MATHEMATICAL METHODS in SCIENCE and ENGINEERING

Proceedings of the 1st International Conference on Mathematical Methods & Computational Techniques in Science & Engineering (MMCTSE 2014)

Athens, Greece
November 28-30, 2014

MATHEMATICAL METHODS in SCIENCE and ENGINEERING

Proceedings of the 1st International Conference on Mathematical Methods & Computational Techniques in Science & Engineering (MMCTSE 2014)

Athens, Greece
November 28-30, 2014

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

Series: Mathematics and Computers in Science and Engineering Series | 37

ISSN: 2227-4588
ISBN: 978-1-61804-256-9

MATHEMATICAL METHODS in SCIENCE and ENGINEERING

Proceedings of the 1st International Conference on Mathematical Methods & Computational Techniques in Science & Engineering (MMCTSE 2014)

Athens, Greece
November 28-30, 2014

Organizing Committee

Editors:
Professor Nikos Mastorakis, Technical University of Sofia, Sofia, Bulgaria
Professor Peter Revesz, University of Nebraska-Lincoln, USA
Professor Panos M. Pardalos, University of Florida, USA
Professor Cornelia Aida Bulucea, University of Craiova, Romania
Professor Atsushi Fukasawa, Institute of Statistical Mathematics, Japan

Organizing Committee:
Prof. Kleanthis Psarris, The City University of New York, USA (General Chair)
Prof. George Vachtsevanos, Georgia Institute of Technology, Atlanta, Georgia, USA (Co-Chair)
Prof. Valeri Mladenov, Technical University of Sofia, Bulgaria (Publications Chair)
Prof. Aida Bulucea, University of Craiova, Craiova, Romania (Publicity Chair)
Prof. Imre Rudas, Obuda University, Budapest, Hungary (Tutorials Chair)
Prof. Nikos Mastorakis, Technical University of Sofia, Bulgaria and HNA, Greece (Workshops Chair)
Prof. Olga Martin. Politehnica University of Bucharest, Romania (Special Sessions Chair)

Steering Committee:
Prof. Theodore B. Trafalis, University of Oklahoma, USA
Prof. Charles A. Long, Professor Emeritus, University of Wisconsin, Stevens Point, Wisconsin, USA
Prof. Maria Isabel García-Planas, Universitat Politècnica de Catalunya, Spain
Prof. Reinhard Neck, Klagenfurt University, Klagenfurt, Austria
Prof. Myriam Lazard, Institut Superieur d' Ingenierie de la Conception, Saint Die, France
Prof. Zoran Bojkovic, University of Belgrade, Serbia
Prof. Claudio Talarico, Gonzaga University, Spokane, USA

International Scientific 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, CNRS/Verimag, France)
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. Martin 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)
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. 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
Dr. Michelle Luke (Univ. Berkeley, USA)
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. Kleanthis 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. Valeri Mladenov (Technical University of Sofia, Bulgaria)
Prof. Hiroshi Sakaki (Meisei University, Tokyo, Japan)
Prof. Zoran S. Bojkovic (Technical University of Belgrade, Serbia)
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. C. G. Helmis (University of Athens, Greece)
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. 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. 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. 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. 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. 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. Josef Boercsoek (Universitat Kassel, Germany)
Prof. Eyad H. Abed (University of Maryland, Maryland, USA)
Prof. F. Castanie (TeSA, Toulouse, France)
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. T.-W. Lee (Arizona State University, AZ, USA)
Prof. Le Yi Wang (Wayne State University, Detroit, 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. Ion Chryssoverghi (National Technical University of Athens, Greece)
Prof. Dr. Aydin Akan (Istanbul University, Turkey)
Prof. Sarka Necasova (Academy of Sciences, Prague, Czech Republic)
Prof. C. D. Memos (National Technical University of Athens, Greece)
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. 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

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

Table of Contents

ISBN: 978-1-61804-256-9

ISBN: 978-1-61804-256-9

ISBN: 978-1-61804-256-9

Plenary Lecture 1

Cubic Spline Interpolation by Solving a Single Recurrence Equation instead of a Triangular Matrix

Professor Peter Revesz
Computer Science and Engineering
University of Nebraska-Lincoln
USA
E-mail: revesz.nebraska@gmail.com

Abstract: The cubic spline interpolation method is probably the most widely-used polynomial interpolation method for functions of one variable. However, the cubic spline method requires solving a triangular matrix-vector equation with an O(n) computational time complexity where n is the number of data measurements. Even an O(n) time complexity may be too much in some time-ciritical applications, such as continuously estimating and updating the flight paths of moving objects. This paper shows that under certain boundary conditions the triangular matrix solving step of the cubic spline method could be entirely eliminated and instead the coefficients of the unknown cubic polynomials can be found by solving a single recurrence equation in much faster time.

Brief Biography of the Speaker: Peter Revesz holds a Ph.D. degree in Computer Science from Brown University. He was a postdoctoral fellow at the University of Toronto before joining the University of Nebraska-Lincoln, where he is a professor in the Department of Computer Science and Engineering. His current research interests are bioinformatics, geoinformatics, databases and data mining. He is the author of several books, including the textbook Introduction to Databases: From Biological to Spatio-Temporal (Springer, 2010). He held visiting appointments at the IBM T.J. Watson Research Center, INRIA, the University of Hasselt, the Max Planck Institute for Computer Science, the University of Athens, and the U.S. Department of State. He is a recipient of a National Science Foundation CAREER award, and a J. William Fulbright, an Alexander von Humboldt, and a Jefferson Science Fellowship.

ISBN: 978-1-61804-256-9

Plenary Lecture 2

Mathematical Models for Fostering the Sustainable Energy Conversion in Electric Power Systems

Professor Cornelia Aida Bulucea
Faculty of Electrical Engineering
University of Craiova
ROMANIA
E-mail: abulucea@em.ucv.ro

Abstract: To address meaningfully many of the problems facing electric power systems, conditions for the performance of sustainable electric systems must be formulated. Correspondingly, mathematical models can help understand the efficiencies of electrical systems and guide improvement efforts. Costs should reflect value, which is doubtless associated with sustainability aspects, and the benefits of using mathematical models to improve the sustainability of systems which convert electric energy are fostered.
In line with this idea, modelling the three-phase electrical transformer could attempt to optimize the efficiency of energy use within the power transformer operation. Hence, the structural diagram method applied to three-phase electric transformers is illustrating the efficiency of energy use within the power transformer operation, by highlighting the interactions and the feedback loops among the different variables (electric currents and magnetic fluxes) which describe the power transformer operation.
Following the notion of sustainable electrically driven systems, mathematical patterns illustrate energy conversion processes during the operation of electric railway vehicles with traction synchronous and induction motors, highlighting the chain of interactions within the main electric equipment. In order to support transport systems' sustainability the operation of electric railway vehicles has been addressed, on electrically driven railway systems supplied from a d.c. or an a.c. contact line. This presentation supports the findings that electric traction drive systems using synchronous motors fed by current inverters, and induction motors fed by variable voltage variable frequency (VVVF) inverters enhance the sustainable operation of electric railway trains.

Brief Biography of the Speaker: Cornelia Aida Bulucea is currently an Associate Professor in Electrotechnics, Electrical Machines and Environmental Electric Equipment in the Faculty of Electrical Engineering, University of Craiova, Romania. She is graduate from the Faculty of Electrical Engineering Craiova and she received the Ph.D degree from Bucharest Polytechnic Institute. In Publishing House she is author of four books in electrical engineering area. Research work is focused on improved solutions for electrical networks on basis of new electric equipment, and environmental impact assessment of electric transportation systems. She has extensive experience in both experimental and theoretical research work, certified by over 70

journal and conference research papers and 15 research projects from industry. Due to WSEAS recognition as huge scientific Forum she participated over time in nineteen WSEAS International Conferences, presenting papers and chairing sessions. She was Plenary Speaker in the 13th International Conference on Electric Power Systems, High Voltages, Electric Machines (POWER'13), Chania, Crete Island, Greece, August 27-29, 2013, in the 5th IASME/WSEAS International Conference on ENERGY&ENVIRONMENT (EE'10), held by the University of Cambridge, UK, February 23-25, 2010, in the 4th IASME/WSEAS International Conference on ENERGY&ENVIRONMENT (EE'09),), held by the University of Cambridge, Cambridge UK, February 24-26, 2009, in the 8th WSEAS International Conference on POWER SYSTEMS (PS'08), held by the University of Cantabria, Santander, Spain, September 23-25, 2008. She is very proud by her over 30 papers published in the WSEAS Conferences Books and in the WSEAS TRANSACTIONS ON ENVIRONMENT AND DEVELOPMENT, WSEAS TRANSACTIONS ON POWER SYSTEMS, WSEAS TRANSACTIONS ON CIRCUITS AND SYSTEMS and WSEAS TRANSACTIONS ON ADVANCES IN ENGINEERING EDUCATION.

ISBN: 978-1-61804-256-9

Plenary Lecture 3

Synchronization of Neural Systems and Sensing of Acoustic Events in the Time - Space Domain

Professor Atsushi Fukasawa
Institute of Statistical Mathematics
Japan
E-mail: takizawa@ism.ac.jp

Abstract: Synchronization is an essential condition to organize large systems for computation, control, and telecommunication. Synchronization provides the system with stability and reliability. This condition should be applied to neural systems corresponding to the above functions. This lecture focuses on the study of neural systems operating under synchronous condition.
He will first present electro-physical modeling of an excitatory neuron. This model is given as the common structure of excitatory cells of paramecium (unicellular organism), neuron, and active semiconductor device. The commonality of these elements gives a unified structure with three zones in electrolyte or solid state semiconductor.
He will then present the principle of systematization of synchronous neural systems. This system was composed of recurrent connection of neurons. The system was applied for sensing of acoustic events in time - space domain of 2D plane and 3D space.

Brief Biography of the Speaker: Atsushi Fukasawa received the Master of Arts degree in Electrical communication and the Ph.D. degree from Waseda University in 1967 and 1983. He joined Graduate School of Science and Technology, Chiba University as a professor in 1997. He received the Award of the Agency of Science and Technology, Japan in 1982, and Ohm (publisher) Prize in 1994. He received Telecommunication System Technology Prize from the Foundation of Telecommunication Association, Japan in 2004. He is a senior member of the IEEE.

ISBN: 978-1-61804-256-9

Plenary Lecture 4

Simulation and Design of a Lactose to Ethanol Conversion Unit in Cyprus

Professor Vassilis Gekas
Co-authors: George Botsaris, Alexandros Koulouris, Ioanna Christodoulidou, Photis Papademas
Department of Agriculture, Biotechnology and Food Science
Cyprus University of Technology
Limassol, Cyprus
E-mail: vassilis.gekas@cut.ac.cy

Abstract: Whey is produced as a byproduct following the halloumi cheese manufacturing process. Conversion of lactose, contained into the whey, to bioethanol would have a twofold benefit; (i) at getting rid of an environmental pollution problem andalso (ii) producing a biofuel thus contributing to the renewable energy balance of Cyprus. SuperPro Designer®, a simulation software, was used to run the model simulations because it contains a set of unit procedures that can be customized to the specific modeling needs throughout the lactose-to-ethanol conversion processes. This paper will attempt to first discuss quantitative and qualitative data of lactose production from whey, followed by the application of ethanol-plant simulation models that will be applied in order to convert lactose into biofuel. Finally, an example of an economic analysis generated by SuperPro® Designer will be presented assessing the financial feasibility of the proposed operation. The optimal conditions which such a unit can operate are also demonstrated in an attempt in increase the efficiency and efficacy of the proposed operation.

Brief Biography of the Speaker: VG was born in 1948 in Larissa of Greece, and graduated from the School of Chemical Engineering at the National Technical University of Athens (NTUA) in 1971. He fulfilled his military obligations 1971-1973 in the Greek army and then he worked in the Greek Industry for ten years. .He has specialized in the area of Food Engineering at the University of Lund in Sweden . obtaining his Master Thesis (1986) in “Lactose Hydrolysis towards a mixture of glucose and galactose” and his Ph D thesis (1987) in “Mass Transfer Studies in Ultrafiltration and in Bioreactors”. He did a postdoctoral sejour at the Paul Sabatier University of Toulouse, France, in the academic year 1990/1991, He became Docent (Assistant/Associate) Professor in Sweden in the year 1992. On 23d April 1997, he was elected at the rank of the Professor in the subject of Transport Phenomena at the Department of Environmental Engineering in the Technical University of Crete working there in the period 1998-2010. In the year 2010 he became a Professor at the Department of Agriculture, Biotechnology and Food Science and Technology of the Cyprus University of Technology at the rank of Professor and at the subject of Food Technology and Engineering. His research interests focus on the unit operations concerning the recovery of bioactive substances from aromatic herbs. According to the Web of Science his articles in peer review journal are 69 with a citation index, excluding self-references, 1619 and an h-index of 23. He is also the author of books in DEnglish and Greek.

ISBN: 978-1-61804-256-9

Plenary Lecture 5

The Engine Power and the Exhaust Gas Emissions on an Outboard Engine

Prof. Charalampos Arapatsakos
Department of Production and Management Engineering
Democritus University of Thrace
V. Sofias Street, 67100, Xanthi
GREECE
E-mail: xarapat@pme.duth.gr

Abstract: Outboard engine is a propulsion system for small boats to speedboats. It's consisted of a self contained unit that includes the engine, a gearbox and a propeller. It is called outboard because its entire structure remains on the boat exterior. The primary difference of an outboard engine in operation to other small engines is the inclusion of a driveshaft and propeller and cooling system which relies on water rather than air. Regarding on engine burning times, it is called two or four stroke engine. Each of these two types of engine have their own advantages and disadvantages. However, the most popular type of engine is the four stroke one due to its technological development. Particularly, the propulsion system of the four stroke engine has been designed to be placed on the vessels' transom and it is consisted of a self contained unit that includes the engine, a gearbox and a propeller. In addition to movement outboard engines provide steering control of the boat, as they are designed to rotate on their mounting material and thus control the direction of thrust. The aim of this work, it was built a construction that allows the function of an outboard engine in conditions similar to the factual. In order to measure the performance of the engine power, a prototype measurement procedure was developed. According to this procedure the measurement of the force is made by a direct connection between the engine's rpm and applied load. During the measurements operating characteristics of the engine, as well as the exhaust gases, were recorded. For the measurement of the emitted pollutants, a laboratory protocol and measurement standards defined by 40 CFR 1045 were used.

Brief Biography of the Speaker: Dr Charalampos Arapatsakos is a Greek citizen, who has been born in Athens. He has studied Mechanical Engineering and PhD. He is Professor on Democritus University of Thrace in Greece. Prof C. Arapatsakos has participated in many research programs about renewable sources of energy, gas emissions and antipollution technology. His research domains are mainly on biofuels and their use in internal combustion engines, the power variation from the use of biofuels, the gas emissions, mechanical damages, internal combustion engines, antipollution technology, renewable sources of energy, gas emissions, vehicle design, elements of machines, resistance of materials, technical mechanics, heat transmission.

ISBN: 978-1-61804-256-9

Cubic Spline Interpolation by Solving a Recurrence Equation Instead of a Tridiagonal Matrix

Peter Z. Revesz
Department of Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, Nebraska 68588-0115
Email: revesz@cse.unl.edu
http://cse.unl.edu/revesz
Telephone: (1+) 402 472–3488

***Abstract*—The cubic spline interpolation method is probably the most widely-used polynomial interpolation method for functions of one variable. However, the cubic spline method requires solving a tridiagonal matrix-vector equation with an O(n) computational time complexity where n is the number of data measurements. Even an O(n) time complexity may be too much in some time-ciritical applications, such as continuously estimating and updating the flight paths of moving objects. This paper shows that under certain boundary conditions the tridiagonal matrix solving step of the cubic spline method could be entirely eliminated and instead the coefficients of the unknown cubic polynomials can be found by solving a single recurrence equation in much faster time.**

I. Introduction

Cubic spline interpolation is a widely-used polynomial intepolation method for functions of one variable [2]. Cubic splines can be described as follows. Let f be a function from $\mathcal{R}$ to $\mathcal{R}$. Suppose we know about f only its value at locations $x_0 < \ldots < x_n$. Let $f(x_i) = a_i$. Piecewise cubic spline interpolation of f is the problem of finding the b_i, c_i and d_i coefficients of the cubic polynomials S_i for $0 \leq i \leq n-1$ written in the form:

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \quad (1)$$

where each piece S_i interpolates the interval $[x_i, x_{i+1}]$ and fits the adjacent pieces by satisfying certain smoothness conditions. Taking once and twice the derivative of Equation (1) yields, respectively the equations:

$$S_i'(x) = b_i + 2c_i(x - x_i) + 3d_i(x - x_i)^2 \quad (2)$$

$$S_i''(x) = 2c_i + 6d_i(x - x_i) \quad (3)$$

Equations (1-3) imply that $S_i(x_i) = a_i$, $S_i'(x_i) = b_i$ and $S_i''(x_i) = 2c_i$. For a smooth fit between the adjacent pieces the cubic spline interpolation requires that the following conditions hold for $0 \leq i \leq n-2$:

$$S_i(x_{i+1}) = S_{i+1}(x_{i+1}) = a_{i+1}, \quad (4)$$

$$S_i'(x_{i+1}) = S_{i+1}'(x_{i+1}) = b_{i+1} \quad (5)$$

$$S_i''(x_{i+1}) = S_{i+1}''(x_{i+1}) = 2c_{i+1} \quad (6)$$

This paper is organized as follows. Section II review the usual solution for cubic splines by solving a tridiagonal matrix. Section **??**

II. The Tridiagonal Matrix-Based Solution

In this section we review the usual tridiagonal matrix-based solution for cubic splines. Let $h_i = x_{i+1} - x_i$. Substituting Equations (1-3) into Equations (4-6), respectively, yields:

$$a_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = a_{i+1} \quad (7)$$

$$b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1} \quad (8)$$

$$c_i + 3d_i h_i = c_{i+1} \quad (9)$$

Equation (9) yields a value for d_i, which we can substitute into Equations (7-8). Hence Equations (7-9) can be rewritten as:

$$a_{i+1} - a_i = b_i h_i + \frac{2c_i + c_{i+1}}{3} h_i^2 \quad (10)$$

$$b_{i+1} - b_i = (c_i + c_{i+1}) h_i \quad (11)$$

$$d_i = \frac{1}{3h_i}(c_{i+1} - c_i). \quad (12)$$

Solving Equation (10) for b_i yields:

$$b_i = (a_{i+1} - a_i)\frac{1}{h_i} - \frac{2c_i + c_{i+1}}{3} h_i \quad (13)$$

which implies for $j \leq n-3$ the condition:

$$b_{i+1} = (a_{i+2} - a_{i+1})\frac{1}{h_{i+1}} - \frac{2c_{i+1} + c_{i+2}}{3} h_{i+1} \quad (14)$$

ISBN: 978-1-61804-256-9

Substituting into Equation (11) the values for b_i and b_{i+1} from Equations (13-14) yields:

$$(a_{i+1}-a_i)\frac{1}{h_i}-(2c_i+c_{i+1})\frac{h_i}{3}+(c_i+c_{i+1})h_i = (a_{i+2}-a_{i+1})\frac{1}{h_{i+1}}-(2c_{i+1}+c_{i+2})\frac{h_{i+1}}{3}$$

The above can be rewritten as:

$$h_ic_i+2(h_i+h_{i+1})c_{i+1}+h_{i+1}c_{i+2} = \frac{3}{h_i}a_i-\left(\frac{3}{h_i}+\frac{3}{h_{i+1}}\right)a_{i+1}+\frac{3}{h_{i+1}}a_{i+2}$$

The above holds for $0 \le i \le n-3$. However, changing the index downward by one the following holds for $1 \le j \le n-2$:

$$h_{i-1}c_{i-1}+2(h_{i-1}+h_i)c_i+h_ic_{i+1} = \frac{3}{h_{i-1}}a_{i-1}-\left(\frac{3}{h_{i-1}}+\frac{3}{h_i}\right)a_i+\frac{3}{h_i}a_{i+1} \quad (15)$$

The above is a system of $n-1$ linear equations for the unknowns c_i for $0 \le i \le n$. By Equation (3) $S_0''(x_0) = 2c_0$ and by extending Equation (6) to $j = n-1$, $S_{n-1}''(x_n) = 2c_n$.

The cubic spline interpolation allows us to specify several possible boundary conditions regarding the values of c_0, c_n. A commonly used boundary condition called a natural cubic spline assumes that $c_0 = c_n = 0$, which is equivalent to setting the second derivative of the splines at the ends to zero. Alternatively, in the clamped cubic spline interpolation, the assumed boundary condition is $b_0 = f'(x_0)$ and $b_n = f'(x_n)$ where the derivatives of the f at x_0 and x_n are known constants.

In addition, in sovling a cubic spline a uniform sampling is also commonly assumed and available, that is, each h_i has the same constant value h. Then dividing Equation (15) by h yields:

$$c_{i-1}+4c_i+c_{i+1} = \frac{3}{h^2}(a_{i-1}-2a_i+a_{i+1}) \quad (16)$$

Since the values of a_i are known, the values of c_i can be found by solving the tridiagonal matrix-vector equation $Ax = B$. Under the natural cubic spline interpolation, we have:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{bmatrix}$$

the vector of unknowns is:

$$x = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}$$

and the vector of constants is:

$$B = \begin{bmatrix} 0 \\ \frac{3}{h^2}(a_0-2a_1+a_2) \\ \vdots \\ \frac{3}{h^2}(a_{n-2}-2a_{n-1}+a_n) \\ 0 \end{bmatrix}.$$

Similarly, under the clamped spline interpolation we have:

$$A = \begin{bmatrix} 2 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 2 \end{bmatrix}$$

the same vector of unknowns:

$$x = \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}$$

and the following vector of constants:

$$B = \begin{bmatrix} \frac{3}{h^2}(a_1-a_0)-\frac{3}{h}f'(x_0) \\ \frac{3}{h^2}(a_0-2a_1+a_2) \\ \vdots \\ \frac{3}{h^2}(a_{n-2}-2a_{n-1}+a_n) \\ \frac{3}{h}f'(x_n)-\frac{3}{h^2}(a_n-a_{n-1}) \end{bmatrix}.$$

Both the natural cubic spline and the clamped cubic spline boundary conditions yield a system of $n+1$ linear equations with only $n+1$ unknowns. Such a system normally yields a unique solution except in some special cases. Moreover, either system is a tridiagonal matrix sytem that can be solved in $O(n)$ time. Once the c_i values are found, the d_i and the b_i values also can be found by Equations (12) and (13), respectively. Computating the b_i and d_i coefficients can be done also within $O(n)$ time.

III. A New Recurrence Equation-Based Solution

In our solution to the cubic spline interpolation problem, we chose a boundary condition that requires solving the following tridiagonal system where x_i are rational variables, d_i are rational constants and $r \neq 0$ is a rational constant, and A is:

$$A = \begin{bmatrix} r & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Furthermore,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_{n-1} \\ e_n \end{bmatrix}.$$

A. Relationship to Clamped and Natural Cubic Splines

Our new matrix is closely related to clamped cubic splines. Consider the first equation for the clamped cubic spline, which can be written as:

$$2c_0 + c_1 = \frac{3}{h}\left(\frac{(a_1 - a_0)}{h} - f'(x_0)\right)$$

The above equation becomes the following after multiplying by $r/2$:

$$rc_0 + \frac{r}{2}c_1 = \frac{3r}{2h}\left(\frac{(a_1 - a_0)}{h} - f'(x_0)\right)$$

Adding $(1 - r/2)c_1$ yields:

$$rc_0 + c_1 = \frac{3r}{2h}\left(\frac{(a_1 - a_0)}{h} - f'(x_0)\right) + \left(1 - \frac{r}{2}\right)c_1$$

Hence the first row of our new matrix A is equivalent to first row of the clamped cubic spline for any $r \neq 0$ if e_1 is:

$$e_1 = \frac{3r}{2h}\left(\frac{(a_1 - a_0)}{h} - f'(x_0)\right) + \left(1 - \frac{r}{2}\right)\tilde{c_1}.$$

where $\tilde{c_1}$ is an estimate for the value of c_1.

