{\sin \beta_2} - H_1 z_2 \cos \alpha, \quad (162)$$

$$z_1' = (1 + bH_1) z_1 z_4 \frac{\cos \alpha}{\sin \alpha} -$$

$$- (1 + bH_1) z_1 z_3 \frac{\cos \alpha \cos \beta_1}{\sin \alpha \sin \beta_1} +$$

$$+ (1 + bH_1) z_1 z_2 \frac{\cos \alpha}{\sin \alpha} \frac{1}{\sin \beta_1} \frac{\cos \beta_2}{\sin \beta_2} - H_1 z_1 \cos \alpha, \quad (163)$$

$$\beta_1' = (1 + bH_1) z_3 \frac{\cos \alpha}{\sin \alpha}, \quad (164)$$

$$\beta_2' = - (1 + bH_1) z_2 \frac{\cos \alpha}{\sin \alpha \sin \beta_1}, \quad (165)$$

$$\beta_3' = (1 + bH_1) z_1 \frac{\cos \alpha}{\sin \alpha \sin \beta_1 \sin \beta_2}. \quad (166)$$

We see that the eighth-order system (159)–(166) (which can be considered on the tangent bundle TS^4 of the four-dimensional sphere S^4), contains an independent seventh-order system (159)–(165) on its own seven-dimensional manifold.

For the complete integration of system (159)–(166), we need, in general, seven independent first integrals. However, after the change of variables

$$\begin{pmatrix} z_4 \\ z_3 \\ z_2 \\ z_1 \end{pmatrix} \rightarrow \begin{pmatrix} w_4 \\ w_3 \\ w_2 \\ w_1 \end{pmatrix},$$

$$w_4 = z_4, \quad w_3 = \sqrt{z_1^2 + z_2^2 + z_3^2},$$

$$w_2 = \frac{z_2}{z_1}, \quad w_1 = \frac{z_3}{\sqrt{z_1^2 + z_2^2}}, \quad (167)$$

system (159)–(166) splits as follows:

$$\alpha' = -(1 + bH_1)w_4 + b \sin \alpha, \quad (168)$$

$$w_4' = \sin \alpha \cos \alpha - (1 + bH_1)w_3^2 \frac{\cos \alpha}{\sin \alpha} - H_1 w_4 \cos \alpha, \quad (169)$$

$$w_3' = (1 + bH_1)w_3 w_4 \frac{\cos \alpha}{\sin \alpha} - H_1 w_3 \cos \alpha, \quad (170)$$

$$w_2' = d_2(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3) \times \frac{1 + w_2^2 \cos \beta_2}{w_2 \sin \beta_2}, \quad (171)$$

$$\beta_2' = d_2(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3),$$

$$w_1' = d_1(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3) \times \frac{1 + w_1^2 \cos \beta_1}{w_1 \sin \beta_1}, \quad (172)$$

$$\beta_1' = d_1(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3),$$

$$\beta_3' = d_3(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3), \quad (173)$$

where

$$d_1(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3) = (1 + bH_1)Z_3(w_4, w_3, w_2, w_1) \frac{\cos \alpha}{\sin \alpha},$$

$$d_2(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3) = -(1 + bH_1)Z_2(w_4, w_3, w_2, w_1) \frac{\cos \alpha}{\sin \alpha \sin \beta_1}, \quad (174)$$

$$d_3(w_4, w_3, w_2, w_1; \alpha, \beta_1, \beta_2, \beta_3) = (1 + bH_1)Z_1(w_4, w_3, w_2, w_1) \frac{\cos \alpha}{\sin \alpha \sin \beta_1 \sin \beta_2},$$

herewith,

$$z_k = Z_k(w_4, w_3, w_2, w_1), \quad k = 1, 2, 3, \quad (175)$$

are the functions, due to the change of variables (167).

We see that the eighth-order system splits into independent subsystems of lower orders: system (168)–(170) of order 3 and each of system (171), (172) (certainly, after a choice of the independent variables) of order 2. Thus, for the complete integration of system (168)–(173), it suffice to find two independent first integrals of system (168)–(170), one first integral of each system (171), (172), and an additional first integral that attaches Eq. (173).

Note that system (168)–(170) can be considered on the tangent bundle TS^2 of the two-dimensional sphere S^2 .

C. Complete list of invariant relation

System (168)–(170) has the form of a system of equations that appears in the dynamics of a three-dimensional (3D-) rigid body in a nonconservative field

First, to the third-order system (168)–(170), we put in correspondence the nonautonomous second-order system

$$\frac{dw_4}{d\alpha} = \frac{\sin \alpha \cos \alpha - (1 + bH_1)w_3^2 \cos \alpha / \sin \alpha - H_1 w_4 \cos \alpha}{-(1 + bH_1)w_4 + b \sin \alpha},$$

$$\frac{dw_3}{d\alpha} = \frac{(1 + bH_1)w_3 w_4 \cos \alpha / \sin \alpha - H_1 w_3 \cos \alpha}{-(1 + bH_1)w_4 + b \sin \alpha}. \quad (176)$$

Using the substitution $\tau = \sin \alpha$, we rewrite system (176) in the algebraic form:

$$\frac{dw_4}{d\tau} = \frac{\tau - (1 + bH_1)w_3^2/\tau - H_1 w_4}{-(1 + bH_1)w_4 + b\tau}, \quad (177)$$

$$\frac{dw_3}{d\tau} = \frac{(1 + bH_1)w_3 w_4/\tau - H_1 w_3}{-(1 + bH_1)w_4 + b\tau}.$$

Further, introducing the homogeneous variables by the formulas

$$w_3 = u_1\tau, \quad w_4 = u_2\tau, \quad (178)$$

we reduce system (177) to the following form:

$$\tau \frac{du_2}{d\tau} + u_2 = \frac{1 - (1 + bH_1)u_1^2 - H_1 u_2}{-(1 + bH_1)u_2 + b}, \quad (179)$$

$$\tau \frac{du_1}{d\tau} + u_1 = \frac{(1 + bH_1)u_1 u_2 - H_1 u_1}{-(1 + bH_1)u_2 + b},$$

which is equivalent to

$$\tau \frac{du_2}{d\tau} = \frac{(1 + bH_1)(u_2^2 - u_1^2) - (b + H_1)u_2 + 1}{-(1 + bH_1)u_2 + b}, \quad (180)$$

$$\tau \frac{du_1}{d\tau} = \frac{2(1 + bH_1)u_1 u_2 - (b + H_1)u_1}{-(1 + bH_1)u_2 + b}.$$

To the second-order system (180), we put in correspondence the nonautonomous first-order equation

$$\frac{du_2}{du_1} = \frac{1 - (1 + bH_1)(u_1^2 - u_2^2) - (b + H_1)u_2}{2(1 + bH_1)u_1 u_2 - (b + H_1)u_1}, \quad (181)$$

which can be easily reduce to the exact-differential form:

$$d \left(\frac{(1 + bH_1)(u_2^2 + u_1^2) - (b + H_1)u_2 + 1}{u_1} \right) = 0. \quad (182)$$

Thus, Eq. (181) has the following first integral:

$$\frac{(1 + bH_1)(u_2^2 + u_1^2) - (b + H_1)u_2 + 1}{u_1} = C_1 = \text{const}, \quad (183)$$

which in the original variables has the form

$$\frac{(1 + bH_1)(w_4^2 + w_3^2) - (b + H_1)w_4 \sin \alpha + \sin^2 \alpha}{w_3 \sin \alpha} = C_1 = \text{const}. \quad (184)$$

Remark 3. Consider system (168)–(170) with variable dissipation with zero mean, which becomes conservative for $b = H_1$:

$$\begin{aligned} \alpha' &= -(1 + b^2)w_4 + b \sin \alpha, \\ w_4' &= \sin \alpha \cos \alpha - (1 + b^2)w_3^2 \frac{\cos \alpha}{\sin \alpha} - bw_4 \cos \alpha, \\ w_3' &= (1 + b^2)w_3w_4 \frac{\cos \alpha}{\sin \alpha} - bw_3 \cos \alpha. \end{aligned} \quad (185)$$

It possesses the following two analytic first integrals:

$$(1 + b^2)(w_4^2 + w_3^2) - 2bw_4 \sin \alpha + \sin^2 \alpha = C_1^* = \text{const}, \quad (186)$$

$$w_3 \sin \alpha = C_2^* = \text{const}. \quad (187)$$

Obviously, the ratio of the two first integrals (186), (187) is also a first integral of system (185). But for $b \neq H_1$ none of the functions

$$(1 + bH_1)(w_4^2 + w_3^2) - (b + H_1)w_4 \sin \alpha + \sin^2 \alpha \quad (188)$$

and (187) is a first integral of system (168)–(170). However, the ratio of the functions (188), (187) is a first integral of system (168)–(170) for any b, H_1 .

We find the explicit form of the additional first integral of the third-order system (168)–(170). First, we transform the invariant relation (183) for $u_1 \neq 0$ as follows:

$$\begin{aligned} \left(u_2 - \frac{b + H_1}{2(1 + bH_1)}\right)^2 + \left(u_1 - \frac{C_1}{2(1 + bH_1)}\right)^2 &= \\ &= \frac{(b - H_1)^2 + C_1^2 - 4}{4(1 + bH_1)^2}. \end{aligned} \quad (189)$$

We see that the parameters of this invariant relation must satisfy the condition

$$(b - H_1)^2 + C_1^2 - 4 \geq 0, \quad (190)$$

and the phase space of system (168)–(170) is stratified into the family of surfaces defined by Eq. (189).

Thus, due to relation (183), the first equation of system (180) has the form

$$\begin{aligned} \tau \frac{du_2}{d\tau} &= \\ &= \frac{2(1 + bH_1)u_2^2 - 2(b + H_1)u_2 + 2 - C_1U_1(C_1, u_2)}{b - (1 + bH_1)u_2}, \end{aligned} \quad (191)$$

where

$$U_1(C_1, u_2) = \frac{1}{2(1 + bH_1)} \{C_1 \pm U_2(C_1, u_2)\}, \quad (192)$$

$$U_2(C_1, u_2) = \sqrt{C_1^2 - 4(1 + bH_1)(1 - (b + H_1)u_2 + (1 + bH_1)u_2^2)},$$

and the integration constant C_1 is defined by condition (190).

Therefore, the quadrature for the search for an additional first integral of system (168)–(170) becomes

$$\begin{aligned} \int \frac{d\tau}{\tau} &= \\ &= \int \frac{(b - (1 + bH_1)u_2)du_2}{2A^1 - C_1 \{C_1 \pm U_2(C_1, u_2)\} / (2(1 + bH_1))}, \end{aligned} \quad (193)$$

$$A^1 = 1 - (b + H_1)u_2 + (1 + bH_1)u_2^2.$$

Obviously, the left-hand side (up to an additive constant) is equal to

$$\ln |\sin \alpha|. \quad (194)$$

If

$$u_2 - \frac{b + H_1}{2(1 + bH_1)} = r_1, \quad b_1^2 = (b - H_1)^2 + C_1^2 - 4, \quad (195)$$

then the right-hand side of Eq. (193) becomes

$$\begin{aligned} &-\frac{1}{4} \int \frac{d(b_1^2 - 4(1 + bH_1)r_1^2)}{(b_1^2 - 4(1 + bH_1)r_1^2) \pm C_1 \sqrt{b_1^2 - 4(1 + bH_1)r_1^2} -} \\ &\quad -(b - H_1)(1 + bH_1) \times \\ &\quad \times \int \frac{dr_1}{(b_1^2 - 4(1 + bH_1)r_1^2) \pm C_1 \sqrt{b_1^2 - 4(1 + bH_1)r_1^2}} = \\ &= -\frac{1}{2} \ln \left| \frac{\sqrt{b_1^2 - 4(1 + bH_1)r_1^2}}{C_1} \pm 1 \right| \pm \frac{b - H_1}{2} I_1, \end{aligned} \quad (196)$$

where

$$\begin{aligned} I_1 &= \int \frac{dr_3}{\sqrt{b_1^2 - r_3^2}(r_3 \pm C_1)}, \\ r_3 &= \sqrt{b_1^2 - 4(1 + bH_1)r_1^2}. \end{aligned} \quad (197)$$

In the calculation of integral (197), the following three cases are possible:

I. $|b - H_1| > 2$.

$$\begin{aligned} I_1 &= -\frac{1}{2\sqrt{(b - H_1)^2 - 4}} \times \\ &\times \ln \left| \frac{\sqrt{(b - H_1)^2 - 4} + \sqrt{b_1^2 - r_3^2}}{r_3 \pm C_1} \pm \frac{C_1}{\sqrt{(b - H_1)^2 - 4}} \right| + \\ &\quad + \frac{1}{2\sqrt{(b - H_1)^2 - 4}} \times \\ &\times \ln \left| \frac{\sqrt{(b - H_1)^2 - 4} - \sqrt{b_1^2 - r_3^2}}{r_3 \pm C_1} \mp \frac{C_1}{\sqrt{(b - H_1)^2 - 4}} \right| + \\ &\quad + \text{const}. \end{aligned} \quad (198)$$

II. $|b - H_1| < 2$.

$$I_1 = \frac{1}{\sqrt{4 - (b - H_1)^2}} \arcsin \frac{\pm C_1 r_3 + b_1^2}{b_1(r_3 \pm C_1)} + \text{const}. \quad (199)$$

III. $|b - H_1| = 2$.

$$I_1 = \mp \frac{\sqrt{b_1^2 - r_3^2}}{C_1(r_3 \pm C_1)} + \text{const}. \quad (200)$$

Returning to the variable

$$r_1 = \frac{w_3}{\sin \alpha} - \frac{b + H_1}{2(1 + bH_1)}, \quad (201)$$

we have the following final form of I_1 :

I. $|b - H_1| > 2$.

$$I_1 = -\frac{1}{2\sqrt{(b - H_1)^2 - 4}} \times$$

$$\begin{aligned} & \times \ln \left| \frac{\sqrt{(b-H_1)^2 - 4} \pm 2(1+bH_1)r_1}{\sqrt{b_1^2 - 4(1+bH_1)^2 r_1^2 \pm C_1}} \pm \frac{C_1}{\sqrt{(b-H_1)^2 - 4}} \right| + \\ & \quad + \frac{1}{2\sqrt{(b-H_1)^2 - 4}} \times \\ & \times \ln \left| \frac{\sqrt{(b-H_1)^2 - 4} \mp 2(1+bH_1)r_1}{\sqrt{b_1^2 - 4(1+bH_1)^2 r_1^2 \pm C_1}} \mp \frac{C_1}{\sqrt{(b-H_1)^2 - 4}} \right| + \\ & \quad + \text{const.} \end{aligned} \tag{202}$$

II. $|b - H_1| < 2$.

$$\begin{aligned} I_1 &= \frac{1}{\sqrt{4 - (b - H_1)^2}} \times \\ & \times \arcsin \frac{\pm C_1 \sqrt{b_1^2 - 4(1+bH_1)^2 r_1^2} + b_1^2}{b_1(\sqrt{b_1^2 - 4(1+bH_1)^2 r_1^2} \pm C_1)} + \text{const.} \end{aligned} \tag{203}$$

III. $|b - H_1| = 2$.

$$I_1 = \mp \frac{2(1+bH_1)r_1}{C_1(\sqrt{b_1^2 - 4(1+bH_1)^2 r_1^2} \pm C_1)} + \text{const.} \tag{204}$$

Thus, we have found an additional first integral for the third-order system (168)–(170) and we have the complete set of first integrals that are transcendental functions of their phase variables.

Remark 4. Formally, in the expression of the found first integral, we must substitute instead of C_1 the left-hand side of the first integral (183).

Then the obtained additional first integral has the following structure (similar to the transcendental first integral from planar dynamics):

$$\ln |\sin \alpha| + G_2 \left(\sin \alpha, \frac{w_4}{\sin \alpha}, \frac{w_3}{\sin \alpha} \right) = C_2 = \text{const.} \tag{205}$$

Thus, to integrate the eighth-order system (168)–(173), we have already found two independent first integrals. For the complete integration, as was mentioned above, it suffices to find one first integral for each (potentially separated) system (171), (172), and an additional first integral that attaches Eq. (173).

To find a first integral of each (potentially separated) system (171), (172), we put in correspondence the following nonautonomous first-order equation:

$$\frac{dw_s}{d\beta_s} = \frac{1 + w_s^2 \cos \beta_s}{w_s \sin \beta_s}, s = 1, 2. \tag{206}$$

After integration we obtain the required invariant relation

$$\frac{\sqrt{1 + w_s^2}}{\sin \beta_s} = C_{s+2} = \text{const}, s = 1, 2. \tag{207}$$

Further, to obtain an additional first integral that attaches Eq. (173), to Eqs. (173) and (171) we put in correspondence the following nonautonomous equation:

$$\frac{dw_2}{d\beta_3} = -(1 + w_2^2) \cos \beta_2. \tag{208}$$

Since

$$C_4 \cos \beta_2 = \pm \sqrt{C_4^2 - 1 - w_2^2}, \tag{209}$$

by (207), we have

$$\frac{dw_2}{d\beta_3} = \mp \frac{1}{C_4} (1 + w_2^2) \sqrt{C_4^2 - 1 - w_2^2}. \tag{210}$$

Integrating this relation, we arrive at the following quadrature:

$$\mp(\beta_3 + C_5) = \int \frac{C_4 dw_2}{(1 + w_2^2) \sqrt{C_4^2 - 1 - w_2^2}}, \tag{211}$$

$C_5 = \text{const.}$

Integration leads to the relation

$$\mp \text{tg}(\beta_3 + C_5) = \frac{C_4 w_2}{\sqrt{C_4^2 - 1 - w_2^2}}, C_5 = \text{const.} \tag{212}$$

Finally, we have the following additional first integral that attaches Eq. (173):

$$\text{arctg} \frac{C_4 w_2}{\sqrt{C_4^2 - 1 - w_2^2}} \pm \beta_3 = C_5, C_5 = \text{const.} \tag{213}$$

Thus, in the case considered, the system of dynamical equations (17)–(21), (24)–(33) under condition (148) has twelve invariant relations: the analytic nonintegrable constraint of the form (38), the cyclic first integrals of the form (36) and (37), the first integral of the form (184), the first integral expressed by relations (198)–(205), which is a transcendental function of the phase variables (in the sense of complex analysis) expressed through a finite combination of functions, and the transcendental first integrals of the form (207) and (213).

Theorem 3. *System (17)–(21), (24)–(33) under conditions (38), (148), (37) possesses twelve invariant relations (complete set); five of them are transcendental functions from the point of view of complex analysis. All relations are expressed through finite combinations of elementary functions.*

D. Topological analogies

Consider the following seventh-order system:

$$\begin{aligned} & \ddot{\xi} + (b_* - H_{1*}) \dot{\xi} \cos \xi + \sin \xi \cos \xi - \\ & - [\eta_1^2 + \eta_2^2 \sin^2 \eta_1 + \eta_3^2 \sin^2 \eta_1 \sin^2 \eta_2] \frac{\sin \xi}{\cos \xi} = 0, \\ & \ddot{\eta}_1 + (b_* - H_{1*}) \dot{\eta}_1 \cos \xi + \dot{\xi} \dot{\eta}_1 \frac{1 + \cos^2 \xi}{\cos \xi \sin \xi} - \\ & - (\eta_2^2 + \eta_3^2 \sin^2 \eta_2) \sin \eta_1 \cos \eta_1 = 0, \\ & \ddot{\eta}_2 + (b_* - H_{1*}) \dot{\eta}_2 \cos \xi + \dot{\xi} \dot{\eta}_2 \frac{1 + \cos^2 \xi}{\cos \xi \sin \xi} + \\ & + 2\dot{\eta}_1 \dot{\eta}_2 \frac{\cos \eta_1}{\sin \eta_1} - \eta_3^2 \sin \eta_2 \cos \eta_2 = 0, \\ & \ddot{\eta}_3 + (b_* - H_{1*}) \dot{\eta}_3 \cos \xi + \dot{\xi} \dot{\eta}_3 \frac{1 + \cos^2 \xi}{\cos \xi \sin \xi} + \\ & + 2\dot{\eta}_1 \dot{\eta}_3 \frac{\cos \eta_1}{\sin \eta_1} + 2\dot{\eta}_2 \dot{\eta}_3 \frac{\cos \eta_2}{\sin \eta_2} = 0, \\ & b_* > 0, H_{1*} > 0. \end{aligned} \tag{214}$$

This system describes a five-dimensional pendulum in a field of a running medium for which the moment of forces depends on the angular velocity, i.e., a mechanical system in a nonconservative field. Generally speaking, the order of this

system must be equal to 8, but the phase variable η_3 is a cyclic variable, which leads to the stratification of the phase space and reduced the order of the system.

The phase space of this system is the tangent bundle

$$TS^3\{\dot{\xi}, \dot{\eta}_1, \dot{\eta}_2, \dot{\eta}_3, \xi, \eta_1, \eta_2, \eta_3\} \quad (215)$$

of the four-dimensional sphere $S^4\{\xi, \eta_1, \eta_2, \eta_3\}$. The equation that transforms system (140) into the system on the tangent bundle of the three-dimensional sphere

$$\dot{\eta}_3 \equiv 0, \quad (216)$$

and the equations of great circles

$$\dot{\eta}_1 \equiv 0, \quad \dot{\eta}_2 \equiv 0, \quad \dot{\eta}_3 \equiv 0 \quad (217)$$

define families of integral manifolds.

It is easy to verify that system (214) is equivalent to the dynamical system with variable dissipation with zero mean on the tangent bundle (215) of the four-dimensional sphere. Moreover, the following theorem holds.

Theorem 4. *System (17)–(21), (24)–(33) under conditions (38), (148), (37) is equivalent to the dynamical system (214).*

Indeed, it suffices to set $\alpha = \xi$, $\beta_1 = \eta_1$, $\beta_2 = \eta_2$, $\beta_3 = \eta_3$, $b = -b_*$, $H_1 = -H_{1*}$.

VI. CONCLUSION

In the previous studies of the author, the problems on the motion of the lower-dimensional solid were already considered in a nonconservative force field in the presence of the following force. This study opens a new cycle of works on integration of a multidimensional solid in the nonconservative field because previously, as was already specified we considered only such motions of a solid when the field of external forces was the potential.

ACKNOWLEDGMENT

This work was supported by the Russian Foundation for Basic Research, project no. 12-01-00020-a.

REFERENCES

- [1] M. V. Shamolin, *Methods of analysis of dynamical systems with various dissipation in rigid body dynamics*, Moscow, Russian Federation: Ekzamen, 2007.
- [2] M. V. Shamolin, *Some questions of the qualitative theory of ordinary differential equations and dynamics of a rigid body interacting with a medium*, Journal of Mathematical Sciences, Vol. 110, No. 2, 2002, p. 2526–2555.
- [3] M. V. Shamolin, *Foundations of differential and topological diagnostics*, Journal of Mathematical Sciences, Vol. 114, No. 1, 2003, p. 976–1024.
- [4] M. V. Shamolin, *New integrable cases and families of portraits in the plane and spatial dynamics of a rigid body interacting with a medium*, Journal of Mathematical Sciences, Vol. 114, No. 1, 2003, p. 919–975.
- [5] M. V. Shamolin, *Classes of variable dissipation systems with nonzero mean in the dynamics of a rigid body*, Journal of Mathematical Sciences, Vol. 122, No. 1, 2004, p. 2841–2915.
- [6] M. V. Shamolin, *Structural stable vector fields in rigid body dynamics*, Proc. of 8th Conf. on Dynamical Systems (Theory and Applications) (DSTA 2005), Lodz, Poland, Dec. 12–15, 2005; Tech. Univ. Lodz, 2005, Vol. 1, p. 429–436.
- [7] M. V. Shamolin, *The cases of integrability in terms of transcendental functions in dynamics of a rigid body interacting with a medium*, Proc. of 9th Conf. on Dynamical Systems (Theory and Applications) (DSTA 2007), Lodz, Poland, Dec. 17–20, 2007; Tech. Univ. Lodz, 2007, Vol. 1, p. 415–422.
- [8] M. V. Shamolin, *Methods of analysis of dynamic systems with various dissipation in dynamics of a rigid body*, ENOC-2008, CD-Proc., June 30–July 4, 2008, Saint Petersburg, Russia, 6 p.
- [9] M. V. Shamolin, *Some methods of analysis of the dynamical systems with various dissipation in dynamics of a rigid body*, PAMM (Proc. Appl. Math. Mech.), **8**, 10137–10138 (2008) / DOI 10.1002/pamm.200810137.
- [10] M. V. Shamolin, *Dynamical systems with variable dissipation: methods and applications*, Proc. of 10th Conf. on Dynamical Systems (Theory and Applications) (DSTA 2009), Lodz, Poland, Dec. 7–10, 2009; Tech. Univ. Lodz, 2009, p. 91–104.
- [11] M. V. Shamolin, *New cases of integrability in dynamics of a rigid body with the cone form of its shape interacting with a medium*, PAMM (Proc. Appl. Math. Mech.), **9**, 139–140 (2009) / DOI 10.1002/pamm.200910044.
- [12] M. V. Shamolin, *The various cases of complete integrability in dynamics of a rigid body interacting with a medium*, Multibody Dynamics, ECCOMAS Thematic Conf. Warsaw, Poland, 29 June–2 July 2009, CD-Proc.; Polish Acad. Sci., Warsaw, 2009, 20 p.
- [13] M. V. Shamolin, *Dynamical systems with various dissipation: background, methods, applications* // CD-Proc. of XXXVIII Summer School-Conf. "Advances Problems in Mechanics" (APM 2010), July 1–5, 2010, St. Petersburg (Repino), Russia; St. Petersburg, IPME, 2010, p. 612–621.
- [14] M. V. Shamolin, *Integrability and nonintegrability in terms of transcendental functions in dynamics of a rigid body*, PAMM (Proc. Appl. Math. Mech.), **10**, 63–64 (2010) / DOI 10.1002/pamm.201010024.
- [15] M. V. Shamolin, *Cases of complete integrability in transcendental functions in dynamics and certain invariant indices*, CD-Proc. 5th Int. Sci. Conf. on Physics and Control PHYSCON 2011, Leon, Spain, September 5–8, 2011. Leon, Spain, 5 p.
- [16] M. V. Shamolin, *Variety of the cases of integrability in dynamics of a 2D-, 3D-, and 4D-rigid body interacting with a medium*, Proc. of 11th Conf. on Dynamical Systems (Theory and Applications) (DSTA 2011), Lodz, Poland, Dec. 5–8, 2011; Tech. Univ. Lodz, 2011, p. 11–24.
- [17] M. V. Shamolin, *Cases of integrability in dynamics of a rigid body interacting with a resistant medium*, CD-proc., 23th International Congress of Theoretical and Applied Mechanics, August 19–24, 2012, Beijing, China; Beijing, China Science Literature Publishing House, 2012, 2 p.
- [18] M. V. Shamolin, *Variety of the cases of integrability in dynamics of a 2D-, and 3D-rigid body interacting with a medium*, 8th ESMC 2012, CD-Materials (Graz, Austria, July 9–13, 2012), Graz, Graz, Austria, 2012, 2 p.

Fuzzy- multi agent hybrid system for decision support of consumers of energy from renewable sources

Otilia Dragomir, Florin Dragomir, Eugenia Minca

Abstract—This paper purposes an intelligent decision support system for low voltage grids with distributed power generation from renewable energy sources (InDeSEn). The added value of this innovative software tool consists in integrating decision theory and artificial intelligence concepts in monitoring, supervising, forecasting and control actions, allowing prosumers of energy from renewable sources: to control electricity consumption of the used devices, to reduce their monthly bills, carbon emissions, energy demand during peak periods and to use more efficient the energy from renewable energy sources. The application is accessible to users anytime, through the web interface attached, providing both: information for general use and technical information.

Keywords— fuzzy logic control, fuzzy- multi agent, intelligent decision support system, renewable energy, smart grid

I. INTRODUCTION

SMART grids could be described as an upgraded electricity network to which two-way digital communication between supplier and consumer, intelligent metering and monitoring systems have been added. Intelligent metering is usually an inherent part of them.

The benefits of smart grids are widely acknowledged. Smart grids can manage direct interaction and communication among consumers, households or companies, other grid users and energy suppliers. They open up unprecedented possibilities for consumers to directly control and manage their individual consumption patterns, providing, in turn, strong incentives for efficient energy use if combined with time-dependent electricity prices. Improved and more targeted management of the grid translates into a grid that is more secure and cheaper to operate.

Smart grids will be the backbone of the future decarbonized power system. They will enable the integration of vast amounts of both on-shore and off-shore renewable energy and electric vehicles while maintaining availability for conventional power generation and power system adequacy.

European context. The current energy policy of the

European Union (EU) considers the security of supply, competitiveness and sustainability as central goals. In order to achieve these targets, through European strategies [1] are imposed a series of constraints ("objective 20-20-20"): 20% reduction in emissions of greenhouse gases compared to 1990, providing 20% of entire EU energy consumption by renewable energy sources (RES) and a reduction in energy use by 20% compared to a similar scenario in which no action regarding sustainability has been taken. To achieve these objectives and generate a "sustainable growth" a policy of encouraging distributed generation from RES (PD- RES), such as solar power must be followed.

Intense concerns at European level regarding the PD- RES were materialized by setting up a giant cluster of projects called Integration of Renewable Energy Sources and Distributed Generation into the European Electricity Grid - IRED cluster [2]. The studies that were conducted by this consortium highlighted the need for an energy management system at micro to macro level, the existing control strategies not being always successfully applied.

Romanian context. The share of RES in the electricity production in Romania is currently 17.8%. EU has set for Romania a 24% target for energy generation from RES by 2020, but there have also been identified needs for investments and large operating costs as main barriers for the successful implementation of an increased generating capacity. Compatibility with EU objectives in the field of clean energy and national levels is achieved through regional policy [3]. European policies had a national resonance since 2003, when the draft for the project strategy for the use of renewable energy [4] was proposed.

Two new important trends on the Romanian national energy market are to be noticed: firstly, the consumers involvement in the complex process of efficient management of the energy from RES and secondly the increased attention, paid to both: the technical plan and to the organizational and economical plan, for the energy production from RES.

Unfortunately Romania follows a centralized approach of the regional policy and although the country is covered adequately with electricity networks and the potential development for RES is high, its aging infrastructure (30% of it was built in the 1960s) causes significant losses along the energy supply chain.

Otilia Dragomir, Florin Dragomir and Eugenia Minca are with Automation, Computer Science and Electrical Engineering Department, Valahia University of Targoviste, Targoviste, Romania, (e-mail: drg_otilia@yahoo.com, drg_florin@yahoo.com, eugenia.minca@gmail.com)

In addition, for a large number of energy resources, the current energy systems are hardly scalable. The European Commission believes that the current energy infrastructure is inadequate to connect and serve all of Europe and recognizes the challenges [1].

In this global and local context, this paper purposes an intelligent decision support system for low voltage grids with distributed power generation from renewable energy sources (InDeSEn).

The added value and the originality of this innovative software tool consists in integrating decision theory and artificial intelligence concepts in monitoring, supervising, forecasting and control actions, allowing prosumers (producers and consumers of energy from RES): to control electricity consumption of the used devices, to reduce their monthly bills, carbon emissions, energy demand during peak periods and to use more efficient the energy from renewable energy sources. The application is accessible to users anytime, through the web interface attached, providing both: information for general use and technical information.

First of all, it is presented a state of the art of the main interest directions in the field as well as the problematic in this area. The second section describes the conceptual framework of the InDeSEn, from hardware and software point of view. The third part is dedicated to the implementation and the results analysis of the software tool. Precisely, the identified criteria used for choosing a decision support system, made in fuzzy logic and intelligent agents context are tested on a hybrid system, fuzzy- multi-agent. The tool design of the proposed hybrid system is built using MATLAB. The article ends with conclusions and work in progress.

II. INDeSEn FRAMEWORK

A. State of the art and the problematic

The problematic in the area of low voltage grids with distributed power generation from renewable energy sources is vast.

The main interest directions in the field, identified by reviewing the scientific literature are: to increase conversion efficiency [5] and [6], to identify new opportunities, complementary to the limitations of existing solutions [7], to limit the variability of the power production through integration of RES [8], to balance the insufficient number of providers reported to market needs [9], to supply with energy the increasing number of consumers [10], to provide and promote sustainable buildings and housing projects [11], to identify the new renewable energy sources in order to reduce risks, social and economic costs [12], to understand the effects associated with the use of RES and the increasing complexity of systems [13], to identify measures and solutions for ensuring the sustainability of renewable resources [14], to resize the impact of renewable energy for the final beneficiaries, in terms of cost / benefit ratio [15] and [16], to highlight the importance of being efficiently informed in order to make better decisions about energy options [17].

Following the analysis performed, the electrical network stability isn't considered as a service ensured in any country. However it is a typical problem for PD-RES microgrid connection to national electricity network because the system must be robust and the parameters oscillations must be amortized.

In addition, in networks distribution isn't taken in consideration the balance between productions and consume from RES as viable solution, for off-grid functioning of microgrids,

To support this interest, worldwide there are **only a few tools** which are able to perform dynamic assessment of electrical networks; none of them take into account the characteristics and the specificities of PD- RES microgrid.

The GARPUR (www.garpur-project.eu/) project designs, develops, assesses and evaluates new reliability criteria for transmission grids, based on a probabilistic approach. PEGASE project (<http://fp7-pegase.eu/>) aims at improving the performance of various computational tools for transmission network management and to propose novel routes about the sharing of dynamic models. REALISEGRID develops a set of criteria, metrics, methods and tools to assess how the transmission infrastructure should be optimally developed to support the achievement of a reliable, competitive and sustainable electricity supply in the EU (realisegrid.rse-web.it/).

In the next sections we will presents the main characteristics of the proposed solution, from the hardware and software point of view.

B. InDeSEn framework

To overcome the mentioned barriers, InDeSEn proposes to re-define the role of consumer of energy in "prosumer" in the context of a reorganized decentralized energy market, now reported to intelligent grids (smart grids). Integration of interactive technologies in a decision support system for the PD- RES microgrids energy management optimizes: functioning from an economical point of view, active control of distributed generation, controlled consumption, loading the storage equipment.

The innovative aspects of the proposal consist in: implementing the systems for producing electricity with RES and identifying the operating conditions in the public low-voltage grid, creating a database regarding the dynamic mode operation of distributed energy systems, and developing the functional models of low voltage hybrid networks that integrate renewable energy sources.

The original features of the project are: implementing a dynamic software platform that can handle real-time data of the PD-SER microgrids operation, improving the specific algorithms used in power management of the microgrids by integrating intelligent agents of the elements of artificial intelligence in order to optimize decisions and to assist consumers-producers, creating a database with information associated with different types of RES, using advanced technologies for online data acquisition and optimizing online

communication of this type, using friendly interfaces such as display panels, used for working with virtual instruments, harmonizing the Romanian technological solutions with the international operation on-grid or off-grid.

C. InDeSen architecture: software and hardware

InDeSen integrates the following software and hardware components:

- software components: monitoring module; diagnosis module; prediction module; knowledge base; decision module; control module.
- hardware components: energy monitoring devices and power control units

The main functions implemented in the software application – the intelligent support system are: monitoring, diagnosis and prediction. (Fig. 1).

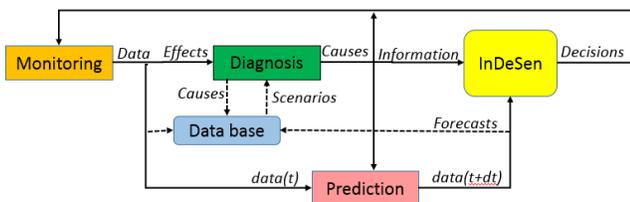


Fig. 1 Functions of InDeSen

Monitoring. At this level, the significant parameters that characterize the electrical energy management the electrical energy quality and the environment factors (specific to the primary resources of renewable energy) are measured in real time. These values are stored in a data base. Data resulted from the monitoring represent, at this level, inputs. In addition, by primary processing is determined the performance report of the PV panels and the balance between the momentary and the rated power of the local network.

Diagnosis. In order to assess the operational status of the installations and electricity consumption and identify the causes of the state of emergency or alarm in the network, within the proposed scheme will be implemented the diagnosis function. The diagnosis function evaluates the functioning state of the generating/consuming installations and identifies the damage causes or an alarm in the network. These data will influence the dynamic of the trust of each renewable energy producer and also the price of the green energy so as to develop clean technologies.

Prediction. Once we have an updated knowledge base, enriched with new scenarios (input/output data) and with a minimal intelligence level established in the design stage, the prediction function can be used. This one provide short, medium or long time forecasts in respect of the consuming process evolution and energy generation. This information will be used by the Intelligent Decision Support System (INDES) in analyzing the forward decisions. The tendency to determinate the electrical energy generation is based on the cycle (day/night, winter/summer) of the environment factors (solar radiation, temperature, wind speed) and on the decrease

of the efficiency of the installations because of the generators use. The results of the prediction have an important impact not only on the action plan but also on the financial evaluation of the green energy.

The architecture of the hardware infrastructure supporting the decisional system is presented in the block diagram from the Fig. 2.

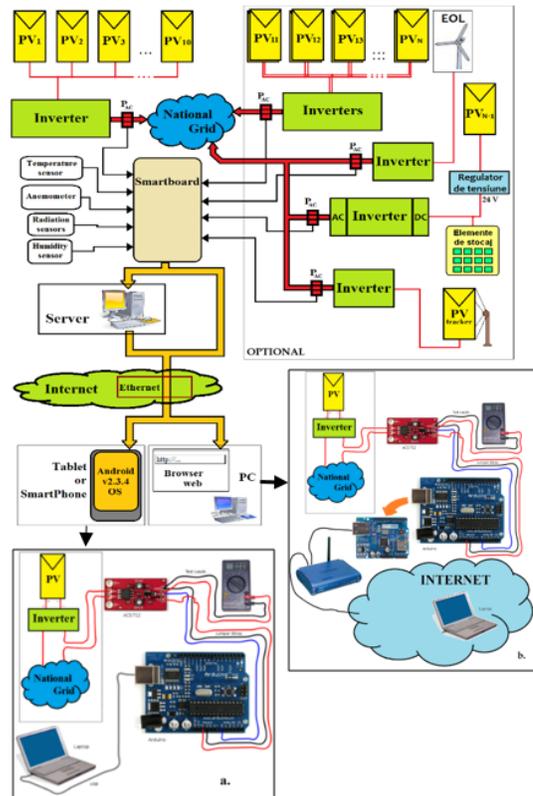


Fig. 2 The hardware infrastructure

Thanks to the initial configuration and components provided in the architecture, the platform manifests flexibility and adaptability in operation. Any electricity generator can be integrated with the solar central, wind, hydroelectric plants or diesel generators. The key characteristics of the platform's modules structure are: photovoltaic panels and wind power station; storage unit; inverters; PC with DAQ and software application and loads.

The adaptive character of the decision support system will permanently ensure low energy lose correlated with the new scenarios that occur in the environment or at the consumers. The damages repair will provide a higher utile energy in the local network, a low energetic transfer from RNE and will minimize the losses in the energy transport. The implemented optimization algorithm will take into account the necessary energy in the case of isolated functioning (islands of energy). Using artificial intelligence techniques in power management actions in a low voltage grid through a software tool is a new technique and also expresses the complexity of the proposed solution.

III. INDeSen AS FUZZY- MULTI AGENT HYBRID SYSTEM

A. System design

In this section we will present only a part of the application InSeSen, due to the length restrictions imposed by the publisher. It is about the design a fuzzy multi- agent system able to assist the users' decisions.

The exploration and the assessment of criteria used for choosing a decision support system are made in fuzzy logic and intelligent agents context. In this respect, firstly are presented the criteria used for choosing the best parameters, in relation with each step of the tool design. Precisely, the identified criteria used for choosing a decision support system, made in fuzzy logic and intelligent agents context are tested on a hybrid system, fuzzy- multi-agent. The tool design of the proposed hybrid system is build using MATLAB.

The flow diagram showing the steps of tool design is presented in the fig. 3. It is related only to the temperature sensors information modeling and processing. The application InDeSen also integrates modules for luminance, air quality, humidity, noise and sound level, safety, movement etc.

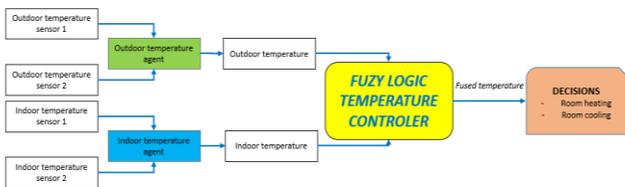


Fig. 3 The flow diagram

The measured data monitored by temperature sensors, located indoor and outdoor are the main sources for a deeper understanding of the system behavior and environmental status.

The tool has a **hierarchical structure**, meaning that information is firstly selected with a heuristic approach based on fusion of physical sensors and implemented using local intelligent agents (*outdoor temperature agent* and *indoor temperature agent*). This selection presumes that inputs are independent and give no priority or importance of the selected input variables.

On the second level of the hierarchical structure in placed the fuzzy logic temperature controller (FLC). This one can be viewed as an approach combining conventional precise mathematical control and humanlike decision-making.

The architecture of the temperature decision support module is presented in fig. 4. There are five primary distinctive panels integrated in the graphical user interface (GUI) for building the FLC corresponding with: the inputs/ outputs fuzzyfication and membership functions (MFs) editor, editing, building and viewing the fuzzy inference systems and the last one, the defuzzyfication and output surface viewer.

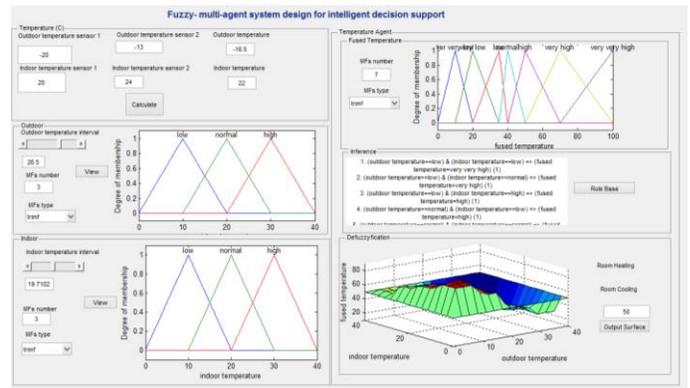


Fig. 4 The graphical user interface of the temperature decision support module

B. Fuzzyfication

The structure of a fuzzy controller is essentially the structure of a Mamdani technical fuzzy controller. It has two inputs (*outdoor temperature* and *indoor temperature*) and one output (*fused temperature*).

The GUI proposed give to the user the possibility to choose the number of the membership functions for each input/output and these ones shapes: triangular, Gaussian etc. The universe of discourse is: [0 40] Celsius degrees for *indoor temperature* and [-30 40] Celsius degrees for *outdoor temperature* and [0 100] Celsius degrees for *fused temperature*.

The users' choices are made by clicking in the GUI: the outdoor/indoor/ fused temperature interval slider, editing the MFs number and selecting the MFs type (Fig. 4).

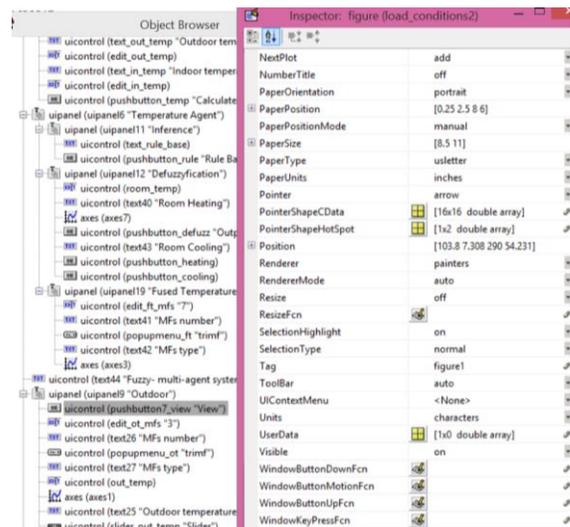


Fig. 4. The source code

Considering these options, in the source code of universe of discourse of each variable is divided into fuzzy regions. The method used is a trial and error one. The number of regions equals with (2N+1) [19], where N represent the number of MFs selected by user. For each fuzzy region is assigned the membership function with the shape imposed by user in the

GUI.

Fig. 5(a), 5(b), 5(c) show an example where *indoor temperature* is divided in 3 fuzzy regions, triangular shapes, the *outdoor temperature* is divided in 5 fuzzy regions, Gaussian shapes, and the *fused temperature* is divided in 7 fuzzy regions, triangular shapes.

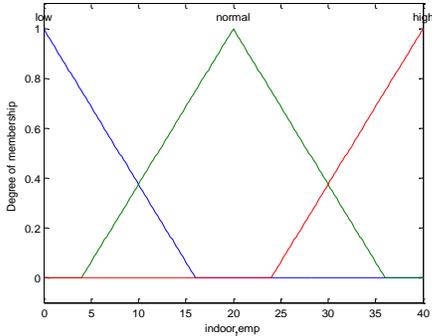


Fig. 5a. The indoor temperature input divided in 3 fuzzy regions, triangular shapes

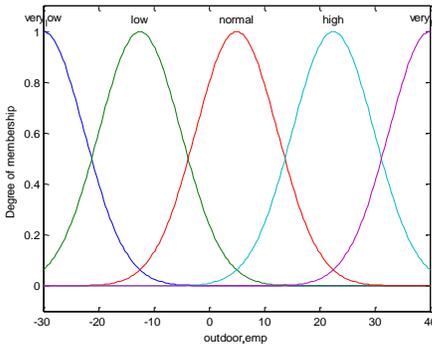


Fig. 5b. The outdoor temperature input divided in 5 fuzzy regions, Gaussian shapes

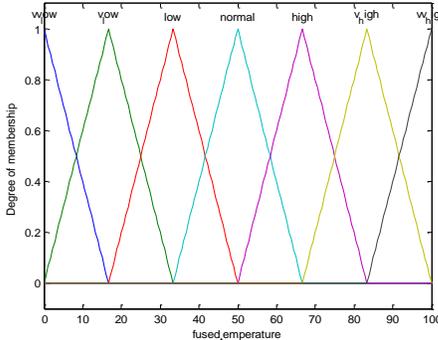


Fig. 5c. The fused temperature output divided in 7 fuzzy regions, triangular shapes

It was found that, in our case, the smallest number giving satisfactory results is 3 MFs, triangular shape for each input and 7 MFs, triangular shape for the output.

C. Inference system

Based on the descriptions of inputs and output variables, made in fuzzyfication panel, the rule base editor panel allows constructing the rule statements in *if-then* format.

We have built our FLC rule base in relation with Hagar’s

method [20]. The fuzzy rule decision table, in indexed format, used for FLC inference, using our reasoning and Hagar’s, is shown in the Table 1.

Table1. Rule base in indexed format

Indoor temp. MFs	Outdoor temp. MFs	Fused tem. MFs		Weight	Inference Method
		Dragomir	Hagras		
1	1	7	6	1	1
1	2	6	4	1	1
1	3	5	4	1	1
2	1	5	4	1	1
2	2	4	2	1	1
2	3	3	1	1	1
3	1	3	2	1	1
3	2	2	4	1	1
3	3	1	1	1	1

Legend: 1-Low , 2-Normal, 3-High for MFs associated with inputs and 1-Very Very Low, 2- Very Low, 3- Low, 4-Normal, 5- High, 6- Very High, 7- Very Very High for MFs associated with the FLC’s output. The AND inference method is denoted with 1 and the OR inference method is denoted with 2.

Having two inputs variables, 3 membership functions each, the number of rules equals with $3^2=9$. The representation in a matrix of inference rules facilitates checking there are not contradictory rules. The accepted modification without a deterioration of system performance implies modifying/removing the neighbors’ rules and replacing them with an average value.

D. Defuzzyfication

The fuzzy output: *fused temperature* is represented as a surface. It represents the decision to be made. It’s crisp format is obtained applying standard center of-gravity defuzzyfication method (Fig. 6a and 6b).

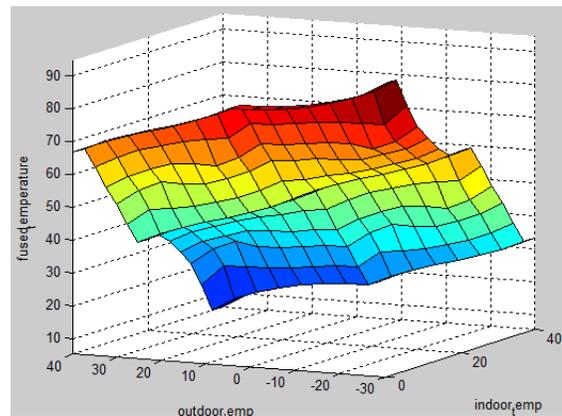


Fig. 6a. The fused temperature output divided, Dragomir approach

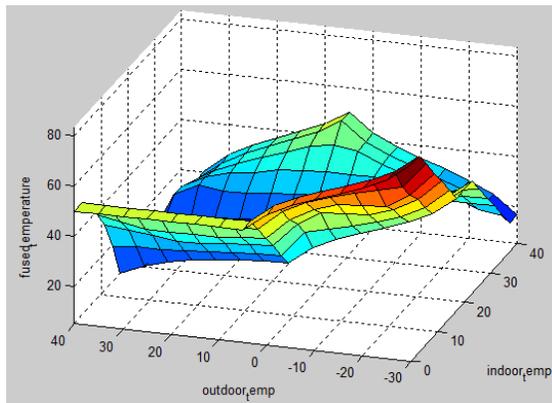


Fig. 6b. The fused temperature output divided, Hagra's approach

The tests results show differences between the analyzed methods: Dragomir's is an offline rule extraction method who needs an expert to supply his expertise formalized in a set of desired values. Hagra's method is an online learning and control method.

On the other hand for InDeSEn intelligent decision system, the GUI presented is only a local control module who will be integrated in a hierarchical structure with cascade control loop.

IV. CONCLUSIONS AND WORK IN PROGRESS

Novelty of this proposal consist in creating an intelligent decision support system for the low voltage grid with distributed power generation from renewable energy resources, allowing customers to control electricity consumption of the used devices, to reduce their monthly bills and carbon emissions, also reducing the demand during peak periods. InDeSEn informs customers about network status and consequently they will be able to program their appliances, as washing machines, during off-peak hours in a proactive manner.

The beneficiaries of InDeSEn software innovative platform are: green energy producers; universities, schools and research institutions - students and specialists in training, giving them the useful tool in their work in a new field of academic and professional specialization; individual electrical energy consumers; researchers from the energetic field and environment agencies; Ministry of Finance and Ministry of Environment and Sustainable Development which will establish and coordinate more effectively and faster exchange of information between policy makers and the public will be able to correlate better environmental protection plans with the development of RES and last, but not least the citizens as consumers that will be informed about the advantages of the systems with RES.

The work is still in progress in two directions: validation of the software in respect with the energy quality standards and integrating advanced control algorithms to energy distribution in smart grids with DP-RES.

ACKNOWLEDGMENT

This work was supported by a grant of the Romanian National Authority for Scientific Research, CNDS-UEFISCDI, project code PN-II-PT-PCCA-2011-3.2-1616.

REFERENCES

- [1] Strategia Europa 2020 , http://ec.europa.eu/europe2020/europe-2020-in-a-nutshell/priorities/index_ro.htm
- [2] IRED cluster, <http://www.ired-cluster.org/>
- [3] Politica de coeziune 2014-2020, http://ec.europa.eu/regional_policy/activity/energy/index_ro.cfm
- [4] Autoritatea Nationala de Reglementare in domeniul Energiei, <http://www.anre.ro/documente.php?id=393>
- [5] G. Boyle, Renewable Energy: Power for a Sustainable Future, Oxford University Press., 2012.
- [6] B. Sorensen, Renewable Energy, Fourth Edition: Physics, Engineering, Environmental Impacts, Economics & Planning, Academic Press., 2010.
- [7] G. Masters, Renewable and Efficient Electric Power Systems, New Jersey: Wiley-IEEE Press, 2004.
- [8] G. Boyle, Renewable Energy: Power for a Sustainable Future, Oxford University Press., 2012.
- [9] H. Kohl, The Development, In R. Wengenmayr, & T. Buhrke, Renewable Energy, pp. 4-14, Weinheim: Wiley-VCH, 2008.
- [10] A.V. Da Rosa A., Fundamentals of Renewable Energy Processes, Oxford: Academic Press, 2012.
- [11] W. Kemp, The Renewable Energy Handbook, Revised Edition: The Updated Comprehensive Guide to Renewable Energy and Independent Living, Tamworth: Aztex Press, 2009.
- [12] D. MacKay, Sustainable Energy - Without the Hot Air, Cambridge: UIT Cambridge Ltd., 2009.
- [13] M.S. Kaltschmitt, Renewable Energy: Technology, Economics and Environment, Berlin: Springer, 2010.
- [14] R. Bryce, Power Hungry: The Myths of "Green" Energy and the Real Fuels of the Future, New York: PublicAffairs, 2011
- [15] D. Chiras, The Homeowner's Guide to Renewable Energy: Achieving Energy Independence Through Solar, Wind, Biomass, and Hydropower, Gabriola Island: New Society Publishers, 2011.
- [16] G. Boyle, Renewable Electricity and the Grid, London, 2007.
- [17] R. Rapier, Power Plays: Energy Options in the Age of Peak Oil, Apress., 2012
- [18] Program CE – Altener – “Photovoltaic Training Courses for Candidate Countries”, proiect SolTrain
- [19] L. Wang, J. Mendel, “Generating fuzzy rules by learning from examples”, *IEEE Transactions on Systems Man, and Cybernetics*, vol. 22, no. 6, 1992.
- [20] H. Hagra's, V. Callaghan, M. Calley, G. Clarke, “A hierarchical fuzzy-genetic multiagent architecture for intelligent buildings online learning, adaptation and control”, *Information Science 150* , pp. 33- 57, 2003.

Otilia Dragomir is currently working as Assistant Professor, in Valahia University Targoviste Electrical Engineering, Electronics and Information Technology Faculty- FIEETI, Automation, Computer Science and Electrical Engineering Department-DAIIE. She had received PhD's degree in branch automation from Universite Franche Comte de Besancon, France and Politehnica University of Bucharest, Romania. She has twelve years of teaching Experience. Her specialization is in field of artificial intelligence and control systems.

Florin Dragomir is currently working as Assistant Professor, in Valahia University FIEETI- DAIIE He had received PhD's degree in branch automation from Politehnica University of Bucharest, Romania. He has twelve years of teaching Experience. His specialization is in field of microprocessors and real time systems.

Eugenia Minca is currently working as Professor, in Valahia University Targoviste, FIEETI- DAIIE. She had received PhD's degree in branch automation from Universite Franche Comte de Besancon, France. She has 20 years of Teaching Experience. Her specialization is in field of control systems.

A hybrid system for identification of elastic, isotropic thin plate parameters applying Lamb waves and artificial neural networks

Zenon Waszczyszyn and Ewa Pabisek

Abstract— A new hybrid computational system for material identification is presented, developed for the identification of homogeneous, elastic, isotropic plate parameters. Attention is focused on the construction of dispersion curves related to the Lamb wave propagation. The main idea of the hybrid system lies in the separation of two essential basic computational stages, corresponding to the direct or inverse analyses. This was made possible due to the use of the Lamb wave monitoring technique and artificial neural networks for the inverse analysis.

Keywords and acronyms – Artificial Neural Networks (ANN), Dispersion Curves (DCs), Hybrid Computational System (HCS), Guided Wave Monitoring (WM), Structure Health Monitoring (SHM).

I. INTRODUCTION

The presented paper is related to the area of Structure Health Monitoring/Masurement (SHM). The research and new technologies in this area are in current development, see [1-6]. An important role for SHM is played by systems which on the basis of monitoring or measurements can reflect the actual state (health) of structures. This permits the control of structures and early warning against failures or hazardous events. Non-destructive methods of structure examination and ‘on line’ methods of information transmission are especially valuable for SHM.

From among the non-destructive methods, the application of ultrasonic waves is worth emphasising for the evaluation of material properties and detection of various defects, see e.g. [7-11]. In the presented paper we discuss the application of the Lamb Waves (LWs) propagation in a thin, elastic and homogeneous plate. These waves are guided in the vibration plane and are propagated over a comparatively long distance.

The basis of the mathematical analysis are the Dispersion Curve (DCs), which can be derived as the function $k(f)$, where k is the wavenumber and f is the frequency of vibrations.

This work is supported by the Polish National Science Centre. Grant No. UMO-2011/01/B/ST8/ 07210, AGH No.18.18.130.384, “Structural Health Monitoring by means of inverse problem solution under uncertainty” is gratefully acknowledged for financial support.

Z. Waszczyszyn, Professor Emeritus, is with the Institute of Computational Civil Engineering, Cracow University of Technology, Poland, (phone number: 0048-12-628-2566, e-mail: zenwa@L5.pk.edu.pl).

E.Pabisek is Associate Professor with the Institute of Computational Civil Engineering, Cracow University of Technology, Poland, (e-mail: epabisek@L5.pk.edu.pl).

Unfortunately, the implicit formulation of DCs is analytically impossible so only numerical methods can be used, see, [1, 7].

An approach commonly applied to the identification of vector **par** components is to find experimentally a DC_{exp} and simulate numerically a corresponding DC_{sim} . Then a measure of distance $\|k_{sim} - k_{exp}\|$ is minimised applying various computer methods for the identification of searched parameters.

Such an approach can be called the classical approach, which is explored for the analysis of the Lamb dispersion curves. Apart from the application of general computer methods based on FEM, BEM and FDM, special, computationally efficient methods are worth mentioning, e.g. LISA/SIM (Local Interaction Simulation Approach/Sharp Interface Model) and EFIT (Elastodynamic Finite Integration Technique), see [3], application of spectral FEs [5], and SAM (Semi-Analytical Methods), cf. [9, 12, 13].

In the presented paper we are continuing the problem analysed in [13]. In this paper, the identification of ‘a priori’ unknown plate parameters is carried out. The parameters are adopted as the vector $\mathbf{par} = \{ E, \nu, \rho, d \}$ with the following components: E, ν – elasticity modulus and Poisson ratio of plate material, ρ, d – plate density and thickness, respectively.

In our paper [13] the classical method of numerical simulation of DCs was applied in a Hybrid Computational System (HMS). Such a system can be used for the analysis of more complicated problems for the design and control of structural defects as well as for the design of special devices improving the material parameters in intelligent structures

The system analysed in [11] was composed of two parts called Stage A and Stage B. Stage A corresponds to the direct analysis. It is related to the Guided Wave Measurement technique for plate testing and transformation of time signals into the transform space. As the input the approximate dispersion curve k_{exp} is simulated. Then Stage B is applied for the selection of vibration modes and numerical simulation.

The corresponding dispersion curve k_{sim} takes into account the Lamb equations.

The novelty of the presented paper lies in the introduction of Artificial Neural Networks (ANS) as a good tool for the inverse analysis. It was stated in many numerically analysed problems, see [7, 8, 13 -15], that ANS were applied also in SHM problems. The network, formulated for the inverse analysis, can be trained ‘off line’. Then it can be explored in HCS for the separation of the direct and inverse analyses. In such a way the “numerically costly” methods (referring to the number of operations) can be eliminated and iterative contact between Parts A and B is shortened to one substitution of the experimental results from Part A into B, see [16].

In the subsequent Sections of the presented paper the basics of the Lamb waves are briefly presented. Then the essential steps of GWM are discussed. The main attention is focused on hybrid identification of plate parameters based on the introduction of an ANN into part B of HCS.

Another novelty in the presented paper is the application of dimensionless approach to the computation of LDCs proposed by Armirkulova in [17], instead of other semi-analytical algorithms.

At the end of our paper a case study devoted to the identification of thin aluminium plate parameters, taken from [11], is discussed.

II. LAMB WAVES IN THE ANALYSIS OF ISOTROPIC, HOMOGENEOUS ELASTIC PLATES

A. Some basics on Lamb Waves (LWs)

The 3D waves in elastic solids can be reduced to 2D vibrations plane, in which the LWs are propagated, see Fig. 1. The plane of vibration (x_1, x_3) is perpendicular to the tested elastic plate midsurface of dimensions $L \times d, 2 \times b$, where plate thickness is $d = 2h$.

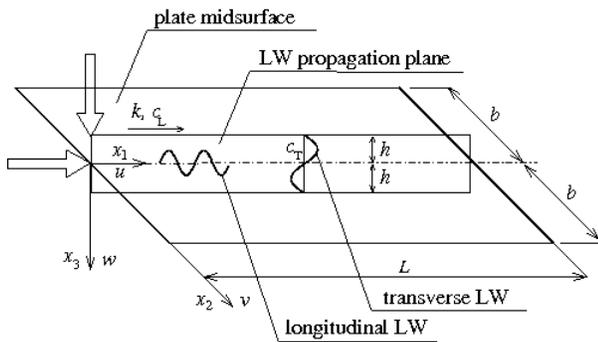


Fig. 1. Lamb wave propagation plane with longitudinal and transverse/shear LWs

The following assumptions are adopted to formulate the Lamb dispersion equations

- i) plain strain, i.e. the displacement v disappears

$$v(\mathbf{x}) = 0 \text{ for all the points } \mathbf{x},$$

- ii) upper and lower plate surfaces are stressless

$$\sigma_{33} = \sigma_{31} = 0 \text{ for } x_3 = \pm h. \tag{2}$$

The Lamb waves have symmetric and anti-symmetric modes of vibrations. They are related to the influence of internal stresses, which cause either the radial-in-plane or out-in-plane motions, see [7].

From the condition of nonzero value of the homogeneous algebraic equations related to the boundary conditions (1), the following equations can be derived, see e.g. [1, 7], relative either to the symmetric or anti-symmetric modes of vibrations:

$$(S): \frac{\tan \beta h}{\tan \alpha h} = -\frac{4k^2 \alpha \beta}{(k^2 - \beta^2)^2}, \tag{3S}$$

$$(A): \frac{\tan \beta h}{\tan \alpha h} = -\frac{(k^2 - \beta^2)^2}{4k^2 \alpha \beta}. \tag{3A}$$

B. Dimensionless Lamb equations

Instead of (3), Armirkulova formulated the dimensionless dispersion equations in the implicit form, see her MSc. Thesis [17]:

$$(S): (\xi^2 - x^2)^2 \sin x \cos y + 4xy \xi^2 \cos x \sin y = 0, \tag{4S}$$

$$(A): (\xi^2 - x^2)^2 \sin y \cos x + 4xy \xi^2 \cos y \sin x = 0, \tag{4A}$$

where the main dimensionless variables are used:

$$x = (\Omega^2 - \xi^2)^{1/2}, \quad y = (\Omega^2 \kappa^{-2} - \xi^2)^{1/2}. \tag{5}$$

In (6) the dimensionless frequency Ω , wavenumber ξ and velocity ratio κ are introduced:

$$\Omega = \omega h / c_T, \quad \xi = kh, \quad \kappa = c_L / c_T. \tag{6}$$

Function $\Omega(\xi)$ can be derived in an analytical form, but in further computations its inverse form $\xi(\Omega)$ is needed, see Fig. 2. The corresponding curve was formulated numerically for discrete points, owing to one-to-one correspondence of coordinates (ξ, Ω) .

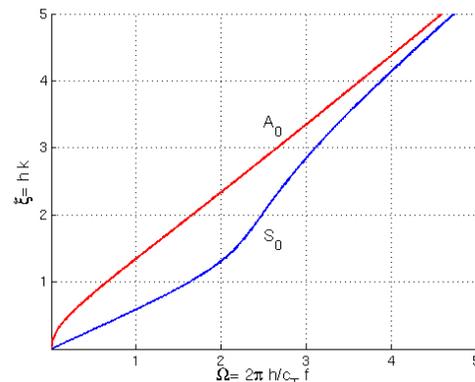


Fig.2. Dimensionless dispersion curves for Lamb modes A_0 and B_0

four essential steps can be pointed out, see [2]. The steps are sketched in Fig. 3, underlining Step IV, see also [2, 16].

III. FOUR ESSENTIAL STEPS IN HCS

In the Guided Wave Monitoring or Measurements (GWMs)

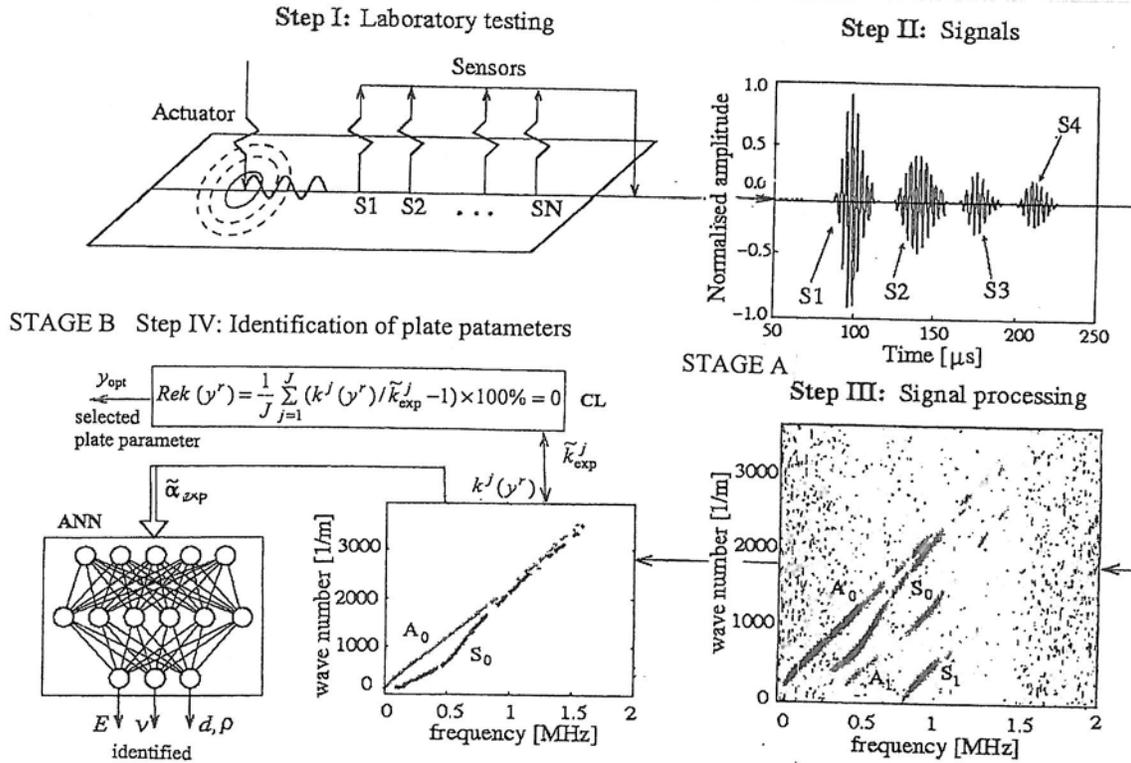


Fig. 3. Four essential steps for identification of plate parameters, applying Lamb wave measurement technique and identification procedures in step IV: CL – Classical approach, ANN – artificial neural network

IV. HYBRID COMPUTATIONAL SYSTEM

A. Some remarks on simulation and identification problems in mechanical systems

Simulation and identification problems of mechanical systems can be classified from the point of view of regression analysis, see [19]. Adopting the following scheme

$$\text{Input – Transformer – Output} \equiv \text{I – T – O}, \quad (7)$$

different regression problems can be formulated. The direct (forward) analysis, called simulation, takes place if we know I and T and the output O is searched.