The last row of the new matrix allows fixing the value of c_n. This is a generalization of natural cubic spline which fixes the value to be 0.

B. A Recurrence Equation-Based Solution

In this section, we solve the new system using the value $r = 2+\sqrt{3} \approx 3.732$. In that case, the first three equations can be written as:

$$\begin{aligned} rx_1 + \ x_2 &= e_1 \\ x_1 + 4x_2 + \ x_3 &= e_2 \\ x_2 + 4x_3 + x_4 &= e_3 \end{aligned}$$

Multiplying the second row by r, subtracting from it the first row, and then dividing it by r gives:

$$\begin{aligned} rx_1 + \ x_2 &= e_1 \\ rx_2 + \ x_3 &= e_2 - \frac{e_1}{r} \\ x_2 + 4x_3 + x_4 &= e_3 \end{aligned}$$

Multiplying now the third row by r, subtracting from it the second row, and then dividing it by r gives:

$$\begin{aligned} rx_1 + \ x_2 &= e_1 \\ rx_2 + \ x_3 &= e_2 - \frac{e_1}{r} \\ rx_3 + x_4 &= e_3 - \frac{e_2}{r} + \frac{e_1}{r^2} \end{aligned}$$

Continuing this process until the last row, we get:

$$\begin{aligned} rx_{n-3} + x_{n-2} &= e_{n-3} - \tfrac{e_{n-4}}{r} + \tfrac{e_{n-5}}{r^2} - \ldots + (-1)^{n-4}\tfrac{e_1}{r^{n-4}} \\ rx_{n-2} + x_{n-1} &= e_{n-2} - \tfrac{e_{n-3}}{r} + \tfrac{e_{n-4}}{r^2} - \ldots + (-1)^{n-3}\tfrac{e_1}{r^{n-3}} \\ rx_{n-1} + \ x_n &= e_{n-1} - \tfrac{e_{n-2}}{r} + \tfrac{e_{n-3}}{r^2} - \ldots + (-1)^{n-2}\tfrac{e_1}{r^{n-2}} \\ x_n &= e_n \end{aligned}$$

Dividing each row except the last one by r yields:

$$\begin{aligned} x_{n-3} + \tfrac{x_{n-2}}{r} &= \tfrac{e_{n-3}}{r} - \tfrac{e_{n-4}}{r^2} + \ldots + (-1)^{n-4}\tfrac{e_1}{r^{n-3}} \\ x_{n-2} + \tfrac{x_{n-1}}{r} &= \tfrac{e_{n-2}}{r} - \tfrac{e_{n-3}}{r^2} + \tfrac{e_{n-4}}{r^3} - \ldots + (-1)^{n-3}\tfrac{e_1}{r^{n-2}} \\ x_{n-1} + \tfrac{x_n}{r} &= \tfrac{e_{n-1}}{r} - \tfrac{e_{n-2}}{r^2} + \tfrac{e_{n-3}}{r^3} - \ldots + (-1)^{n-2}\tfrac{e_1}{r^{n-1}} \\ x_n &= e_n \end{aligned}$$

Note that each row $1 \leq i \leq n-1$ will be the following:

$$x_i + \frac{x_{i-1}}{r} = \sum_{0 \leq k \leq (i-1)} (-1)^k \frac{e_{i-k}}{r^{k+1}}$$

We define the values for α_0, α_i for $1 < i \leq n-1$, and α_n, respectively, as follows:

$$\begin{aligned} \alpha_0 &= 0 \\ \alpha_i &= \frac{e_i - \alpha_{i-1}}{r} = \sum_{0 \leq k \leq (i-1)} (-1)^k \frac{e_{i-k}}{r^{k+1}} \\ \alpha_n &= e_n \end{aligned} \tag{17}$$

The solution to the linear equation system can be described in terms of the α constants as follows:

ISBN: 978-1-61804-256-9

$$\vdots$$

$$x_{n-3} = \alpha_{n-3} - \frac{\alpha_{n-2}}{r} + \frac{\alpha_{n-1}}{r^2} - \frac{\alpha_n}{r^3}$$

$$x_{n-2} = \alpha_{n-2} - \frac{\alpha_{n-1}}{r} + \frac{\alpha_n}{r^2}$$

$$x_{n-1} = \alpha_{n-1} - \frac{\alpha_n}{r}$$

$$x_n = \alpha_n$$

Therefore, x_i for each row $1 \leq i \leq n$ will be:

$$\begin{aligned} x_n &= \alpha_n \\ x_i &= \alpha_{i-1} - \frac{x_{i+1}}{r} \end{aligned} \tag{18}$$

The above can be solved in closed form as follows:

$$x_i = \sum_{0 \leq k \leq (n-i)} \left(\frac{-1}{r}\right)^k \alpha_{i+k} \tag{19}$$

Note that no matter what exactly are the initial values for e, we have pre-solved the system. This can lead to a faster evaluation of the cubic spline than solving the tridiagonal system each time. We need only $O(n)$ multiplications and subtractions to compute the values of all the x_i. Moreover, when any new measurement is made, the conventional tridiagonal matrix-based algorithm requires a complete redo of the entire computation in $O(n)$ time. In contrast, Equation (18) leads to a faster update because to each x_i for $i \leq n$ we need to add only the term:

$$\left(\frac{-1}{r}\right)^{n+1-i} \alpha_{n+1}.$$

We also need to make $x_{n+1} = \alpha_{n+1}$.. Afterward updating the other α_i constants can be done also similarly efficiently.

C. A Moving Object Example

Suppose that an object is released from a height of 400 feet with zero initial velocity. Suppose also that we measure the object's position to be 384, 336 and 256 feet from earth at one, two and three seconds after release. We also suspect that the object is in free fall with a gravitational acceleration of $32ft/sec^2$ at one second after release and at three seconds after release. Find a cubic spline approximation for the object's position at all times from the release to three seconds after.

We will measure the distance traveled from the release point. The cubic polynomials we need to find for the intervals $[0,1], [1,2]$ and $[2,3]$ can be expressed as follows:

$$\begin{cases} S_0(x) = a_0 + b_0 x + c_0 x^2 + d_0 x^3 \\ S_1(x) = a_1 + b_1(x-1) + c_1(x-1)^2 + d_1(x-1)^3 \\ S_2(x) = a_2 + b_2(x-2) + c_2(x-2)^2 + d_2(x-2)^3 \end{cases}$$

We have $n = 4$, $a_0 = 400$, $a_1 = 384$, $a_2 = 336$, $a_3 = 256$ and the uniform step size is $h = 1$. By our assumptions of zero initial velocity $f'(0) = 0$ and free fall at one second $c_1 = -16$ and free fall at four seconds $c_3 = -16$, which implies $e_4 = -16$. The matrix A and the vectors x and B are:

$$A = \begin{bmatrix} r & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad x = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} \text{ and } B = \begin{bmatrix} -16r - 16 \\ -96 \\ -96 \\ -16 \end{bmatrix}$$

$$\text{because } B = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix} = \begin{bmatrix} \frac{3r}{2}(-16) + \left(1 - \frac{r}{2}\right)(-16) \\ 3(400 - (2 \times 384) + 336) \\ 3(384 - (2 \times 336) + 256) \\ -16 \end{bmatrix}$$

By Equation (17), we have:

$$\alpha_1 = \frac{e_1}{r} = -16 - \frac{16}{r}$$

$$\alpha_2 = \frac{e_2 - \alpha_1}{r} = -16 - \frac{16}{r}$$

$$\alpha_3 = \frac{e_3 - \alpha_2}{r} = -16 - \frac{16}{r}$$

$$\alpha_4 = e_4 = -16$$

By Equation (18) we also have when calculating in reverse order:

$$c_3 = \alpha_4 = -16$$

$$c_2 = \alpha_3 - \frac{c_3}{r} = -16$$

$$c_1 = \alpha_2 - \frac{c_2}{r} = -16$$

$$c_0 = \alpha_1 - \frac{c_1}{r} = -16$$

Solving for the b_i coefficients by Equation (13) gives:

ISBN: 978-1-61804-256-9

$b_0 = \frac{1}{1}(384 - 400) - \frac{1}{3}(-16 - 32) = 0$

$b_1 = \frac{1}{1}(336 - 384) - \frac{1}{3}(-16 - 32) = -32$

$b_2 = \frac{1}{1}(256 - 336) - \frac{1}{3}(-16 - 32) = -64$

Solving for the d_i coefficients by Equation (12) gives:

$$d_0 = \frac{1}{3}(-16 - (-16)) = 0$$

$$d_1 = \frac{1}{3}(-16 - (-16)) = 0$$

$$d_2 = \frac{1}{3}(-16 - (-16)) = 0$$

The above values show that an object in free fall has an increasing velocity but its acceleration remains constant. Using the above values, the cubic spine interpolation can be described as:

$$\begin{cases} S_0(x) = 400 - 16x^2 \\ S_1(x) = 384 - 32(x-1) - 16(x-1)^2 = 400 - 16x^2 \\ S_2(x) = 336 - 64(x-2) - 16(x-2)^2 = 400 - 16x^2 \end{cases}$$

Hence in each piece the cubic spline interpolation gives $400 - 16x^2$, which agrees with the expected physics equation for the position of a moving object that strarts with zero velocity from an elevation of 400 feet and freely falls downward with an acceleration of $32ft/sec^2$.

IV. Conclusion

The general method described in this paper can be used in a wide variety of applications which require interpolation of a function of one variable. For example, interpolation of measurement data can generate constraint databases that can be efficiently queried using constraint query languages [5], [6]. The simple one-variable function interpolation can be also extended to higher dimensions yielding interpolations of higher-dimensional functions that describe surfaces [4] and three-dimensional spatio-temporal or moving objects [1], [3]. This extension remains an interesting future work.

References

[1] S. Anderson and P. Z. Revesz, Efficient MaxCount and threshold operators of moving objects, *Geoinformatica*, 13 (4), 355–396, 2009.

[2] R. L. Burden and J. D. Faires, *Numerical Analysis*, 9th ed. New York, USA: Springer, 2014.

[3] J. Chomicki and P. Z. Revesz, Constraint-based interoperability of spatiotemporal databases, *Geoinformatica*, 3 (3), 211-243, 1999.

[4] L. Li and P. Z. Revesz, Interpolation methods for spatio-temporal geographic data, *Computers, Environment and Urban Systems*, 28 (3), 201-227, 2004.

[5] P. C. Kanellakis, G. M. Kuper and P. Z. Revesz, Constraint query languages, *Journal of Computer and System Sciences*, 51 (1), 26-52, 1995.

[6] P. Z. Revesz, *Introduction to Databases: From Biological to Spatio-Temporal*, New York, USA: Springer, 2010.

Geometric construction of wavefronts

S. Thanasoulas, D. A. Pliakis, T. Papakostas, P. Soupios

Abstract—The propagation of scalar waves in an inhomogeneous isotropic medium is formulated mathematically as the problem of the wave equation in curved space equipped with metric conformal to the euclidean metric. The geometric optics approximation studies the motion of a free particle in this space and hence the rays represent the geodesic motion. In the geodesic motion the rays move while remain orthogomal to geodesic spheres: we provide here curvature estimates and a method of construction of geodesic spheres taking into account the conformal factor (assumed an analytic function) of the metric which is directly related to the refractive index or the physical characteristics of the medium.

I. Scalar wave quation in an inhomogeneous medium

We want to study the propagation of waves in an inhomogeneous medium. These are described as solutions of wave equation

$$\Delta_3 u = e^{\sigma} u_{tt}$$

with specific initial data. Howvere we are interested in monochromatic waves, i.e.

$$u(\underline{x}, t) = e^{i\omega t} v(\underline{x})$$

with v satisfying the usual radiation conditions in far distances. Moreover the equation is put it in the form of Schródinger equation

$$(\Delta_\sigma + V)\, v = -\omega^2 e^{\sigma} v$$

where

$$V = \frac{3}{4}\left(|\nabla\sigma|^2 - \frac{3}{4}\Delta_0\sigma\right),$$

and Δ_σ is the Laplace-Beltrami for the metric e^σ. High frequancy solutions of equations of the above form are searched as superpositions

$$u = \sum_{j=1}^{N} \chi_j u,$$

where $\{\chi_j\}$ is a partition of unity supported in geodesic balls and moreover

$$\chi_j u = A_j(\underline{x}) e^{i\omega\phi_j(\underline{x})}$$

and

$$A_j \sim \sum_{k=1}^{\infty} A_{j,k}\omega^{-k}$$

We derive the usual equations of geometric optics, eikonal and transport:

$$|\nabla\phi|^2 = \frac{1}{\omega^2} V + e^{\sigma} \qquad \text{(EE)}$$

This project is implemented through the Operational Program Education and Lifelong Learning, Action Archimedes III and is co-financed by the European Union (European Social Fund) and Greek national funds (National Strategic Reference Framework 20072013), project title: Interdisciplinary study for exploring, understanding and management of groundwater resources. Pilot field investigation North-Western and Central Crete (AQUADAM).

$$\nabla\phi \cdot \nabla A_k - \Delta_\sigma \phi A_k = \frac{1}{i\omega}\Delta_\sigma A_{k-1} \qquad \text{(TE)}$$

II. Inhomogeneous isotropic media

We recall the standard constructions of differential geometry that are employed in the solution of the eikonal equation. The crucial geometric objects are the geodesic spheres which are the level sets of the distance function.

The space we consider is a domain in three dimensional space $\Omega \subset \mathbf{R}^3$ equipped with the metric $e^\sigma \times e, e = dx_1^2 + dx_2^2 + dx_3^2$ where the conformal factor $\sigma : \Omega \to \mathbf{R}$ is smooth . The Ricci curvature completely determines the curvature of the space Ω and it is given by the following formula:

$$R_{ij} = \frac{1}{2}\left(|\nabla\sigma|^2 - \Delta\sigma\right)\delta_{ij} + \frac{1}{2}\left(\sigma_i\sigma_j - \sigma_{ij}\right)$$

while the scalar curvature R is given by

$$R = 2|\nabla\sigma|^2 - \frac{3}{4}\Delta\sigma$$

A. Solution of the eikonal equation

The geodesic spheres in Ω centered at the point $\underline{O} \in \mathbf{R}^3$ and of radius r are characterized as the 2 dimensional surfaces:

$$S^2_{\underline{O},r} = \{\underline{x} \in \Omega : d(\underline{x}, \underline{O}) = r\}$$

where $d(\underline{O}, \cdot)$ is the length of the geodesic curve emanating from the point $\underline{O}$ with unit speed. Alternatively the geodesic spheres are the level sets of the distance function from a given point $O \in \mathbf{R}^3$ $\rho(\underline{x}) = d(\underline{x}, O)$ and solves the eikonal equation

$$e^\sigma |\nabla\rho|^2 = 1$$

with initial values that the small eucliean sphere

$$|\underline{x}| = \epsilon, \quad \rho = \epsilon$$

The solution is given through the method of characteristics and the solution of the following system:

$$\frac{d\underline{x}}{ds} = 2e^\sigma \underline{p}, \quad \frac{d\underline{p}}{ds} = -e^\sigma |p|^2 \nabla\sigma$$

This system has the following integral

$$e^\sigma |\underline{p}|^2 = E$$

which leads to

$$\frac{d\underline{x}}{ds} = 2e^\sigma \underline{p}, \quad \frac{d\underline{p}}{ds} = -E\nabla\sigma$$

This system is solved with Euler schemes and the solution of the eikonal equation is obtained by an inversion, a method that is computationally very costly. Specifically in a grid of N cells this system shows complexity $O(N^3)$.

ISBN: 978-1-61804-256-9

B. Alternative method: perturbed spherical fronst

Therefore we proceed to the construction of the geodesic spheres as perturbation of the euclidean sphere through the solution of a Schrödinger equation derived in differntial geometry, the potential is gieven through the norm of the second fundamental form. Specifically let $\varphi : \mathbf{S}^2 \to \mathbf{R}$ be a function with its graph defining the preturbation of the sphere. Then this satisfies

$$\Delta_2\varphi + |A|^2\varphi = 2(h - h_0) + Q$$

where A, h are the second fundamental form and mean curvature of the perturbed sphere while $h_0 = \frac{2}{r}$. The term Q involves qubic terms with factors depending on $\phi A, \nabla A, \nabla\phi, \nabla^2\phi$.

A Schauder fixed point theorem is used to show the existence of a solution of this nonlinear equation provided suitable a priori estimates are obtained, ([GT]). However we solve it numerically with an iterative algorithm. For this we derive a priori estimates weighted with the conformal factor using the Gauss-Codazzi equations. This method shows complexity of order $O(N^2)$

1) Gauss, Codazzi equations: The geodesic spheres are characterized by the intrinsic curvature function, denoted by κ and their second funedamental form A which is a symmetric 2-covariant tensor field. The metric $e^\sigma \cdot e$ is written in geodesic polar coordinates centered at a point $\underline{O}$ as:

$$g = dr^2 + \gamma(r)$$

where $\gamma(r)$ is a riemannian metric on the geodesic sphere $S^2_{\underline{O},r}$ with second fundamental form, mean curvature and curvature function respectively A, h, κ. The coordinates are defined by the preceding system of ODE 's and jacobian between the two system of coordinates involve the Jacobian vectro fields and hence the curvature of $E^\sigma e$ given above. Accordingly we have the first variation equation for the metric γ:

$$\frac{d\gamma}{dr} = 2A, \quad \text{(FV)}$$

while the second variation equations are for the radial direction standing for $0-$index:

$$\frac{dA_{ij}}{dr} = R_{i0j0} - A_{im}A^m_j, \quad \text{(SV)}$$

These imply in particular that the radial variation of the volume is given by $\alpha = \det(\gamma))$ that

$$\frac{d\alpha}{dr} = 2h\alpha, \quad (\text{FV}_0)$$

The Gauss equations that relate the curvature of γ

$$\kappa = \overline{R}_{1212} = \overline{R}_{11} = \overline{R}_{22} = -\overline{R}_{12}$$

to the ambient curvature

$$\kappa - \det(A) = R_{1212}, \qquad (\text{G})$$
$$\kappa + A_{ij}h - A_{im}A^m_j = R_{ij}, \qquad (\text{G}_{01})$$
$$\kappa - \det(A) = R - R_{00}, \qquad (\text{G}_{02})$$

Combinig we derive the crucial identity

$$|A|^2 \leq C(h^2 + |Ric| + R)$$

Moreover we have the Codazzi equations that give that:

$$A_{jm;i} - A_{im;j} = R_{m0ij}$$

These give us the following Hodge type system on the geodesic sphere for A:

$$\text{curl}(A) = A_{mij} = A_{mi;j} - A_{mj;i} = A_{m0ij}$$
$$\text{div}(A) = A^j_{i;j} - A_i = R_{0i}$$

Using the preceding and the commutation rules for covariant differentation we can prove the inequality for a cut-off function ζ on the unit sphere supported in a given disc in the geodesic sphere:

$$\int_S |\nabla(\zeta A)|^2 \leq C\int_S |\nabla\zeta|^2|A|^2 + |Ric|^2\zeta^2$$
$$\int_S |\nabla(\zeta h)|^2 \leq C\int_S |\nabla\zeta|^2h^2 + |Ric|^2\zeta^2$$

The latter allows to derive bounds for

$$\sup_{\tilde{\Omega}} |A|, \quad \sup_{\tilde{\Omega}} |h|, \quad \inf_{\tilde{\Omega}} |A|, \quad \inf_{\tilde{\Omega}} |h|$$

in sets $\tilde{\Omega}$ that are characterized by the fact that ambient curvature or the conformal factor has specific values for $\vartheta \in (0, 1)$:

$$\tilde{\Omega}_{\eta,\vartheta,r}(\psi) = \{\underline{x} \in S^2_{\underline{O},r} : \vartheta\eta \leq \sigma(\underline{x} \leq \eta\}$$

where σ approximates the values of the conformal factor near the sphere $S^2_{\underline{O},r}$. Similarly we obtain estimates for the solution $\sup_{\tilde{\Omega}} |\nabla^j\varphi|$ as given in [P2],[PM],[PP].

III. The construction

We will approximate the geodesic balls which for r close in the σ sense, for $M > 0$ sufficiently big:

$$r|\nabla\sigma| \leq 10^{-M}|\sigma|, \qquad r^2|\nabla^2\sigma| \leq 10^{-M}|\sigma|$$

We start with the round sphere and contruct *bumps over dics* on it. Therefore we fix the parameters

- Bump size ϵ_j
- Bump center coordinates $\{\underline{\nu}_j\}_{j=1}^N \subset S^2, \varphi_j, N$ where

$$\underline{\nu}_j = \frac{1}{1 + \frac{\eta_j^2}{4}}(2\eta_{1j}, 2\eta_{2j}, 1 - \frac{\eta_j^2}{4})$$

for the vectors $\{\underline{\eta}_j\}_{j=1}^N$ in the disc $\{\underline{\eta} \in \mathbf{R}^2 : |\underline{\eta}| < 2\}$.

A. One bump

However instead of using the mean curvvature equation we select to solve the the Schrödinger equation on a disc

$$\Delta_v + av = h$$

with Dirichlet boundary values on the disc $\{\underline{\xi} \in \mathbf{R}^2 : |\underline{\xi} \leq \epsilon_j\}$. where we use the polynomial functions :

$$h_j : \mathbf{R}^2 \to \mathbf{R}, h_j(r;\xi) = \sum_{\ell_1+\ell_2=m_j}^{d_j} h_{\ell_1\ell_2}(r)\xi_1^{\ell_1}\xi_2^{\ell_2}$$

ISBN: 978-1-61804-256-9

for $0 \leq r \leq \epsilon$ that approximate the mean curvature mear the disc at the northern pole. The bump is given by the function $u = r^2 - \frac{\xi^2}{4} + v$ near an ϵ_j disc of the sphere centered around the northern pole The bumped sphere has the parametrization near the north pole

$$\underline{X}(r;\underline{\xi}) = \kappa(\xi_1, \xi_2, u)$$

Finally we form the vector field for

$$\underline{N}(r;\underline{\xi}) = (\kappa(u\xi_1+\psi\xi_2)-r^2u_1, \kappa(\psi\xi_1-u\xi_2)+r^2u_2, \kappa\underline{\xi}^2-r^2)$$

where

$$\psi = -(\underline{\xi} \times \nabla u) \cdot \underline{e}_3 = \xi_2 u_1 - \xi_1 u_2$$

and

$$\kappa = \frac{2r^2}{\underline{\xi}^2 + 4r^2}$$

a) Several bumps: We form the rotation matrix

$$\mathscr{R}(\underline{\nu}_j) = -\cos\theta_j \mathbf{1} + \sin\theta_j L_j + \cos\theta_j \Pi(\underline{\nu}_j)$$

where the projector matrix has entries

$$\Pi_{rs}(\underline{\nu}_j) = \nu_{rj}\nu_{sj}, \quad r,s = 1,2,3$$

and the infinitesimal rotation

$$L_j = \begin{pmatrix} 0 & \nu_{3j} & -\nu_{2j} \\ -\nu_{3j} & 0 & \nu_{1j} \\ \nu_{2j} & -\nu_{1j} & 0 \end{pmatrix}$$

Then we produce more bumps by rotating the vectors $\underline{X}$ by $\underline{X} \mapsto \mathscr{R}(\underline{\nu}_j)\underline{X}, \underline{N} \mapsto \mathscr{R}(\underline{\nu}_j)\underline{N}$.