The inverse analysis (identification) is more complicated. The problem discussed was formulated with respect to the internal identification. In such a regression problem I and O are known and parameters of T are searched.

In the presented paper the forward analysis is carried out by the guided wave technique and numerical simulation

process in order to find the Lamb dispersion curves.

The neural networks, which are a soft computational tool, were proved in the analysis of many inverse engineering problems, cf. [14, 15]. ANNs turned out to be numerically efficient especially in the inverse analysis of structural mechanics and solids. Thus, artificial neural networks have complementary features to hard features of computational simulations, e.g. to FEM. This is a foundation for formulating different HMS, see [20].

B. Parameter identification of thin plate parameters

Stage A

Direct analysis, called Stage A, corresponds to three essential steps, shown in Fig. 3. These steps are related to carrying out a laboratory test with excitation and propagation of the LWs (Essential Step I).

The signals are transmitted from sensors and preprocessed in Essential Step II.

Then, in Step III, the time signals are transformed into 2-B scans. Applying 2D-FFT and searching the local maxima in

the transform space, the points related to different vibration modes points at dispersion curves are found, see [18]. Thus, the experimental set of points can be formulated:

$$\mathcal{D}_{\text{exp}} = (f^j, k^j | m), \quad (8)$$

where: $j = 1, 2, \dots, J$ – numbers of dispersion points, m – number of vibration mode.

Data from (8) serve formulating an approximate experimental curve with vectors of basic functions $\mathbf{BF}_{\text{ref}}(f)$ and corresponding basic parameters \mathbf{a}_{exp} , cf. [12]:

$$k_{\text{exp}} = \mathbf{BF}_{\text{ref}}(f)^T \mathbf{a}_{\text{exp}}. \quad (9)$$

On the basis of data (8), approximate experimental curves can be formulated. The parameters of these curves are computed by means of the Least Squared Method (LSM):

$$\tilde{k}_{\text{exp}} = (f | \mathbf{BF}_{\text{ref}}(f), \mathcal{D}_{\text{exp}}) \xrightarrow{\text{LSM}} \tilde{\mathbf{a}}_{\text{exp}}. \quad (10)$$

In the presented paper only the anti-symmetric basic mode of vibrations A_0 was adopted, according to suggestions in [11]. The computed values of vector $\tilde{\mathbf{a}}_{\text{exp}}$ can be called internal parameters which control the transition from Stage A into Stage B.

Stage B

In the presented paper, Stage B of computer identification is related to Essential Step IV in [21]. In Stage B a possibility of the identification of the inverse analysis is carried out in a classical way, marked in Fig. 3 as CL.

Such an approach was used in our paper [13]. The application of ANN approach, in which the network is trained ‘off line’, is much more efficient numerically, cf. [16]. In this sense the presented paper is a continuation of paper [11] approach that needed iterations and it enabled fast identification of the parameters of a single only. In this respect the presented paper is a development of paper [13].

V. APPLICATION OF ANN FOR IDENTIFICATION OF PLATE PARAMETERS

Let us discuss Stage B, marked as ANN in the Essential Step IV in Fig. 3.

A. Formulation of ANN

The ANN corresponds to the MultiLayer Perceptron (MLP) network, see [21] of the following architecture:

$$\text{MLP: } \mathbf{\alpha}_{(I \times 1)} - \mathbf{H}_{(H \times 1)} - \mathbf{spar}_{(3 \times 1)}, \quad (11)$$

where: I – number of inputs, H – number of sigmoid functions in the hidden layer designed.

The corresponding set P , composed of P pattern pairs, was adopted for the training of the Multi Layered Perceptron (MLP) neural network, see [21].

$$\mathcal{P} = \{ \mathbf{a}^p, \mathbf{t}^p = \mathbf{spar}^p \}_{p=1}^P, \quad (12)$$

which fulfil Armikulova’s dimensionless equations (4).

The patterns for the training set are selected from the assumed ranges of the output vector components $\mathbf{spar}^p = \{E, \nu, d\}$. The number of patterns equals:

$$P = PE \times P\nu \times Pd \quad (13)$$

for the fixed value of the plate density $\rho = \rho_{\text{ref}}$. The outputs are related to the target values $\{t^p\}_{p=1}^{Pt}$ for $t^p = E^p, \nu^p, d^p$ within the ranges $\{t^p\} \in (t_{\text{min}}, t_{\text{max}})$.

After the training and testing of MLP, the relative errors of neural approximation can be computed:

$$\text{Rey} = \frac{1}{P} \sum_{p=1}^P (1 - y^p / t^p) \times 100\%, \quad (14)$$

where: y and t are computed and target values as components of the vector \mathbf{spar} .

The identified values of the plate parameter identification can be computed by the MLP trained ‘off line’:

$$\tilde{\mathbf{a}}_{\text{exp}} \xrightarrow{\text{MLP}} \mathbf{spar}_{\text{MLP}}. \quad (15)$$

The solution (17) is obtained by patterns corresponding to the exact Lamb dispersion curves. Thus, we can compute:

$$\mathbf{spar}_{\text{MLP}} = \mathbf{spar}_{\text{ident}} \approx \mathbf{spar}_{\text{exact}}, \quad (16)$$

with the accuracy corresponding to the MLP network errors (14).

Consequently, we can conclude that the final identification of plate parameters is nearly equal to the exact value of the identified parameters. It has been proved by many computer verifications.

VI. CASE STUDY

From among many results of numerical validation, only one case study is presented, as corresponding to the results obtained in [13]. Two case studies were analysed by the same MLP neural network, generated for the following ranges of the plate parameters:

$$E \in \{65.0, 77.5\} \text{ GPa}, \nu \in \{0.25, 0.35\}, \\ d \in \{1.5, 4.5\} \text{ mm}, \rho_{\text{ref}} = 2700 \text{ kg/m}^3, \quad (17)$$

Assuming a uniform distribution of material parameters within ranges (17), the corresponding number of selected patterns was adopted:

$$P = PE \times Pv \times Pd = 20 \times 15 \times 10 = 3000. \quad (18)$$

Then the numerical analysis was applied to design the MLP neural network of the following architecture:

$$\text{MLP: } 4-6-3. \quad (19)$$

After the numerical analysis the following basis functions were designed in [13]:

$$\tilde{k} = \alpha_{(1 \times 4)}^T \cdot \mathbf{BF}_{(4 \times 1)}(f) = \alpha_1 + \alpha_2 f + \alpha_3 f^2 + \alpha_4 (1/f). \quad (20)$$

The relative network errors computed by formula (16) gave the network errors for the fixed value of plate density $\rho_{\text{ref}} = 2700 \text{ kg/m}^3$:

$$ReE = 3.72 \times 10^{-5}, Re\nu = 3.73 \times 10^{-5}, Red = 10.5 \times 10^{-5}. \quad (21)$$

The experimental data discussed in [13] were taken from [11]. These data correspond to $J = 1722$ points at the experimental curve shown in Fig. 4.

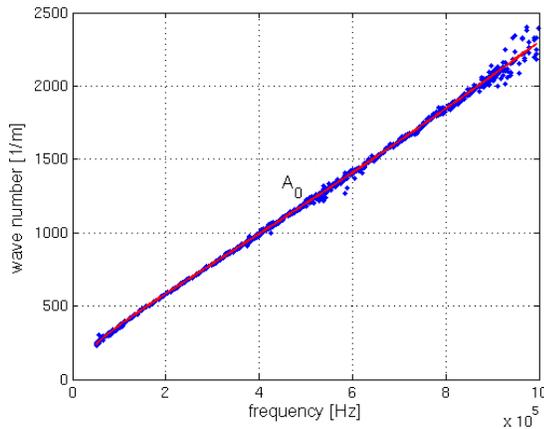


Fig. 4. A part of Lamb dispersion curve mode A_0 taken from [11]

The following vector of approximate experimental curve parameters was computed:

$$\tilde{\alpha}_{\text{exp}} = \{181.7, 194.2, 1.01, -20.0\}. \quad (24)$$

for $\rho_{\text{ref}} = \rho_{\text{fix}} = 2700 \text{ kg/m}^3$.

Their substitution as inputs into the trained MLP gave the outputs, i.e. the value of identified plate parameters:

$$E_{\text{ident}} = 65.92 \text{ GPa}, \nu_{\text{ident}} = 0.297, d_{\text{ident}} = 3.998 \text{ mm} \\ \rho_{\text{fix}} = 2700 \text{ kg/m}^3. \quad (26)$$

In the paper [13] it was reported that only Young's

modulus E was identified iteratively for the fixed values of other parameters:

$$E_{\text{ident}} = 65.45 \text{ GPa}, \nu_{\text{fix}} = 0.3, d_{\text{fix}} = 4.0 \text{ mm}, \\ \rho_{\text{fix}} = 2700 \text{ kg/m}^3. \quad (26)$$

VII. FINAL CONCLUSIONS AND REMARKS

1. The proposed hybrid method is non-iterative. This means that after the formulation of a neural network only one substitution of data obtained from the approximate experimental dispersion curves (in general they should fulfil Lamb equations) is needed. Thus, we can obtain quite accurately the identified values of Lamb DCs. In [9] the identification of curves needed about 14-17 iterations, obtained my means of FEM.

2. Looking at the results of the verification carried out in the case studies, it can be seen that the hybrid computational system gives nearly accurate results of plate parameters identification.

3. What is worth mentioning is the possibility of identification of three selected material parameters, i.e. $\{E, \nu, d\}$ or $\{E, \nu, \rho\}$, unlike in the papers by [9, 12], where only values of Young's modulus E were identified.

4. At present we are trying to generalise the discussed hybrid approach to the identification of orthotropic and, especially, composite laminate parameters, see [16].

REFERENCES

- [1] G.L. Rose GL, *Ultrasonic Waves in Solid Media*, Cambridge University Press, 1999.
- [2] A. Raghavan and C.E.S. Cesnik, "Review of Guided-Wave Structural Health Monitoring," *The Shock and Vibration Digest*, vol. 38, pp. 91-114, 2007.
- [3] L. Ambrozinski, P. Packo, Stepinski and T. Uhl, "Ultrasonic guided waves based method for SHM: simulations and an experimental test," *5th World Conference on Structural Control and Monitoring 5WCSCM* pp. 10443-10452, 2010.
- [4] M.M. Ettore, S. Alampalli, *Infrastructure Health Monitoring in Civil Engineering: Theory and Components*, vol. 12. CRS Press, Taylor& Francis Group, 2012.
- [5] W. Ostachowicz, P. Kudela, M. Krawczyk, A. Żak. *Guided Waves in Structures for SHM*, J. Wiley&Sons, 2012.
- [6] W. Ostachowicz and J.A.Guemes (eds), *New trends in Structure Health Monitoring*, CISM Courses and Lectures, vol. 542, Springer, 2013.
- [7] Z. Su and L. Ye, "Identification of damage using Lamb waves: from fundamentals to applications," In: F. Pfeifer and P. Wriggers (Series Editors) *Lecture Notes in Applied and computational Mechanics*, vol. 48, Berlin - Heidelberg, Springer, 2009.
- [8] G. R. Liu, K.Y. Lam and X. Han, "Determination of elastic constant of anisotropic laminated plates using elastic waves and a progressive neural network", *J. Soun Vibr.*, vol. 252, pp. 239-259, 2002.
- [9] M. Sale, M. Rizo P and Z. Marzani, "Semi-analytical formulation for guided waves-based reconstruction of elastic moduli." *Mechanical Systems and Signal Processing*, vol. 25, pp. 2241-2256, 2011.
- [10] Packo, L. Pieczonka, L. Ambrozinski and T. Uhl, "Elastic constants identification", *Proceedings of the 9th International workshop on Structural Health Monitoring*, Stanford, September. vol 1, pp. 10-12, ed. Fu-Kuo Chang ISBN 978-1-60595-115-7, 2013.
- [11] P. Packo, L. Ambroziński, L. Pieczonka, W.J. Staszewski, T. Uhl. The influence of experimental and numerical parameters on the dispersion curves used for elastic constants identification. *Inverse Problems in Science and Engineering*, submitted for publishing, 2014.
- [12] P. Packo, T. Uhl and W.J. Staszewski, "A general semi-analytical finite

- difference method for dispersion curves calculation and numerical analysis in Lamb wave propagation,” submitted for publication, *Inverse Problems Sci. Eng.*, 2014.
- [13] E. Pabisek, Z. Waszczyszyn and L. Ambrozinski, “A semi-analytical method for identification of thin elastic parameters basing on LWM, *Compu. Assist. Mech. Engn. Sci.*, vol. 21, pp. 5-14, 2014.
- [14] Z. Waszczyszyn, “Artificial neural networks in civil and structural engineering: ten years of research in Poland,” *Comp. Assist. Mech. Engn. Sci.*, vol. 13, pp. 489-512, 2006.
- [15] Z. Waszczyszyn, “Artificial neural networks in civil and structural engineering: another five years of research in Poland,” *Comp. Assist. Mech. Engn. Sci.*, vol. 18, pp. 131-146, 2011.
- [16] Z. Waszczyszyn, P. Nazarko, P. Packo, L. Ambrozinski, “Development of a Hybrid Computational System for identification of elastic material in composite lamina,” paper submitted to *COMPADYN 2015*, Crete Island, Greece, May 25-27, 2015
- [17] V. Amirkulova, 2011 “Dispersion relations for elastic wave in plates and rods,” M. Sc. Thesis, The State Univ of New Jersey, 2011.
- [18] D. Alleyne D and P. Cawley, “A two-dimensional Fourier transform method for the measurement of propagating multimode signals.” *The Journal of the Acoustical Society of America ASA*, vol. 89, pp. 1159-1168, 1991.
- [19] Th. Paez, “Neural networks in mechanical system simulation, identification and assessment,” *Shock and Vibration*, vol.1, pp. 177-199, 1993.
- [20] E. Pabisek, *Hybrid systems integrating FEM and ANN for the analysis of selected problems of structural and materials mechanics*, (in Polish), Monographs 369, TU Cracow, 2008.
- [21] [4] S. Haykin, *Neural networks: a comprehensible foundations*, 2nd ed, Prentice–Hall, 1999.

ADAPTIVE SPLINE PROCESSING OF DISCRETE FLOW

Burova Irina, Dem'yanovich Yu.K.

Abstract—An algorithm of approximation for discrete function f by Lagrange splines associated with the adaptive grid is proposed. The conditions, in which the algorithm is more affective than the analogous algorithm with equidistant grid, are done.

Index Terms—Splines, Interpolation, Approximation, Numerical flow, Adaptive grid.

I. INTRODUCTION

The adaptive methods of solution for a lot of problems are connected with usage of large flows of numerical information associated with some grid.

In particular the adaptive grids are used for enlargement of accuracy for solution of problems of mathematical physics (see [Lebedev, Lisejkin, Hakimzijanov (2002)], [Terekhov, Vassilevski (2013)]); their application leads up to reduction of size of numerical information flows.

The another approach to treatment of information flows is wavelet decomposition; the last one represent the origin flow as a main flow and an auxiliary flow (wavelet flow) such that it is able to use the main flow instead of the origin flow. The essential property of the decomposition is opportunity to restore the origin flow if it is necessary (see [Mallat (2002)]).

Now there are algorithmic basis for construction of spline-wavelets of Lagrange type associated with irregular grids; therefore the union of the both approaches becomes actual thing. Some variants of adaptive grids (with a priori fixed quantity of used knots) for Lagrange splines are proposed previously (see [Dem'yanovich, Hodakovskii (2008)]).

In the offered report the algorithm of approximation for discrete function u (i.e. function defined on an irregular grid) by Lagrange splines associated with the adaptive grid is proposed. The conditions, in which the algorithm is more affective than the analogous algorithm with pseudo-equidistant grid, are done.

II. AUXILIARY STATEMENTS

We discuss the grid Ξ on real interval (α, β) ,

$$\Xi : \dots < \xi_{-2} < \xi_{-1} < \xi_0 < \xi_1 < \xi_2 \dots,$$

$$\lim_{i \rightarrow -\infty} \xi_i = \alpha, \quad \lim_{i \rightarrow +\infty} \xi_i = \beta.$$

I. Burova is with the Mathematics and Mechanics Faculty, St. Petersburg State University, St. Petersburg, Russia e-mail: i.g.burova@spbu.ru, burovaig@mail.ru.

Yu.K. Dem'yanovich is with the Mathematics and Mechanics Faculty, St. Petersburg State University, St. Petersburg, Russia e-mail: Yuri.Demjanovich@gmail.com

Let $f(t)$ be function defined for $t \in \Xi$, and there is positive c such that

$$f(t) \geq c > 0 \quad \forall t \in \Xi. \quad (1)$$

If $a \in \Xi$, then $a = \xi_i$ for some integer $i \in \mathbf{Z}$. By definition, put $a^- \stackrel{\text{def}}{=} \xi_{i-1}$, $a^+ \stackrel{\text{def}}{=} \xi_{i+1}$.

Suppose $a, b \in \Xi$, $a < b$; consider the set $\{a, b\} = \{\xi_s \mid a \leq \xi_s \leq b, s \in \mathbf{Z}\}$ named by *grid segment*. We discuss a linear space $C\{a, b\}$ that is the set of functions $u(t)$ defined on the grid Ξ :

$$\|u\|_{C\{a,b\}} = \max_{t \in \{a,b\}} |u(t)|.$$

Assume that

$$\varepsilon \in (\varepsilon^*, \varepsilon^{**}), \quad (2)$$

where

$$\varepsilon^* = \max_{\xi \in \{a,b^-\}} \max_{t \in \{\xi, \xi^+\}} f(t)(\xi^+ - \xi),$$

$$\varepsilon^{**} = (b - a) \|f\|_{C\{a,b\}}. \quad (3)$$

Lemma 1. *If conditions (1), (2) – (3) are true, then there exist the unique positive integer $K = K(f, \varepsilon, \Xi)$ and the unique grid*

$$\tilde{X} = \tilde{X}(f, \varepsilon, \Xi) : a = \tilde{x}_0 < \tilde{x}_1 < \dots < \tilde{x}_K \leq \tilde{x}_{K+1} = b \quad (4)$$

such that

$$\max_{t \in \{\tilde{x}_s, \tilde{x}_{s+1}\}} f(t)(\tilde{x}_{s+1} - \tilde{x}_s) \leq \varepsilon < \max_{t \in \{\tilde{x}_s, \tilde{x}_{s+1}^+\}} f(t)(\tilde{x}_{s+1}^+ - \tilde{x}_s) \quad (5)$$

$$\forall s \in \{0, 1, \dots, K - 1\},$$

$$\max_{t \in \{\tilde{x}_K, b\}} f(t)(b - \tilde{x}_K) \leq \varepsilon, \quad \tilde{X} \in X. \quad (6)$$

The net (4) with properties (5) – (6) is called *the net of adaptive type*.

The Lemma 1 can be proved by mathematical induction over the number of knots.

Summing the relations (5), we obtain

$$\sum_{s=0}^{K-1} \max_{t \in \{\tilde{x}_s, \tilde{x}_{s+1}\}} f(t)(\tilde{x}_{s+1} - \tilde{x}_s) \leq K\varepsilon <$$

$$< \sum_{s=0}^{K-1} \max_{t \in \{\tilde{x}_s, \tilde{x}_{s+1}^+\}} f(t)(\tilde{x}_{s+1}^+ - \tilde{x}_s). \quad (7)$$

III. PSEUDO-EQUIDISTANT GRID

Now suppose that

$$\varepsilon \in (\bar{\varepsilon}^*, \varepsilon^{**}), \tag{8}$$

where

$$\begin{aligned} \bar{\varepsilon}^* &\stackrel{\text{def}}{=} \max_{\xi \in \{a, b^-\}} (\xi^+ - \xi) \|f\|_{C\{a, b\}}, \\ \varepsilon^{**} &\stackrel{\text{def}}{=} (b - a) \|f\|_{C\{a, b\}}. \end{aligned} \tag{9}$$

Then we find the numbers

$$N = N(f, \varepsilon, \Xi) = \lceil \varepsilon^{**} / \varepsilon \rceil \leq N + 1, \tag{10}$$

and

$$h = h(f, \varepsilon, \Xi) \stackrel{\text{def}}{=} \frac{b - a}{N + 1}. \tag{11}$$

Consider now the grid

$$\bar{X} = \bar{X}(f, \varepsilon, \Xi): \quad a = \bar{x}_0 < \bar{x}_1 < \dots < \bar{x}_N = b, \quad \bar{X} \subset X, \tag{12}$$

where

$$\bar{x}_{s+1} - \bar{x}_s \leq h < \bar{x}_{s+1}^+ - \bar{x}_s, \quad s \in \{0, 1, \dots, N - 1\}, \tag{13}$$

$$\bar{x}_{N+1} - \bar{x}_N \leq h. \tag{14}$$

The grid (12) with properties (13) – (14) is called *pseudo-equidistant grid*.

By (10) we have

$$N \leq \frac{b - a}{\varepsilon} \|f\|_{C\{a, b\}} < N + 1; \tag{15}$$

therefore

$$(b - a) \|f\|_{C\{a, b\}} - \varepsilon < N\varepsilon \leq (b - a) \|f\|_{C\{a, b\}}. \tag{16}$$

By (15) we get $\frac{b-a}{N+1} \|f\|_{C\{a, b\}} < \varepsilon$; hence $h \|f\|_{C\{a, b\}} < \varepsilon$. By left inequality (13) and by inequality (14) we obtain

$$\max_{t \in \{\bar{x}_s, \bar{x}_{s+1}\}} f(t) (\bar{x}_{s+1} - \bar{x}_s) \leq \varepsilon, \quad s \in \{0, 1, \dots, N\}. \tag{17}$$

By mathematical induction over the number of knots the next assertion can be proved.

Lemma 2. *If the relations (8) – (9) are true, then there exists the unique grid (12) with properties (13) – (14), and the relations (16) – (17) are fulfilled.*

IV. RELATIVE QUANTITY OF KNOTS

Theorem 1. *Suppose the hypothesis of lemmas 1 and 2 are fulfilled. Then*

$$\begin{aligned} &\frac{(b - a) \|f\|_{C\{a, b\}} - \varepsilon}{\sum_{s=0}^{K-1} \max_{t \in \{\bar{x}_s, \bar{x}_{s+1}\}} f(t) (\bar{x}_{s+1}^+ - \bar{x}_s)} < \frac{N}{K} \leq \\ &\leq \frac{(b - a) \|f\|_{C\{a, b\}}}{\sum_{s=0}^{K-1} \max_{t \in \{\bar{x}_s, \bar{x}_{s+1}\}} f(t) (\bar{x}_{s+1} - \bar{x}_s)}. \end{aligned} \tag{18}$$

Proof. Inequality (18) follows from the relations (7) and (16).

V. LIMIT RELATIONS

Suppose the function $f(t)$ is continuous in the segment $[a, b]$, and satisfies to the property

$$f(t) \geq c > 0 \quad \forall t \in [a, b]. \tag{19}$$

Consider the sequence of grids $\Xi(\eta)$,

$$\Xi(\lambda) : \dots < \xi_{-2}(\lambda) < \xi_{-1}(\lambda) < \xi_0(\lambda) < \xi_1(\lambda) < \xi_2(\lambda) \dots, \tag{20}$$

depending on parameter $\lambda > 0$; suppose $a, b \in \Xi(\lambda)$.

By definition, put

$$\{a, b\}_\lambda = \Xi(\lambda) \cap [a, b], \quad h_\lambda = \max_{\xi \in \{a, b^-\}_\lambda} (\xi^+ - \xi).$$

Theorem 2. *If $f \in C[a, b]$, the condition (19) is valid, and*

$$\lim_{\lambda \rightarrow +0} h_\lambda = 0, \tag{21}$$

then

$$\lim_{\varepsilon \rightarrow +0} \lim_{\lambda \rightarrow +0} \frac{K}{N} = \frac{1}{b-a} \int_a^b f(t) dt. \tag{22}$$

Proof. It is clear to see that

$$\lim_{\lambda \rightarrow +0} \|f\|_{C\{a, b\}_\lambda} = \|f\|_{C[a, b]}.$$

After passage to the limit $\lambda \rightarrow +0$ in the relations (5) – (6) we can pass to limit under condition $\varepsilon \rightarrow +0$; as a result we get the integral $\int_a^b f(t) dt$ instead of sums in the relation (18). The formula (22) is proved.

VI. APPROXIMATION OF THE DISCRETE FLOW

Suppose the function $u(t)$ is defined for $t \in \{y, z\}$, where

$$\{y, z\} : \quad y = \xi_0 < \xi_1 < \dots < \xi_{M-1} < \xi_M = z. \tag{23}$$

Define the function $\tilde{u}(t)$ by formula

$$\tilde{u}(t) \stackrel{\text{def}}{=} u(y) + \frac{u(z) - u(y)}{z - y} (t - y), \quad t \in [y, z].$$

The summation analog of Leibnitz formula can be written in the form

$$w(\theta) - w(\tau) = \sum_{s=i}^{k-1} \frac{w(\xi_{s+1}) - w(\xi_s)}{\xi_{s+1} - \xi_s} (\xi_{s+1} - \xi_s), \tag{24}$$

where $\theta = \xi_k, \tau = \xi_i$.

Using the formula (24), it is very simple to obtain the next assertion.

Lemma 3. *If $t \in \{y, z^-\}, t = \xi_k$, then*

$$\begin{aligned} &u(t) - \tilde{u}(t) = \\ &= \sum_{j=0}^{M-1} (\xi_{j+1} - \xi_j) \sum_{i=0}^{k-1} \left[D_{\Xi} u(\xi_i) - D_{\Xi} u(\xi_j) \right] \frac{\xi_{i+1} - \xi_i}{\xi_M - \xi_0}; \end{aligned} \tag{25}$$

here

$$D_{\Xi} u(\xi) = \frac{u(\xi^+) - u(\xi)}{\xi^+ - \xi}, \quad \xi \in \{y, z^-\}.$$

Theorem 3. If $t \in \{y, z\}$, then the next inequality

$$|u(t) - \tilde{u}(t)| \leq 2(z - y) \max_{\xi \in \{y, z^-\}} |D_{\Xi} u(\xi)| \quad (26)$$

is correct.

Proof. If $t \in \{y, z\}$, $t = \xi_k$, the formula (25) gives

$$|u(t) - \tilde{u}(t)| \leq \frac{1}{\xi_M - \xi_0} \times$$

$$\times \sum_{j=0}^{M-1} (\xi_{j+1} - \xi_j) \sum_{i=0}^{k-1} |D_{\Xi} u(\xi_i) - D_{\Xi} u(\xi_j)| (\xi_{i+1} - \xi_i). \quad (27)$$

Using the relation (27), we obtain (26).

Using the analog (24) of Leibnitz formula, we easily get the following statement.

Lemma 4. If $t \in \{y, z\}$, $t = \xi_k$, then

$$u(t) - \tilde{u}(t) = \sum_{j=0}^{M-1} (\xi_{j+1} - \xi_j) \sum_{i=0}^{k-1} (\xi_{i+1} - \xi_i) \times \\ \times \sum_{p=j}^{i-1} D_{\Xi}^2 u(\xi_{p+1}) (\xi_{p+1} - \xi_p) / (\xi_M - \xi_0), \quad (28)$$

where

$$D_{\Xi}^2 u(\xi) \stackrel{\text{def}}{=} \frac{D_{\Xi} u(\xi) - D_{\Xi} u(\xi^-)}{\xi - \xi^-}, \quad \xi \in \{y^+, z^-\}. \quad (29)$$

Theorem 4. If $t \in \{y, z\}$, then

$$|u(t) - \tilde{u}(t)| \leq (z - y)^2 \max_{\xi \in \{y^+, z^-\}} |D_{\Xi}^2 u(\xi)|, \quad t \in \{y, z\}. \quad (30)$$

Proof. By (28) – (29) we have

$$|u(t) - \tilde{u}(t)| \leq \\ \leq \sum_{j=0}^{M-1} (\xi_{j+1} - \xi_j) \sum_{i=0}^{k-1} (\xi_{i+1} - \xi_i) \left| \sum_{p=j}^{i-1} (\xi_{p+1} - \xi_p) \right| / (\xi_M - \xi_0) \times \\ \times \max_{\xi \in \{y^+, z^-\}} |D_{\Xi}^2 u(\xi)|. \quad (31)$$

The evaluation (30) follows from (31).

Consider the grid \hat{X} , $\hat{X} \subset \Xi$, defined by relation

$$\hat{X}: \quad a = \hat{x}_0 < \hat{x}_1 < \hat{x}_2 < \dots < \hat{x}_{\hat{K}} < \hat{x}_{\hat{K}+1} = b, \quad \hat{X} \subset \Xi.$$

Let $\tilde{u}(t)$ be piecewise linear interpolation of the function $u \in C\{a, b\}$:

$$\tilde{u}(t) = u(\hat{x}_j) + \frac{u(\hat{x}_{j+1}) - u(\hat{x}_j)}{\hat{x}_{j+1} - \hat{x}_j} (t - \hat{x}_j) \quad \forall t \in [\hat{x}_j, \hat{x}_{j+1}), \\ j \in \{0, 1, \dots, \hat{K}\}.$$

Theorem 5. If $t \in \{\hat{x}_j, \hat{x}_{j+1}\}$, then

$$|u(t) - \tilde{u}(t)| \leq 2(\hat{x}_{j+1} - \hat{x}_j) \max_{\xi \in \{\hat{x}_j, \hat{x}_{j+1}\}} |D_{\Xi} u(\xi)|, \quad (32)$$

$$|u(t) - \tilde{u}(t)| \leq (\hat{x}_{j+1} - \hat{x}_j)^2 \max_{\xi \in \{\hat{x}_j^+, \hat{x}_{j+1}^-\}} |D_{\Xi}^2 u(\xi)|. \quad (33)$$

If $u \in C^1[a, b]$, then

$$|u(t) - \tilde{u}(t)| \leq 2 \max_{\xi \in [\hat{x}_j, \hat{x}_{j+1}]} |u'(\xi)| (\hat{x}_{j+1} - \hat{x}_j), \quad (34)$$

and if $u \in C^2[a, b]$, then

$$|u(t) - \tilde{u}(t)| \leq \max_{\zeta \in [\hat{x}_j, \hat{x}_{j+1}]} |u''(\zeta)| (\hat{x}_{j+1} - \hat{x}_j)^2 \quad \forall t \in (\hat{x}_j, \hat{x}_{j+1}). \quad (35)$$

Proof. The inequalities (32) – (33) follow from relations (26) and (30). Using passage to the limit $\max_{\xi \in \{y, z^-\}} (\xi^+ - \xi) \rightarrow +0$ in (32) – (33) we get (34) – (35).

VII. ON NUMBER OF KNOTS FOR GRID OF ADAPTIVE TYPE

Theorem 6. If the condition $|D_{\Xi} u(t)| \geq c > 0 \quad \forall t \in \{y, z\}$ is true, and the grid \tilde{X} be the same as $\tilde{X}(|D_{\Xi} u(t)|, \eta/2, \Xi)$, then

1) the number $K'_{u, \Xi}(\eta) \stackrel{\text{def}}{=} K(|D_{\Xi} u(t)|, \eta/2, \Xi)$ satisfies to the relations

$$2 \sum_{s=0}^{K-1} \max_{t \in \{\tilde{x}_s, \tilde{x}_{s+1}\}} |D_{\Xi} u(t)| (\tilde{x}_{s+1} - \tilde{x}_s) / \eta \leq K'_{u, \Xi}(\eta) < \\ < 2 \sum_{s=0}^{K-1} \max_{t \in \{\tilde{x}_s^+, \tilde{x}_{s+1}^+\}} |D_{\Xi} u(t)| (\tilde{x}_{s+1}^+ - \tilde{x}_s) / \eta, \quad (36)$$

2) the inequality

$$|u(t) - \tilde{u}(t)| \leq \eta \quad \forall t \in \{y, z\} \quad (37)$$

is true, 3) if $u \in C^1[a, b]$, $|u'(t)| \geq c > 0 \quad \forall t \in [a, b]$, and sequence (20) satisfies the condition (21), then

$$\lim_{\eta' \rightarrow +0} \lim_{\lambda \rightarrow +0} K'_{u, \Xi(\lambda)}(\eta') \eta' = 2 \int_a^b |u'(t)| dt. \quad (38)$$

Proof. The formula (36) follows from (7), where $f(t) = |D_{\Xi} u(t)|$. The inequality (37) can be obtained from (26) and (5), where $f(t) = |D_{\Xi} u(t)|$, $\varepsilon = \eta/2$. The relation (38) follows from (36) by sequential passages to the limits: first we have $\lambda \rightarrow +0$, and then we pass η to zero.

Theorem 7. Suppose the condition

$$|D_{\Xi}^2 u(t)| \geq c > 0 \quad \forall t \in \{y, z\} \quad (39)$$

is fulfilled. If the grid \hat{X} coincides with $\tilde{X}(\sqrt{|D_{\Xi}^2 u(t)|}, \eta, \Xi)$, then

1) the quantity of knots $K''_{u, \Xi}(\eta) = K(\sqrt{|D_{\Xi}^2 u(t)|}, \eta, \Xi)$ satisfies to relations

$$\sum_{s=0}^{K-1} \max_{t \in \{\tilde{x}_s, \tilde{x}_{s+1}\}} \sqrt{|D_{\Xi}^2 u(t)|} (\tilde{x}_{s+1} - \tilde{x}_s) / \eta \leq K''_{u, \Xi}(\eta) < \\ < \sum_{s=0}^{K-1} \max_{t \in \{\tilde{x}_s^+, \tilde{x}_{s+1}^+\}} \sqrt{|D_{\Xi}^2 u(t)|} (\tilde{x}_{s+1}^+ - \tilde{x}_s) / \eta. \quad (40)$$

2) the inequality

$$|u(t) - \tilde{u}(t)| \leq \eta^2 \quad \forall t \in \{y, z\} \quad (41)$$

is true,

3) if $u \in C^2[a, b]$, $|u''(t)| \geq c > 0 \forall t \in [a, b]$, and (21) is fulfilled, then

$$\lim_{\eta' \rightarrow +0} \lim_{\lambda \rightarrow +0} K''_{u, \Xi(\lambda)}(\eta')\eta' = \int_a^b \sqrt{|u''(t)|} dt. \quad (42)$$

Proof. The formula (40) follows from (7), where $f(t) = \sqrt{|D_{\Xi}^2 u(t)|}$. The inequality (41) can be obtained from (30) and (5), where $f(t) \stackrel{\text{def}}{=} \sqrt{|D_{\Xi} u(t)|}$, $\varepsilon \stackrel{\text{def}}{=} \eta$. Finally, the formula (42) follows from (40) by sequential passages to the limits (see the proof of Theorem 6).

VIII. ON THE QUANTITY OF KNOTS OF PSEUDO-EQUIDISTANT GRID

Theorem 8. If the grid \hat{X} is the same as the grid $\bar{X}(|D_{\Xi} u|, \eta/2, \Xi)$, then

1) the number $N'_{u, \Xi}(\eta) = N(|D_{\Xi} u|, \eta/2, \Xi)$ of inner knots of the grid satisfies to the relation

$$2(b-a)\|D_{\Xi} u\|_{C\{a, b\}}/\eta - 1 < N'_{u, \Xi}(\eta) \leq 2(b-a)\|D_{\Xi} u\|_{C\{a, b\}}/\eta. \quad (43)$$

2) the inequality

$$|u(t) - \tilde{u}(t)| \leq \eta \quad \forall t \in \{a, b\} \quad (44)$$

is fulfilled.

Proof. Suppose that $\hat{X} \stackrel{\text{def}}{=} \bar{X}(|D_{\Xi} u|, \eta/2, \Xi)$. Using the formula (16), we get relation (43). The inequality (44) follows from (26) and (17), where $f = |D_{\Xi} u|$ and $\varepsilon = \eta/2$.

Theorem 9. If the grid \hat{X} is equal to the grid $\bar{X}(\sqrt{|D_{\Xi}^2 u|}, \eta, \Xi)$, then

1) the number $N''_{u, \Xi}(\eta) = N(\sqrt{|D_{\Xi}^2 u|}, \eta/2, \Xi)$ of inner knots of the grid satisfies to relation

$$(b-a)\| |D_{\Xi}^2 u|^{1/2} \|_{C\{a, b\}}/\eta - 1 < N''_{u, \Xi}(\eta) \leq (b-a)\| |D_{\Xi}^2 u|^{1/2} \|_{C\{a, b\}}/\eta, \quad (45)$$

2) the inequality

$$|u(t) - \tilde{u}(t)| \leq \eta^2 \quad \forall t \in \{a, b\} \quad (46)$$

is true.

Proof. Applying the formula (16) with $\hat{X} \stackrel{\text{def}}{=} \bar{X}(\sqrt{|D_{\Xi}^2 u|}, \eta, \Xi)$, we get the relation (45). The inequality 46 follows from (30) and (17) if $f = \sqrt{|D_{\Xi}^2 u|}$, $\varepsilon = \eta$.

IX. RELATIVE CHARACTERISTIC OF THE QUANTITIES OF KNOTS FOR DIFFERENT GRIDS UNDER CONDITION OF THE SAME APPROXIMATION

Theorem 10. The inequality $|\tilde{u}(t) - u(t)| \leq \eta$ is true for each of two variants of grids: $\hat{X} = \bar{X}(|D_{\Xi} u|, \eta/2, \Xi)$ and $\hat{X} = \bar{X}(\sqrt{|D_{\Xi}^2 u|}, \eta, \Xi)$. In addition we have

$$\begin{aligned} & \frac{(b-a)\|D_{\Xi} u\|_{C\{a, b\}} - \eta/2}{\sum_{s=0}^{K-1} \max_{t \in \{\tilde{x}_s, \tilde{x}_{s+1}^+\}} |D_{\Xi} u(t)|(\tilde{x}_{s+1}^+ - \tilde{x}_s)} < \frac{N'_{u, \Xi}(\eta)}{K'_{u, \Xi}(\eta)} \leq \\ & \leq \frac{(b-a)\|D_{\Xi} u\|_{C\{a, b\}}}{\sum_{s=0}^{K-1} \max_{t \in \{\tilde{x}_s, \tilde{x}_{s+1}^+\}} |D_{\Xi} u(t)|(\tilde{x}_{s+1}^+ - \tilde{x}_s)}. \end{aligned} \quad (47)$$

Proof. Using the inequality (18), we put $f = |D_{\Xi} u|$ and $\varepsilon = \eta/2$. As a result we get (47)

Theorem 11. If the family of grids (20) has property (21), $u \in C^1[a, b]$ and $\|u'\|_{C[a, b]} \neq 0$, then

$$\lim_{\eta \rightarrow +0} \lim_{\lambda \rightarrow +0} \frac{N'_{u, \Xi(\lambda)}(\eta)}{K'_{u, \Xi(\lambda)}(\eta)} = \frac{\frac{1}{b-a} \int_a^b |u'(t)| dt}{\|u'\|_{C[a, b]}}. \quad (48)$$

Proof. Passing on to the limit $\lambda \rightarrow +0$ in the relation (47) and then to the limit $\eta \rightarrow +0$, we get (48).

Theorem 12. If $\hat{X} = \bar{X}(\sqrt{|D_{\Xi}^2 u|}, \eta, \Xi)$ or if $\hat{X} = \tilde{X}(\sqrt{|D_{\Xi}^2 u|}, \eta, \Xi)$, then in the both cases the evaluation $|\tilde{u}(t) - u(t)| \leq \eta^2$ is correct. Moreover we have

$$\begin{aligned} & \frac{(b-a)\| |D_{\Xi}^2 u|^{1/2} \|_{C\{a, b\}} - \eta}{\sum_{s=0}^{K-1} \max_{t \in \{\tilde{x}_s, \tilde{x}_{s+1}^+\}} \sqrt{|D_{\Xi}^2 u(t)|}(\tilde{x}_{s+1}^+ - \tilde{x}_s)} < \frac{N''_{u, \Xi}(\eta)}{K''_{u, \Xi}(\eta)} \leq \\ & \leq \frac{(b-a)\| |D_{\Xi}^2 u|^{1/2} \|_{C\{a, b\}}}{\sum_{s=0}^{K-1} \max_{t \in \{\tilde{x}_s, \tilde{x}_{s+1}^+\}} \sqrt{|D_{\Xi}^2 u(t)|}(\tilde{x}_{s+1}^+ - \tilde{x}_s)}. \end{aligned} \quad (49)$$

Proof. Applying the inequality (18) with $f = \sqrt{|D_{\Xi}^2 u|}$ and $\varepsilon = \eta$, we get (49)

Theorem 13. If the family of grids (20) has the property (21) and the function $u \in C^2[a, b]$ satisfies to the relation $\|u''\|_{C[a, b]} \neq 0$, then

$$\lim_{\eta \rightarrow +0} \lim_{\lambda \rightarrow +0} \frac{N''_{u, \Xi(\lambda)}(\eta)}{K''_{u, \Xi(\lambda)}(\eta)} = \frac{\frac{1}{b-a} \int_a^b \sqrt{|u''(t)|} dt}{\|\sqrt{|u''}\|_{C[a, b]}}. \quad (50)$$

Proof. Passing on to the limit $\lambda \rightarrow +0$ in the relation (49) and then to the limit $\eta \rightarrow +0$, we get (50).

Acknowledgment

The work is partly supported by RFFI, Grants No 14-01-00069, 15-01-08847

REFERENCES

- [Lebedev, Lisejkin, Hakimzjanov (2002)] A.S.Lebedev, V.D.Lisejkin, G.S.Hakimzjanov. Development of methods for construction adaptive grids. *Computer Technologies*. Vol. 7. No.3, 2002, pp.29-43 (in Russian).
- [Terekhov, Vassilevski (2013)] K.Terekhov, Yu.Vassilevski. Two-phase water flooding simulations on dynamic adaptive octree grids with two-point nonlinear fluxes. *Journal of Numerical Analysis and Mathematical Modelling*. Vol.28. No.3. 2013, pp.267-288. (Russian)
- [Mallat (2002)] Stephane Mallat. A wavelet tour of signal processing. Academic Press. 2002.
- [Dem'yanovich, Hodakovskii (2008)] Yu.K.Dem'yanovich, V.A.Hodakovskii Introduction in the wavelet theory. SPb. 2008 (in Russian).

Introduction and Simulation of a New Model of Phantom by Monte Carlo to Obtain Depth Dose

Seyed Alireza Mousavi Shirazi

Abstract— In this investigation the simulation of a new model of liver phantom is studied by applying the Monte Carlo simulation. This phantom has identical compositions compared with existing compositions in a human liver tissue so that each of the materials in an adult liver tissue (including water, protein, glucose and glycogen) is decomposed to constituent elements of it, based on mass percentage and density of every element, then the accurate mass of every analyzed element in human liver tissue (such as H, O, C and N) is corresponded to mass of components of the mentioned phantom. The depth dose in all components and different layers of this phantom are computed by Monte Carlo simulation. In addition in this investigation another method using mathematical equations and computer programming is introduced so that it computes the deposited energy in the liver phantom components. The results of both MCNP code and analytical approximations to survey agreeing are compared together. The results show that the depth dose computed by MCNP code agrees as well with analytical approximations for neutron energy below 15MeV.

Keywords— Analytical method; Depth dose; Monte Carlo; Neutron; Phantom.

I. INTRODUCTION

THE radiotherapy course is applicable in treatment of liver cancer. During clinical therapy, it is always indispensable to stop absorption of excess dose by normal tissue. On the other hand, measurement and assessment of the depth dose and its calibration is an important matter [1].

It is considered that reactor based epithermal neutron beams with near optimum characteristics are currently available and can mostly be constructed at existing reactors [2, 3]. The boric acid solution moderator might be appropriate for the spectrum measurement of an epithermal neutron irradiation field.

Thus computation and modeling of the delivered dose by Monte Carlo method before practical treatment is recommended. An appropriate software tool for this aim is MCNPX Code [4, 5].

In this study, one of the main objectives is to study the interaction of neutrons in a real liver tissue and also deduction of incident neutron energy emitted from clinical neutron source for a vast spectrum of neutrons so that it can specify the accurate amounts of depth dose and neutron energy reached components of a real liver tissue [6, 7, 8].

In real state the average width of liver tissue across for adult human is approximately 21-22.5 cm and the vertical height of the organ at the greatest height is estimated to be 15-17.5 cm and the depth is 10-12.5 cm from the front to back [9].

After analyzing the structural materials of liver tissue (weighing 2kg) based on mass percentage and densities of them to their constituent elements consisting of H, O, C, S and N, the amounts of H and O are incorporated into the water and the amounts of other elements including C, S and N are calculated accordingly as accurately as possible. Therefore the mass, thickness and sphere radius relating to every analyzed element result from calculations according to Table 1:

Table.1 Mass, thickness and sphere radius belonging to every analyzed element

Elements	Mass (gr)	Outside radius of related sphere (cm)	Thickness (cm)
C	316.937	3.437	0.7
S	6.085	3.456	0.019
N	102.943	27.004	23.547
H	198.900	incorporated into the water sphere with radius: 2.737cm	

In this research to simulate a liver phantom by MCNPX Code, a spherical phantom is considered so that it is comprised

a sphere of water that its radius is averagely considered: 2.737cm. This water sphere is covered with a layer of carbon that according to Table.1 the hypothetical outside radius and thickness of this layer are respectively: 3.437cm and 7mm. There is an external thin layer of sulfur with thickness: 197.2 μ m around the carbon layer. The entire assembly is encased in a spherical shell of carbon as reflector with inside radius: 27.004 cm and with thickness: 3mm. This layer is as a reflector to decrease escaping the fast neutrons. The blank space between sulfur layer and carbon shell is filled with nitrogen gas. This blank space has inside and outside radiuses: 3.456cm and 27.004cm respectively. It must be noted that the spherical shape of phantom is different of real geometry aspect so that the different shapes of phantom like cubic format for body parts of human in some study have already been applied. But in general, in such kind of research phantom, the main objective is to simulate the nuclear and atomic interactions in the material [10, 11]. This phantom might also have a variable radius. It means that it might be bigger or smaller than applied dimensions in this investigation and might reality be equivalent to dimensions of livers belonging to either minor or major [12].

In the present work, neutrons are emitted from an external source and after passing through carbon and slowing down, the deposited energy in the phantom materials are computed by the MCNPX Code. The absorbed energy in the liver phantom is also computed by analytical computations as well. This includes generation of random numbers along with applying the neutron diffusion equations for neutron source [13, 14]. The results of two methods are compared.

The results of computation provide assurance that whether the absorbed dose in cancerous and healthy tissues are in accord with requirements [15, 16]. As Am-Be source produces a wide spectrum of neutron energies thus is not appropriate for this purpose [17].

II. MATERIALS AND METHODS

A. Simulation by MCNPX Code

According to Fig.1 and Fig.2 a phantom is considered. The Fig.1 illustrates the geometric view of the phantom that is introduced into MCNPX Code and the Fig.2 shows the

simulated view of the phantom which has been simulated by MCNPX Code. The incident neutron after passing the layers and lots of collision (with other nuclei) arrives at the water of phantom.

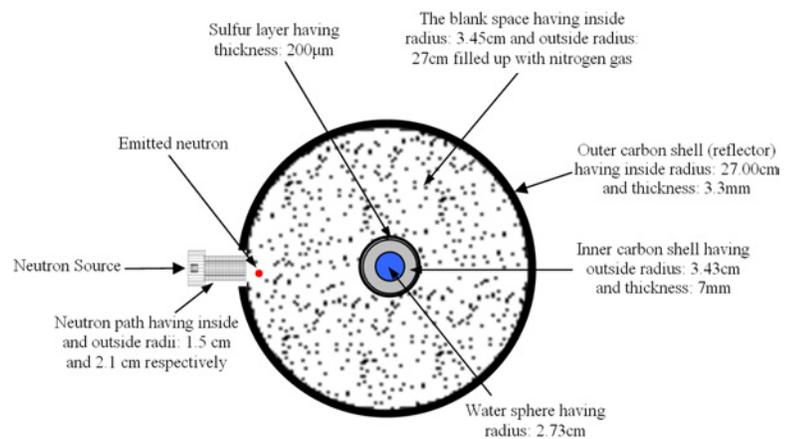


Fig. 1 The schematic of the spherical liver phantom

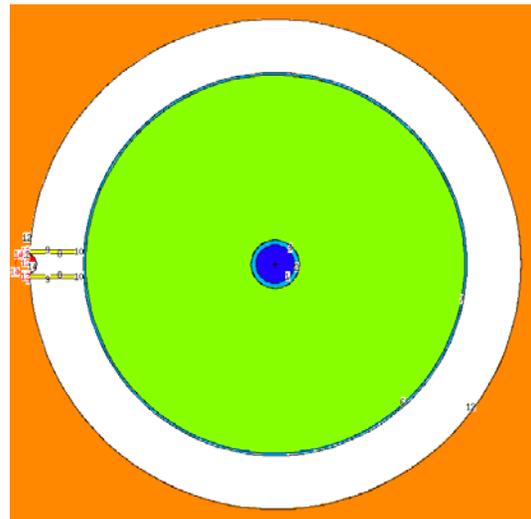


Fig. 2 The simulated geometry by MCNPX Code

Both composition and geometrical data belonging to liver phantom have been inputted into the MCNPX Code. In the simulation, neutron slowing down has been taken into account as well.

B. Analytical method

It is necessary to obtain data on neutron scattering cross section and its angular distribution because of computing the neutron penetration in liver phantom. Since neutrons passing through the cited phantom and the knowledge about neutron angular distribution after scattering are indispensable. Thus, use of the random sampling techniques can help to compute the probability of those neutrons which may be scattered in a

definite angle.

In the case of neutron transport in matter, it is subjected to three major types of interactions with carbon and hydrogen nuclei. These are elastic scattering, inelastic scattering, and radioactive capture. Each type of interactions takes place with a specific probability with a preponderant proportion into the related total cross section according to the related equation [11]:

The probability of inelastic scattering in carbon (for both the first and the second excitation levels in: 4.43MeV and 7.65MeV respectively):

$$P = \frac{\sum_C}{\sum_{tot}} \quad (1)$$

The inelastic scattering of neutron in carbon is important. There are two excited statuses in carbon nucleus, namely 4.43MeV and 7.65MeV. Thus significant value of energy is absorbed in the recoiled nucleus and E_R is less than the energy of inelastic scattered neutrons [18].

The transferred energy to a recoiled nucleus with mass number (A) during the collision of a neutron with energy E_n is computed with Eq.2:

$$E_r = \frac{2A}{(A+1)^2} E_n \cos^2 \psi \quad (2)$$

Since remarkable percentage of the liver tissue is made up of hydrogen, the interaction of neutron with hydrogen must be studied as well. Approximately 85%-95% of neutron energy transferred to liver tissue is attributed to its interaction with hydrogen. For example, for $E_n > 15\text{MeV}$ the (n,α) reaction makes the considerable fraction of depth dose in the tissue.

At 14MeV the contribution of recoiled proton (hydrogen nucleus) is related to energy deposition phenomenon that is 66% of total and the rest is due to α particle plus other heavy nuclei [18].

The computer programming which has been developed in the present work computes the transferred energy from incident neutrons to the phantom based on initial neutron energy, scattering angle (ψ) and mass number of the target nuclei. Equation 3 describes the neutron energy transfer from high energy to the medium:

$$E_{th} = E_n e^{-n\xi} \quad (3)$$

III. Results and Discussion

The derived graphs for depth dose in the components of the mentioned phantom (per emitted neutrons) in E_n-E_R by both Monte Carlo simulation with nps:10⁶ and analytical method are as Figs.3-5:

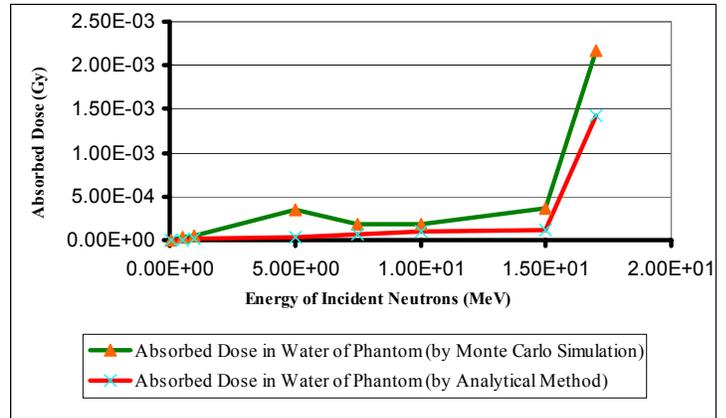


Fig. 3 The depth dose in the water of phantom

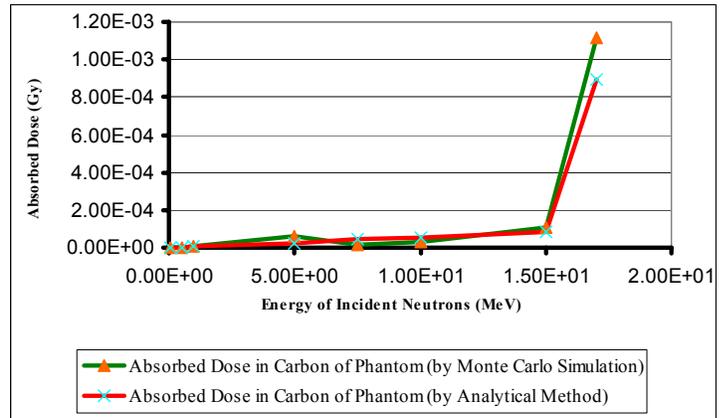


Fig. 4 The absorbed energy in the layer of carbon

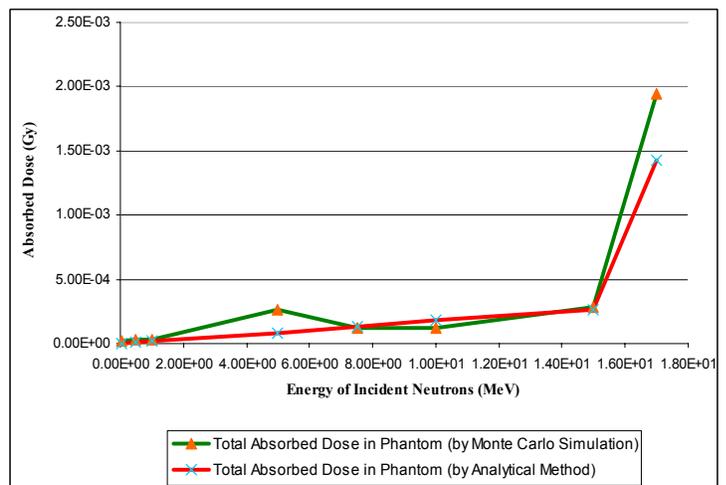


Fig. 5 The total depth dose in the phantom obtained from both Monte Carlo simulation and analytical method

IV. CONCLUSION

According to Figs.3-5, it is observed that within neutron energy range of 0.001eV–15MeV the calculated depth doses by analytical method are approximately similar to obtained results by MCNPX Code and the derived graphs of both methods for neutron energy below 15MeV agree together as well. For neutron energy above 15MeV, the results of these two methods produce significant error.

ACKNOWLEDGMENT

This paper is related to a research project entitled: “Presenting a New Model of Phantom and Its Simulation by MCNPX Code for Dosimetry and Also Presenting an Analytical Method to Compare the Obtained Results” that by sponsorship and financial supporting the “Islamic Azad University-South Tehran Branch” has been carried out.

REFERENCES

- [1] Emiliano C, Pozzi C, Thorp S, Brockman J, Miller M, Nigg D, Hawthorne F (2011) Intercalibration of physical neutron dosimetry for the RA-3 and MURR thermal neutron sources for BNCT small-animal research. *Appl Rad Isot* **69**, 1921-1923.
- [2] Rafiei Karahroudi M, Mousavi Shirazi S. A (2015) Study of power distribution in the CZP, HFP and normal operation states of VVER-1000 (Bushehr) nuclear reactor core by coupling nuclear codes. *Annals of Nuclear Energy* **75**, 38-43.
- [3] Sheibani J, Mousavi Shirazi, S.A., Rahimi, M.F (2014) Studying the effects of compound nucleus energy on coefficient of surface tension in fusion reactions using proximity potential formalism. *J FUSION ENERG* **33**, 74-82.
- [4] Harling OK, Riley KJ (2003) Fission reactor neutron sources for neutron capture therapy-a critical review. *J Neurooncol* **62** (1-2), 7-17.
- [5] Mousavi Shirazi S. A (2012) The simulation of a model by SIMULINK of MATLAB for determining the best ranges for velocity and delay time of control rod movement in LWR reactors. *Progress in Nuclear Energy* **54**, 64-67.
- [6] Mousavi Shirazi S. A, Rastayesh S (2011) The comparative investigation and calculation of thermo-neutronic parameters on two gens II and III nuclear reactors with same powers. *World Academy of Science, Engineering and Technology (WASET)* **5**, 99-103.
- [7] Mousavi Shirazi S. A, Aghanajafi C, Sadoughi S, Sharifloo N (2010) Design, construction and simulation of a multipurpose system for precision movement of control rods in nuclear reactors. *Annals of Nuclear Energy* **37**, 1659-1665.
- [8] Mousavi Shirazi S. A, Shafeie Lilehkouhi, M. S (2012) The assessment of radioisotopes and radiomedicines in the MNSR reactor of Isfahan and obtaining the burnup by applying the obtained information. *Proc. Conf. Asia-Pacific Power and Energy Engineering (APPEEC)*, Shanghai, 1-4.
- [9] Mousavi Shirazi S. A, Sardari D (2012) Design and Simulation of a New Model for Treatment by NCT. *Sci. Technol. Nucl. Install*, 1-7, doi:10.1155/2012/213640.
- [10] Reginatto M (2009) What can we learn about the spectrum of high-energy stray neutron fields from Bonner sphere measurements? *Rad Measur* **44**, 692-699.
- [11] Dhairyawan M, Nagarajan P, Venkataraman G (1980) Response functions of spherically moderated neutron detectors. *Nuc Instr Meth* **169**, 115-120.
- [12] Rafiei Karahroudi M, Mousavi Shirazi S. A, Sepanloo K (2013) Optimization of designing the core fuel loading pattern in a VVER-1000 nuclear power reactor using the genetic algorithm. *Annals of Nuclear Energy* **57**, 142-150.
- [13] Vega H, Hernández V, Manzanares E, Sánchez M, Iñiguez M, Barquero R, Villafañe M, Arteaga t, Rodriguez J (2006) Neutron spectrometry using artificial neural networks. *Rad Measur* **41**, 425-431.
- [14] Rafiei Karahroudi M, Mousavi Shirazi S. A (2014) Obtaining the neutronic and thermal hydraulic parameters of the VVER-1000 Bushehr nuclear reactor core by coupling nuclear codes. *Kerntechnik* **79** (6), 528-531.
- [15] Trompier F, Battaglini P, Tikunov D, Clairand I (2008) Dosimetric response of human bone tissue to photons and fission neutrons. *Rad Measur* **43**, 837-840.
- [16] Bartesaghi G, Burian J, Gambarini G, Marek M, Negri A, Vierebl L (2009) Evaluation of all dose components in the LVR-15 reactor epithermal neutron beam using Fricke gel dosimeter layers. *Appl Rad Isot* **67**, 199-201.
- [17] *Annals of the ICRP (1977) Recommendations of the International Commission on Radiological Protection (ICRP). Publication 26.* Pergamon Press, New York.
- [18] Mousavi Shirazi S. A, Taheri A (2010) A NEW METHOD FOR NEUTRON CAPTURE THERAPY (NCT) AND RELATED SIMULATION BY MCNP4C CODE. *AIP Conf*, Kuala Lumpur, Malaysia, 77-83.

Seyed Alireza Mousavi Shirazi (Corresponding and the sole author); M.Phil graduated in nuclear energy engineering.

Simulation study of using shift registers based on 16th Degree Primitive Polynomials

Mirella Amelia Mioc

Abstract—Almost all of the major applications in the specific Fields of Communication used a well-known device called Linear Feedback Shift Register. Usually LFSR functions in a Galois Field $GF(2^n)$, meaning that all the operations are done with arithmetic modulo 2 degree Irreducible and especially Primitive Polynomials. Storing data in Galois Fields allows effective and manageable manipulation, mainly in computer cryptographic applications. The analysis of functioning for Primitive Polynomials of 16th degree shows that almost all obtained results are in the same time distribution.

Keywords—Cryptosystem, Irreducible polynomials, Pseudo-Random Sequence, Primitive Polynomials, Shift Registers.

I. INTRODUCTION

A code-breaking machine appeared as one of the first forms of a shift register early in the 40's, in Colossus. It was a five-stage device built of vacuum tubes and thyratrons. Many different implementation forms were developed along the years.

The LFSR (Linear Feedback Shift Register) is the basis of the stream ciphers and most often used in hardware designs. string of memory cells stored a string of bits and a clock pulse can advance the bits with one position in that string. For each clock pulse it is produced the new bit in the string using the XOR of certain positions. The basis of every LFSR is developed with a polynomial, which can be irreducible or primitive [4], [15]. A primitive polynomial satisfies some additional mathematical conditions and determines for the LFSR to have its maximum possible period, meaning $(2^n - 1)$, where n is the number of cells of the shift register or the length. LFSR can be built based on XOR (exclusive OR) circuits or XNOR (exclusive denied OR). The difference of status is, of course, the equivalent status will be 1, where it was 0. For an n bits LFSR, all the registers will be configured as shift registers, but only the last significant register will determine the feedback. An n bits register will always have $n + 1$ signals.

Shift registers are a form of sequential logic like counters.

Always the shift registers produce a discrete delay of a digital signal or waveform. Considering that a shift register has n stages, the waveform is delayed by n discrete clock times.

Usually the naming of the shift register follows a type of convention shown normally in digital logic, with the least significant bit on the left. According to the communication

protocol, the signals will be addressed, not the registers. There are $n+1$ signals for each n -bit register.

Always the next state of an LFSR is uniquely determined from the previous one by the feedback network. Any LFSR will generate a sequence of different states starting with the initial one, called seed.

A feedback shift register is composed of:

- a shift register
- a feedback function.

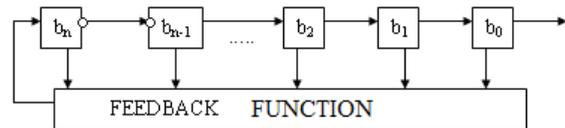


Fig. 1 Basis scheme for a Feedback Shift Register

An LFSR can be represented as a polynomial of variable x referred to as the generator polynomial or the characteristic polynomial. The input bit is given from a linear function of the initial status for a special shift register called Linear Feedback Shift Register (LFSR). The initial value of the register is called seed and the sequence produced is completely determined by the initial status.

Because the register has a finite number of possible statuses, after a period the sequence will be repeated. If the feedback function is very good chosen the produced sequence will be random and the cycle will be very long.

Reference [8] shows two possibilities to implement a LFSR:

- Fibonacci Form
- Galois.

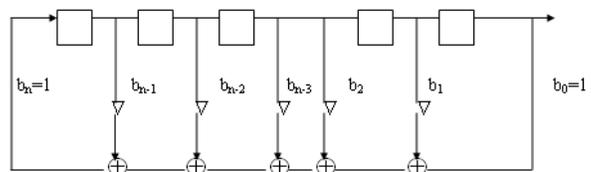


Fig. 2 Fibonacci implementation

In Fibonacci form the weight for any status is 0, when there isn't any connection and 1 for sending back.

Exceptions of this are the first and the last one, both connected, so always on 1.

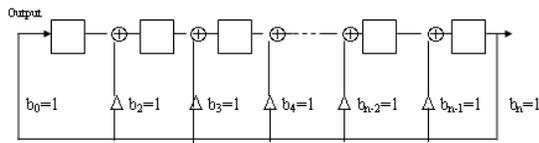


Fig. 3 Galois implementation

In Galois implementation there is a Shift Register, whose content is modified each step at a binary value sent to the output.

In Galois configuration the single bit shifted out is XOR ed with several bits in the shift register and in conventional configuration each new bit input to the shift register is the XOR of several bits in the register. Comparing the two scheme of representation it is shown that the weight order in Galois is opposite the one in Fibonacci. From the hardware point of view, because of the reduced number of XOR gates in feedback, Galois implementation is fastest than Fibonacci and so it is much more used. Some other used names for these two implementations are Simple Shift Register Generator (SSRG) for Fibonacci implementation and Multiple-Return Shift Register Generator (MRSRG) for the Galois one.

From the utilization point of view there are two kinds of LFSR: the well-known LFSR, that is an “in-tapping” LFSR and the “out-tapping” LFSR. The “in-tapping” LFSR is usually called a MISR (Multiple Input Shift Register). Cycle codes belong to algebraically codes for errors detecting. This experiment develops an analysis of a Linear Feedback Shift Register and a Multiple Input – output Shift Register. By using a primitive polynomial in the polynomials modulo 2 as modular polynomial in the polynomial multiplication it can be created a Galois Field of order 2^n with a polynomial beginning with x^n .

The most popular and widely used application of Galois Field is in Cryptography. Because all the data are represented as a vector in a finite field, encryption and decryption became easily to manipulate and straightforward by using mathematical arithmetic.

Such kind of field can be denoted as $GF(2^n)$ or $GF(n)$ and one of the famous applications for that is in the Rijndael Algorithm (AES), where $n=8$.

Beginning with 2000 Rijndael cryptosystem is officially the Advanced Encryption System (AES) [6].

The old DES (Data Encryption Standard) was broken from Electronic Frontier Foundation in three days [9]. The two authors Joan Daemen and Vincent Rijmen from Holland chose to use a Galois Field $GF(2^8)$ with the following generator polynomial.

$$P(x)=x^8+x^4+x^3+x+1$$

or ‘11B’ in hexadecimal representation.

All arithmetical operations are developed in a Galois group.

The Shift Register Cryptosystems variant has been developed from the evolution of the encrypting techniques [15]. Such a cryptosystem is based upon generating a sequence in a finite field and for obtaining it a Feedback Shift Register is used. There are some methods for using LFSR to build secure ciphers. For increasing the strength of the output from an LFSR, often it is used another LFSR for controlling how often it is stepped. Another technique uses three LFSRs with different periods and it is known as the Geffe generator. Usually it is necessary to combine the methods for obtaining more elaborate constructions. Almost all applications of using shift registers representing generator polynomials need to be developed in a finite field.

Evariste Galois demonstrated that a field is an algebra with both addition and multiplication forming a group. Some ground information from Algebra demonstrated the importance of working with irreducible polynomials and primitive polynomials. Also the importance of using shift registers in cryptosystems based on irreducible polynomials is demonstrated in increasing the security obtained.

Applications for using The Linear Feedback Shift Registers are in a variety of Fields:

- Testing [1], [18];
- Pattern Generators;
- Optimized counters [2];
- Checksums;
- Data Integrity;
- Data Encryption/ Decryption;
- Built-in Self-Test (BIST);
- Digital Signal Processing;
- Pseudo-random Number Generation (PN);
- Scrambler and/Descrambler;
- Signature Analyzer [3];
- Error Detection and Correction;
- Wireless communications.

II. MATHEMATICAL BACKGROUND

This finite field (FF) or Galois Field (GF) in abstract algebra is a field that contains only finitely many elements.

Finite fields are important in algebraic theory, number theory, Galois theory, cryptography and coding theory [5], [20].

It’s possible to classify the finite fields by size.

So, for each prime p and positive integer k there is exactly one finite field up to isomorphism of size p^k .

Each finite field of size q is the splitting field of the polynomial $x^q - x$.

A cyclic group is similarly the multiplicative group of the field.

Finite fields have applications in many areas of mathematics and computer science, including coding theory and others.

For most applications of $GF(2^n)$ to cryptography, the value of n is large and it is impossible to construct a complete look-up table for the field. In transmission of data the binary n -tuple representation $(a_0, a_1, \dots, a_{n-1})$ is used. The discrete log problem is that when is given the binary n -tuple representation

of an element in $GF(2^n)$ and it will find its power representation.

For security reasons it was demonstrated that the maximum number of pseudo-random sequences is obtained by using irreducible polynomials [19].

III. EXPERIMENTAL RESULTS AND MATHEMATICAL CALCULUS

The main subject of analyze the functioning of linear feedback shift register (LFSR) and multiple input output shift register (MISR) has the irreducible or primitive polynomials for degree 4, 8 and 16 [13].

All the analyze is based on the three possible implementations for LFSR [21].

First of all were developed programs for simulating the functioning for the three different types of implementations for comparing the obtained results for 4th degree irreducible polynomials [10].

Reference [11] shows a complete analyze and presentation of the functioning of LFSR for 8th degree irreducible polynomials.

Basic information concerning the comparative analysis of a LFSR and MISR has been specified in [12].

No	Octal	Binary
1	219913	1000100000001011
2	234313	10011100011001011
3	233303	10011011011000011
4	307107	11000111001000111
5	201735	10000001111011101
6	272201	10111010010000001
7	242413	10100010100001011
8	270155	10111000001101101
9	305667	11000101110110111
10	236107	10011110001000111
11	307527	11000111101010111
12	306357	11000110011101111
13	302157	11000010001101111
14	210205	10001000010000101

Table I. The 16th degree Primitive Polynomials

- $(X^{16}+X^5+X^3+X^2+1)$;
- $(X^{16}+X^{14}+X^{13}+X^{11}+1)$;
- $(X^{16}+X^5+X^4+X^3+1)$;
- $(X^{16}+X^{13}+X^{12}+X^{11}+1)$;
- $(X^{16}+X^5+X^4+X^3+X^2+X+1)$;
- $(X^{16}+X^{15}+X^{14}+X^{13}+X^{12}+X^{11}+1)$;
- $(X^{16}+X^6+X^4+X+1)$;
- $(X^{16}+X^{15}+X^{12}+X^{10}+1)$;
- $(X^{16}+X^7+X^5+X^4+X^3+X^2+1)$;
- $(X^{16}+X^{14}+X^{13}+X^{12}+X^{11}+X^9+1)$;
- $(X^{16}+X^7+X^6+X^4+X^2+X+1)$;
- $(X^{16}+X^{15}+X^{14}+X^{12}+X^{10}+X^9+1)$;
- $(X^{16}+X^8+X^5+X^3+X^2+X+1)$;
- $(X^{16}+X^8+X^5+X^4+X^3+X^2+1)$;
- $(X^{16}+X^{14}+X^{13}+X^{12}+X^{11}+X^8+1)$;
- $(X^{16}+X^8+X^6+X^3+X^2+X+1)$;
- $(X^{16}+X^{15}+X^{14}+X^{13}+X^{10}+X^8+1)$;

- $(X^{16}+X^8+X^6+X^4+X^3+X^2+1)$;
- $(X^{16}+X^{14}+X^{13}+X^{12}+X^{10}+X^8+1)$;
- $(X^{16}+X^{11}+X^9+X^8+1)$;
- $(X^{16}+X^8+X^7+X^5+X^3+X^2+1)$;
- $(X^{16}+X^{14}+X^{13}+X^{11}+X^9+X^8+1)$;
- $(X^{16}+X^8+X^7+X^5+X^4+X^3+X^2+X+1)$;

Table II. Some common 16th Degree Polynomials

It was developed a simulation program for the functioning on LFSR of 16th degree for the Galois implementation. In the following it will be presented an analyze for the 14 selected primitive polynomials. A list with the positions which will influence the future state is called tap sequence. For example, for the next scheme this is [16, 14, 13, 11].

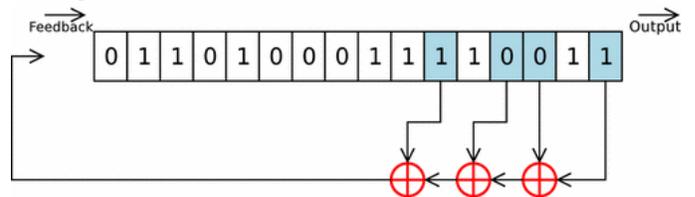


Fig. 4 Scheme for the polynomial with [16, 14, 13, 11] tap sequence

This sequence can be represented by a polynomial mod 2 with coefficients only 1 and 0, called Feedback Polynomial or Characteristic Polynomial.

For the above scheme this polynomial is:

$$P(X) = X^{16} + X^{14} + X^{13} + X^{11} + 1 ;$$

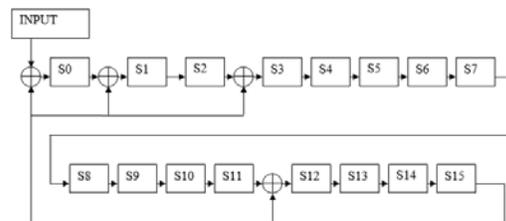


Fig. 5 Galois Implementation for the Polynomial $X^{16} + X^{14} + X^{13} + X^{11} + 1$

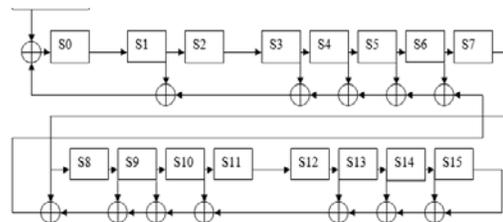


Fig. 6 Fibonacci Implementation for the Polynomial $X^{16} + X^{12} + X^3 + X + 1$

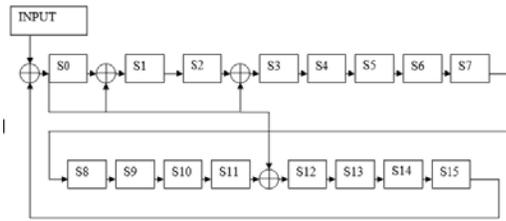


Fig. 7 Ring Implementation for the Polynomial $X^6+X^{12}+X^3+X+1$

Reference [21] shows three types of scheme for a 4th degree polynomial.

Similar to it, were developed the three different implementation for the Primitive Polynomial $X^{16}+X^{12}+X^3+X+1$.

It can be specified that these schemes are according to the well-known Galois Form, Fibonacci Representation and some other Form, rarely used, called Ring Implementation.

All of these Implementations have the same function, because describe a linear feedback shift register. In the experimental work has been analyzed the behavior for 14 Primitive Polynomial degree 16 randomly selected .For each of these polynomial it has been developed by a program the simulation of the specific functioning.

Because the goal of this experimental work was to compare the different obtained results, a few rows of input data of a different length have been selected.

First of all in this analyze it was verified with the simulation program the obtained times and they have been compare for each program

All the analyze submit the results obtained in the case of use Galois implementation. Times have been measured from 10 runs for the same string of input data and calculated average time.

For each Primitive Polynomial 6 different input data were used.

In the following tables all the obtained results are shown by carrying out their time. So, for the 14 selected Polynomials and for the input data lengths of 20, 30, 40, 50, 100 and 1000 the corresponding times are presented (display).

The following table contains the time measured in seconds.

	Time 20	Time 30	Time 40
Prel 1	0.00003471	0.00005425	0.00006220
Prel 2	0.00003864	0.00005288	0.00007746
Prel 3	0.00003761	0.00005422	0.00007319
Prel 4	0.00003562	0.00005112	0.00006738
Prel 5	0.00003662	0.00005137	0.00007063
Prel 6	0.00003763	0.00005172	0.00007026
Prel 7	0.00004377	0.00005093	0.00007052

Prel 8	0.00003557	0.00005649	0.00006912
Prel 9	0.000039022	0.000053175	0.00007273
Prel 10	0.00003513	0.00005102	0.00007181
Prel 11	0.00003442	0.00005404	0.00006725
Prel 12	0.00004232	0.00005098	0.00006978
Prel 13	0.00003514	0.00005345	0.00006794
Prel 14	0.00003489	0.00006187	0.00007052

Table III. Results of the main program

	Time 50	Time 100	Time 1000
Prel 1	0.00023055	0.00022300	0.00173524
Prel 2	0.00010348	0.00022504	0.00215812
Prel 3	0.00008803	0.00022427	0.00224510
Prel 4	0.00008655	0.00022607	0.00242678
Prel 5	0.00009081	0.00023949	0.00211611
Prel 6	0.00008708	0.00022512	0.00226576
Prel 7	0.00008897	0.00021033	0.00216120
Prel 8	0.00008610	0.00021813	0.00209186
Prel 9	0.0001895	0.00209705	0.00348371
Prel 10	0.00008959	0.00028245	0.00209705
Prel 11	0.00008359	0.00028720	0.00214601
Prel 12	0.00008981	0.00023582	0.00203451
Prel 13	0.00008942	0.00023207	0.00214814
Prel 14	0.00008507	0.00021405	0.00236788

Table IV. Results of the main program

Generator Polynomial	Coef. no.
$X^{16}+X^{12}+X^3+X+1$	5
$X^{16}+X^{13}+X^{12}+X^{11}+X^7+X^6+X^3+X+1$	9
$X^{16}+X^{13}+X^{12}+X^{10}+X^9+X^7+X^6+X+1$	9
$X^{16}+X^{15}+X^{11}+X^{10}+X^9+X^6+X^2+X+1$	9
$X^{16}+X^9+X^8+X^7+X^6+X^4+X^3+X^2+1$	9
$X^{16}+X^{14}+X^{13}+X^{12}+X^{10}+X^7+1$	7
$X^{16}+X^{14}+X^{10}+X^8+X^3+X+1$	7
$X^{16}+X^{14}+X^{13}+X^{11}+X^6+X^5+X^3+X^2+1$	9
$X^{16}+X^{15}+X^{11}+X^9+X^8+X^7+X^5+X^4+X^2+X+1$	11
$X^{16}+X^{13}+X^{12}+X^{11}+X^{10}+X^6+X^2+X+1$	9

$X^{16}+X^{15}+X^{11}+X^{10}+X^9+X^8+X^6+X^4+X^2+X+1$	11
$X^{16}+X^{15}+X^{11}+X^{10}+X^7+X^6+X^5+X^3+X^2+X+1$	11
$X^{16}+X^{15}+X^{10}+X^6+X^5+X^3+X^2+X+1$	9
$X^{16}+X^{12}+X^7+X^2+1$	5

Table V. Coefficients number of the Generator Polynomials

In the following all of the steps that simulates operation using the first Primitive Polynomial and .the selected input data are showed. The used operations are shifting and XOR.

110101010101010101

Checking input data: 1 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1

Initial Status 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Step 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Step 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0

Step 2 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0

Step 3 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0

Step 4 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0

Step 5 1 0 1 0 1 1 0 0 0 0 0 0 0 0 0

Step 6 0 1 0 1 0 1 1 0 0 0 0 0 0 0 0

Step 7 1 0 1 0 1 0 1 1 0 0 0 0 0 0 0

Step 8 0 1 0 1 0 1 0 1 1 0 0 0 0 0 0

Step 9 1 0 1 0 1 0 1 0 1 1 0 0 0 0 0

Step 10 0 1 0 1 0 1 0 1 0 1 1 0 0 0 0

Step 11 1 0 1 0 1 0 1 0 1 0 1 1 0 0 0

Step 12 0 1 0 1 0 1 0 1 0 1 0 1 1 0 0

Step 13 1 0 1 0 1 0 1 0 1 0 1 0 1 1 0

Step 14 0 1 0 1 0 1 0 1 0 1 0 1 0 1 1

Step 15 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1

Step 16 1 0 0 0 0 1 0 1 0 1 0 1 1 1 0

Step 17 0 0 1 0 0 1 0 1 0 1 0 0 1 1 0

Step 18 0 0 0 1 0 0 1 0 1 0 1 0 0 1 1

Step 19 0 1 0 1 0 1 0 0 1 0 1 0 0 0 1

Runing Time : 0.00003421 seconds

The next graphics show the obtained results from the execution of the main program for each of all 14 degree 16th primitive polynomials for three different situations depending on the lengths of the entrance data polynomial.

The lengths of the input polynomials were 20. 30. 40, 50, 100 and 1000 bits. The maximum number of sequences is $2^{16}-1$ [17].

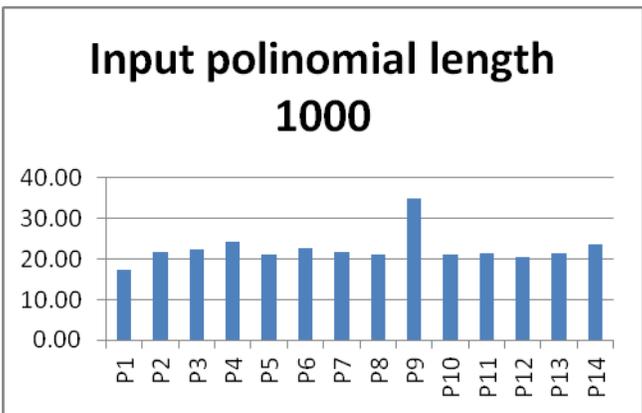


Fig. 8 Graphic containing the results for 1000 bits

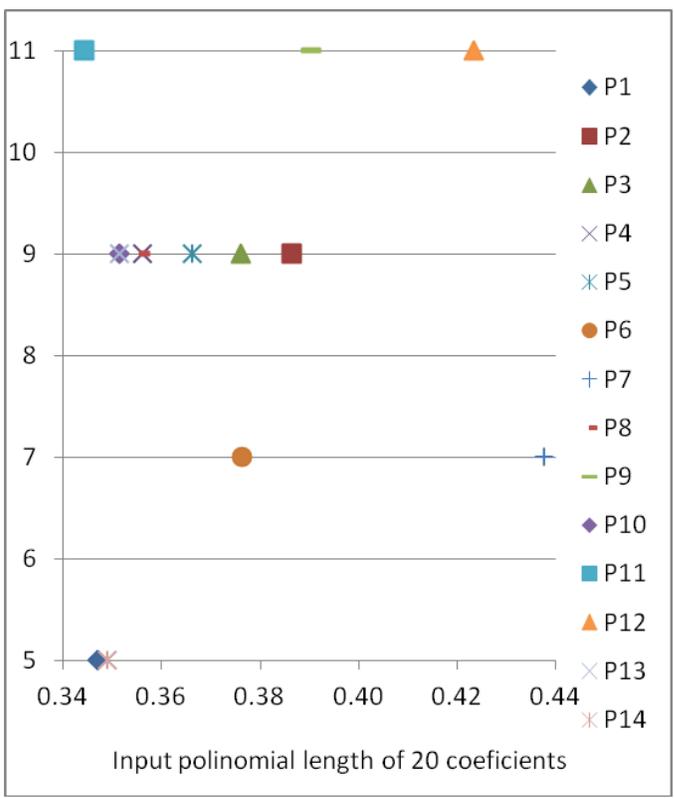


Fig. 9 Graphic containing the results for 20 bits

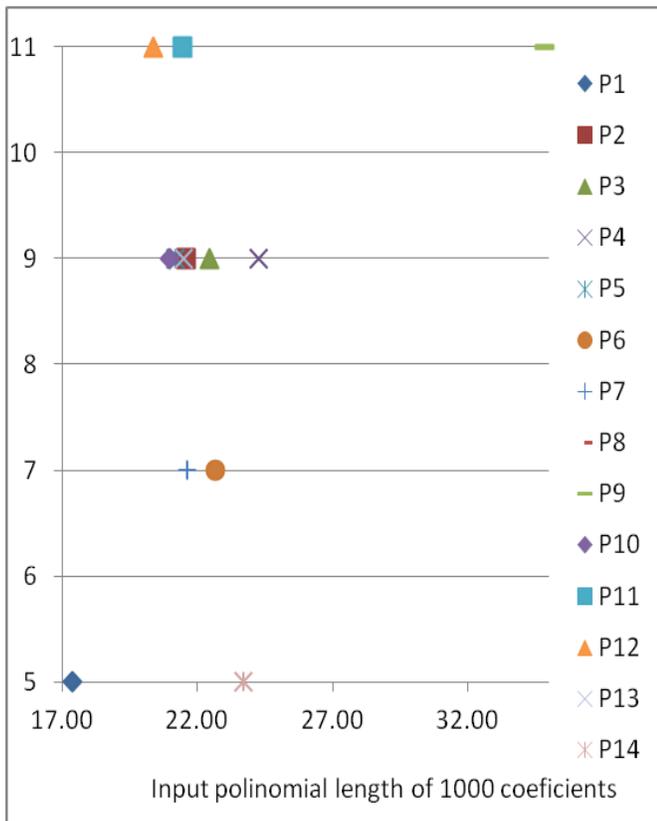


Fig. 10 Graphic containing the results for 1000 bits

Distribution obtained depending on the length of the input string shows that time depends of input length, but for lengths even close together, the times are also close (this can be seen in Fig. 9 for 20 bits inputs).

Time does not change so much depending on which one of the 14 different 16th degree primitive polynomials has been used.

IV. CONCLUSION

Entire analysis of functioning for primitive polynomials of 16th degree proves that almost all obtained results are in the same distribution of time. The aspect of security was taken into consideration, so that the used polynomials are all irreducible or primitive polynomials. A shift register is a device whose function is to shift its contents into adjacent positions within the register or, for the end position, out of the register.

The main practical uses for a shift register are the convert between parallel and serial data and the delay of a serial bit stream. The total number of the generated random state is depending on the feedback polynomial.

LFSR based on the PN Sequence Generator is a technique used for different applications in Cryptography and also in communication channel for designing encoder and decoder. Such kind of analyze can be done from the hardware or software point of view. Reference [14] shows an interesting simulation problem for long bit LFSR on FPGA referring 8,

16 and 32 Bit. Some similar problems are presented and analyzed in this paper.

This study focuses on a comparative study of different types of implementations for a Linear Feed-back Shift Register for 16th degree primitive or irreducible polynomials. The results of all these experiments were used for obtaining some graphics showing the time distribution.

REFERENCES

- [1] Abramovici M., Breuer M. A., Friedman A. D., Digital Systems Testing and Testable Design, Computer Science Press,1990;
- [2] Alfke P., Efficient Shift Registers, LFSR, Counters, and Long Pseudo-Random Sequence Generators, XAPP 052, July 7,1996.
- [3] Alvarez R., Martinez F.-M., Vincent J.-F., Zamora A., A Matricial Public Key Cryptosystem with Digital Signature, *WSEAS TRANSACTIONS on MATHEMATICS* Manuscript , vol.7,No.4,2008, pp. 195-204.
- [4] Angheloiu I., Gyorfı E., Patriciu V.V., Securitatea și protecția informației în sistemele electronice de calcul, Ed. Militară, București, 1986.
- [5] Berlekamp E. R., Algebraic Coding Theory, McGraw-Hill, New York, 1968.
- [6] Daemen J., Rijmen V., "The Design of Rijndael: AES - The Advanced Encryption Standard", Springer-Verlag, 2002.
- [7] Golomb S. W., Shift Register Sequences, Holden-Day, San Francisco, Calif., 1967.
- [8] Goresky M., Klapper A., Fibonacci And Galois Representations of Feedback with Carry Shift Registers, December 4, 2004;
- [9] Matsui M., The First Experimental Cryptanalysis of the Data Encryption Standard .In *Advances in Cryptology, Proceedings of Crypto'94*, LNCS 839, Y. Desmedt, Ed., Springer- Verlag, 1994.
- [10] Mioc M. A., Simulation study of the functioning of LFSR for grade 4 Irreducible Polynomials, *WSEAS Conference ISPRA*, 21-23 February, 2009.
- [11] Mioc M. A., Study of Using Shift Registers in Cryptosystems for Grade 8 Irreducible Polynomials, *WSEAS Conference SMO*, 23-25 September 2008.
- [12] Mioc M. A., An analyze of functioning for a linear feed-back shift register and a multiple input-output shift register, *Buletinul Stiintific al Universitatii „Politehnica” din Timisoara, Seria ELECTRONICA si TELECOMUNICATII*, Transactions on electronics and communications, Tom 50(64), Fascicola 2, 2005.
- [13] Mioc M. A., A complete analyze of using Shift Registers in Cryptosystems for Grade 4, 8 and 16 Irreducible Polynomials”, *WSEAS Transactions on Computers*, Volume 7, Issue 10, ISSN: 1109-2750, October, 2008.
- [14] Panda A. K.,Rajput P., Shukla B., FPGA Implementation of 8, 16 and 32 Bit LFSR with

Maximum Length Feedback Polynomial Using VHDL,
Rajkot, India, 2012.

- [15] Schneier B., *Applied Cryptology: Protocols, Algorithms, and Source Code in C*, John Wiley and Sons, New York, 1996;
- [16] Shannon C.E., *Mathematical Theory of Communication*, 1948.
- [17] Solomon G., *Shift register sequences*, Aegean Park Press, Laguna Hills, Canada, 1967.
- [18] Tsui F., *LSI/VLSI Testability Design*, McGraw-Hill Book Company, 1987.
- [19] Udar S., Kagaris D., *LFSR Reseeding with Irreducible Polynomials*, *IOLTS 2007*, pp. 293-298
- [20] Van Lint J.H., *Introduction to Coding Theory*, 2nd ed., Springer-Verlag, USA, 1992.
- [21] Vlăduțiu M., Crișan M., *Tehnica testării echipamentelor automate de prelucrare a datelor*, Editura Facla, Timișoara, 1989.

Topological optimization of lake aeration process

Mohamed Abdelwahed

Abstract—This work deals with the optimization of the injectors location in the aeration process of water reservoir. This application is used to combat oxygen depletion in water due to the eutrophication. A simplified model is then used for illustrate the direct simulation of the air dynaminc effect on water. For the optimization problem, we used the topological sensitivity analysis method.

Keywords—Two phase flow, Three dimensional Navier-Stokes equations, topological optimization, topological sensitivity.

I. INTRODUCTION

THE mechanical aeration process in water reservoirs is one of the most used techniques to combat eutrophication. It consists on pumping a source of compressed air in the reservoir bottom via injectors in order to create a dynamic and aerate the water by bringing it in contact with the surface air. We focus in this work in the first hand to the direct problem. It concerns the numerical simulation of the resulting two phase water air-bubbles flow. Different models can be used to describe this problem [2], [4], [5], [9]. Using the fact that the water phase is dominant. We used a simplified model in which the water phase is governed by the Navier-Stokes equations and the aeration effects are taken into account through a local boundary condition for the velocity on the injector holes. Our discretization method is based on three dimensional mixed finite element method $P^1 + \text{bubble}/P^1$ [3]. The Uzawa algorithm is used to solve the obtained matrix system.

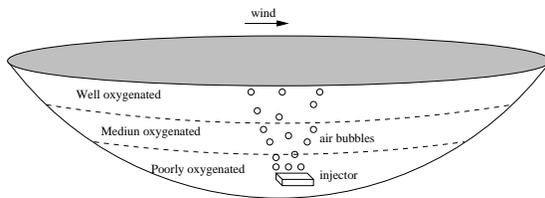


Fig. 1. Aeration process

M. Abdelwahed is with the Department of Mathematics, College of Science, King Saud University, Riyadh 11451, Kingdom of Saudi Arabia e-mail: mabdelwahed@ksu.edu.sa.

In the other hand, we look at the inverse problem: find the optimal injectors location generating the best motion in the fluid with respect to the aeration purpose. The optimal injectors location is characterized as the solution to a topological optimization problem. The topological sensitivity analysis is used to solve this problem [6], [7], [8], [10]. The main idea is to compute the asymptotic topological expansion with respect to the insertion of an injector.

The paper is organized as follows. The used model, its numerical analysis and a direct numerical simulation is presented in section 2. Section 3 is devoted to a topological sensitivity analysis for the Quasi-Stokes equations. The obtained results are valid for a large class of cost functions. Finally, we illustrate the efficiency of the proposed method by a numerical test.

II. DIRECT SIMULATION

Let Ω be a three dimensional flow domain representing the eutrophized water reservoir. The used model is based on three dimensional Navier-Stokes equations for water flow in which we integrate the effect of momentum released by the injected bubbles by adding a local boundary condition for the velocity on the injector holes.

In the presence of an injector $\omega_{inj} \subset \Omega$, the velocity $u(x, t)$ and the pressure $p(x, t)$ solve the following system

$$\left\{ \begin{array}{ll} \text{Find } u \text{ and } p \text{ solutions of} & \\ \frac{\partial u}{\partial t} + u \cdot \nabla u - \nu \Delta u + \nabla p = \mathcal{G} & \text{in } \Omega_i \times [0, T] \\ \text{div } u = 0 & \text{in } \Omega_i \times]0, T[\\ u^0 = u_0 & \text{in } \Omega_i \\ u = u_d & \text{on } \Gamma_i \times]0, T[\end{array} \right. \quad (1)$$

where $\Omega_i = \Omega \setminus \overline{\omega_{inj}}$ is the water reservoir domain in the presence of the injector ω_{inj} , ν is the water viscosity, \mathcal{G} is the gravitational force, T is the final time of simulation, u_0 is the initial velocity field. $\Gamma_i = \Gamma_s \cup \Gamma_w \cup \partial\omega_{inj}$ the boundary of Ω_i where Γ_s : the surface in contact with the atmosphere,

Γ_w : the bottom water reservoir boundary, ω_{inj} : the injector boundary.

$$u_d = \begin{cases} u_{wind} & \text{on } \Gamma_s, \\ 0 & \text{on } \Gamma_w, \\ u_{inj} & \text{on } \partial\omega_{inj}. \end{cases} \quad (2)$$

Using characteristics method in (1), we obtain

$$\begin{cases} \text{Find } u^{n+1} \text{ and } p^{n+1} \text{ solutions of} \\ \alpha u^{n+1} + \nu \Delta u^{n+1} + \nabla p^{n+1} = F^{n+1} & \text{in } \Omega_i \\ \text{div } u^{n+1} = 0 & \text{in } \Omega_i \\ u^{n+1} = u_d & \text{on } \Gamma_i, \end{cases} \quad (3)$$

where $\alpha = \frac{1}{\Delta t}$, $F^{n+1} = \frac{1}{\Delta t} u^n \circ \chi^n + \mathcal{G}$, u^{n+1} and p^{n+1} are the approximations of u and p on time $t^{n+1} = (n+1)\Delta t$ and $\chi^n(x) = X^n(t^{n+1}; x)$ represents the position at time t^{n+1} of the particle of fluid which is at point x at time t^n .

System (3) is solved iteratively for $n = 0, 1, \dots$. At each time step, we have to solve a steady state problem of Quasi-Stokes type having the following generic form.

For $F \in L^2(\Omega_i)^3$, and $u_d \in H^{\frac{1}{2}}(\Gamma_i)^3$ such that $\int_{\Gamma_i} u_d \cdot n \, ds = 0$, find u in $H^1(\Omega_i)^3$ and p in $L_0^2(\Omega_i)$ solutions of the problem :

$$\begin{cases} \alpha u - \nu \Delta u + \nabla p = F & \text{in } \Omega_i \\ \text{div } u = 0 & \text{in } \Omega_i \\ u = u_d & \text{on } \Gamma_i. \end{cases} \quad (4)$$

Using ‘P1 +bubble/P1’ mixed finite element method (see [3]) for the space approximation, we derive a linear matrix system. The resolution is based on Uzawa method and conjugate gradient algorithm [1].

For the numerical simulation, we used the following boundary conditions: $u_{wind} = 0.01m/s$ the wind velocity at the surface, no slip condition at the bottom and $u_{inj} = 0.1m/s$ the injection velocity on the injector. We present in figure 2 the numerical simulation of the aeration effect on the water flow in a three dimensional domain containing one injector obtained for $T = 10mn$. This result shows that the aeration effect is located in the region between the injector and the top surface. Then we have to use more injectors in order to aerate all the water reservoir domain. For this reason, we are interested in the following optimization problem: for a given injectors number, find their optimal location generating the best motion in the water.

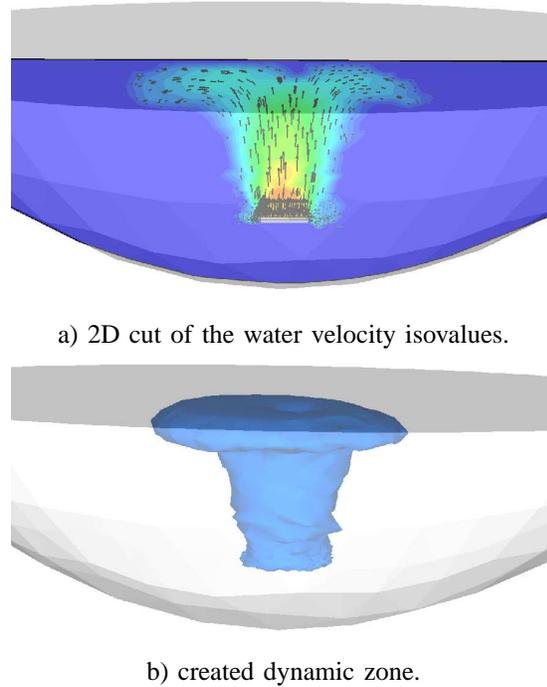


Fig. 2. Numerical simulation of the aeration process

III. OPTIMIZATION PROBLEM

Our aim in this section is to design an efficient method to optimize the injectors location in order to generate the best motion of the fluid.

For the sake of simplicity, we shall assume that the injectors are well separated and have the geometry form $\omega_{z_k, \varepsilon} = z_k + \varepsilon \omega^k$, $1 \leq k \leq m$, where ε is the shared diameter and $\omega^k \subset \mathbb{R}^3$ are bounded and smooth domains containing the origin. The points $z_k \in \Omega$, $1 \leq k \leq m$ determine the location of the injectors. The domains ω^k describe the injectors geometries.

Here we limit ourselves to the steady state system described by the Quasi-Stokes equations (4). Then, in the presence of injectors, the velocity u_ε and the pressure p_ε satisfy the following equations

$$\begin{cases} \alpha u_\varepsilon - \nu \Delta u_\varepsilon + \nabla p_\varepsilon = F & \text{in } \Omega \setminus \bigcup_{k=1}^m \overline{\omega_{z_k, \varepsilon}} \\ \nabla \cdot u_\varepsilon = 0 & \text{in } \Omega \setminus \bigcup_{k=1}^m \overline{\omega_{z_k, \varepsilon}} \\ u_\varepsilon = u_d & \text{on } \Gamma \\ u_\varepsilon = u_{inj}^k & \text{on } \partial\omega_{z_k, \varepsilon}, \quad 1 \leq k \leq m, \end{cases} \quad (5)$$

where u_{inj}^k is the injection velocity of the injector $\omega_{z_k, \varepsilon}$.

Consider now a design function j having the form

$$j(\Omega \setminus \bigcup_{k=1}^m \overline{\omega_{z_k, \varepsilon}}) = J_\varepsilon(u_\varepsilon), \quad (6)$$

where J_ε is a given cost function describing the optimization criteria and u_ε is the solution of (5). Our identification problem can be formulated as a topological optimization problem: find the optimal location of the injectors $\omega_{z_k, \varepsilon} = z_k + \varepsilon \omega^k$, $1 \leq k \leq m$, inside the water reservoir domain Ω minimizing the function j :

$$(\mathcal{P}) \begin{cases} \text{Find } z_k^* \in \Omega, 1 \leq k \leq m, \text{ such that :} \\ j(\Omega \setminus \cup_{k=1}^m \overline{\omega_{z_k^*, \varepsilon}}) \\ = \min_{\omega_{z_k, \varepsilon} \subset \Omega} j(\Omega \setminus \cup_{k=1}^m \overline{\omega_{z_k, \varepsilon}}). \end{cases}$$

To solve this optimization problem (\mathcal{P}) we shall use the topological gradient method. It consists in studying the variation of the design function j with respect to a small topological perturbation of the domain Ω .

IV. TOPOLOGICAL SENSITIVITY ANALYSIS

In this section we derive a topological asymptotic expansion of the design function j with respect to the insertion of a small injector $\omega_{z, \varepsilon} = z + \varepsilon \omega$ inside the domain Ω . Next we assume that J_ε satisfies the following assumptions.

Hypothesis

- i) J_0 is differentiable with respect to u , its derivative being denoted by $DJ_0(u)$.
- ii) There exists a real number δJ such that $\forall \varepsilon \geq 0$

$$J_\varepsilon(u_\varepsilon) - J_0(u_0) = DJ_0(u_0)(\hat{u}_\varepsilon - u_0) + \varepsilon \delta J + o(\varepsilon), \quad (7)$$

where \hat{u}_ε denotes the extension of u_ε in Ω defined by $u_\varepsilon = u_{inj}$ in $\omega_{z, \varepsilon}$.

We are now ready to derive the topological asymptotic expansion of the design function j . It consists in computing the variation $j(\Omega \setminus \overline{\omega_{z, \varepsilon}}) - j(\Omega)$ when inserting a small injector inside the domain. The asymptotic expansion described in Theorem 4.1 is valid for arbitrary shaped holes and all cost function verifying the above Hypothesis.

Theorem 4.1: If Hypothesis i) and ii) holds, the function j has the following asymptotic expansion

$$j(\Omega \setminus \overline{\omega_{z, \varepsilon}}) - j(\Omega) = \varepsilon \left[\left(- \int_{\partial \omega} \eta(y) \, ds(y) \right) \cdot v_0(z) + \delta J \right] + o(\varepsilon),$$

where v_0 is the solution to the adjoint problem

$$\begin{cases} \alpha v_0 - \nu \Delta v_0 + \nabla q_0 = -DJ_0(u_0) & \text{in } \Omega \\ \nabla \cdot v_0 = 0 & \text{in } \Omega \\ v_0 = 0 & \text{on } \Gamma. \end{cases}$$

and $\eta \in H^{-1/2}(\partial \omega)^3$ is the solution to the boundary integral equation.

$$\int_{\partial \omega} E(y-x) \eta(x) \, ds(x) = u_{inj} - u_0(z) \quad \forall y \in \partial \omega. \quad (8)$$

with (E, Π) the fundamental solution of the Stokes equations

$$E(y) = \frac{1}{8\pi\nu r} (I + e_r e_r^T), \quad \Pi(y) = \frac{y}{4\pi r^3},$$

with $r = \|y\|$, $e_r = y/r$ and e_r^T is the transposed vector of e_r .

Corollary 4.1: If $\omega = B(0, 1)$, the density η is given explicitly : $\eta(y) = -\frac{3\nu}{2} u_0(z) \forall y \in \partial \omega$ and under the hypothesis of theorem 4.1, we have

$$j(\Omega \setminus \overline{\omega_{z, \varepsilon}}) - j(\Omega) = \varepsilon \left[6\pi\nu u_0(z) \cdot v_0(z) + \delta J \right] + o(\varepsilon).$$

V. NUMERICAL RESULTS

Aeration is considered as the best remedial action against eutrophication. This process consists in inserting some injector holes ω_k in the bottom layer of the reservoir in order to create a dynamic and aerate the water. We suppose that a ‘‘good’’ water reservoir aeration can be described by a target velocity \mathcal{U}_g (see figure 4). Our aim is to determine the optimal location in Ω of some injector holes ω_k , $1 \leq k \leq m$ in order to minimize the function

$$J_\varepsilon(u_\varepsilon) = \int_{\Omega_m} |u_\varepsilon - \mathcal{U}_g|^2 \, dx, \quad (9)$$

where $\Omega_m \subset \Omega$ is the measurement domain (the top layer).

The following Proposition describes the variation of the associated design function j with respect to the insertion of a small injector $\omega_{z, \varepsilon} = z + \varepsilon B(0, 1)$ inside the domain Ω .

Proposition 5.1: The cost function J_ε defined in (9) satisfies the Hypothesis i) and ii) with

$$DJ_0(u_0)(w) = 2 \int_{\Omega_m} (u_0 - \mathcal{U}_g) w \, dx, \quad \forall w \in \mathcal{V}_0 \text{ and } \delta J = 0. \quad (10)$$

Then, the design function j has the following expansion

$$j(\Omega \setminus \overline{\omega_{z, \varepsilon}}) - j(\Omega) = 6\pi\nu u_0(z) \cdot v_0(z) + o(\varepsilon).$$

The optimal location of the injectors $\omega_k = z_k + \varepsilon B(0, 1)$, $1 \leq k \leq m$ is obtained using the following topological optimization algorithm.

The algorithm :

- Initialization: choose $\Omega_0 = \Omega_b$, and set $k = 0$.
- Repeat until target is reached:
 - compute u_k and v_k , respectively solutions to direct and adjoint problems in Ω_k ,
 - compute the topological sensitivity δj_k ,
 - set $\Omega_{k+1} = \{x \in \Omega_k, \delta j_k(x) \geq c_{k+1}\}$ where c_{k+1} is chosen in such a way that the cost function decreases,
 - $k \leftarrow k + 1$.

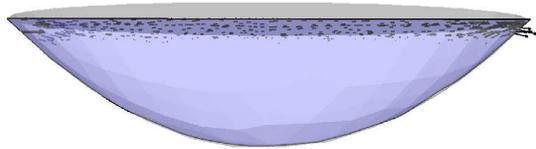


Fig. 3. The initial flow u^0

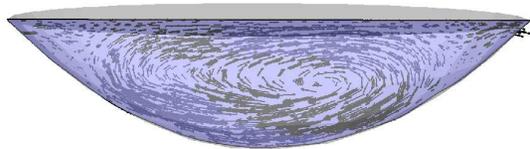


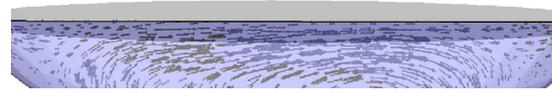
Fig. 4. The wanted flow U_g

This algorithm can be seen as a descent method where the descent direction is determined by the topological sensitivity δj_k and the step length is given by the volume variation $\Omega_k \setminus \Omega_{k+1}$.

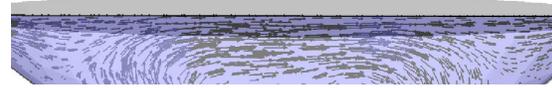
We propose an adaptation of the previous algorithm to our context. We consider the set $\{x \in \Omega_k; \delta j_k(x) < c_{k+1}\}$ Each connected component of this set is a hole created by the algorithm. Our idea is to replace each hole by an injector located at the local minimum of $\delta j_k(x)$.

For this numerical test, we consider in figure 4 a constructed solution representing the velocity field U_g . This solution is obtained by the dynamic aeration process using more than 1000 injectors located at the bottom layer Ω_b . We aim to find the optimal location of a fixed number of injectors m in order to approximate the wanted flow U_g .

Using our algorithm with 25 injectors (i.e. $m = 25$), we show in figure 7 the obtained flow during the optimization process at iterations 1, 3 and 5. The optimal injectors location is given un figure 6. Figure 5 shows a vertical cut of the wanted and obtained flows in the measurement domain Ω_m .



The wanted velocity U_g in Ω_m



The obtained velocity $u|_{\Omega_m}$

Fig. 5. Velocities field obtained in Ω_m (measurement domain) at the end of the optimization process.

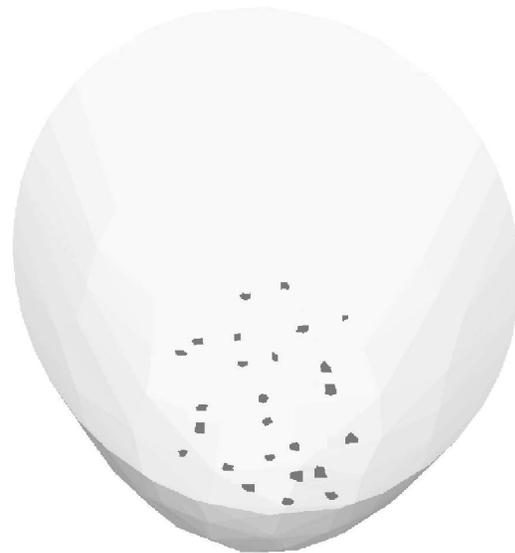
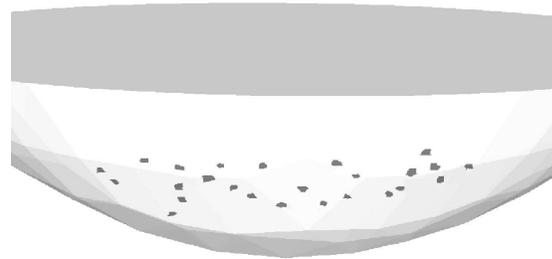


Fig. 6. Injectors location obtained during the optimization process: lateral view (left) and top view (right)

We remark that we obtain approximately the same flow.

This work can be considered as a preliminary step to study the transient Navier-Stokes problem.

REFERENCES

[1] M. Abdelwahed, M. Amara, Numerical analysis of a two phase flow, Journql of Computational methods, 9(3), 2012.

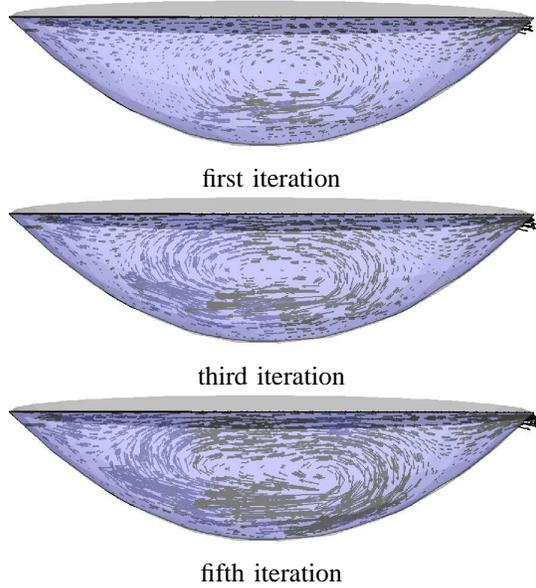


Fig. 7. Velocities field obtained during the optimization process.

- [2] M. Abdelwahed, M. Amara, D. Ouazar, A virtual numerical simulator for aeration effects in lake eutrophication, *Int. J. Comput. Fluid. Dynamics*. 16(2), 119-128; 2002.
- [3] D. Arnold, F. Brezzi, M. Fortin, A stable finite element for the Stokes equations, *Calcolo*, 21(4), 337-344, 1984.
- [4] E. Clement, Dispersion de bulles et modifications du mouvement de la phase porteuse dans des écoulements tourbillonnaires, Ph.D, Institut Nationale polytechnique de Toulouse,, 1999.
- [5] D. Legende, Quelques aspects des forces hydrodynamiques et des transferts de chaleur sur une bulle sphérique, Ph.D, Institut Nationale polytechnique de Toulouse, 1996.
- [6] Ph. Guillaume, K. Sid Idris, Topological sensitivity and shape optimization for the Stokes equations, *SIAM J. Control Optim.* 43(1), 1-31, 2004.
- [7] M. Hassine, S. Jan, M. Masmoudi, From differential calculus to 0 – 1 topological optimization, *SIAM J. Control Optim.*, 45(6), 1965-1987, 2004.
- [8] M. Hassine, S. Jan, M. Masmoudi, The topological asymptotic expansion for the Quasi-Stokes problem, *ESAIM, COCV J.*, 10(4), 478-504, 2004.
- [9] M. Ishii, Thermo-fluid dynamic theory of a two-phase flow, *Collection de la direction des études de recherche d'électricité de france*, 1975.
- [10] J. Sokolowski, A. Zochowski, On the topological derivative in shape optimization, *SIAM J. Control Optim.*, 37(4), 1251-1272, 1999.

Rigid non-Archimedean spaces and applications

NIKOLAJ GLAZUNOV
National Aviation University
Department of Electronics
1 Cosmonaut Komarov Avenue, 03680 Kiev
UKRAINE
glanm@yahoo.com

Abstract: Rigid non-Archimedean spaces, problems, methods and results are presented. At first we give an elementary algebraic and algebraic geometry introduction to formal schemes, groups and stacks and to rigid non-Archimedean spaces in the framework of local one dimensional complete regular rings of any characteristics, modules over rings and trees. Next we give a brief account of certain results and problems which arise in the theory of rigidity and which are connected with some problems of mathematical physics.

Key-Words: Rigid non-Archimedean space, module over a ring, lattice, formal scheme, rigidity, group action, Picard-Fuchs equation, Gauss-Manin connection, Frobenius lift, Frobenius structure, effective convergence bounds

1 Introduction

The aim of this paper is to give a brief account of certain results and problems which arise in the theory of rigidity and which are connected with some problems of mathematical physics. In the framework the rigid problems, methods and results in arithmetic geometry and dynamics are presented. At first we give an elementary algebraic and algebraic geometry introduction to rigid non-Archimedean spaces in the framework of local one dimensional complete regular rings of any characteristics, modules over rings, trees and formal schemes. Next we review Frobenius structures on connections and Newton strata in the loop group of a reductive group.

2 Formal groups and formal stacks

Here we present results on two-dimensional commutative formal groups and on formal stacks

2.1 On two-dimensional commutative formal groups

Let F be a commutative formal group law of n variables over commutative ring R with unit. In the case $n = 1$, following to the known results by M. Lazard, there is only one 1-bud of the form $x + y + \alpha xy$.

Proposition 1 *Let $n = 2$, $A = \mathbb{Z}_p[\alpha, \beta]$ be the ring of polynomials with integer p -adic coefficients from*

α, β . 1-buds are

$$F(x, y) = \begin{cases} x_1 + y_1 + \alpha x_1 y_1 \\ x_2 + y_2 + \beta x_2 y_2, \end{cases}$$

$$F_a(x, y) = \begin{cases} x_1 + y_1 + \alpha x_1 y_1 \\ x_2 + y_2 + \beta x_1 y_1, \end{cases}$$

$$F_b(x, y) = \begin{cases} x_1 + y_1 + \alpha x_2 y_2 \\ x_2 + y_2 + \beta x_2 y_2, \end{cases}$$

$$F_c(x, y) = \begin{cases} x_1 + y_1 + \alpha(x_1 + x_2)(y_1 + y_2) \\ x_2 + y_2 + \beta(x_1 + x_2)(y_1 + y_2), \end{cases}$$

Remark 2 *1-buds given in Proposition 1 are also two-dimensional formal group laws, whose coefficients under terms of degrees ≥ 3 are zeros.*

Remark 3 *These group laws define classes of group laws. In particular, the class F_a contains under values of parameters $\alpha = 0, \beta = -1$, the Witt group, that corresponds to prime number $p = 2$.*

2.2 Formal stacks

Let now the ring R is the field k . Recall, that formal k -scheme is formal k -functor, that is the limit of directed inductive system of finite k -schemes, and a formal group is a group object in the category of formal k -schemes. The notion of a stack, as one of category theory variants of moduli space is defined by P. Deligne and D. Mumford.

Proposition 4 *There exist formal stacks, that are categories that are bundled on formal groupoids and that satisfy axioms of decent theory.*

Let R be a complete discrete valuation ring with quotient field K and perfect residue field k of characteristic p . Under Calabi-Yau variety over K we understand smooth projective scheme \mathcal{V} over K of dimension n with trivial canonical bundle $\omega_X = \Omega_X^n$ [2, 1]. A weak Néron model of the variety X is called smooth proper scheme \mathcal{V} of finite type over R with the isomorphism $\mathcal{V} \otimes_R K \simeq X$, that satisfies next property: for every finite unramified extension $R' \supset R$ with quotient field K' , the canonical mapping $\mathcal{V}(R') \rightarrow X(K')$ is bijection [10].

3 Discrete valuation rings, modules, trees and formal schemes

At first we formulate very briefly some elementary (and probably well known) results on connections among local one dimensional complete regular rings, trees and formal schemes. We follow to [3, 4, 5, 6, 7, 8, 9, 19]. Let A be a local one dimensional complete regular ring with maximal ideal π , K its field of fractions with the multiplicative group K^* , V a two dimensional vector space over K , M a module of the rank 2 over A (a two-dimensional lattice in the space V). Denote by $S(M)$ the symmetric algebra of the module M . The main example is the case of the ring $A = \mathbb{Z}_p$ of integer p -adic numbers, $K = \mathbb{Q}_p$ the field of p -adic numbers, $\pi = p$ the prime number, M a module of the $\text{rank}_{\mathbb{Z}_p} M = 2$ over \mathbb{Z}_p .

Definition 5 *Let K be a locally compact non-Archimedean field, A its valuation ring, \mathfrak{m} the maximal ideal of A . A free module of rank n over A is called a lattice in K^n .*

It is well known

Lemma 6 *Let L be the A -submodule of the module K^n such that L is finitely generated over A and the set L generates the space K^n over K . Then L is the lattice in K^n .*

3.1 The Bruhat-Tits tree

Two modules M and M' of the rank 2 over A are called similar if $M' = xM$, $x \in K^*$. Denote by \mathcal{T} the set of classes of similar modules.

Definition 7 *Let X be the graph whose vertices are equivalence classes $[M]$ of similar modules M of the rank 2 over A in V , where two vertices x and y are joint by an edge if $x = [M]$ and $y = [M']$ with $M' \subset M$, $M' \not\subset \pi M$, $M/M' \simeq A/\pi A$.*

Two modules are called adjacent if the length $l(M/M') = 1$, i.e. $M/M' \simeq A/\pi A$.

Theorem 8 *The graph X is a homogeneous or a regular tree. We will denote the tree by \mathcal{T} .*

Corollary 9 *In the case of $A = \mathbb{Z}_p$ the degree of \mathcal{T} is equal $p + 1$ or that \mathcal{T} is a $(p + 1)$ -regular tree.*

3.2 Ends and projective spaces

Let x_1, x_2, \dots be an infinite, non-backtracking sequences of adjacent vertices of the tree \mathcal{T} , a subtree of \mathcal{T} . Let \mathcal{S} be a subtree of \mathcal{T} such that \mathcal{S} is isomorphic of the tree $\bullet - \bullet - \bullet - \bullet - \dots \infty$. The subtree \mathcal{S} is called the end of the tree \mathcal{T} .

By $\partial\mathcal{T}$ denote the set of ends of \mathcal{T} and by $\mathbb{P}^1(A)$ denote the one-dimensional projective space over A .

Theorem 10 $\partial\mathcal{T} \simeq \mathbb{P}^1(A)$.

Recall that the generic fiber of $\mathbb{P}^1(A)$ is the one-dimensional projective space \mathbb{P}^1_K over K .

3.3 Action of groups on trees

Recall that a group G acts on a set X if there is a map $G \times X \rightarrow X, (g; x) \mapsto gx$ such that the following are true: (i) For e the identity of G , $ex = x$; (ii) For $h; g \in G, x \in X, h(gx) = (hg)x$. On the space V act the projective linear group $PGL_2(K)$ and its subgroups. This action extends to the action on the tree \mathcal{T} .

Theorem 11 *Let a group G acts on a tree \mathcal{T} without fixed points and without inversions. Then G is the free group.*

Let a group $G \subset PGL_2(K)$ acts on \mathcal{T} discretely and freely. Construct follow to [7] a subtree \mathcal{T}_G of \mathcal{T} .

Theorem 12 *If the group G has finite number of generators then \mathcal{T}_G/G is finite.*

3.4 Schemes and formal schemes

For the above mentioned symmetric algebra $S(M)$ of the module M define the corresponding scheme $\mathbb{P}(M)$ by the formula

$$\mathbb{P}(M) = \text{Proj } S(M). \tag{1}$$

For each module $M \hookrightarrow V$ there is the birational isomorphism $\mathbb{P}(M) = \mathbb{P}^1(A) \otimes_A K \xrightarrow{\varphi_M} \mathbb{P}^1_K$. For two modules $M = (u, v)$ and $M' = (u', v')$, $u' = u, v' = \pi v$, construct $\mathbb{P}(M) \times_A \mathbb{P}(M')$ and the closed subset that is defined by the equation $u'v = \pi uv'$. Now let \mathcal{S} be a finite subtree of \mathcal{T} . It is possible to construct many formal schemes [3, 4, 7, 8, 9, 19] from these data. We indicate here the formal scheme \mathcal{P} that is the formal completion $(\mathbb{P}(\mathcal{S}))_0$ of the scheme $\mathbb{P}(\mathcal{S})$ along its closed fibre $\mathbb{P}(\mathcal{S})_0$ only.

4 On Frobenius structures on connections

Let p be a prime, n a positive integer, and \mathbb{F}_q the finite field with $q = p^n$ elements. Let \mathbb{Q}_q denote the unique unramified extension of degree n of the field of p -adic numbers. Let U be an open dense subscheme of the projective space $\mathbb{P}_{\mathbb{Q}_q}^1$ with nonempty complement Z . Let V be the rigid analytic subspace of $\mathbb{P}_{\mathbb{Q}_q}^1$ which is the complement of the union of the open disks of radius 1 around the points of Z . A Frobenius structure on \mathcal{E} with respect to σ is an isomorphism $\mathcal{F} : \sigma^*\mathcal{E} \simeq \mathcal{E}$ of vector bundles with connection defined on some strict neighborhood of V .

A meromorphic connection on \mathbb{P}^1 over a p -adic field admits a Frobenius structure defined over a suitable rigid analytic subspace. The effective convergence bounds for Frobenius structures on connections that improve the previous bound in question of the paper by K.S. Kedlaya [12] is given in the paper [13]. The techniques used are computational [13]. This is a good place to see the interplay between matrix representation of a Frobenius structure and a Gauss-Manin connection.

5 Newton strata in the loop group of a reductive group

Let G be a split connected reductive group over \mathbb{F}_p , let T be a split maximal torus of G and let LG be the loop group of G by Faltings [14].

Let R be a \mathbb{F}_q -algebra and K be the sub-group scheme of LG with $K(R) = G(R[[z]])$. Let σ be the Frobenius of k over \mathbb{F}_q and also of $k((z))$ over $\mathbb{F}_q((z))$. For algebraically closed k , the set of σ -conjugacy classes $[b] = \{g^{-1}b\sigma(g) | g \in G(k((z)))\}$ of elements $b \in LG(k)$ is classified by two invariants, the Kottwitz point $\kappa_G(b)$ and the Newton point ν .

The author of the paper [15] proves the following two main results.

Theorem 13 *Let S be an integral locally noetherian scheme and let $b \in LG(S)$. Let $j \in J(\nu)$ be a break point of the Newton point ν of b at the generic point of S . Let U_j be the open subscheme of S defined by the condition that a point x of S lies in U_j if and only if $\text{pr}_{(j)}(\nu_b(x)) = \text{pr}_{(j)}(\nu)$. Then U_j is an affine S -scheme.*

Theorem 14 *Let $\mu_1 \preceq \mu_2 \in X_*(T)$ be dominant coweights. Let $S_{\mu_1, \mu_2} = \bigcup_{\mu_1 \preceq \mu' \preceq \mu_2} Kz^{\mu'}K$. Let b be a σ -conjugacy class with $\kappa_G(b) = \bar{\mu}_1 = \bar{\mu}_2$ as elements of $\pi_1(G)$ and with $\nu_b \preceq \mu_2$. Then the Newton stratum $N_b = [b] \cap S_{\mu_1, \mu_2}$ is non-empty and*

pure of codimension $\langle \rho, \mu_2 - \nu_b \rangle + \frac{1}{2} \text{def}(b)$ in S_{μ_1, μ_2} . The closure of N_b is the union of all $N_{b'}$ for $[b']$ with $\kappa_G(b') = \bar{\mu}_1$ and $\nu_{b'} < \nu_b$.

Here ρ is the half-sum of the positive roots of G and the defect $\text{def}(b)$ is defined as $\text{rk}G - \text{rk}_{\mathbb{F}_q} J_b$ where J_b is the reductive group over \mathbb{F}_q with $J_b(k((z))) = \{g \in LG(\bar{k}) | gb = b\sigma(g)\}$ for every field k containing \mathbb{F}_q and with algebraically closed \bar{k} .

The proof of Theorem 13 is based on a generalization of some results by Vasiu [16]. An interesting feature of E. Viehmann method in the prove of Theorem 14 is the using of various results on the Newton stratification on loop groups as Theorem 13 and the dimension formula for affine Deligne-Lusztig varieties by Görtz, Haines, Kottwitz, Reuman [17] together with results on lengths of chains of Newton points by Chai [18].

6 Conclusion

Rigid problems, methods and results in arithmetic algebraic geometry and in dynamics have presented. Diverse notions of rigidity and respective novel results are reviewed.

References:

- [1] S.T. Yau, *A survey of Calabi-Yau manifolds* / Yau Shing-Tung // Surveys in differential geometry. Vol. XIII. Geometry, analysis, and algebraic geometry: forty years of the Journal of Differential Geometry, Scholarpedia, Surv. Differ. Geom. 4 (8). Somerville. MA: Int. Press, 2009, 277318.
- [2] A.N. Rudakov, I.R. Shafarevich, Surfaces of type $K3$ over fields of finite characteristic, *Journal of Soviet Mathematics*, 1983, 22:4, 147–1533
- [3] I. Shafarevich, *Foundations of Algebraic Geometry*, Moscow: Nauka, 1988, v.1, v.2 (in Russian).
- [4] R. Hartshorne, *Algebraic Geometry*, Springer – Verlag, Berlin–Heidelberg–New York 1977.
- [5] J.-P. Serre, *Trees*, Springer –Verlag, Berlin–Heidelberg–New York 2003.
- [6] J. Tate, Rigid analytic spaces, *Invent. Math.* 12, 1971, pp. 257–289.
- [7] D. Mumford, An analytic construction of degenerating curves over complete local rings, *Compositio Mathematica*. 24, 1972, pp. 129–174.
- [8] M. Raynaud, Géométrie analytique rigide d’après Tate, Kiehl, ..., table ronde a’analyse non-archimédienne, *Bull. Soc. math. France* 39-40, 1974, pp. 319–327.

- [9] M. Demazure, *Lectures on p - divisible groups*, LNM 302, Springer Verlag, Berlin 1972.
- [10] S. Bosch , W. L'utkebohmert , M. Raynaud, *Neron Models*, Ergebnisse der Mathematik 21, Springer-Verlag, 1990.
- [11] N.M. Glazunov , On norm maps and "universal norms" of formal groups over integer rings of local fields, *Continuous and Distributed Systems. Theory and Applications*, Springer. 2014. - P. 73 80.
- [12] K. Kedlaya, Effective p -adic cohomology for cyclic threefolds. In: *Computational Algebraic and Analytic Geometry of Low dimensional Varieties*. Amer. Math. Soc. Vol. 572
- [13] K. Kedlaya, J. Tuitman, Effective convergence bounds for Frobenius structures on connections, *Rend. Semin. Mat. Univ. Padova* 128, 2012, pp. 7–16.
- [14] G. Faltings , *Journ. Eur. Math. Soc. (JEMS)* 5, 2003, 41 - 68 .
- [15] E. Viehmann, Newton strata in the loop group of a reductive group, *Am. J. Math.*, 135, No. 2, 2013, 499-518.
- [16] Vasiu, *Ann. Sci. `Ecole Norm. Sup.* (4) 39, no. 2, 2006, 245 - 300.
- [17] G'ortz, Haines, Kottwitz, Reuman, *Ann. Sci. Ecole Norm. Sup.* (4) 39, 2006, 467 - 511.
- [18] Chai, *Journ. Amer. Math. Soc.*, 13, 2003, 209-241.
- [19] N.M. Glazunov, *Development of methods to justification of conjectures of formal theories (in Russian)*, LAP, Germany 2014.
- [20] N. Glazunov, Crystalline cohomology and their applications, *Algebra and Number Theory: Modern Problems and Application: XII International Conference, Tula, RFFI*. 2014, pp. 52-54.

Math modeling of underground water infiltration in exhausted gas deposit

Irina N. Polshkova

Abstract— The studying goal is the state of groundwater flow, what is changing during the development of gas field, as well as basing of possibility for reserves assessment of related mineral water. As the main settlement in the hydrogeological scheme showing the dynamic of ground water flow approximation adopted single-layer aquifer, located under the Cenomanian horizon roof, taken as impenetrable border (top). The lower boundary of the aquifer thickness is in the state of prop from regional underground water system. During exhausting gas deposit the gas pressure in the gas cap falls. Pore space, the volume of which depends on the capacity of the Cenomanian deposits horizon becomes available for infiltration of underground water. There are two models, both processing simultaneously showing changes of hydrodynamic flow state. Two processes are represented with math models – gas-water contact raise and the process of pressure decrease in watered thickness as underground water fills the pore space that is being freed. The modeling results are: the surface map of gas-water contact, map of Cenomanian groundwater pressures head and graph of changes in time - the value of influx from below and the appropriate value of outflow into capacity.

Keywords— Math model, underground water pressure, groundwater flow dynamics, gas deposit

I. INTRODUCTION

The contemporary level of hydrogeology development and computing machinery application makes it possible to solve a number of scientific and applied tasks whose setting enables to adequately reflect real natural and anthropogenic processes. Math modeling opportunities make it possible to move from the simplified calculation schemes, used in analytical calculations, to special filtration math models in multi-layer system of real hydrogeological object and that is principally new level of natural and anthropogenic processes research.

They turned out to be widespread to maximum extent when it had become possible to generate any combination of input data and automation of processing multi-variant model calculations.

Math modeling provides the most complete implementation of three basic objectives of science: description, explanation and forecasting [1]. Actually, math model of hydrogeological object gives schematic description of processes going on within it, explains mechanisms of these processes with interaction of clearly highlighted factors and finally provides

for opportunity of quantitative forecasting and development of the processes being considered in space and time.

All these opportunities have been materialised in author system of math and software «Aquasoft»[2].

Research of underground water infiltration dynamics in exhausted gas deposit has been carried out on the basis of materials and data presented in reports [3], [4].

This math model has been created with the aim to reproduce the process of pressure decrease in watered/saturated thickness as gas-water contact rises in the course of gas deposit mining works.

Classic variant of math modeling method assumes multi-variant solutions of the task when sought pressure fields are being reproduced at model in the process of studying of parameters sensibility of aquiferous and dividing layers as well as all external disturbing factors.

Adequacy degree of the model is assessed by matching mode observation data and absolute marks of underground water levels obtained as results of modeling for different time steps. Besides, it is desirable to have work mode data on flow volumes that characterize water exchange between underground and surface waters.

In the given case the process of lowering pressure is not the result of water extraction from aquiferous layers, that is set in the model in the form of border conditions of second kind with known negative debits. That is why ground water infiltration dynamics reproduction has been implemented in two stages. Gas-water contact movement is reproduced at the model at the first stage, and the following factors have been monitored as model adequacy criteria:

- gas-water contact rise velocity;
- volume of infiltrated water – total and annual.

Information about absolute gas-water contact surface marks is thought to be the most trustworthy but the second parameter is a result of some calculations and therefore need consider only size volume of infiltrated water, not its exact value.

At the second stage the process of pressure lowering has been reproduced as a result of filling of aquifer capacity with underground water in accordance with gradual raising of gas-water contact.

The main result of first stage modeling due to gas extraction is the volume of freed aquifer capacity in all the settlement points in each step of a time. These values are typical of negative debits, and they are set at as time-dependent boundary conditions of the second kind at the basic model.