IV. Conclusion

The wave equation in an inhomogeneous equation in the high frequency limit requires the construction of the phase function as solution of the eikonal equation. The eikonal equation is solved by integrating the characteristic ODE's and then inverting the solution, this method is costly. Here instead we construct the phase function by considering the phase fronts as perturbations of the sperical ones and hence we use a nonlinear elliptic equation that we solve iteratively. We derived a priori estimates for this equation that facilitate the design of a parallel scheme for its solution.

References

[BW] Born M., Wolf E., Principles of Optics, CUP, 7th edition, (1999)

[Bl1] Bleistein N, Mathematical models for wave phenomena, Academic press, New York

[Bl2] Bleistein N., mathematics of modeling, Migration and Inversion with Gaussian beams, (2008)

[CVS] Cerveny V., Soares J. E., *Fresnel volume ray tracing*, Geophysics, **57**, (1992), 902-915.

[GT] Gilbarg D. Trudinger N., Elliptic Partial differential equations of the second order, (1983), Springer Berlin Heidelberg New York

[KO] Kravstov, Yu.A., Orlov, Yu.I. Geometric optics of inhomogeneous media, (1980) Nauka, Moscow (in Russian).

[N] Nolet, G., *Seismic wave propagation and seismic tomography* in *Seismic Tomography with Applications in Global Seismology and Exploration Geophysics*, , ed. Nolet G, (1987), S23, Reidel, Dordrecht.

[PM] Pliakis D., Minardi S., *A phase retrieval method*, JOSA A, **26**(1), (2009), 99-107

[P1] Pliakis D., *Generalized Hardy's inequalities*, Asian J. Math, **18**, (3)

[P2] Pliakis D., *Nodal volume estimates*, available at `arXiv:1304.7143`

[PP] T Papakostas, D Pliakis *Local interior estimates for the Euler-Einstein system*, Journal of Physics: Conference Series 453 (1), 012010

[S] Soupios, P. M., Papazachos, C. B., Juhlin, C. Tsokas, G. N., *Nonlinear Three Dimensional Traveltime Inversion of Crosshole Data With an Application in the Area of Middle Urals*, Geophysics, **66**, (2001), 627-636.

[V] Virieux J., *Seismic Ray tracing* Seismic Mod. Eart Str., Ed. E. Boschi, G. Ekström, A. Morelli, ING, (1995), 223-304

[VF] Virieux, J., and Farra, V., *Ray tracing in 3-D complex isotropic media: An analysis of the problem.* Geophysics **56**, (1991), 2057-2069.

[V2] Vidale, J.E., *Finite-difference calculations of travel times in three dimensions* Geophysics, **55**, (1990),521-526.

[WE] Wielandt E, *On the validity of the ray approximation for interpreting delay times*, in *Seismic Tomography with Applications in Global Seismology and Exploration Geophysics*, ed. Nolet G, (1987) Reidel, Dordrecht.

Spyros Thanasoulas is PhD candidate in Physics in Queen's University of Belfast. He is a specialist in network security working in the Colorado State University. He is also founding partner of ADITAL Software.

Demetrios A. Pliakis studied Physics in Thessaloniki, Mathematical Physics in Marseille and received his Doctorate in Partial Differential equations from the University of Crete under C. Callias. He was a Marie Curie Fellows in Berlin (Humboldt Uni.) and in TEI Crete. He is a founding partner of ADITAL Software

Taxiarchis Papakostas studied Physics in Thessaloniki and Mathematical Physics in Un. Libre Bruxelles where he received his doctoral degree on Einstein's equations. His Professor of Applied Mathematics in the department of Electrical engineering of TEI Crete.

Pantelis Soupios studied Geology in Thessaliniki where he received his Doctorate in Seismic Tomography. Since 2002 he is in department of Environmental Engineering of TEI Crete where he is now Professor of Geophysics and Hydrogeology. He has written over 100 papers in Geophysics.

The linear complexity and the autocorrelation of generalized cyclotomic binary sequences of length $2^n p^m$

Vladimir Edemskiy, Olga Antonova

***Abstract*—In this article, we generalize results about binary sequences of length $2p^m$ and evaluate the linear complexity and autocorrelation properties of generalized cyclotomic binary sequences of length $2^n p^m$. We show that in most cases these sequences have high linear complexity and poor autocorrelation performance.**

***Index Terms*—Autocorrelation, linear complexity, generalized cyclotomic binary sequences**

I. Introduction

THE linear complexity of a sequence is an important characteristic of its quality. The linear complexity may be defined as the length of the shortest linear feedback shift register that is capable of generating the sequence [11]. Knowledge of just $2L$ consecutive digits of the sequence is sufficient to enable the remainder of the sequence to be constructed. Thus, it is reasonable to suggest that 'good' sequences have $L > N/2$ (where N denotes the period of the sequence) [1]. Autocorrelation is also important for many practical applications [1], [6]. Ideally, good sequences combine autocorrelation properties of a random sequence with high linear complexity.

Classical cyclotomic classes and generalized cyclotomic classes can be used to construct binary sequences, which are called classical cyclotomic sequences and generalized cyclotomic sequences respectively [1]. C. Ding and T. Helleseth first bring in a new generalized cyclotomy of order 2 with respect to $p_1^{e_1} \dots p_t^{e_t}$, which includes classical cyclotomy as a special case, and subsequently show how to design binary sequences based on this new construction. T. Yan et al. [13], Y. J. Kim et al. [10], and S. Y. Jin et al. [8] studied the linear complexity and autocorrelation properties of generalized cyclotomic binary sequences of length p^m (see, also the article [4]). J. W. Zhang et al. [14] proposed two generalized cyclotomic sequences of length $2p^m$ with high linear complexity. The results of J. W. Zhang et al. were generalized in [5]. Later, P. Ke et al. [9] represented new generalized cyclotomic binary sequences with period $2p^m$, which included constructions proposed by J. W. Zhang [14] as a special case. P.Ke et al. [9] discussed sequences given by defining a vector. We believe that the investigation of sequences of length $2p^m$ defined by two vectors, and generalized cyclotomic sequences of length $2^n p^m$ is interesting from a theoretical point of view. In this paper we evaluate the linear complexity and the autocorrelation function of sequences of length $2p^m$ defined by two vectors, and generalized cyclotomic sequences of length $2^n p^m$. We show that in this case the pattern noted by P. Ke et al [9] persists. In most examined cases new generalized cyclotomic binary sequences have high linear complexity, but do not have desirable autocorrelation properties.

V. Edemskiy is with the Department of Applied Mathematics and Information Science, Novgorod State University, Veliky Novgorod, Russia, 173003 e-mail: Vladimir.Edemsky@novsu.ru.

O. Antonova is with Novgorod State University. e-mail: antonova2906@yandex.ru

II. Generalized cyclotomic sequences

Let p be an odd prime, and let θ be a primitive root modulo p^m, where m is a positive integer [7]. Denote $H_0 = \langle \theta^2 \rangle \pmod{p^m}$ and $H_1 = \theta H_0 \pmod{p^m}$, where H_0 and H_1 are generalized cyclotomic classes of order two with respect to p^m [2]. Here and hereafter $a \bmod p$ denotes the least nonnegative integer that is congruent to a modulo p.

If A is a subset of $\mathbb{Z}_{p^m}$, then let us put by definition $bA = \{ba \mod p^m | a \in A\}$ and

$b + A = \{(b + a) \mod p^m | a \in A\}$, where $b \in \mathbb{Z}$.

Then, we have partitions

$$\mathbb{Z}_{p^m}^* = H_0 \cup H_1 \text{ and } \mathbb{Z}_{p^m} = \bigcup_{k=0}^{m-1} (p^k H_0 \cup p^k H_1) \cup \{0\}. \quad (1)$$

Let $N = 2^n p^m$, where n is a positive integer. The ring residue classes $\mathbb{Z}_N \cong \mathbb{Z}_{2^n} \times \mathbb{Z}_{p^m}$ relative to isomorphism $\phi(a) = (a \bmod 2^n, a \bmod p^m)$ [7]. Let $L_j = (i_0^{(j)}, i_1^{(j)}, \dots, i_{m-1}^{(j)})$, $j = 0, 1, \dots, 2^n - 1$ be a set of any fixed binary vectors in $\mathbb{Z}_2^m$ and $E_j = \bigcup_{k=0}^{m-1} p^k H_{i_k^{(j)}}$, $C_j = \phi^{-1}(\{j\} \times E_j)$.

By definition, put

$$C = \bigcup_{j=0}^{2^n-1} C_j \cup \{0, 2p^m, \dots, (2^n - 2)p^m\} \text{ and } \widetilde{C} = \mathbb{Z}_N \setminus C.$$

The generalized cyclotomic binary sequence $S = \{s_i\}$ of length $2^n p^m$ is then defined by

$$s_i = \begin{cases} 1, & \text{if } i \mod N \in C; \\ 0, & \text{if } i \mod N \in \widetilde{C}. \end{cases} \quad (2)$$

The sequence S is balanced by (1) and the definition.

If g is an odd from integers θ and $\theta + p^m$, then g is a primitive root modulo $2p^m$ [7]. Let $D_j = \{g^{j+2t} \bmod 2p^m; t = 0, \dots, p^{m-1}(p-1)/2 - 1\}$, $j = 0, 1$ be cyclotomic classes modulo $2p^m$, and let $\text{ind}_\theta 2$ be a discrete logarithm

ISBN: 978-1-61804-256-9

of 2 base θ in the field $GF(p)$. It is easy to see that $D_j = \phi^{-1}(\{1\} \times H_j)$ and $2D_j = \phi^{-1}\big(\{0\} \times H_{(j+\mathrm{ind}_\theta 2) \mod 2}\big)$. Hence, binary sequences proposed by J.W. Zhang [14] and P. Ke [9] are special cases of S. They were obtained for $L_1 = (1, \ldots, 1)$ and $L_0 = (1, \ldots, 1)$ or $L_0 = (0, \ldots, 0)$ [14], also for $L_0 = L_1 = \mathcal{L}$ [9].

In the next sections we derive the linear complexity and the autocorrelation function of S.

III. The linear complexity of generalized cyclotomic sequences of length $2^n p^m$

It is well known ([1], [11]) that if $\{s_i\}$ is a binary sequence with period N, then the minimal polynomial $m(x)$ and the linear complexity L of this sequence is defined by

$$m(x) = (x^N - 1)/\big(\gcd(x^N - 1, S(x))\big), \qquad L = N - \deg\big(\gcd(x^N - 1, S(x))\big), \quad (3)$$

where $S(x) = s_0 + s_1 x + \ldots + s_{N-1}x^{N-1}$, $S(x) \in GF(2)[x]$.

In our case $N = 2^n p^m$, hence we have $x^{2^n p^m} - 1 = (x^{p^m} - 1)^{2^n}$ in the ring $GF(2)[x]$ and

$$L = N - \deg\big(\gcd\left((x^{p^m} - 1)^{2^n}, S(x)\right)\big). \quad (4)$$

Let α be a primitive root of order p^m of unity in the extension of the field $GF(2)$. Then, according to (3) and (4), in order to find the minimal polynomial and the linear complexity of S it is sufficient to find the zeros of $S(x)$ in the set $\{\alpha^v, v = 0, 1, \ldots, p^m - 1\}$ and determine their multiplicity.

Earlier we derived values of $S(\alpha^v)$ [4]. Now we use the same technique. Let us introduce auxiliary polynomials $S_k(x) = \sum_{i \in p^k H_0} x^i$, $k = 0, 1, \ldots, m-1$.

Lemma 1: If $v \in \mathbb{Z}$, then

$$\sum_{u \in \phi^{-1}(\{l\} \times p^k H_f)} \alpha^{vu} = S_k(\alpha^{v\theta^f})$$

for all $l = 0, 1, ..., 2^n - 1$ and $k = 0, 1, ..., m-1$.

Proof: By definition of α we have $\alpha^u = \alpha^{u \pmod{p^m}}$. Since $\{u \mod p^m \mid u \in \phi^{-1}\big(\{l\} \times p^k H_f\big)\} = p^k H_f$ and $H_f = \theta^f H_0$, then Lemma 1 is proved. ∎

Lemma 2: If $v \in p^h \mathbb{Z}^*_{p^m}$, $h = 0, 1, \ldots, m-1$, then

$$S_k(\alpha^v) + S_k(\alpha^{v\theta}) = \begin{cases} 1, & \text{if } h = m - k - 1; \\ 0, & \text{else.} \end{cases}$$

for $k = 0, 1, \ldots, m-1$.

Proof: By (1) and the definition of the auxiliary polynomial we have $S_k(\alpha^v) + S_k(\alpha^{v\theta}) = \sum_{i \in p^k \mathbb{Z}^*_{p^m}} \alpha^{vi}$. Then

$$S_k(\alpha^v) + S_k(\alpha^{v\theta}) = \begin{cases} 0, & \text{if } k + h \geq m; \\ \sum_{j \in \mathbb{Z}^*_{p^{m-k-h}}} \alpha^{p^{k+h} j}, & \text{if } k + h < m. \end{cases}$$

By the condition, $\alpha^{p^m} - 1 = 0$ and order of α equals p^m, hence we obtain $\sum_{j \in \mathbb{Z}_{p^{m-t}}} \alpha^{jp^t} = 1$. Then

$$\sum_{j \in \mathbb{Z}^*_{jp^{m-t}}} \alpha^{jp^t} = \begin{cases} 1, & \text{if } t = m-1; \\ 0, & \text{else.} \end{cases}$$

Lemma 2 follows from the last formula. ∎

Let

$$I = \{k | \sum_{l=0}^{2^n - 1} i_k^{(l)} \equiv 1 \pmod 2 \text{ and } k = 0, 1, \ldots, m-1\},$$

i.e., the number of zeros and unities in the set $\{i_k^{(l)}, l = 0, \ldots, 2^n - 1\}$ for $k \in I$ is odd. By I^* denote the complement of the set I in the set $\{0, 1, \ldots, m-1\}$.

Theorem 3: Let generalized cyclotomic sequence S be defined by (2). Then $S(\alpha^v) = 0$ for $v = 1, ..., p^m - 1$ if and only if $v \in \bigcup_{k \in I} p^{m-k-1} \mathbb{Z}^*_{p^m}$ for $n = 1$ and $v \in \bigcup_{k \in I^*} p^{m-k-1} \mathbb{Z}^*_{p^m}$ for $n > 1$.

Proof: By Lemma 1 and (1)

$$S(\alpha^v) = 2^{n-1} + \sum_{l=0}^{2^n - 1} \sum_{k=0}^{m-1} S_k\left(\alpha^{v\theta^{i_k^{(l)}}}\right)$$

or

$$S(\alpha^v) = 2^{n-1} + \sum_{k \in I} \big(S_k(\alpha^v) + S_k(\alpha^{v\theta})\big)$$

by Lemma 1 and the choice of I.

Therefore, $S(\alpha^v) = 0$ if and only if $\sum_{k \in I} S_k(\alpha^v) + S_k(\alpha^{v\theta}) = 1$ for $n = 1$ and $\sum_{k \in I} S_k(\alpha^v) + S_k(\alpha^{v\theta}) = 0$ for $n > 1$. In the first case, by Lemma 2, $v \in p^{m-k-1}\mathbb{Z}^*_{p^m}$ for $k \in I$, and in the second case $v \in p^{m-k-1}\mathbb{Z}^*_{p^m}$ for $k \notin I$. Theorem 3 is proved. ∎

Corollary 4:

$$|\{v | S(\alpha^v) = 0, v = 0, 1, \ldots, p^m - 1\}| = \begin{cases} \sum_{k \in I} p^k (p-1), & \text{if } n = 1; \\ \sum_{k \in I^*} p^k (p-1) + 1, & \text{if } n > 1. \end{cases}$$

Corollary 5: Let generalized cyclotomic sequence S be defined by (2). If $I = \varnothing$ and $n = 1$, then $L = 2p^m$. Also, if $I = \{0, 1, \ldots, m-1\}$ and $n > 1$, then $L \geq 2^n p^m - 2^n$.

All sequences satisfying conditions of Corollary 5 have high linear complexity. Moreover, if the set of vectors $\{L_j\}$ defining the sequence is such that $m - 1 \notin I$ for $n = 1$ and $m - 1 \in I$ for $n > 1$, then $\sum_{k \in I(I^*)} p^k (p-1) \leq p^{m-1} - 1$. By Corollary 4 and (4) we see that $L \geq 2^n p^m - 2^n p^{m-1}$, i.e., $L > N/2$.

In order to refine the estimate of the linear complexity, we investigate the multiplicity of the zeros α^v of $S(x)$. For this purpose let us examine the derivative of $S(x)$. Since $\left(\sum_{i \in \phi^{-1}(\{l\} \times H_{i_k^{(l)}})} x^i\right)' = 0$ when l is even, then

$$S'(\alpha^v) = \alpha^{-v} \sum_{t=0}^{2^{n-1}-1} \sum_{k=0}^{m-1} \sum_{i \in \phi^{-1}(\{2t+1\} \times H_{i_k^{(2t+1)}})} \alpha^{v\theta^{i_k^{(2t+1)}}}. \quad (5)$$

It is obvious from (5) that the analysis of $S'(\alpha^v)$ substantially differs in cases $n = 1$ and $n > 1$.

First, let $n > 1$. Define the set

$$J = \{k | \sum_{t=0}^{2^{n-1}-1} i_k^{(2t+1)} \equiv 1 \pmod 2 \text{ and } k = 0, 1, \ldots, n-1\},$$

that is the number of zeros and unities in the set $\{i_k^{(2t+1)}, t = 0, \ldots, 2^{n-1} - 1\}$ is odd for $k \in J$. By J^* denote the complement of J in $\{0, 1, \ldots, m-1\}$.

Lemma 6: If $n > 1$ and α^v is a zero of $S(x)$, then α^v is a multiple zero if and only if $v \in \bigcup_{k\in I^*\cap J^*} p^{m-k-1}\mathbb{Z}_{p^m}$.

Proof: By (5) and the definition of J, we obtain

$$S^{'}(\alpha^v) = \alpha^{-v}\sum_{k\in J}\left(S_k(\alpha^v) + S_k(\alpha^{v\theta})\right),$$

similar as in Theorem 3. Then by Lemma 2, it follows that $S^{'}(\alpha^v) = 0$ if and only if $v \in p^{m-k-1}\mathbb{Z}_{p^m}$ for $k \in J^*$. So, the statement of Lemma 6 follows from Theorem 3. ∎

From Theorem 3 and Lemma 6, we get the following estimate:

$$L \geq 2^n p^m - \sum_{k\in I^*\setminus J^*} p^k(p-1) - 2^n \sum_{k\in I^*\cap J^*} p^k(p-1) - 2^n.$$

Hence, if $n > 1$, then it is easy to find out for which defining vectors the sequence S has high linear complexity.

Let $n = 1$ and symbols be the same as before. Without loss of generality, we can assume that $S_{m-1}(\alpha) \neq 0$.

Theorem 7: If generalized cyclotomic sequence S is defined by (2), then

1) $L = 2p^m - \sum_{k\in I} p^k(p-1)$ and $m(x) = \left(x^{2p^m}-1\right)/G(x)$, if $p \equiv \pm 3 \pmod 8$. Here $G(x) = \prod_{k\in I}\left(x^{p^{k+1}}-1\right)/\left(x^{p^k}-1\right)$.

2) $L = 2p^m - 1.5\sum_{k\in I} p^k(p-1)$ and $m(x) = \left(x^{2p^m}-1\right)/\left(G(x)\prod_{k\in I}\prod_{u\in p^{m-k-1}H_{f_k}}(x-\alpha^v)\right)$, if $p \equiv \pm 1 \pmod 8$.

Here

$$f_k = \begin{cases} i_k^{(1)}, & \text{if } (m-k) \text{ is even and } p \equiv -1 \pmod 8; \\ 1 - i_k^{(1)}, & \text{else.} \end{cases}$$

Proof: If $n = 1$ then $S^{'}(\alpha^v) = \alpha^{-v}\sum_{k=0}^{m-1} S_k(\alpha^{v\theta_k^{(1)}})$. By Lemma 2 [4] we have

$$S^{'}(\alpha^v) = \alpha^{-v}\left(S_{m-1}\left(\alpha^{\theta^{i_k^{(1)}+f}}\right) + (m-k-1)(p-1)/2\right),$$

if $v \in p^{m-k-1}H_f$ or

$$S^{'}(\alpha^v) = \alpha^{-v}\left(T\left(\beta^{\theta^{i_k^{(1)}+f}}\right) + (m-k-1)(p-1)/2\right),$$

where $\beta = \alpha^{p^{m-1}}$ and $T(x) = \sum_{j\in H_0 (\mathrm{mod}\ p)} x^j$. The values $T(\beta^v)$ were derived by C. Ding et al [3]. Specifically, if $p \equiv \pm 3 \pmod 8$, then $T(\beta^v) \notin \{0, 1\}$, that is $S^{'}(\alpha^v) \neq 0$, and the first statement follows from Theorem 3. In the case $p \equiv \pm 1 \pmod 8$, according to our assumption $T(\beta) = 1$ and $T(\beta^\theta) = 0$, therefore $S^{'}(\alpha^v) = 0$, if $v \in p^{m-k-1}H_f$ for

$$f = \begin{cases} i_k^{(1)}, & \text{if } (m-k) \text{ is even and } p \equiv -1 \pmod 8; \\ 1 - i_k^{(1)}, & \text{else.} \end{cases}$$

Applying (3) and (4), we conclude the proof of Theorem 7. ∎

Theorems 1 and 2 [14], Theorem 1 [9] are special cases of Theorem 7 ($I = \{0, 1, \ldots, m-1\}, I = \varnothing$, respectively).

In the second case of Theorem 7, if the set of vectors $\{L_k\}$ defining the sequence S is such that $m-1 \in I$, then $L \leq p^m$, i.e., in this case sequences do not have high linear complexity. All results from section 2 have been verified by subjecting various sequences to Berlekamp-Massey algorithm for small values of p, m, and n.

IV. Autocorrelation of generalized cyclotomic sequences of length $2^n p^m$

The periodic autocorrelation function $C_S(\tau)$ of a binary sequence $\{s_i\}$ of period N is defined by

$$C_S(\tau) = \sum_{i=0}^{N-1}(-1)^{s_{i+\tau}+s_i}.$$

In this section, we evaluate the autocorrelation function of generalized cyclotomic binary sequences defined by (2). We measure the autocorrelation function by using known methods [8], [12] and generalized cyclotomic numbers of order 2 with respect to p^h for $h \geq 1$ [2],

$$(u, v)^{(p^h)} = |(H_v^{(p^h)} + 1) \cap H_u^{(p^h)}|,$$

where $H_j^{(p^h)} = \{a \pmod{p^h} | a \in H_j\}$. C. Ding and T. Helleseth [2] showed that

1. If $p \equiv 1 \pmod 4$, then

$$(0,0)^{(p^h)} = p^{h-1}(p-5)/4,$$
$$(0,1)^{(p^h)} = (1,0)^{(p^h)} = (1,1)^{(p^h)} = p^{h-1}(p-1)/4;$$

2. If $p \equiv 3 \pmod 4$, then

$$(0,0)^{(p^h)} = (1,0)^{(p^h)} = (1,1)^{(p^h)} = p^{h-1}(p-3)/4,$$
$$(0,1)^{(p^h)} = p^{h-1}(p+1)/4.$$

As before, let $E_j = \bigcup_{k=0}^{m-1} p^k H_{i_k^{(j)}}$. First we define the difference function $d(j, l, \tau) = |E_j \cap (E_l + \tau)|$.