Irina N. Polshkova is with Water Problem Institute of Russian Academy of Sciences, Moscow, Russia (e-mail: z_irpol1@mail.ru).

Model adequacy has been controlled with factual data obtained from wells temporal observations, for wells located below initial gas-water contact.

II. STUDY METHODS

A. Preliminary Model of Water-Gas Contact Rise

As hydrogeological scheme of underground water hydrodynamic flow for calculation purposes we used the scheme of one-layer aquiferous stratum located under the roof of Cenomanian layer considered as impenetrable frontier (from above). Lower boundary of the watered thick, being an element of regional water-pressure system, is in the state of prop. As gas deposit is being exhausted, pressure in gas cap, lying above the watered thick, falls and in the thinned out space with rather high porosity parameters (less porous space occupied with non-wasted restrained gas) occurs watered thick surface rise (Fig.1).

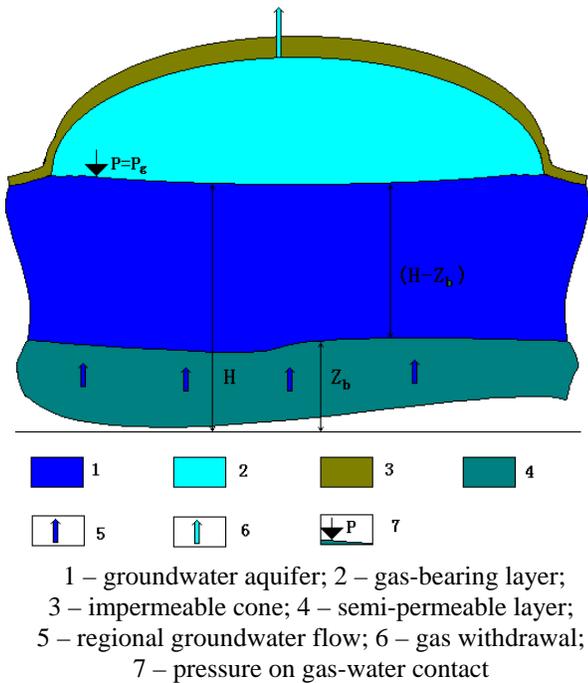


Fig.1 hydrogeological schematization.

Watered thick level rise with free space can be calculated by confined-unconfined filtration equation with taking into account effective thickness of aquiferous layer, calculated by ratio of absolute magnitudes of Cenomanian layer roof marks, gas-water contact level and lower boundary of watered thick Z_b. This equation representing sum of weighted debits of interconnected flows of underground waters applied to this one-layer system, has the following structure:

$$\frac{\partial}{\partial x}(\hat{T}_x \frac{\partial H}{\partial x}) + \frac{\partial}{\partial y}(\hat{T}_y \frac{\partial H}{\partial y}) + Q_2 + (Z_b - H) \frac{k_z}{M_z} = \hat{\mu} \frac{\partial H}{\partial t} \quad (1)$$

where x,y – coordinates of a point within area;

H – sought piezometric pressure of underground water in aquiferous layer;

$\hat{T}_x(\hat{T}_y)$ - effective magnitude of water conductivity coefficient of the layer in the direction of axis OX (OY):

$$\hat{T}_x(\hat{T}_y) = \begin{cases} k_{x(y)}(Z_t - Z_b) & \text{if } H > Z_b; \\ k_{x(y)}(H - Z_b) & \text{if } Z_t \leq H \leq Z_b; \\ 0 & \text{if } H < Z_b, \end{cases}$$

where $k_{x(y)}$ - filtration coefficient of the layer in the direction of axis X (Y);

Z_t - absolute mark of aquiferous layer top;

Z_b - absolute mark of aquiferous layer bottom,

$\hat{\mu}$ - effective value of layer capacity (elastic or gravitational) assuming that gravitational water debit coefficient, in the situation of lowering underground water level, equals to the saturation lack coefficient in the situation of rising level:

$$\hat{\mu}_n = \begin{cases} \mu_{\text{elast}} & \text{if } H > Z_t; \\ \mu_{\text{grav}} & \text{if } Z_t \leq H \leq Z_b; \\ 0 & \text{if } H \leq Z_b. \end{cases}$$

Q₂ - boundary condition of the 2-nd type – known value of weighted debit of the source (drainage),

k_z - coefficient of vertical interaction between regional (local) pressure system and watered thick with free surface;

M_z – thickness of hypothetic dividing layer between regional (local) pressure system and calculated watered thick;

Equation in finite differentials is solved by the method of minimising balance flows discrepancies [6].

Initial condition for modeling nonstationary filtration process in the given system is the absolute mark of initial gas-water contact level.

As it has been shown balance components of horizontal flows are quite insignificant and a nonstationary process of rise gas-water contact is described with the following equation

$$(Z_b - H) \frac{k_z}{M_z} = \hat{\mu} \frac{\partial H(x, y)}{\partial t} \quad (2),$$

where

$$Q = (Z_b - H) \frac{k_z}{M_z} S - \text{annual volume of invaded water,}$$

S – area via which influx occurs is restricted with the Cenomanian layer roof perimeter,

Z_b – absolute mark of watered thickness contact with regional pressure system of underground waters, H – absolute mark of gas-water contact level.

Parameters of hydrogeological medium for the given hydrodynamic system, according to the above equation, are the following values:

- capacity of aquiferous part of the layer -effective porosity coefficient $\hat{\mu}$ from work [4];

- absolute mark of contact surface between watered thickness and regional pressure system, i.e. surface at the level of regional prop is being effected - Z_b ;

- parameter of vertical conductivity sets the interconnection between watered thick and regional underground water flow located deeper than watered thick

$$G = \frac{k_z}{M_z} S ;$$

- parameter of lateral water conductivity in aquiferous part of Cenomanian layer.

Modelled functions are gas-water contact rise value and volume of invaded water counting from the beginning of extraction works on the deposit i.e. from -1972.

Territory of southern part of «Medvezhie» deposit is the area of 47×34 km, to which there is a corresponding math grid model with the size of 58×43 net points. Average value of initial gas-water contact level equalling 1131.3 m (that was determined across 65 wells) has been taken as initial conditions.

The model has been used for reproducing of nonstationary geofiltration process relatively to rise of gas-water contact level with time discretization step in one year.

Dynamics of gas-water contact rise depending on all of the hydrogeology parameters has been assessed on the model.

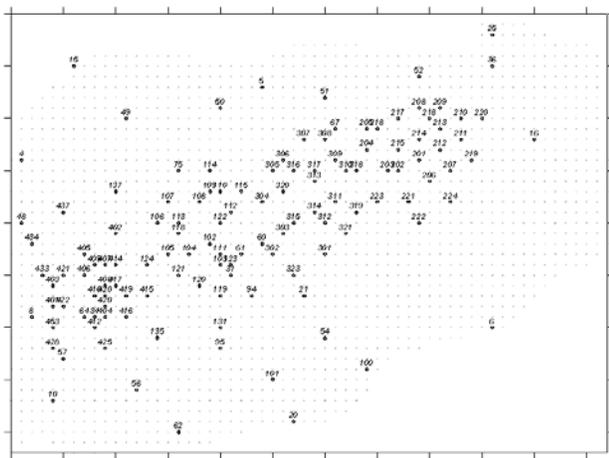


Fig. 2 placement of monitoring wells on grid model

As it turned out in the process of modeling, capacity parameter exerts maximum influence on the velocity of gas-water contact rise, and volume of invaded water is depends on the depth Z_b of the regional contact .

Coefficient of conductivity between watered thickness of Cenomanian layer and regional underground water flow influences both functions and it is given as generalized value in direct proportion to vertical filtration coefficient and layer interaction area (along the cupola perimeter) and inversely proportional to the strength of hypothetic dividing thick. Results of task solution for the chosen parameters range are given in Table 1 [7].

It can be seen from the table that modeling process is manageable and changes of functions are quite logical. As it has been supposed, range of lateral water conductivity

parameters change (from 100 to 500 m²/day) does not significantly influence the modelled process, and that fact confirms the assumption the main process is the process of outflow into freed capacity.

Table 1 Correlation among functions and parameters basing on results of math modeling

№ var	Lateral water conduct. TX=TY m ² /day	Capac. μ	Vertical conduct. G=(K/M)S m ² /day	Region contact depth Z _b m	Gas-water contact rise m	Invaded water volume (th. m ³ /year)
1	300.	0.2	10	-1190.	19.4	4261.6
2	300	0.2	5	-1190.	10.7	2263.1
3	300	0.1	5	-1190.	19.3	2130.9
4	300	0.1	1.5	-1190.	6.7	695.8
5	300	0.05	1.5	-1190.	12.6	670.7
6	300	0.02	1.5	-1190.	26.4	602.9
7	300	0.02	1.5	-1180.	22	500.2
8	100	0.02	1.5	-1180.	21.9	500.2
9	500	0.02	1.5	-1180.	21.8	500.3

However, it is worth noticing that the model has to significant degree qualitative character since the value of aquifer capacity was set one and the same for all points of modelled area. Column “gas-water contact rise” shows the average value of lifting gas-water contact, obtained on a preliminary model that is illustrated at fig.3

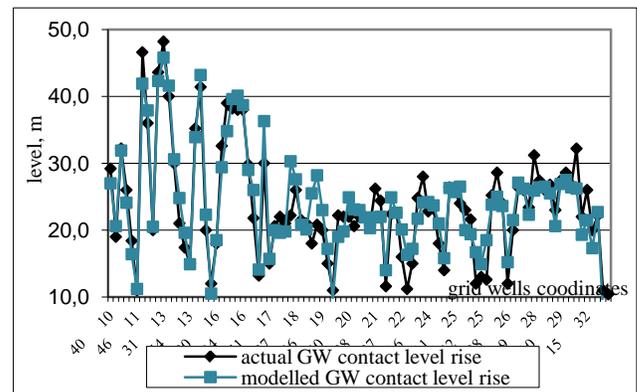


Fig.3 comparison of values between actual and modelled gas-water contact rise (1993)

The factual average value of gas-water contact rise is 23.7 m [4], and the value of 22 meters has been obtained on the model. This result can be explained by low value of gravitational capacity – 0.02, set on the model. Invaded water volume, according to the same data [4],[5] at the end of 1983 and 1984 are 441 and 569 thousand.m³/year, correspondingly.

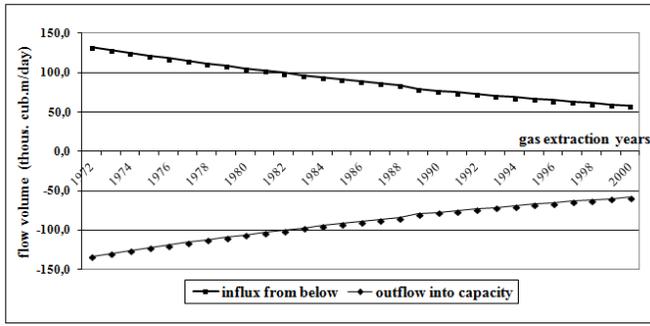


Fig. 4 balance of main flows during gas-water contact rise

Minimum value of about 500 thousands m³/year has been obtained on the model. Moreover, it is noticeable on the model that as pressure in gas cap falls year by year, invaded water volume decreases as well.

The results of task solving confirm assumption that dynamic groundwater flow is described by equation (2) and as can see from Fig. 4, balance component of lateral flow are quite insignificant, because illustrated flows are nearly equal modulo. As far as filling capacity possible volume to invading decreases groundwater influx decreases respectively, however the value of model parameters remain constant in time.

As result of modeling it can be assumed that initial thickness of aquifer, equalling to difference between absolute marks of regional contact depth and initial gas-water contact level, is (1180-1131.3) – about 50 meters, and that is determined with enough degree of precision by results of task solution.

B. Basic Model of Lowering Head Pressure

The described model, reproducing gas-water contact rise, is preliminary.

Basic model reproduces the process of lowering head pressure in watered part of Cenomanian layer which is schematized as one-layer system interacting on the bottom of the model with regional flow of underground waters in the form of stationary border conditions of the third kind - $Q_3 = (H_{bound} - H)G_{bound}$.

Pressure lowering process occurs due to filling up porous space that is being freed after the gas extraction in accordance with equation (1).

The volumes of capacity filled with water are the main modeling results at the preliminary stage and they were written into model database when solving task.

Field of this parameter has a meaning of hypothetic water extraction which realises the process of lowering pressure in watered part of Cenomanian aquifer in classical task setting and is given on the main model as nonstationary border condition of second kind at every step k in time:

$$Q_i^k(\mu_i) = \hat{\mu} \frac{\partial H_i^k}{\partial t^k} \equiv Q_2^k.$$

Nonstationary geofiltration process for one-layer task in this case is described with the following equation:

$$\frac{\partial}{\partial x} (\hat{T}_x \frac{\partial H}{\partial x}) + \frac{\partial}{\partial y} (\hat{T}_y \frac{\partial H}{\partial y}) + Q_2 + (H_{bound} - H)G_{bound} = \hat{\mu} \frac{\partial H}{\partial t}$$

where \hat{T}_x, \hat{T}_y - parameters of lateral water conductivity analogous to definition in equation (1), are set, as result, from preliminary model,

$\mu = 0,02$ - gravitational capacity of aquifer,

Q_2 - nonstationary border conditions of second type,

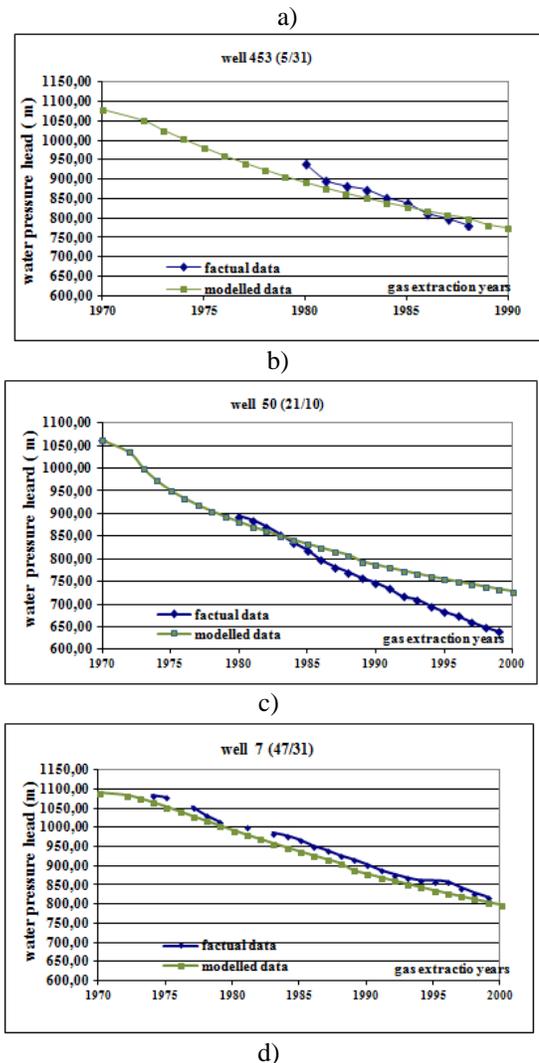
$Q_3 = (H_{bound} - H)G_{bound}$ - stationary border conditions of third type:

H_{bound} - absolute mark of underground water level along the Cenomanian layer cupola (according to preliminary model results),

G_{bound} - parameter of vertical conductivity, sets interconnection with regional hydrodynamic flow (according to preliminary model results).

Pressure lowering process has been reproduced from 1972 to 2000. Model adequacy to real process is traced for 9 hydrogeological observation wells existing in watered thick.

On fig. 5 you can see the results of modeling water-pressure head lowering compared with observation data in time.



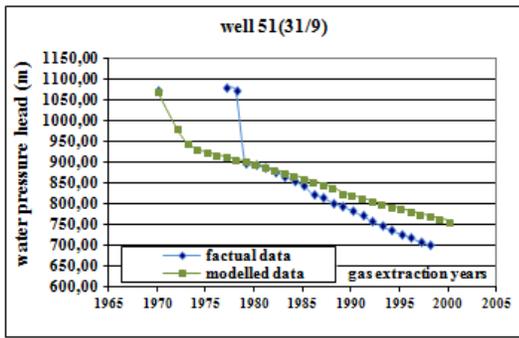


Fig. 5 comparison of modeling results and factual water-pressure head lowering data in watered thick during gas-water contact rise

You can see on Fig. 6 model solutions for 30th year of gas deposit exploitation.

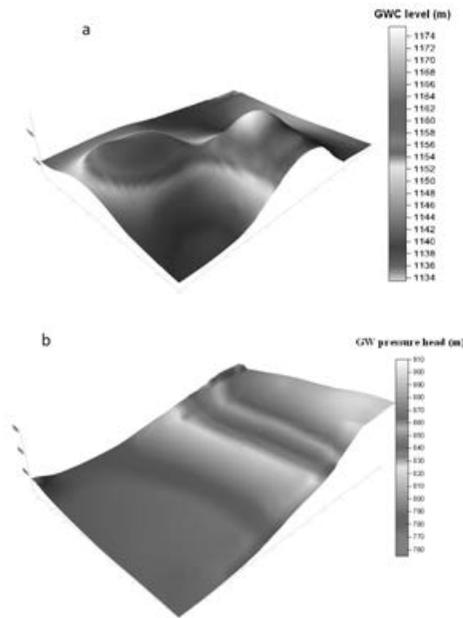


Fig. 6 modeling results: (a) – gas-water contact surface – preliminary model, (b) – underground water pressures field of cenomanian – main model

Two cupolas of gas-water contact correspond to two parts of gas deposit.

III. CONCLUSION

Thus, the following method of solving similar tasks can be determined:

1. At first, the model reproduces the nonstationary process of lifting the GWC and changed volume of invaded water. Rise of GWC is modeled by boundary conditions of the third kind, because of the prop at the lower bound of aquifer from regional flow through a hypothetical separating layer.

The criterion of adequacy-comparative graphs of gas-water contact levels rise and dynamic amounts of structural balance, as comparison with an independent calculation of invading water. Result – fields of capacity flows at every step in time.

2. The corresponding capacity flows are transferred to the main model from grid data database of auxiliary model at every calculation time step. To check model adequacy it is necessary to have a few hydrogeological wells in watered thick located below the initial level of gas-water contact. Major attention is given to specifying water conductivity parameter on the basis of level lowering data in monitoring wells. Adequacy criterion – comparative graphs across the wells and balance tables.

3. Underground water (usually mineralized to some degree) reserves can be assessed on the main model.

ACKNOWLEDGMENT

The first author is thankful to her parents.

REFERENCES

- [1] Work book on forecasting, M., "Mysl", 1982, 430 p.
- [2] Polshkova I.N. System of special software aimed at calculating filtration and mass transfer processes in underground waters -«Aquasoft» - certificate of registration № 2006610658 of 17.02.2006.
- [3] Skorobogatov V.A. and others. Report on the subject P.1.1.II.11 «Hydrogeological aspects of watering the largest mined gas deposits of the northern part of Tyumen region», VNIIGAZ, 1998.
- [4] Remizov and others. Methods of analysing mining cenomanian deposits of the northern part of Tyumen region according to working wells investigation (during the period of active manifestation of water pressure mode) M., IRZ Gazprom. Information review, 1999.
- [5] Polshkova I.N. Specifics of implementing the system of special math provision for automated grid models of basins and deposits of underground waters. – Moscow, VSEGINGEO, 1994, dissertation abstract ... PhD.
- [6] Polshkova I.N., Silaev V.S., Gilfanova T.N. Math model of geofiltration process in watered thick during gas deposit exploitation. – Collection of scientific works «Issues of developing underground industrial waters», Makhachkala 2003.

I. Polshkova Leningrad State University, Physics Faculty, Doctor of Science Engineering (Hydrogeology), Leading Researcher with Water Problem Institute of Russian Academy of Sciences

Main research field:

studying of underground water dynamics processes, software implement for underground water hydrodynamic math models, qualitative and quantitative underground water assessment and forecasting, transboundary models, math and software “Aquasoft” .

Mathematical modelling of groundwater flow coupled with internal flow in drainage pipe situated in a bounded shallow aquifer

I. David, C. Grădinaru, C. Gabor, I. Vlad, C. Stefanescu

Abstract — *The main objective of the paper is to present a relatively simple modelling method for estimate the influence of hydraulic head losses of the internal non linear flow along a horizontal drainage tube (or along of an vertical partially penetrating well) on the groundwater inflow distribution along the drainage tube. As coupling condition the flux continuity on the drainage tube wall of both external linear groundwater flow and internal non linear flow in the drainage tube will be considered.*

The horizontal and vertical drainage tube are modelled as 3 D line sinks elements distributed along their axis while 2 D BEM has been used for modelling the groundwater flow with given mixed boundary conditions on the external boundary which bounded the groundwater flow domain.

The elaborated mathematical modelling was tested through numerical calculus of the inflow distribution along a drainage tube of finite length with and without inner hydraulic head loss along the drainage tube..

Keywords — groundwater flow, coupled internal/external flow by drainage pipe, line sinks strength, BEM

I. INTRODUCTORY REMARKS

The line sink analytical element method (AEM) for modelling ground-water flow in shallow regional aquifer which incorporate drainage flow features as well have been studied in numerous authors [1], [2], [3], [4], [5]. Several results are obtained [6], modelling radial collector well with laterals using semi-analytical methods (e.g. conformal mapping and 2D line strength distributions) and in [7], [8], [9]) for modeling groundwater flow generated by ground water recovery and recharge systems in a bounded flow domain using coupled AEM and BEM.

I. David, „POLITECHNICA“ University of Timisoara, Department of Hydrotechnical Engineering, George Enescu 1/A, 300022 Timisoara, Romania and University of Applied Sciences Giessen, Germany (corresponding author e-mail: ioan.david@gmx.net; ioan.david@upt.ro)

C. Grădinaru, „POLITECHNICA“ University and S.C. GAUSS S.R.L. Timisoara (e-mail: cristian.gradinaru@gauss.ro)

I. Vlad, „POLITECHNICA“ University and S.C. GAUSS S.R.L. Timisoara (e-mail: ioan.vlad@gauss.ro)

C. Gabor, „POLITECHNICA“ University and S.C. GAUSS S.R.L. Timisoara (e-mail: Cristian.gabor@gauss.ro)

C. Stefanescu, „POLITECHNICA“ University of Timisoara, „POLITECHNICA“ University of Timisoara, (e-mail: achim_camelia@yahoo.co.uk)

The partially penetrating well (pW) is one of the most important technical application which was discussed also in the above mentioned papers. The solutions obtained for pW or horizontal drainage tube or radial collectors of well with laterals can be achieved by assuming that the flow is generated by a distribution of sources or line sinks of unknown strength along the well axis, whose values can be determined from the boundary conditions such as the given head on the well screen. The modeling procedure for pW, which can be traced back to [1] was analysed and improved in [2] who developed a numerical procedure to determine the total discharge of the pW. The pW is divided into N intervals (i.e. line sink strengths) distributed along the well axis having a specific discharge as a constant strength for each interval i.e. $\psi_i = q_i$ (i=1,2, ...N). So the total discharge Q of the pW can be calculate as a simple sum $Q = \sum \psi_i \Delta L_i = \sum q_i \Delta L_i$. Both authors remarked that this method is only accurate for wells with a much larger length of the well-screen than the well-radius. It is to be mentioned that in the case of large diameter, when the well is represented by a cylindrical equipotential surface, the appropriate sink distribution along the well axis become oscillatory [2]. In [4] and [5] a numerical procedure for modeling pW and horizontal canal/drain using third order strength line sinks, a combination of a linearly varying sink density with a square root density. This line sink distribution of high order strength, which depend on several unknown parameters allows a relatively good approach for calculating the well screen as a cylindrical equipotential surface and the specific discharge distribution as well, especially for a well with smaller radius. A new method using Analytical and Boundary Elements (AEM/BEM) based on 3-D Integral Representation for Numerical Solution of Potential Problems in Heterogeneous Media Containing Singularities was developed in [10].

In connection with mathematical numerical modeling of groundwater flow problems it is to mentioned that known high performance FDM/FVM based softwares (e.g. MODFLOW or EEFLOW) [11], [12] can not include correctly the singular flow conditions close to wells or drainage pipes (logarithmical singularities in the case of 2-D flow or polar singularities in the case

of 3-D flow) and so these singular flow conditions are with insufficient accuracy represented. We also should mention that none of these modelling methods and papers take into account the inner hydraulic head loss along the pW or drainage tube, and consequently the real flow conditions resulting through the coupled interior/exterior flows along the pW or along the horizontal drain pipes. In the same context it is to underline also that in the above mentioned publications the head losses through the drainage pipe wall or pW wall is not take into account .

For a correct calculation of drainage systems (Ldr) or partially penetrated wells (pW), it is necessary to take into account the internal hydraulic head loss. That because the inner head losses influence the piezometric head distribution along the well or drain pipe changing the piezometric head distribution and consequently the groundwater inflow distribution will be influenced. So coupled flow conditions appear on the wall of the drainage pipe or pW.

The main objective of the paper is to present a relatively simple modelling method for estimate the influence of hydraulic head losses of the internal non linear flow along a horizontal drainage tube (or along of an vertical partially penetrating well) on the groundwater inflow distribution along the drainage tube.

II. MATHEMATICAL REPRESENTATION OF the GROUNDWATER FLOW SYSTEM

The Scheme of the considered flow system i.e. a bounded groundwater domain comprising local 3D flow features such a partially penetrating Well (pW), horizontal drainage pipe of finite length (L_{dr}) well with laterals (W_1) is shown in Figure 2. The closed boundary $C_0 = C_{0H} \cup C_{0q} \cup C_{0\Sigma}$ with its given boundary conditions: given head (i.e. Dirichlet boundary conditions) on C_{0H} , C_{0q} as inflow/outflow boundary (i.e. Neumann boundary conditions) and $C_{0\Sigma}$ as impervious boundary are modelled mathematically means of integral representations specifically for use further the indirect boundary element method (IBEM). The drainage pipe of finite length L denoted further (L_{dr}) and the partially penetrating well denoted (pW) are modelled as 3D line strength distribution along the axis of the drainage pipes. A recharge area from precipitations D_σ is also considered.

The potential function $\phi(M)$ generated from 2D and local 3-D flow-system (i.e. Dp and pW) situated in a large extended plane domain D_0^+ inside of its external boundary C_0 (Figure 2) can be derived by superposition of all partial potentials generated from line strengths and boundary elements which represent the contribution of the boundary conditions, recharge area and of the horizontal and vertical components of the considered flow system.

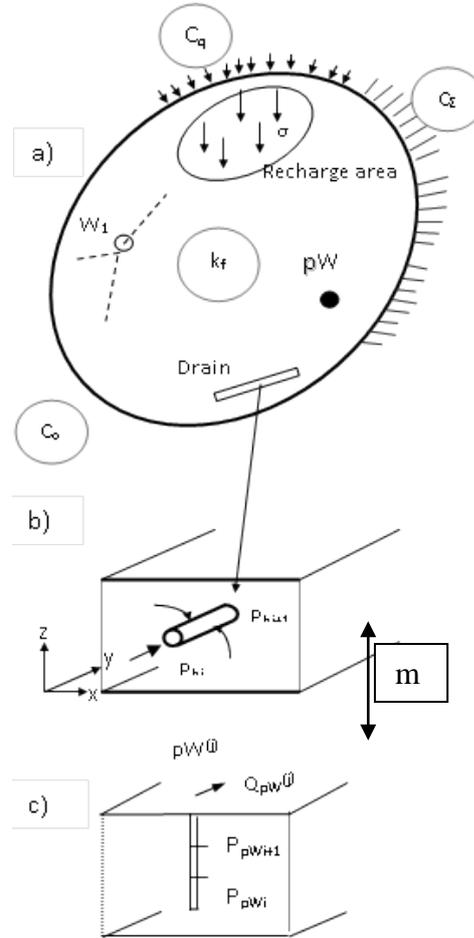


Fig. 1. Scheme of the bounded groundwater flow domain
 a) Plan view
 b) horizontal drainage pipe
 c) partially penetrating Well

The potential function can be expressed in the following integral form:

$$\phi(M) = -\frac{1}{2\pi} \left[\int_{C_0} \psi(P) G(M,P) dl + \int_{D_\sigma} \sigma(P_\sigma) G(M,P_\sigma) d\Omega + \int_{L_{dr}} \psi_{L_{dr}}(P_{L_{dr}}) G(M,P_{L_{dr}}) dl + \sum_{k=1}^{N_{pW}} \psi_{kpW} G_k(M,P_w) \right] + c \quad (1)$$

$$M \in D_0^{**} \cup D_\sigma \quad ; \quad D_0^{**} = D_0^+ - \{pW \cup DL_{dr}\}, P \in C_0$$

The specific discharge in the normal direction to boundary or pipe axis "n" i.e. $q_n(M)$ can be determine from (1) using the Darcy's law:

$$q_n(M) = -\frac{\partial \phi(M)}{\partial n_M} = -\frac{1}{2\pi} \left[\int_{C_0} \psi(P) F(M,P) dl + \int_{D_\sigma} \sigma(P_\sigma) F(M,P_\sigma) d\Omega + \int_{L_{dr}} \psi_{L_{dr}}(P_{L_{dr}}) F(M,P_{L_{dr}}) dl + \sum_{k=1}^{N_{pW}} \psi_{kpW} F_k(M,P_w) \right] \quad (2)$$

$$M \in D_0^+ \cup D_\sigma, P \in C_0$$

The first term in both equations represent the effects of the boundary C_o with its given boundary conditions: given head on C_{oH} , given inflow/outflow on C_{oq} and $C_{o\Sigma}$ as impervious boundary. Mathematically this term is modelled by means of 2D indirect integral representation in which $\psi(P)$ is the unknown density distributed along the boundary C_o , $G(M,P)$ is the known logarithmic potential and $F(M,P)$ its derivation in direction "n" located in $M(x,y,z)$. For the numerical implementation, these terms are expressed with constant boundary elements i.e. unknown constant strength $\psi(P)$ for each boundary element.

The second term represent the effect of the recharge from the precipitation on the area D_σ , with a recharge rate distribution of $\sigma(P)$ (Fig.1).

The third terms represent the effect of the horizontal drainage pipe L_{dr} , with the strength density ψ_{Ldr} distributed along its axes.

The last terms in (1) and (2) represent the potential and specific discharges respectively, generated by partially penetrating well (pW) (Fig.2). For the mathematical modelling the well length L_w is divided in k well elements (Fig.2 c) with length L_{WEk} . For each element an unknown line sink constant strength ψ_k on the well axis will be considered. The functions F_k and G_k incorporate the effects which are result through the multi images method of the pW line sink elements in relation to the impervious top and bottom of the aquifer for the line strength. The multi images procedure ensures the achievement of the boundary conditions on the impervious top and bottom of the groundwater layer.

The second term which represents the effect of the horizontal drainage pipe of finite length L_{dr} was implemented using a simplified procedure: 2D horizontal line sinks with constant strengths $\psi_{Ldr}(P)$ on each element of the drainage pipe L_{dr} and an additional head loss to model the 3D effects in the vertical plane in the neighbourhood of the drainage pipe (Figure 2 b). The additional head loss $\Delta h_{v(Ldr)}$ in a point P of the L_{dr} can be calculated [6] as

$$\Delta h_{v(Ldr)}(P) = \frac{q_{Ldr}(P) m}{T \pi} \ln \left[\frac{m}{\pi d_0} \frac{1}{\sin\left(\frac{\pi a}{m}\right)} \right] \quad (3)$$

where $T=k_f m$ is the transmissivity of the groundwater layer having a thickness m , k_f is the filtration coefficient and $q_{Ldr}(P)$ is the specific inflow rate into drainage pipe at the point P , d_0 the pipe diameter and "a" the distance from the drain axis to the bottom of the aquifer.

The last terms of (1) and (2) represent the potential and specific discharges respectively, generated by partially penetrating recharge wells (pW). For modelling the length L_w of the pW is divided into elements (pWE_k) with length L_{WEk} and the line sink distributions on the well axis has an unknown constant strength ψ_k for each well element (pWE_k),

(Figure 2). The functions F_k and G_k incorporate the image terms needed in order to satisfy the boundary conditions on the impervious top and bottom of the aquifer.

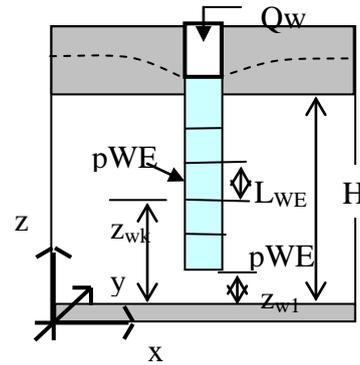


Fig. 2. Scheme of the partially penetrating recharge well (pW) discretized with well elements (pWE_k)

III. IMPLEMENTATION AND PROVE THE PROPOSED METHOD

For the implementation of the obtained mathematical representations it is to observe that unknown strength distributions along the boundary, along the drainage pipes and along the (pW) elements as well as integration constants c can be determined from the equation system which is obtained from the integral representations (1) and (2), taking into account the above described numerical approach and the given boundary conditions. For the L_{dr} and pW the boundary conditions are given heads on the drainage pipe and pWell wall (envelop/screen) respectively. The given heads $H_{0(Ldr)}(P)$ along the L_{dr} will be modified for including the 3D effects in the neighbourhood of pipe and the inner head loss along the pipe as follows

$$H_{0(c)}^*(P) = H_{0(c)}(P) \pm \Delta h_{v(c)}(P) \pm h_{v(c)}(P) \quad (4)$$

where $\Delta h_{v(c)}(P)$ is calculated with (3). The inner head loss in the drain pipe $h_{v(c)}(P)$ can be calculated for a line segment $[0-l_p]$ with the known hydraulic head loss formula

$$h_{v(c)}(l) = \frac{8}{\pi g d_0^5} \int_0^{l_p} \left[\lambda - 2 d_0 \frac{q(x)}{Q(x)} \right] Q^2(x) dx \quad (6)$$

where λ is the Darcy pipe frictional coefficient (i.e. Colebrook-formula), $q(x)$ the specific discharge and $Q(x)$ the total discharge at x , ($0 \leq x \leq l_p$), along the pipe.

For the purpose of these calculations a Computer Programme has been developed. It was confirmed that the line sink strength distribution along the well and drain axis can be replaced with the specific discharge, e.g. for pW-elements

$$q_{pWE_{k,k+1}} \cong \psi_{pWE_{k,k+1}} \quad (7)$$

So the total discharge is found by adding up the strengths:

$$Q_{pW} = \sum_{i=1}^{N_{WE}} q_{pWEk,k+1} L_{WEk} = \sum_{i=1}^{N_{WE}} \psi_{pWEk,k+1} L_{WEk} \quad (8)$$

In Figure 3 is depicted the specific discharge distribution along a drainage pipe of a radial collector well with three horizontal laterals (radial drainage pipes) placed in the centre of circular groundwater flow domain.

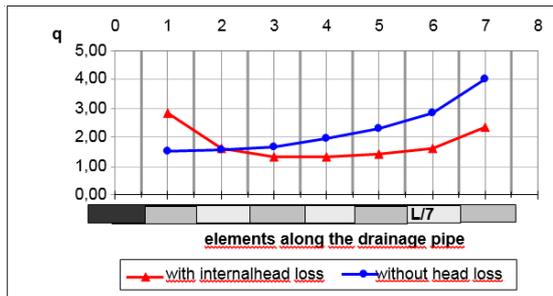


Fig. 3 Inflow rate distribution along the drainage pipe

The parameters for this numerical example (Figure 3) are: aquifer depth $m=10m$; distance from the vertical shaft centre to the end of the collector $L_{CWL}=50m$; length of an collector drainage pipe $L_{dr}=6,25m$ (7 inflow elements); diameter of the circular flow domain $400m$; diameter of the collector drainage pipe $d_0=0,25m$; drawdown in the shaft $S_0=10m$; hydraulic conductivity of the aquifer $k_f=0,003m/s$ and the roughness of the pipe wall of $3mm$. We observe that the inner head loss can have a very important influence on the specific discharge distribution and on the total discharge as well ($Q_{with\ hl} = 0,78 Q_{without\ hl}$). On the basis of numerous examples, we come to the conclusion, that for drainage pipe of collector wells the proposed modelling method based on the simplified assumption (i.e. 2D line sinks and complementary head loss in the vertical plan) can be accepted in all practical cases. The results are very accuracy for the discharge distribution and for the total discharge as well.

IV. CONCLUSIONS

The obtained results confirmed that for a correct calculation of drainage pipes or partially penetrated wells (pW), it is necessary to take into account the internal hydraulic head significant changes loss along the pipe.

In the paper has been shown that the hydraulic inner head losses in the drainage pipe can leads to the significant changes of the the groundwater inflow rate distribution along the drainage pipe.

So for correct calculus the coupled flow conditions must be taken into account on the wall of the drainage pipe or pW.

The paper present and proved a mathematical modelling method for estimate the influence of the

hydraulic head losses of the internal non-linear flow along a drainage pipe on the inflow rate distribution along the drainage pipe.

We consider that it is appropriate to propose that the coupling of the nonlinear flow within the drainage tube with the external flow in the aquifer is required to be implemented also in performance comercial softwares such as MODFLOW numerical performance or FEEFLOW.

References

[1] M. Muskat. (1937): The Flow of Homogeneous Fluids through Porous Media. cGraw-Hill, Ann Arbor, Michigan
 [2] G. Dagan, G. (1978): A Note on Packer, Slug, and Recovery Tests in Unconfined Aquifers. Water Resources Research, Vol. 14, No. 5
 [3] O.D.L. Strack, (1989): Groundwater Mechanics. Prentice Hall, Englewood Cliffs, New Jersey
 [4] H.M. Haitjema (1995): Analytic Element Modeling of Groundwater Flow. Academic Press, New York London Toronto
 [5]H. M. Haitjema, V. Kelso, H. Luther (2000): Analytic Element Modeling of Groundwater Flow and High Performance Computing. U.S. EPA/600/S-00/001, May 2000
 [6] I. David (1977): Grundwasserfassungsanlagen mit Filterrohren. Technischer Bericht Nr. 19, TH Darmstadt, Institut für Hydraulik und Hydrologie
 [7] I. David, H. Gerdes (1995): Incorporation of local three-dimensional flow in plane BEM. Boundary Element XVII. Computational Mechanics Publications. Southampton Boston
 [8] I. David, H. Gerdes (1998): Coupling Analytical Element, BEM and FEM to develop a model for groundwater flow, Computational Mechanics Publication, Computational Methods in Water Resources, Vol.1, 362-370
 [9] I. David, H. Gerdes, (2002): An environmentally friendly method for artificial groundwater recharge in wooded areas. Proceedings of the International Conference "Preventing and Fighting Hydrological Disasters", "Politehnica" University Timisoara, Romania
 [10] I. David, Analytical and Boundary Elements based Integral Representation for Numerical Solution of 3-D Potential Problems in Heterogeneous Media Containing Singularities. Proceedings of the 12th WSEAS International Conference (MACMESE '10), University of Algarve, Faro, Portugal, November 3-5, 2010, pg.350-357
 [11] A. W. Harbaugh, "MODFLOW, The U.S. Geological Survey Modular Ground-Water Model—the Ground-Water Flow Process", Chapter 16 of Book 6. Modeling techniques, Section A. Ground Water, 2005
 [12] H. J. Diersch, "FEFLOW Finite Element Subsurface Flow and Transport Simulation System, Reference Manual". WASY GmbH, Berlin, 2005

Generalized Real Numbers Pendulums and Transport Logistic Applications

A. P. Buslaev and A. G. Tatashev

Abstract—A discrete dynamical system, called a real numbers bipendulum, is considered in the paper. This system is the model of relocation of particles on an abstract graph. The behavior of the system is formalized. Competitions resolution rules and movement schedules, i.e. logistic, are given. The movement schedules are given with aid of real numbers, represented in system of base, equal to the graph cardinality. The main problem is investigation of the pendulum behavior depending on rationality or irrationality of logistics. A chaotic pendulum is considered parallel to the logistic pendulum.

In the case of the chaotic pendulums, the plan of tomorrow behavior of a particle is played today. A special case of the egoistic pendulum is considered. In this case plans of the particles are time shifts of N -ary representation of a number called a phase pendulum.

Keywords: Discrete dynamical systems; Markov processes; Number theory; Ergodic theory; Classical Russian literature.

1. FORMULATION OF PROBLEM

1.1. "You are sitting wrong"

Suppose there are N vertices V_0, \dots, V_{N-1} and M particles P_0, \dots, P_{N-1} , Fig1. Each particle is in one of N vertices at every time instant. At the next time instant, after the supreme verdict "you are sitting wrong", particles are trying to change seats, Fig1. If no conflicts take place, then the seats are changed in accordance with a given rule. Here there are logistic plans of particles. This is the so called democratic jumping. Otherwise conflicts are solved in accordance with given rules. We have a dynamical system. Basic essence of this system is an endless search of the correct dislocation. However one of the Russian literature classics [1] dared to say that *any music orchestra cannot be obtained in this manner*. Formally speaking, the system state can be described with a binary matrix. The rows of the matrix correspond to particles. Each row contains a single "one". The index of the column, containing this "one", is equal to the index of the vertex, containing the particles at present time. This index is equal to one of numbers $0, 1, 2, \dots, N-1$.

1.2. Plan logistics

A real number a^j is given. This number is called the P_j particles plan, $j = 0, 1, 2, \dots, M-1$. This number is

This work was supported by Ministry of Education and Science of the Russian Federation, project No. 14.740.11.0397

A. P. Buslaev is with the Department of Mathematics, MADI, Russian Federation apal2006@yandex.ru

A. G. Tatashev is with the Department of Math. Cybernetics, Moscow Tech. Univ. of Communications and Informatics, and the Department of Mathematics, Moscow Automobile and Road State Tech. Univ. a-tatashev@yandex.ru

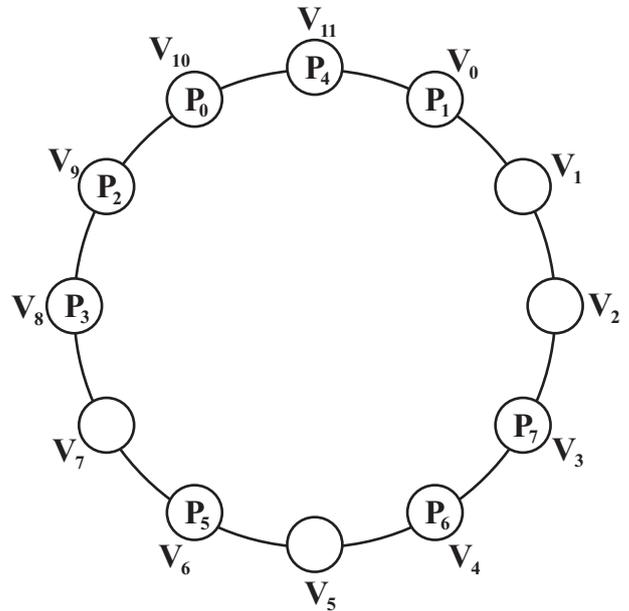


Fig. 1. Round table and dislocation of particles

represented in N -ary system

$$a^j = 0.a_1^{(j)} a_2^{(j)} \dots a_k^{(j)} \dots,$$

where each digit value is equal to one of the number $0, 1, \dots, N-1$. We assume that the number a^j is recorded on the tape, correspondingly to the particle P_j , $j = 0, 1, \dots, M-1$. Each particle reads a digit recorded on its tape at every discrete time instant $T = 1, 2, \dots$. This digit determines the index of the vertex such that the particle tries to pass to this vertex, Fig.2.

1.3 Democratic chaos

Random walks are considered instead of the destroyed system of logistics. In the case of these random walks the next dislocation of particles is determined flip the rest of their coins. Each digit of the number a^j is played before the particle reads this digit, and the value of the digit is equal to i with probability $P_{i,j}$, $j = 0, 1, \dots, M-1$; the numbers $P_{i,j}$ are given, $0 < P_{i,j} < 1, 0 \leq j \leq M-1, P_{0,j} + P_{1,j} + \dots + P_{N,j} = 1$. The main problem of our paper is to investigate the behavior of the dynamical system in the cases of given strategies and rational or irrational plans.

1.4 Simulation of activity

At the initial time instant $T = 1$, the particle P_j reads the first digit of its tape at the initial time instant $T = 1$. The

today	time table				
$a_1^{(0)}$	$a_2^{(0)}$	$a_3^{(0)}$	$a_4^{(0)}$	$a_5^{(0)}$	}
$a_1^{(1)}$	$a_2^{(1)}$	$a_3^{(1)}$	$a_4^{(1)}$	$a_5^{(1)}$	
$a_1^{(2)}$	$a_2^{(2)}$	$a_3^{(2)}$	$a_4^{(2)}$	$a_5^{(2)}$	
$a_1^{(3)}$	$a_2^{(3)}$	$a_3^{(3)}$	$a_4^{(3)}$	$a_5^{(3)}$	

Fig. 2. Turing tapes and plan logistics

particle is in the vertex with index $a_1^{(j)}$, $j = 0, 1, \dots, M - 1$. If no *conflict competition* takes place at time $T = 1$, then each particle will be, at time $T = 2$, in the vertex such that the index of this vertex is equal to the second digit of the plan. If a conflict takes place at time $T = 1$, then one of the competing particles, winning the competition, will be, at time $T = 2$, in the vertex, determined by the plan, and the losing particle, does not move. The tape of winning particle will read the third digit at time $T = 2$. The behavior of the system at time $T \geq 2$ is similar. A *competition takes place* if, at present time, there are particles which try to come from the vertex V_i to the vertex V_j , and particles which try to come from the vertex V_j to the vertex V_i , $0 \leq i, j \leq N - 1$. If $s_{i,j}$ particles try to move from the vertex V_i to the vertex V_j and $s_{j,i}$ particles try to move from the vertex V_j to the vertex V_i , then the particle, trying to move from the vertex V_i to the vertex V_j , win the competition with probability

$$\frac{s_{i,j}}{s_{i,j} + s_{j,i}}$$

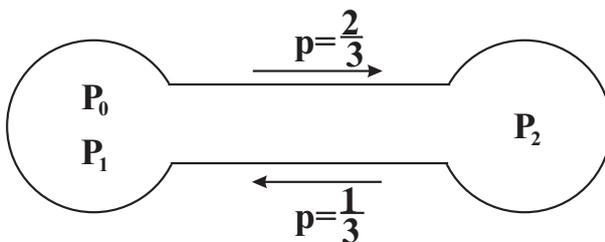


Fig. 3. Conflict resolution

1.5 Quantitative and qualitative characteristics

Denote by $D_i(T)$ the number of transitions of the particle P_i tape on the time interval $[0; T]$, $i = 0, 1, \dots, M - 1$; $T = 1, 2, \dots$; $H(t)$ is the number of conflicts, on time interval $(0; T]$; $H_i(t)$ is the number of conflicts, losing by the particle P_i in time interval $(0; T]$, $T > 0$.

The limit

$$w_i = \lim_{T \rightarrow \infty} \frac{D_i(T)}{T}, \quad i = 0, 1, \dots, M - 1,$$

is called the *velocity of the particle P_i tape*, $i = 0, 1, \dots, M - 1$ if this limit exists.

The limit

$$h = \lim_{T \rightarrow \infty} \frac{H(T)}{T}, \quad i = 0, 1, \dots, M - 1,$$

is called the *intensity of conflicts* if this limit exists.

The limit

$$h_i = \lim_{T \rightarrow \infty} \frac{H_i(T)}{T}, \quad i = 0, 1, \dots, M - 1,$$

is called the *intensity of conflicts losing by the particle P_j* if this limit exists.

The limits (1) – (3) depend on the process realization. These limits can exist or not exist depending on the realization. The system is in the state of system after a time instant T_{syn} if, after the instant T_{syn} , no conflicts take place, and each particle comes to the next state at every instant.

2 RATIONAL OR IRRATIONAL LOGISTIC PLANS

If the number a_j is rational, then this number can be represented as a periodic fraction

$$a_j^{(j)} = 0.a_1^{(j)} a_2^{(j)} \dots a_{k_j}^{(j)} (a_{k_j+1}^{(j)} a_{k_j+2}^{(j)} \dots a_{k_j+l_j}^{(j)}),$$

where k_j is the length of the aperiodic part of the number a_j representation, and l_j is the length of the repeating part of the representation, $j = 0, 1, \dots, M - 1$.

2.1. An example of a rational bipendulum

Consider an example. Suppose $N = M = 2$; a_0 and a_1 are the numbers $1/3$ and $1/5$, which are represented in the binary system as periodic fractions

$$a_0 = \frac{1}{3} = 0.(01), \quad a_1 = 0.(0011).$$

Proposition 1. Suppose

$$a_0 = 0.(01), \quad a_1 = 0.(0011).$$

Limits (1), (2) and (3) exist with probability 1, and

$$h = \frac{2}{5}, \quad h_1 = h_2 = \frac{1}{5}, \quad w_1 = w_2 = \frac{4}{5}.$$

Proof. Consider a Markov chain are vectors (i_0, i_1) , where i_j is the number of the period digit, which the particle P_j reads, $j = 1, 2$, $1 \leq i_0 \leq 2$, $1 \leq i_1 \leq 4$. There are 8 states of the chain

$$E_1 = (1, 1), \quad E_2 = (1, 2), \quad E_3 = (1, 3), \quad E_4 = (1, 4),$$

$$E_5 = (2, 1), E_6 = (2, 2), E_7 = (2, 3), E_8 = (2, 4).$$

Denote by p_{ij} the probability of transition from the state E_i to the state E_j , $1 \leq i, j \leq 8$. The transitions from the state E_i to the state E_j , $1 \leq i, j \leq 8$. The transition probabilities matrix has the form

$$P = (p_{ij}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The state E_3 and E_5 are inessential. The other states form single communication class. Hence steady probabilities of these states exist, [3,4]. Denote by p_i the steady probability of the state E_i , $i = 1, 2, 4, 6, 7, 8$. The steady probabilities satisfy the system of equations

$$\begin{aligned} p_1 &= \frac{p_4}{2} + p_8, \quad p_2 = \frac{p_6}{2}, \\ p_4 &= p_7, \quad p_6 = p_1, \quad p_7 = p_2 + \frac{p_6}{2}, \\ p_8 &= p_4/2, \quad p_1 + p_2 + p_4 + p_6 + p_7 + p_8 = 1. \end{aligned}$$

The solution of this system is

$$p_1 = p_4 = p_6 = p_7 = \frac{1}{5}, \quad p_2 = p_8 = \frac{1}{10}.$$

Suppose the chain comes to the state E_i in the time interval $(0, T]$. The value $\theta_i(T)/T$ tends to steady probability of the state with probability 1 [4]

$$\lim_{T \rightarrow \infty} \frac{\theta_i(T)}{T} = p_i.$$

Conflicts take place only in the states E_4 and E_6 . The number of conflicts in the time interval $(0, T]$ equals to

$$H(T) = \theta_4(T) + \theta_6(T).$$

Therefore, with probability 1,

$$h = \lim_{T \rightarrow \infty} \frac{H(T)}{T} = p_4 + p_6 = \frac{2}{5}.$$

Since the particle P_i loses each conflict with probability 1, then we have

$$\lim_{T \rightarrow \infty} \frac{H_i(T)}{H(T)} = \frac{4}{5}, \quad i = 1, 2.$$

Proposition 1 has been proved.

2.2 Properties of the real value pendulum

(1) For some states of a real value bipendulum generated by irrational numbers (irrational bipendulum) limits (1) - (3) do not exist.

(2) Numeric characteristics of a rational real valued pendulum (1) - (3) are defined correctly. The proof is immediate from the definition so that these characteristics on the interval $[0.5, 1]$. Better estimations is a problem to be solved.

(3) For irrational pendulums typical behavior system is described by democratic chaos.

2.3 Generalize pendulums fluctuations with a phase shift

Common pendulum is considered, provided that logistical plans of particles are time shifts generated by the same number (father, mother, source).

(1) If father is rational with probability is equal to one and finite expectation time pendulum is going to synergy state.

(2) Synergy expectation time of a rational bipendulum can be lower estimated by border reaching problem during random walk of an integer valued cell on a right line.

(3) Irrational bipendulum of wellfamous irrational numbers, as can be seen below, are well defined by democratic chaos structure.

3 APPROXIMATION OF IRRATIONAL BIPENDULUMS WITH RANDOM WALKS BIPENDILUMS

Suppose $N = M = 2$,

$$p_{00} = p_{01} = q, \quad p_{10} = p_{11} = p.$$

Therefore any digit of each tape is equal to 1 with probability p , and the digit is equal to 1 with probability q , $p + q = 1$. Consider the stochastic process $X(T)$, $T = 2, 3, \dots$, which, at each time instant, is in one of 5 states G_i , $i = 1, 2, \dots, 5$. The process $X(T)$ is in the state G_1 if both the particles are in the vertex P_0 , and no conflict took place at time $T - 1$. The process $X(T)$ is in the state G_2 if both the particles are in the vertex P_0 , and a conflict took place at time $T - 1$. The process $X(T)$ is in the state G_3 if both the particles are in different vertices. The process $X(T)$ is in the state G_4 if both the particles are in the vertex P_1 , and no conflict took place at time $T - 1$. The process $X(T)$ is in the state G_5 if both the particles are in the vertex P_1 , and a conflict took place at time $T - 1$.

Suppose $p_i(T)$ is the probability of the stochastic process $X(T)$ is in the state E_i , and

$$p_i = \lim_{T \rightarrow \infty} p_i(T), \quad i = 1, 2, \dots, 5.$$

The following theorems have been proved.

Theorem 1. *The stochastic process $X(T)$ is a Markov chain. There exist steady state probabilities p_1, p_2, \dots, p_5 . These probabilities satisfy the system of equations*

$$p_1 = q^2 p_1 + q^2 p_3 + q^2 p_4, \tag{4}$$

$$p_2 = \frac{pq}{2} \cdot p_3, \tag{5}$$

$$p_3 = 2pqp_1 + qp_2 + pqp_3 + 2pqp_4 + pp_5, \tag{6}$$

$$p_4 = p^2 p_1 + pp_2 + p^2 p_3 + 2pqp_4 + pp_5, \tag{7}$$

$$p_5 = \frac{pq}{2} \cdot p_3, \tag{8}$$

$$p_1 + p_2 + \dots + p_5 = 1. \tag{9}$$

The average number of conflicts per a time unit is equal to pqp_3 . The average number of conflicts, where the particle P_j loses, per a time unit, $j = 0, 1$, is equal to $pqp_3/2$.

Theorem 2. Suppose $p = q = 1/2$. Then the average number of conflicts per a time unit equals $1/20, j=0,1$.

The proof of Theorem 1 is based on representation of the dynamical system in the form of the Markov chain. Steady state probabilities of this chain have been found. Having found these state probabilities, we can find the tape velocities. Theorem 2 follows from Theorem 1.

4. IRRATIONAL PENDULUMS COMPUTER SIMULATION

4.1. Irrational plans

If plan of particles are irrational numbers, then the system cannot be described with finite Markov chain.

Simulation experiments have been implemented. The plans were irrational numbers such as $\sqrt{2}(\text{mod } 1), \sqrt{3}(\text{mod } 1), \pi - 3, \sqrt{5}(\text{mod } 1)$.

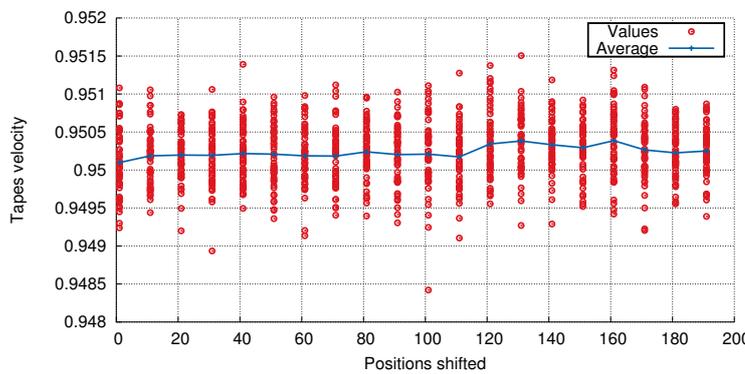


Fig. 4. Phases logistic pendulum $\sqrt{2}(\text{mod}1) - \sqrt{3}(\text{mod}1)$

4.2. Chaotic behavior of the system in the case of irrational plans

Let us describe results of experiments in the case $N = M = 2$ (bipendulum). The results of experiments show that the velocity of particles is equal to $\frac{19}{20}$ as in the case of the chaotic pendulum. The simulation experiments are stable in the case of rational plans. The simulation experiments can be unstable in the case of irrational plans.

4.3. Phase pendulums

Let us get the plan a_1 , shifting plan a_0 onto a fixed number of positions. If plans are rational numbers, then the system comes to the state of synergy after a time interval with a finite expectation. If plans are irrational numbers, then the system behave such as it comes to the state of synergy, with probability 1, after a time interval T_{syn} , but the expectation of T_{syn} is infinite.

The behavior of the bipendulum has been investigated with plans $a_0 = \sqrt{2}$ and a_1 such that we get a_1 , shifting a_0 onto c positions. The dependence of the average velocities on the time interval $(0, T)$ is shown in Fig. 5. We suppose that the system comes to the state of synergy with probability 1 for a time interval. However the duration of this interval is large if c is large.

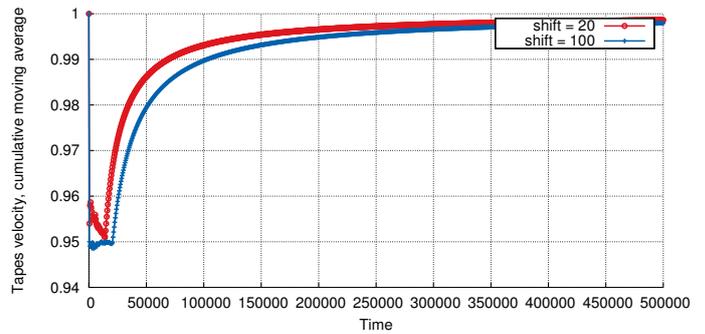


Fig. 5. Phases shift $\sqrt{2} \text{ mod } 1$

5. COMMENTS, DRAWBACKS AND FURTHER RESEARCH

(1) General transport-logistical problem is described in article [5]. There were considered flows on linear networks (neckless type) and plane flows (chainmail).

(2) Rational pendulum as $N = M = 2$ (bipendulum) is introduced and considered in [8]. Some aspects are described.

(3) In [9] are described algebras connected with dynamic system analysis problem (real valued pendulum) in case of rational logistical plans. The Bernoulli algebra divide the triangle of rational numbers to subalgebras such that the tape velocity of the bipendulum equals 1 (*the synergy*). If the plans belong to different subalgebras, then the tape velocity can be less than 1. This problem is studied.

(4) From computational experiments with irrational bipendulum can be seen remarkable behavior distinctions comparing with rational pendulum. In particular, for *classical* irrational examples velocity bipendulum estimation with half a million characters is equal to correspond characteristic in case of a democratic chaos. Behavior of a system is described according to the next rules *each character equiprobably is 0 or 1 independent from other characters*.

(5) Phase pendulum fluctuations give foreseen results, so that synergic state is always to happen, however in irrational case, with remarkable lower velocity.

(6) It is interesting to find exact low border of a real valued pendulum main numeric characteristic. For instance in the case when quantity of vertex and number are equal. This problem is solved only for rational pendulum.

(7) It is proved from random walk theory that on two dimension cell any point is reached with 100% probability during infinite expectation time. In case of a random walk on a cell with dimension more than two, the particle returns to given point with probability less than one. We may consider that phase pendulums with more than two particles and generative irrational logistics are not going to be synchronized for finite time.

(8) Since the plans are given with real numbers, we use constructive approaches to determine irrational numbers. These approaches to determine irrational numbers. These methods were not used in analysis of models, investigated in [5 - 7]. The theory of Markov chains and the theory of

Markov chains and the theory of random walks are used as in [5 -7].

I. ACKNOWLEDGMENTS

This work was supported by the Ministry of Education and Science of Russian Federation under Grant No. 2.723.2014/K

REFERENCES

- [1] Krylov I. A. Fables. Moscow, Detskaya literatura, 1989.
- [2] Feller W. An introduction to probability theory and its applications. Vol. 1. John Willey, New York, 1970.
- [3] Kemeny J. G., Snell J.L. Finite Markov chains. Springer Verlag. New York, Heidelberg, Tokio, 1976.
- [4] Borovkov A. A. Probability theory. Moscow, Nauka, 1986.
- [5] Kozlov V. V., Buslaev A. P., Tatashev A. G., Yashina M.V. Monotonic walks of particles on a chainmail and coloured matrices. Proceedings of the 14th International Conference on Computational and Mathematical Methods in Science and Engineering, CMSSE 2014, Cadiz Spain, June 3 – 7 2014, vol. 3, pp. 801 – 805.
- [6] Kozlov V. V., Buslaev A. P., Tatashev A. G. Monotonic walks on a necklace and coloured dynamic vector. International Journal of Computer Mathematics (2014). DOI: 1080 00207150.2014/915964
- [7] Kozlov V. V., Buslaev A. P., Tatashev A. G. A dynamical communication system on a network. Journal of Computational and Applied Mathematics, vol. 275 (2015), pp. 247 – 261.
- [8] Kozlov V. V., Buslaev A. P., Tatashev A. G. On real-valued oscillations of bipendulum. Applied mathematical Letters (2015) pp. 1 – 6. DOI 10.1016/j.aml. 2015.02.003
- [9] Kozlov V. V., Buslaev A. P., Tatashev A. G. Bernoulli algebra on common fractions. International Journal of Computer Mathematics, in print, 2015, pp. 1 – 6.

Scorpion Envenomation in Naama, Algeria

Schehrazad Selmane

Abstract—In Algeria scorpion envenomation represents a real public health problem with a population at risk of scorpion stings estimated at 68% of the national population. A total of 903,461 scorpion sting cases and 1996 deaths were recorded by health services between 1991 and 2012.

The physical-geographic and climate conditions make the province of Naama a conducive environment for scorpion species and an endemic zone for scorpion envenomation. A total of 22,498 scorpion stings and 66 deaths were recorded by the Department of Public Health of the province between 1999 and 2013.

An early warning system is an essential tool for preparedness and effectiveness of scorpion stings control; it could help determine the appropriate number of antivenom vials necessary in health facilities and anticipate the demand for antivenoms and symptomatic drugs so that they can be distributed in advance. To this end, we performed a regression analysis to estimate the relationship between scorpion sting cases and climate conditions. The obtained results showed that the scorpion activity in Naama province is climate dependent phenomenon; the temperature and precipitation are the main factors; they were used to derive the best predictive model for scorpion sting cases. If we know beforehand the change on climate variables, we can use regression model to predict the number of scorpion sting cases using those climate variables.

Index Terms—Climate, Correlation, Naama province, Precipitation, Regression analysis, Scorpion, Scorpion Sting, Temperature.

I. INTRODUCTION

SCORPION stings represent a public health problem in many tropical and subtropical regions. North Saharan Africa, Sahelian Africa, South Africa, Near and Middle-East, South India, Mexico and South Latin America, East of the Andes are identified as at risk areas involving 2.3 billion at risk population. The annual number of scorpion stings exceeds 1.2 million leading to more than 3250 deaths worldwide [1].

Scorpions are venomous arthropods of the class Arachnida. They are grouped into six families, 70 genera and more than 1500 species. Only 25 species are deadly to humans, and most potentially lethal to human belong to the family Buthidae which primarily is distributed in Africa and Southeast Asia. Scorpions are easily recognizable because of their morphological structures; they are of 13 to 220 mm length. They are primarily nocturnal, fearful of nature, not aggressive and lucifugous. They withstand aggressive environmental factors either cold or hot. They are active in the spring and summer. The longevity of the adult varies from 2 to 10 years or even twenty years. They feed essentially on insects and on spiders, preferring the alive or freshly killed prey. The big scorpions eat invertebrates, small lizards, snakes and even small mice.

S. Selmane is with L'IFORCE. Faculty of Mathematics. University of Science and Technology Houari Boumediene, Algiers, ALGERIA e-mail: cselmane@usthb.dz

Scorpions are cannibals inter/intra species and even the mother can eat its young. They can stay almost two years without food and water. They are found in diverse habitats: under stones, rocks, tree bark and old buildings. They look dark corners where they dig burrows. On the other hand certain scorpions affect the neighborhood of houses, take place between sheets, in shoes, in kitchens and bathrooms. They detect their prey by senses of contact and sound, and similar to the way seismologists locate earthquakes. They use their venom to kill or paralyze their prey so it can be eaten. The sting of most scorpions can be very painful, like a bee sting, although most are not lethal. Scorpion stings should always be treated as a medical emergency that requires treatment as soon as possible [1], [14].

In Algeria scorpion envenomation is a real public health problem. Twenty-eight species and fourteen genera of scorpions were identified in the country and the most important health threatening scorpions found belong to the Buthidae family. They include *Androctonus australis* and *Leiurus quinquestriatus*, and are found mostly in the southern highlands and in the Atlas and Hoggar mountain ranges [14]. The population at risk of scorpion stings is estimated at 68% of the total national population. Among the 48 provinces of the country, 39 provinces are affected by the scorpion envenomation accidents. Fourteen provinces belonging to Highlands and Sahara account together for almost 90% of patients stung and the entire deaths. The incidence varies between less than 7 scorpion stings per 100,000 inhabitants in the northern provinces and more than 1000 scorpion stings per 100,000 inhabitants in those of the South [11].

Naama ranks among the endemic provinces of the country and records every year a high incidence of scorpion stings. A total of 22,498 scorpion stings and 64 deaths were recorded by the Department of Public Health of the province between 1999 and 2013. The public health authorities of the province are faced to scorpionism, and consequently, they are required to establish prevention and control strategies. An early warning system is an essential tool for preparedness and effectiveness of scorpion stings control; it could help determine the appropriate number of antivenom vials necessary in health facilities and anticipate the demand for antivenoms and symptomatic drugs so that they can be distributed in advance in this endemic province.

As far as we know the first mathematical approach on predicting scorpion sting incidence is due to Chowell and al; they analyzed the significance of climate variables to predict the incidence of scorpion stings in humans in the state of Colima (Mexico) using multiple linear regression [2]. Other studies on other regions on the influence of climate factors on scorpion envenomation following the statistical approach conducted by Chowell and al have been performed using

simple statistical analysis and correlation between scorpion stings and climate variables [6], [13].

The scorpion envenomation surveillance in Algeria is based on a passive system. Neither the analysis nor interpretation of data were undertaken; the only performed statistical approach to scorpionism is due to Selmane and El hadj [12].

In the aim to estimate the effects of climate variables on scorpion envenomation in Naama, we performed a regression analysis to estimate the relationship between scorpion sting cases (the dependent variable) and climate conditions (the independent variables). The obtained results showed that the scorpion activity in Naama province is climate dependent phenomenon; the temperature and precipitation are the main factors; they were used to derive the best predictive model of scorpion sting cases.

II. MATERIALS AND METHODS

A. Scorpionism in Algeria

Scorpion stings are common in Algeria and represent an actual public health problem. Health services have recorded between 1991 and 2012 a total number of 903,461 scorpion stings and 1996 related deaths. The number of stings doubled between 1996 and 1999 and from 1999 to 2012 a weak fluctuation of this number is perceived (Fig. 1.). Unlike, the number of deaths have halved. The geographical distribution of the incidence per 100,000 inhabitants of scorpion stings for the year 2012 (Fig. 2.) mapped using MapInfo Professional 11.0, shows that the incidence predominate in Highlands and Sahara, which together account for almost 90% of patients stung [11].

B. Study Area : Naama Province

Naama is one of the 48 provinces of Algeria. It is situated in the west between the Tell Atlas and the Saharan Atlas at 33° 16' N and 0° 19' W of the equator and more than 1,000 meters above sea level. The province is made up of seven districts gathering twelve municipalities over a land size of about 29,950 km² with an estimated population of 238,087 as 2013, that is, a population density of 8 inhabitants per km² [9]. The climate is split into two main seasons; cold and relatively wet season which extends from November to April and a hot and dry season which extends from May to October. However, this climate is marked by irregularities. This is significant not only from one year to another, but also in the distribution between the different months. Rainfalls remain low and irregular; it is heterogeneous in time and space [8].

C. Data

The study period comprises 96 months from January 2003 to December 2010 and consists of two different monthly data sets : epidemiological data and meteorological data; the climates conditions being assumed to be of great influence on scorpion distribution and activity. Monthly scorpion sting cases were obtained from the Department of Public Health of the province of Naama, and monthly mean temperature (in °C), period of sunshine (in hours), precipitation amount (in mm), wind

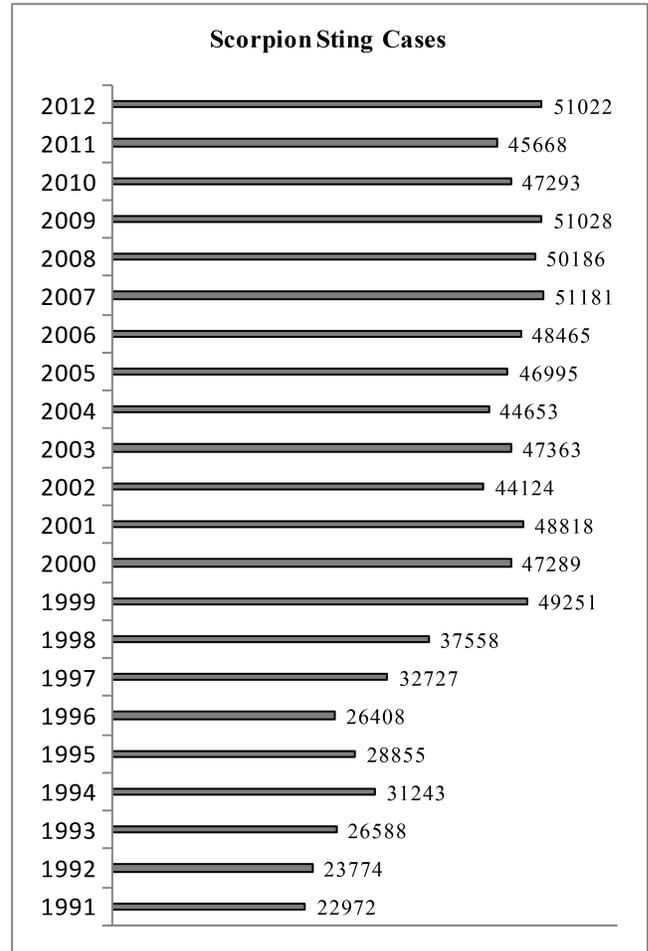


Fig. 1. The annual evolution of recorded scorpion sting cases in Algeria.

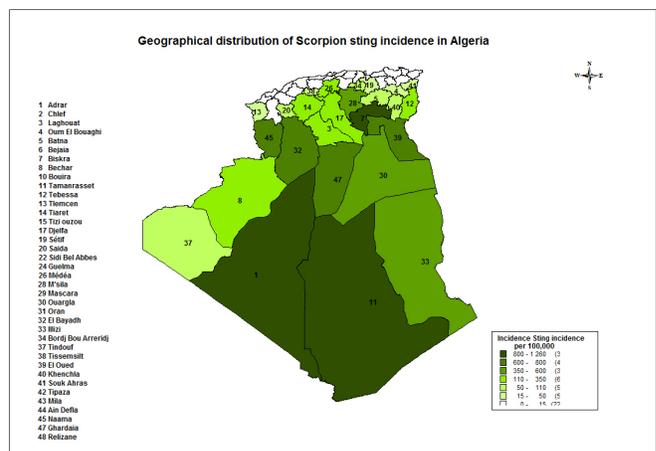


Fig. 2. Geographical distribution of scorpion sting incidence in Algeria.

speed (in m/s), and relative humidity (in %) recorded by the weather station of Naama (Latitude : $33^{\circ} 16' N$, Longitude : $0^{\circ} 18' W$, Altitude: $1166 m$) were extracted from the National Office of Meteorology [8].

D. Statistical Modeling Method

Descriptive statistics are performed to quantitatively describe the main features of the data. Time series analysis of scorpion sting cases and climate factors are also performed in order to extract meaningful statistics and other characteristics of the data and also to analysis of temporal trends of the variables. To find any significantly relationship between the scorpion sting variable and the climate variables, first, the scatterplots and Pearson product-moment correlation coefficient are drawn up, then a regression analysis is undertaken.

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. A regression model relates the dependent variable, Y to a specified function of independent variables, X , and unknown parameters, β :

$$Y \approx f(X, \beta).$$

The approximation is usually formalized as

$$E(Y|X) = f(X, \beta)$$

where $E(Y|X)$ is the average value of the dependent variable when the independent variables are fixed. One method of parameter estimation is ordinary least squares; which consists to minimize the sum of squared residuals [3].

A best regression model has to fulfil the following features :

- The value of R-square should be more than 60 percent. Higher the R-square value, better the data fitted.
- Most of the independent variables should be individually significant to influence the dependent variable (this matter can be checked using t-test).
- The independent variables should be jointly significant to influence or explain dependent variable (This can be checked using F-test).
- No serial correlation in the residual (can be tested using Bruesch-Godfrey serial correlation LM test).
- No heteroscesticty in the residual (can be tested using Bruesch-Pegan-Godfrey Test).
- Residuals should be normally distributed (can be tested using Jarque Bera statistics).

When all these features are met; the model can be used for forecasting [3].

All performed computations and generated figures were carried out with Eviews 7 software.

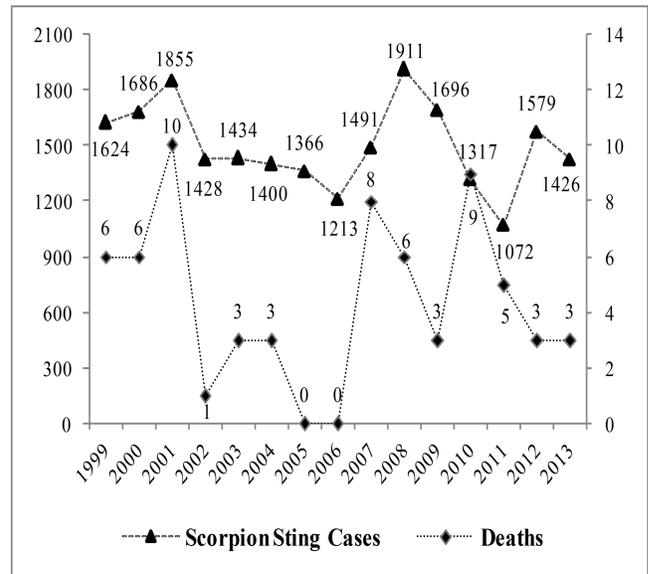


Fig. 3. Evolution of annual recorded scorpion sting cases and deaths in Naama province from 1999 to 2013.

III. DATA ANALYSIS AND RESULTS

A. Annual evolution of recorded scorpion sting cases

A total of 22,498 scorpion stings and 66 deaths were recorded by the Department of Public Health of the province of Naama between 1999 and 2013; the yearly distribution is plotted in Fig. 3. The highest total yearly scorpion sting cases occurred in the years 2001 and 2008 with 1851 and 1911 respectively and the highest number of deaths were notified in 2001 with 10 deaths and in 2010 with 9 deaths [11]. We note pronounced fluctuations on the yearly evolution; this is to be expected due the fact that scorpion activity is related to climate and the latter is marked by irregularities.

B. Geographical distribution of scorpion envenomations in Naama

The geographical distribution of the incidence per 100,000 inhabitants of scorpion stings for the year 2013 by municipality for Naama province is mapped using MapInfo Professional 11.0 (see Fig. 4.). Almost half of scorpion sting cases occurred in Mechria (27.1%) and Ain Sefra (21.9%); the incidence per 100,000 inhabitants is 517.08 for Mechria and 487.33 for Ain Sefra. The highest incidence was recorded in Mekmen Ben Amar (1304) and Sfisifa (1501).

C. Scorpion sting cases and population size by municipality

To estimate the importance of the heterogeneity of the province, we analyzed the correlation between the total number of scorpion stings, the population size, the population density, and the incidence per municipality for the year 2013. The number of scorpion stings showed a high degree of correlation with the population size ($r = 0.943$) (Fig. 5.) and a high degree of correlation with population density ($r = 0.909$).

For all municipalities, the total number of recorded scorpion

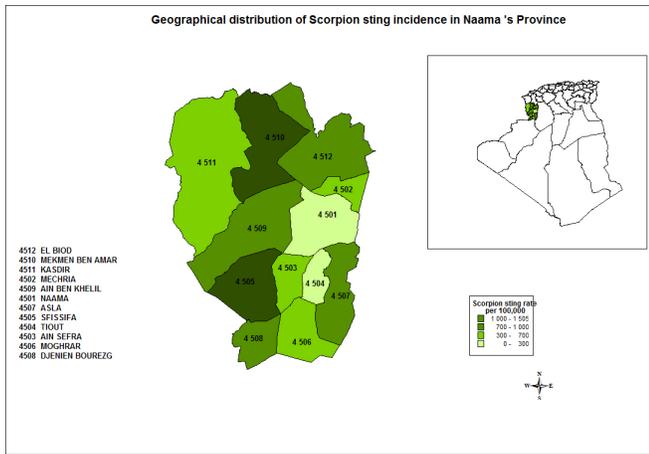


Fig. 4. Geographical distribution of scorpion sting incidence in Naama's province.

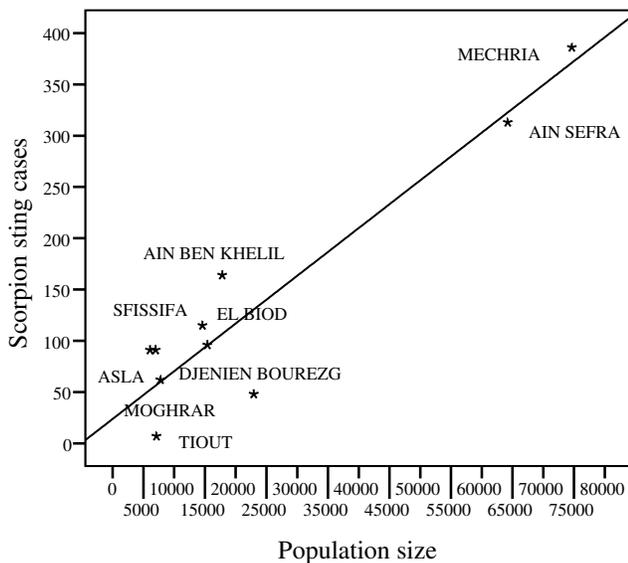


Fig. 5. The scatterplot of the total number of recorded scorpion stings and the total population size by municipality. The straight line represents the linear regression equation (1).

stings (S_{Mun}) and the population size (P_{Mun}) are related as follows :

$$S_{Mun} = 0.005 P_{Mun} \quad (1)$$

an equation that explains 86.5% of the observed variance.

D. Descriptive statistics of the variables

The descriptive statistics of the monthly data for the study period (2003 – 2010) and the Pearson product-moment correlation coefficient (r) between scorpion sting cases (S) and climate variables are displayed in Table I. The climate variables with strong positive correlation coefficient with scorpion sting cases are mean temperature (T), mean maximum temperature ($MaxT$) and mean minimum ($MinT$). This confirms the increasing activity of scorpion with increasing the environment

TABLE I
DESCRIPTIVE STATISTICS OF THE VARIABLES

Variables	Minimum	Maximum	Mean	SD	r
S	0	584	124,89	152,35	
T	3,50	30,40	16,82	8,11	0.891**
$MinT$	-3	22,30	10,20	7,27	0.887**
$MaxT$	7,90	38,70	23,43	9,02	0.888**
P	0	157,30	19,23	22,15	-0.153
RH	23	82	50,36	15,63	-0.799**
$MinRH$	4	63	26,53	13,85	-0.740**
$MaxRH$	40	96	74,61	14,71	-0.854**
I	134,90	361,90	251,34	51,18	0.609**
W	1,40	5,10	3,12	0,86	0.182
$MaxW$	6,30	16,70	12,30	2,55	0.530**

** The correlation is significant at the 0.01 level (bilateral).

temperature. There is strong negative correlation between scorpion sting cases and relative humidity (RH), maximum relative humidity ($MaxRH$) and minimum relative humidity ($MinRH$). The correlation between sunshine time (I) (resp. maximum wind speed ($MaxW$)) and the scorpion sting cases is mild ($r = 0.609$) (resp. $r = 0.530$). The correlation between accumulated precipitation (P) amount (resp. wind speed (W)) and the scorpion sting cases is very weaker ($r = -0.153$) (resp. ($r = 0.182$)).

The coefficient of variation CV ($CV = SD/Mean = 1,22$ where SD is the standard deviance) is closer to 1, which means the greater the variability of scorpion data.

E. Time Series Analysis

Scorpion stings are recorded throughout the year and the epidemiological year starts from March to April (lowest scorpion stings cases) with peaks in July-August, to resume its lowest rate toward November-December (Fig. 6 (A)). The monthly peaks are observed in July (27.9% of cases) and in August (27.9% of cases), accounting alone for more than half of cases (55.8% of cases). The maximum recorded scorpion sting cases during the study period occurred in July and August 2008, with 584 and 570 cases respectively; for these dates highest temperature recorded was $37.9^{\circ}C$ in July 2008 and $37^{\circ}C$ in August 2008. Most of the cases (70.4%) were notified during the summer period followed by Autumn period (16.7%), then Spring period (12.3%) (Fig. 6 (B)).

The monthly recorded scorpion sting cases with monthly mean maximum temperature and mean minimum temperature and with monthly accumulated precipitation are plotted in Fig. 7. Temperature follows the same trends with scorpion sting cases. The highest accumulated precipitations were recorded in October 2008 and in September 2005 with an amount of 157.3 mm and 93.2 mm respectively and the corresponding recorded scorpion sting cases were 76 (average number is 79 and minimum is 40) and 134 (average number is 169 and minimum is 130) recorded cases respectively; this is against the stated conclusion in [2].

F. Regression Analysis

Temperature, relative humidity, and sunshine time are highly pairwise correlated; the Pearson product-moment correla-

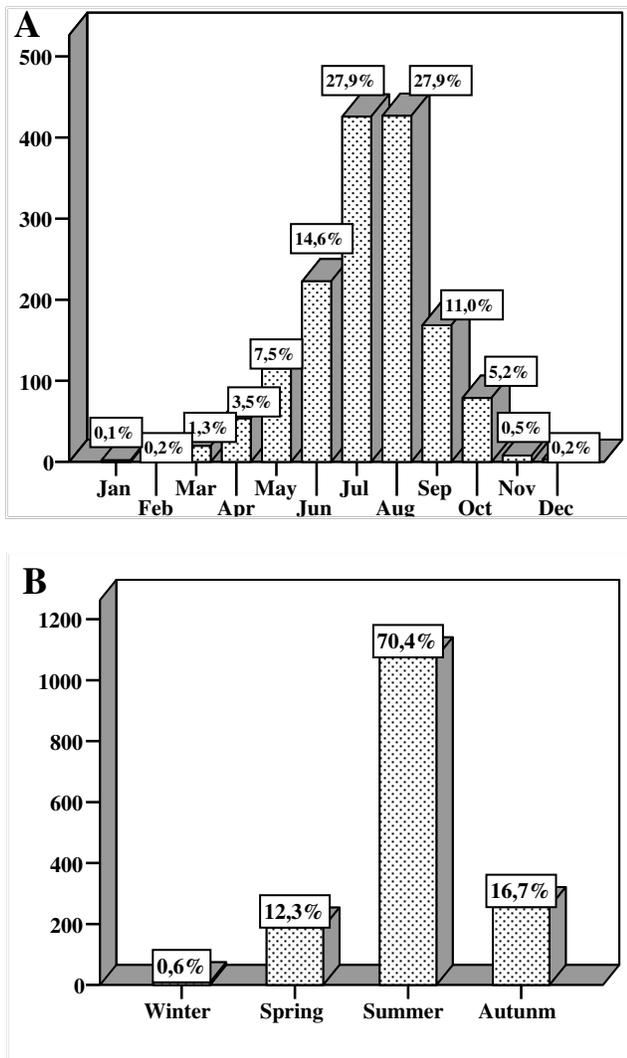


Fig. 6. **A** Monthly average distribution of recorded scorpion sting cases. **B** Seasonal average distribution of recorded scorpion sting cases in Naama province for the period 2003 – 2010.

tion coefficient between temperature and relative humidity is $r = -0.901$, between temperature and sunshine time is $r = 0.734$ and between relative humidity and sunshine is $r = -0.823$. Therefore these variables will impart nearly exactly the same information to a regression model. To avoid the multicollinearity and the unreliability of the regression model's regression coefficients related to these highly pairwise correlated variables, we included into the model only the temperature. The choice of the temperature is justified by the fact that the activity of scorpions increases with increasing temperature [12].

The scatterplot (Fig. 8. A) between the monthly scorpion sting cases and the monthly mean temperature shows a quadratic relationship. We therefore performed a regression analysis to regard (S) as dependent variable and T and T^2 as independent variables. Even though the model is good, it cannot be used for forecasting; the residuals were heteroscedastic and the residuals were not normally distributed.

The scatterplot (Fig. 8. B) between the monthly squared

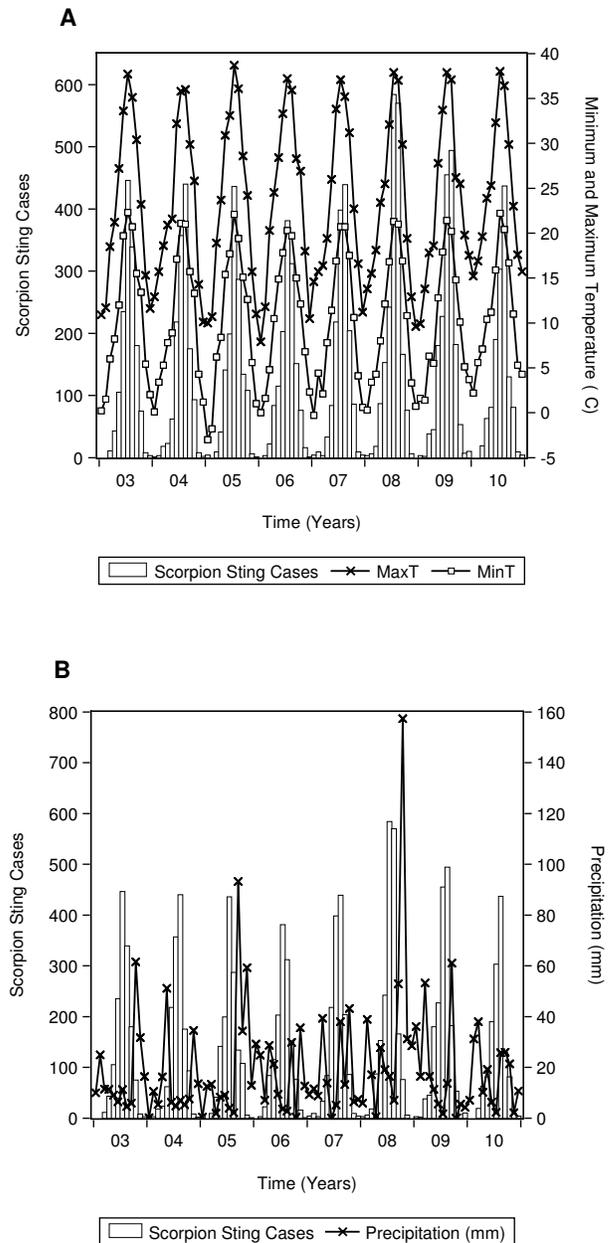


Fig. 7. Time series of the monthly recorded scorpion sting cases (bars); **A** with monthly average of the maximum and minimum temperature and **B** with monthly accumulated precipitation in Naama province for the period 2003 – 2010.

root of scorpion sting cases and the monthly square of mean temperature shows a linear relationship; ($S^{1/2}$) is strongly correlated with T^2 with Pearson product-moment correlation coefficient $r = 0.978$. We therefore performed a regression analysis to regard ($S^{1/2}$) as dependent variable and T^2 and all the other climate variables as well as a trend variable to account for non-climatic factors such human behavior, degradation of the environment and other factors that could influence the number of sting cases, as independent variables. The choice of the model was based on the coefficient of

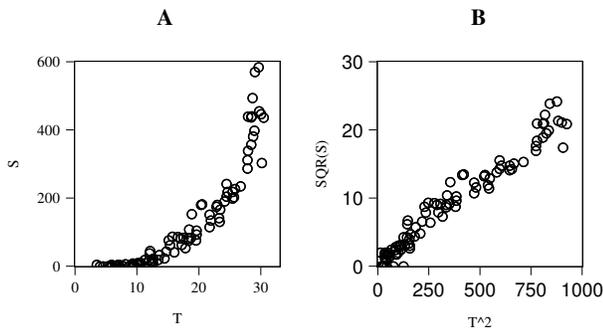


Fig. 8. The scatterplot : (A) of recorded scorpion sting cases and temperature, (B) of squared root of scorpion sting cases and square of temperature.

TABLE II
MODEL OUTCOMES

Dependent Variable: SQR(S)
Method: Least Squares
Sample: 2003Month1 2010Month12
Included observations: 96

Variable	Coefficient	Std. Error	t-Statistic	Prob.
T^2	0.023998	0.000351	68.39514	0.0000
P	0.021780	0.005399	4.034130	0.0001
R-squared	0.959746	Akaike info criterion	3.528154	
Adjusted R-squared	0.959317	Durbin-Watson stat	1.510372	
S.E. of regression	1.397706	Log likelihood	-167.3514	

determination R^2 , the Akaike information criterion (AIC), and Standard Error (SE). A model with higher R^2 , smallest AIC, and lower standard error and fulfilling the features of the best regression model for forecasting corresponds to the model incorporating only T^2 and P as independent variables, and the squared root of S as dependent variable. The model outcomes are displayed in Table II. The estimation equation is given by :

$$S^{1/2} = 0.0239 T^2 + 0.0218 P \quad (2)$$

- The value of R^2 is more than 60% hence the model is acceptably fitted. It indicates also that 95.93% variance in the dependent variable can be explained jointly by the temperature and precipitation; the remaining 4.07 percent variation in the dependent variable can be explained by residuals or other variables other than the selected independent variables.
- The independent variables T^2 and P are individually significant to influence the dependent variable; their corresponding $p - value$ are less than 5%.
- Temperature and precipitation are jointly significant to influence the dependent variable; the $p - value$ of F-statistic is less than to 5%.
- The residuals are normally distributed; Jarque-Bera value 4.644 is less than 5.99 and the corresponding $p - value = 0.098$ is more than 5%.
- Moreover the residuals are not serially correlated; the Bruesch-Godfrey serial correlation LM test shows that

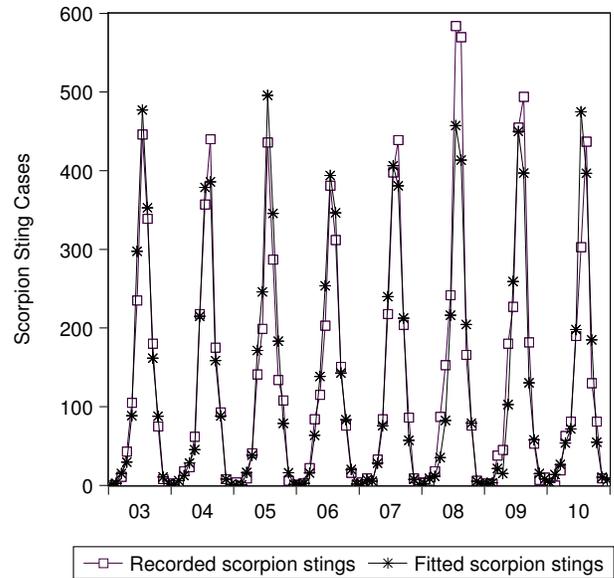


Fig. 9. Recorded scorpion sting cases versus simulated scorpion sting cases during the period 2003 – 2010.

corresponding $p - value = 0.0532$ is more than 5%.

- Finally, according to Bruesch-Pegan-Godfrey Test, the residuals are not heteroscedastic; the corresponding $p - value = 0.061$ is more than 5%.

All features of a best regression model are fulfilled, thus the model can be used for forecasting.

The simulated scorpion sting cases for the period 2003 – 2010 are closely approximated to the recorded data (see Fig. 9.) with correlation $r = 0.968$.

The predicted number of scorpion stings for the year 2011 was computed using the model equation (2) and temperature and precipitation for 2011 [8] and was compared with recorded scorpion stings for the same year and plotted in Fig. 10. The correlation between simulated and recorded scorpion sting cases for this year is very strong ($r = 0.996$).

IV. DISCUSSION

Using the monthly recorded scorpion sting data for the period 2003 – 2010 for Naama province, the linkage between scorpion stings and weather conditions was demonstrated using a regression analysis. The temperature and precipitation are the retained climate factors for this province. This raises optimism for forecasting scorpion stings provided that appropriate climate information are at our disposal. If we know beforehand the change in the climate variables, we can use the built regression model to estimate how much the change in the value of those variables influence the number of cases of scorpion stings. This could be used to help health authorities determine the appropriate number of antivenom vials necessary for the province in advance. This study represents also an important step to find a way to help in the designing of a control strategy.

In conclusion, our study shows optimism for weather-based forecasting of scorpion stings. It represents an important

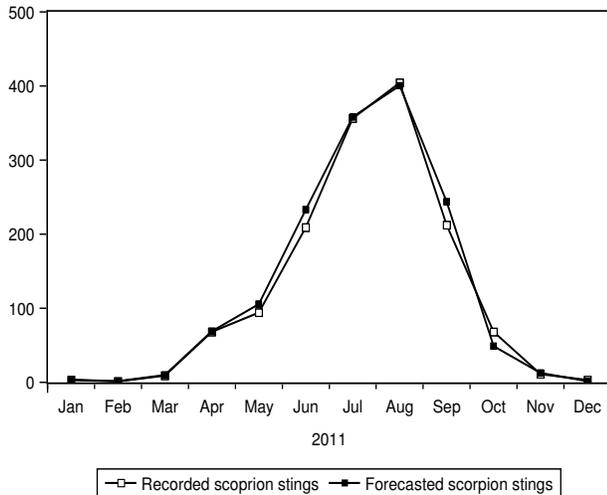


Fig. 10. Forecasted versus recorded scorpion sting cases in 2011.

support for designing of intervention strategies. However, further studies are needed to explore whether other independent variables, such as land cover index, can improve the prediction. As the epidemiology of scorpion envenomation is determined, besides scorpions, by man and environment, the modeling incorporating environmental conditions and human behavior is to be undertaken.

ACKNOWLEDGMENT

The author is deeply thankful to the referees for their valuable comments.

REFERENCES

- [1] J. P. Chippaux, M. Goyffon, *Epidemiology of scorpionism: A global appraisal*, Acta Trop, 107(2), 71-79 (2008).
- [2] G. Chowell, J. M. Hyman, P. Díaz-Duenas, N. W. Hengartner, *Predicting scorpion sting incidence in an endemic region using climatologically variables*, Int J Environ Health Res, 15(6), 425-435 (2005).
- [3] J. Fox, *Applied Regression Analysis, Linear Models, and Related Methods*, SAGE Publications, Social Science (1997).
- [4] M. Goyffon, *Le rôle de l'homme dans l'expansion territoriale de quelques espèces de scorpions*, Bulletin de la Société Zoologique de France Evolution et Zoologie, 117(1): 15-19 (1992)
- [5] M. Goyffon, J. P. Chippaux, *Les envenimations scorpioniques en Afrique*, Rev. Medicopharmaceutique 53(4), 17-21 (2009)
- [6] H. Kassiri, K. Shemshad, A. Kassiri, M. Shemshad, A. Valipor, A. Teimori *Epidemiological and Climatological Factors Influencing on Scorpion Envenoming in Baghmalek County, Iran*, Academic Journal of Entomology 6 (2): 47-54 (2013).
- [7] J. Neter, M. Kutner, W. Wasserman, *Applied Linear Statistical Models*, (1996)
- [8] *Office national de météorologie*. URL : <http://www.onm.meteo.dz>
- [9] *Office national des statistiques*. URL : <http://www.ons.dz>
- [10] *Prise en charge de l'envenimation scorpionique*, Comité national de lutte contre l'envenimation scorpionique (2009).
- [11] *Relevé épidémiologique mensuel*, Institut national de la santé publique. URL: <http://www.and.s.dz/insp/scorpionisme.html>.
- [12] S. Selmane, M. El Hadj, *sss*
- [13] S. Taj, M. Vazirian, B. Vazirianzadeh, S. Bigdeli, Z. Salehzadeh, *Effects of climatological variables on scorpion sting incidence in Ramshir area south west of Iran*, Journal of Experimental Zoology, India Vol. 15 No. 2 pp. 575-577 (2012).
- [14] M. Vachon, *Etude sur les scorpions*, 479p. Institut Pasteur d'Algérie. Alger (1952)

Mathematical Modelling of Groundwater flow in Aquifers which Contain Extraction/infiltration Cavity of Arbitrary Shape, Using the Theory of Functions of a Complex Variable

I. David, C. Ștefănescu, C. Grădinaru, I. Vlad, C. Gabor

Abstract — It is known that aquifers, i.e. groundwater reservoirs with large plane extension in relation to the depth, are in water resources exploitation of major importance for both domestic uses and industrial uses. In that regard may be mentioned groundwater balance problems at ecologic lakes, ponds, groundwater recharge pits, drainage pits, foundation pits, as well as groundwater extraction systems like wells or wells with laterals and so on. All these systems will be referred further as cavity i.e. a cavern in aquifer.

In the present paper a general mathematical representation of the plane groundwater flow in an aquifer which contain an extraction or recharge cavity of arbitrary shape is deduced using the theory of the analytical functions of a complex variable. The mathematical representations take into account asymptotic conditions determined by an pre-existing initial uniform groundwater flow which has a important influence on the flow processes by the cavity.

It will be deduced formulas which allow a rapid analysis of the groundwater balance in the modelled region taking into account the dependence of the infiltration rate or drainage rate of the cavity from the the system parameter and from the pre-existing uniform groundwater flow. The obtained mathematical representations and formulns can be applied for cavities with particular shape using conformal mapping.

Keywords — groundwater modelling, aquifer, infiltration/extraction cavity, complex variable functions,

I. David, „POLITEHNICA“ University of Timisoara, Department of Hydrotechnical Engineering, George Enescu 1/A, 300022 Timisoara, Romania and Univerity of Applied Sciences Giessen, Germany (corresponding author e-mail: ioan.david@gmx.net; ioan.david@upt.ro)

C. Ștefănescu „POLITEHNICA“ University of Timisoara (e-mail: achim.camelia@yahoo.co.uk)

C. Grădinaru, „POLITEHNICA“ University and S.C. GAUSS S.R.L. Timisoara (e-mail: cristian.gradinaru@gauss.ro)

I. Vlad, „POLITEHNICA“ University and S.C. GAUSS S.R.L. Timisoara (e-mail: ioan.vlad@gauss.ro)

C. Gabor, „POLITEHNICA“ University and S.C. GAUSS S.R.L. Timisoara (e-mail: Cristian.gabor@gauss.ro)

I. INTRODUCTORY REMARKS

The main purpose in this paper is to present a general and unified method for the study of bidimensional groundwater flow of groundwater in an aquifer, with has large plane extension in relation to the depth and contains a cavity of arbitrary shape limited by a closed contour C_0 . In the cavity i.e. inside of the contour C_0 is free water and from the cavity is extracted or infiltrated a flow rate (Q) of water. Concequently the cavity will be referred further as extraction/infiltration cavity of arbitrary form. The cavity can be referred/represent in practical view ecologic lake, pond, groundwater recharge pits, drainage pits, foundation pits, groundwater extraction well or well with laterals and so on.

It should be noted that although currently for modelling groundwater flow in aquifer can be used numerical methods based on performant software like MODFLOW or FEEFLOW [1], [2]. However the analytical methods retain its advantages because they provide the possibility to obtain closed solutions in the form of functions and formulas. That represents a substantial advantage for a rapid and efficient analysis of the investigated system processes, their dependence from the relevant geometric and physical parameters.

It will consider an pre-existing uniform groundwater flow with a rate of infiltration at large distances from cavity equal to v_0 or as an groundwater basin in which case $v_0 = 0$ which are for the practical applications the most significant initial aquifer piezometric conditions.

Analytical and computational representations derived in this paper refers to plane i.e. bidimensional groundwater flow in porous media accepting the Darcy's linear infiltration law.

Therefore solving motion can be done using analytic functions of a complex variable ($z = x + iy$) [3], [4].

In Fig. 1. is represented the groundwater flow schema in the complex plane (z) called physical plane

corresponding to the case of a cavity limited by an closed contour C_0 arbitrary shape situated in an pre-existing uniform groundwater flow, with a rate of infiltration at large distances from cavity constant and equal to v_0 , whose direction is determined by the angle α (Fig. 1).

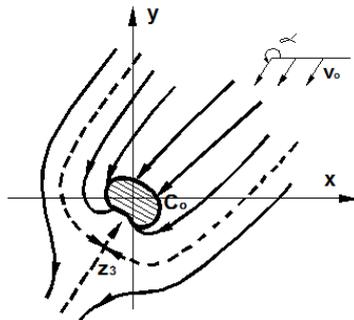


Fig.1. Groundwater flow scheme in the presence of a drain cavity in a plane parallel groundwater flow

This schematic feature plane groundwater flow (two-dimensional) is available for both groundwater layer type confined and unconfined, while the Dupuit - Forcheimer assumptions are satisfied, i.e. quasi plane groundwater flow when the vertical component of filtration velocity is negligible compared to the horizontal plane component [1], [2], [3], [4].

Study approach of groundwater flow generated by an extraction cavity C_0 of arbitrary shape by mathematical modeling was allows a general theoretical approach of ecological groundwater flow in the presence of lakes, ponds, groundwater recharge pits, drainage pits, foundation pits, groundwater extraction systems without being necessary to specify its particular shape.

The results obtained in the paper can be applied for various particular forms of cavities such as ecologic lakes, ponds, groundwater recharge pits, drainage pits, foundation pits, as well as groundwater extraction systems like wells or drains of finite length or wells with radial drainage using conformal mapping means analytic functions of a complex variable.

II. MATHEMATICAL FORMULATION AND GENERAL EQUATIONS OF GROUNDWATER FLOW

According to with the considerations of the preceding paragraph the groundwater flow takes place in an aquifer (porous media) which occupies the outer domain D of the extraction cavity limited by the closed contour " C_0 " (Figure 2). With z , ($z = x + iy$) is denoted the complex plane of the physical groundwater flow.

The groundwater flow is bidimensional in an horizontal plane (x,y) and the Darcy's law is considered valid. Following the basic equation of groundwater flow is of the form [3],[4]:

$$\vec{v} = \text{grad}\Phi(x, y) - \nabla\Phi; \quad x,y \in D^- \quad (1)$$

where:

\vec{v} – the groundwater flow velocity (flow rate);
 Φ – the potential function (velocity potential);

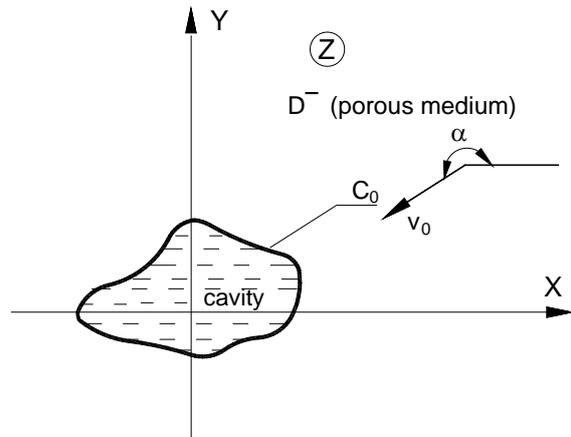


Fig.2 Scheme of the physical groundwater flow (complex plane "z")

The relationship between the potential function of the groundwater flow $\Phi(x, y)$ and the piezometric head $h(x, y)$ is given by:

$$\Phi(x, y) = -k \cdot h(x, y) + c \quad (2)$$

where:

k – the Darcy filtration coefficient
 h – piezometric head
 c – an undetermined constant.

The continuity equation in steady flow and incompressible fluid has the form:

$$\text{div}\vec{v} = \nabla \cdot \vec{v} = 0 \quad (3)$$

From the two equations is obtained the general differential equation of bidimensional groundwater flow known as:

$$\nabla \cdot \nabla\Phi = \Delta\Phi = 0 \quad (4)$$

a partial differential equation of Laplace type.

Therefore potential function $\Phi(x,y)$ (the potential of the groundwater flow velocity) is a harmonic function defined in the flow field D^- . The form and uniqueness of the solution of the equation (4) will be determined by the characteristic boundary conditions of considered flow.

Known the properties of harmonic functions for the study of the considered class of groundwater flows can be used analytical functions of a complex variable $F(z)$, attributing of its real part the physical meaning of potential function $\Phi(x, y)$ and of its imaginary part the meaning of flow function denoted $\Psi(x,y)$:

$$\begin{aligned} \operatorname{Re}\{f(z)\} &= \phi(x, y), \quad \operatorname{Im}\{f(z)\} = \psi(x, y), \\ z &\in D^- \end{aligned} \quad (5)$$

The analytical function of a complex variable $F(z)$ is called the complex potential of flow [4], [5].

The potential function $\phi(x,y)$ and the flow function $\Psi(x,y)$ checks the Laplace equation (4) with partial derivatives.

It also introduces another function of a complex variable the analytic function $W(z)$ called the complex velocity function, which is related to the $F(z)$ by:

$$W(z) = \frac{dF(z)}{dz} = v_x - iv_y, \quad z \in D^- \quad (6)$$

Where v_x și v_y are the flow velocity components.

From these considerations above described, results that for study of such groundwater flow is sufficient to know one of the functions $\phi(x, y)$, $F(z)$ or $W(z)$. Therefore to solve a given plane groundwater flow problem it is sufficient to determine one of the three functions. The uniqueness of the solution is determined by the boundary conditions associated with specific groundwater flow considered. In the following we will first determine the complex velocity $W(z)$, then following through relation (6) can be obtained the complex potential $F(z)$ and through relation (5) the potential function $\phi(x, y)$ as the real part of $F(z)$.

To determine the complex velocity $W(z)$ will be searched the solution to equation (4) that satisfied the asymptotic conditions and the characteristic boundary conditions for the groundwater flow generated by the extraction cavity limited by the contour C_0 . The extraction flow rate from the cavity is Q and the cavity is situated in an pre-existing uniform groundwater flow having a velocity of v_0 . Therefore the complex velocity $W(z)$ must satisfy the following conditions:

- $W(z)$ should be a holomorphic function in D^- , namely

$$\frac{\partial W(z)}{\partial \bar{z}} = 0, \quad z \in D^-, \quad z=x+iy, \quad \bar{z}=x-iy \quad (7)$$

- The contour C_0 is an equipotential line i.e.:

$$\operatorname{Re}_{z \in C_0} \{F(z)\} = \phi(x, y) = \text{const}; \quad (8)$$

In an equivalent formulation can be used the tangential component (v_τ) of the velocity which satisfy the following condition:

$$v_\tau = \frac{\partial \phi}{\partial \tau} = 0, \quad (8')$$

Taking into account (6) the condition (8') may be formulated directly with the complex velocity $W(z)$:

$$\operatorname{Re}\{W(z)dz\} = 0, \quad z \in C_0 \quad (8'')$$

- The draining effect i.e. extraction of a constant flow rate Q from the cavity is expressed as:

$$\int_{C_0} W(z)dz = -iQ \quad (9)$$

- The asymptotic condition i.e. the complex velocity of the groundwater flow at large distances from the cavity will be expressed as:

$$\lim_{z \rightarrow \infty} W(z) = W_\infty = V_0 e^{-i\alpha} \quad (10)$$

III. MATHEMATICAL REPRESENTATION OF GROUNDWATER FLOW USING COMPLEX VELOCITY AND COMPLEX POTENTIAL FUNCTIONS

Determination of complex flow velocity $W(z)$

To determine the complex velocity $W(z)$ as mathematical problem formulated in paragraph 2 will be considered a new complex plan (ζ) in which the exterior domain of a circle of radius ρ_0 as canonical domain Δ^- will be denoted. Further a conformal mapping will be considered which maps the external domain D^- of the physical complex plan (z) on the domain Δ^- in the plan (ζ). The conformal mapping is defined by a holomorphic function:

$$\zeta = f(z) \quad (11)$$

which is inversable having an invers denoted:

$$z = f^{-1}(\zeta) \quad (11')$$

By (11) the domain D^- exterior of cavity contour C_0 in the physical complex plane (z) is represented on the domain Δ^- exterior of circle K_0 of radius ρ_0 in the complex plane (ζ):

$$\Delta^- = \{\zeta : |\zeta| \geq \rho_0\} \quad (12)$$

Without loss of generality of the mathematical problem the transformation (11) can be considered so that the point at infinity and flow direction remain the same in both plans. Such a transformation there is always, according to the theorem of Riemann [5].

Denoting $W(\zeta)$ the complex flow velocity in the complex plane (ζ) and using the conformal mapping

properties and taking into account the asymptotic and boundary conditions formulated above for the flow in physical plan these can be transcribed for $W(\zeta)$ in the complex plane (ζ) using (11) and (11'). Thus the correspondences in the two complex plans are (Fig.2):

$$\begin{aligned} z &\leftrightarrow \zeta \\ D^- &\leftrightarrow \Delta^- \\ C_0 &\leftrightarrow K_0 \end{aligned} \quad (13)$$

The relationship between the complex flow velocity in the two complex plans, original physical complex plane (z) and her image plane (ζ) is given by the relationship:

$$W(\zeta) = W(z) \frac{dz}{d\zeta} \quad \text{si} \quad W(z) = W(\zeta) \frac{d\zeta}{dz} \quad (14)$$

On the contour K_0 occurs:

$$\lim_{\zeta \in \Delta^- \rightarrow \tau \in K_0} W(\zeta) = W(\tau) \quad (15)$$

With these clarifications conditions (7-10) for $W(\zeta)$ in the transformed plane ζ have the following forms:

$$\frac{\partial W(\zeta)}{\partial \bar{\zeta}} = 0, \quad \zeta \in \Delta^- \quad (7')$$

$$\text{Re}\{W(\zeta)d\zeta\} = 0, \quad \zeta \in K_0 \quad (8''')$$

$$\int_{K_0} W(\zeta)d\zeta = -iQ \quad (9')$$

$$\begin{aligned} W_\infty^* &= \lim_{\zeta \rightarrow \infty} W(\zeta) = \\ &= \lim_{z \rightarrow \infty} W(z) \left(\frac{dz}{d\zeta}\right)_{z \rightarrow \infty} = \\ &= W_\infty \cdot b_0 = V_0 \cdot b_0 e^{-i\alpha} \end{aligned} \quad (10')$$

$$\text{where } b_0 = \left(\frac{dz}{d\zeta}\right)_{z \rightarrow \infty}$$

The function $W(\zeta)$ will be determined by applying the Cauchy integral formula for holomorphic functions like in [5], [6], [7]:

$$W(\zeta) = \frac{1}{2\pi i} \int_K \frac{W(\tau)d\tau}{\tau - \zeta}, \quad \zeta \in \Delta^-, \quad \tau \in K \quad (16)$$

where K is a closed contour in D^- located outside the circle K_0 in the complex plane (ζ).

Through the continuum deformation of the contour K , by passing to the limit on the to the circle K_0 and taking into account the property of complex

flow velocity at the large distances from the cavity, expressed by (10') yields the following integral representation of complex flow velocity in the plane (ζ) [6], [7],[8]:

$$W_\infty^* - \frac{1}{2\pi i} \int_{K_0} \frac{W(\tau)d\tau}{\tau - \zeta} = \begin{cases} W(\zeta), & \zeta \in \Delta^-, \\ \frac{1}{2}W(\tau_0), & \zeta \in \Delta^- \rightarrow \tau_0 \in K_0 \\ 0, & \zeta \in \Delta^+ \end{cases} \quad (17)$$

Where Δ^- is the interior domain of circle K_0 .

Using these integral representations on obtine the complex potential in the physical plan of the groundwater flow generated by a extraction cavity, arranged in an initial pre-existing groundwater plan parallel flow:

$$\begin{aligned} F(z) &= -\frac{Q}{2\pi} \ln \frac{f(z)}{|f(z \in C_0)|} + \\ &+ v_0 b_0 \left[e^{-i\alpha} \cdot f(z) - e^{i\alpha} \cdot \frac{|f(z \in C_0)|^2}{f(z)} \right] + c \end{aligned} \quad (18)$$

In practical applications are useful the potential function $\Phi(x, y)$.e. the real part of the complex potential $F(z)$ and the flow function Ψ i.e. the imaginary part of the complex potential $F(z)$ expressed as:

$$\phi = \text{Re}\{F(z)\} = V_0 \cdot b_0 \cdot \left(\rho - \frac{\rho_0^2}{\rho}\right) \cdot \cos(\theta - \alpha) - \frac{Q}{2\pi} \ln \rho \quad (32)$$

$$\psi = \text{Im}\{F(z)\} = -\frac{Q}{2\pi} \cdot \theta + v_0 \cdot b_0 \cdot \left(\rho + \frac{\rho_0^2}{\rho}\right) \cdot \sin(\theta - \alpha) \quad (19)$$

where we used the notations:

$$b_0 = \left(\frac{dz}{d\zeta}\right)_{\zeta \rightarrow \infty}; \quad \rho = |f(z)|; \quad \theta = \arg f(z) \quad (20)$$

$f(z)$ is the conformal mapping analitic function according (11), (11').

IV. CHARACTERISTIC FORMS OF GROUNDWATER FLOW SYSTEM

Based on mathematical representations obtained in paragraph 3 may highlight a number of properties of parallel flow in the presence of a drained cavity without the need to customize its concrete form. To this end it will first determine the transit Q_i flow rate crossing the cavity contour C_0 , which depends obviously from the extraction flow rate Q_C of cavity and the uniform initial flow characteristics i.e. the initial velocity v_0 and its direction given by α .

For

$$0 \leq Q \leq 4\pi \cdot v_0 \cdot b_0 \cdot \rho_0 \quad (21)$$

the plane forms of the flow are presented in Figure 3.a,b,c,d.

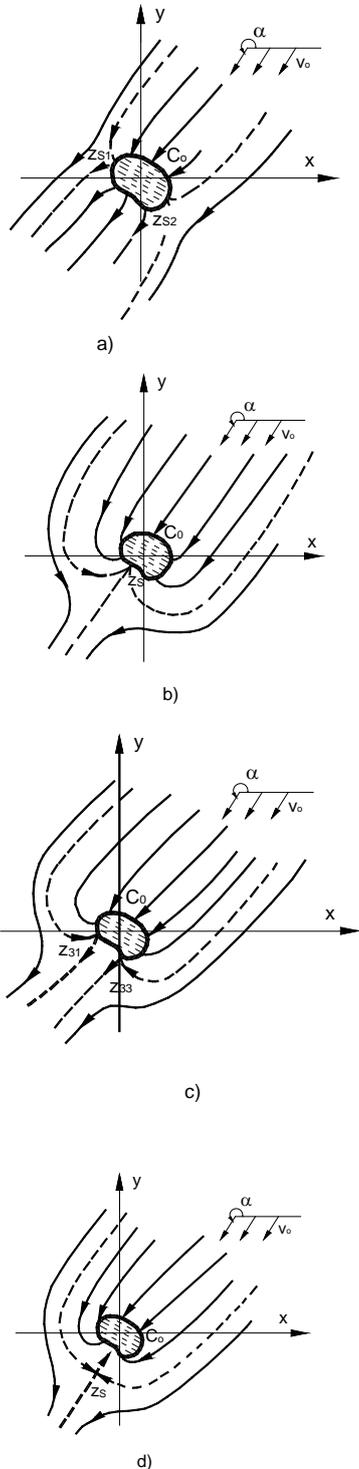


Fig. 3. Characteristic forms of flow groundwater plane in the presence of a cavity draining

The appearance of motion is of the form shown in Fig. 3.a, there are two stagnation points on the

contour C0. In this case the flow rate which enter in the cavity is equal to that which pass the cavity and that which leaves the cavity (Qt).

The other limiting case when the extracted flow rate is captured (drained) has the value:

$$Q = 4\pi v_0 \cdot b_0 \cdot \rho_0 \quad (22)$$

If exist a single point of stagnation (ie two stagnation points coincide in Zs) the flow aspect is sketched in Figure 3.b.

Current case when Q is strictly within the limits:

$$0 < Q < 4 \cdot \pi \cdot v_0 \cdot b_0 \cdot \rho_0 \quad (23)$$

In tis case the physical aspect of the flow is sketched in Figure 3.c.

The case when the flow exceeds the maximum extracted flow rate i.e.:

$$Q \geq 4\pi \cdot v_0 \cdot b_0 \cdot \rho_0 \quad (24)$$

If dont exist stagnation points on the contour C the groundwater flow have the form sketched in Figure 3.d

Note:

If Q has the opposite sign (-Q) thee cavity is an infiltration basin through which the aquifer is supplied from a lake/pool limited by contour C0.

If Q = 0 that is extracted from the cavity flow is zero, representations obtained it refers to a plane parallel flow in the presence of a simple cavity i.e lake.

V. DETERMINATION OF EXTRACTION FLOW RATE FROM THE CAVITY

In practical applications, including the above presented flow forms, an important role has the extraction flow rate Qc from the cavity which is closely related to hydraulic and geometric parameters of the cavity.

Using the connection and the dependence relationship is given by the equation (2) which connects the potential function $\phi(x,y)$ to piezometric head $h(x,y)$ can be expressed a direct link between piezometric head $h(x,y)$, geometrical parameters of the cavity and parameters and potential function of the cavity:

$$\begin{aligned} -kh(x, y) + c &= \\ &= v_0 \cdot b_0 \cdot \left(\rho - \frac{\rho_0^2}{\rho} \right) \cdot \cos(\theta - \alpha) - \frac{Q_c}{2\pi m} \ln \rho \end{aligned} \quad (25)$$

where the following notations are used:

$$b_0 = \left(\frac{dz}{d\zeta}\right)_{z,\zeta \rightarrow \infty}; \rho_0 = |f(z \in C_0)|,$$

$$\rho = |f(z \in D^-)|, \quad (26)$$

$$\theta = \arg f(z), z=x+iy$$

It should be noted that the expression (25) Q_C is the total flow rate of the cavity calculated for the entire depth (m) of the aquifer. The relation between Q with Q_C/m is evident: $Q=Q_C/m$.

To determine the total flow rate of the cavity Q_C will be used the relation (25) and the hydraulic boundary conditions of the considered groundwater flow:

- $h = H_R$ at large distances of drained (extraction) cavity denoted by R_0 (influence radius tion)
- $h = H_0$ on the boundary of the cavity

Also take into account that the pre-existing groundwater parallel flow velocity (v_0) which can be expressed using hydraulic gradient (I_0) and the known Darcy's law:

$$v_0=k \cdot I_0 \quad (27)$$

With these clarifications from (25) can be obtained the total extraction flow rate of the drainage cavity located in an initial groundwater flow in the form:

$$Q_C = \frac{2 \cdot \pi \cdot m \cdot [k \cdot (H_R - H_0) - I_0 \cdot R_0 \cdot U]}{\ln \frac{\rho_\alpha}{\rho_0}} \quad (28)$$

or in a dimensionless form:

$$\frac{Q}{k \cdot m \cdot \Delta H} = \frac{2\pi \cdot \left(1 - \frac{I_0 \cdot R_0}{\Delta H} \cdot U\right)}{\ln \frac{\rho_\alpha}{\rho_0}} \quad (29)$$

In (28) and (29) the following notations have been used:

$$\rho_\alpha = \left|f(z_\alpha = R_0 \cdot e^{i(\alpha-\pi)})\right|; U = \frac{b_0}{R_0} \cdot \left(\rho_\alpha - \frac{\rho_0^2}{\rho_\alpha}\right)$$

where

$$\Delta h = H_R - H_0 \quad (30)$$

These formulas allow the calculation of the extracted flow rate from cavity of arbitrary form limited with a closed contour C_0 when the cavity is located in a groundwater layer with given hydraulic parameters: coefficient of filtration k , thickness of the layer (m), slope I_0 of the initial groundwater flow and the upstream influence radius R_0 of drainage cavity

For a given particularity form of the cavity an analytical conformal mapping function will be

determined according to (11). This function determine the geometric parameters that appear in the flow rate formula (28) or (29).

VI. CONCLUSIONS

In the present paper a general mathematical representations and calculus formulas for plane groundwater flow in an aquifer which containe an extraction cavity of arbitrary shape are deduced using the theory of analytical functions of a complex variable. These representations and formulas refer to the complex potential of the flow, to the potential function and flow line function of the flow as well as to the extracted flow rate formula from the cavity . On the basis of these general representations characteristically flow pattern of the groundwater flow system, consisting of a cavity arbitrary shape situated in an pre-existing parallel groundwater flow will be analysed. The results have a general validity independent from cavity shape. For a particular shape the corresponding conformal mapping function must be determined which maps the external domain of the cavity on the external domain of the canonical circle.

The extraction cavity can represent several practical shapes which currently are used in technical applications. So the in the paper obtained results can be used in engineering planning and management of a large number of groundwater flow problems like groundwater balance in ecologic lakes, ponds, groundwater recharge pits, drainage pits, foundation pits and so on. which can be modelled as extraction cavities or recharge cavities. Such concrete examples will be presented in a next paper.

REFERENCES

- [1] A. W. Harbaugh, "MODFLOW, The U.S. Geological Survey Modular Ground-Water Model—the Ground-Water Flow Process", Chapter 16 of Book 6. Modeling techniques, Section A. Ground Water, 2005
- [2] H. J. Diersch, "FEFLOW Finite Element Subsurface Flow and Transport Simulation System, Reference Manual". WASY GmbH, Berlin, 2005
- [3] J. Bear, " Dynamics of fluids in porous media". New York: American Elsevier Publishing Company, 1972
- [4] I.David "Grundwasserhydraulik. Strömung- und Transportvorgänge". Vieweg Verlag, 2005
- [5] A. I. Markushevich, "Theory of Functions of a Complex Variable, Second Edition, AMS CHELLSEA PUBLISHING, American Mathematical , Revised edition, 2005
- [6] I. David, Sur la généralisation d' un théorème du mouvement plan en milieux poreux ayant une cavité, *Comptes Rendus Acad. Sc. Paris*, t. 269, 809 – 811, 29 oct. 1969;
- [7] I. David, Sur une extension du theoreme du cercle a la solution d'un probleme de Riemann non-homogene aux singularites donnees, *Buetine Scientifique IPT, Seria Math-Mec*, Tom 18 (32), fasc.2/1973
- [8] I. David, Analytical and Boundary Elements based Integral Representation for Numerical Solution of 3-D Potential Problems in Heterogeneous Media Containing Singularities. *Proceedings of the 12th WSEAS International Conference (MACMESE '10)*, University of Algarve, Faro, Portugal, November 3-5, 2010, pg.350-357

One Approach to the Design of TS Fuzzy Fault Detection Filters

Dušan Krokavec, Anna Filasová and Vratislav Hladký

Abstract—One principle for designing robust Takagi-Sugeno fuzzy fault detection filter, dedicated to a class of continuous-time nonlinear MIMO system, is treated in this paper. The problem addressed can be designated as an approach exploiting fuzzy reference model to reflect the problem as an H_∞ optimization task, guaranteeing the fault detection performance and the state observer stability. The conditions are outlined in the terms of linear matrix inequalities to possess a stable structure closest to optimal asymptotic properties. Simulation results illustrate the design procedures and demonstrate the performance of the proposed detection scheme.

Index Terms—Robust fault detection filters, TS fuzzy observers, H_∞ performance, convex optimization, linear matrix inequalities.

I. INTRODUCTION

The fault detection filters, usually relying on the use of particular type of state observers, are mostly used to produce the fault residuals in fault tolerant control systems. Because it is generally not possible in residuals to decouple totally fault effects from the perturbation influence, the H_∞ approach is used to tackle in part this conflict and to create residuals that are as a rule zero in the fault free case, maximally sensitive to faults, as well as robust to disturbances [3], [5]. Since faults are detected usually by setting a threshold on the generated residual signal, determination of an actual threshold is often formulated in adaptive frames [4], [7].

The nonlinear system theory principles, exploiting known Lipchitz conditions, has emerged as a method capable of use in the state estimation based residual generator design for nonlinear systems [11], [20], [23], although Lipschitz conditions may be strong restrictive in many cases. Possible alternative to this conception is the Takagi–Sugeno (TS) fuzzy modeling approach [18] which benefits from the advantages of local system dynamics approximation techniques, where the nonlinear system is represented as a collection of fuzzy rules, where each rule describes the local system dynamics by a linear system model. The main advantage of the TS type fuzzy models, in addition to linear sub-models, is their description permitting utilization of the nonlinear system state space representation. In light of the above, TS fuzzy observer projection as well as robust TS fuzzy fault detection filters (FDF) design have attracted interest in fault detection praxis (see, e.g., [9] an the reference therein).

The work presented in this paper was supported by VEGA, the Grant Agency of the Ministry of Education and the Academy of Science of Slovak Republic, under Grant No. 1/0348/14. This support is very gratefully acknowledged.

Authors are with the Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9/B, 042 00 Košice, Slovakia, e-mail: dusan.krokavec@tuke.sk, anna.filasova@tuke.sk, vratislav.hladky@tuke.sk

Different TS fuzzy state observers [12], [15], [19] as well as FDF structures based on the TS fuzzy system model were designed [2], [21], [22], usually using technique based on linear matrix inequalities (LMIs). To achieve robustness, TS fuzzy observers can be combined with classical sliding mode technique, where the fault reconstruction is also attained. The new trends are based upon the idea of reducing the robust fault detection problem to a standard H_∞ model-matching problem with main goal to discriminate the effect between the fault and the unknown disturbances in the residual signals.

The main contribution of the paper is to present an extension principle for designing robust TS fuzzy FDFs for continuous-time nonlinear multi-input multi-output (MIMO) systems approximated by the TS model. Following the basic idea and concerning with the residual reference models [1], [6] applied originally in FDFs design for systems with time-delays, the desired specification of the TS fuzzy reference residual model in the form of a cross-bonds matrix for different particular channels is introduced in the paper. Using the reference residual model, the robust TS fuzzy FDF design problem is formulated as an H_∞ optimization task in the presence of bounded disturbances. Such reference residual model is used to reduce the design problem to a standard formulation [12], considering both the robustness against disturbances and the sensitivity to faults, as well as guaranteeing the fault detection performance and the observer stability.

The remainder of this paper is organized as follows. In Sec. II the TS fuzzy model for given class of nonlinear continuous-time systems is briefly described and some basic properties of TS fuzzy models are presented. The TS fuzzy reference model design problem is formulated in Sec. III and, subsequently, the robust TS fault detection filter design conditions are presented in Sec. IV. In Sec. V, a numerical example is given to confirm the validity of the proposed fault detection scheme and, finally, Sec. VI draws some conclusions and remarks.

Throughout the paper, the following notations are used: \mathbf{x}^T , \mathbf{X}^T denotes the transpose of a vector \mathbf{x} and a matrix \mathbf{X} , respectively, for a square matrix $\mathbf{X} > 0$ (respectively $\mathbf{X} < 0$) means that \mathbf{X} is a symmetric positive definite matrix (respectively, negative definite matrix), $\|\mathbf{X}\|_\infty$ designs the H_∞ norm of the matrix \mathbf{X} , the symbol \mathbf{I}_n represents the n -th order unit matrix, \mathbb{R} denotes the set of real numbers, \mathbb{R}^n , $\mathbb{R}^{n \times r}$ refers to the set of all n -dimensional real vectors and $n \times r$ dimensional real matrices, respectively and $L_2(0, +\infty)$ is the space of square-integrable vector functions over $(0, +\infty)$.

II. TS FUZZY FAULT DETECTION FILTER

The systems under consideration enter into a class of MIMO nonlinear dynamic continuous-time systems, described by using the TS approach as follows

$$\dot{\mathbf{q}}(t) = \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) (\mathbf{A}_i \mathbf{q}(t) + \mathbf{B}_i \mathbf{u}(t) + \mathbf{B}_{f_i} \mathbf{f}(t) + \mathbf{B}_{d_i} \mathbf{d}(t)) \quad (1)$$

$$\mathbf{y}(t) = \mathbf{C} \mathbf{q}(t) \quad (2)$$

where $\mathbf{q}(t) \in \mathbb{R}^n$, $\mathbf{u}(t) \in \mathbb{R}^r$, $\mathbf{y}(t) \in \mathbb{R}^m$ are vectors of the state, input, and output variables, respectively, $\mathbf{C} \in \mathbb{R}^{m \times n}$, $\mathbf{A}_i \in \mathbb{R}^{n \times n}$, $\mathbf{B}_i \in \mathbb{R}^{n \times r}$, $\mathbf{B}_{f_i} \in \mathbb{R}^{n \times r_f}$, $\mathbf{B}_{d_i} \in \mathbb{R}^{n \times r_d}$ for $i = 1, 2, \dots, s$ are known real matrices and $\mathbf{d}(t) \in \mathbb{R}^{r_d}$ is the unknown disturbance that belongs to $L_2(0, +\infty)$. It is considered that a fault $\mathbf{f}(t)$ may occur at an uncertain time instant, the size of the fault is unknown but bounded and the elements of the fault vector $\mathbf{f}(t) \in \mathbb{R}^{r_f}$ are zero to the time until a fault is occurred. Since the problem of interest is to design a robust, observer based fault detection filter using TS fuzzy model of the nonlinear system, it is considered that all matrix pairs $(\mathbf{A}_i, \mathbf{C})$, $i = 1, 2, \dots, s$, are observable.

The membership function $h_i(\boldsymbol{\theta}(t))$ is the averaging weight of the i -th fuzzy rule and satisfies the following conditions

$$0 \leq h_i(\boldsymbol{\theta}(t)) \leq 1, \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) = 1 \text{ for all } i \in \langle 1, \dots, s \rangle \quad (3)$$

Assuming that each nonlinear term in the nonlinear system description is bounded in an associated sector within it operates, as well as that the number of nonlinear terms is p and the number of sector functions varies from 2 to k , then the number of TS linear sub-models s falls within the range of values $\langle 2^k, p^k \rangle$. Note, $h_i(\boldsymbol{\theta}(t))$ for $i = 1, 2, \dots, s$ are calculated from all combinations of sector functions.

The vector $\boldsymbol{\theta}(t) \in \mathbb{R}^o$ of the structure

$$\boldsymbol{\theta}(t) = [\theta_1(t) \quad \theta_2(t) \quad \dots \quad \theta_o(t)] \quad (4)$$

is the vector of premise variables. A premise variable generally represents any measurable or observable variable occurring in the associate sector nonlinear term. Most often the premise variable is a state variable or a function of the system state variables. It is supposed in the following that all premise variables are measured and none of them is a function of the input variables defined in $\mathbf{u}(t)$.

Using the TS fuzzy model, the conclusion part of a fuzzy rule consists no longer of the fuzzy set [16], but determines a function requiring as arguments the system state variables, while this function is local for the fuzzy region that is described by the premise part of a rule. As a result, the nonlinear system is so described locally (in fuzzy regions) by linear models, and at the region boundaries an interpolation is used between the corresponding local models. More details can be found, e.g., in [14], [19].