Lemma 8: If $\tau \in p^f H_a$, $f = 0, 1, \ldots, m-1; a = 0, 1$, then

$$d(j,l,\tau) = \sum_{k=0: i_k^{(j)} = j_k^{(l)}}^{f-1} p^{m-k-1}(p-1)/2 + \left(p^{m-f-1}(p \pm 1) + \delta\right)/4,$$

where

$$\delta = \begin{cases} -4, \text{ if } a = i_f^{(j)}, a \neq i_f^{(l)}, p \equiv 3 \pmod 4 \\ \qquad \text{or } a = i_f^{(j)} = i_f^{(l)}, p \equiv 1 \pmod 4; \\ -2, \text{ if } i_f^{(j)} = i_f^{(l)}, p \equiv 3 \pmod 4 \\ \qquad \text{or } i_f^{(j)} \neq i_f^{(l)}, p \equiv 1 \pmod 4; \\ \ 0, \text{ if } a = i_f^{(l)}, a \neq i_f^{(j)}, p \equiv 3 \pmod 4 \\ \qquad \text{or } a \neq i_f^{(j)}, i_f^{(j)} = i_f^{(l)}, p \equiv 1 \pmod 4). \end{cases}$$

Here we use minus sign if $i_f^{(j)} = i_f^{(l)}$ and plus sign otherwise.

Proof: To simplify the proof, we denote $i_k^{(j)}, i_k^{(l)}$ as h_k and g_k respectively. Then $d(j, l, \tau) = \sum_{k=0}^{m-1} |E_j \cap (p^k H_{g_k} + \tau)|$.

Let us break the last sum into three summands and examine each of them separately.

1) Let $k < f$, then $p^k H_{g_k} + \tau = p^k H_{g_k}$ [4] and by (1)

$$\sum_{k=0}^{f-1} |E_j \cap (p^k H_{g_k} + \tau)| = \sum_{k=0}^{f-1} |p^k H_{h_k} \cap p^k H_{g_k}| = \sum_{k=0: h_k = g_k}^{f-1} p^{m-k-1}(p-1)/2.$$

2) Let $k = f$, then

$$|E_j \cap (p^f H_{g_f} + \tau)| = |p^f H_{h_f} \cap (p^f H_{g_f} + \tau)| + \sum_{k=f+1}^{m-1} |p^k H_{h_k} \cap (p^f H_{g_f} + \tau)|.$$

The first summand from the last sum equals $(g_f + a, h_f + a)^{(p^{m-f})}$. By Lemma 2 [13]

$$|p^k H_{h_k} \cap (p^f H_{g_f} + \tau)| = \begin{cases} (p^{m-k} - p^{m-k-1})/2, & \text{if } a = g_f, p \equiv 1 \pmod 4 \\ & \text{or } a \neq g_f, p \equiv 3 \pmod 4; \\ 0, & \text{else.} \end{cases}$$

Therefore,

$$|E_j \cap (p^f H_{g_f} + \tau)| = \begin{cases} (0, h_f + a)^{(p^{m-f})} + (p^{m-f-1} - 1)/2, \\ \quad \text{if } a = g_f, p \equiv 1 \pmod 4; \\ (1, h_f + a)^{(p^{m-f})} + (p^{m-f-1} - 1)/2, \\ \quad \text{if } a \neq g_f, p \equiv 3 \pmod 4; \\ (g_f + a, h_f + a)^{(p^{m-f})}, \quad \text{else.} \end{cases}$$

3) Let $k > f$, then $p^k H_{g_k} + \tau \subset p^f H_a$ [9] and

$$\sum_{k=f+1}^{m-1} |E_j \cap (p^k H_{g_k} + \tau)| = \sum_{k=f+1}^{m-1} |p^f H_{h_f} \cap (p^k H_{g_k} + \tau)| = (p^{m-f-1} - 1)/2,$$

if $a = h_f$ and $\sum_{k=f+1}^{m-1} |E_j \cap (p^k H_{g_k} + \tau)| = 0$ if $a \neq h_f$.

Summing up results, we obtain the statement of Lemma 8. ∎

Like Y. Sun and H. Shen showed [12], it is simple to see that

$$|C \cap (C + \tau)| = \sum_{u=0}^{2^n-1} |C_u \cap (C_{(u-\tau) \mod 2^n} + \tau)| + |C \cap \{0, \ldots, (2^n - 2)p^m\}| + |\{0, \ldots, (2^n - 2)p^m\} \cap (C + \tau)|,$$

and

$$|C_j \cap (C_{(j-\tau) \pmod{2^n}}| = |E_j \cap (E_{(j-\tau) \mod 2^n} + \tau)| = d\big(j, (j - \tau) \mod 2^n, \tau\big).$$

Since $C_S(\tau) = 4|C \cap (C + \tau)| - N$, then Lemma 8 shows that for $m \geq 1$ the sequence S has poor autocorrelation properties. Applying Lemma 8, we can derive the autocorrelation function for the given set of defining vectors. In particular, from Lemma 8 we can simply obtain Theorem 2 [9].

Consider the autocorrelation function for $\tau \equiv 0 (\mathrm{mod}\ 2^n)$.

Theorem 9: If sequence S is defined by (2) and $\tau \in \phi^{-1}(\{0\} \times p^f H_a)$, $f = 0, 1, \ldots, m-1; a = 0, 1$, then

$$C_S(\tau) = 2^n(p^m - p^{m-f} - p^{m-f-1}) + A,$$

where $A = 0$, if $p \equiv 3 (\mathrm{mod}\ 4)$ and

$$A = |\{t | i_f^{(2t)} = a, t = 0, 1, \ldots, 2^{m-1} - 1\}| - |\{t | i_f^{(2t+1)} = a, t = 0, 1, \ldots, 2^{m-1} - 1\}|,$$

if $p \equiv 1 (\mathrm{mod}\ 4)$.

Proof: Under the conditions of Theorem 9 we have

$$|C \cap (C + \tau)| = \sum_{u=0}^{2^n-1} |E_u \cap (E_u + \tau)| + |C \cap (\{0, \ldots, (2^n - 2)p^m\} + \tau)| + |\{0, \ldots, (2^n - 2)p^m\} \cap (C + \tau)|. \quad (6)$$

By definition, put $A_1 = |\{t | i_f^{(2t+1)} = a, t = 0, 1, \ldots, 2^{m-1} - 1\}|$ and $A_2 = |\{t | i_f^{(2t)} = a, t = 0, 1, \ldots, 2^{m-1} - 1\}|$. From Lemma 8 it follows that

$$\sum_{u=0}^{2^n-1} |E_u \cap (E_u + \tau)| = \sum_{u=0}^{2^n-1} d(u, u, \tau) = 2^{n-2}(2p^m - p^{m-f} - p^{m-f-1}) - B, \quad (7)$$

where $B = \begin{cases} A_1 + A_2, & \text{if } p \equiv 1 \pmod 4; \\ -2^{n-1}, & \text{if } p \equiv 3 \pmod 4. \end{cases}$.

If $\tau \in \phi^{-1}(\{0\} \times p^f H_a)$, then

$$|C \cap (\{0, 2p^m, \ldots, (2^n - 2)p^m\} + \tau)| = \sum_{j=0}^{2^{n-1}-1} |E_{2j} \cap (\{0, 2p^m, \ldots, (2^n - 2)p^m\} + \tau)| = \sum_{j=0}^{2^{n-1}-1} |E_{2j} \cap \{\tau\}|.$$

Similarly, $|\{0, 2p^m, \ldots, (2^n - 2)p^m\} \cap (C + \tau)| = \sum_{j=0}^{2^{n-1}-1} |\{0\} \cap (E_{2j} + \tau)|$.

If $p \equiv 1 \pmod 4$, then $-1 \in H_0$ and $-1 \in H_1$ if $p \equiv 3 \pmod 4$ [9]. Hence, $|E_{2j} \cap \{\tau\}|$ is equal to 1, if $a = i_f^{(2j)}$ and zero if $a \neq i_f^{(2j)}$. Next,

$$|\{0\} \cap (E_{2j} + \tau)| = \begin{cases} 1, & \text{if } a = i_f^{(2j)}, p \equiv 1 \pmod 4 \\ & \text{or } a \neq i_f^{(2j)}, p \equiv 3 \pmod 4; \\ 0, & \text{else.} \end{cases}$$

Therefore,

$$|C \cap (\{0, 2p^m, \ldots, (2^n - 2)p^m\} + \tau)| + |\{0, 2p^m, \ldots, (2^n - 2)p^m\} \cap (C + \tau)| = \begin{cases} 2A_2, & \text{if } p \equiv 1 \pmod 4; \\ 2^{n-1}, & \text{if } p \equiv 3 \pmod 4. \end{cases} \quad (8)$$

Since $C_S(\tau) = 4|C \cap (C + \tau)| - N$, we obtain Theorem 9 by summing up (6), (7) and (8). ∎

Theorem 9 shows that for $m > 1$ the autocorrelation function of S is far from ideal (see [9]).

REFERENCES

[1] T. Cusick, C. Ding, and A. Renvall, *Stream ciphers and number theory*. N.-Holl. Math. Libr. vol.55, 1998.

[2] C. Ding, T. Helleseth. "Generalized cyclotomy and its applications". *Finite Fields Appl.*, vol.4, pp. 467–474, 1999

[3] C. Ding, T. Helleseth, and W. Shan. "On the linear complexity of Legendre sequences". *IEEE Trans. Inform. Theory*, vol. 44, pp. 1276–1278, 1998

[4] V. Edemskiy. "About computation of the linear complexity of generalized cyclotomic sequences with period p^{n+1}". *Des. Codes Cryptogr.*, vol. 61, pp. 251–260, 2011.

[5] V. Edemskiy, O. Antonova. "The linear complexity of generalized cyclotomic sequences with period $2p^n$". *AAECC*, vol. 25, iss. 3, pp. 213–223, 2014.
[6] S. W. Golomb, G. Gong, *Signal Design for Good Correlation: For-Wireless Communications, Cryptography and Radar Applications*, Cambridge, 2005.
[7] K. Ireland and M. Rosen, *A Classical Introduction to Modern Number Theory*. Springer, 1982.
[8] S. Y. Jin, Y. J. Kim, and H. Y. Song. "Autocorrelation of new generalized cyclotomic sequences of period p^n". *IEICE Trans. Fundam.*, vol.E93-A, pp. 2345–2348, 2010.
[9] P. Ke, J. Zhang, and S. Zhang. "On the linear complexity and the autocorrelation of generalized cyclotomic binary sequences of length $2p^m$". *Des. Codes Cryptogr.,* vol.67, no. 3, pp.325–339, 2013
[10] Y. J. Kim, H. Y. Song. "Linear complexity of prime n-square sequences". *In: IEEE International Symposium on Information Theory*, Toronto, Canada, pp. 2405–408, 2008
[11] R. Lidl, H. Neid, *Finite Fields.* Addison-Wesley, 1983.
[12] Y. Sun, H. Shen. "New Binary Sequences of Length $4p$ with Optimal Autocorrelation Magnitude". *Ars Combinatoria (A Canadian Journal of Combinatorics).* vol.89, pp.255–262, 2008
[13] T. Yan, S. Li, and G. Xiao. "On the linear complexity of generalized cyclotomic sequences with the period p^m". *Appl. Math. Lett.,* vol.21, pp. 187–193, 2008
[14] J. W. Zhang, C. A. Zhao, and X. Ma. "Linear complexity of generalized cyclotomic binary sequences of length $2p^m$". *AAECC.,* vol. 21, pp. 93–108, 2010.

ISBN: 978-1-61804-256-9

Harmonic Mappings Related to the Bounded Boundary Rotation

Melike Aydoğan[1], H. Esra Özkan Uçar[2] and Yaşar Polatoğlu[3]

Abstract—The aim of this paper to give investigation of the class of harmonic mapping related to the bounded boundary rotation. The class of bounded boundary rotation is generalized to the convex function. For this aim we use subordination techniques to obtain known as well as new results related to the Libera-type problem [1].

Keywords—Harmonic mapping, bounded rotation, bounded radius rotation, starlike function, convex function.

I. Introduction

Let Ω be the class of functions $\phi(z)$ which are regular in $\mathbb{D}$ and satisfying the conditions $\phi(0)=0$, $|\phi(z)|<1$ for all $z\in\mathbb{D}$.

Next, denote by $\mathcal{P}$ the family of functions $p(z)=1+p_1z+p_2z^2+\cdots$ regular in $\mathbb{D}$, such that $p(z)$ is in $\mathcal{P}$ if and only if

$$p(z)=\frac{1+\phi(z)}{1-\phi(z)} \tag{I.1}$$

for some functions $\phi(z)\in\Omega$, and every $z\in\mathbb{D}$.

$\mathcal{P}_k$ denotes the class of functions $p(0)=1$ and analytic in $\mathbb{D}$ with the representation

$$p(z)=\int_0^{2\pi}\frac{1+ze^{-it}}{1-ze^{-it}}d\mu(t), \tag{I.2}$$

where $\mu(t)$ is defined by

$$\int_0^{2\pi}d\mu(t)=2 \quad \text{and} \quad \int_0^{2\pi}|d\mu(t)|\leq k. \tag{I.3}$$

Clearly $p_2=p$. From the (I.2), one can easily find that $p(z)\in\mathcal{P}_k$ can also be written as

$$p(z)=\left(\frac{k}{4}+\frac{1}{2}\right)p_1(z)-\left(\frac{k}{4}-\frac{1}{2}\right)p_2(z), \tag{I.4}$$

where $p_1(z),p_2(z)\in\mathcal{P}$.

Moreover, let $\mathcal{A}$ be the the class of all analytic functions of the form $s(z)=z+c_2z^2+c_3z^3+\cdots$ which are regular in $\mathbb{D}$. Let $\mathcal{C}$ denote the family of functions $s(z)\in\mathcal{A}$ such that $s(z)$ is in $\mathcal{C}$ if and only if

$$\left(1+z\frac{s''(z)}{s(z)}\right)=p(z) \tag{I.5}$$

for some $p(z)\in\mathcal{P}$, and all $z\in\mathbb{D}$. The class $\mathcal{C}$ is called the class of convex functions, and let $s(z)$ be an element of $\mathcal{A}$ if the equality

$$\left(z\frac{s'(z)}{s(z)}\right)=p(z)$$

is satisfied for some $p(z)\in\mathcal{P}$ and every $z\in\mathbb{D}$, then $s(z)$ is called starlike functions, the class of functions is denoted by $\mathcal{S}^*$. Let $s(z)$ be an element of $\mathcal{A}$ and which maps $\mathbb{D}$ conformally onto an image domain of boundary rotation at most $k\pi$, the class of such functions is denoted by $V(k)$. The concept of functions of bounded boundary rotation original from Loewner [3] in 1917 but he did not use the present terminology. Paatero [2], who systematically developed their properties and made an exhaustive study of the class $V(k)$. Paatero [2] has shown that $s(z)\in V(k)$ if and only if

$$s'(z)=Exp\left[-\int_0^{2\pi}\log(1-ze^{-it})d\mu(t)\right], \tag{I.6}$$

where $\mu(t)$ is given in (I.3). For a fixed $k\geq 2$ it can also be expressed as

$$\int_0^{2\pi}\left|Re\frac{(zs'(z))'}{s(z)}\right|d\theta\leq k\pi, \quad z=re^{i\theta}. \tag{I.7}$$

Clearly, if $k_1<k_2$ then $V(k_1)\subset V(k_2)$, that is the class $V(k)$ obviously expand as k increases. $V(2)$ is simply the class of $\mathcal{C}$ convex univalent functions and Paatero [2] showed that $V(4)\subset\mathcal{S}$, where $\mathcal{S}$ is the class of normalized univalent functions. Later Pinchuk [4] proved that functions in $V(k)$ are close-to-convex in $\mathbb{D}$ if $2\leq k\leq 4$ and hence univalent. A function $s(z)$ analytic in $\mathbb{D}$ is said to be close-to-convex, if there exists a function $\ell(z)\in\mathcal{C}$ such that

$$Re\left(\frac{s'(z)}{\ell(z)}\right)>0 \tag{I.8}$$

for all $z\in\mathbb{D}$. Let $s(z)$ be an element of $\mathcal{A}$ if $s(z)$ having the representation

$$s(z)=zExp\left[-\int_0^{2\pi}\log(1-ze^{-it})d\mu(t)\right], \tag{I.9}$$

where $\mu(t)$ is given in (I.3), then $s(z)$ is called bounded radius rotation. The class of these functions is denoted by $R(k)$. Pinchuk [4] also showed that Alexander type relation between the classes $V(k)$ and $R(k)$ is

$$s(z)\in V(k) \quad \text{if and only if} \quad (zs'(z))\in R(k). \tag{I.10}$$

Pinchuk [4] has shown that the classes $V(k)$ and $R(k)$ can be defined by using class $\mathcal{P}_k$ as given below

$$s(z)\in V(k) \quad \text{if and only if} \quad \frac{(zs'(z))'}{s'(z)}\in\mathcal{P}_k, \tag{I.11}$$

$$s(z)\in R(k) \quad \text{if and only if} \quad z\frac{s'(z)}{s(z)}\in\mathcal{P}_k. \tag{I.12}$$

Let $s_1(z)$ and $s_2(z)$ be elements of $\mathcal{A}$. If there exist a function $\phi(z)\in\Omega$ such that $s_1(z)=s_2(\phi(z))$ for all $z\in\mathbb{D}$, then

ISBN: 978-1-61804-256-9

we say that $s_1(z)$ is subordinate to $s_2(z)$, and we write $s_1(z) \prec s_2(z)$. If $s_2(z)$ is univalent in $\mathbb{D}$, then $s_1(z) \prec s_2(z)$ if and only if $s_1(\mathbb{D}) \subset s_2(\mathbb{D}), s_1(0) = s_2(0)$ implies $s_1(\mathbb{D}_r) \subset s_2(\mathbb{D}_r)$, where $\mathbb{D}_r = \{z| \quad |z| < r, \quad 0 < r < 1\}$ (Subordination principle and Lindelöf principle [5]).

Finally, a planar harmonic mapping f in the open unit disc $\mathbb{D}$ is a complex-valued harmonic function which maps $\mathbb{D}$ onto the same planar domain $f(\mathbb{D})$. Since $\mathbb{D}$ is a simply connected domain, the mapping f has a canonical decomposition $f = h(z) + \overline{g(z)}$, where $h(z)$ and $g(z)$ are analytic in $\mathbb{D}$ and have the following power series

$$h(z) = \sum_{n=0}^{\infty} a_n z^n, \qquad g(z) = \sum_{n=0}^{\infty} b_n z^n, a_n, b_n \in \mathbb{C}$$

We call $h(z)$ is analytic part of f and $g(z)$ is co-analytic part of f. An elegant and complete treatment theory of harmonic mapping is given in Duren's monograph [6]. Lewy [6] proved that the harmonic mapping f is locally univalent in $\mathbb{D}$ if and only if its Jacobian $J_f = \left(|h'(z)|^2 - |g'(z)|^2\right)$ is different form zero in 1936. In view of this result locally univalent harmonic mapping in the open unit disc $\mathbb{D}$ are either sense-preserving if $|h'(z)| > |g'(z)|$ in $\mathbb{D}$ or sense-reversing if $|h'(z)| < |g'(z)|$ in $\mathbb{D}$. In this paper we will restrict ourselves to the study of sense-preserving harmonic mappings. We also note that $f = h(z) + \overline{g(z)}$ is sense-preserving in $\mathbb{D}$ if and only if $h'(z)$ does not 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 $\mathbb{D}$ with $a_0 = b_0 = 0$ and $a_1 = 1$ will be denoted by $\mathcal{S}_H$. Thus $\mathcal{S}_H$ contains standart class $\mathcal{S}$ of univalent functions. The family of all mappings $f \in \mathcal{S}_H$ with the additional property $g'(0) = 0$, i.e, $b_1 = 0$ is denoted by $\mathcal{S}_H^0$. Hence it is clear that $\mathcal{S} \subset \mathcal{S}_H^0 \subset \mathcal{S}_H$.

We will investigate the following class

$$\mathcal{S}_{HV(k)} = \Big\{ f = h(z) + \overline{g(z)} \quad | \quad \frac{\frac{1}{b_1} g'(z)}{h'(z)} \in P_k \quad h(z) \in V(k) \Big\}.$$

For our proofs we need the following lemma and theorems.