For the purpose of residual generation, fault detection filters (FDF) are proposed in the standard structure adhering

that the TS fuzzy observer

$$\dot{\mathbf{q}}_e(t) = \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) (\mathbf{A}_i \mathbf{q}_e(t) + \mathbf{B}_i \mathbf{u}(t) + \mathbf{J}_i (\mathbf{y}(t) - \mathbf{y}_e(t))) \quad (5)$$

$$\mathbf{y}_e(t) = \mathbf{C} \mathbf{q}_e(t) \quad (6)$$

is designed together with the residual generator

$$\mathbf{r}(t) = \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) \mathbf{V}_i \mathbf{C} (\mathbf{y}(t) - \mathbf{y}_e(t)) \quad (7)$$

where $\mathbf{q}_e(t) \in \mathbb{R}^n$ is estimation of the system state vector, $\mathbf{y}_e(t) \in \mathbb{R}^m$ is the observed output vector, $\mathbf{r}(t) \in \mathbb{R}^{m_r}$ is the residual signal and $\mathbf{J}_i \in \mathbb{R}^{n \times m}$, $\mathbf{V}_i \in \mathbb{R}^{m_r \times m}$, $i = 1, 2, \dots, s$, are the set of the observer gain matrices and residual signal weighting matrices, respectively.

Introducing the observer state error

$$\mathbf{e}(t) = \mathbf{q}(t) - \mathbf{q}_e(t) \quad (8)$$

and taking the time derivative of $\mathbf{e}(t)$, the dynamics of FDF can be expressed as

$$\dot{\mathbf{e}}(t) = \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) ((\mathbf{A}_i - \mathbf{J}_i \mathbf{C}) \mathbf{e}(t) + \mathbf{B}_{f_i} \mathbf{f}(t) + \mathbf{B}_{d_i} \mathbf{d}(t)) \quad (9)$$

$$\mathbf{r}(t) = \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) \mathbf{V}_i \mathbf{C} \mathbf{e}(t) = \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) \mathbf{H}_i \mathbf{e}(t) \quad (10)$$

where

$$\mathbf{H}_i = \mathbf{V}_i \mathbf{C} \quad (11)$$

$\mathbf{H}_i \in \mathbb{R}^{m_r \times n}$, $\mathbf{V}_i \in \mathbb{R}^{m_r \times m}$.

In general, the goal is to design the FDF matrix parameter sets $\mathbf{J}_i \in \mathbb{R}^{n \times m}$, $\mathbf{V}_i \in \mathbb{R}^{m_r \times m}$, $i = 1, 2, \dots, s$, in such a way that

$$\|\mathbf{r}(t)\|_{\infty}^2 \leq \gamma^o \|\mathbf{d}(t)\|_{\infty}^2 \quad (12)$$

and the square of the H_{∞} norm $\gamma^o \in \mathbb{R}$ is as small as possible.

Formulating more detailed, to boost the accuracy of fault detection, it would be wantable to craft residuals with sensitivity to faults under robustness to disturbances. One of the options is the use of H_{∞}/H_{∞} optimization principle in residual generator design [10], but the restriction of this method is mainly the necessity of existence of a full rank direct-feedthrough matrix from a fault to the residual.

III. REFERENCE MODEL

The reference model in the proposed structure provides a pattern that partly separates interactions, represented by cross-bonds in $\mathbf{d}(t)$ and $\mathbf{f}(t)$, from the observer data model. Thus, by formalizing a reference model in an appropriate mathematical framework, minimally this elementary property is ensured for residual generation and the other related functions are keep together in defined layer of complexity and extensibility. Note, only the reference model structure is based on certain simplifications.

Lemma 1: The cross-bonds in $\mathbf{d}(t)$ and $\mathbf{f}(t)$ is minimized in the reference model

$$\dot{e}^\circ(t) = \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) \left((\mathbf{A}_i - \mathbf{J}_i^\circ \mathbf{C}) e^\circ(t) + \mathbf{G}_i^\circ \mathbf{g}^\circ(t) \right) \quad (13)$$

$$\mathbf{r}^\circ(t) = \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) \mathbf{V}_i^\circ \mathbf{C} e^\circ(t) = \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) \mathbf{H}_i^\circ e^\circ(t) \quad (14)$$

where

$$\mathbf{N}^{\circ T} = [\mathbf{I}_{r_g} \quad \mathbf{I}_{r_d}] \quad (15)$$

$$\mathbf{G}_i^\circ = [\mathbf{B}_{di} \quad -\mathbf{B}_{fi}] \mathbf{N}^\circ, \quad \mathbf{g}^\circ(t) = \mathbf{d}(t) - \mathbf{f}(t) \quad (16)$$

if for $\mathbf{G}_i^\circ \in \mathbb{R}^{n \times r_g}$, $\mathbf{g}^\circ(t) \in \mathbb{R}^{r_g}$ and

$$\int_0^t \mathbf{r}^{\circ T}(x) \mathbf{r}^\circ(x) dx < \gamma^\circ \int_0^t \mathbf{g}^{\circ T}(x) \mathbf{g}^\circ(x) dx \quad (17)$$

the square of the H_∞ norm $\gamma^\circ \in \mathbb{R}$ is as small as possible.

Proof: It is regarded in this proof, for the sake of simplicity, that $r_f = r_d = r_g$. If now (9) is written as

$$\dot{e}(t) = \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) \left\{ (\mathbf{A}_i - \mathbf{J}_i \mathbf{C}) e(t) + [\mathbf{B}_{di} - \mathbf{B}_{fi}] \begin{bmatrix} \mathbf{d}(t) \\ -\mathbf{f}(t) \end{bmatrix} \right\} \quad (18)$$

and the same cross-bonds between $\mathbf{d}(t)$ and $\mathbf{f}(t)$ are considered, then it can be introduced an reference model structure of the form

$$\dot{e}^\circ(t) = \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) \left\{ (\mathbf{A}_i - \mathbf{J}_i^\circ \mathbf{C}) e^\circ(t) + [\mathbf{B}_{di} - \mathbf{B}_{fi}] \mathbf{T}^\circ \begin{bmatrix} \mathbf{d}(t) \\ -\mathbf{f}(t) \end{bmatrix} \right\} \quad (19)$$

where, by using the structure of $\mathbf{N}^{\circ T}$ defined in (15), the cross-bonds matrix \mathbf{T}° was selected as

$$\mathbf{T}^\circ = \begin{bmatrix} \mathbf{I}_{r_g} & \mathbf{I}_{r_g} \\ \mathbf{I}_{r_g} & \mathbf{I}_{r_g} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{r_g} \\ \mathbf{I}_{r_g} \end{bmatrix} [\mathbf{I}_{r_g} \quad \mathbf{I}_{r_g}] = \mathbf{N}^\circ \mathbf{N}^{\circ T} \quad (20)$$

Applying (16), the reference model takes the form (13), (14). This concludes the proof. ■

Note that minimizing $\mathbf{g}^\circ(t)$ means maximizing the influence of $\mathbf{f}(t)$ and minimizing the influence of $\mathbf{d}(t)$ on the residual signal in the proposed reference model structure.

The following design conditions are now proven for design of the sets of reference model parameters.

Theorem 1: The reference model (13), (14) is asymptotically stable with the quadratic performance γ° if there exist a symmetric positive definite matrix $\mathbf{P}^\circ \in \mathbb{R}^{n \times n}$, matrices $\mathbf{Y}_i^\circ \in \mathbb{R}^{n \times m}$, $\mathbf{V}_i^\circ \in \mathbb{R}^{m_r \times n}$, $i = 1, 2, \dots, s$ and a positive scalar $\gamma^\circ \in \mathbb{R}$ such that for all i

$$\mathbf{P}^\circ = \mathbf{P}^{\circ T} > 0, \quad \gamma^\circ > 0 \quad (21)$$

$$\begin{bmatrix} \mathbf{P}^\circ \mathbf{A}_i + \mathbf{A}_i^T \mathbf{P}^\circ - \mathbf{Y}_i^\circ \mathbf{C} - \mathbf{C}^T \mathbf{Y}_i^{\circ T} & * & * \\ \mathbf{G}_i^{\circ T} \mathbf{P}^\circ & -\gamma^\circ \mathbf{I}_{r_g} & * \\ \mathbf{V}_i^\circ \mathbf{C} & \mathbf{0} & -\mathbf{I}_{m_r} \end{bmatrix} < 0 \quad (22)$$

When the above conditions hold, the reference model gain matrices are given as

$$\mathbf{J}_i^\circ = (\mathbf{P}^\circ)^{-1} \mathbf{Y}_i^\circ, \quad \mathbf{H}_i^\circ = \mathbf{V}_i^\circ \mathbf{C} \text{ for all } i \quad (23)$$

Hereafter, * denotes the symmetric item in a symmetric matrix.

Proof: Defining the Lyapunov function candidate of the form

$$v(e^\circ(t)) = e^{\circ T}(t) \mathbf{P}^\circ e^\circ(t) + (\gamma^\circ)^{-1} \int_0^t (\mathbf{r}^{\circ T}(x) \mathbf{r}^\circ(x) - \gamma^{\circ 2} \mathbf{g}^{\circ T}(x) \mathbf{g}^\circ(x)) dx \quad (24)$$

then, after evaluation the derivative of (24), it is obtained

$$\dot{v}(e^\circ(t)) = \dot{e}^{\circ T}(t) \mathbf{P}^\circ e^\circ(t) + e^{\circ T}(t) \mathbf{P}^\circ \dot{e}^\circ(t) + (\gamma^\circ)^{-1} \mathbf{r}^{\circ T}(t) \mathbf{r}^\circ(t) - \gamma^\circ \mathbf{g}^{\circ T}(t) \mathbf{g}^\circ(t) < 0 \quad (25)$$

Such formulation of the stability criterion means that the double summation through membership function occurs in the calculation of the product $\mathbf{r}^{\circ T}(t) \mathbf{r}^\circ(t)$ in (25). Since $\sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) = 1$, then the following approximation can be applied (the proof see, e.g., in [13])

$$\begin{aligned} \mathbf{r}^{\circ T}(t) \mathbf{r}^\circ(t) &= \\ &= e^{\circ T}(t) \sum_{i=1}^s \sum_{j=1}^s h_i(\boldsymbol{\theta}(t)) h_j(\boldsymbol{\theta}(t)) \mathbf{H}_i^{\circ T} \mathbf{H}_j^\circ e^\circ(t) \leq \\ &\leq e^{\circ T}(t) \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) \mathbf{H}_i^{\circ T} \mathbf{H}_i^\circ e^\circ(t) \end{aligned} \quad (26)$$

Therefore, the substitutions of (13) and (26) in (25) gives

$$\begin{aligned} \dot{v}(e^\circ(t)) &\leq -\gamma^\circ \mathbf{g}^{\circ T}(t) \mathbf{g}^\circ(t) + \\ &+ \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) e^{\circ T}(t) (\mathbf{A}_{ei}^{\circ T} \mathbf{P}^\circ + \mathbf{P}^\circ \mathbf{A}_{ei}^\circ + \gamma^{\circ -1} \mathbf{H}_i^{\circ T} \mathbf{H}_i^\circ) e^\circ(t) + \\ &+ \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) e^{\circ T}(t) (\mathbf{P}^\circ \mathbf{G}_i^\circ \mathbf{g}^\circ(t) + \mathbf{g}^{\circ T}(t) \mathbf{G}_i^{\circ T} \mathbf{P}^\circ) e^\circ(t) < 0 \end{aligned} \quad (27)$$

where

$$\mathbf{A}_{ei}^\circ = \mathbf{A}_i^\circ - \mathbf{J}_i^\circ \mathbf{C} \quad (28)$$

Defining the composite vector

$$e_c^{\circ T}(t) = [e^{\circ T}(t) \quad \mathbf{g}^{\circ T}(t)] \quad (29)$$

(26) can be rewritten as

$$\dot{v}(e^\circ(t)) = \sum_{i=1}^s h_i(\boldsymbol{\theta}(t)) e_c^{\circ T}(t) \mathbf{P}_{ci}^\circ e_c^\circ(t) < 0 \quad (30)$$

where, for all i ,

$$\mathbf{P}_{ci}^\circ = \begin{bmatrix} \mathbf{P}^\circ \mathbf{A}_{ci} + \mathbf{A}_{ci}^T \mathbf{P}^\circ + (\gamma^\circ)^{-1} \mathbf{H}_i^{\circ T} \mathbf{H}_i^\circ & * \\ \mathbf{G}_i^{\circ T} \mathbf{P}^\circ & -\gamma^\circ \mathbf{I}_{r_g} \end{bmatrix} < 0 \quad (31)$$

Using the Schur complement property the inequality (31) can be rewritten as

$$\mathbf{P}_{ci}^\circ = \begin{bmatrix} \mathbf{P}^\circ \mathbf{A}_{ci} + \mathbf{A}_{ci}^T \mathbf{P}^\circ & * & * \\ \mathbf{G}_i^{\circ T} \mathbf{P}^\circ & -\gamma^\circ \mathbf{I}_{r_g} & * \\ \mathbf{H}_i^\circ & \mathbf{0} & -\gamma^\circ \mathbf{I}_{m_r} \end{bmatrix} < 0 \quad (32)$$

and since (28) implies

$$\mathbf{P}^\circ \mathbf{A}_{ei}^\circ = \mathbf{P}^\circ \mathbf{A}_i^\circ - \mathbf{P}^\circ \mathbf{J}_i^\circ \mathbf{C} \quad (33)$$

with the notations (11) and

$$\mathbf{Y}_i^\circ = \mathbf{P}^\circ \mathbf{J}_i^\circ \quad (34)$$

then (32) implies (22). This concludes the proof. ■

IV. FAULT DETECTION FILTER DESIGN

Since residuals generated by (13), (14) are, in general, not totally decoupled from the unknown input $d(t)$, using the parameters (23), the model (19), (20) can be interpreted as a reference, defining "minimal" virtual threshold for the residual signal evaluation. The paper proposes a way to include (23) into FDF design as a hidden threshold and not only as a difference in the residual signal.

To design the generated residual $r(t)$ associated with the reference model residual, the overall FDF model, incorporating the state estimate error (9), (10) as well as the reference model (13), (14) can be expressed as

$$\dot{e}^\bullet(t) = \sum_{i=1}^s h_i(\theta(t)) ((A_i^\bullet - J_i^\bullet C^\bullet) e^\bullet(t) + G_i^\bullet g^\bullet(t)) \quad (35)$$

$$r^\bullet(t) = \sum_{i=1}^s h_i(\theta(t)) (V_i^\bullet - V_i^*) C^\bullet e^\bullet(t) \quad (36)$$

where

$$e^\bullet(t) = \begin{bmatrix} e(t) \\ e^\circ(t) \end{bmatrix}, \quad g^\bullet(t) = \begin{bmatrix} f(t) \\ d(t) \end{bmatrix}, \quad G_i^\bullet = \begin{bmatrix} B_{fi} & B_{di} \\ B_{fi} & B_{di} \end{bmatrix} \quad (37)$$

$$A_i^\bullet = \text{diag} [A_i \quad A_i], \quad C^\bullet = \text{diag} [C \quad J^\circ C] \quad (38)$$

$$J_i^\bullet = \text{diag} [J_i \quad I_n], \quad V_i^\bullet = [V_i \quad 0], \quad V_i^* = [0 \quad V_i^\circ] \quad (39)$$

$$H_i^\bullet = (V_i^\bullet - V_i^*) C^\bullet, \quad C^{\bullet T} = [C^T \quad C^T] \quad (40)$$

$e^\bullet(t) \in \mathbb{R}^{2n}$, $g^\bullet(t) \in \mathbb{R}^{r_f+r_d}$, $G_i^\bullet \in \mathbb{R}^{2n \times (r_f+r_d)}$, $A_i^\bullet \in \mathbb{R}^{2n \times 2n}$, $C^\bullet \in \mathbb{R}^{(m+n) \times 2n}$, $J_i^\bullet \in \mathbb{R}^{2n \times (m+n)}$, $V_i^\bullet, V_i^* \in \mathbb{R}^{m_r \times m}$, $H_i^\bullet, C^* \in \mathbb{R}^{m_r \times 2n}$.

Note, these matrix structures were defined in order to use structured LMI variables in the proposed design conditions. These are formulated in the sense of existence of a robust FDF of the type (5)-(7) which achieves the asymptotic stability as well as the H_∞ performance simultaneously.

Theorem 2: The residual filter (5)-(7), associated with the reference model (13), (14) is stable with the quadratic performance γ^\bullet if there exist symmetric positive definite matrices $P_1^\bullet, P_2^\bullet \in \mathbb{R}^{n \times n}$, matrices $Y_{1i}^\bullet \in \mathbb{R}^{n \times m}$, $V_{1i}^\bullet \in \mathbb{R}^{m_r \times m}$, $i = 1, 2, \dots, s$ and a positive scalar $\gamma^\bullet \in \mathbb{R}$ such that for all i

$$P_1^\bullet = P_1^{\bullet T} > 0, \quad P_2^\bullet = P_2^{\bullet T} > 0, \quad \gamma^\bullet > 0 \quad (41)$$

$$\begin{bmatrix} P^\bullet A_i^\bullet + A_i^{\bullet T} P^\bullet - Y_i^\bullet C^\bullet - C^{\bullet T} Y_i^{\bullet T} & * & * \\ G_i^{\bullet T} P^\bullet & -\gamma^\bullet I_{2r_g} & * \\ V_i^\bullet C^\bullet - V_i^* C^* & 0 & -\gamma^\bullet I_{m_r} \end{bmatrix} < 0 \quad (42)$$

where $P^\bullet \in \mathbb{R}^{2n \times 2n}$, $Y_i^\bullet \in \mathbb{R}^{2n \times (m+n)}$ are structured matrix variables of the form

$$P^\bullet = \text{diag} [P_1^\bullet \quad P_2^\bullet], \quad Y_i^\bullet = \text{diag} [Y_{1i}^\bullet \quad P_2^\bullet] \quad (43)$$

$$V_i^\bullet = [V_i \quad 0] \quad (44)$$

When the above conditions hold, then

$$J_i = P_1^{\bullet -1} Y_{1i}^\bullet, \quad V_i = V_i^\bullet [I_m \quad 0]^T \quad \text{for all } i \quad (45)$$

Proof: Since the Lyapunov function candidate can be defined as

$$v(e^\bullet(t)) = e^{\bullet T}(t) P^\bullet e^\bullet(t) + (\gamma^\bullet)^{-1} \int_0^t (r^{\bullet T}(x) r^\bullet(x) - \gamma^{\bullet 2} g^{\bullet T}(x) g^\bullet(x)) dx \quad (46)$$

its time derivative is

$$\dot{v}(e^\bullet(t)) = \dot{e}^{\bullet T}(t) P^\bullet e^\bullet(t) + e^{\bullet T}(t) P^\bullet \dot{e}^\bullet(t) + (\gamma^\bullet)^{-1} r^{\bullet T}(t) r^\bullet(t) - \gamma^{\bullet 2} g^{\bullet T}(t) g^\bullet(t) < 0 \quad (47)$$

where the structure of A_i^\bullet , C^\bullet implies the structure of P^\bullet . Since (27), (28) and (48), (49) have the same structure, where

$$A_{ei}^\bullet = A_i^\bullet - J_i^\bullet C^\bullet \quad (48)$$

and from the analogy with (26) it yields

$$r^{\bullet T}(t) r^\bullet(t) \leq e^{\bullet T}(t) \sum_{i=1}^s h_i(\theta(t)) H_i^{\bullet T} H_i^\bullet e^\bullet(t) \quad (49)$$

then, evidently,

$$\dot{v}(e^\bullet(t)) \leq \sum_{i=1}^s h_i(\theta(t)) e_c^{\bullet T}(t) P_{ci}^\bullet e_c^\bullet(t) < 0 \quad (50)$$

if

$$P_{ci}^\bullet = \begin{bmatrix} P^\bullet A_{ci}^\bullet + A_{ci}^{\bullet T} P^\bullet + (\gamma^\bullet)^{-1} H_i^{\bullet T} H_i^\bullet & * \\ G_i^{\bullet T} P^\bullet & -\gamma^\bullet I_{2r_g} \end{bmatrix} < 0 \quad (51)$$

Rewriting (51) using the Schur complement property as follows

$$\begin{bmatrix} P^\bullet A_{ci}^\bullet + A_{ci}^{\bullet T} P^\bullet & * & * \\ G_i^{\bullet T} P^\bullet & -\gamma^\bullet I_{2r_g} & * \\ H_i^\bullet & 0 & -\gamma^\bullet I_{m_r} \end{bmatrix} < 0 \quad (52)$$

and since (49) implies

$$P^\bullet A_{ei}^\bullet = P^\bullet A_i^\bullet - P^\bullet J_i^\bullet C^\bullet \quad (53)$$

then with the notations (40) and

$$P^\bullet J_i^\bullet = \text{diag} [P_1^\bullet J_i \quad P_2^\bullet] = \text{diag} [Y_{1i}^\bullet \quad P_2^\bullet] \quad (54)$$

$$Y_{1i}^\bullet = P_1^\bullet J_i \quad (55)$$

(52) implies (42). This concludes the proof. \blacksquare

V. ILLUSTRATIVE EXAMPLE

The nonlinear dynamics of the hydrostatic transmission was taken from [8] and this model was used in the design and simulations. Its dynamics is represented by a nonlinear fourth order state-space model

$$\begin{aligned} \dot{q}_1(t) &= -a_{11}q_1(t) + b_{11}u_1(t) + b_{d1}d(t) \\ \dot{q}_2(t) &= -a_{22}q_2(t) + b_{22}u_2(t) + b_{d2}d(t) \\ \dot{q}_3(t) &= a_{31}q_1(t)p(t) - a_{33}q_3(t) - a_{34}q_2(t)q_4(t) + b_{d3}d(t) \\ \dot{q}_4(t) &= a_{43}q_2(t)q_3(t) - a_{44}q_4(t) + b_{d4}d(t) \end{aligned}$$

where $q_1(t)$ is the normalized hydraulic pump angle, $q_2(t)$ is the normalized hydraulic motor angle, $q_3(t)$ is the pressure difference [bar], $q_4(t)$ is the hydraulic motor speed [rad/s], $p(t)$ is the speed of hydraulic pump [rad/s], $u_1(t)$ is the

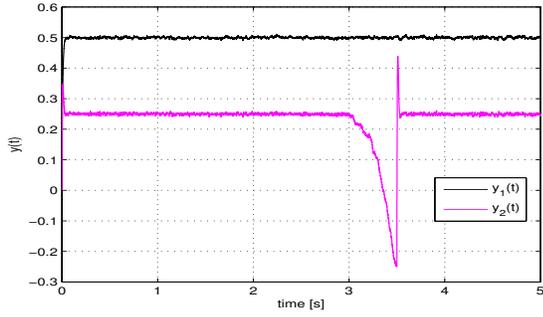


Fig. 1. System output response for the system in the fault regime

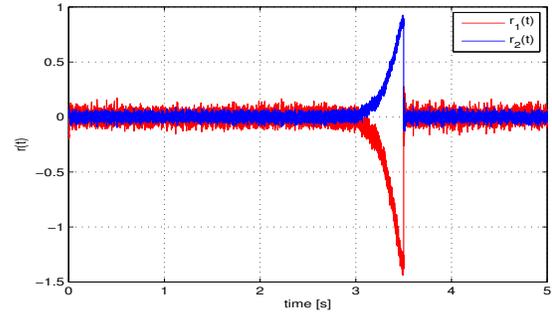


Fig. 2. Residual filter response for the system in the fault regime

normalized control signal of the hydraulic pump, and $u_2(t)$ is the normalized control signal of the hydraulic motor. It is supposed that the external variable $p(t)$, as well as the second state variable $q_2(t)$ are measurable. In given working point the model parameters are

$$\begin{aligned} a_{11} &= 7.6923, & a_{22} &= 4.5455, & a_{33} &= 7.6054 \cdot 10^{-4}, \\ a_{31} &= 0.7877, & a_{34} &= 0.9235, & b_{11} &= 1.8590 \cdot 10^3, \\ a_{43} &= 12.1967, & a_{44} &= 0.4143, & b_{22} &= 1.2879 \cdot 10^3, \\ b_{d1} &= 1.00 \cdot 10^3, & b_{d2} &= 0.80 \cdot 10^3, & b_{d3} &= 0.07 \cdot 10^3, \\ b_{d4} &= 0.01 \cdot 10^3, & \mu_d &= 0, & \sigma_d^2 &= 10^{-5}. \end{aligned}$$

Since the variables $p(t) \in \langle c_1, c_2 \rangle = \langle 105, 300 \rangle$ and $q_2(t) \in \langle d_1, d_2 \rangle = \langle 0.0001, 1 \rangle$ are bounded on the prescribed sectors then vector of the premise variables can be chosen as follows

$$\theta(t) = [\theta_1(t) \quad \theta_2(t)] = [q_2(t) \quad p(t)]$$

Thus, the set of nonlinear sector functions

$$w_{11}(q_2(t)) = \frac{d_1 - q_2(t)}{d_1 - d_2}, \quad w_{12}(q_2(t)) = 1 - w_{11}(q_2(t))$$

$$w_{21}(p(t)) = \frac{c_1 - p(t)}{c_1 - c_2}, \quad w_{22}(p(t)) = 1 - w_{21}(p(t))$$

implies the set of normalized membership functions

$$h_1(t) = w_{11}(q_2(t))w_{21}(p(t)), \quad h_2(t) = w_{11}(q_2(t))w_{22}(p(t))$$

$$h_3(t) = w_{12}(q_2(t))w_{21}(p(t)), \quad h_4(t) = w_{12}(q_2(t))w_{22}(p(t))$$

The transformation of the nonlinear differential equation systems into TS fuzzy system gives

$$\mathbf{A}_i = \begin{bmatrix} -a_{11} & 0 & 0 & 0 \\ 0 & -a_{22} & 0 & 0 \\ a_{31}c_k & 0 & -a_{31} & -a_{34}d_l \\ 0 & 0 & a_{43}d_l & -a_{44} \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} a_{11} & 0 \\ 0 & b_{22} \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B}_d = \begin{bmatrix} b_{d1} \\ b_{d2} \\ b_{d3} \\ b_{d4} \end{bmatrix}, \quad \mathbf{C}^T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

with the associations

$$\begin{aligned} i = 1 &\leftarrow (l = 1, k = 1) & i = 2 &\leftarrow (l = 1, k = 2) \\ i = 3 &\leftarrow (l = 2, k = 1) & i = 4 &\leftarrow (l = 2, k = 2) \end{aligned}$$

Note, the necessary but not sufficient condition to obtain system observability is that the state variable $q_2(t)$ is measurable. Considering this condition, the system output matrix \mathbf{C} was chosen as given above.

Supposing that faults are affecting the second actuator, i.e., $\mathbf{B}_f = \mathbf{B}_2$ and choosing the maximal rank of residual model $m_r = m = 2$, then, exploiting Self-Dual-Minimization (SeDuMi) package for Matlab [17], the reference model design conditions (21), (22) were feasible with the following design parameters

$$\mathbf{J}_1^\circ = 10^3 \begin{bmatrix} -2.0487 & 1.4501 \\ 1.9037 & -1.3536 \\ 0.1930 & -0.0659 \\ 0.0234 & -0.0168 \end{bmatrix}, \quad \mathbf{J}_2^\circ = 10^3 \begin{bmatrix} -1.8759 & 1.5017 \\ 1.7436 & -1.4010 \\ 0.1668 & -0.0755 \\ 0.0234 & -0.0106 \end{bmatrix}$$

$$\mathbf{J}_3^\circ = 10^3 \begin{bmatrix} -0.3766 & 4.0399 \\ 0.3488 & -3.7619 \\ 0.0411 & -0.2512 \\ 0.0060 & -0.0437 \end{bmatrix}, \quad \mathbf{J}_4^\circ = 10^3 \begin{bmatrix} -0.1709 & 4.1239 \\ 0.1579 & -3.8395 \\ 0.0112 & -0.2645 \\ 0.0047 & -0.0385 \end{bmatrix}$$

$$\mathbf{V}_1^\circ = \frac{1}{10^6} \begin{bmatrix} 0.0006 & -0.5067 \\ -0.0104 & -0.0959 \end{bmatrix}, \quad \mathbf{V}_2^\circ = \frac{1}{10^5} \begin{bmatrix} 0.0556 & 0.8547 \\ -0.0028 & -0.0045 \end{bmatrix}$$

$$\mathbf{V}_3^\circ = \frac{1}{10^6} \begin{bmatrix} 0.0862 & 0.8390 \\ -0.0437 & 0.2176 \end{bmatrix}, \quad \mathbf{V}_4^\circ = \frac{1}{10^5} \begin{bmatrix} -0.0279 & 0.6291 \\ 0.0006 & 0.0625 \end{bmatrix}$$

where $\gamma^\circ = 0.9848$. Subsequently, solving (41), (42) with respect to the LMI matrix variables $\mathbf{P}_1^\bullet, \mathbf{P}_2^\bullet, \mathbf{Y}_{1i}^\bullet, \mathbf{V}_i^\bullet, \delta^\bullet$ within the reference model parameters $\mathbf{J}_i^\circ, \mathbf{V}_i^\circ, i = 1, 2, 3, 4$, the feasible solution offers the results $\gamma^\bullet = 0.4117$,

$$\mathbf{J}_1 = 10^3 \begin{bmatrix} 0.0793 & 3.8859 \\ 1.0590 & 0.6706 \\ 0.0913 & 0.0907 \\ 0.0134 & 0.0081 \end{bmatrix}, \quad \mathbf{J}_2 = 10^3 \begin{bmatrix} 0.4077 & 4.3822 \\ 1.1198 & 0.8057 \\ 0.0947 & 0.1008 \\ 0.0143 & 0.0157 \end{bmatrix}$$

$$\mathbf{J}_3 = 10^3 \begin{bmatrix} -0.3700 & 7.5517 \\ 0.9996 & 1.2591 \\ 0.0873 & 0.1749 \\ 0.0125 & 0.0168 \end{bmatrix}, \quad \mathbf{J}_4 = 10^3 \begin{bmatrix} -0.3847 & 7.6388 \\ 0.9739 & 1.2707 \\ 0.0849 & 0.1753 \\ 0.0124 & 0.0228 \end{bmatrix}$$

$$\mathbf{V}_1 = \frac{1}{10^2} \begin{bmatrix} -0.1260 & -0.1012 \\ 0.0159 & -0.0145 \end{bmatrix}, \quad \mathbf{V}_2 = \frac{1}{10^4} \begin{bmatrix} -0.2042 & -0.7317 \\ -0.1741 & 0.0238 \end{bmatrix}$$

$$\mathbf{V}_3 = \frac{1}{10^3} \begin{bmatrix} 0.1332 & 0.1832 \\ -0.0365 & -0.1036 \end{bmatrix}, \quad \mathbf{V}_4 = \frac{1}{10^4} \begin{bmatrix} -0.9652 & -0.1236 \\ 0.1877 & 0.0094 \end{bmatrix}$$

As can be seen, the state observer gain parameters differ from the observer parameters of the reference model. The most interesting, however, is a substantial difference in the weight matrices of the robust TS fuzzy fault detection filter.

The simulation for the fault detection performance was done under the system fuzzy control law in the forced regime

$$\mathbf{u}(t) = \sum_{j=1}^s h_j(\boldsymbol{\theta}(t)) (\mathbf{K}_j \mathbf{q}(t) + \mathbf{W}_j \mathbf{w}(t))$$

where the controller parameters were designed as follows (all details concerning the used design methodology can be found in [13])

$$\mathbf{K}_1 = \begin{bmatrix} 0.0834 & 0.0000 & 0.0844 & 0.0001 \\ 0.0000 & 0.0902 & 0.0000 & 0.0000 \end{bmatrix}$$

$$\mathbf{K}_2 = \begin{bmatrix} 0.0855 & 0.0000 & 0.0869 & 0.0001 \\ 0.0000 & 0.0898 & 0.0000 & 0.0000 \end{bmatrix}$$

$$\mathbf{K}_3 = \begin{bmatrix} 0.1353 & 0.0000 & 0.1933 & 0.0004 \\ 0.0000 & 0.0907 & 0.0000 & 0.0000 \end{bmatrix}$$

$$\mathbf{K}_4 = \begin{bmatrix} 0.1350 & 0.0000 & 0.1926 & 0.0004 \\ 0.0000 & 0.0903 & 0.0000 & 0.0000 \end{bmatrix}$$

$$\mathbf{W}_1 = \begin{bmatrix} 0.0000 & 0.0861 \\ 0.0937 & 0.0000 \end{bmatrix}, \quad \mathbf{W}_2 = \begin{bmatrix} 0.0000 & 0.0884 \\ 0.0933 & 0.0000 \end{bmatrix}$$

$$\mathbf{W}_3 = \begin{bmatrix} 0.0000 & 0.1936 \\ 0.0942 & 0.0000 \end{bmatrix}, \quad \mathbf{W}_4 = \begin{bmatrix} 0.0000 & 0.1927 \\ 0.0938 & 0.0000 \end{bmatrix}$$

The working point was set by $p(t) = 105$ and the desired steady-state output $\mathbf{w}^T(t) = [0.50 \quad 0.25]$, the fault was modeled as the short-circuit outage of the second actuator during $t \in \langle 3.0, 3.5 \rangle$ [s]. Fig. 1 shows the system output response to the fault, the corresponding residual signals are presented in Fig. 2. Comparing with the previous results [12], it can be concluded that H_∞ norm of the residual transfer function with respect to disturbance was substantially reduced and the residual absolute sensitivity to faults is substantially increased.

VI. CONCLUDING REMARKS

Newly introduced robust TS fuzzy fault detection filter design method, as augmentation of the residual generators synthesis for one class of nonlinear systems, is presented in the paper. This is achieved by application of TS fuzzy reasoning, relating to multi-model approximation, as in the observer structure as well as in the residual signals frame and is supported by the TS fuzzy residual reference model. Design conditions are derived in terms of optimization over LMI constraints using standard numerical optimization procedures to achieve, simultaneously, the fuzzy observer asymptotic stability and the optimal residual signal H_∞ performance with respect to the unknown disturbances. Since TS fuzzy fault detection filter design task is generally singular for a non-square system output matrix \mathbf{C} , to obtain more regular conditions any further extensions could be included in the design conditions in future research.

REFERENCES

- [1] L. Bai, Z. Tian, and S. Shi, Robust fault detection for a class of nonlinear time-delay systems, *J. Franklin Institute*, vol. 344, no. 6, pp. 873-888, 2007.
- [2] M. Chadli, State and an LMI approach to design observer for unknown inputs Takagi-Sugeno fuzzy models, *Asian J. Control*, vol.12, no.4, pp. 524530, 2010.
- [3] W. Chen and M. Saif, Observer-based strategies for actuator fault detection, isolation and estimation for certain class of uncertain nonlinear systems, *IET Control Theory Appl.*, vol. 1, no. 6, pp. 1672-1680, 2007.
- [4] S. Ding, *Model-Based Fault Diagnosis Techniques. Design Schemes, Algorithms, and Tools*, Berlin: Springer, 2008.
- [5] J. Guo, X. Huang, and Y. Cui, Design and analysis of robust fault detection filter using LMI tools, *Computers and Mathematics with Applications*, vol. 57, no. 11-12, pp. 1743-1747, 2009.
- [6] Z. Gao and B. Jiang, Delay-dependent robust fault detection for a class of nonlinear time-delay systems, in *Proc. 2nd Int. Symp. Systems and Control in Aerospace and Astronautics ISSCAA 2008*, Shenzhen, China, pp. 1-6, 2008.
- [7] Z. Gao, X. Shi, and S.X. Ding, Fuzzy state/disturbance observer design for T-S fuzzy systems with application to sensor fault estimation, *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 38, no. 3, pp. 875-880, 2008.
- [8] P. Gerland, D. Gross, N. Schulte, and A. Kroll, Robust adaptive fault detection using global state information and application to mobile working machines, in *Proc. 1st Int. Conf. Control and Fault Tolerant Systems SysTol10*, Nice, France, pp. 813-818, 2010.
- [9] D. Ichalal, B. Marx, D. Maquin, and J. Ragot, Observer design and fault tolerant control of Takagi-Sugeno nonlinear systems with unmeasurable premise variables, in *Fault Diagnosis in Robotic and Industrial Systems*, G.G. Rigatos, Ed., Seattle, WA: CreateSpace, ch. 5, 21p, 2012.
- [10] M. Hou and R.J. Patton, An LMI approach to H_∞/H_∞ fault detection observers, in *Proc. UKACC Int. Conf. CONTROL'96*, Exeter, UK, pp. 305-310, 1996.
- [11] A.J. Koshkouei and A.S.I. Zinober, Partial Lipschitz nonlinear sliding mode observers, in *Proc. 7th Mediterranean Conf. Control and Automation MED99*, Haifa, Israel, pp. 2350-2359, 1999.
- [12] D. Krokavec and A. Filasová, On observer-based residual generator design for a class of nonlinear systems, in *Proc. 9th Int. Symp. Applied Machine Intelligence and Informatics SAMI 2011*, Smolenice, Slovakia, pp. 95-99, 2011.
- [13] D. Krokavec and A. Filasová, Optimal fuzzy control for a class of nonlinear systems, *Mathematical Problems in Engineering*, vol. 2012, ID 481942, 29p, 2012.
- [14] D. Krokavec and A. Filasová, Actuator faults reconstruction using reduced-order fuzzy observer structures, in *Proc. 12th European Control Conf. ECC'13*, Zürich, Switzerland, pp. 4299-4304, 2013.
- [15] Z. Lendek, T.M. Guerra, R. Babuška, and B. De Schutter, *Stability Analysis and Nonlinear Observer Design Using Takagi-Sugeno Fuzzy Models*, Berlin: Springer-Verlag, 2010.
- [16] K.M. Passino and S. Yurkovich, *Fuzzy Control*, Berkeley, CA: Addison-Wesley Longman, 1998.
- [17] D. Peaucelle, D. Henrion, Y. Labet, and K. Taitz, *User's Guide for SeDuMi Interface 1.04*, Toulouse: LAAS-CNRS, 2002.
- [18] T. Takagi and M. Sugeno, Fuzzy identification of systems and its applications to modeling and control, *IEEE Trans. Systems, Man, and Cybernetics*, vol. 15, no. 1, pp. 116132, 1985.
- [19] K. Tanaka and H.O. Wang, *Fuzzy Control Systems Design and Analysis. A Linear Matrix Inequality Approach*, New York, NY: John Wiley & Sons, 2001.
- [20] F.E. Thau, Observing the state of nonlinear dynamical systems, *Int. J. Control*, vol. 17, no. 3, pp. 471-479, 1973.
- [21] D. Xu, B. Jiang, and P. Shi, Nonlinear actuator fault estimation observer. An inverse system approach via a TS fuzzy model, *Int. J. Appl. Math. Comput. Sci.*, vol. 22, no. 1, pp. 183-196, 2012.
- [22] F. Yang and Y. Li, Set-membership fuzzy filtering for nonlinear discrete-time systems, *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 1, pp. 116-124, 2010.
- [23] X. Zhang, M.M. Polycarpou, and T. Parisini, Fault diagnosis of a class of nonlinear uncertain systems with Lipschitz nonlinearities using adaptive estimation, *Automatica*, vol. 46, no. 2, pp. 290-299, 2010.

Modeling and Simulation of a 12kW Direct Driven PM Synchronous Generator of Wind Power

A.Senthil Kumar

Energy for Utilization of Non-Traditional Energy
VSB-Technical University of Ostrava
Ostrava- Poruba, Czech Republic
vastham@gmail.com

Thomas Cermak and Stanislav Misak

Energy for Utilization of Non-Traditional Energy
VSB-Technical University of Ostrava
Ostrava-Poruba, Czech Republic

Abstract— Wind energy, a kind of non-conventional energy source and environmental fresh energy, is widely studied nowadays. Directly driven wind turbine permanent magnet synchronous generator (D-PMSG) plays a vital role in the latest small-scale wind Power technology. This paper presents the modeling and simulation of a 12kW direct driven Permanent Magnet Synchronous Generator of Wind Power. This model consists of Weibull parameter estimation, wind turbine model, PMSG model and load model. The present modeling and simulation results are validated with sample experiments results are conducted in VSB-Technical university of Ostrava.

Index Terms— Wind turbine, PMSG, renewable energy, Weibull parameter, wind power.

I. INTRODUCTION

Nowadays, most electrical energy is generated by burning huge fossil fuels and special weather conditions such as acid rain and snow, climate change, urban smog, regional haze, several tornados, etc., have happened around the whole world. It is now clear that the installation of a number of wind turbine generators can effectively reduce environmental pollution, fossil fuel consumption, and the costs of overall electricity generation. Although wind is only an intermittent source of energy, it represents a reliable energy resource from a long-term energy policy viewpoint. Among various renewable energy resources, wind power energy is one of the most popular and promising energy resources in the whole world today. Global winds caused by pressure difference across the earth surface due to uneven heating of the earth by solar radiation. In a simple flow model, air rises at the equator and sinks at the poles. The atmospheric winds are affected by many factors such as the pressure, gradient, gravitational forces, inertia of air, the earth rotation, and friction with the earth surface [1, 2].

The wind power mainly depends on geographic and weather conditions and varies from time to time. Therefore it is necessary to construct a system that can generate maximum power for all operation condition. Recently, PMSG is used for wind power generating system because of its advantages such as better reliability, lower maintenance and more efficient. Industry generally remains hesitant to changes in the paradigm of reliable electric power supply that has been in

place for most of the last century. This is mainly because the new paradigm mainly involves renewable energy, which is mainly captured from naturally intermittent renewable sources. For instance, solar electricity only functions at night from battery storage systems, and wind being stochastic in nature means power generation from it is only possible when the wind is blowing. This raises the challenge of intermittent generation integration issues [3].

Wind power is an environmental friendly energy source with no fuel costs. The use of wind power is increasing every year and more and more countries invest in large wind turbines for part of their energy supply. As the amount of the wind turbines increases it is important that the technology of wind turbines keep evolving [4]. There are several different types of electrical system available for converting the wind power to electricity but no single technology is dominating the market. Modern technology of wind and solar radiation is largely developed in the last twenty years. These two categories of resources inherently represent small wind turbines with rotor surface to 200m². Nowadays, it quite commonly uses these technologies for obtaining electricity not only in places where there is no public distribution site. These energy can be used for lighting, power appliances, hot water and heating. MVE generated clean energy and can help reduce CO² emissions. The island's power grid systems, or as it is sometimes referred to, these plants are currently experiencing in Europe and in the Czech Republic boom. Basic insular power or if insular systems that are not connected into the distribution grid, are referred to as an island or even off-grid. Its application is found primarily where it is not possible to build a traditional public electricity connection (cottages, RD, etc.) or mobile installations (caravans, boats, mobile homes, mobile units etc.).

VSB-TUO University was established in early 2010 wind turbine with a synchronous generator with permanent magnets. The control system allows the wind turbine to operate in an autonomous mode, where the performance of the plant led to a general load, also can work in the general supply of load voltage bus and last but not least, it is possible to deliver power from wind power into the grid via the inverter with regenerative unit. Prior to the implementation of wind power were a number of measurements in the laboratory in order to

find the optimum settings for each component of the power and control system and also to verify the operating characteristics of synchronous generator with permanent magnets for its various operating modes. Parallel testing of synchronous generator with permanent magnets mathematical model was created in the generator environment matlab that was verified based on the results of experimental measurements. Thus created a mathematical model of a synchronous generator with permanent magnets will serve as support for real time measurement of individual components in the design of autonomous networks, where the performance of wind power will be accumulated using car batteries, together with the performance of the newly built photovoltaic power plants. The rechargeable battery will subsequently fed through power converters general load simulating consumption with varying degrees of priority power.

A direct driven generator is secure from losses, maintenance and costs associated with a gearbox. Also, it reduces the torsional constraints on the shaft imposed by eigen frequency oscillations. Gearless wind turbines are becoming increasingly popular, see for instance [5]. Here, a radial flux machine is used but other designs have been used for wind turbines [6-8]. Furthermore, different designs, for instance to have an ironless stator, have been suggested [9]. The use of PMs instead of electromagnets is motivated by the simpler rotor construction, i.e. no field coils have to be electrified. Furthermore, the efficiency is improved, as rotor losses are practically eliminated. In this article, simulation and analysis of 12 kW direct driven PMSG of wind power are installed in VSB-Technical university of ostrava are presented. The sample experiments results are validated with simulated work are verified.

II. SYSTEM UNDER STUDY

The proposed system in installed in VSB-TUO, Ostrava as shown in Fig.1 and detailed block diagram contains such as weibull parameter estimation, wind turbine model, PMSG model and load as given in Fig.2.



Fig.1. wind power plant at VSB-Technical University of Ostrava

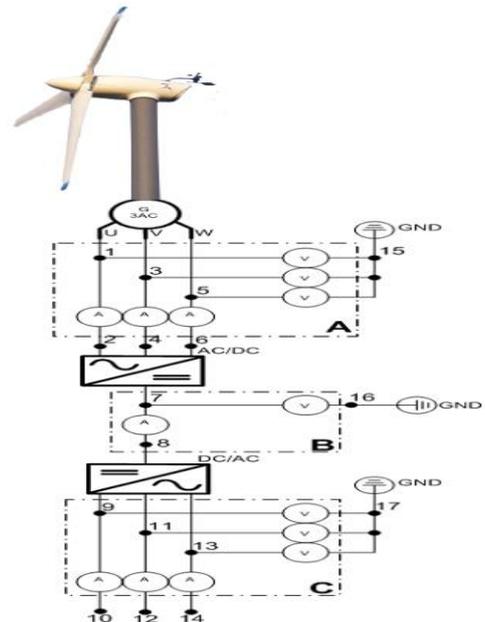


Fig.2 block diagram for wind system

A. Weibull parameter estimation

Graphical method for estimating Weibull parameters, namely, shape parameter (k) and scale parameter (c). The Weibull distribution is an important distribution especially for reliability and maintainability analysis. The suitable values for both shape parameter and scale parameters of Weibull distribution are important for selecting locations of installing wind turbine generators. The scale parameter of Weibull distribution also important to determine whether a wind farm is good or not. The presented method is the analytical methods and computational experiments on the presented methods are reported. Wind speed data for the continuous six months as shown in Fig.3. Fig3 shows the variation of wind speed versus measurement or count in time series.

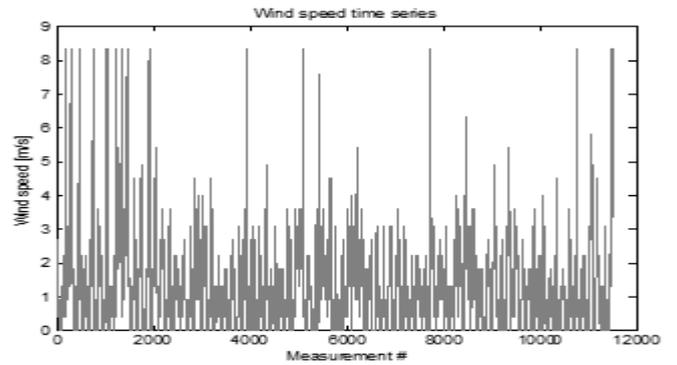


Fig.3 wind speed versus measurement in time series

The probability density function of Weibull distribution is given by [10],

$$f(v) = \frac{k}{c} \left(\frac{v}{c}\right)^{k-1} e^{-\left(\frac{v}{c}\right)^k} \tag{1}$$

Many literature surveys for this study shows that, shape parameter of Weibull distribution range from 1.2 to 2.75 for most wind condition in the world. Cumulative distribution function (CDF) of Weibull distribution is given by

$$F(v) = 1 - e^{-\left(\frac{v}{c}\right)^k} \quad (2)$$

From equation 1 and 2, we plot of the Weibull probability density function and Weibull cumulative distribution function with the same values of k as the probability density function as shown in Fig.4.

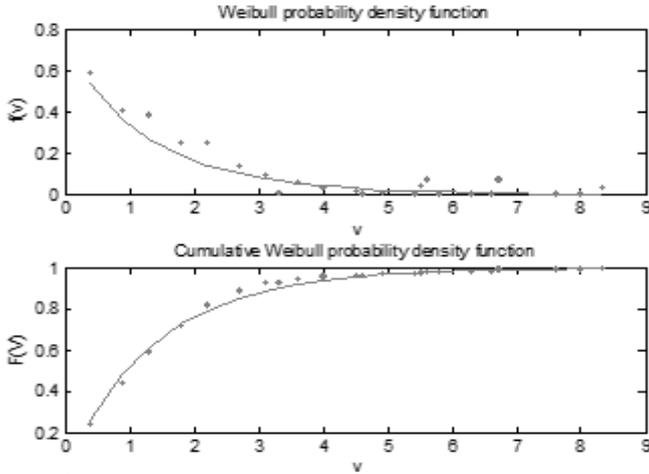


Fig.4. Probability density and cumulative probability density function with wind speed

Graphical method is calculated by using cumulative distribution function. Equation 3 can be applying twice the logarithm. Therefore, this can be rewritten as.

$$\ln(\ln(1 - F(v))) = k \ln c - k \ln v \quad (3)$$

Equation 3 is an equation of a straight line. To plot F(v) versus v as shown in fig.5.

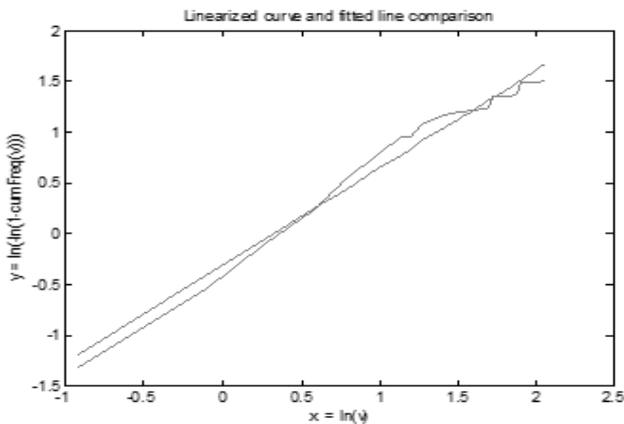


Fig.5 F(v) versus v

B. Wind turbine model

The proposed model of the wind turbine used in this paper is based on the actual value system. The conventional wind model was found not to be suited to the actual value system. In order to make wind turbine compatible with real value components, several modifications have been made in the Simulink wind turbine model. First, the wind-turbine model was changed from a per unit system to the actual value system. Next, the original model represents a variable pitch model,

while for this paper; the model was changed to represent a fixed pitch turbine. For this purpose, the value of pitch angle is set to zero.

The fixed pitch model is used to isolate the effects of electrical control rather than mechanical control because pitch control is achieved through hydraulic manipulation. Since the power coefficient characteristic of a non-linear curve that reflects the aerodynamic behaviour of a wind turbine is necessary, this curve must be defined. The C_p curve in this paper is taken from the wind-turbine model provided by Matlab- Simulink [11].

The tip speed ratio λ , which is the ratio of linear speed at the tip of the blades to the speed of wind, is expressed as.

$$\lambda = \left(\frac{\omega_T}{v_w}\right) r \quad (4)$$

Where, ω_T . angular velocity of the wind turbine, r- radius of the wind turbine, V_w - wind speed

$$C_p(\lambda, \beta) = c_1(c_2/\lambda - c_3\beta - c_4)e^{-c_5/\lambda} + c_6\lambda_i \quad (5)$$

$$1/\lambda_i = 1/\lambda + 0.08\beta - 0.035/\beta^3 + 1 \quad (6)$$

Where $c_1 = 0.5176$, $c_2 = 116$, $c_3 = 0.4$, $c_4 = 5$, $c_5 = 21$, $c_6 = 0.0068$

Therefore, the new power coefficient is derived as:

$$C_p(\lambda) = 0.5176(116/\lambda - (116*0.035) - 5) e^{-21/\lambda - 21*0.035} + 0.0068\lambda \quad (7)$$

The tip speed ratio indicates the operating condition of a turbine as it takes into account the wind created by the rotation of the rotor blades. The tip speed ratio shows tangential speed at which the rotor blade is rotating compared with the fixed wind speed.

As the wind speed changes, the tip speed ratio and the power coefficient will vary. The power coefficient characteristic has a single maximum at a specific value of the tip speed ratio ($\lambda_{optimum}$). Therefore, if the wind turbine is operating at constant speeds, then the power coefficient will be maximum (C_{pmax}), only at one wind speed. C_p and λ characteristics curve for the wind turbine is illustrated in Fig.6 using a curve fitting technique.

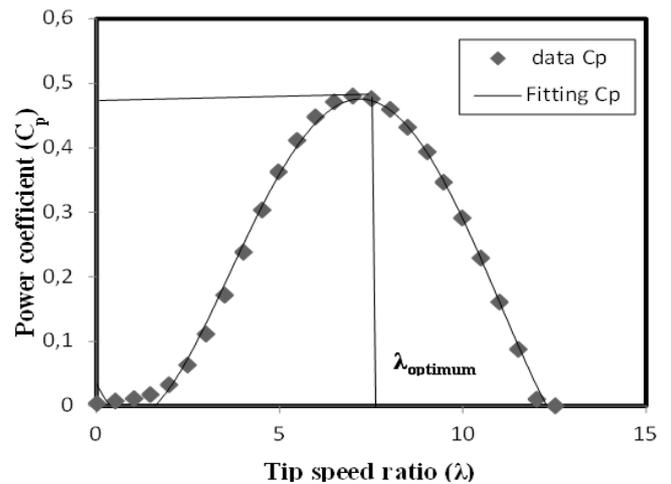


Fig.6. Curve relationship between C_p and tip speed ratio (λ)

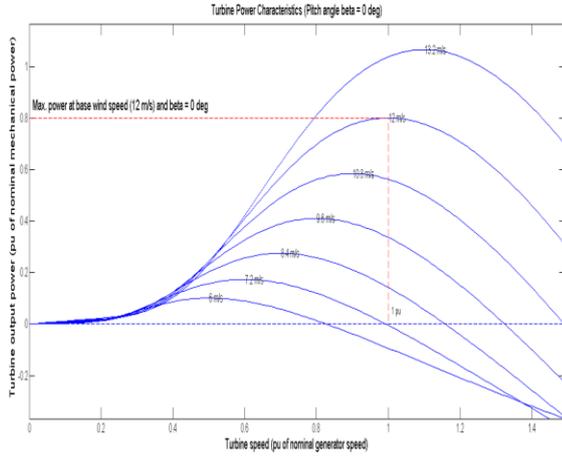


Fig. 7. Power-speed characteristics

Typically, wind turbines are designed to start running at wind speeds somewhere around 4 to 5 m/s. This is called the cut-in wind speed. The wind turbine will be programmed to stop at high wind speeds of 25 m/s, in order to avoid damaging the turbine. This speed is called the cut-out wind speed. Figure 7 shows the turbine power-speed characteristics for various wind velocity values.

C. PMSG Model

Before developing the mathematical model of the PMSG, several important assumptions need to be made: the damping effect in the magnets and in the rotor are negligible; the magnetic saturation effects are neglected; the eddy current and hysteresis losses are neglected; the back electromotive force (EMF) induced in the stator windings are sinusoidal; for simplicity, all the equations of PMSMs are expressed in motor (consumer/load) notation, that is, negative current will be prevailing when the model refers to a generator. Negative current means that at the positive polarity of the terminal of a device the current is out of that terminal.

A directly driven PMSG is dynamically modeled using MATLAB detailed synchronous machine model. In order to define the generator-side control methodology, equations of the generator-side circuit are projected in to a reference frame synchronously rotating and aligned with the rotor, giving [12, 13].

Voltage equation are given by

$$v_q = R_s i_q + \omega_r \lambda_d + p \lambda_q \quad (8)$$

$$v_d = R_s i_d - \omega_r \lambda_q + p \lambda_d \quad (9)$$

Therefore, Flux linkage

$$\lambda_q = L_q i_q \quad (10)$$

$$\lambda_d = L_d i_d + \lambda_f \quad (11)$$

Substituting equations 10 and 11 into equations 8 and 9

$$v_q = R_s i_q + \omega_r [L_d i_d + \lambda_f] + p L_q i_q \quad (12)$$

$$v_d = R_s i_d - \omega_r L_q i_q + p [L_d i_d + \lambda_f] \quad (13)$$

Arranging equations 12 and 13 in matrix form

$$\begin{pmatrix} v_q \\ v_d \end{pmatrix} = \begin{pmatrix} R_s + p L_q & \omega_r L_d \\ -\omega_r L_q & R_s + p L_d \end{pmatrix} \begin{pmatrix} i_q \\ i_d \end{pmatrix} + \begin{pmatrix} \omega_r \lambda_f \\ p \lambda_f \end{pmatrix} \quad (14)$$

The developed torque motor is being given by

$$T_e = \left[\frac{3}{2} \right] \left[\frac{P}{2} \right] [\lambda_d i_q - \lambda_q i_d] \quad (15)$$

Mechanical torque equation is

$$T_e = T_L + B \omega_m + J \frac{d\omega_m}{dt} \quad (16)$$

Solving for the rotor mechanical speed form equ.9

$$\omega_m = \int \left(\frac{T_e - T_L - B \omega_m}{J} \right) dt \quad (17)$$

$$\text{Where } \omega_m = \omega_r \left[\frac{2}{P} \right]$$

In the above equations ω_r is the rotor electrical speed where as ω_m is the rotor mechanical speed

III. RESULT AND DISCUSSION

The model of PMSG implemented in matlab simulink as shown in Fig.8. The presented method and simulation results are conducted in MATLAB/SIMULINK.. In this paper, the base wind speed is considered as 12m/s. Fig. 9 shows the mechanical torque of PMSG, Electromagnetic torque of PMSG, pitch angle, rotor speed (rad/sec) and Per-unit with time.

Fig.10 shows the power, line voltage, line current and rotor speed with time keeping the speed of the wind will be 12m/s.

Experimental measurements, which has been on the service was implemented, measurement of the voltage constant, therefore the measurement of output voltage depending on the speed of the generator. The results from the experimental measurement were also used for verification of mathematical model in matlab-simulink environment. When measuring the generator powered by the dynamometer and the set speed is measured the phase voltage connection to the stars.

The table 1 shows the average value of the voltage measurement in three phases, the last column lists the output values of the mathematical modeling.

Sl.no.	Speed(rev/min)	Voltage(experiment)	Voltage(simulated)
1.	50	117.1	120
2.	100	233.2	230
3.	150	349	350
4.	180	417.8	420
5.	200	464.8	465

Table 1 comparison of the results of experimental measurement and the mathematical model for idle status

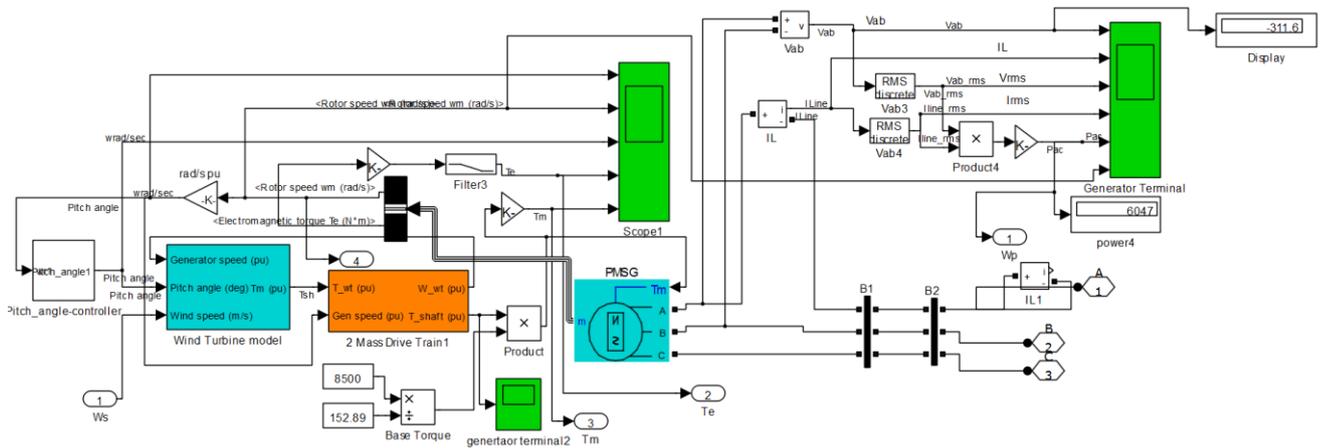


Fig.8. PMSG modeled with simulink

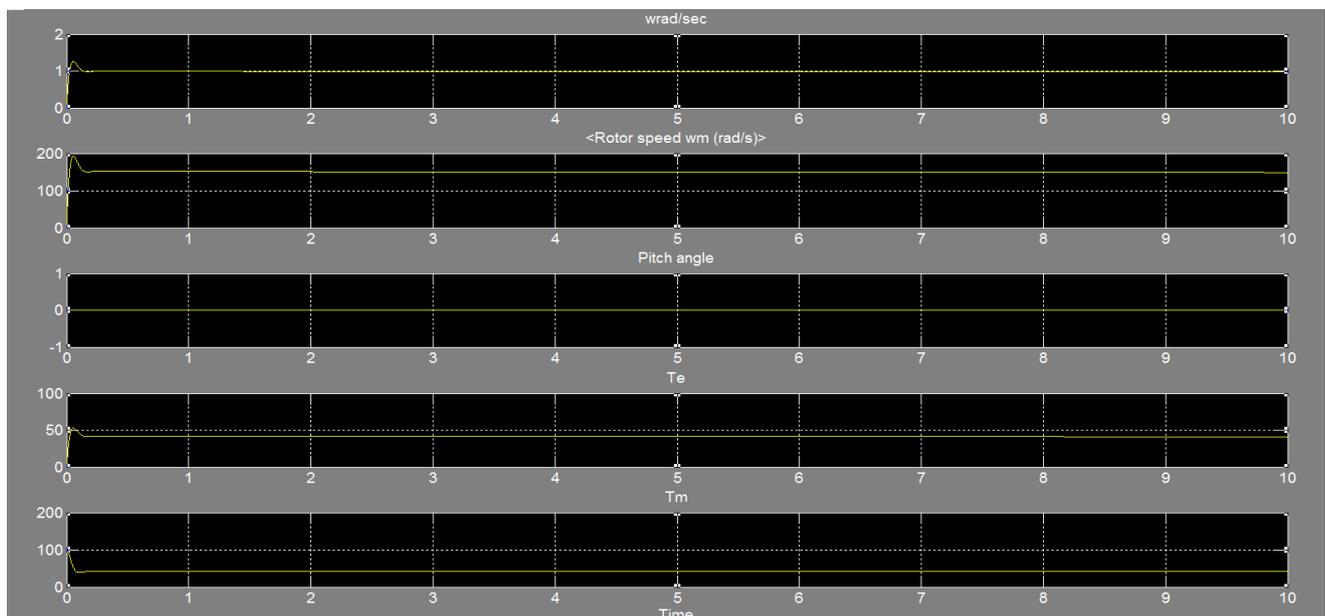


Fig.9. mechanical torque, Electromagnetic torque pitch angle, rotor speed with time

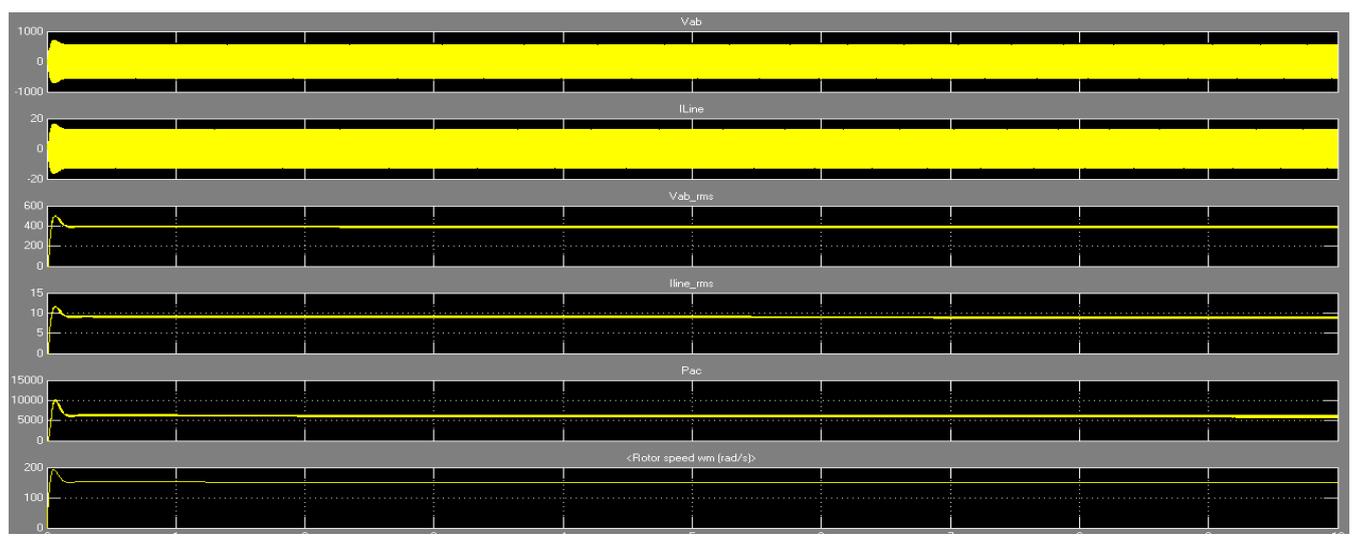


Fig.10. power, line voltage, line current and rotor speed with time

IV. CONCLUSIONS

This paper presented the results of mathematical modeling and simulation of a synchronous generator to the newly built wind power 12 kW in the locality of VSB-TU Ostrava. It consists of wind turbine, drive train, PMSG, pitch angle control and Load model. PMSG and load model are established in $d-q$ model. The sample experiment's results are validated with simulated work are verified.

ACKNOWLEDGMENT

This paper has been elaborated in the framework of the project New creative teams in priorities of scientific research, reg. no. CZ.1.07/2.3.00/30.0055 supported by Operational Programme Education for Competitiveness and co-financed by the European Social Fund and the state budget of the Czech Republic.

REFERENCES

- [1] A.Senthil Kumar, Josiah L Munda, Optimization of Voltage and Frequency Regulation in an Isolated Wind-Driven Six-Phase Self-Excited Induction Generator, Journal of the Energy Institute, Volume 87, issue 3, pp. 235–245, 2014.
- [2] Paritosh Bhattachara, Rakhi Bhattacharjee, A Study on Weibull Distribution for Estimating the Parameters, Journal of Applied Quantitative Methods, Volume 5, No.2, pp.234-241, 2010
- [3] A. K. Akella, R. P. Saini, and M. P. Sharma, "Social, economical and environmental impacts of renewable energy systems," Renewable Energy, vol. 34, no. 2, pp. 390–396, 2009.
- [4] Sandra Eriksson, andreas solum, Mats Leijon and Hans Bernhoff, Simulations and Experiments on a 12kW direct driven PM synchronous generator for wind power, Renewable Energy Journal, vol.33, pp. 674-681, 2008.
- [5] Eriksson S, Bernhoff H. Generator-damped torsional vibrations of a vertical axis wind turbine. Wind Eng vol.29, issue 5, pp.449–62, 2005.
- [6] Muljadi E, Butterfield CP, Wan Y. Axial-flux modular permanentmagnet generator with a toroidal winding for wind-turbine applications. IEEE Trans. Ind Appl., vol. 35, no.4, pp.831–836, 1999.
- [7] Chalmers BJ, Spooner E. An axial-flux permanent-magnet generator for a gearless wind energy system. IEEE Trans Energy Convers, vol.14, no.2, pp.251–7, 1999.
- [8] Chen J, Nayar CV, Xu L. Design and FE analysis of an outer-rotor PM -generator for directly-coupled wind turbine applications. In: IEEE industry applications conference. Thirty-Third IAS annual meeting, Australia, 1998, pp. 387–94.
- [9] Spooner E, Gordon P, Bumby JR, French CD. Lightweight ironless stator PM generators for direct-drive wind turbines. IEE Proc Electric Power, vol.152, no.1, pp.17–26, 2005.
- [10] Kaigui xie, zefu jiang, wenyuan lu. Effect of wind speed on wind turbine power converter reliability, IEEE transaction of Energy conversion , vol.27, pp.96-104, 2012.
- [11] Seyit A, Akdag a, Ali Dinler b, A new method to estimate Weibull parameters for wind energy applications, Energy Conversion and Management, vol. 50, pp. 1761–1766, 2009.
- [12] M. Chinchilla, S. Arnaltes, and J. C. Burgos, "Control of permanentmagnet generators applied to variable-speed wind-energy systems connected to the grid," IEEE Trans. Energy Conversion, vol. 21, no. 1, pp. 130-135, Mar. 2006.
- [13] A. E. Fitzgerald, J. C. Kingsley, and S. D. Umans, Electric Machinery. New York:McGraw-Hill, 1990.

Some Problems of Fuzzy Modeling of Telecommunications Networks

Kirill Garbuzov
Novosibirsk State University
Department of Mechanics and Mathematics
Novosibirsk, Russia, 630090,
Email: gartesk@rambler.ru

Alexey S. Rodionov
ICM&MG SB RAS
Novosibirsk, Russia, 630090,
Email: alrod@sscc.ru

Abstract—Some common ideas and problems concerning the use of fuzzy logic for modeling unreliable networks are discussed on simple example. Approach to measuring fuzziness of different fuzzy parameters types is presented along with rules of their mutual conversions. The difficulty of keeping information at these conversions is shown.

I. INTRODUCTION

Modern communication systems, sensor networks in particular, provide us with many problems concerning calculation and optimization of models corresponding to real infrastructure parts. This also includes some other kinds of networks, like transport.

Many of these models are representing unreliable networks, i.e. networks with failing components. The most common example is probabilistic models, where each component has probability of failure [1]–[3]. But this method has its own cons since we can't always have sufficient statistical data or even have to use expert opinion. Alternative approach is using fuzzy models for description of reliability.

Fuzzy models help avoid problem of insufficient statistical data allowing membership function to have values between 0 and 1. Fuzzy logic was introduced in 1965 by Lotfi A. Zadeh. It has been applied to many fields involving descriptions of uncertain events.

This approach allows us to weaken the formalism of classic probabilistic models and simplify the solution of our problem. There is also a wide choice of ways of representing fuzzy parameters, which

gives us flexibility to build completely different models for different purposes. In last years, we can find a lot of researches concerning usage of fuzzy models for network's analysis and optimization [4]–[6]. In all these and other papers that consider fuzzy models of telecommunication networks, a unified approach for description of uncertainty is used, that is some unique model is chosen, for example triangle numbers or intervals. At the same time, when describing large network, different teams of experts may be recruited that use different techniques and so present results in different measures of uncertainty. For analysis of such mixed data conversion to some unique presentation is needed. This conversion may lead to loss of information thus increasing uncertainty of a result. In this paper we discuss the problem and propose some methods for conversion of uncertainty presentations with minimal loss of information.

II. INFORMATIVENESS OF NETWORK ELEMENTS

Using fuzzy logic in networks modeling allows us to formulate the problems, which can't be considered in standard probabilistic approach [1]. Calculating of network components' informativeness is one of these problems.

Let G be an arbitrary graph with unreliable edges, which are set by fuzzy numbers. Let's call these parameters the *possibility* of presence of each edge. For each edge we can also introduce *fuzziness* - some function showing us how much information we can get from this edge. One of the examples is

logarithmic entropy, which can be easily calculated for most cases of classic fuzzy numbers. At last, we can talk about *informativeness* - a value, which shows us how much influence has the change of fuzziness to the value of the objective function. In practice this problem could stay for the process of maintenance of big network in the state of lack of resources. Knowing the informativeness, we can consider, which elements of our network are more important in terms of specifying their condition.

In this paper in most of the examples we are using the logarithmic entropy as the measure of fuzziness:

$$d(A) = \int_U S(\mu_A(x)) dx;$$

where $S(y) = -y \cdot \ln(y) - (1 - y) \cdot \ln(1 - y)$ - Shannon function.

Let's consider a simple example to show the problem closely. Let each edge of G have a triangular number $\langle \gamma, a, \delta \rangle : 0 < a < 1; a - \gamma \geq 0; a + \delta \leq 1$ as a fuzzy parameter. Operations between them could be defined as:

$$\begin{aligned} &\langle \gamma_1, a_1, \delta_1 \rangle * \langle \gamma_2, a_2, \delta_2 \rangle = \\ &= \langle \max(\gamma_1, \gamma_2), a_1 * a_2, \max(\delta_1, \delta_2) \rangle \end{aligned}$$

Here we can consider these triangular numbers as probabilities of presence for the graph elements with some additional parameters (γ and δ).

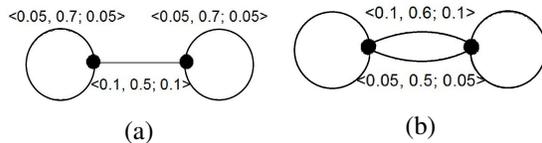


Fig. 1. Informativeness of network elements

On the Fig.1(a) we compare connection possibility (fuzzy analogue of connection probability) of two subgraphs connected by a bridge before and after changing the fuzziness of this bridge. At the beginning, the fuzzy number corresponding to objective function value equals $\langle 0.1, 0.245, 0.1 \rangle$. If we now change the fuzzy parameter of bridge to $\langle 0, 0.5, 0 \rangle$ connection possibility value becomes equal to $\langle 0.05, 0.245, 0.05 \rangle$.

In the case (b) we are calculating the possibility of connection of two subgraphs. Without changing parameters this value equals $\langle 0.1, 0.8, 0.1 \rangle$. Alternately modifying fuzzy numbers corresponding to both edges of the bridge, we get different results. It means that these elements are not equal in terms of informativeness.

Consequently, informativeness of elements can be affected not only by their own fuzzy parameters but also by the structure of the whole network and parameters of other edges. It is obvious that we only need to compare the values of informativeness of elements, so we can try to use different heuristic algorithms.

We are currently checking the hypothesis, which lets us to use not the whole graph but only some subgraphs containing the edges we try to compare. This localization could dramatically decrease the calculation difficulty (e.g. using the minimal chain containing objective edges makes the solution trivial).

III. CONVERSION OF FUZZY NUMBERS

Converting different types of fuzzy parameters is another important problem. If we get network parameters from different sources, they can be represented in different types of fuzzy numbers. So we have to convert them to one common type and keep the most information they initially have. The problem is: how to choose this common type to lose the least amount of informativeness?

Knowing the membership function of the fuzzy number we can easily find the logarithmic entropy. E.g. for triangular numbers $\langle \gamma, a, \delta \rangle$ it equals $\frac{1}{2}(\gamma + \delta)$. Now if we know the measure of informativeness we can set arithmetic operations to correspond to any specific problem (depending on the objective function, network structure etc.). So we have to commit not only the measure but also the set of operations on our fuzzy parameters.

The problem appears when we try to convert these parameters from one type to another e.g. trapezoidal numbers to triangular or triangular numbers to intervals. Let's consider another simple example to show it.

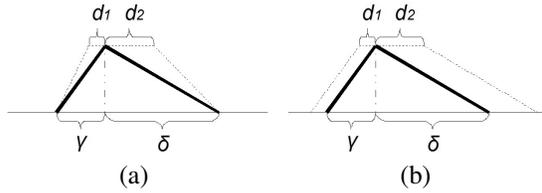


Fig. 2. Converting fuzzy numbers

Let A be a triangular number $\langle \gamma, a, \delta \rangle$. As shown above, its logarithmic entropy equals $\frac{1}{2}(\gamma + \delta)$. Let's build a trapezoidal number by changing the membership function to 1 on the interval with length $d_1 + d_2$ (Fig.2(a)). Its entropy equals $\frac{1}{2}(\gamma + \delta - d_1 - d_2)$. As we see, logarithmic entropy decreases and this can be treated as information loss since we don't use any additional information about network to change that parameter.

If we try to keep the value of one chosen fuzziness measure (Fig.2(b)), we obviously get the changes in other ways of calculating it. So we have to choose this function carefully for the current problem in order to avoid the artificial changes of entropy.

Another important special case is representing fuzzy parameters by intervals, which actually can not be straightly represented as classic fuzzy numbers. Arithmetic of intervals is a special algebraic system, which formalizes operations on intervals. Converting classic fuzzy numbers to intervals and back can be done in different ways.

The easiest idea is to build membership function as constant on the interval with length $\gamma + \delta$. Entropy then equals:

$$d = (\gamma + \delta) (-h \cdot \ln(h) - (1 - h) \cdot \ln(1 - h)).$$

Keeping the entropy value we can get $h \approx 0.8$ or $h \approx 0.2$. But here we obviously lose information about the peak of the initial parameter.

To show the reverse connection between these types of fuzzy numbers we formulate the following lemma:

Lemma 1: Let the fuzziness measure on intervals and triangular numbers be continuous, strictly monotone functions $\phi(\gamma, \delta)$ and $\psi(l)$, and ψ is differentiable on $[0,1]$. Now if for some interval

with length $l = \gamma + \delta$ we have $\phi(\gamma, \delta) \leq \psi(\gamma + \delta)$, then additional fuzziness on converting interval to triangular number is bounded by the value depending only on ψ .

Proof: Let's point out that we only need to proof the case, where the entropy growth, i.e.:

$$\psi(l) = \psi(\gamma + \delta) \leq \phi(\gamma', \delta').$$

Consequently from the condition of lemma:

$$\psi(\gamma + \delta) \leq \psi(\gamma' + \delta'),$$

Then using the mean value theorem:

$$\begin{aligned} \Delta d &\leq \psi(\gamma' + \delta') - \psi(\gamma + \delta) \leq \\ &\leq \psi'(l + \epsilon(l' - l))(l' - l) \leq \\ &\leq \max_{x \in [0,1]} \psi'(x) \end{aligned}$$

Remark: In lemma proof $l' - l$ is bounded by 1, but in practice this value can be much smaller. ■

We can define membership function for intervals as 0.5 on the interval with length l . It gives estimation for entropy difference ($\approx 0.2l$). Lemma gives rough estimation for reverse conversion and provides better results for special cases.

IV. ALGORITHMS EXAMPLES

A. Search of spanning trees

Problem of searching the minimum spanning tree can be easily solved by many algorithms. Getting a single spanning tree is also a trivial task, which can be resolved by either depth-first search or breadth-first search in linear time. These trees can be used for calculating network connection probability.

For any graph, the number of spanning trees can be calculated using Kirchhoff's matrix-tree theorem. In the fuzzy models we can try to find k best spanning trees, or even trees with fuzzy restrictions. Difficulty of such problems can vary depending on our model. For example, we can represent the informal description of heuristic algorithm for searching the set of trees meeting the fuzzy restrictions on connection possibility. It is based on the local search method. On each iteration we try to change one of the edges of current tree without breaking the connection. As a result we get some set of spanning trees without the full search. Depending on the fuzziness measure and initial restrictions we can ignore "unpromising" trees during each iteration.

B. Connection of bipartite graph

Let G be a bipartite graph with a set of fuzzy edges as a bridge between its subgraphs. We need to choose a subset of these edges to have the maximal connection possibility with limited summary weight (represented by another fuzzy parameter in common case).

With additional restrictions we can get different problems of discrete optimization. For example, with prohibition on "boundary" nodes to have more than one incidental edge from the "bridge" set, we get the problem of maximal matching. Here we don't have any additional restrictions, so we get the 0-1 knapsack problem. All initial parameters are considered fuzzy numbers.

Let's look more carefully at the greedy algorithm. Like for the real parameters, it doesn't always give the exact solution. For testing we used triangular numbers with two different sets of operations for weights and possibilities. Since the possibilities of presence must be limited by the fuzzy analogue of segment [0,1], operations were formulated in this way: centers were handled as usual probabilities and boundaries were found as mean values.

For weights we additionally had to define multiplicative inverse element and full order. So we used operations defining the field of triangular numbers:

$$\begin{aligned} \langle \gamma_1, a_1, \delta_1 \rangle + \langle \gamma_2, a_2, \delta_2 \rangle &= \\ &= \langle \gamma_1 \cdot \gamma_2, a_1 + a_2, \delta_1 \cdot \delta_2 \rangle \\ \langle \gamma_1, a_1, \delta_1 \rangle \cdot \langle \gamma_2, a_2, \delta_2 \rangle &= \\ &= \langle e^{\ln(\gamma_1) \cdot \ln(\gamma_2)}, a_1 \cdot a_2, e^{\ln(\delta_1) \cdot \ln(\delta_2)} \rangle \end{aligned}$$

Multiplicative inverse value:

$$\langle \gamma, a, \delta \rangle^{-1} = \left\langle e^{\frac{1}{\ln(\gamma)}}, \frac{1}{a}, e^{\frac{1}{\ln(\delta)}} \right\rangle$$

It is possible to avoid defining some additional operations on weight, e.g. by using any other function growing with possibility growth and decreasing with weight growth instead of common ratio. Algorithm is obviously polynomial: sorting for $O(n \cdot \ln(n))$ and then passing n elements.

```

1: function GREEDY(costs, weights, limit)
2:   for i = 0..n - 1 do
3:     ratios[i] = costs[i] * weights[i]-1
4:     resultVector[i] = false
5:   end for
6:   currSum = 0
7:   cost = 0
8:   while currSum ≤ limit do
9:     currMax = 0
10:    max = -1
11:    for i = 0..n - 1 do
12:      tempVal = ratios[i]
13:      if currMax < tempVal and
!resultVector[i] then
14:        tempSum = currSum +
weights[i]
15:        if tempSum ≤ limit then
16:          currMax = tempVal
17:          max = i
18:        end if
19:      end if
20:    end for
21:    if max ≥ 0 then
22:      currSum+ = weights[max]
23:      cost+ = costs[max]
24:      resultVector[max] = true
25:    else
26:      break
27:    end if
28:  end while
29:  return cost
30: end function

```

In this algorithm we consider all values template, extending FuzzyNumber class (i.e. triangular numbers). All used operations must be defined for the exact types of fuzzy numbers. This approach has to use the multiplicative inversion, which we don't need for brute-force search. And most of the results for the normal greedy algorithm for knapsack problem can be applied in our case as well.

Having the trivial realization, this problem shows flexibility and convenience of fuzzy models. This algorithm was implemented using object-oriented programming language Java, as a part of a big package dedicated to fuzzy networks. First of all,

using interfaces allows to make a template algorithm without specifying type of fuzzy numbers and operations on them. Secondary, it is easy to think of hierarchy of fuzzy numbers, which can also be easily represented using object-oriented paradigm (Fig.3). Any algorithm can be implemented as a separate generic class, and interfaces allow to monitor the correctness of operations in use.

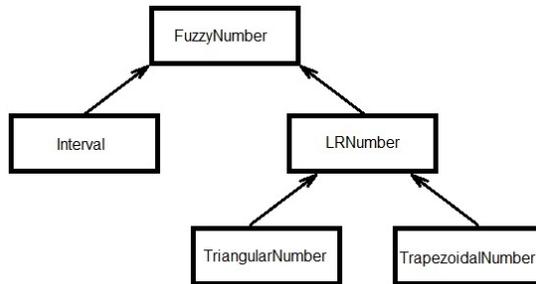


Fig. 3. Example of fuzzy numbers class hierarchy

V. CONCLUSION

The brief results presented above shows some problems in fuzzy modeling of unreliable telecommunications networks. We hope that the problem of using different presentations of uncertainty in description of one unreliable network will attract attention of other researchers. Our initial theoretical results for converting fuzzy parameters and informal descriptions of algorithms on fuzzy models may be useful for automation of constructing fuzzy models of telecommunication and other networks.

Our next goal is comparing network elements' informativeness by using the "localization" method. We also want to check the advantages of switching between different types of fuzzy parameters on the examples of big networks with much fuzzy data represented in three or more ways.

VI. ACKNOWLEDGMENTS

This work is supported by the grant of the Program of basic researches of the Presidium of Russian Academy of Science.

REFERENCES

- [1] C. J. Colbourn, *The Combinatorics of Network Reliability*. New York, NY, USA: Oxford University Press, Inc., 1987.
- [2] L. Jereb, "Network reliability: models, measure and analysis," in *Proceedings of the 6th IFIP Workshop on Performance Modeling and Evaluation of ATM Networks*, 1998, pp. T02/1–T02/10.
- [3] B. Awerbuch and S. Event, "Reliable broadcast protocols in unreliable networks," *Networks*, vol. 16, no. 4, pp. 381–396, 1986. [Online]. Available: <http://dx.doi.org/10.1002/net.3230160405>
- [4] P. Rafiee and S. G. Latif, "Evaluating the reliability of communication networks wan using their fuzzy fault tree analysis – a case study," *The Journal of Mathematics and Computer Science (JMCS)*, vol. 2, no. 2, pp. 262–270, 2011. [Online]. Available: http://www.tjmcs.com/includes/files/articles/Vol2_Iss2_262
- [5] D. S. Jena and M. K. Deepthi, "Fuzzy reliability analysis in interconnection networks," *International Journal of Computational Engineering*, vol. 1, no. 1, pp. 262–270, 2011. [Online]. Available: www.ijceronline.com/papers/vol1_issue1/c0110115022.pdf
- [6] R. Sujatha and B. Praba, "Fuzzy reliability of a network using fuzzy sum of disjoint product technique," *International Journal of Recent Trends in Engineering*, vol. 2, no. 2, pp. 10–12, November 2009.

A Mathematical Model of Hierarchical Organization

Satoshi IKEDA
 Miyazaki University
 Computer Science and System Engineering
 1-1 Gakuen Kibanadai, Miyazaki, 889-2192
 Japan
 bisu@cs.miyazaki-u.ac.jp

Makoto SAKAMOTO
 Miyazaki University
 Computer Science and System Engineering
 1-1 Gakuen Kibanadai, Miyazaki, 889-2192
 Japan
 sakamoto@cs.miyazaki-u.ac.jp

Takao Ito
 University of Hiroshima
 Graduate School of Engineering
 1-4-1 kagamiyama, Higashi-Hiroshima,739-8527
 Japan
 itotakao@hiroshima-u.ac.jp

Abstract: Organization theory suggests that to streamline the management of a large organization begins by dividing it into several sections. This paper offers two or more evaluation measures that are required in order for an organization to specialize.

Key-Words: Combinatorial Optimization, Primary business, Peripheral, External, Internal, Profitability

1 Introduction

Two typical but different types of organization exist: one is classical organization with a pyramid shaped hierarchical structure, the other is a network organization with a non-hierarchical structure. Fig.1 shows an example of the classical type organization model which is expressed by a rooted tree. To attain an efficient operating organization, it is necessary to determine where the members should be assigned. This is a crucial issue in traditional organization theories that is relevant to “clarifying the limit of authority” and “layering of the organizations”.

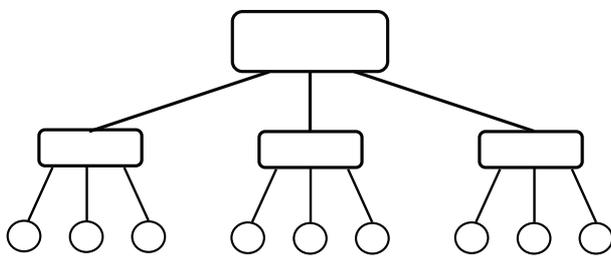


Figure 1: Hierarchical organization as a rooted tree.

In the organization sciences including business administration, economics, public administration, sociology, psychology, etc., the behavioral patterns and values are required in a specific organization have been discussed from the standpoint of the superiority in competing with others, or the possibility to succeed. For this reason, the problem of the organi-

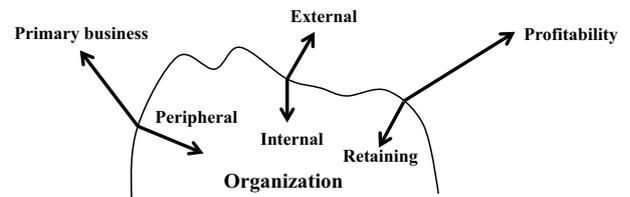


Figure 2: Three different behaviors in a modern large organization.

zations has primarily been preoccupied with qualitative discussions. Fig.2 illustrates three different categories of the members’ behaviors. They are primary-business-oriented behavior vs. peripheral-business-oriented behavior, external-contribution-oriented behavior vs. internal-contribution-oriented behavior, and profitability-oriented behavior vs. retaining-oriented behavior.

The primary trust of members is devoted to performing tasks indirectly related to the organizational primary business itself, such as the training members in public responsibility and compliance. As described in traditional organization theories, a legal remedy for compliance and organizational sustainability is required, which can be considered qualitatively. But in reality, the amount of effort to retain an organization is much higher than the amount of effort to perform its primary business. Thus, the primary task of an organization can be neglected. Accordingly deciding the ra-

tio of internal and external contributions is considered difficult and ambiguous. Internal contribution denotes the effort required to retain organizations and enhance organizational survival, whereas external contribution refers to the efforts to fulfill the mission of their own organization which perform against external. The reason for this ambiguity is that the problem cannot be dealt with extant quantitatively based frameworks of traditional organization theories.

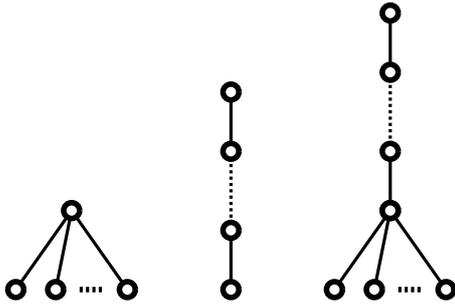


Figure 3: Three types of efficient trees when evaluation criteria is one only.

Recently, a mathematical model for evaluating the hierarchical organization quantitatively was proposed [[2, 3]]. This research demonstrated that the shape of the hierarchical organization that maximizes the evaluation value of organization, can be classified into three types (see Fig. 3) depending on the capacity value of the members when number of the evaluation criteria is only one. According to Fig.3, the hierarchical organization having only one evaluation criterion is an undifferentiated organization. In other words, it is not differentiated into several departments. This means that two or more evaluation criteria are required for an organization to specialize. In this paper, we propose a new mathematical model which defines internal and external contributions for the organizations. Consequently, the evaluation function in the new model is adapted by the sum of the contribution to the external of all members in a given organization. Here, the external contribution shows that members' behaviors directly increases and improves. However the internal contribution involved in retaining and the survival of firm is not directly related to the business of the organization originally in a hierarchical organization. In addition, the old model is compatible if the middle management in hierarchical organization uses their own capabilities in term of internal contribution.

2 Mathematical Model

Suppose that $G = (V(G), E(G))$ is a graph. Throughout this article, a graph is always finite, undi-

rected and simple, with order $n = |V(G)| (n \geq 2)$ and size $m = |E(G)|$.

For $u \in V(G)$, $G - U$ is obtained from G by deleting all the vertices in $V(G) \cap U$ and their incident edges. If $U = \{v\}$ is a singleton, we write $G - v$ rather than $G - \{v\}$. As above, $G - \{e\}$ and $G + \{e\}$ are abbreviated to $G - e$ and $G + e$ for $e \in E(G)$.

For $u \in V(G)$, by $N(u) = \{v \mid \{u, v\} \in E(G)\}$, we denote the set of vertices adjacent to u , and call $\deg(u) = |N(u)|$ the degree of $u \in V(G)$. We refer to a path in $G = (V(G), E(G))$ by the sequence of its vertices and write

$$G(x_0, x_k) = x_0x_1 \cdots x_k$$

for $x_i \in V(G) (i = 0, 1, \dots, k)$ and $x_jx_{j+1} \in E(G) (j = 0, 1, \dots, k - 1)$, where x_i are all distinct, and calling $G(x_0, x_k)$ a path from x_0 to x_k in G , and the number of edges of the path is its length. The above path $G(x_0, x_k)$ has length k .

Assume that $P = x_0x_1 \cdots x_{k-1}$ is a path and $k \geq 3$, then

$$C \equiv P + x_{k-1}x_0$$

is called a cycle. On the other hand, an acyclic graph, i.e., one not containing any cycles, is called a forest. A connected forest is called a tree. Thus, a forest is a graph whose components are trees. Sometimes we consider one vertex of a tree as special, and then such vertex is called the root of the tree, while the vertices of degree 1 in a tree, but not the root of the tree, are its leaves. A tree graph T with fixed root r is written as T_r , and then the set of T_r 's leaves is written as $L(T_r)$. That is,

$$L(T_r) = \{v \in E(T_r) \mid \deg(v) = 1, v \neq r\}.$$

Writing $x \preceq y$ for $x \in T_r(r, y)$, we then define a partial ordering on $V(T_r)$, the tree-order associated with T_r . This ordering will be considered as the expression "depth": if $x \prec y$, we say x lies below y in T_r , see Fig.4.

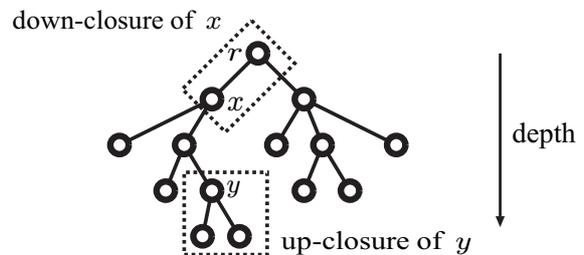


Figure 4: $x \prec y$ in T_r , down-closure of x , and up-closure of y .

We call

$$[x] \equiv \{v \in V(T_r) \mid v \preceq x\}$$

and

$$\lfloor y \rceil \equiv \{v \in V(T_r) \mid v \succeq y\}$$

the down-closure of x and the up-closure of y in T_r . Note that the root r is the least element, and that the leaves of T_r are its maximal elements in this partial order.

Suppose that $\Sigma = \{\sigma_1, \sigma_2, \dots, \sigma_n\} (n \geq 2)$ and $\mathcal{A} (\#\mathcal{A} \geq 1)$ are finite sets. Throughout this paper, Σ is interpreted as the set of members of a given organization, which consists of $\sigma_1, \sigma_2, \dots, \sigma_n$. And \mathcal{A} is the set of the evaluation measures.

For a given Σ , we call $(\Sigma, \{\phi_i\}_{i \in \mathcal{A}})$ an evaluation system, if

$$\phi_i : \Sigma \rightarrow \mathbb{R}^+ \equiv \{x \in \mathbb{R} \mid x > 0\} \quad \text{for } i \in \mathcal{A}.$$

We call $\phi_i(\sigma)$ the personal ability of $\sigma \in \Sigma$ with respect to an evaluation measure $i \in \mathcal{A}$.

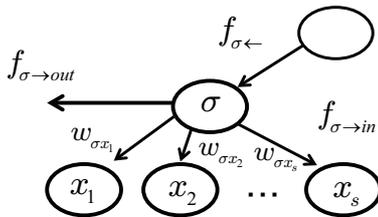


Figure 5: Input ($f_{\sigma \leftarrow}$) and output ($f_{\sigma \rightarrow in} + f_{\sigma \rightarrow out}$) for $\sigma \in \Sigma$.

In order for an organization to achieve its purpose to aim at, it is also necessary that appropriate instructions are transmitted to subordinates from superiors. Thus, for a fixed organization tree T_r with $V(T_r) = \Sigma$, we considered that the value of the output of σ written by $f_{\sigma \rightarrow}$, is determined as the interaction of “ability value of subordinate $\phi(\sigma)$ ” and “accuracy of instruction from superior $f_{\sigma \leftarrow}$ ”. In this paper, it is assumed that the total output $f_{\sigma \rightarrow}$ for $\sigma \in \Sigma$ become the value obtained by multiplying his own ability $\phi(\sigma)$ to his input $f_{\sigma \leftarrow}$ as the instruction from his superior. That is

$$f_{\sigma \rightarrow} = \phi(\sigma) \times f_{\sigma \leftarrow}.$$

Further, it is assumed that the output $f_{\sigma \rightarrow}$ for $\sigma \in \Sigma$ can be classified into the external output $f_{\sigma \rightarrow out}$ which works on the outside of an organization, and the internal output $f_{\sigma \rightarrow in}$ which works on the inside. The former $f_{\sigma \rightarrow out}$ is a direct effort to be intended to achieve the purpose of the organization by approach to the outside. On the other hand, the latter $f_{\sigma \rightarrow in}$ is an indirect effort to be intended to achieve the objectives of organization by assisting his own subordinates relevant to its maintenance and management. As a result, both of efforts contribute to the achievement of the purpose

that the organization aims at. However, in order to increase allover activities of the organization, the decision problem of whether to put a big weight on either external or internal output is difficult at the individual level of the members. Throughout this paper, the ratio of external output $f_{\sigma \rightarrow out}$ and internal output $f_{\sigma \rightarrow in}$ is assumed to be constant regardless of the members. That is

$$\text{external output} : \text{internal output} = \alpha : 1 - \alpha \quad (2.1)$$

for $\alpha \in [0, 1]$ and any $\sigma \in \Sigma$. We call α the external output coefficient and call $1 - \alpha$ the internal output coefficient.

For the subordinate $\sigma \in \Sigma$ who received instructions from his superior, it is necessary to transmit appropriate instructions $f_{\sigma \rightarrow in}$ to his own subordinates as superior, while σ as subordinate carries out the instructions $f_{\sigma \rightarrow out}$. For a given organizational structure tree T_r , we assume that the value of the input for subordinate $x \in N(\sigma)$ with $x \succ \sigma$ is obtained by multiplying its weight $w_{\sigma x}$ to $f_{\sigma \rightarrow in}$, see Fig.5. Therefore, the total contribution of $\sigma \in \Sigma$ for the organization can be expressed by

$$f_{\sigma \rightarrow in} \sum_{x \in N(\sigma), x \succ \sigma} w_{\sigma x} + f_{\sigma \rightarrow out},$$

where

$$0 \leq w_{\sigma x} \leq 1$$

for any $\sigma \notin L(T_r), x \in N(\sigma), x \succ \sigma$ and

$$1 \leq \sum_{x \in N(\sigma), x \succ \sigma} w_{\sigma x} \leq \text{deg}^* \sigma. \quad (2.2)$$

Here for an arbitrarily fixed rooted tree T_r with $V(T_r) = \Sigma$ and $\sigma \in \Sigma$,

$$\text{deg}^* \sigma = \begin{cases} \text{deg } \sigma & \text{if } \sigma \in \{r\} \cup L(T_r), \\ \text{deg } \sigma - 1 & \text{otherwise.} \end{cases}$$

Putting $\sum_{x \in N(\sigma), x \succ \sigma} w_{\sigma x} = 1$ in (2.2). This case

corresponds to the organization model which $\sigma \in \Sigma$ as superior instruct his/her subordinates individually. Since this instructor $\sigma \in \Sigma$ only assigns his/her total amount of instruction $f_{\sigma \rightarrow in}$ to his subordinates, the organizational management of this instructor is inefficient. On the other hand, this counter case,

$\sum_{x \in N(\sigma), x \succ \sigma} w_{\sigma x} = \text{deg}^* \sigma$ in (2.2) implies that

$$w_{\sigma x} = 1 \quad \text{for any } x \in N(\sigma), x \succ \sigma.$$

That is

$$f_{x\leftarrow} = f_{\sigma\rightarrow in} \quad \text{for any } x \in N(\sigma), x \succ \sigma.$$

This case corresponds to the organization model which $\sigma \in \Sigma$ as superior complete his/her indication to all subordinates with only one instruction. For example, the organization that intention transmission is performed only by meetings, this is true. In this case, since an internal output is proportional to the number of participants of meeting, the organizational management of this instructor is efficient. In this way,

$\sum_{x \in N(\sigma), x \succ \sigma} w_{\sigma x}$ means an efficient of organizational management of $\sigma \in \Sigma$.

To summarize the above, for a given evaluation system $(\Sigma, \{\phi_i\}_{i \in \mathcal{A}})$, a rooted tree T_r with $V(T_r) = \Sigma$, $\sigma \in \Sigma$, $i \in \mathcal{A}$ and an external output coefficient $\alpha \in [0, 1]$, $f_{\sigma\leftarrow}^{\alpha, i}$ and $f_{\sigma\rightarrow}^{\alpha, i}$ denote the input and the output of $\sigma \in \Sigma$, respectively. And $f_{\sigma\rightarrow out}^{\alpha, i}$ and $f_{\sigma\rightarrow in}^{\alpha, i}$ denote the external and the internal outputs of $\sigma \in \Sigma$ for the organization, respectively. Then we define the follows;

$$f_{\sigma\rightarrow}^{\alpha, i} = \phi_i(\sigma) f_{\sigma\leftarrow}^{\alpha, i}$$

and

$$f_{\sigma\leftarrow}^{\alpha, i} = \begin{cases} 1 & \text{if } \sigma = r \text{ (root),} \\ w_{p(\sigma)\sigma}^i f_{p(\sigma)\rightarrow in}^{\alpha, i} & \text{otherwise.} \end{cases}$$

Where $\{w_{xy}^i\}_{i \in \mathcal{A}, x \in \Sigma \setminus L(T_r), y \in N(x), y \succ x}$ denote the weights from $x \in \Sigma$ to $y \in N(x)$ ($y \succ x$) with respect to $i \in \mathcal{A}$ in an arbitrarily fixed T_r . And $p(\sigma)$ denotes the parent node (as superior) of $\sigma \in \Sigma$ on T_r .

Under the assumption (2.1), we define

$$f_{\sigma\rightarrow}^{\alpha, i} = f_{\sigma\rightarrow in}^{\alpha, i} + f_{\sigma\rightarrow out}^{\alpha, i}$$

$$f_{\sigma\rightarrow in}^{\alpha, i} = \begin{cases} 0 & \text{if } \sigma \in L(T_r), \\ (1 - \alpha) f_{\sigma\rightarrow}^{\alpha, i} & \text{otherwise,} \end{cases}$$

and

$$f_{\sigma\rightarrow out}^{\alpha, i} = \begin{cases} f_{\sigma\rightarrow}^{\alpha, i} & \text{if } \sigma \in L(T_r) \\ \alpha f_{\sigma\rightarrow}^{\alpha, i} & \text{otherwise.} \end{cases}$$

On a given rooted tree T_r , we define σ 's total contribution $F_{\sigma\rightarrow}^{\alpha, i}$ with respect to $i \in \mathcal{A}$ and $\alpha \in [0, 1]$ by

$$F_{\sigma\rightarrow}^{\alpha, i} = f_{\sigma\rightarrow in}^{\alpha, i} \sum_{x \in N(\sigma), x \succ \sigma} w_{\sigma x}^i + f_{\sigma\rightarrow out}^{\alpha, i}.$$

Here

$$0 \leq w_{\sigma x}^i \leq 1$$

for any $\sigma \in \Sigma \setminus L(T_r)$, $x \in N(\sigma)$, $x \succ \sigma$, $i \in \mathcal{A}$ and

$$1 \leq \sum_{x \in N(\sigma), x \succ \sigma} w_{\sigma x}^i \leq \deg \sigma^*.$$

Remark that $F_{\sigma\rightarrow}^{\alpha, i} = f_{\sigma\rightarrow out}^{\alpha, i}$ for $\sigma \in L(T_r)$, since $f_{\sigma\rightarrow in}^{\alpha, i} = 0$ for $\sigma \in L(T_r)$. Therefore, for convinience of defining $F_{\sigma\rightarrow}^{\alpha, i}$ for any $\sigma \in \Sigma$, we assume formally that

$$\sum_{x \in N(\sigma), x \succ \sigma} w_{\sigma x}^i = 1 \quad \text{for } \sigma \in L(T_r).$$

Then, for any $\sigma \in \Sigma$ and $i \in \mathcal{A}$, we see that

$$f_{\sigma\rightarrow}^{\alpha, i} \leq F_{\sigma\rightarrow}^{\alpha, i}$$

in general and the equality is attained only in the case of

$$\sum_{x \in N(\sigma), x \succ \sigma} w_{\sigma x}^i = 1 \quad \text{for any } \sigma \in \Sigma.$$

In this way, this total contribution value $F_{\sigma\rightarrow}^{\alpha, i}$ depends on the value of $\sum_{x \in N(\sigma), x \succ \sigma} w_{\sigma x}^i$. Therefore, we call

$\sum_{x \in N(\sigma), x \succ \sigma} w_{\sigma x}^i$ the efficiency coefficient of $\sigma \in \Sigma$ with respect to $i \in \mathcal{A}$.

Throughout this paper, we assume that $\{w_{\sigma x}^i\}_{x \in N(\sigma), x \succ \sigma}$ for $\sigma \in \Sigma \setminus L(T_r)$ is a sequence depending only on $\deg^* \sigma$ and $i \in \mathcal{A}$. That is, for $\sigma, \sigma' \in \Sigma \setminus L(T_r)$,

$$\deg^* \sigma = \deg^* \sigma'$$

implies that $(w_{\sigma'x}^i)_{x \in N(\sigma'), x \succ \sigma'}$ is a permutation of $(w_{\sigma x}^i)_{x \in N(\sigma), x \succ \sigma}$. Thus, for a fixed $\deg^* \sigma$ and $i \in \mathcal{A}$, the selection that we can do is which weight to assign whom. We call the way of determination of a weights' policy. For any weights' policy, we assume that if $\deg^* \sigma \geq \deg^* \sigma'$ for $\sigma, \sigma' \in \Sigma$,

$$w_{\sigma x_1}^i \geq w_{\sigma x_2}^i \geq \dots \geq w_{\sigma x_{\deg^* \sigma}}^i$$

and

$$w_{\sigma'x'_1}^i \geq w_{\sigma'x'_2}^i \geq \dots \geq w_{\sigma'x'_{\deg^* \sigma'}}^i,$$

then $(w_{\sigma x}^i)_{x \in N(\sigma), x \succ \sigma}$ and $(w_{\sigma'x'}^i)_{x' \in N(\sigma'), x' \succ \sigma'}$ satisfy

$$w_{\sigma x_j}^i \leq w_{\sigma'x'_j}^i \quad \text{for } j = 1, 2, \dots, \deg^* \sigma'.$$

For a given evaluation system $(\Sigma, \{\phi_i\}_{i \in \mathcal{A}})$, an external output coefficient $\alpha \in [0, 1]$ and a weights'

policy, let T_r be a rooted tree graph with $V(T_r) = \Sigma$. Then we will evaluate the rooted tree T_r as organization model by

$$\Phi^{(\alpha)}(T_r) = \max_{\{w_{xy}^i\}} \sum_{i \in \mathcal{A}} \sum_{\sigma \in \Sigma} f_{\sigma \rightarrow out}^{\alpha, i}$$

Here $\{w_{xy}^i\}$ shall be taken about all the possible combinations under the given weights' policy. We call $\Phi^{(\alpha)}(T_r)$ the ability value of T_r with respect to $(\Sigma, \{\phi_i\}_{i \in \mathcal{A}})$ and $\alpha \in [0, 1]$. Under a given weights' policy, we say that T_r is an efficient tree for a given external outpt coefficient $\alpha \in [0, 1]$, if $\max_{T \in \mathcal{T}} \Phi^{(\alpha)}(T)$ is attained by $\Phi^{(\alpha)}(T_r)$. Here \mathcal{T} denotes the set of rooted tree graph with $V(T) = \Sigma$.

One of our interest is to find the efficient organization structure tree T_r for fixed the external outpt coefficient $\alpha \in [0, 1]$, or to find a pair of α and $T_r \in \mathcal{T}$ which maximize its ability value. However, by the definition, if $\alpha = 1$ then we see that

$$\Phi^{(1)}(T_r) = \sum_{i \in \mathcal{A}} f_{r \rightarrow out}^{1, i}$$

for any weights' policy. That is, when $\alpha = 1$, there are obvious efficient trees only. Therefore, throughout this paper, we assume that $\alpha \in [0, 1)$.

3 Suitability of Hierarchical Model

Firstly, we will show that this hierarchical model has a suitability for a special case of $\sharp \mathcal{A} = 1$.

Theorem 1 *Under a given weights' policy, suppose that T_r is an efficient tree for a given $(\Sigma, \{\phi(\sigma)\})$ and a given $\alpha \in [0, 1)$. Then we see that $x \prec y$ for $x, y \in \Sigma$ implies $\phi(x) \geq \phi(y)$.*

Proof of theorem 1 For a given $(\Sigma, \{\phi\})$ and $\alpha \in [0, 1)$, let T_r be an efficient tree under a given weights' policy. Assume that $x \prec y$ in T_r and that T'_r is the tree by interchanging x and y in T_r . Then we get

$$\begin{aligned} & \Phi(T_r) - \Phi(T'_r) \\ &= \sum_{\ell \succeq x \text{ on } T_r} f_{\ell \rightarrow out}^{\alpha} - \sum_{\ell' \succeq y \text{ on } T'_r} f_{\ell' \rightarrow out}^{\alpha} \\ &= \left(1 - \frac{\phi(y)}{\phi(x)}\right) \sum_{\ell \succeq x, \ell \not\succeq y \text{ on } T_r} f_{\ell \rightarrow out}^{\alpha} \end{aligned}$$

Since we see that $\Phi(T_r) - \Phi(T'_r) \geq 0$ by the assumption, thus we get

$$1 - \frac{\phi(y)}{\phi(x)} \geq 0,$$

which implies $\phi(x) \geq \phi(y)$. \square

Theorem 1 is intended to satisfy the most fundamental image that we are holding about an organization. In this sense, our efficient trees are suitable for a hierarchical model. Theorem 1 can be slightly modified as follows:

Corollary 2 *Under a given weights' policy, suppose that T_r is an efficient tree for a given $(\Sigma, \{\phi_i\}_{i \in \mathcal{A}})$ and a given $\alpha \in [0, 1)$. Then we see that*

$$\phi_i(x) \geq \phi_i(y) \quad \text{for any } i \in \mathcal{A}$$

implies $x \not\prec y$ in T_r .

4 Examples

Let us set $\Sigma = \{1, 2, 3, 4\}$, $\mathcal{A} = \{a, b\}$ and put

$$\begin{aligned} \phi_a(1) = \phi_a(2) = 4, & \quad \phi_a(3) = \phi_a(4) = 1/2, \\ \phi_b(1) = \phi_b(2) = 1/2, & \quad \phi_b(3) = \phi_b(4) = 4. \end{aligned}$$

For an evaluation system $(\Sigma, \{\phi_i\}_{i \in \mathcal{A}})$ described above, we consider two settings that only those weights' policies differs from with each other. The first setting is

$$w_{\sigma x}^i = 1$$

for any $\sigma \in \Sigma \setminus L(T_r)$, $x \in N(\sigma)$, $x \succ \sigma$, $i \in \mathcal{A}$, which implies the efficiency coefficient of σ equals to $\deg^* \sigma$. That is, $\sum_{x \in N(\sigma), x \succ \sigma} w_{\sigma x}^i = \deg^* \sigma$ for any $\sigma \in \Sigma$ and $i \in \mathcal{A}$. Then we see that

an efficient tree is $\begin{cases} \text{(a) in Fig.6} & \text{if } 0 \leq \alpha \leq \frac{1}{5}, \\ \text{(b) in Fig.6} & \text{otherwise.} \end{cases}$

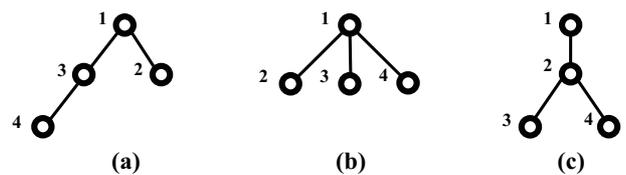


Figure 6: Various efficient trees.

The second setting is

$$w_{\sigma x}^i = \frac{1}{\deg^* \sigma}$$

for any $\sigma \in \Sigma \setminus L(T_r)$, $x \in N(\sigma)$, $x \succ \sigma$, $i \in \mathcal{A}$, which implies the efficiency coefficient of $\sigma \in \Sigma$ equals to 1. That is, $\sum_{x \in N(\sigma), x \succ \sigma} w_{\sigma x}^i = 1$ for any $\sigma \in \Sigma$ and $i \in \mathcal{A}$. Then we see that

an efficient tree is $\begin{cases} \text{(b) in Fig.6} & \text{if } 0 \leq \alpha \leq \frac{29}{78}, \\ \text{(c) in Fig.6} & \text{otherwise.} \end{cases}$

5 Preliminary Results

As previously noted, organization theory posits that in order to streamline the management of a large organization, it should be divided into several sections, as shown in Fig.7. When we call a certain group *the section* of its organization, two or more sections must exist in the organization, and at least one of those sections must contain two or more members. For example, we consider that (b) and (c) in Fig.6 are not departmentalized in this sense.

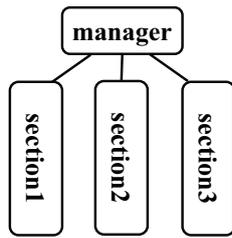


Figure 7: Departmentalized organization structure.

Intuitively, if an organization that has several sections, as depicted in Fig.7 is effective, it is thought that each section plays with their own role. In other words, two or more evaluation measures must be required in order for an organization to specialize. In fact, under some special assumptions when the number of evaluation measures $\#A = 1$, we have proved that the shape of the hierarchical organization that maximizes the evaluation value of organization, can be classified into three types depending on the personal ability values of the members, see Fig.8. We found that three types of appearing herein consist of one section fundamentally, therefore these are not departmentalized.

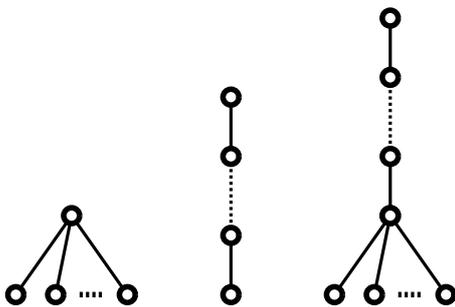


Figure 8: Three types of efficient trees when evaluation criteria is only one.

In this section, we will introduce the outline of preliminary results of the above-mentioned. Through out this section, we assume that the number of evaluation measures $\#A = 1$, the external output coefficient $\alpha = 0$ and the efficiency coefficient of $\sigma \in \Sigma$ is equal

to $\text{deg}^* \sigma$ for any $\sigma \in \Sigma$. That is, its weights' policy is the following.

$$w_{\sigma x} = 1 \quad \text{for any } \sigma \in \Sigma \setminus L(T_r), x \in N(\sigma), x \succ \sigma. \tag{5.1}$$

Note that under this weights' policy, we get

$$\Phi^{(0)}(T_r) = \sum_{i \in A} \sum_{\ell \in L(T_r)} \prod_{v \in T_r(r, \ell)} \phi(v).$$

For a given (Σ, ϕ) and its organizational structure tree T_r , let us define

$$\text{DEG}_2(T_r) = \{\sigma \in \Sigma \mid \text{deg}^*(\sigma) \geq 2\}.$$

The person who manages several sections on T_r directly always needs to be a element of $\text{DEG}_2(T_r)$. When the number of evaluation measure $\#A$ is one, we claim that this set $\text{DEG}_2(T_r)$ has few elements and that such organization cannot configure any sections.

Theorem 3 (Ikeda et.al.[2, 3]) *Under the weights' policy (5.1) and the extenal output coefficient $\alpha = 0$, assume that T_r is an efficient tree for a given (Σ, ϕ) . Then we see the followings.*

- (a) $\# \text{DEG}_2(T_r)$ is equal to 0 or 1.
- (b) Putting $\text{DEG}_2(T_r) = \{x\}$ when $\# \text{DEG}_2(T_r) = 1$, then we see that

$$\{y \in \Sigma \mid y \succ x \text{ on } T_r\} = L(T_r).$$

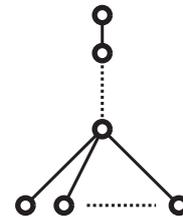


Figure 9: General form of an efficient tree with $\#A = 1$.

The typical form of an efficient tree is a path graph and a star graph. The general form of the most efficient tree which theorem 3 stressed is shown in Fig.9. We found that the upper-half is a path graph, and the lower-half is the star graph. Since these most efficient trees consist of only one section, see the beginning of this chapter, thus we found that they are not departmentalized.

6 Main Results

In this section, some necessary conditions that non-departmentalized organization is optimal will be discussed. Through-out this section, assume that the number of evaluation measures $\#\mathcal{A} = 1$.

Firstly, we will examine what kind of situation would be better will be examined if the structural tree of the organization branch. For a given (Σ, ϕ) , assume that

$$\begin{aligned} \Sigma &= \{\sigma_1, \sigma_2, \dots, \sigma_n\}, \\ S &= \{s_0, s_1, \dots, s_\ell, s_{\ell+1}, \dots, s_{\ell+m}\} \subseteq \Sigma \\ \phi(\sigma_1) &\geq \phi(\sigma_2) \geq \dots \geq \phi(\sigma_n), \end{aligned}$$

and

$$\phi(s_0) \geq \dots \phi(s_\ell) \geq \phi(s_{\ell+1}) \geq \dots \geq \phi(s_{\ell+m}).$$

Let us $\tilde{S} = \{\tilde{s}_0, \tilde{s}_1, \dots, \tilde{s}_\ell, \tilde{s}_{\ell+1}, \dots, \tilde{s}_{\ell+m}\}$ be a permutation of S associated with bijection map $\pi : \tilde{S} \rightarrow S$. That is, $s = \pi(\tilde{s}) \in S$ for $\tilde{s} \in \tilde{S}$ and $S = \tilde{S}$. We assume that

$$\begin{aligned} \phi(\tilde{s}_0) &\geq \phi(\tilde{s}_1) \geq \dots \geq \phi(\tilde{s}_\ell), \\ \phi(\tilde{s}_0) &\geq \phi(\tilde{s}_{\ell+1}) \geq \dots \geq \phi(\tilde{s}_{\ell+m}). \end{aligned}$$

Note that $\tilde{s}_0 = \pi(\tilde{s}_0) = s_0$ by the assumption. Let us set two organizational structure trees T_{s_0} and \tilde{T}_{s_0} as shown in Fig.10.

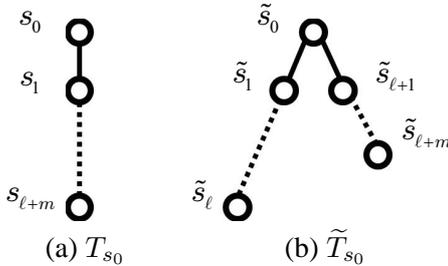


Figure 10: Two organizational structure trees.

For two organizational structure trees T_1 and T_2 with $V(T_1) = V(T_2)$, let us define

$$\Delta_{T_1}^{T_2} f_{\sigma \rightarrow out}^{\alpha, i} = f_{\sigma \in T_2 \rightarrow out}^{\alpha, i} - f_{\sigma \in T_1 \rightarrow out}^{\alpha, i},$$

then we have the following for above T_{s_0} and \tilde{T}_{s_0} .

Lemma 4 For a given $(\Sigma, \{\phi\})$, an external output coefficient $\alpha \in [0, 1)$ and a weights' policy, we assume that

$$\phi(\sigma_n) \geq \frac{1}{1 - \alpha}$$

and

$$\Delta_{\tilde{T}_{s_0}}^{T_{s_0}} \left(f_{\tilde{s}_\ell \rightarrow out}^\alpha + f_{\tilde{s}_{\ell+m} \rightarrow out}^\alpha \right) \geq 0.$$

Then we see that

$$\Phi^{(\alpha)}(T_{s_0}) \geq \Phi^{(\alpha)}(\tilde{T}_{s_0}).$$

Proof of lemma4 Put

$$\tilde{\phi}(x) = (1 - \alpha)\phi(x) \quad \text{for } x \in \Sigma.$$

Then, by the assumption we get

$$\tilde{\phi}(x) \geq 1 \quad \text{for any } x \in \Sigma.$$

Together with $w_{s_0 \tilde{s}_1} \leq 1$ and

$$\{s_1, s_2, \dots, \pi^{-1}(\tilde{s}_i)\} \supseteq \{\tilde{s}_1, \tilde{s}_2, \dots, \tilde{s}_i\}$$

for $i \in \{1, 2, \dots, \ell - 1\}$, we see that

$$\begin{aligned} &\Delta_{\tilde{T}_{s_0}}^{T_{s_0}} f_{\tilde{s}_i \rightarrow out}^\alpha \\ &= \alpha \phi(s_0) \left[\prod_{\sigma \neq s_0: \sigma \preceq \tilde{s}_i \text{ in } T_{s_0}} \tilde{\phi}(\sigma) - w_{s_0 \tilde{s}_1} \left(\prod_{\sigma \neq s_0: \sigma \preceq \tilde{s}_i \text{ in } \tilde{T}_{s_0}} \tilde{\phi}(\sigma) \right) \right] \\ &\geq 0 \end{aligned}$$

for $i \in \{1, 2, \dots, \ell - 1\}$, which implies

$$\sum_{i=1}^{\ell-1} \Delta_{\tilde{T}_{s_0}}^{T_{s_0}} f_{\tilde{s}_i \rightarrow out}^\alpha \geq 0.$$

Together with $\Delta_{\tilde{T}_{s_0}}^{T_{s_0}} f_{\tilde{s}_0}^\alpha = 0$, we have

$$\sum_{i=0}^{\ell-1} \Delta_{\tilde{T}_{s_0}}^{T_{s_0}} f_{\tilde{s}_i \rightarrow out}^\alpha \geq 0.$$

Similarly, we have $\sum_{i=\ell+1}^{\ell+m-1} \Delta_{\tilde{T}_{s_0}}^{T_{s_0}} f_{\tilde{s}_i \rightarrow out}^\alpha \geq 0$. Since

$$\begin{aligned} &\Phi^{(\alpha)}(T_{s_0}) - \Phi^{(\alpha)}(\tilde{T}_{s_0}) \\ &= \sum_{i=0}^{\ell-1} \Delta_{\tilde{T}_{s_0}}^{T_{s_0}} f_{\tilde{s}_i \rightarrow out}^\alpha + \sum_{i=\ell+1}^{\ell+m-1} \Delta_{\tilde{T}_{s_0}}^{T_{s_0}} f_{\tilde{s}_i \rightarrow out}^\alpha \\ &\quad + \Delta_{\tilde{T}_{s_0}}^{T_{s_0}} \left(f_{\tilde{s}_\ell \rightarrow out}^\alpha + f_{\tilde{s}_{\ell+m} \rightarrow out}^\alpha \right), \end{aligned}$$

thus, we get

$$\Phi^{(\alpha)}(T_{s_0}) - \Phi^{(\alpha)}(\tilde{T}_{s_0}) \geq 0$$

if $\Delta_{\tilde{T}_{s_0}}^{T_{s_0}} \left(f_{\tilde{s}_\ell \rightarrow out}^\alpha + f_{\tilde{s}_{\ell+m} \rightarrow out}^\alpha \right) \geq 0$ holds. \square

Lemma 5 Suppose that T_{s_0}, \tilde{T}_{s_0} and S are the same as lemma1. Then we have the followings.

(a) For any $\alpha \in [0, 1)$, if $\phi(\sigma_n) \geq \frac{2}{1-\alpha}$ then $\Phi^{(\alpha)}(T_{s_0}) \geq \Phi^{(\alpha)}(\tilde{T}_{s_0})$ holds.

(b) For any $\alpha \in \left[\frac{3-\sqrt{5}}{2}, 1\right)$, if $\phi(\sigma_n) \geq \frac{1}{(1-\alpha)^2}$ then $\Phi^{(\alpha)}(T_{s_0}) \geq \Phi^{(\alpha)}(\tilde{T}_{s_0})$ holds.

Remark 6 $\frac{3-\sqrt{5}}{2} \approx 0.382$. When $\alpha = \frac{3-\sqrt{5}}{2}$ then $\frac{1}{1-\alpha} = \frac{1+\sqrt{5}}{2}$ (golden number).

Remark 7 For $\alpha \in \left[\frac{1}{2}, 1\right)$, we see that $\frac{1}{(1-\alpha)^2} \leq \frac{1}{(1-\alpha)(1-1/2)} = \frac{2}{1-\alpha}$.

Proof of lemma5 By lemma4, we have only to show $\Delta_{\tilde{T}_{s_0}}^{T_{s_0}}(f_{\tilde{s}_\ell \rightarrow out}^\alpha + f_{\tilde{s}_{\ell+m} \rightarrow out}^\alpha) \geq 0$.

Without loss of generality, we assume that $\phi(\tilde{s}_\ell) \geq \phi(\tilde{s}_{\ell+m})$. That is $\tilde{s}_{\ell+m} = s_{\ell+m}$. Then we see

$$\begin{aligned} & \Delta_{\tilde{T}_{s_0}}^{T_{s_0}}(f_{\tilde{s}_\ell \rightarrow out}^\alpha + f_{\tilde{s}_{\ell+m} \rightarrow out}^\alpha) \\ &= \alpha\phi(s_0)\tilde{\phi}(s_1)\cdots\tilde{\phi}(\tilde{s}_\ell) + \phi(s_0)\tilde{\phi}(s_1)\cdots\tilde{\phi}(\tilde{s}_{\ell+m}) \\ & \quad - w_{s_0\tilde{s}_1}\phi(s_0)\tilde{\phi}(\tilde{s}_1)\cdots\tilde{\phi}(\tilde{s}_\ell) \\ & \quad - w_{s_0\tilde{s}_{\ell+1}}\phi(s_0)\tilde{\phi}(\tilde{s}_{\ell+1})\cdots\tilde{\phi}(\tilde{s}_{\ell+m}). \end{aligned}$$

Therefore we get

$$\begin{aligned} & \Delta_{\tilde{T}_{s_0}}^{T_{s_0}}(f_{\tilde{s}_\ell \rightarrow out}^\alpha + f_{\tilde{s}_{\ell+m} \rightarrow out}^\alpha) \\ & \geq \phi(s_0) \left[\alpha\tilde{\phi}(s_1)\tilde{\phi}(s_2)\cdots\tilde{\phi}(\tilde{s}_\ell) \right. \\ & \quad + \left(\prod_{i=1}^{\ell} \tilde{\phi}(\tilde{s}_i) - w_{s_0\tilde{s}_1} \right) \left(\prod_{i=1}^m \tilde{\phi}(\tilde{s}_{\ell+i}) - w_{s_0\tilde{s}_{\ell+1}} \right) \\ & \quad \left. - w_{s_0\tilde{s}_1} w_{s_0\tilde{s}_{\ell+1}} \right]. \end{aligned}$$

Firstly, we assume $\phi(s_n) \geq \frac{2}{1-\alpha}$, then $\tilde{\phi}(x) \geq 2$ holds for $x \in \Sigma$. Since $w_{s_0\tilde{s}_1} w_{s_0\tilde{s}_{\ell+1}} \leq 1$, we have

$$\begin{aligned} & \left(\prod_{i=1}^{\ell} \tilde{\phi}(\tilde{s}_i) - w_{s_0\tilde{s}_1} \right) \left(\prod_{i=1}^m \tilde{\phi}(\tilde{s}_{\ell+i}) - w_{s_0\tilde{s}_{\ell+1}} \right) \\ & \quad - w_{s_0\tilde{s}_1} w_{s_0\tilde{s}_{\ell+1}} \geq 0, \end{aligned}$$

which implies

$$\begin{aligned} & \Delta_{\tilde{T}_{s_0}}^{T_{s_0}}(f_{\tilde{s}_\ell \rightarrow out}^\alpha + f_{\tilde{s}_{\ell+m} \rightarrow out}^\alpha) \\ & \geq \alpha\phi(s_0)\tilde{\phi}(s_1)\tilde{\phi}(s_2)\cdots\tilde{\phi}(\tilde{s}_\ell) \geq 0. \end{aligned}$$

Thus, we get lemma5 (a).

Next, we assume $\phi(s_n) \geq \frac{1}{(1-\alpha)^2}$, then $\tilde{\phi}(x) \geq \frac{1}{1-\alpha}$ holds for $x \in \Sigma$. Thus, we see that

$$\begin{aligned} & \Delta_{\tilde{T}_{s_0}}^{T_{s_0}}(f_{\tilde{s}_\ell \rightarrow out}^\alpha + f_{\tilde{s}_{\ell+m} \rightarrow out}^\alpha) \\ &= \alpha\phi(s_0)\tilde{\phi}(s_1)\cdots\tilde{\phi}(\tilde{s}_\ell) \\ & \quad + \phi(s_0)\tilde{\phi}(s_1)\cdots\tilde{\phi}(\tilde{s}_{\ell+m}) \\ & \quad - w_{s_0\tilde{s}_1}\phi(s_0)\tilde{\phi}(\tilde{s}_1)\cdots\tilde{\phi}(\tilde{s}_\ell) \\ & \quad - w_{s_0\tilde{s}_{\ell+1}}\phi(s_0)\tilde{\phi}(\tilde{s}_{\ell+1})\cdots\tilde{\phi}(\tilde{s}_{\ell+m}) \\ & \geq \frac{\alpha}{1-\alpha}\phi(s_0)\tilde{\phi}(s_1)\cdots\tilde{\phi}(\tilde{s}_\ell) \\ & \quad + \phi(s_0)\tilde{\phi}(s_1)\cdots\tilde{\phi}(\tilde{s}_\ell)\tilde{\phi}(\tilde{s}_{\ell+1})\cdots\tilde{\phi}(\tilde{s}_{\ell+m}) \\ & \quad - w_{s_0\tilde{s}_1}\phi(s_0)\tilde{\phi}(\tilde{s}_1)\cdots\tilde{\phi}(\tilde{s}_\ell) \\ & \quad - w_{s_0\tilde{s}_{\ell+1}}\phi(s_0)\tilde{\phi}(\tilde{s}_{\ell+1})\cdots\tilde{\phi}(\tilde{s}_{\ell+m}) \\ & \geq \phi(s_0) \left[\frac{\alpha}{(1-\alpha)^2} - w_{s_0\tilde{s}_1} w_{s_0\tilde{s}_{\ell+1}} \right. \\ & \quad \left. + \left(\prod_{i=1}^{\ell} \tilde{\phi}(\tilde{s}_i) - w_{s_0\tilde{s}_1} \right) \left(\prod_{i=1}^m \tilde{\phi}(\tilde{s}_{\ell+i}) - w_{s_0\tilde{s}_{\ell+1}} \right) \right]. \end{aligned}$$

Since $w_{s_0\tilde{s}_1} w_{s_0\tilde{s}_{\ell+1}} \leq 1$ and

$$\left(\prod_{i=1}^{\ell} \tilde{\phi}(\tilde{s}_i) - w_{s_0\tilde{s}_1} \right) \left(\prod_{i=1}^m \tilde{\phi}(\tilde{s}_{\ell+i}) - w_{s_0\tilde{s}_{\ell+1}} \right) \geq 0,$$

we get

$$\Delta_{\tilde{T}_{s_0}}^{T_{s_0}}(f_{\tilde{s}_\ell \rightarrow out}^\alpha + f_{\tilde{s}_{\ell+m} \rightarrow out}^\alpha) \geq \phi(s_0) \left(\frac{\alpha}{(1-\alpha)^2} - 1 \right),$$

which implies lemma5 (b). \square

We will be extended lemma5 to the case of $\deg^* s_0 \geq 3$. Suppose that T_r is a rooted tree with $V(T_r) = \Sigma$. We assume that

$$\deg^* x = 1 \quad \text{for any } x \in \Sigma, x \succ s_0.$$

And that for $k \geq 3$ and $i \in \{0, 1, 2, \dots, k-1\}$,

$$\{s \in N(s_0) | s \succ s_0\} = \{\tilde{s}_1, \tilde{s}_{\ell_1+1}, \tilde{s}_{\ell_2+1}, \dots, \tilde{s}_{\ell_{k-1}+1}\},$$

$$\phi(\tilde{s}_{\ell_i+1}) \geq \phi(\tilde{s}_{\ell_i+2}) \geq \dots \geq \phi(\tilde{s}_{\ell_i+1}) \quad (\text{where } \ell_0 = 0),$$

$$w_{s_0 \tilde{s}_1} \geq w_{s_0 \tilde{s}_2} \geq \dots \geq w_{s_0 \tilde{s}_k},$$

$$S = \{x \in \Sigma | x \succeq \tilde{s}_1, x \succeq \tilde{s}_2\} \cup \{s_0\}$$

$$= \{s_0, s_1, \dots, s_{\ell_1}, \dots, s_{\ell_1+2}, \dots, s_{\ell_2}\},$$

and

$$\phi(s_0) \geq \phi(s_1) \geq \dots \geq \phi(s_{\ell_1})$$

$$\geq \phi(s_{\ell_1+1}) \geq \dots \geq \phi(s_{\ell_2}).$$

Let us put T_{s_0} and \tilde{T}_{s_0} as Fig.11;

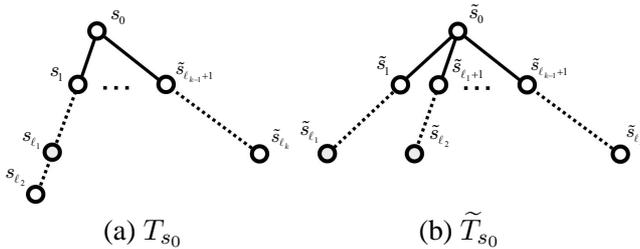


Figure 11: Two organizational structure trees.