Lemma I.1. *([7])Let $\phi(z)$ be a regular in the open unit disc $\mathbb{D}$ with $\phi(0) = 0$, then if $|\phi(z)|$ attains its maximum value on the circle $|z| = r$ at the point z_0 one has $z_0\phi'(z_0) = m\phi(z_0)$, $m \geq 1$.*

Theorem I.2. *([8]) Let $s(z)$ be an element of $V(k)$ then*

$$\frac{(1-r)^{\frac{1}{2}k-1}}{(1+r)^{\frac{1}{2}k+1}} \leq |s'(z)| \leq \frac{(1+r)^{\frac{1}{2}k-1}}{(1-r)^{\frac{1}{2}k+1}}, \qquad \text{(I.13)}$$

$$M_2(a,b,c,r) \leq |s(z)| \leq M_1(a,b,c,r), \qquad \text{(I.14)}$$

where

$$M_1(a,b,c,r) = \frac{2^{b-1}}{a}\left[G(a,b,c,-1) - r_1^a G(a,b,c,r_1^{-1})\right],$$

$$M_2(a,b,c,r) = \frac{2^{b-1}}{a}\left[G(a,b,c,-1) - r_1^a G(a,b,c,-r_1)\right],$$

$$a = \frac{k}{2}, \quad b = 0, \quad c = \frac{k}{2} + 1, \quad r_1 = \frac{1-r}{1+r},$$

$$G(a,b,c,r) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(c-a)} \int_0^r u^{a-1}(1-u)^{c-a-1}(1-zu)^{-b} du.$$

Theorem I.3. *([9]) Let $k \geq 2$ and $s(z) \in V(k)$ such that $\frac{s(z)}{z} \neq 0$ in $\mathbb{D}$. Then*

$$G(z) = \left(\int_0^z \frac{s(t)}{t} dt\right) \in V(k).$$

Theorem I.4. *([10]) Let $s(z) \in V(k)$, $2 \leq k < \infty$. Let $x \in \mathbb{D}$ and*

$$F(z) = \frac{s\left(\frac{z+x}{1+\bar{x}z}\right) - s(x)}{s'(x)(1-|x|^2)}.$$

Then $F(z) \in V(k)$ and

$$\left| z\frac{s''(z)}{s'(z)} - \frac{2r^2}{1-r^2} \right| \leq \frac{kr}{1-r^2}.$$

II. Conclusion

Lemma II.1. *Let $p(z) \in P_k$ then*

$$\left| p(z) - \frac{1+r^2}{1-r^2} \right| \leq \frac{kr}{1-r^2}.$$

Proof: Using Theorem I.4, then we write

$$\left| z\frac{s"(z)}{s'(z)} - \frac{2r^2}{1-r^2} \right| \leq \frac{kr}{1-r^2}$$

after the simple calculations we get

$$\left| \left(1 + z\frac{s"(z)}{s'(z)}\right) - \frac{1+r^2}{1-r^2} \right| \leq \frac{kr}{1-r^2}.$$

In this step by using the definition of the class $V(k)$ we can write

$$\left| p(z) - \frac{1+r^2}{1-r^2} \right| \leq \frac{kr}{1-r^2}.$$

■

Theorem II.2. *Let $f = h(z) + \overline{g(z)}$ be an element of $\mathcal{S}_{HV(k)}$, then $\frac{g(z)}{h(z)} \in V(k)$.*

Proof: Using the subordination principle and definition of the class $\mathcal{S}_{HV(k)}$, then we write $w(\mathbb{D}_r)$ as

$$\begin{cases} \frac{g'(z)}{h'(z)} \left| \left[b_1 \left(\left(\frac{k}{4} + \frac{1}{2}\right) \frac{1+\phi_1(z)}{1-\phi_1(z)} - \left(\frac{k}{4} - \frac{1}{2}\right) \frac{1+\phi_2(z)}{1-\phi_2(z)} \right) \right] - \frac{1+r^2}{1-r^2} \right| \\ \leq \frac{kr}{1-r^2}, \quad k > 2 \\ \\ \frac{g'(z)}{h'(z)} \left| b_1 \frac{1+\phi_1(z)}{1-\phi_1(z)} - \frac{1+r^2}{1-r^2} \right| \leq \frac{2r}{1-r^2}, \quad k = 2 \end{cases} \qquad \text{(II.1)}$$

where $\phi_1, \phi_2 \in \Omega, 0 < r < 1$. Now we define the function $p_k(z)$ by

$$\frac{g(z)}{h(z)} = b_1\left[\left(\frac{k}{4}+\frac{1}{2}\right)\frac{1+\phi_1(z)}{1-\phi_1(z)} - \left(\frac{k}{4}-\frac{1}{2}\right)\frac{1+\phi_2(z)}{1-\phi_2(z)}\right] = b_1 p_k(z),$$

then $p_k(z)$ is analytic and $p_k(0) = 1$, i.e., $\phi_1(0) = 0$, $\phi_2(0) = 0$

$$\frac{g'(z)}{h'(z)} = \begin{cases} b_1\left(\frac{k}{4}+\frac{1}{2}\right)\left[\frac{1+\phi_1(z)}{1-\phi_1(z)} + \frac{2z\phi_1'(z)}{(1-\phi_1(z))^2}\cdot\frac{h(z)}{zh'(z)}\right] \\ -b_1\left(\frac{k}{4}-\frac{1}{2}\right)\left[\frac{1+\phi_2(z)}{1-\phi_2(z)} + \frac{2z\phi_2'(z)}{(1-\phi_2(z))}\cdot\frac{h(z)}{zh'(z)}\right] & , k>2 \\ b_1\frac{1+\phi_1(z)}{1-\phi_1(z)} + \frac{2z\phi_1'(z)}{(1-\phi_1(z))^2}\cdot\frac{h(z)}{zh'(z)} & , k=2. \end{cases} \quad \text{(II.2)}$$

In this step, if we use Theorem I.2 and Lemma II.1, then (II.2) can be written in the following manner

$$w(z_0) = \frac{g'(z_0)}{h'(z_0)} = \begin{cases} b_1\left(\frac{k}{4}+\frac{1}{2}\right)\left[\frac{1+\phi_1(z_0)}{1-\phi_1(z_0)} + \frac{2m\phi_1(z_0)}{(1-\phi_1(z_0))^2}\cdot\frac{1+kre^{i\theta}+r^2}{1-r^2}\right] \\ -b_1\left(\frac{k}{4}-\frac{1}{2}\right)\left[\frac{1+\phi_2(z_0)}{1-\phi_2(z_0)} + \frac{2z_0\phi_2'(z_0)}{(1-\phi_2(z_0))^2}\cdot\frac{1+kre^{i\theta}+r^2}{1-r^2}\right] \notin w(\mathbb{D}_r) & , k>2 \\ (|\phi_1(z)| \text{ attains maximum value on the circle } |z|=r \text{ at the point} z_0) \\ b_1\left(\frac{k}{4}+\frac{1}{2}\right)\left[\frac{1+\phi_1(z_0)}{1-\phi_1(z_0)} + \frac{2z\phi_1(z_0)}{(1-\phi_1(z_0))^2}\cdot\frac{1+kre^{i\theta}+r^2}{1-r^2}\right] \\ -b_1\left(\frac{k}{4}-\frac{1}{2}\right)\left[\frac{1+\phi_2(z_0)}{1-\phi_2(z_0)} + \frac{2m\phi_2(z_0)}{(1-\phi_2(z_0))^2}\cdot\frac{1+kre^{i\theta}}{1-r^2}\right] \notin w(\mathbb{D}_r) & , k>2 \\ (|\phi_2(z)| \text{ attains maximum value on the circle } |z|=r \text{ at the point} z_0) \\ b_1\left(\frac{k}{4}+\frac{1}{2}\right)\left[\frac{1+\phi_1(z_0)}{1-\phi_1(z_0)} + \frac{2m\phi_1(z_0)}{(1-\phi_1(z_0))^2}\cdot\frac{1+kre^{i\theta}+r^2}{1-r^2}\right] \\ -b_1\left(\frac{k}{4}-\frac{1}{2}\right)\left[\frac{1+\phi_2(z_0)}{1-\phi_2(z_0)} + \frac{2m\phi_2(z_0)}{(1-\phi_2(z_0))^2}\cdot\frac{1+kre^{i\theta}+r^2}{1-r^2}\right] \notin w(\mathbb{D}_r) & , k>2 \\ (|\phi_1(z)| \, and \, |\phi_2(z)| \text{ attains maximum value on the circle } |z|=r \text{at the point} z_0) \\ b_1\frac{1+\phi_1(z_0)}{1-\phi_1(z_0)} + \frac{2m\phi_1(z_0)}{(1-\phi_1(z_0))^2}\cdot\frac{1+kre^{i\theta}+r^2}{1-r^2} & , k=2. \end{cases} \quad \text{(II.3)}$$

But this contradicts to (II.2) because $|\phi(z_0)| = 1$, $m \geq 1$. So our assumption $|\phi(z_0)| = 1$ is wrong, i.e. $|\phi(z)| < 1$ for all $z \in D$, therefore we have $\frac{g(z)}{h(z)} \prec b_1 p_k(z)$. ■

Corollary II.3. *If $f = h(z) + \overline{g(z)} \in \mathcal{S}_{HV(k)}$, then*

$$M_2(a,b,c,r)\frac{1-kr+r^2}{1-r^2} \leq |g(z)| \leq M_1(a,b,c,r)\frac{1+kr+r^2}{1-r^2},$$

$$\frac{(1-kr+r^2)(1-r)^{\frac{k}{2}-2}}{(1+r)^{\frac{k}{2}+2}} \leq |g'(z)| \leq \frac{(1+kr+r^2)(1+r)^{\frac{k}{2}-2}}{(1-r)^{\frac{k}{2}+2}}$$

where $M_1(a,b,c,r)$ and $M_2(a,b,c,r)$ are given Theorem I.2.

Proof: Using the definition $\mathcal{S}_{HV(k)}$ we can write

$$|h(z)|\frac{1-kr+r^2}{1-r^2} \leq |g(z)| \leq \frac{1+kr+r^2}{1-r^2}|h(z)|,$$

$$|h'(z)|\frac{1-kr+r^2}{1-r^2} \leq |g'(z)| \leq \frac{1+kr+r^2}{1-r^2}|h'(z)|.$$

In this step if we use Theorem I.2 then we obtain desired result. ■

Corollary II.4. *Let $f = h(z)+\overline{g(z)}$ be an element of $\mathcal{S}_{HV(k)}$, then*

$$\frac{(1-r)^{k-2}}{(1+r)^{k+2}}\left(1-\left(\frac{1+kr+r^2}{1-r^2}\right)^2\right) \leq J_f \leq \frac{(1+r)^{k-2}}{(1-r)^{k+2}}\left(1-\left(\frac{1-kr+r^2}{1-r^2}\right)^2\right).$$

Proof: Since

$$J_f = |h'(z)|^2 - |g'(z)|^2 = |h'(z)|^2\left(1-\left|\frac{g'(z)}{h'(z)}\right|^2\right)$$

and in this step we use Corollary II.3, then we obtain desired result. ■

Corollary II.5. *Let $f = h(z) + \overline{g(z)} \in \mathcal{S}_{HV(k)}$ then*

$$|f| \leq \int_0^1 \frac{(1+r)^{\frac{1}{2}}k-1}{(1-r)^{\frac{1}{2}}k+1}\frac{kr-2r^2}{1-r^2}dr.$$

Proof: Using the inequality

$$(|h'(z)| - |g'(z)|)\,|dz| \leq |df| \leq (|h'(z)| + |g'(z)|)\,|dz|$$

we get

$$|h'(z)|\left(1-\left|\frac{g'(z)}{h'(z)}\right|\right)|dz| \leq |df| \leq |h'(z)|\left(1+\left|\frac{g'(z)}{h'(z)}\right|\right)|dz|$$

after simple calculations we can get the result easily. ■

REFERENCES

[1] P. Duren, Harmonic mappings in the plane, Cambridge University press, Cambridge Tracts in Mathematics 156, 2004.

[2] A.W. Goodman, Univalent functions, Mariner publishing company, Inc. Tampa, Florida, 1983.

[3] Richard R. Hall and St. Ruscheweyh, Indian J. Pure. Appl. Math. 16(11)1985, 1317-1325.

[4] I. S. Jack, Functions starlike and convex of order α, J. London Math. Soc. 2(3)1971, 369-374.

[5] R. J. Libera, Some classes of regular univalent functions, Proc. Amer. Math. Soc. 24(1965), 755-758.

[6] C. Loewner, Untersuchungen ber die Verzerrung bei konformen Abbildungen des Einheitskreises $|z| < 1$, Ber verh Sachs Ges Wiss Leipzig, 69(1917), 89-16.

[7] K. I. Noor, On subclasses of close-to-convex functions of higher order, Internat. J. Math. and Math. Sci. Vol. 15(2)1992, 279-290.

[8] V. Paatero, ber die konforme Abbildungen von Gebieten deren Rnder von beschrnkter Drehung sind, Ann. Acad. Sci. Fenn. Ser. A 33(9)1931.

[9] B. Pinchuk, Functions with bounded boundary rotation, Isr. J. Math. 10(1971), 7-16.

[10] M. S. Robertson, Coefficient of functions with bounded boundary rotation, Canad. J. Math. 21(1969),1477-1482.

Estimating the Flight Path of Moving Objects Based on Acceleration Data

Peter Z. Revesz
Department of Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, Nebraska 68588-0115
Email: revesz@cse.unl.edu
http://cse.unl.edu/revesz
Telephone: (1+) 402 472–3488

***Abstract*—Inertial navigation is the problem of estimating the flight path of a moving object based on only acceleration measurements. This paper describes and compares two approaches for intertial navigation. Both approaches estimate the flight path of the moving object using cubic spline interpolation, but they find the coefficients of the cubic spline pieces by different methods. The first approach uses a tridiagonal matrix, while the second approach uses recurrence equations. They also require different boundary conditions. While both approaches work in O(n) time where n is the number of given acceleration measurements, the recurrence equation-based method can be easier updated when a new measurement data is obtained.**

I. Introduction

Inertial navigation is the problem of estimating the flight path of a moving object based on only acceleration measurements. With the wide-spread availability of GPS sensors, inertial navigation is still important when the GPS system is not accessible, for example, when the moving object is a submarine deep in the ocean or when the GPS system is deliberately disrupted in the course of combat. Understanding inertial navigation is also important for biology because several animal species, including different kinds of birds, seem to use inertial navigation to find they way.

The problem of inertial navigation is more challenging than the simpler problem of estimating the flight path of a moving object based on data on its position at either sporadic or regular periodic time intervals. This simpler problem may be solved using several interpolation methods. For example, the problem can be solved using cubic spline interpolation for functions of one time variable [2]. Cubic splines can be described as follows.

Let $f(t)$ be a function from $\mathcal{R}$ to $\mathcal{R}$. Suppose we know about f only its value at locations $t_0 < \ldots < t_n$. Let $f(t_i) = a_i$. Piecewise cubic spline interpolation of $f(t)$ is the problem of finding the b_i, c_i and d_i coefficients of the cubic polynomials S_i for $0 \leq i \leq n-1$ written in the form:

$$S_i(t) = a_i + b_i(t - t_i) + c_i(t - t_i)^2 + d_i(t - t_i)^3 \quad (1)$$

where each piece S_i interpolates the interval $[t_i, t_{i+1}]$ and fits the adjacent pieces by satisfying certain smoothness conditions. Taking once and twice the derivative of Equation (1) yields, respectively, the equations:

$$S_i'(t) = b_i + 2c_i(t - t_i) + 3d_i(t - t_i)^2 \quad (2)$$

$$S_i''(t) = 2c_i + 6d_i(t - t_i) \quad (3)$$

Equations (1-3) imply that $S_i(t_i) = a_i$, $S_i'(t_i) = b_i$ and $S_i''(t_i) = 2c_i$. For a smooth fit between the adjacent pieces, the cubic spline interpolation requires that the following conditions hold for $0 \leq i \leq n-2$:

$$S_i(t_{i+1}) = S_{i+1}(t_{i+1}) = a_{i+1}, \quad (4)$$

$$S_i'(t_{i+1}) = S_{i+1}'(t_{i+1}) = b_{i+1} \quad (5)$$

$$S_i''(t_{i+1}) = S_{i+1}''(t_{i+1}) = 2c_{i+1} \quad (6)$$

This paper is organized as follows. Section II describes the cubic splines interpolation method using the tridiagonal matrix approach. Section III describes an alternative recurrence equation-based approach. Section IV presents an example of cubic spline interpolation of a moving object and compares the two approaches. Section V describes the generalization of the two approaches to objects that move in 3D space. Finally, Section VI gives some conclusions and describes several open problems and future work.

II. A Tridiagonal Matrix-Based Solution

In this section we present a cubic spline interpolation using a tridiagonal matrix-based approach. Let $h_i = t_{i+1} - t_i$. Substituting Equations (1-3) into Equations (4-6), respectively, yields:

$$a_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = a_{i+1} \quad (7)$$

$$b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1} \quad (8)$$

$$c_i + 3d_i h_i = c_{i+1} \quad (9)$$

ISBN: 978-1-61804-256-9

Equation (9) yields a value for d_i, which we can substitute into Equations (7-8). Hence Equations (7-9) can be rewritten as:

$$a_{i+1} - a_i = b_i h_i + \frac{2c_i + c_{i+1}}{3} h_i^2 \tag{10}$$

$$b_{i+1} - b_i = (c_i + c_{i+1}) h_i \tag{11}$$

$$d_i = \frac{1}{3h_i}(c_{i+1} - c_i). \tag{12}$$

Solving Equation (10) for b_i yields:

$$b_i = (a_{i+1} - a_i)\frac{1}{h_i} - \frac{2c_i + c_{i+1}}{3} h_i \tag{13}$$

which implies for $j \leq n-3$ the condition:

$$b_{i+1} = (a_{i+2} - a_{i+1})\frac{1}{h_{i+1}} - \frac{2c_{i+1} + c_{i+2}}{3} h_{i+1} \tag{14}$$

Substituting into Equation (11) the values for b_i and b_{i+1} from Equations (13-14) yields:

$$(a_{i+1} - a_i)\frac{1}{h_i} - (2c_i + c_{i+1})\frac{h_i}{3} + (c_i + c_{i+1})h_i = (a_{i+2} - a_{i+1})\frac{1}{h_{i+1}} - (2c_{i+1} + c_{i+2})\frac{h_{i+1}}{3}$$

The above can be rewritten as:

$$\frac{3}{h_i}a_i - \left(\frac{3}{h_i} + \frac{3}{h_{i+1}}\right)a_{i+1} + \frac{3}{h_{i+1}}a_{i+2} = h_i c_i + 2(h_i + h_{i+1})c_{i+1} + h_{i+1}c_{i+2}$$

The above holds for $0 \leq i \leq n-3$. However, changing the index downward by one the following holds for $1 \leq j \leq n-2$:

$$\frac{3}{h_{i-1}}a_{i-1} - \left(\frac{3}{h_{i-1}} + \frac{3}{h_i}\right)a_i + \frac{3}{h_i}a_{i+1} \\ h_{i-1}c_{i-1} + 2(h_{i-1} + h_i)c_i + h_i c_{i+1} = \tag{15}$$

The above is a system of $n-1$ linear equations for the unknown position values a_i for $1 \leq i \leq n$ in terms of the measured acceleration values $2c_i$ for $0 \leq i \leq n$. By Equation (3) $S_0''(t_0) = 2c_0$ and by extending Equation (6) to $i = n-1$, $S_{n-1}''(t_n) = 2c_n$.

The cubic spline interpolation allows us to specify several possible boundary conditions regarding the values of a_0 and a_n. A commonly used boundary condition, called a natural cubic spline, assumes that $a_0 = a_n = 0$, which is equivalent to saying that the moving object starts at position 0 and returns to it at the end of its flight. This is a natural condition because birds can be expected to return to their nests and airplanes can be expected to return to their hangars. Hence this is used as a common default condition when there is no better boundary value available. However, we can assume any boundary value for $f(t_0) = a_0$ and $f(t_n) = a_n$ if they are known.

In solving a cubic spline, a uniform sampling is also commonly assumed to be available. This is natural to assume because accelerometers can send a signal every few seconds. In that case each h_i has the same constant value h. Then multiplying Equation (15) by $h/3$ yields:

$$a_{i-1} - 2a_i + a_{i+1} = \frac{h^2}{3}(c_{i-1} + 4c_i + c_{i+1}) \tag{16}$$

Since the values of c_i are known, the values of a_i can be found by solving a particular tridiagonal matrix-vector equation $Ax = B$. The matrices can be represented as follows:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{bmatrix}$$

the vector of unknowns is:

$$x = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix}$$

and the vector of known constants is:

$$B = \begin{bmatrix} f(t_0) \\ \frac{h^2}{3}(c_0 + 4c_1 + c_2) \\ \vdots \\ \frac{h^2}{3}(c_{n-2} + 4c_{n-1} + c_n) \\ f(t_n) \end{bmatrix}.$$

The above describes a system of $n+1$ linear equations with $n+1$ unknowns. Such a system normally yields a unique solution except in some special cases. Moreover, such a tridiagonal matrix system can be solved in $O(n)$ time. Once the a_i values are found, the d_i and the b_i values also can be found by Equations (12) and (13), respectively. Computating the b_i and d_i coefficients can be done also within $O(n)$ time.

The above solution to the inertial navigation problem seems new, although the reverse problem of finding the acceleration values given the position values is a straightforward cubic spline problem. The novelty of the above approach is in Equation (16), which highlights that three consecutive a variables

could be considered the unknowns and can be expressed by three consecutive c constants.

III. Alternative Recurrence Equation Solution

Instead of using a tridiagonal matrix, in this section we give a more direct and effective method for solving the problem of interpolating the location of a moving object described by a function $f(t)$ when we know only the acceleration of the object at times $t_0 < \ldots < t_n$. The measured acceleration value at any time t_i is twice the value of the corresponding constant c_i, that is, $f''(t_i) = 2c_i$. Hence in this case we need to find a piecewise cubic spline interpolation of $f(t)$ by finding the a_i, b_i and d_i coefficients of the cubic polynomials S_i for $0 \leq i \leq n-1$ written in the form of Equation (1). At first note that Equation (11) implies:

$$b_i = b_{i-1} + (c_{i-1} + c_i)h_{i-1} \tag{17}$$

The above can be used to express any b_i for $i > 0$ in terms of the initial velocity b_0 and the c_i coefficients, the known constants, as follows:

$$b_i = b_0 + \sum_{1 \leq k \leq i} (b_k - b_{k-1}) = b_0 + \sum_{1 \leq k \leq i} (c_{k-1} + c_k)h_{k-1}$$

Further, we can rewrite Equation (10) as:

$$a_i = a_{i-1} + b_{i-1}h_{i-1} + \frac{2c_{i-1} + c_i}{3}h_{i-1}^2 \tag{18}$$

The above can be used to express each a_i for $i > 0$ in terms of the b_i and c_i constants as follows:

$$a_i = a_0 + \sum_{1 \leq j \leq i} (a_j - a_{j-1}) =$$

$$a_0 + \sum_{1 \leq j \leq i} \left(b_{j-1}h_{j-1} + \frac{2c_{j-1} + c_j}{3}h_{j-1}^2 \right) \tag{19}$$

Clearly, we can find first all the b_i in $O(n)$ time, and then we can compute all the a_i also in $O(n)$ time. All the d_i can be also computed in $O(n)$ time using Equation (12). Hence in this case also the piecewise cubic interpolation can be found in $O(n)$ time.

IV. Example of An Object in Free Fall

Suppose that an object is released from a height of 400 feet and falls to the ground in five seconds. Suppose also that we measure the object's acceleration at every second until five seconds after release to be always $-32ft/sec^2$ due to the gravitational pull of the earth. Find a cubic spline approximation for the object's position at all times from the release to five seconds after.

As the object falls to the earth, its elevation is decreasing. Hence the gravitational force is considered with a negative sign. The cubic polynomials we need to find for the intervals $[0,1], [1,2], [2,3], [3,4]$ and $[4,5]$ can be expressed as follows:

$$\begin{cases} S_0(t) = a_0 + b_0 t + c_0 t^2 + d_0 t^3 \\ S_1(t) = a_1 + b_1(t-1) + c_1(t-1)^2 + d_1(t-1)^3 \\ S_2(t) = a_2 + b_2(t-2) + c_2(t-2)^2 + d_2(t-2)^3 \\ S_3(t) = a_3 + b_3(t-3) + c_3(t-3)^2 + d_3(t-3)^3 \\ S_4(t) = a_4 + b_4(t-4) + c_4(t-4)^2 + d_3(t-4)^3 \end{cases}$$

We have $n = 6$, $c_0 = -16$, $c_1 = -16$, $c_2 = -16$, $c_3 = -16$ $c_4 = -16$ $c_5 = -16$ and the uniform time step size is $h = 1$ second. By our assumption $f(0) = 400$ and $f(4) = 0$. Hence matrix A and the vectors x and B are:

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$x = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix}$$

and

$$B = \begin{bmatrix} 400 \\ \frac{1}{3}\Big(-16 + 4(-16) - 16 \Big) = -32 \\ \frac{1}{3}\Big(-16 + 4(-16) - 16 \Big) = -32 \\ \frac{1}{3}\Big(-16 + 4(-16) - 16 \Big) = -32 \\ \frac{1}{3}\Big(-16 + 4(-16) - 16 \Big) = -32 \\ 0 \end{bmatrix}$$

We can solve the above tridiagonal linear system to be:

$$a_0 = 400$$
$$a_1 = 384$$
$$a_2 = 336$$
$$a_3 = 256$$
$$a_4 = 144$$
$$a_5 = 0$$

Solving for the b_i coefficients by Equation (13) gives:

$b_0 = \frac{1}{1}(384 - 400) - \frac{1}{3}(-16 - 32) = 0$

$b_1 = \frac{1}{1}(336 - 384) - \frac{1}{3}(-16 - 32) = -32$

$b_2 = \frac{1}{1}(256 - 336) - \frac{1}{3}(-16 - 32) = -64$

$b_3 = \frac{1}{1}(144 - 256) - \frac{1}{3}(-16 - 32) = -96$

$b_4 = \frac{1}{1}(0 - 144) - \frac{1}{3}(-16 - 32) = - - 128$

ISBN: 978-1-61804-256-9

Solving for the d_i coefficients by Equation (12) gives:

$$d_0 = \frac{1}{3}(-16 - (-16)) = 0$$

$$d_1 = \frac{1}{3}(-16 - (-16)) = 0$$

$$d_2 = \frac{1}{3}(-16 - (-16)) = 0$$

$$d_3 = \frac{1}{3}(-16 - (-16)) = 0$$

$$d_4 = \frac{1}{3}(-16 - (-16)) = 0$$

The above values show that an object in free fall travels a quadratically increasing distance. Using the calculated values, we can now describe the cubic spine interpolation as follows:

$$\begin{cases} S_0(x) = 400 - 16t^2 \\ S_1(x) = 384 - 32(t-1) - 16(t-1)^2 \\ S_2(x) = 336 - 64(t-2) - 16(t-2)^2 \\ S_3(x) = 256 - 96(t-3) - 16(t-3)^2 \\ S_4(x) = 144 - 128(t-4) - 16(t-4)^2 \end{cases}$$

It can be calculated that in each piece the cubic spline interpolation can be simplified to $400 - 16t^2$, which agrees with the physics equation for the position of a free falling object that starts with zero velocity from an elevation of 400 feet above the surface of the earth.

Let us next consider the calculation of the same problem using the alternative method. Since the initial velocity is $b_0 = 0$, we can calculate by Equation (17) that:

$$\begin{aligned} b_1 &= 0 + (-16 + (-16)) = -32 \\ b_2 &= -32 + (-16 + (-16)) = -64 \\ b_3 &= -64 + (-16 + (-16)) = -96 \\ b_4 &= -96 + (-16 + (-16)) = -128 \end{aligned}$$

Similarly to the previous approach, Equation (12) can be used to calculate the d_i constants. Hence we get the same solution as with the previous method.

In comparing the two approaches, we see that they require different boundary conditions. For the first method, the tridiagonal system required only the initial and the final position of the moving object. The second method required the initial position and the initial velocity. While both methods work in $O(n)$ time where n is the number of past acceleration measurements, the recurrence equation-based method can be updated easier when a new measurement data is obtained. Hence it may be more practical in time-critical applications.

V. Objects Moving in 3D Space

A moving object, such as an airplane, can fly in 3-dimensional space along latitude, longitude as well as elevation. To model the flight of the airplane, it is possible to describe its movement by a parametric solution consisting of separate functions $f_x(t)$, $f_y(t)$ and $f_z(t)$ for the movement along the x, the y and the z-axis, respectively. Accelerometers signal separately the movement along these three dimensions. Hence it is possible to find a separate cubic spline interpolation for the the functions $f_x(t)$, $f_y(t)$ and $f_z(t)$. Moreover, it is possible to use different kinds of boundary conditions for each of the separate interpolations. For example, to interpolate the elevation function $f_z(t)$, one may use the initial conditions $f_z(t_0) = f_z(t_n) = 0$ when an object is expected to start and finish its movement on the ground, while for $f_x(t)$ an initial position different from zero and some initial velocity may be used.

VI. Conclusion

Inertial navigation relies heavily on the accuracy of accelerometers that need to signal at periodic time intervals the acceleration values in all three dimensions. Another problem is speed. Even an $O(n)$ method is too slow when the object is traveling at very high speeds. In that case, we need a solution that can be easily updated with each new accelerometer measurement. The balancing of computational efficiency with computational accuracy is a challenging problem. We are currently developing methods that describe a trade-off in these two variables. A related problem is to find the flight path of moving objects given their speeds at regular time intervals instead of their accelerations. We also developed some approaches to that problem.

We also implemented the cubic spline interpolation method in the MLPQ constraint database system [6]. The advantage of the implementation is that the moving object representation can be queried using constraint query languages [4], which are extensions of SQL and Datalog. This approach was used successfully in dealing with other interpolation data, such as real estate prices [5] and other moving objects [1], [3]. The MLPQ system also provides a convenient user-friendly graphical user interface that enables animation and other visualizations of moving objects.

References

[1] S. Anderson and P. Z. Revesz, Efficient MaxCount and threshold operators of moving objects, *Geoinformatica*, 13 (4), 355–396, 2009.

[2] R. L. Burden and J. D. Faires, *Numerical Analysis*, 9th ed. New York, USA: Springer, 2014.

[3] J. Chomicki and P. Z. Revesz, Constraint-based interoperability of spatiotemporal databases, *Geoinformatica*, 3 (3), 211-243, 1999.

[4] P. C. Kanellakis, G. M. Kuper and P. Z. Revesz, Constraint query languages, *Journal of Computer and System Sciences*, 51 (1), 26-52, 1995.

[5] L. Li and P. Z. Revesz, Interpolation methods for spatio-temporal geographic data, *Computers, Environment and Urban Systems*, 28 (3), 201-227, 2004.

[6] P. Z. Revesz, *Introduction to Databases: From Biological to Spatio-Temporal*, New York, USA: Springer, 2010.

Notes about the linear complexity of sequences over the finite field of order four

Vladimir Edemskiy, Andrey Ivanov

***Abstract*—We derive the linear complexity of sequences over the finite field of four elements. We consider sequences constructed from Legendre sequences, Hall sequences and twin-prime sequences using the technique proposed by Ting, Ding, Lim, Kim et al.**

***Index Terms*—Balanced quaternary sequences, finite field, linear complexity**

I. Introduction

THE linear complexity and the autocorrelation are important parameters of pseudo-random sequences significant for practical applications [1], [5]. Ting, Ding, Lim, Kim et al. constructed new balanced quaternary sequences with optimal autocorrelation values [8], [12], [13] using the interval structure and inverse Gray map. In the same way, we can build the sequences over the finite field of four elements.

In this paper we investigate the linear complexity of above mentioned sequences over the finite field of four elements. For the application of sequences over the finite field, see [10], for instance.

II. The linear complexity of Tang and Ding sequences

Let $a = a_0, \dots, a_{N-1}$ and $b = b_0, \dots, b_{N-1}$ be binary sequences of period N, $N \equiv 3(\mathrm{mod}4)$. Define sequences c and d as

$$c_i = \begin{cases} a_{i/2}, & \text{if } i \equiv 0(\mathrm{mod}2), \\ a_{(i+N)/2}, & \text{if } i \equiv 1(\mathrm{mod}2). \end{cases}$$

$$d_i = \begin{cases} b_{i/2}, & \text{if } i \equiv 0(\mathrm{mod}2), \\ b_{(i+N)/2} + 1, & \text{if } i \equiv 1(\mathrm{mod}2), \end{cases} \quad (1)$$

i.e. $c = I(a, L^{1/2}a)$ and $d = I(b, L^{1/2}b + 1)$, where I and L denote the interleaved operator and the left cyclic shift operator respectively [13].

The well-known Gray mapping $\phi : \mathbb{Z}_4 \to \mathbb{Z}_2 \times \mathbb{Z}_2$ is defined as

$$\phi(0) = (0,0),\ \phi(1) = (0,1),\ \phi(2) = (1,1),\ \phi(3) = (1,0).$$

In their paper [13], Tang and Ding proved that a sequence $u : u_i = \phi^{-1}(c_i, d_i)$ is balanced quaternary sequence with optimal autocorrelation values if a, b are binary sequences with optimal autocorrelation value.

V. Edemskiy is with the Department of Applied Mathematics and Information Science, Novgorod State University, Veliky Novgorod, Russia, 173003 e-mail: Vladimir.Edemsky@novsu.ru.

A. Ivanov is with Novgorod State University, e-mail: dk@live.ru.

Let $\mathbb{F}_4 = \{0, 1, \mu, \mu + 1\}$ be a finite field of four elements. If we view $\mathbb{F}_4$ as a vector space over $\mathbb{F}_2$ with basis $\mu,\ 1$, then we can define a sequence v by inverse Gray map as

$$v_i = \begin{cases} 0, & \text{if } (c_i, d_i) = (0,0), \\ 1, & \text{if } (c_i, d_i) = (0,1), \\ \mu + 1, & \text{if } (c_i, d_i) = (1,1), \\ \mu, & \text{if } (c_i, d_i) = (1,0). \end{cases} \quad (2)$$

In this section we investigate the linear complexity of sequences constructed by (2) when a, b are Legendre sequences, Hall sequences and twin-prime sequences.

A. Subsidiary lemmas

The minimal polynomial $m(x)$ and the linear complexity LC of v are given by the following equations [1]:

$$m(x) = (x^{2N} - 1)/\gcd\big(x^{2N} - 1, s_v(x)\big),$$
$$LC = 2N - \deg\gcd\big(x^{2N} - 1, s_v(x)\big), \quad (3)$$

where $s_v(x)$ is the generating polynomial of v. Thus, $s_v(x) = \sum_{i=0}^{2N-1} v_i x^i$.

Lemma 1: Let v be defined by (2). Then

$$s_v(x) = \mu s_c(x) + s_d(x),$$

where $s_c(x) = \sum_{i=0}^{2N-1} c_i x^i$ and $s_d(x) = \sum_{i=0}^{2N-1} d_i x^i$.

We see that the statement of Lemma 1 follows from (2). The next statements were proved earlier in [11], [14].

Lemma 2: [14] (i) If $c = I(a, L^{1/2}a)$ then $s_c(x) = (1 + x^N)s_a(x^2)$;

(ii) If $d = I(b, L^{1/2}b + 1)$ then $s_d(x) = (1 + x^N)s_b(x^2) + x(x^{2N} - 1)/(x^2 - 1)$.

(iii) If $b = L^m a$ then $s_b(x^2) = x^{2N-2m}s_a(x^2)$.

Thus, by Lemmas 1 and 2 we have

$$\gcd(x^{2N}-1, s_v(x)) = \frac{x^N - 1}{x - 1}\gcd\Big(\frac{x^N - 1}{x - 1}, \mu s_a(x^2) + s_b(x^2)\Big). \quad (4)$$

So, by (3) and (4), the greatest possible value of the linear complexity v defined by (2) is equal to $N + 1$.

Let $w(x^2) = \mu s_a(x) + s_b(x)$, and let α be a primitive N-th root of unity in the extension of the field $\mathbb{F}_4$ ($\alpha = \mu$ for $N = 3$). Then, by (3) and (4), to compute the minimal polynomial and the linear complexity of v it is sufficient to know the roots of polynomial $w(x)$ in the set $\{\alpha^l, l = 0, 1, \dots, N - 1\}$.

ISBN: 978-1-61804-256-9

B. The linear complexity of sequences obtained from Legendre sequences

Let $\mathbf{QR}_p$ and $\mathbf{NQR}_p$ denote all the nonzero squares and non-squares in $\mathbb{Z}_p$, respectively. Here p is a prime. The Legendre sequence l or $l^{'}$ is defined as

$$l_i = \begin{cases} 1, & \text{if } i \in \mathbf{QR}_p, \\ 0, & \text{if } i \in \{0\} \cup \mathbf{NQR}_p. \end{cases}$$

$$\text{or } l_i^{'} = \begin{cases} 1, & \text{if } i \in \{0\} \cup \mathbf{QR}_p, \\ 0, & \text{if } i \in \mathbf{NQR}_p. \end{cases}$$

It is well known that Legendre sequences have optimal autocorrelation value if $p \equiv 3(\mathrm{mod} 4)$.

The linear complexity of Legendre sequences was studied in [2]. In particular, it was shown that with an appropriate choice of α we can assume that

$$s_l(\alpha^j) = \begin{cases} 1, & \text{if } j \in \mathbf{QR}_p, \\ 0, & \text{if } j \in \mathbf{NQR}_p \end{cases} \tag{5}$$

for $p \equiv 7(\mathrm{mod}\ 8)$, and

$$s_l(\alpha^j) = \begin{cases} \mu, & \text{if } j \in \mathbf{QR}_p, \\ \mu + 1, & \text{if } j \in \mathbf{NQR}_p \end{cases} \tag{6}$$

for $p \equiv 3(\mathrm{mod}\ 8)$.

Let $t(x) = \prod_{j \in \mathbf{QR}_p}(x - \alpha^j)$. Our first contribution in this paper is the following.

Theorem 3: Let $c = I(l, L^{1/2}l)$, $d = I(L^m l, L^{m+1/2}l + 1), m = 0, \ldots, p-1$, and let v be defined by (2). Then:

(i) $LC = (p+3)/2$ and $m(x) = (x-1)^2 t(x)$ if $p \equiv 7\ (\mathrm{mod}\ 8)$.

(ii) $LC = p + 1$ and $m(x) = (x^p - 1)(x - 1)$ if $p \equiv 3(\mathrm{mod}\ 8)$ and $m = 0$ for $p = 3$.

(iii) $LC = 3$ and $m(x) = (x-1)^2(x - (\mu+1)^m)$ if $p = 3, m = 1, 2$.

Proof: If $b = L^m l$ then $s_b(x^2) = x^{2p-2m} s_l(x^2)$ by Lemma 2. Hence, in this case $w(x^2) = \mu s_a(x) + s_b(x) = \mu s_l(x^2)\big(1 + \mu^{-1}x^{2p-2m}\big)$ and $1 + \mu^{-1}\alpha^{-2mj} \neq 0, j = 1, \ldots, p-1$ for $p \neq 3$.

We consider three cases.

(i) Let $p \equiv 7(\mathrm{mod}\ 8)$. Then $2 \in \mathbf{QR}_p$ [7] and by (5) we have that $v(\alpha^{2j}) = 0, j = 1, \ldots, p-1$ iff $j \in \mathbf{NQR}_p$. So,

$$\gcd(x^{2N} - 1, s_v(x)) = \frac{x^N - 1}{x - 1} \prod_{j \in \mathbf{NQR}_p} (x - \alpha^j).$$

By (3), $m(x) = (x-1)^2 t(x)$. Hence, $LC = (p+3)/2$. This completes the proof of the first case.

(ii) Let $p \equiv 3(\mathrm{mod}\ 8)$ and $p \neq 3$. Then $v(\alpha^{2j}) \neq 0, j = 1, \ldots, p-1$ by (6). We see that in this case the statement of Theorem 3 follows from (3)-(4).

(iii) We can make sure that Theorem 3 holds for $p = 3$ by computing the value $1 + \mu^{-4mj-1}, m = 0, 1, 2; j = 1, 2$. ∎

Remark 4: For cryptographic applications one needs sequences with high linear complexity, i.e $LC > N/2$. In the case of Tang and Ding sequences the last inequality means that $LC = p + 1$. Then always $m(x) = (x^p - 1)(x - 1)$ by (3)-(4). Later we will omit the expression for $m(x)$.

Theorem 5: Let $c = I(l, L^{1/2}l)$, $d = I(L^m l^{'}, L^{m+1/2}l^{'} + 1), m = 0, \ldots, p-1$, and let v be defined by (2). Then:

(i) $LC = (p+3)/2$ if $p \equiv 3(\mathrm{mod}\ 8)$ and $m = 0$ or $p = 3, m = 2$.

(ii) $LC = p + 1$ if $p \equiv 7(\mathrm{mod}\ 8)$ or $p \equiv 3(\mathrm{mod}\ 8)$ and $m \neq 0$ for $p \neq 3$ or $m = 1$ for $p = 3$.

We prove Theorem 5 similarly as Theorem 3.

The results of computing the linear complexity by Berlekamp-Massey algorithm when $p = 3, 7, 11, 19, 23, \ldots$ confirm Theorems 3 and 5.

C. The linear complexity of sequences obtained from Legendre and Hall sequences or Hall sequences

Denote by $H_k, k = 0, \ldots, 5$ cyclotomic classes of order 6 modulo p, i.e. $H_k = \{g^{k+6t} \bmod p, t = 0, \ldots, R-1\}$. Let $p = A^2 + 27 = 6R + 1$ be a prime, $A \equiv 1(\mathrm{mod}\ 3)$ and let g be a primitive root modulo p. Let g be (and, always can be) selected such that $3 \in H_1$ [6].

Let $D = H_0 \cup H_1 \cup H_3$ be a Hall difference set [6], and let h be a Hall sequence, i.e.

$$h_i = \begin{cases} 1, & \text{if } j \bmod p \in D, \\ 0, & \text{else.} \end{cases}$$

Put, by definition $D_k = g^k D, k = 0, \ldots, 5$. Denote by $h^{(k)}$ the characteristic sequence D_k, i.e. D_k is the support of the sequence $h^{(k)}$. Then $h^{(k)}$ has optimal autocorrelation value $\{-1\}$.

The polynomial $s_h(x)$ was studied in [9] and in [4]. It is easy to verify that $s_h(\alpha^j) = s_h(\alpha^n)$ if j and n belong to the same cyclotomic class and $s_{h^{(k)}}(\alpha^j) = s_h(\alpha^{jg^k})$.

Lemma 6: [4], [9] Let h be a Hall sequence. Then there exist the primitive p-th root α of unity such that:

(i)

$$s_h(\alpha^j) = \begin{cases} 1, & \text{if } j \in H_0, \\ 0, & \text{if } j \in H_1 \cup \cdots \cup H_5. \end{cases} \tag{7}$$

for $p \equiv 7(\mathrm{mod}\ 8)$;

(ii)

$$s_h(\alpha^j) = \begin{cases} 1, & \text{if } j \in H_0 \cup H_1 \cup H_3 \cup H_4, \\ \mu, & \text{if } j \in H_2, \\ \mu + 1, & \text{if } j \in H_5, \end{cases} \tag{8}$$

for $p \equiv 3(\mathrm{mod}\ 8)$.

Proof: The first statement is proved in [9].

For $p \equiv 3(\mathrm{mod}\ 8)$ the values $\sum_{f \in H_k} \alpha^f, k = 0, 1, \ldots, 5$ were computed in [4]. Using this, we obtain the statement of Lemma 6. ∎

The linear complexity of sequences over $\mathbb{F}_4$ obtained from Legendre and Hall sequences or Hall sequences we investigate below.

Theorem 7: Let $c = I(l, L^{1/2}l)$, $d = I(L^m h^{(k)}, L^{m+1/2}h^{(k)} + 1), m = 0, \ldots, p-1$, and let v be defined by (2). Then:

(i) $LC = p+1$ if $p \equiv 3 (\mathrm{mod}\ 8)$ and $m \neq 0$.

(ii) $LC = (p+3)/2$ if $m = 0$, $p \equiv 3 (\mathrm{mod}\ 8)$ and $k = 1, 3, 5$ or $p \equiv 7 (\mathrm{mod}\ 8)$ and $k = 0, 2, 4$.

(iii) $LC = 2(p+2)/3$ if $m = 0$, $p \equiv 3 (\mathrm{mod}\ 8)$ and $k = 0, 2, 4$ or $p \equiv 7 (\mathrm{mod}\ 8)$ and $k = 1, 3, 5$.

Proof: Under the conditions of Theorem 7 we have

$$w(\alpha^{2j}) = \mu s_l(\alpha^{2j}) + \alpha^{-4mj} s_h(\alpha^{2jg^k}). \quad (9)$$

Thus, if $m \neq 0$ and $p \equiv 3 (\mathrm{mod}\ 8)$ then $w(\alpha^{2j}) \neq 0, j = 1, \ldots, p-1$ by (6) and (8). Hence, from (4) and (3) we obtain that $LC = p+1$.

Let $m = 0$ and $p \equiv 3 (\mathrm{mod}\ 8)$. The values $s_l(\alpha^j)$ and $s_h(\alpha^j)$ in this case are given by (5) and (7). After summing over (9), we have

$$|\{j : w(\alpha^{2j}) = 0\}| = \begin{cases} (p-1)/2, & \text{if } k = 1, 3, 5, \\ (p-1)/3, & \text{if } k = 0, 2, 4 \end{cases}$$
$$\text{for } m = 0 \text{ and } p \equiv 3 (\mathrm{mod}\ 8).$$

Similarly, we have

$$|\{j : w(\alpha^{2j}) = 0\}| = \begin{cases} (p-1)/2, & \text{if } k = 0, 2, 4, \\ (p-1)/3, & \text{if } k = 1, 3, 5 \end{cases}$$
$$\text{for } m = 0, \ldots, N-1 \text{ and } p \equiv 7 (\mathrm{mod}\ 8).$$

We see that the statement of Theorem 7 follows from (3) - (4). ■

Theorem 8: Let $c = I(h, L^{1/2}h)$, $d = I(L^m h^{(k)}, L^{m+1/2} h^{(k)} + 1), m = 0, \ldots, p-1$, and let v be defined by (2). Then:

1. $LC = p+1$ if $p \equiv 3 (\mathrm{mod}\ 8)$ and $m \neq 0$ or $p \equiv 3 (\mathrm{mod}\ 8)$ and $m = k = 0$.
2. $LC = 2(p+2)/3$ if $m = 0$, $p \equiv 3 (\mathrm{mod}\ 8)$ and $k = 1, 2, 4, 5$.
3. $LC = (5p+7)/6$ if $m = 0$, $p \equiv 3 (\mathrm{mod}\ 8)$ and $k = 3$.
4. $LC = (p+5)/3$ if $p \equiv 7 (\mathrm{mod}\ 8)$ and $k = 1, \ldots, 5$.
5. $LC = (p+11)/6$ if $p \equiv 3 (\mathrm{mod}\ 8)$ and $k = 0$.

Theorem 8 we prove similarly as Theorem 7. The results of computing the linear complexity by Berlekamp-Massey algorithm when $p = 31, 43, 127, 283, \ldots$ confirm Theorem 7 and 8.

D. The linear complexity of sequences obtained from twin-prime sequences

Let a be a twin-prime sequence with period $N = p(p+2)$, both p and $p+2$ are primes, and let $b = L^m a$. In this case we have $w(x^2) = \mu s_a(x^2) + x^{2N-2m} s_a(x^2)$ by Lemma 2 and $w(\alpha^{2j}) = \mu s_a(\alpha^{2j})\big(1 + \alpha^{-2mj}\mu^{-1}\big)$, at the same time $1 + \alpha^{-2mj}\mu^{-1} \neq 0$ for $j = 1, \ldots, N-1$ and $p \neq 3$. Thus, by (4) for $p \neq 3$ we have

$$\gcd(x^{2N} - 1, s_v(x)) = \frac{x^N - 1}{x - 1} \gcd\Big(\frac{x^N - 1}{x - 1}, s_a(x^2)\Big). \quad (10)$$

The linear complexity of twin-prime sequences and the values $s_a(\alpha^j)$ were computed in [3]. In particular, from [3] and (10) we obtain the next statement.

Lemma 9: Let v be defined by (2), where a is a twin-prime sequence and $c = I(a, L^{1/2}a)$, $d = I(L^m a, L^{1/2+m} a + 1)$. Then $LC = p(p+2)+1$ iff $p \equiv 1 (\mathrm{mod}\ 8)$ or $p \equiv -3 \ (\mathrm{mod}\ 8)$.

For example, the conditions of Lemma 9 are satisfied for $p = 17, 29$.

III. The linear complexity of Lim et al. sequences

In their paper [12], Lim et al. proved that if a, b are binary sequences with optimal autocorrelation value and

$$e_i = \begin{cases} a_i, & \text{if } i \equiv 0 (\mathrm{mod} 2), \\ a_i, & \text{if } i \equiv 1 (\mathrm{mod} 2). \end{cases}$$
$$f_i = \begin{cases} b_i, & \text{if } i \equiv 0 (\mathrm{mod} 2), \\ b_i + 1, & \text{if } i \equiv 1 (\mathrm{mod} 2), \end{cases} \quad (11)$$

then a sequence $u : u_i = \phi^{-1}(e_i, f_i)$ is a balanced quaternary sequence with period $2N$ and optimal autocorrelation values.