Then, by the assumptions of weights' policy,

$$w_{s_0 s_1} (\text{in } T_{s_0}) \geq w_{s_0 \tilde{s}_{\ell_1+1}} (\text{in } \tilde{T}_{s_0}) \geq w_{s_0 \tilde{s}_1} (\text{in } \tilde{T}_{s_0})$$

and for $i \in \{3, 4, \dots, k\}$

$$w_{s_0 s_{\ell_i+1}} (\text{in } T_{s_0}) \geq w_{s_0 s_{\ell_i+1}} (\text{in } \tilde{T}_{s_0}).$$

Thus, we have lemma8.

Lemma 8 For T_{s_0} and \tilde{T}_{s_0} in Fig.11, we have the followings.

(a) For any $\alpha \in [0, 1)$, if $\phi(\sigma_n) \geq \frac{2}{1-\alpha}$, then $\Phi^{(\alpha)}(T_{s_0}) \geq \Phi^{(\alpha)}(\tilde{T}_{s_0})$ holds.

(b) For any $\alpha \in \left[\frac{3-\sqrt{5}}{2}, 1\right)$, if $\phi(\sigma_n) \geq \frac{1}{1-\alpha}$, then $\Phi^{(\alpha)}(T_{s_0}) \geq \Phi^{(\alpha)}(\tilde{T}_{s_0})$ holds.

By using lemma 8 in order from the vertex closer to the leaves repeatedly, we obtain the followings.

Theorem 9 For a given evaluation system $(\Sigma, \{\phi\})$, a given weights' policy, suppose that $\alpha \in [0, 1)$ denotes the external output coefficient and that $\Sigma = \{\sigma_1, \sigma_2, \dots, \sigma_n\}$. Assume that

$$\phi(\sigma_1) \geq \phi(\sigma_2) \geq \dots \geq \phi(\sigma_n) \geq \frac{2}{1-\alpha}$$

hold, then we see that an efficient tree is the path graph in Fig.12.

Theorem 10 For a given evaluation system $(\Sigma, \{\phi\})$, a given weights' policy, suppose that $\alpha \in [0, 1)$ denotes the external output coefficient and that $\Sigma = \{\sigma_1, \sigma_2, \dots, \sigma_n\}$. Assume that $\alpha \in \left[\frac{3-\sqrt{5}}{2}, 1\right)$ and

$$\phi(\sigma_1) \geq \phi(\sigma_2) \geq \dots \geq \phi(\sigma_n) \geq \frac{1}{(1-\alpha)^2}$$

hold, then we see that an efficient tree is the path graph in Fig.12.

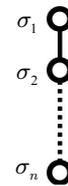


Figure 12: An efficient organization structure as a path graph.

A tree graph as organization structure is not a substantial organization. Thus, when the evaluation measure is one, theorem 9 and theorem 10 show that an organization with a hierarchical structure is not suitable for members with high ability and that the external output coefficient α is large means the path type of organization.

For example, in the organization which considers individual achievements as important, its members should take the action such as increasing their external output coefficients. Therefore in order for the organization which considers individual achievements as important to maintain the organization of a hierarchical type, it is necessary to multiple evaluation measures.

Theorem 11 For a given evaluation system $(\Sigma, \{\phi\})$, a given weights' policy, suppose that T_r is the path graph in Fig.12. Assume that

$$\phi(\sigma) \geq 1 \quad \text{for } \sigma \in \Sigma = \{\sigma_1, \sigma_2, \dots, \sigma_n\} \quad (n \geq 3).$$

Then we see that

$$\Phi^{(\beta)}(T_r) \geq \Phi^{(\alpha)}(T_r) \quad \text{for } \alpha > \beta.$$

By theorem 11, we have the following directly.

Corollary 12 Suppose that T_r is the path graph in Fig.12, then we see that

$$\Phi^{(0)}(T_r) \geq \Phi^{(\alpha)}(T_r) \quad \text{for any } \alpha \in [0, 1).$$

Proof of theorem11 For $\alpha > \beta$, put

$$\Delta_k = (\alpha - \beta)\phi(\sigma_k)f_{\sigma_k \leftarrow}^\alpha \quad \text{for } k \in \{1, 2, \dots, n-1\},$$

then we see that

$$\begin{aligned} & \Phi^{(\beta)}(T_r) - \Phi^{(\alpha)}(T_r) \\ &= \sum_{i=1}^{n-1} \Delta_i \left[\sum_{j=0}^{n-i-2} \beta(1-\beta)^j \prod_{\ell=i+1}^{i+j+1} \phi(\sigma_\ell) \right. \\ & \quad \left. + (1-\beta)^{n-i-1} \prod_{\ell=i+1}^n \phi(\sigma_\ell) - 1 \right]. \end{aligned}$$

Since $\phi(\sigma) \geq 1$ for any $\sigma \in \Sigma$ by the assumption, we get

$$\begin{aligned} & \Phi^{(\beta)}(T_r) - \Phi^{(\alpha)}(T_r) \\ & \geq \sum_{i=1}^{n-1} \Delta_i \left[\sum_{j=0}^{n-i-2} \beta(1-\beta)^j + (1-\beta)^{n-i-1} - 1 \right] \\ & = 0. \quad \square \end{aligned}$$

Theorem 11 and corollary 12 imply that the most efficient external output coefficient for members with high ability in the path type organization is zero. That is, an individual's action in the organization which thinks individual achievements as important is contradictory to making the activity of the whole organization into the maximum.

Acknowledgment: This work was supported by JSPS KAKENHI Grant Number 24510217.

References:

- [1] Alvesson Mats, "A Flat Pyramid: A Symbolic Processing of Organizational Structure", International Studies of Management and Organization, 19-4,5-23,1989/1990
- [2] S.Ikeda, "An Efficient Structure of Organization", International Interdisciplinary Workshop on Robotics, Ecosystem, and Management,pp.72-78, 2010
- [3] S.Ikeda, T.Ito and M.Sakamoto, "Discovering the efficient organization structure: horizontal versus vertical", Artificial Life and Robotics, vol.15, 4, pp.478-481, 2011
- [4] A.Brown, "Organization of Industry",1947
- [5] Krackhardt David and Jeffrey R. Hanson, "Informal Networks: The company Behind the Chart", Harvard Business review,71-4,104-111,1993
- [6] P.R.Lawrence and J.W Lorsch "Organization and environment: Managing defferentiation and integration", Harvard University, Boston M.A., 1967

[7] Reinhard Diestel,"Graph Theory(Third Edition)",Springer,2000

Advance Trends of Hybrid Electric Vehicles



Shahram Javadi

Electrical Engineering Department, Islamic AZAD University, Central Tehran Branch, Tehran, Iran
Corresponding Email: sh.javadi@iauctb.ac.ir

Abstract

Nowadays hybrid electric vehicles are the most efficient technology in transportation industry. They have demonstrated the capability of reducing the energy consumption while maintaining vehicle performance [1]. Plug-in hybrid electric vehicles (PHEVs) are powered by conventional or alternative fuels as well as electric power stored in a battery. Using electricity from the grid to run the vehicle some of the time costs less and reduces petroleum consumption compared with conventional vehicles. PHEVs might also reduce emissions, depending on the electricity source. A PHEV has an internal combustion engine and an electric motor, which uses energy stored in batteries. PHEVs have larger battery packs than HEVs. This makes it possible to drive using only electricity for some distance (about 10 to 40 miles), commonly referred to as the "all-electric range" of the vehicle. PHEV batteries can be charged by an outside electric power source, by the internal combustion engine, or through regenerative braking. During braking, the electric motor acts as a generator, charging the battery. PHEV fuel consumption depends on the distance driven between battery charges. For example, if the vehicle is never plugged in to charge, fuel economy will be about the same as a similarly sized hybrid electric vehicle. If the vehicle is driven less than its all-electric range and plugged in, it is possible to use only electric power. There are two main designs for combining the power from the electric motor and the engine: parallel and series. These options exist among HEVs also.

Keywords: Hybrid Electric Vehicles

1. Introduction

A hybrid vehicle is a vehicle that uses two or more distinct power sources to move the vehicle. The term most commonly refers to hybrid electric vehicles (HEVs), which combine an internal combustion engine and one or more electric motors. However, other mechanisms to capture and use energy are included

2. All-Electric Vehicles (EVs)

An all-electric vehicle (EVs) uses a battery to store the electrical energy that powers the motor. EV batteries are charged by plugging the vehicle into an electric power source. EVs are sometimes referred to as battery electric vehicles (BEVs).

Electricity production may contribute to air pollution, but EVs are considered zero-emission vehicles, because their motors produce no exhaust.

Because EVs use no other fuel, they help eliminate petroleum consumption.

Heavy-duty vehicles are available now, and more light-duty EVs are beginning to enter the market. EVs are more expensive than similar conventional and hybrid vehicles, but owners can offset costs through fuel savings, tax credits, or incentives.

EVs have a shorter range per charge than conventional vehicles have per tank of gas. The custom-order, all-electric Tesla Roadster has a 220-mile range. Less expensive vehicles under development are targeting a 100-mile range.

According to the U.S. Department of Transportation Federal Highway Administration, 100 miles is sufficient for over 90% of all household vehicle trips in the United States.

For long trips, it's necessary to charge the vehicle or swap the battery en route.

3. Batteries

Energy storage systems, usually batteries, are essential for electric drive vehicles.

Most near-term PHEVs and EVs will use lithium-ion batteries. They have a high power-to-weight ratio, high energy efficiency, good high-temperature performance, and low self-discharge. Some components of these batteries can be recycled.

Nickel-metal hydride batteries have been successful in EVs and are widely used in HEVs. Challenges with these batteries are high cost, high self-discharge and heat generation at high temperatures, and hydrogen loss.

Lead-acid batteries can be designed to be high power and are inexpensive, safe, and reliable. Drawbacks include low specific energy, poor cold-temperature performance, and short calendar and life cycle.

Lithium-polymer batteries with high specific energy, initially developed for EVs, also can provide high specific power for HEVs. They could become commercially viable if the cost were lowered and life cycle improved.

Ultra capacitors store energy in a polarized liquid between an electrode and an electrolyte. They provide vehicles additional power during acceleration and hill climbing and help recover braking energy.

The battery-recycling market is currently small. As the market grows, the recycling infrastructure will likely grow with it.

For long-distance travel, where fast charging is not available, battery swapping might be a solution.

4. Plug-in Hybrid Electric Vehicles

PHEVs get better fuel economy than similar HEVs and conventional vehicles. They can drive at slow and high speeds using only electricity, so they get about 40% better fuel economy than HEVs. Fuel economy above that of HEVs varies based on how often the vehicle is driven on only electricity.

PHEVs have lower emissions than HEVs and similar conventional vehicles. Their emissions are projected to be lower than HEV emissions, because they are driven on electricity some of the time. Most categories of emissions are lower for electricity generated from power plants than from engines running on gasoline or diesel.

PHEVs are less expensive to operate than HEVs or conventional vehicles. When operating on electricity, a PHEV can be expected to cost \$0.02 to

\$0.04 per mile (based on average U.S. electricity price). When operating on gasoline, the same vehicle will cost \$0.05 to \$0.07 per mile; conventional vehicles cost \$0.10 to \$0.15 per mile to operate.

PHEVs reduce U.S. reliance on imported petroleum. They use electricity produced from coal, nuclear power, natural gas, and renewable sources. Some PHEVs use renewable and domestically produced alternative fuels instead of gasoline or diesel.

PHEVs can fuel up at gas stations or charge at home or public charging stations.

5. Logic Controls of Vehicle Components

All behavior and parts of the vehicle has been modeled using the MATLAB/SIMULINK software. Details of modeling the parts are described in this session.

4.1. Driver control logic

The goal of driver controller is to simulate the behaviour of a real driver. This logic in accordance with requirements of the road (input to model) accelerates the vehicle or presses the brake pedal. To achieve this goal the models compares the difference between actual and desired speed. Two controllers are used to generate the percent throttle and breaking as illustrated in figures 1 and 2.

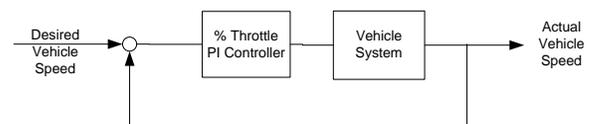


Fig.1.Percent Throttle Closed-Loop PI controller

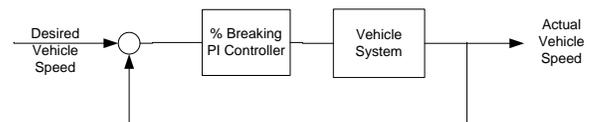


Fig.2.Percent Braking Closed-Loop PI controller

4.2. Power management logic

Because one of the most important aspects of hybrid vehicles is their efficiency and it is because of power management module, modeling of this logic can be the most important part of this article. To model this logic it is enough to find the power desired. This value can be calculated based on the percent of throttle and brake using the equation 1.

$$P_{desired} = P_{max} \times \%Throttle \quad (1)$$

The goal of coupling of generator to engine is to assist engine and improve efficiency. The logic used to assist engine and improve efficiency can be seen in table 1 and table 2 [4].

Table.1 Control logic for engine assisting mode

Engine mode
Transmission gear >1
Desired speed> Actual speed
Percent throttle >50%
Desired power > Max engine power available

Table.2 Control logic for regenerative braking mode

Generator braking mode
Desired speed< Actual speed
Percent throttle =0%
Percent braking >5%
Vehicle speed>16km/h
Desired speed< Max engine available power

4.3. Brake logic

As mentioned in table 2, regenerative braking only activates when vehicle speed is greater than 16km/h and when percent of braking is more than 5% mechanical brakes will be activated. Figure 3 shows the logic control of brake.

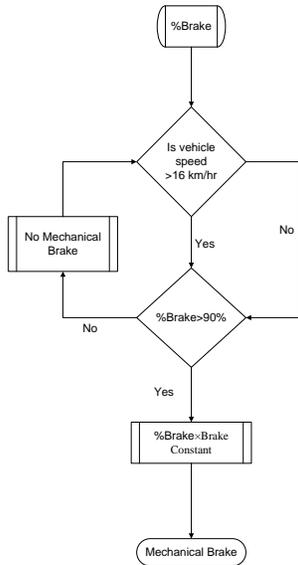


Fig.3. Brake logic control

For this hybrid electric vehicle, two simulation software have been used. As it is discussed earlier powertrain components and logic controls are modeled in MATLAB and mechanical components are modeled in MSC ADAMS. In this session these components are described.

4.4. Drive cycle

One of the most important inputs for analysing the hybrid electric vehicle is the drive cycle. It is simulated using the drive cycle used in ADVISOR software[5],[6] because output of the model should be validated with valid software using the same inputs. In figure 4 schematic of the model is illustrated.

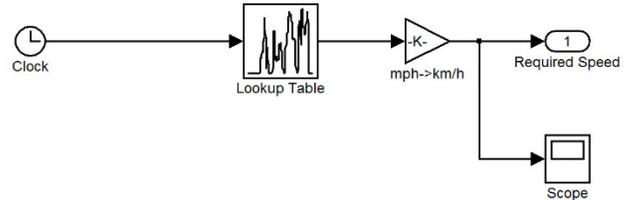


Fig.4. Drive cycle schematic

4.5. Driving control

Driving control uses the drive cycle as the main input to accelerate or activate the brake. This module prepares the most important input for other units like power management control that is described in next session. A model simulation of this part is shown in figure 5.

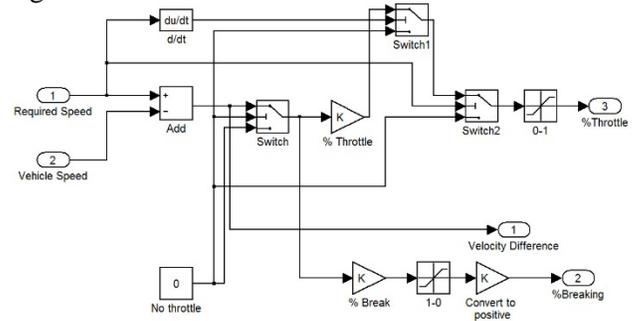


Fig.5. Driving control logic

The output of this control system is braking percent and throttle percent and velocity difference. All these outputs are inputs to power controller that controls the produced power and regenerative braking.

4.6. Power management control

This module controls the power and manages it for best efficiency. For example when the vehicle is in stationary status the engine goes off and when the driver decides to move the engine goes on. In all cases the power is calculated by this module using maximum power, multiply by the percent throttle[7]. It is shown in figure 6.

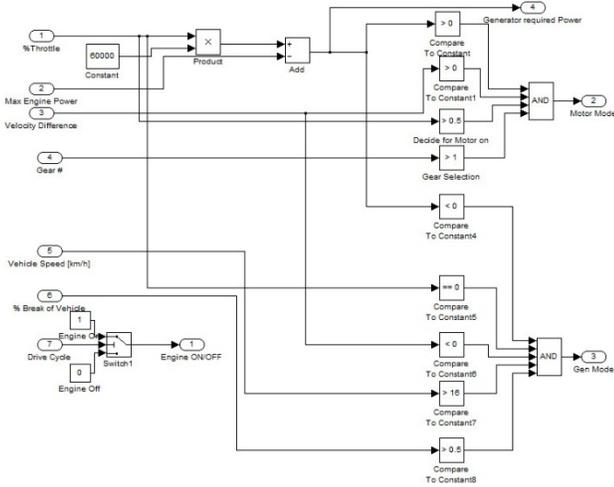


Fig.6. Power controller

4.7.Engine

The main task of this module is to calculate the engine output torque based on the throttle percent and speed .

Table 3. Engine Speed VS Full throttle engine torque[8]

Engine Speed [RPM]	100% Throttle Engine Torque [lb-ft]
800	56.9
1273	58.2
1745	59.5
2218	60.7
2691	62
3146	63.2
3636	64.5
4109	65.7
4582	67
5055	64.3
5527	61.5
6000	58.6

Table 4. Engine speed VS closed throttle engine torque[8]

Engine Speed [RPM]	100% Throttle Engine Torque [lb-ft]
800	-5.15
1273	-8.58
1745	-12.29
2218	-16.28
2691	-20.57
3146	-25.13
3636	-29.97
4109	-35.11
4582	-40.52
5055	-46.23
5527	-52.2
6000	-58.47

This module also calculates the engine fuel consumption as one of the most important criteria for the vehicle.(see table 5)

Table 5: fuel consumption rate [g/s][8]

Engine Torque [lbs]	Engine Speed [rpm]			
	800	2218	4109	5527
5.6	0.0962	0.1883	0.3112	0.4641
22.3	0.2371	0.4927	0.8334	1.3207
44.7	0.5591	1.0325	1.6637	2.6182

61.4	1.0663	1.8193	2.9448	3.8801
------	--------	--------	--------	--------

The engine subsystem model is shown in figure 7.

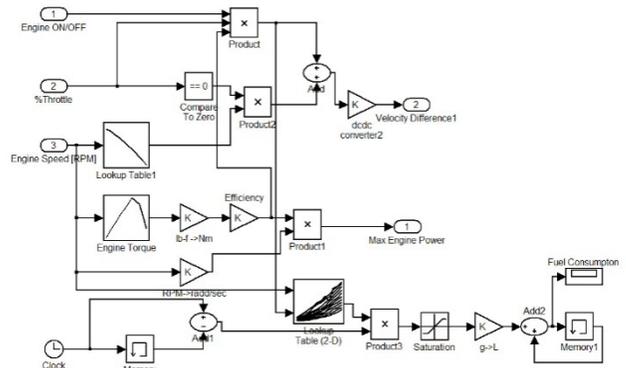


Fig.7. Engine subsystem

4.8 Motor/Generator

This subsystem (See Figure 8) calculates the Motor/Generator torque using motor and generator map (see table 5) and efficiency map (see table 6).

Table 5: Motor and Generator Torque[9]

Shaft Speed [rpm]	Max Motor Torque [Nm]	Max Gen Torque [Nm]
0	46.5	-46.5
500	46.5	-46.5
1000	46.5	-46.5
1500	46.5	-46.5
2000	46.5	-46.5
2500	38.2	-38.2
3500	27.3	-27.3
4000	23.9	-23.9
4500	21.2	-21.2
5000	19.1	-19.1
5500	17.4	-17.4
6000	15.9	-15.9
6500	14.7	-14.7

Table 6: Motor and Generator efficiency map[9]

Speed [rpm]	Motor/ Gen Torque [Nm]				
	-36	-12	0	20	43.5
0	54.17	66.49	63.07	78.1	63.88
2000	83.36	90.39	81.05	90.42	87.95
4000	93.49	93.45	86.24	95.67	95.88
6000	94.07	93.27	80.69	95.4	95.4
8000	91.27	91.27	76.08	93.49	93.49

Another decision that is taken from this module is a decision if the system is in motor mode or regenerator mode. It means that the system is producing power or charging the battery.

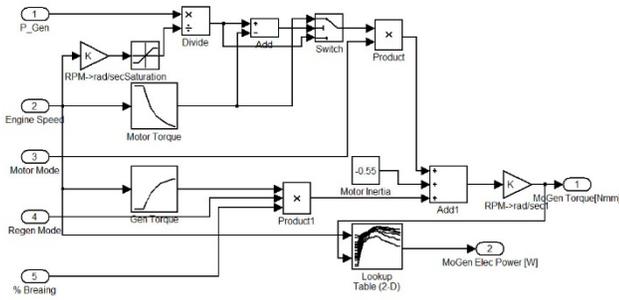


Fig.8. Motor/Generator Subsystem

4.9. Mechanical brake

As it is discussed earlier, mechanical brake (See Figure 9) supply the whole torque required when the velocity of the vehicle is less than 16 km/hr and when the velocity is higher than 16km/hr the regenerative mode is activated through braking system.

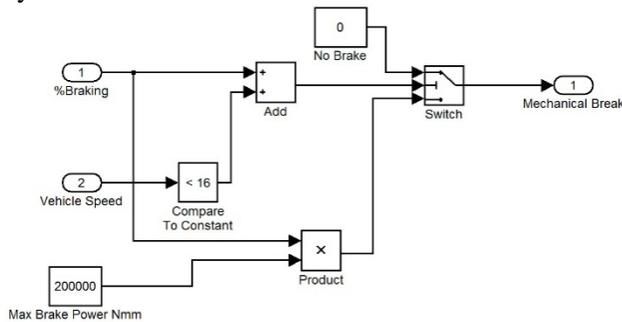


Fig.9. Mechanical Brake

4.10. Battery subsystem

In this model the state of charge [10] and initial level at the beginning is defined and based on the power generated or consumed it is updated during the vehicle run (Figure 10).

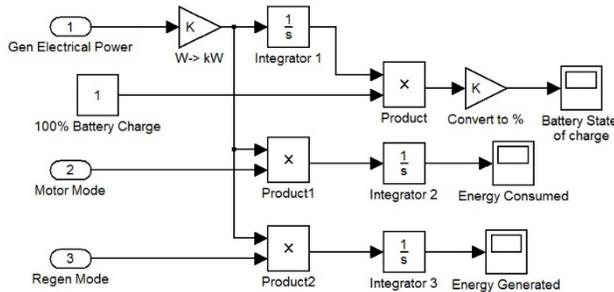


Fig.10. Battery System

5. Charging Electric Drive Vehicles

To keep PHEVs and EVs running, they need to be charged.

Charging equipment, or Electric Vehicle Supply Equipment (EVSE), for PHEVs and EVs is

classified by the maximum amount of power provided to the battery. Charging times range from 30 minutes to 20 hours, depending on how empty the battery is, the type of battery and the type of charging equipment. There are three levels of charging:

- **Level 1:** Level 1 equipment provides charging through a 120 volt (V), alternating-current (AC) plug and requires a dedicated circuit. Equipment is portable and does not require installation. One end of the cord has a standard, three-prong household plug. The other end connects to the vehicle. Reaching a full charge can take 8 to 20 hours.
- **Level 2:** Level 2 equipment offers charging through a 240 V, AC plug and requires installation and a dedicated 40 amp circuit. Most homes have 240 V service available. Reaching a full charge can take 3 to 8 hours.
- **Level 3:** Level 3 equipment is still in development. It will operate at a higher voltage and current than Level 2. Reaching a full charge could take less than 30 minutes.
- **DC Fast Charging:** Direct-current (DC) fast charging equipment (480 V) provides 50 kW to the battery. This option enables charging along heavy traffic corridors and at public stations. A DC fast charge can take less than 30 minutes to fully charge a battery.
- **Inductive Charging:** Inductive charging equipment installed for EVs in the early 1990s is still being used in certain areas. Some companies are working on inductive charging options for future EVs

7. Charging at Home

For consumers to widely accept using EVs and PHEVs, they need affordable, convenient, and compatible options to charge their electric drive vehicles at home.

The Electric Power Research Institute anticipates most EV and PHEV owners will charge their vehicles overnight at home. For this reason, Level 1 (120 volts) and Level 2 (240 volts) charging equipment will be the primary options for homeowners.

Currently available Level 2 charging equipment costs about \$1,500 to \$2,500 (installed) before a 50% federal tax credit (up to \$2,000) and potential state incentives. Nissan and Tesla have information on Level 2 equipment for their vehicles.

Installation contractors can inform homeowners if their home has adequate electrical capacity for vehicle charging. Most people will prefer Level 2 equipment for faster charging, but older homes might have insufficient electric capacity. Homeowners can add circuits to accommodate the capacity needed for Level 2 charging.

8. Charging in Public

For fleet drivers and consumers to charge their EVs and PHEVs in public, charging stations must be deployed and integrated with consideration of daily commutes and typical driving habits.

Public charging stations increase the useful range of EVs and reduce the amount of gasoline consumed by PHEVs.

The majority of EV and PHEV owners will charge at single-family homes. In urban areas, though, residents of high-density housing have only on-street or garage parking. So public chargers must be available to give EVs and PHEVs broad appeal in cities.

General public charging will use Level 2 or DC fast charging to enable faster charging. The public charging infrastructure should consist of charging locations where vehicle owners are highly concentrated, such as shopping centers, city parking lots and garages, airports, hotels, government offices, and other businesses. Widespread public charging infrastructure will help facilitate the penetration of EVs and PHEVs and help address consumer "range anxiety" for vehicles with limited range.

The ability to charge at work can double the daily feasible commuting distance for an EV or a PHEV driver and allows fleets to charge their vehicles overnight.

While ample unused electric generation capacity exists to charge electric drive vehicles overnight, a large number of vehicles using public charging stations during times of peak load could strain the electric grid. This issue can be addressed partially through time-of-use pricing, which would charge more for electricity during periods of peak demand, providing an incentive to charge off-peak.

9. Maintenance and Safety

HEVs and PHEVs have internal combustion engines, so maintenance requirements are similar to those of conventional vehicles. The electrical system (battery, motor, and associated electronics) does not

require scheduled maintenance. Regenerative braking reduces brake wear, extending the life of brake systems.

EVs typically require less maintenance than conventional vehicles because:

The battery, motor, and associated electronics require no regular maintenance. There are no fluids to change, aside from brake fluid. Brake wear is significantly reduced due to regenerative braking. There are far fewer moving parts compared to a conventional gasoline engine.

Electric drive vehicles must meet the same safety standards required for conventional vehicles sold in the United States. The exception is neighborhood electric vehicles, which are subject to less-stringent standards because they are typically limited to roadways specified by state and local regulations.

Emergency response for electric drive vehicles is not significantly different from conventional vehicles. Electric drive vehicles are designed with cut-off switches to isolate the battery and disable the electric system, and all high-voltage power lines are colored orange.

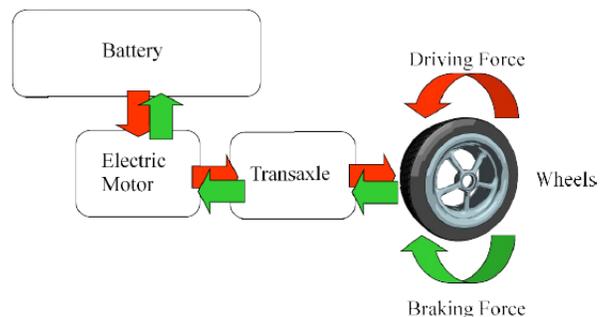
EVs tend to have a lower center of gravity than conventional vehicles, making them less likely to roll over.

10. Types of System

10.1. Traditional IC engine based car

Full electric vehicles have a large battery pack instead of fuel tank and instead of an IC engine there is an electric motor.

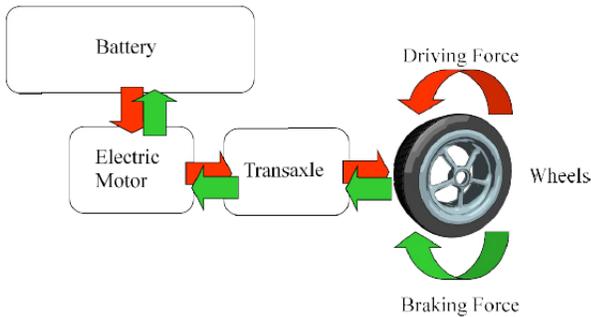
Electrical energy is taken from the battery pack and used in electric motor to produce mechanical energy.



10.2. Full electric vehicle

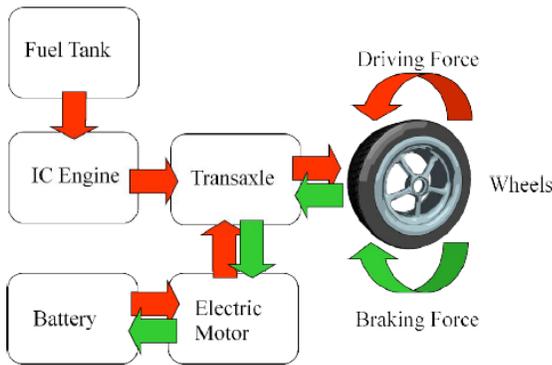
Full electric vehicles have a large battery pack instead of fuel tank and instead of an IC engine there is an electric

motor. Electrical energy is taken from the battery pack and used in electric motor to produce mechanical energy



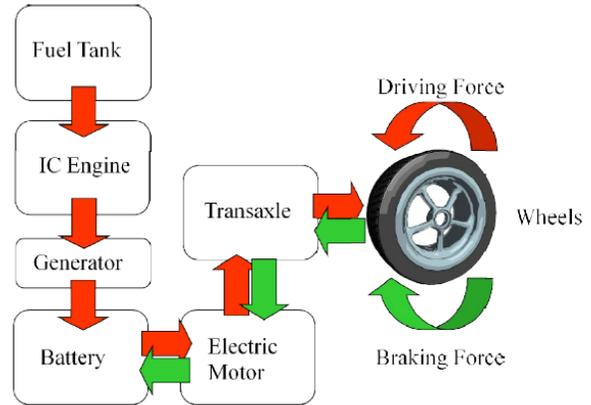
10.3. Parallel hybrid electric vehicle

It is important to notice that in hybrid electric vehicles all the energy used by the car is taken from the liquid fuel, gasoline or diesel, stored in the fuel tank. The battery works only as an energy buffer which temporarily stores energy and then feeds it back to the driving wheels.



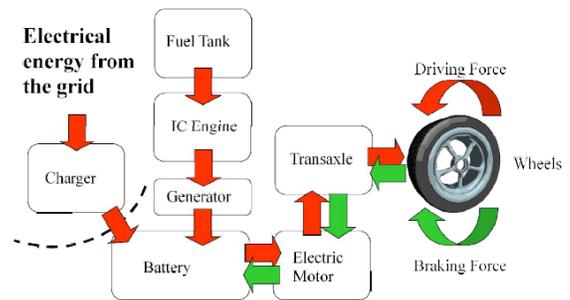
10.4. Serial hybrid electric vehicle

Series-hybrid vehicles are driven by the electric motor with no mechanical connection to the engine. Instead there is an engine tuned for running a generator when the battery pack energy supply isn't sufficient for demands.



10.5. Plug-in hybrid electric vehicle

A plug-in hybrid electric vehicle (PHEV), plug-in hybrid vehicle (PHV), or plug-in hybrid is a hybrid electric vehicle which utilizes rechargeable batteries, or another energy storage device, that can be restored to full charge by connecting a plug to an external electric power source (usually a normal electric wall socket). A PHEV shares the characteristics of both a conventional hybrid electric vehicle, having an electric motor and an internal combustion engine (ICE); and of an all-electric vehicle, having a plug to connect to the electrical grid. Most PHEVs on the road today are passenger cars, but there are also PHEV versions of commercial vehicles and vans, utility trucks, buses, trains, motorcycles, scooters, and military vehicles.



CONCLUSION

It is concluded that MATLAB/ADAMS vehicle results matches very well with results of ADVISOR data. So the model can be used to forecast the behaviour of the vehicle for any drive cycles. For furtherer analysis it is recommended that a better state of charge of battery should be used to calculate better and more precise criteria.

Acknowledgements

The author thank Islamic Azad University, Central Tehran Branch for its financially support.

REFERENCES

- [1] M.Takimoto: "Experience", SAE Paper, No. 2002-21-0068
- [2] J. Peons, Energy Future Coalition, "Challenge and Opportunity: Charting a new Energy Future, Appendix A, Working Group Reports : Transportation", 2007
- [3] Aoki, K., Kuroda, S., Kajiwara, S., Sato, H. et al., "Development of Integrated Motor Assist Hybrid System: Development of the 'Insight'. A Personal Hybrid Coupe, "SAE Technical Paper 2000-01-2216, 2000, doi:10.4271/2000-01-2216
- [4] Karen L. Butler, "A Matlab – Based Modeling and simulation", IEEE Transaction on Vehicular technology, Vol. 48, No. 6, 1999
- [5] MC NREL, Engineer at Ford Company, "Development of data for representation of New York City Cycle" Aug, 17, 1998
- [6] Nigel Clark, West Virginia University, "Cycle designed for heavy duty vehicles and used for dynamometer" Apr, 11, 1998
- [7] Chan-Chiao Lin, "Power management strategy for parallel hybrid vehicle." IEEE Transaction on control systems technology Vol. 11 No. 6, 2003
- [8] K.Kelly: "Advisor Data file: FC_INSIGHT.m", National renewable energy laboratory, 2000
- [9] Published Honda Insight article: "Advisor Data file: MC_INSIGHT_draft.m", National renewable energy laboratory, 2000
- [10] Gregory L. Plett, "Extended Kalman filtering for battery management systems of LiPB-based HEV battery packs", Journal of power sources 134, 2004
- [11] Ken Kelly of NREL, "Honda Insight properties definition" Advisor Software VEH_INSIGHT.m, Oct, 26, 1999

Study of a neutron transport problem by the variational iteration method

Professor Dr. Olga Martin
Applied Sciences Faculty
Polytechnic University of Bucharest
ROMANIA
omartin_ro@yahoo.com

Abstract. A new approach of the variational iterative method (VIM) is used for solving the integral-differential equation of particle transport theory in an isotropic medium. This technique provides a sequence of functions and the study of its convergence to the exact solution of the boundary problem is presented. The accuracy and computational efficiency of the algorithm will be illustrated by a numerical example.

Keywords: integral-differential equations, variational calculus, iterative methods, sequences of functions.

1 Introduction

The resolution of problem concerning the particle transport phenomena in physics and astrophysics is the subject of several works: [1], [5], [16], [17]-[22], [24], [26]. The following methods have been proposed in these papers: Fourier transform, Laplace transform, the least squares, the finite element method, Monte Carlo method, truncated series of Chebyshev polynomials, the fictitious domain method. A special attention has been given to the task of searching methods based on analytical procedures that generate accurate results to the transport problems. We mention: the spectral method, [12], [16], the method based on the Fourier transform proposed by Case and Hazeltine [5], the spherical harmonics method, that expands the flux in Legendre polynomials, [1] and the techniques that solve the inverse radiative-transfer problems of homogeneous equations with the binomial scattering law and Henyey – Greenstein scattering law, [26]. Existence of solution for the particle transport problem was proved using the functional analysis methods, [4], [19], [25]. The homotopy, which is a concept of the topology and differential geometry, was used to obtain the approximate solutions for a wide class of differential, integral and integral-differential equations. We mention here the homotopy analysis method (HAM) established in 1992 by Liao, [8], [9], [15], [22], [23], [30] and the

homotopy perturbation method, [2], [3], [7], [13], [14], [20].

In recent years a special attention has been devoted to the variational iterative method proposed by He, as a modification of a general Lagrange multiplier method [27]-[29], [31]. VIM is used to solve an integral-differential equation of the n -th order in [9] and for the solving a fourth-order Volterra's integral-differential equations in [10]. Also, the solution of the some integral-differential equations was found with this method by He and Wang, who have chosen the initial approximate solution as a function with unknown constants, [11].

The algorithm presented in this paper transforms a boundary value problem for a stationary transport equation in a boundary problem for an integral-differential system that is solved using the techniques of the variational iterative method. We prove the convergence of this new method, presenting a reformulation of VIM that is adapted for the transport equation in absorbing medium.

2 Problem formulation

Let us consider the integral-differential equation of transport theory for the stationary case:

$$\mu \frac{\partial \varphi(x, \mu)}{\partial x} + \sigma \varphi(x, \mu) = \frac{\sigma_s}{2} \int_{-1}^1 \varphi(x, \mu') d\mu' + f(x, \mu) \quad (1)$$

$$\forall (x, \mu) \in \Omega = D_1 \times D_2 = [0, 1] \times [-1, 1],$$

$$D_2 = D_2' \cup D_2'' = [-1, 0] \cup [0, 1]$$

with the following boundary conditions:

$$\begin{aligned} \varphi(0, \mu) &= 0, \mu \geq 0 \\ \varphi(1, \mu) &= 0, \mu < 0 \end{aligned} \quad (2)$$

where

- $\varphi(x, \mu)$ is the density of neutrons, which migrate inside of an isotropic medium toward a direction that makes the angle β with x axis and $\mu = \cos \beta$;

- σ_s is the scattering coefficient, σ_a is the absorption coefficient, $\sigma = \sigma_s + \sigma_a$, for which there is a constant σ_0 with the property

$$\sigma - \sigma_s = \sigma_a \geq \sigma_0 > 0 \quad (3)$$

- $f(x, \mu)$ is a given particles source function.

To solve the problem (1) – (2) we present the following algorithm.

We split the equation (1) in two equations, using the notations, [21]:

$$\begin{aligned} \varphi^+ &= \varphi(x, \mu) \text{ if } \mu > 0; \\ \varphi^- &= \varphi(x, \mu) \text{ if } \mu < 0 \end{aligned} \quad (4)$$

Denoting $\mu' = -\mu$, we get that

$$\int_{-1}^0 \varphi(x, \mu') d\mu' = \int_0^1 \varphi(x, -\mu) d\mu = \int_0^1 \varphi^- d\mu$$

and thus the equation (1) can be written for $\sigma = 1$ in the form

$$\begin{aligned} \mu \frac{\partial \varphi^+}{\partial x} + \varphi^+ &= \frac{\sigma_s}{2} \int_0^1 (\varphi^+ + \varphi^-) d\mu' + f^+ \\ -\mu \frac{\partial \varphi^-}{\partial x} + \varphi^- &= \frac{\sigma_s}{2} \int_0^1 (\varphi^+ + \varphi^-) d\mu' + f^- \end{aligned} \quad (5)$$

Adding and subtracting the equations (5) and introducing the notations:

$$\begin{aligned} u &= \frac{1}{2}(\varphi^+ + \varphi^-), \quad v = \frac{1}{2}(\varphi^+ - \varphi^-) \\ g &= \frac{1}{2}(f^+ + f^-), \quad r = \frac{1}{2}(f^+ - f^-) \end{aligned} \quad (6)$$

we obtain the following system

$$\mu \frac{\partial v(x, \mu)}{\partial x} + u(x, \mu) = \sigma_s \int_{-1}^1 u(x, \mu') d\mu' + g(x, \mu) \quad (7)$$

$$\mu \frac{\partial u(x, \mu)}{\partial x} + v(x, \mu) = r(x, \mu)$$

accompanied by the boundary conditions:

$$\begin{aligned} u + v &= 0 \quad \text{for } x = 0, \\ u - v &= 0 \quad \text{for } x = 1. \end{aligned} \quad (8)$$

If the second equation (7) is solved with respect to v , we get:

$$v = r - \mu \cdot \frac{\partial u}{\partial x}$$

Replacing v in first equation of (7), this becomes

$$u - \sigma_s \int_0^1 u d\mu' + \mu \cdot \frac{\partial r}{\partial x} - \mu^2 \cdot \frac{\partial^2 u}{\partial x^2} = g \quad (9)$$

Let A be the operator

$$A = \mu^2 \frac{\partial^2}{\partial x^2} - 1 + \sigma_s \int_0^1 d\mu \quad (10)$$

and from (9) – (10), we get

$$Au + F = 0 \tag{11}$$

where

$$F = g - \mu \frac{\partial r}{\partial x},$$

and

$$u(0, \mu) - \mu \frac{\partial u}{\partial x}(0, \mu) = -r(0, \mu) \tag{12}$$

$$u(1, \mu) + \mu \frac{\partial u}{\partial x}(1, \mu) = r(1, \mu), \mu \in [0, 1]$$

3 Description of the variational iteration method

To obtain the solution of the integral-differential equation (11), we rewrite this in the following form: using the linear operators L and R and the nonlinear operator N :

$$Lu + Ru + Nu + F = 0 \tag{13}$$

where L and R are the linear operators:

$$Lu = \mu^2 \frac{\partial^2 u}{\partial x^2}, \quad Ru = -u \tag{14}$$

and N is the nonlinear operator:

$$Nu = \sigma_s \int_0^1 u(x, \mu') d\mu', \tag{15}$$

According to the variational iteration method (VIM), in x direction is considered the correction functional:

$$u_{m+1}(x, \mu) = u_m(x, \mu) + \int_0^x \lambda(\xi) (Lu_m(\xi, \mu) + R\tilde{u}_m + N\tilde{u}_m(\xi, \mu) + F(\xi, \mu)) d\xi, \tag{16}$$

$$\forall (x, \mu) \in (0, 1] \times (0, 1]$$

This recurrence formula is accompanied by the following conditions:

$$u_m(0, \mu) = 0, \quad u_{m,x}(0, \mu) = \gamma(\mu), \quad \forall \mu \in (0, 1],$$

$$m = 0, 1, \dots \tag{17}$$

where

$$u_{m,x}(x, \mu) = \frac{\partial u_m(x, \mu)}{\partial x}, \quad \gamma(\mu) = \frac{r(0, \mu)}{\mu} \tag{18}$$

To obtain the successive approximations, u_{m+1} , we determine the Lagrange multiplier λ , using the variational techniques. The function \tilde{u}_m is a restricted variation in x direction, which means $\delta\tilde{u}_m = 0$, [9].

Note that $\delta u_m(0, \mu) = 0, \delta u_{m,x}(0, \mu) = 0$ and $\delta F = 0$, because the values of u_m , its partial derivative with respect to x at a point $(0, \mu)$ and $F(x, \mu)$ are known. Taking in (16) the variation, $\delta u_{m+1}(x, \mu) = 0$, we obtain via the integration by parts:

$$\begin{aligned} \delta u_{m+1}(x, \mu) &= \delta u_m(x, \mu) + \\ &+ \delta \int_0^x \lambda(\xi) \left[\mu^2 \frac{\partial^2}{\partial \xi^2} u_m(\xi, \mu) - \tilde{u}_m(\xi, \mu) + \right. \\ &\quad \left. + \sigma_s \int_0^1 \tilde{u}_m(\xi, \mu') d\mu' + F(\xi, \mu) \right] d\xi = \\ &= \left(1 - \mu^2 \lambda'(\xi) \Big|_{\xi=x} \right) \delta u_m(x, \mu) + \\ &+ \lambda(\xi) \Big|_{\xi=x} \mu^2 \delta u_{m,x}(x, \mu) + \\ &+ \int_0^x \left[\mu^2 \lambda''(\xi) \delta u_m(\xi, \mu) \right] d\xi = 0 \end{aligned} \tag{19}$$

So, $\lambda(\xi)$ verifies the stationary conditions:

$$\lambda(\xi) \Big|_{\xi=x} = 0, \quad \mu^2 \lambda'(\xi) \Big|_{\xi=x} = 1, \quad \lambda''(\xi) = 0 \tag{20}$$

and the Lagrange multiplier will be identified as

$$\lambda(\xi) = \frac{\xi - x}{\mu^2} \tag{21}$$

The iteration formula becomes

$$u_{m+1}(x, \mu) = u_m(x, \mu) + \frac{1}{\mu^2} \int_0^x (\xi - x) \left[\mu^2 \frac{\partial^2 u_m}{\partial \xi^2}(\xi, \mu) - u_m(\xi, \mu) + \sigma_s \int_0^1 u_m(\xi, \mu') d\mu' + F(\xi, \mu) \right] d\xi$$

$m = 0, 1, \dots$ (22)

The sequence $\{u_m(x, \mu)\}_m$ defined by (22) converges to the exact solution of the problem (11), (12), if we choose a suitable initial approximation $u_0(x, \mu)$.

4 Convergence study

Theorem 1. If $u_m \in C^2(\Omega), \forall m \in \mathbf{N}^*$, then the iteration formula (22) is equivalent to the following relation:

$$L[u_{m+1}(x, \mu) - u_m(x, \mu)] = -G(u_m(x, \mu)) \quad (23)$$

where

$$u_{m,xx}(x, \mu) = \frac{\partial^2 u_m(x, \mu)}{\partial x^2} \text{ and } G(u_m(x, \mu)) = Lu_m(x, \mu) + Ru_m(x, \mu) + Nu_m(x, \mu) + F(x, \mu) \quad (24)$$

■ Let us consider that u_m and u_{m+1} satisfy (16):

$$u_{m+1}(x, \mu) = u_m(x, \mu) + \int_0^x \lambda(\xi) G(u_m(\xi, \mu)) d\xi = u_m(x, \mu) + \frac{1}{\mu^2} \int_0^x (\xi - x) G(u_m(\xi, \mu)) d\xi, \quad m = 0, 1, \dots$$

(25)

where

$$Lu_m(\xi, \mu) = \mu^2 \frac{\partial^2 u_m(\xi, \mu)}{\partial \xi^2} \text{ and } Ru_m(\xi, \mu) = -u_m(\xi, \mu), \quad Nu_m(\xi, \mu) = \sigma_s \int_0^1 u_m(\xi, \mu') d\mu' \quad (26)$$

Applying twice the derivative of (25) with respect to x , taking into account (26) and the conditions (20), we find successively

$$\mu^2 (u_{m+1}(x, \mu) - u_m(x, \mu))_{,xx} = - \int_0^x G(u_m(\xi, \mu)) d\xi$$

$$\mu^2 (u_{m+1}(x, \mu) - u_m(x, \mu))_{,xx} = -G(u_m(x, \mu)) \quad (27)$$

Finally, using the definition (26) of L , we get

$$L[u_{m+1}(x, \mu) - u_m(x, \mu)] = -G(u_m(x, \mu)) \quad (28)$$

Conversely, we prove now: if u_m and u_{m+1} verify (23), then the iterative formula (22) is satisfied.

Multiplying (23) by $\lambda \neq 0$, integrating both sides from zero to x and using the integration by parts and (20), we obtain

$$\mu^2 \int_0^x \lambda(\xi) [u_{m+1,\xi\xi}(\xi, \mu) - u_{m,\xi\xi}(\xi, \mu)] d\xi = - \int_0^x \lambda(\xi) G(u_m(\xi, \mu)) d\xi \quad (29)$$

$$x[u_{m+1,\xi}(0, \mu) - u_{m,\xi}(0, \mu)] - \int_0^x [u_{m+1,\xi}(\xi, \mu) - u_{m,\xi}(\xi, \mu)] d\xi = - \int_0^x \lambda(\xi) G(u_m(\xi, \mu)) d\xi \quad (30)$$

$$-[u_{m+1}(x, \mu) - u_m(x, \mu)] = - \int_0^x \lambda(\xi) G(u_m, u_{m,\xi\xi}) d\xi \quad (31)$$

which lead to the iterative formula (22). ■

Theorem 2. If $u_m \in C^2(\Omega), \forall m \in \mathbf{N}^*$ verifies the iterative formula (22) and $\lim_{m \rightarrow \infty} u_m(x, \mu) = \tau(x, \mu)$ on $\overline{\Omega}$, then $\tau(x, \mu) = u_{ex}(x, \mu)$, where $u_{ex}(x, \mu)$ is the exact solution of the problem (11), (12).

■ According to the properties from hypothesis, the following equalities are valid:

$$\begin{aligned} \lim_{m \rightarrow \infty} L[u_{m+1}(x, \mu) - u_m(x, \mu)] &= \\ = L \lim_{m \rightarrow \infty} [u_{m+1}(x, \mu) - u_m(x, \mu)] &= 0, \forall (x, \mu) \in \overline{\Omega} \end{aligned} \tag{32}$$

In view of Th. 2, we have

$$\lim_{m \rightarrow \infty} G(u_m(x, \mu)) = 0, \forall (x, \mu) \in \overline{\Omega} \tag{33}$$

From the property of continuity of u_m and its derivatives, (33) leads to

$$\begin{aligned} \mu^2 \frac{\partial^2}{\partial x^2} (\lim_{m \rightarrow \infty} u_m(x, \mu)) - \lim_{m \rightarrow \infty} u_m(x, \mu) + \\ + \sigma_s \int_0^1 \lim_{m \rightarrow \infty} u_m(x, \mu') d\mu' + F(x, \mu) = 0 \end{aligned} \tag{34}$$

Using the uniform convergence of the sequence $\{u_m\}_m$ on $\overline{\Omega}$, we get

$$\mu^2 \frac{\partial^2 \tau(x, \mu)}{\partial x^2} - \tau(x, \mu) + \sigma_s \int_0^1 \tau(x, \mu') d\mu' + F(x, \mu) = 0 \tag{35}$$

On the other hand, the boundary conditions become

$$\tau(0, \mu) = \lim_{m \rightarrow \infty} u_m(0, \mu) = 0 \tag{36}$$

$$\tau_{,x}(0, \mu) = \lim_{m \rightarrow \infty} u_{m,x}(0, \mu) = \gamma(\mu) \tag{37}$$

It follows from (35) – (37) that $\tau(x, \mu)$ is the exact solution of the problem (11), (12). ■

5 Numerical example

Let us consider the following problem

$$\mu \frac{\partial \varphi(x, \mu)}{\partial x} + \sigma \varphi(x, \mu) = \frac{\sigma_s}{2} \int_{-1}^1 \varphi(x, \mu') d\mu' + f(x, \mu) \tag{38}$$

$$\forall (x, \mu) \in D_1 \times D_2 = [0, 1] \times [-1, 1]$$

where

$$f(x, \mu) = 2\pi\mu^3 \cos 2\pi x + \left(\mu^2 - \frac{\sigma_s}{3}\right) \sin 2\pi x \tag{39}$$

According to the notations (6), the functions g and r will be of the form

$$g(x, \mu) = \left(\mu^2 - \frac{\sigma_s}{3}\right) \sin 2\pi x, \tag{40}$$

$$r(x, \mu) = 2\pi\mu^3 \cos 2\pi x$$

The problem (11), (12) becomes

$$\mu^2 \frac{\partial^2 u}{\partial x^2} - u + \sigma_s \int_0^1 u(x, \mu') d\mu' + F(x, \mu) = 0 \tag{41}$$

$$F(x, \mu) = \left(\mu^2 + 4\pi^2 \mu^4 - \frac{\sigma_s}{3}\right) \sin 2\pi x = \alpha(\mu) \sin 2\pi x \tag{42}$$

with the following conditions

$$u(0, \mu) = 0, \frac{\partial u}{\partial x}(0, \mu) = \frac{r(0, \mu)}{\mu} = 2\pi\mu^2 \tag{43}$$

The exact solution is

$$u_{ex}(x, \mu) = \mu^2 \sin 2\pi x = \varphi(x, \mu), v(x, \mu) = 0 \tag{44}$$

Using the method presented in the section 4, we obtain the iterative formula

$$\begin{aligned} u_{m+1}(x, \mu) = & u_m(x, \mu) + \\ & + \frac{1}{\mu} \int_0^x \frac{\xi - x}{\mu^2} \left[\mu^2 \frac{\partial^2}{\partial \xi^2} u_m(\xi, \mu) - u_m(\xi, \mu) + \right. \\ & \left. + \sigma_s \int_0^1 u_m(\xi, \mu') d\mu' + F(\xi, \mu) \right] d\xi \\ & m = 0, 1, \dots \end{aligned} \tag{45}$$

Let us consider the initial approximation

$$u_0(x, \mu) = 2\pi\mu^2 x \tag{46}$$

that verifies (43).

First iteration

$$u_1(x, \mu) = 2\pi \mu^2 x + \int_0^x 2\pi \xi \frac{\xi - x}{\mu^2} \left(\frac{\sigma_s}{3} - \mu^2 \right) d\xi + \frac{\alpha(\mu)^x}{\mu^2} \int_0^x (\xi - x) \sin 2\pi \xi d\xi \tag{47}$$

Using an integration by parts for the last term of the sum, we get

$$u_1(x, \mu) = 2\pi \mu^2 x + 2\pi \left(\frac{\sigma_s}{3\mu^2} - 1 \right) \frac{x^3}{6} - \frac{\alpha(\mu)}{4\pi^2 \mu^2} \sin 2\pi x \tag{48}$$

We now expand the function $\sin 2\pi x$ into a MacLaurin series and with the approximation:

$$\sin 2\pi x \approx 2\pi x - \frac{(2\pi x)^3}{3!} \tag{49}$$

we obtain

$$u_1(x, \mu) = 2\pi \mu^2 x + 2\pi \left(1 - \frac{\sigma_s}{3\mu^2} \right) \frac{x^3}{6} - x \left(2\pi \mu^2 + \frac{1}{2\pi} - \frac{\sigma_s}{6\pi \mu^2} \right) + \left(\mu^2 + \frac{1}{4\pi^2} - \frac{\sigma_s}{12\pi^2 \mu^2} \right) \left(2\pi x - \frac{(2\pi x)^3}{6} \right) = \mu^2 \left(2\pi x - \frac{(2\pi x)^3}{3!} \right) \tag{50}$$

Second iteration

$$u_2(x, \mu) = u_1(x, \mu) + \int_0^x \frac{\xi - x}{\mu^2} \left[(2\pi)^3 \mu^4 \xi + \left(\frac{\sigma_s}{3} - \mu^2 \right) \left(2\pi \xi - \frac{(2\pi \xi)^3}{6} \right) + \alpha(\mu) \sin 2\pi \xi \right] d\xi \tag{51}$$

After performing the calculations and using the approximation:

$$\sin 2\pi x \approx 2\pi x - \frac{(2\pi x)^3}{3!} + \frac{(2\pi x)^5}{5!} \tag{52}$$

we find the final form of this iteration

$$u_2(x, \mu) = \mu^2 \left(2\pi x - \frac{(2\pi x)^3}{3!} + \frac{(2\pi x)^5}{5!} \right) \tag{53}$$

.....

m - iteration

In the same manner, it follows that

$$u_m(x, \mu) = \mu^2 \left(2\pi x - \frac{(2\pi x)^3}{3!} + \dots + (-1)^{m-1} \frac{(2\pi x)^{2m-1}}{(2m-1)!} \right) \tag{54}$$

and

$$\lim_{m \rightarrow \infty} u_m(x, \mu) = \mu^2 \sin 2\pi x = u_{ex}(x, \mu) \tag{55}$$

It is evident that the numerical approximation shows that u_m is accurate for low values of m , the solution being rapidly convergent by utilizing the VIM.

6 Conclusions

In the case of particle transport problems, the existence of streaming term, $\mu \frac{\partial \varphi(x, \mu)}{\partial x}$ leads to the additional difficulties. We mention that in our paper, the unknown function depends on two variables and not one variable as in other works, which solve the integral- differential equations with iterative methods.

Unlike the previous methods [19], [20] and [22] used to solve these problems, MIV improves the speed of convergence if the guess function u_0 is chosen so that the conditions (17) are satisfied. Also, in this paper we prove the convergence of the

sequence $\{u_m(x, \mu)\}_m$ to the exact solution on Ω . A numerical example confirms the validity and great potential of this new iterative method for the solving a wide variety of integral-differential equations that appear in physics, astrophysics and in the study of viscoelastic structures.

References

- [1] L.Barichello, C. Siewert, On the Equivalence Between the Discrete-ordinates and Spherical-harmonics Methods in Radiative Transfer, Nuclear Sci. Eng. 130 (1998) 665-675
- [2] J. Biazar, H. Ghazvini, Exact solutions for nonlinear Schrodinger equations by He's homotopy perturbation method, Physics Letters A 366 (2007) 79 - 84.
- [3] J. Biazar, H. Ghazvini, Numerical solution for special non-linear Fredholm integral equation by HPM, Appl Math Comp.,198 (2008) 681-687.
- [4] H. Brezis, *Analyse Fonctionnelle*, Dunod, Paris, 1983.
- [5] K.M. Case, R.D. Hazeltine, Three-dimensional Linear Transport Theory, Journal of Mathematical Physics, 11 (1970) 1126-1135.
- [6] J. Halton , Sequential Monte Carlo Techniques for the Solution of Linear Systems, J Sci Comp. 9 (1994) 213-257.
- [7] J. H. He, Homotopy perturbation technique, Meth in Appl Mech and Eng., 178 (1999) 257-262.
- [8] J. H. He, Comparison of homotopy perturbation method and homotopy analysis method, Appl Math Comp., 156 (2004) 527-539.
- [9] J. H. He, Variational iteration method for autonomous ordinary differential systems, Appl. Math. Comput., 114 (2000) 115-123.
- [10] He J. H., Variational principles for some nonlinear PDE with variable coefficients, Chaos, Solitons and Fractals, 19 (2004) 847-851.
- [11] He, J.H., Comparison of homotopy perturbation method and homotopy analysis method, Appl Math Comp., 156 (2004) 527-539.
- [12] A. Kadem, Analytical Solutions for the Neutron Transport Using the Spectral Methods, Int. J. of Mathematics and Mathematical Sciences, (2006) 1-11.
- [13] X. Li, M., Xu, X., Jiang, Homotopy perturbation method to time-fractional diffusion equation with a moving boundary condition, Applied Mathematics and Computation, 208 (2009) 434-439.
- [14] T. Y. Li, X. and Wang, Solving real polynomial system with real homotopy, Math Comp.,60 (1993) 669-680.
- [15] S. J. Liao, Notes on the homotopy analysis method. Some definitions and theorems, Comm Nonlinear Sci Numer Simul.,14 (2009) 983-997.
- [16] E. Lewis, W. F. Miller, Computational Methods of Neutron Transport, American Nuclear Society, Illinois, 1993.
- [17] O. Martin, A numerical solutions for two – dimensional transport equation, Central European J Math., 2 (2004) 191-198.
- [18] O. Martin, Approximate method for solving a neutron transport equation, J Comp Anal Appl, 11(2009) 431-442.
- [19] O. Martin, On approximation in the fictitious domain method for a bi-dimensional transport equation, Num. Meth. Partial Diff. Eqs. , 26 (2009) 1275 – 1290.
- [20] O. Martin, A new homotopy perturbation method for solving a neutron transport equation, Applied Mathematics and Computation, 217 (2011) 8567 – 8574.
- [21] O. Martin, On bounded solution of transport equations solved with a spectral method, Numerical Meth for Partial Diff Eqs, 28 (2012) 1152-1160.
- [22] O. Martin, On the homotopy analysis method for solving a particle transport equation, Applied Mathematical Modelling, 37 (2013) 3959-3967.
- [23] M. Matinfar, M., Saeidy, J. Vahidi, Analytical solution of BVPs for fourth-order integro-differential equations by using HAM, International Journal of Nonlinear Science, 9 (2010) 414 – 421.
- [24] J. Morel, T., Wareing, K. Smith, A linear-discontinuous spatial differencing scheme S_N radiative transfer calculations, J. Comput. Phys., 128 (1996) 445-453.

- [25] A. Pazy, Semigroups of linear operators and applications to partial differential equations, Springer-Verlag, 1983.
- [26] C. E. Siewert, Inverse solutions to radiative-transfer problems based on the binomial or the Henyey-Greenstein scattering law, Journal of Quantitative Spectroscopy & radiative Transfer, 72 (2002) 837-835.
- [27] J. Saberi, A. Ghorbani, Convergence of He's VIM for nonlinear oscillators, Nonlinear Sci. Lett., 1 (2010) 379 – 384.
- [28] E. Shivanian, S. Abbasbandy, Application of variational iteration method for nth-order integral differential equations, Zeitschrift fur Naturforschung A, 64(a) (2009) 439-444.
- [29] N. Sweilam, Fourth order integro-differential equations using variational iteration method, Comput. Math. Appl., 54 (2007) 1086-1091.
- [30] J. Verschelde, R. Cools, Symmetric homotopy construction, J. Comput. Appl. Math., 50 (1994), 575-592.
- [31] S. Q. Wang, J. H. He, Variational iteration method for solving integro-differential equations, Phys. Lett. A, 367 (2007), p.188-191.

Authors Index

Abd-Alla, A.-El-N. N.	282	El-Yagubi, E.	159	Lomidze, I. R.	228
Abd-Elkader, H. M.	300	Erdodi, G.-M.	264	Lopez, J.	215
Abdelwahed, M.	370	Fazilova, L.	150	Magalhães-Mendes, J.	88
Ahmed, A. E.	300	Filasova, A.	406	Maggiore, P.	54, 166, 276
Ahmed, H. E. H.	300	Fjodorova, N.	237	Mann, Z. A.	102
Al-Hossain, A.	325	Fülep, D.	253	Mano, V. M.	40
Alshaikh, F.	282	Gabor, C.	384, 400	Martin, O.	441
Aprausheva, N. N.	309	Garbuzov, K.	418	Matsuno, S.	32
Aydogan, M.	135, 159	Glazunov, N.	375	Mazal, J.	196
Babankov, V. A.	232	Grădinaru, C.	384, 400	Mehta, R.	32
Belmonte, D.	166	Greene, N.	17	Minca, E.	343
Bertoli-Barsotti, L.	37	Gutierrez, A.	215	Mioc, M. A.	363
Bezmen, P.	83, 146	Guzman, Y.	215	Misak, S.	412
Bílková, D.	63	Hadzikadic, M.	270	Musuroi, S.	264
Blaj, C.	130	Hafez, A.	300	Nikov, A.	257
Bourazza, S.	325	Hladky, V.	406	Novič, M.	237
Burova, I.	355	Hodzic, H.	270	Olenev, N. N.	173
Buslaev, A. P.	388	Huang, S. H.	49	Ortobelli, S.	142, 244
Carmichael, T.	270	Ikeda, S.	32, 423	Ouda, A. N.	300
Cata, I.	130	Işık, Y.	247	Pabisek, E.	349
Caviativa, J.	215	Ito, Ta.	32, 423	Pace, L.	54, 276
Cermak, T.	412	Ito, Ts.	32	Pakdemirli, M.	119
Chatzimichail, E. A.	44	Jatsun, S.	83, 146	Paraskakis, E.	44
Coy, L.	215	Javadi, S.	433	Perner, P.	26
Dalla Vedova, M. D. L.	54, 166, 276	Jonakowski, W.	290	Petrescu, D.-I.	264
Darus, M.	159	Kahramaner, Y.	135	Petronio, F.	142
David, I.	384, 400	Kanjer, K.	290	Polandov, I. H.	232
Dem'yanovich, Yu. K.	355	Kartal, S. K.	208	Polatoglu, Y.	135
Dikusar, V. V.	173, 309	Korul, H.	247	Polshkova, I. N.	379
Dobrikov, S. A.	232	Krokavec, D.	406	Prangishvili, A.	124
Dolapçı, I. T.	119	Kumar, A. S.	412	Pribic, J.	290
Dragomir, F.	343	Lando, T.	37, 142, 244	Radulescu, A. T.	107
Dragomir, O.	343	László, I.	253	Radulescu, C. M.	107
Durmagambetov, A.	150	Leblebicioğlu, M. K.	208	Radulescu, G. M. T.	107
El Sehiemy, R. A.	222	Lhaci, A.	130	Radulovic, M.	290

Rajchakit, G.	295	Shen, K.-S.	181	Tichy, T.	142
Rajchakit, M.	202	Shirazi, S. A. M.	359	Toader, D.	130
Rambharose, T.	257	Shonia, O.	124	Tosun, D. C.	247
Reddy, G. V. R.	313	Sokáč, M.	191	Ucar, H. E. O.	135
Riaza, R.	79	Sopta, J.	290	Uchida, Y.	32
Rigas, A. G.	44	Sorandaru, C.	264	Vasiljevic, J.	290
Riznyk, V. V.	115	Sorokin, S. V.	309	Velísková, Y.	191
Rodionov, A. S.	418	Spanou, E. N.	44	Vieira, L. A.	40
Rodonaia, I.	124	Spyroglou, I. I.	44	Vlad, I.	384, 400
Rodonaia, V.	124	Stodola, P.	196	Vukosavljevic, D. N.	290
Sakamoto, M.	32, 423	Stoeva, S.	257	Waszczyszyn, Z.	349
Savin, S.	83, 146	Tatashev, A. G.	388	Wojtowicz, M.	173
Selman, S.	270	Ștefănescu, C.	384, 400	Yemisci, A.	135
Selmane, S.	393	Tereshchenko, V.	163	Zambelli, A. E.	96
Shakhov, V.	139	Tereshchenko, Y.	163	Zsoldos, I.	253
Shamolin, M. V.	328	Testa, E.	276		