Let z be a sequence defined as

$$z_i = \begin{cases} 0, & \text{if } (e_i, f_i) = (0, 0), \\ 1, & \text{if } (e_i, f_i) = (0, 1), \\ \mu + 1, & \text{if } (e_i, f_i) = (1, 1), \\ \mu, & \text{if } (e_i, f_i) = (1, 0). \end{cases} \quad (12)$$

Lemma 10: Let e, f be defined by (11). Then:

(i) $s_e(x) = (1 + x^N) s_a(x)$;

(ii) $s_f(x) = (1 + x^N) s_b(x) + x \frac{x^{2^N} - 1}{x^2 - 1}$.

Proof: From our definition it follows that $s_e(x) = \sum_{u=0}^{N-1} a_{2u} x^{2u} + \sum_{u=0}^{N-1} a_{2u+1} x^{2u+1}$ or $s_e(x) = \sum_{i=0}^{2N-1} a_i x^i$. Since N is a period of a, we obtain $s_e(x) = (1 + x^N) s_a(x)$. The second statement of Lemma 10 we prove similarly. ■

Lemma 11: Let e, f be defined by (11), and let z be defined by (12). Then

$$\gcd(x^{2N} - 1, s_z(x)) = \frac{x^N - 1}{x - 1} \gcd\Big(\frac{x^N - 1}{x - 1}, \mu s_a(x^2) + s_b(x^2)\Big). \quad (13)$$

Proof: By Lemmas 1 and 10 we have $s_z(x) = \mu s_e(x) + s_f(x)$ or $s_z(x) = (1 + x^N)(\mu s_a(x) + s_b(x)) + x \frac{x^{2^N} - 1}{x^2 - 1}$. The statement of Lemma 11 follows from the latest equality. ■

Let sequences v and z be defined by (2) and (12), respectively, for the same pair of binary sequences a, b. By (4) and Lemma 11

$$\gcd(x^{2N} - 1, s_v(x)) = \gcd(x^{2N} - 1, s_z(x)).$$

So, the linear complexities of v and z are equal. Thus, if a, b are Legendre sequences, Hall sequences or twin-prime, then the linear complexity of the sequence z constructed by (12) is defined by Theorems 3-8.

IV. The linear complexity of Kim et al. sequences

Let l', l be Legendre sequences, and let

$$q_i = \begin{cases} l'_i, & \text{if } i \equiv 0 (\mathrm{mod} 2), \\ l_i, & \text{if } i \equiv 1 (\mathrm{mod} 2). \end{cases}$$
$$r_i = \begin{cases} l'_i, & \text{if } i \equiv 0 (\mathrm{mod} 2), \\ l_i + 1, & \text{if } i \equiv 1 (\mathrm{mod} 2). \end{cases} \quad (14)$$

Then the sequence $u : u_i = \phi^{-1}(q_i, r_i)$ is a balanced quaternary sequence with optimal autocorrelation values [8].

Let y be a sequence defined as

$$y_i = \begin{cases} 0, & \text{if } (q_i, r_i) = (0,0), \\ 1, & \text{if } (q_i, r_i) = (0,1), \\ \mu+1, & \text{if } (q_i, r_i) = (1,1), \\ \mu, & \text{if } (q_i, r_i) = (1,0). \end{cases} \tag{15}$$

In [8] the linear complexity of $\{y_i\}$ was investigated over $\mathbb{F}_n$ for $n > 4$.

Lemma 12: Let q, r be defined by (14). Then:

(i) $s_q(x) = (1+x^p)s_l(x) + 1$;

(ii) $s_r(x) = (1+x^p)s_l(x) + 1 + x\frac{x^{2p}-1}{x^2-1}$.

We prove Lemma 12 similarly as Lemma 10.

Theorem 13: Let sequence y be defined by (15). Then $LC = 2p$ and $m(x) = x^{2p} - 1$.

Proof: By Lemma 1 $s_y(x) = \mu s_q(x) + s_r(x)$ hence from Lemma 11 we obtain

$$s_y(x) = (1+x^N)(\mu+1)s_l(x) + \mu + 1 + x\frac{x^{2N}-1}{x^2-1}.$$

From this we can establish that $s_y(1) = \mu$ and $s_y(\alpha^j) = \mu + 1$ for $j = 1, \ldots, p-1$ or $\gcd(x^{2N}-1, s_v(x)) = 1$. The conclusion of this theorem follows from (3). ■

May 18, 2014

REFERENCES

[1] T.W. Cusick, C. Ding, and A. Renvall. *Stream Ciphers and Number Theory*. North-Holland Publishing Co., Amsterdam (1998)

[2] C. Ding, T. Helleseth, and W. Shan. "On the linear complexity of Legendre sequences". *IEEE Trans. Inform. Theory*, vol. 44, pp. 1276-1278,1998

[3] C. Ding. "Linear complexity of generalized cyclotomic binary sequences of order 2". *Finite Fields Appl*., vol. 3, pp. 159-174,1997

[4] V.A. Edemskii. "On the linear complexity of binary sequences on the basis of biquadratic and sextic residue classes". *Discret. Math. Appl.*, vol. 20, no. 1, pp. 75-84, 2010 (*Diskretn. Mat.*, vol.22, no. 1, pp.74-82, 2010)

[5] S.W. Golomb, G. Gong. *Signal Design for Good Correlation: For Wireless Communications, Cryptography and Radar Applications*. Cambridge University Press (2005)

[6] M. Hall M. *Combinatorial Theory*. Wiley, New York (1975)

[7] K. Ireland, M. Rosen M. *A Classical Introduction to Modern Number Theory*. Springer, Berlin (1982)

[8] Y-S. Kim, J-W. Jang, S-H. Kim, and J-S. No. "New Quaternary Sequences with Ideal Autocorrelation Constructed from Legendre Sequences". *IEICE Trans. Fund. Electron.*, vol. E96-A, no. 9, pp. 1872-1882, 2013.

[9] J.H. Kim, H.Y. Song. "On the linear complexity of Hall's sextic residue sequences". *IEEE Trans. Inform. Theory,* vol.47, pp. 2094-2096, 2001

[10] J.J. Komo, L.L. Joiner. *QPSK sequences over* F_4, in: ISIT. Washington. DC, 2001

[11] N. Li, X. Tang. "On the linear complexity of binary sequences of period $4N$ with optimal autocorrelation value/magnitude". *IEEE Trans. Inf. Theory,* vol.57, pp. 7597-7604, 2011

[12] T. Lim, J-S. No, and H. Chung. "New Construction of Quaternary Sequences with Good Correlation Using Binary Sequences with Good Correlation". I*EICE Trans. Fundamentals.* vol.E94-A, no.8), pp. 1701-1705, 2011

[13] X.H. Tang, C. Ding. "New classes of balanced quaternary and almost balanced binary sequences with optimal autocorrelation value". *IEEE Trans. Inf. Theory,* vol.56, pp. 6398-6405, 2010

[14] Q. Wang, X. N. Du. "The linear complexity of binary sequences with optimal autocorrelation". *IEEE Trans. Inf. Theory,* vol.56, no. 6388-6397, 2010

Properties of weak linear spaces

Dan-Mircea Borş and Anca Croitoru

***Abstract*—In this paper different properties and considerations on weak linear spaces are established. Some examples, comparative results and properties of a weak norm are also presented.**

***Keywords*—weak linear space, weak norm, metric.**

I. Introduction

Different problems in computer science, optimization and functional analysis led to the definition of a space that satisfies only a part from the axioms of a linear space. Different kinds of such spaces were introduced in [1], [2], [3], [4], [8], [11], [12], [13], [14] and a survey on all these spaces is given in [10].

In the present work we establish some properties, considerations and examples concerning weak linear spaces. A comparison with almost linear spaces and properties of a weak norm are also presented.

II. Preliminaries

We denote $\mathbb{R}_+ = [0, +\infty)$.

Definition 1. Let T be a nonempty set and $\mathcal{P}(T)$ the family of all subsets of T. A subfamily $\emptyset \neq \mathcal{A} \subseteq \mathcal{P}(T)$ is called *an algebra* of subsets of T if it satisfies for every $A, B \in \mathcal{A}$:

(i) $A \cup B \in \mathcal{A}$.
(ii) $A \backslash B \in \mathcal{A}$.
(iii) $T \in \mathcal{A}$.

We now recall the notions of almost linear space and quasilinear space.

Definition 2. [14] A nonempty set X is called *an almost linear space* if it is endowed with two mappings "+":$X \times X \to X$ and "$\cdot$":$\mathbb{R} \times X \to X$ satisfying:

$(A1)$ $(\forall (x, y, z) \in X^3)((x + y) + z = x + (y + z))$.
$(A2)$ $(\forall (x, y) \in X^2(x + y = y + x)$.
$(A3)$ $(\exists \theta \in X)(\forall x \in X)(x + \theta = x), \theta$ is called *the neutral element*.
$(A4)$ $(\forall x \in X)(1 \cdot x = x)$.
$(A5)$ $(\forall x \in X)(0 \cdot x = \theta)$.
$(A6)$ $(\forall \alpha, x, y) \in K \times X^2)(\alpha(x + y) = \alpha x + \alpha y)$.
$(A7)$ $(\forall (\alpha, \beta, x) \in K^2 \times X)(\alpha(\beta x) = (\alpha\beta)x)$.
$(A8)$ $(\forall (\alpha, \beta, x) \in \mathbb{R}_+^2 \times X)((\alpha + \beta)x = \alpha x + \beta x)$.

Remark 3. [12] Let X be an almost linear space.

I. We denote $(-1) \cdot x$ by $-x$ and $x + (-y)$ by $x - y$. We remark that $x - x$ need not be equal to θ since an element of X does not have an inverse element. Denote $V_X = \{x \in X; x - x = \theta\}$ and $W_X = \{x \in X; x = -x\}$.
II. $(\forall \alpha \in \mathbb{R})(\alpha\theta = \theta)$.
III. $(\forall (\alpha, \beta, x) \in \mathbb{R}_-^2 \times X)((\alpha + \beta)x = \alpha x + \beta x)$.
IV. $(\forall (\alpha, \beta, x) \in \mathbb{R}^2 \times V_X)((\alpha + \beta)x = \alpha x + \beta x)$.
V. V_X is a linear subspace of X.
VI. $W_X = \{x - x; x \in X\}$ and W_X is an almost linear subspace of X.
VII. $V_X \cap W_X = \{\theta\}$
VIII. The following conditions are equivalent:
  (i) X is a linear space.
  (ii) $V_X = X$.
  (iii) $W_X = \{\theta\}$.

Definition 4 ([3])**.** A nonempty set X is called *a quasilinear space* if a relation "$\leq$" and two operations "+": $X \times X \to X$, "$\cdot$": $\mathbb{R} \times X \to X$ are defined on it and satisfy the following properties for every elements $x, y, z, t \in X$ and every real scalars $\alpha, \beta \in \mathbb{R}$:

(Q1) $(\forall x \in X)(x \leq x)$.
(Q2) $(\forall (x, y, z) \in X^3)((x \leq y) \wedge (y \leq z) \Rightarrow (x \leq z))$.
(Q3) $(\forall (x, y) \in X^2)((x \leq y) \wedge (y \leq x) \Rightarrow (x = y))$.
(Q4) $(\forall (x, y) \in X^2)(x + y = y + x)$.
(Q5) $(\forall (x, y, z) \in X^3)((x + y) + z = x + (y + z))$.
(Q6) $(\exists \theta \in X)(\forall x \in X)(x + \theta = x)$, θ is called the neutral element.
(Q7) $(\forall (\alpha, \beta, x) \in \mathbb{R}^2 \times X)(\alpha(\beta x) = (\alpha\beta)x)$.
(Q8) $(\forall (\alpha, x, y) \in \mathbb{R} \times X^2)(\alpha(x + y) = \alpha x + \alpha y)$.
(Q9) $(\forall x \in X)(1 \cdot x = x)$.
(Q10) $(\forall x \in X)(0 \cdot x = \theta)$.
(Q11) $(\forall (\alpha, \beta, x) \in \mathbb{R}^2 \times X)((\alpha + \beta)x \leq \alpha x + \beta x)$.
(Q12) $(\forall (x, y, z, t) \in X^4)((x \leq y) \wedge (z \leq t) \Rightarrow (x + z \leq y + t))$.
(Q13) $(\forall (\alpha, x, y) \in \mathbb{R} \times X^2)((x \leq y) \Rightarrow (\alpha x \leq \alpha y))$.

An element $x' \in X$ is called an inverse of $x \in X$ if $x + x' = \theta$. Obviously, if x has an inverse x', then x' is unique. If every element x in a quasilinear space X has an inverse $x' \in X$, then X becomes a real linear space.

Now we recall the concept of a weak linear space. Borş [4] defined weak linear spaces in 1969, initially named quasivector spaces.

In 1985 quasilinear spaces were defined by Aseev [3] and then studied by many authors. So quasivector spaces have been renamed weak linear spaces [7] and studied in [5], [6], [7], [9].

Definition 5 ([4])**.** Let $(K, +, \cdot)$ be a field with distinct 0 and 1. A nonempty set X is called *a weak linear space* over K (shortly K-wls) if it is endowed with two mappings "+": $X \times X \to X$ and "$\cdot$": $K \times X \to X$, satisfying:

(W1) $(\forall (x, y, z) \in X^3)((x + y) + z = x + (y + z))$.
(W2) $(\forall (x, y) \in X^2)(x + y = y + x)$.

Technical University "Gh. Asachi" Iaşi, Romania, borsdm@yahoo.com
University "Al. I. Cuza" Iaşi, Romania, croitoru@uaic.ro

ISBN: 978-1-61804-256-9

(W3) $(\exists\theta \in X)(\forall x \in X)(x+\theta = x)$.
(W4) $(\forall x \in X)(\exists(-x) \in X)(x+(-x) = \theta)$.
(W5) $(\forall(\alpha, x, y) \in K \times X^2)(\alpha(x+y) = \alpha x + \alpha y)$.
(W6) $(\forall(\alpha, \beta, x) \in K^2 \times X)(\alpha(\beta x) = (a\beta)x)$.
(W7) $(\forall x \in X)(1 \cdot x = x)$.
(W8) $(\forall(\alpha, \beta, \gamma, x) \in K^3 \times X)((\alpha+\beta+\gamma)x = \alpha x + \beta x + \gamma x)$.

If $K = \mathbb{R}$, then X is called *a real weak linear space* (shortly, wls).

Example 6 ([4])**.** I. Every linear space over a field K is a K-wls.

II. Every abelian periodic group $(G, +)$ of order 2 is a K-wls over an arbitrary field K with respect to the mappings "+" and the multiplication "$\cdot$" defined by $\alpha \cdot x = x$, for every $(\alpha, x) \in K \times G$. But G is not a linear space.

III. Let $\mathcal{A}$ be an algebra of subsets of a nonempty set T and $(X, +, \cdot)$ a weak linear space over a field K. Then the set $U = \mathcal{A} \times X$ is a K-wls relative to the operations:

$$(A, x) + (B, y) = (A \triangle B, x + y), \forall (A, x), (B, y) \in U,$$
$$\alpha(A, x) = (A, \alpha x), \forall \alpha \in K, \forall (A, x) \in U,$$

where Δ is the symmetric difference. The neutral element of addition in U is $\theta = (\emptyset, 0)$, where 0 is the neutral element of addition of X and the opposite of $u = (A, x) \in U$ is $-u = (A, -x)$. But U is not a linear space.

IV. Let $(G, +)$ be an abelian periodic group of order 2 and $(X, +, \cdot)$ a K-wls over a filed K. Then $U = G \times X$ is a K-wls (but not linear) with respect to the laws:

$$(a, x) + (b, y) = (a + b, x + y), \forall (a, x), (b, y) \in U,$$
$$\alpha(a, x) = (a, \alpha x), \forall \alpha \in K, (a, x) \in U.$$

In this case, we have $\theta = (e, 0)$ and $-(a, x) = (-a, -x)$, $\forall (a, x) \in U$, where e and 0 are the neutral elements for the addition of G and X respectively.

Remark 7 ([4])**.** Let $(X, +, \cdot)$ be a K-wls and denote $X^\circ = \{0 \cdot x | x \in X\}$, $X_0 = \{x + 0 \cdot x | x \in X\}$.

I. $(\exists(\alpha, \beta, x) \in K^2 \times X)((\alpha+\beta)x \neq \alpha x + \beta x)(0 \cdot x \neq \theta)$.
II. $(\forall \alpha \in K)(\alpha\theta = \theta)$.
III. $(\forall(\alpha, x) \in (K \backslash \{0\}) \times X)((\alpha x = \theta) \Rightarrow (x = \theta))$.
IV. $(\forall n \in \mathbb{N}^*)(\forall(\alpha_1, \ldots, \alpha_n) \in K^n)(\forall x \in X)$

$$(\alpha_1 + \ldots + \alpha_n)x = \begin{cases} \alpha_1 x + \ldots + \alpha_n x, & n \text{ is odd} \\ \alpha_1 x + \ldots + \alpha_n x + 0 \cdot x, & n \text{ is even.} \end{cases}$$

V. X° is a weak linear subspace of X.
VI. X_0 is a linear subspace of X.
VII. $X^\circ \cap X_0 = \{\theta\}$.
VIII. $(\forall x \in X)(0 \cdot x + 0 \cdot x = \theta)$.

As we noticed in [10] there is no relationship between weak linear spaces and quasilinear spaces.

III. Properties of weak linear spaces

In this section some properties of weak linear spaces are established.

We begin by observing that there is no relationship between weak linear spaces and almost linear spaces.

Example 8. I. Let $\mathcal{A}$ be an algebra of subsets of a nonempty set T and X a real linear space. Then the set $U = \mathcal{A} \times X$ is a wls relative to the operations:

$$(A, x) + (B, y) = (A \triangle B, x + y), \forall (A, x), (B, y) \in U,$$
$$\alpha(A, x) = (A, \alpha x), \forall \alpha \in \mathbb{R}, (A, x) \in U,$$

where $\triangle$ is the symmetric difference. The neutral element of U is $\theta = (\emptyset, 0)$, where 0 is the neutral element of X and the inverse of $u = (A, x) \in U$ is $-u = (A, -x)$.

Let be $A \in \mathcal{A}$, $A \neq \emptyset$ and $u = (A, x) \in U$. We have $0 \cdot u = 0 \cdot (A, x) = (A, 0) \neq (\emptyset, 0) = \theta$.

(A5) is not accomplished and so U is not an almost linear space.

II. Let be $X = \mathbb{R}_+$ and the operation defined for every $x, y \in X$ and every $\alpha \in \mathbb{R}$ by

$$x + y = \max\{x, y\}$$
$$\alpha x = \begin{cases} x, & \alpha \neq 0 \\ 0, & \alpha = 0. \end{cases}$$

Then X is an almost linear space with $\theta = 0$, but it is not a weak linear space since (W4) is false.

Proposition 9. *Let X be a weak linear space. Then:*
(1) $0 \cdot x + 0 \cdot x = \theta$.
(2) $(\forall(n, x) \in \mathbb{N}^* \times X)(0 \cdot (2nx) = \theta)$.

Proof. Let $x \in X$ be arbitrary. By (W5) and (W8) we have:

$$x = 1 \cdot x = (1 + 0 + 0)x = x + 0 \cdot x + 0 \cdot x.$$

From (W3) and (W6) it results $0 \cdot (2x) = 0 \cdot (x + x) = 0 \cdot x + 0 \cdot x = \theta$. So

$$(\forall x \in X)(0 \cdot (2x) = \theta). \tag{3}$$

According to (3), replacing x by $2x$, it follows $0 \cdot (4x) = \theta$.

Now, we recurrently obtain (2). □

Proposition 10. *Let X be a weak linear space over K. Then*

(W8)$\Leftrightarrow$(W8′) $(\forall(\alpha, \beta, x) \in K^2 \times X)((\alpha + \beta)x = \alpha x + \beta x + 0x)$.

Proof. (W8)$\Rightarrow$(W8)′ For every $\alpha, \beta \in K$, $x \in X$ we have:

$$(\alpha + \beta)x = (\alpha + \beta + 0)x \overset{(W8)}{=} \alpha x + \beta x + 0 \cdot x.$$

(W8)′$\Rightarrow$(W8) For every $\alpha, \beta, \gamma \in K$, $x \in X$ it results:

$$(\alpha + \beta + \gamma)x = ((\alpha + \beta) + \gamma)x$$
$$\overset{(W8)'}{=} (\alpha + \beta)x + \gamma x + 0 \cdot x \overset{(W8)'}{=}$$
$$\alpha x + \beta x + 0 \cdot x + \gamma x + 0 \cdot x$$
$$\overset{(W2)}{=} \alpha x + \beta x + \gamma x + 0 \cdot x + 0 \cdot x \overset{(1)}{=} \alpha x + \beta x + \gamma x.$$

ISBN: 978-1-61804-256-9

□

Proposition 11. *Let X be a weak linear space over K. Then the following properties hold:*

(i) $(\forall(\alpha,\beta,x)\in K^2\times X)((\alpha+\beta)x=\alpha x+\beta x+0\cdot x)$.
(ii) $(\forall(\alpha,\beta,x)\in K^2\times X)(\alpha x+\beta x=(\alpha+\beta)x+0\cdot x)$.

Proof. (i) $(\alpha+\beta)x=(\alpha+\beta+0)x\overset{(W8)}{=}\alpha x+\beta x+0\cdot x$.

(ii) $(\alpha+\beta)x+0\cdot x=(\alpha+\beta+0)x+0\cdot x\overset{(W8)}{=}(\alpha x+\beta x+0\cdot x)+0\cdot x\overset{(W1)}{=}(\alpha x+\beta x)+(0\cdot x+0\cdot x)\overset{(1)}{=}(\alpha x+\beta x)+\theta=\alpha x+\beta x$. □

In the sequel we define a weak norm on a weak linear space and present some of its properties.

Firstly, we recall the concept of a normed group.

Definition 12. A *norm* on a group $(G,+)$ is a function $\|\cdot\|:G\to\mathbb{R}_+$ satisfying the properties:

(N1) $(\forall x\in G)((\|x\|=0)\Leftrightarrow(x=0))$.
(N2) $(\forall(\alpha,x)\in\mathbb{Z}\times G)(\|\alpha x\|=|\alpha|\|x\|)$.
(N3) $(\forall(x,y)\in G^2)(\|x+y\|\le\|x\|+\|y\|)$.

The ordered pair $(G,\|\cdot\|)$ is called a *normed group*.

We now give the notion of a weak norm on a weak linear space.

Definition 13 ([7])**.** Let X be a weak linear space. A function $w:X\to\mathbb{R}$ is called *a weak norm* on X if it satisfies the following:

(WN1) $(\forall x\in X)((w(x)=0)\Leftrightarrow(x=\theta))$.
(WN2) $(\forall x\in X)(w(-x)=w(x))$.
(WN3) $(\forall(x,y)\in X^2)(w(x+y)\le w(x)+w(y))$.

The couple (X,w) is called *a normed weak linear space* (shortly, *nwls*).

Example 14 ([7])**.** I. Let $(G,\|\cdot\|_G)$ be a normed abelian periodic group of order 2, $(X,\|\cdot\|_X)$ a real normed space and $U=G\times X$ defined as in Example 6-IV.

Let $w:U\to\mathbb{R}$ defined by $w((a,x))=\|a\|_G+\|x\|_X$. Then (U,w) is a normed weak linear space.

II. Let be $X=\mathbb{R}^p(p\in\mathbb{N}^*)$ and $w(x)=\sum_{i=1}^{p}\frac{|x_i|}{1+|x_i|}$ for every $x=(x_1,\dots,x_p)\in\mathbb{R}^p$.

Then w is a weak norm on X, that is not a norm on X.

III. Let be T a nonempty finite set, $\mathcal{A}=\mathcal{P}(T)$, $(U,\|\cdot\|)$ a real normed space and $X=\mathcal{A}\times U$ the weak linear space of Example 8-I. Now, let $w:X\to\mathbb{R}$ be the function defined by $w((A,u))=\operatorname{card}A+\|u\|$, where $\operatorname{card}A$ means the cardinality of A. Then w is a weak norm on X.

Theorem 15. *Let (X,w) be a normed weak linear space and $d:X\times X\to\mathbb{R}$ defined by $(\forall(x,y)\in X^2)$ $(d(x,y)=w(x-y))$. Then d is a metric on X (named the metric induced by the weak norm w) having the following properties:*

(i) $|w(x)-w(y)|\le w(x-y),\forall x,y\in X$;
(ii) $d(x+z,y+z)=d(x,y),\forall x,y,z\in X$.

Proof. From the axioms of Definition 13, it results that d is a metric.

(i) From (WN3) it follows $p(x)\le p(x-y)+p(y)$ and $p(y)\le p(y-x)+p(x)$ for every $x,y\in X$. This implies that

$$|p(x)-p(y)|\le p(x-y),\ \ \forall x,y\in X.$$

(ii) For every $x,y,z\in X$, we have:

$$d(x+z,y+z)=p((x+z)-(y+z))=$$
$$=p(x+z-y-z)=p(x-y)=d(x,y).$$

□

Conclusion

We presented some properties, considerations and examples of weak linear spaces. A comparison with almost linear spaces and properties of a weak norm are also established.

References

[1] Apreutesei, G. – *Hausdorff topology and some operations with subsets*, An. Şt. Univ. "Al. I. Cuza" Iaşi, XLIV (1998), 445-454.

[2] Apreutesei, G. – *The hyperspacial topologies and almost linear spaces*, An. Şt. Univ. "Al. I. Cuza" Iaşi, XLVIII (2002), 3-18.

[3] Aseev, S.M. – *Quasilinear operators and their applications in the theory of multivalued mappings* (in Russian), Tr. MIAN SSSR, Nauka Moskva, 167 (1985), 25-52; Proc. Steklov Inst. Math. 2 (1986), 23-52.

[4] Borş, D.M. – *O generalizare a spaţiilor vectoriale (A generalization of vector spaces)*, Bull. Inst. Polit. Iaşi, XV(XIX), s.I, Fasc. 3-4 (1969) 17-20 (in Romanian).

[5] Borş, D.M. – *Quasivector spaces over fields*, Bull. Inst. Polit. Iaşi, XVII (XXI), s. I, Fasc. 1-2 (1971), 47-52.

[6] Borş, D.M., Ciobanu, L., Gontineac, M., Grămadă, V. – *Quasi-linear functions*, Bull. Inst. Polit. Iaşi, XLIII (XLVII), Fasc. 3-4 (1997), 21-26.

[7] Borş, D.M., Croitoru, A. – *Weak-linear spaces*, submitted for publication.

[8] Climescu, Al. – *Les espaces à scalaires* ≥ 0, Bull. Inst. Polit. Iaşi, XI (XV), Fasc. 3-4 (1965), 1-6.

[9] Croitoru, A., Borş, D.M. – *Remarks on weak linear spaces*, submitted for publication.

[10] Croitoru, A., Apreutesei, G., Borş, D.M. – *Survey on different near linear spaces*, submitted for publication.

[11] Duffin, R.J., Karlovitz, L.A. – *Formulation of linear programs in analysis. I. Approximation theory*, SIAM J. Appl. Math. 16 (1968), 662-675.

[12] Godini, G. – *A framework for best simultaneous appoximation: normed almost linear spaces*, INCREST, Preprint Series in Mathematics, No. 30(1983), Journal of Approximation Theory, 43 (1985), 338-358.

[13] Löhne, A. – *On convex functions with values in semi-linear spaces*, Report of the Institute of Optimization and Stochastics 07, Martin-Luther-University Halle-Wittenberg, Department of Mathematics and Computer Science, 2003.

[14] Mayer, O. – *Algebraische und metrische strukturen in der intervallrechnung und einige anwendungen*, Computing 5 (1970), 144-162.

ISBN: 978-1-61804-256-9

The properties of solutions of the inverse paleotemperature problems

Oleg V. Nagornov, Sergey A. Tyuflin, and Tatiana I. Bukharova

***Abstract*—**The inverse problem on the past surface temperature reconstruction based on the measured borehole temperature in glaciers and rocks is studied. There were many such reconstructions, however, the properties of such solutions have not been derived. We find out that the solution of this problem is not unique and stable. The uniqueness and stability properties take place for the inverse problems that assume solution in the form of the finite segments of the Fourier series.

***Keywords*—**boreholes, climate reconstruction, heat and mass transfer, inverse problems.

I. Introduction

THE studies of the past temperatures at the Earth surface is important problem for prediction of the climate changes. The systematic instrumental temperature measurements took place no more than two centuries. Thus, indirect estimations of the past temperatures present main information on the past climate. It is considered that the measured temperatures in the boreholes can be used to reconstruct the past surface temperatures at the Earth.

The underground temperature distribution is mainly determined by two types of processes [1], [2]. The first is the surface temperature changes and the second is the heat flux from the Earth that is subjected to the long-time geological processes. The surface temperature changes take place at relatively smaller time scale. Therefore, the measured temperature-depth profiles in the borehole contain information on the climatic changes at the surface. The seasonal temperature variations at the surface are noticeable at depth about 10-15 meters while the climatic oscillations reach several hundred meters and more.

The heat and mass transfer in rocks and glaciers is described by the one-dimensional thermal diffusivity equation [1], [3], [4]. The past surface temperature reconstruction is the inverse problem that contains additional re-determination condition. The measured temperature-depth profile presents such condition. We found out that this problem has not the unique and stable solution in general case.

There are several well-known methods of the past surface temperature reconstructions: the Monte-Carlo method [5], [6]; the least-squares inversion method [1] and the singular value decomposition method for rock boreholes [7], [8]. These methods were used for local, regional and global reconstructions [5], [9]-[12]. We show that these temperature reconstructions are not unique and stable algorithms with mathematical point of view.

O. V. Nagornov is with the National Research Nuclear University MEPhI, Moscow, Russian Federation (corresponding author to provide phone: 7-499-3243255; e-mail: ovnagornov@mephi.ru).

S. A. Tyuflin is with the National Research Nuclear University MEPhI, Moscow, Russian Federation.

T. I. Bukharova, is with the National Research Nuclear University MEPhI, Moscow, Russian Federation.

II. Problem statement

The mathematical statement of the inverse problem consists of the thermal conductivity equation that takes into account the vertical advection term, the initial condition, the boundary condition at the bottom of glacier and the re-determination condition. The measured-temperature-depth profile is used as the re-determination condition, $\chi(z)$, where z is vertical coordinate. Then the inverse problem to find the temperature in the past is the solution of the following one-dimensional problem:

$$\begin{cases} T_t + w(z)T_z = a^2 T_{zz}, & 0 < t < t_f, \quad 0 < z < H, \\ T(0,\, t) = U_s + \mu(t), & 0 < t \le t_f, \\ -k \cdot T_z(H,\, t) = q, & 0 < t \le t_f, \\ T(z,\, 0) = U(z), & 0 < z < H, \\ T\left(z, t_f\right) = \chi(z), & 0 < z < H. \end{cases} \tag{1}$$

Here H is the ice sheet thickness, a^2 is the thermal diffusivity, k is the thermal conductivity, $w(z)$ is the vertical ice velocity, q is the geothermal heat flux, $U(z)$ is the steady-state temperature profile associated with this flux. U_s – is the initial temperature on the surface, which characterizes the average temperature that was on the surface in the past before the beginning of sharp temperature variations on the surface. $\mu(t)$ is temperature variations on the surface in time with respect to its initial value U_s from the moment $t=0$ ($\mu(0)=0$) to the time of measurements of the borehole temperature profile t_f.

Let us represent the borehole temperature profile $T(z,t)$ in the form of the superposition of two temperature profiles: the steady-state temperature profile $U(z)$

ISBN: 978-1-61804-256-9

associated with the geothermal heat flow from the Earth and the residual temperature profile *V(z,t)* associated with temperature variations on the surface:

$$T(z,t)=U(z)+V(z,t) \tag{2}$$

Then, the steady-state temperature profile *U(z)* is the solution of the problem specified as

$$\begin{cases} U_{zz}-\dfrac{w(z)}{a^2}U_z=0, & 0<z<H,\\ U(0)=U_s,\\ U_z(H)=-q/k. \end{cases} \tag{3}$$

Let us denote $\theta(z)=\chi(z)-U(z)$ is deviations from the steady-state temperature profile in the measured temperature profile. This deviations are associated with surface temperature changes. Thus, the problem of finding surface temperature history is reduced to the solution of the problem

$$\begin{cases} V_t+w(z)V_z=a^2V_{zz}, & 0<t<t_f, \quad 0<z<H,\\ V(0,\,t)=\mu(t), & 0<t\le t_f,\\ V_z(H,\,t)=0, & 0<t\le t_f,\\ V(z,\,0)=0, & 0<z<H,\\ V(z,t_f)=\theta(z), & 0<z<H. \end{cases} \tag{4}$$

III. Problem Investigation

Let us show that the inverse problem (4) in the general case has no the uniqueness solution.

Lemma 1.
In addition to the trivial solution ($V(z,t)\equiv 0$; $\mu(t)\equiv 0$), the inverse problem

$$\begin{cases} V_t+w(z)V_z=a^2V_{zz}, & 0<t<t_f, \quad 0<z<H,\\ V(0,\,t)=\mu(t), & 0<t\le t_f,\\ V_z(H,\,t)=0, & 0<t\le t_f,\\ V(z,\,0)=0, & 0<z<H,\\ V(z,t_f)=0, & 0<z<H. \end{cases} \tag{5}$$

has a nontrivial solution ($V(z,t)$; $\mu(t)$).

Proof.

Let us assume that $\mu(0)=\mu(t_f)=0$ and

$$\mu(t)=\sum_{m=1}^{\infty}\alpha_m\cdot\sin\left(\frac{\pi mt}{t_f}\right),$$

where α_m are unknown coefficients. Let $V^{(m)}(z,t)$ be a solution of the direct problem specified as

$$\begin{cases} V_t^{(m)}+w(z)V_z^{(m)}=a^2V_{zz}^{(m)}, & 0<t<t_f, \quad 0<z<H,\\ V^{(m)}(0,\,t)=\alpha_m\sin\left(\dfrac{\pi mt}{t_f}\right), & 0<t\le t_f,\\ V_z^{(m)}(H,\,t)=0, & 0<t\le t_f,\\ V^{(m)}(z,\,0)=0, & 0<z<H. \end{cases} \tag{6}$$

The solution of this problem is easily obtained in the form

$$V^{(m)}(z,t)=\alpha_m\cdot\sin\left(\frac{\pi mt}{t_f}\right)-\alpha_m\frac{\pi m}{t_f}\cdot\sum_{n=1}^{\infty}I_ne_n(z)\int_0^t\exp\left(-\lambda_n(t-\tau)\right)\cdot\cos\left(\frac{\pi m\tau}{t_f}\right)d\tau \tag{7}$$

where

$$I_n=\frac{1}{\left\|e_n(z)\right\|^2}\int_0^H e_n(z)\,dz,$$

$e_n(z)$ and λ_n are the eigenfunctions and eigenvalues of the following Sturm–Liouville problem:

$$\begin{cases} a^2Z''(z)-w(z)Z'(z)+\lambda Z=0, & 0<z<H\\ Z(0)=Z'(H)=0. \end{cases} \tag{8}$$

Let us show that the all the eigenvalues of problem (8) are real ($\lambda_n\in\mathbb{R}$). The first of Eqs. (8) is equivalent to the equation

$$a^2\frac{d}{dz}\left(p(z)Z'(z)\right)=\lambda\,p(z)Z(z), \tag{9}$$

where

$$p(z)=\exp\left(-\int_0^z\frac{w(\tilde z)}{a^2}d\tilde z\right).$$

After two integrations by parts and the use of Eq. (9), the term $\lambda\int_0^H p(z)Z(z)\overline{Z}(z)dz$ becomes $\overline{\lambda}\int_0^H p(z)Z(z)\overline{Z}(z)dz$. This means that $\lambda=\overline{\lambda}$ and $\lambda_n\in\mathbb{R}$.

The asymptotic behavior of the eigenvalues is known [13]: $|\lambda_n|\sim C\cdot n^2$, $n\to\infty$. Therefore, the series $\sum_{n=1}^{\infty}\frac{1}{|\lambda_n|}$ converges. Then, the set of the functions $\left\{e^{\lambda_nt}\right\}_{n=1}^{\infty}$ is incomplete in $L_2(0,t_f)$, this is a corollary of Müntz's theorem [14]. Thus, there is a nonzero function $F(t)$ specified at $t\in[0,t_f]$ such that $F(t)$ orthogonal to $\left\{e^{\lambda_nt}\right\}_{n=1}^{\infty}$ in $L_2(0,t_f)$. Let us expand $F(t)$ into the Fourier series at $t\in[0,t_f]$: $F(t)=\sum_{m=1}^{\infty}\beta_m\cdot\sin\left(\frac{\pi mt}{t_f}\right)$. Let us

prove that $V(z,t)=\sum_{m=1}^{\infty}V^{(m)}(z,t)$ is a solution of the problem specified by Eqs. (5) and $\mu(t)=F(t)$, $\alpha_m=\beta_m$. Indeed, $V(z,t)$ satisfies the first of Eqs. (5), as well as the initial and boundary conditions. Let us verify the last condition in Eqs. (5):

$$V(z,t_f)=-\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\frac{\pi\alpha_m m}{t_f}I_n e_n(z)\int_0^{t_f}e^{-\lambda_n(t_f-\tau)}\cos\left(\frac{\pi m\tau}{t_f}\right)d\tau. \quad (10)$$

The integration of Eq. (10) by parts yields

$$V(z,t_f)=\sum_{n=1}^{\infty}I_n e_n(z)\lambda_n e^{-\lambda_n t_f}\int_0^{t_f}e^{\lambda_n\tau}\left(\sum_{m=1}^{\infty}\alpha_m\sin\left(\frac{\pi m\tau}{t_f}\right)\right)d\tau. \quad (11)$$

Since the inner series in Eq. 11 is identically equal to $F(\tau)$ and is orthogonal to $\{e^{\lambda_n t}\}_{n=1}^{\infty}$, $V(z,t_f)=0$. Thus, we find the nontrivial solution $V(z,t)$ and the lemma is proved.

Thus, the solution of the problem (4) without additional constraints is not unique.

IV. Uniqueness and Stability

Let us assume that $\mu(t)=\sum_{k=-m}^{m}\mu_k\cdot\exp(i2\pi k\frac{t}{t_f})$ is a finite segment of the Fourier series. Let us show that the in this case the uniqueness of the function $\mu(t)$ can be proved.

Let us represent $V(z,t)$ from problem (4) as $V(z,t)=\mu(t)+W(z,t)$, then, the problem of finding surface temperature history is represented in the form

$$\begin{cases} W_t+w(z)W_z+f(t)=a^2W_{zz}, & 0<t<t_f, \quad 0<z<H,\\ W(0,t)=0, & 0<t\le t_f,\\ W_z(H,t)=0, & 0<t\le t_f,\\ W(z,0)=0, & 0<z<H,\\ W(z,t_f)=s(z), & 0<z<H. \end{cases} \quad (12)$$

Here $f(t)=\mu'(t)$ and $s(z)=\theta(z)-\mu(t_f)$. Since $\mu(t)$ is a finite segment of the Fourier series, $f(t)$ is a finite segment of the Fourier series too, $f(t)=\sum_{k=-m}^{m}f_k\cdot\exp(i2\pi k\frac{t}{t_f})$.

Since $f(t)=\mu'(t)$; $\mu(0)=0$, the function $\mu(t)$ is uniquely determined from $f(t)\in L_2[0,t_f]$. If the uniqueness of the function $f(t)$, is proved, then the uniqueness of the function $\mu(t)$ can be proved.

To prove uniqueness of the problem (12) it is sufficient to show that $W(z,t)\equiv 0$ and $f(t)\equiv 0$ if $s(z)=0$

The solution of the problem (12) is given by the formula

$$W(z,t)=\sum_{n=1}^{\infty}I_n e_n(z)\int_0^{t}e^{-\lambda_n(t-\tau)}\cdot f(\tau)d\tau. \quad (13)$$

Here $e_n(z)$ and λ_n are the eigenfunctions and eigenvalues of the following Sturm–Liouville problem (8).

It is known that $\{e_n(z)\}_{n=1}^{\infty}$ is the complete orthonormalized set, $\lambda_n\in R$, $\lambda_n\to\infty$, $n\to\infty$. From the condition $W(z,t_f)=0$, it follows that $\forall n\in N: I_n\int_0^{t_f}e^{-\lambda_n(t_f-\tau)}\cdot f(\tau)d\tau=0$. Here $I_n=-\int_0^{H}e^{z^2/2}\cdot e_n(z)dz\neq 0$ on a certain sequence of numbers. It follows from lemma 2.

Lemma 2.

Let $\{e_n(z)\}_{n=1}^{\infty}$ and $\{\lambda_n\}_{n=1}^{\infty}$ be the eigenfunctions and eigenvalues of the Sturm–Liouville problem, respectively. Then, there is a subsequence $\{k_n\}$ such that $I_{k_n}=-\int_0^{H}e^{z^2/2}\cdot e_{k_n}(z)dz\neq 0$ $(\forall\ k_n)$.

Proof.

Let us assume that this is not the case. Therefore, $\exists N\in N: I_n=0,\ \forall n\ge N$, i.e., function $\psi(z)=e^{z^2/2}$ is orthogonal to all the functions $e_n(z)$ with $n\ge N$. Since $\{e_n(z)\}_{n=1}^{\infty}$ is the orthonormalized basis in $L_2(0,H)$, $\psi(z)=\sum_{n=1}^{N-1}I_n\cdot e_n(z)$. However, $\psi(0)=1$ and $e_n(0)=0$ for all n values. Therefore, we arrive at a contradiction and the lemma is proved.

Thus, the integer function $F(\lambda)=\int_0^{t_f}e^{\lambda\tau}\cdot f(\tau)d\tau$ has the infinite number of zeros. When $f(t)$ is a finite segment of a Fourier series, this is possible only for $f(t)\equiv 0$. Therefore, $\mu(t)\equiv 0$. Thus, uniqueness is proved.

Let us show that this solution is stable. Let two solutions $W_1(z,t)$, $f_1(t)$ and $W_2(z,t)$, $f_2(t)$ of problem (12) correspond to close functions $s_1(z)$ and $s_2(z)$. Let us show that if $f(t)$ is a finite segment of a Fourier series, these solutions are closed.

From Eq. (13) and from overdetermination condition $W(z,t_f)=s(z)$, it follows that:

$$s(z)=\int_0^{t_f}K(z,\tau)f(\tau)d\tau. \quad (14)$$

Here $K(z,\tau)=\sum_{n=1}^{\infty}I_n e_n(z)e^{-\lambda_n(t_f-\tau)}$ is kernel of the linear operator.

Eq. (14) is the Fredholm integral equation of the first kind. This is a classical ill-posed problem. If $f(t)$ is a finite segment of a Fourier series, the uniqueness theorem is proved for Eq. (14); i.e., from the condition $s(z)=0$, $z\in[0,H]$, it follows that $f(t)=0$, $t\in[0,t_f]$.

ISBN: 978-1-61804-256-9

Since $\{e_n(z)\}_{n=1}^{\infty}$ is an orthonormalized basis in $L_2(0,H)$, $f(t)=\sum_{k=-m}^{m} f_k \cdot \exp(i2\pi k\frac{t}{t_f})$ Eq. (14) is equivalent to the following system of linear equations:

$$s_n = I_n \sum_{k=-m}^{m} f_k \cdot \int_0^{t_f} e^{-\lambda_n(t_f-\tau)} e^{i2\pi k\frac{\tau}{t_f}} d\tau, \qquad n=1,2,3.... \qquad (15)$$

where $s_n = \frac{1}{\|e_n(z)\|^2}\int_0^H s(z)e_n(z)dz$; f_k, $k=0,\pm1,\pm2,...\pm m$ is unknown.

The number of the equations is infinite, whereas the number of unknowns is *2m+1*, therefore, the system in the general case has no solution at arbitrary s_n values. Thus, the problem under investigation is reduced to the solution of the system of the algebraic equations of the form: **Af=s**, where

$$\mathbf{f}=\begin{bmatrix} f_{-m} \\ f_{-m+1} \\ \vdots \\ f_0 \\ \vdots \\ f_m \end{bmatrix}; \qquad \mathbf{s}=\begin{bmatrix} s_1 \\ \vdots \\ s_n \\ \vdots \end{bmatrix}; \qquad \mathbf{A}=(\alpha_{pq})_{p=1,2...;\ q=1,2,...,2m+1},$$

$$\alpha_{pq} = I_p \int_0^{t_f} e^{-\lambda_p(t_f-\tau)} e^{i2\pi q\frac{\tau}{t_f}}.$$

The uniqueness theorem is proved for the problem under investigation. Therefore, the homogeneous problem has only the trivial solution. Let us prove that this problem has the stability property in the following meaning. Let two solutions $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$ close in the norm correspond to columns $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ close in the norm $\|\mathbf{s}^{(1)}-\mathbf{s}^{(2)}\| = \sup_{i\in N}|s_i^{(1)}-s_i^{(2)}|$.

Let us find the image $\mathrm{Im\,A}$ of the linear operator A that specifies the transformation $R^n \to C_0$, where C_0 - is the space of the number sequences $(s_1, ..., s_n, ...)$ converging to zero (because the Fourier coefficients tend to zero) and is represented by the matrix **A** in a certain basis.

Statement.

If $\mathbf{e}_1,\ldots,\mathbf{e}_n$ constitute a basis in R^n, then the vectors $\mathrm{A}\mathbf{e}_1,\ldots,\mathrm{A}\mathbf{e}_n$ constitute a basis in $\mathrm{Im\,A}$.

Proof.

Let us consider an arbitrary vector $\mathbf{z}\in\mathrm{Im\,A}$. By the definition of $\mathrm{Im\,A}$, $\exists\,\mathbf{x}\in R^n$ such that A**x**=**z**.

Let us expand the vector **x** in the basis $\mathbf{e}_1,\ldots,\mathbf{e}_n$: $\mathbf{x}=x_1\mathbf{e}_1+x_2\mathbf{e}_2+\ldots+x_n\mathbf{e}_n$. Then, since the operator A is linear, $\mathbf{z}=\mathrm{A}\mathbf{x}=x_1\mathrm{A}\mathbf{e}_1+x_2\mathrm{A}\mathbf{e}_2+\ldots+x_n\mathrm{A}\mathbf{e}_n$. Therefore, such an expansion exists.

Let us prove the uniqueness of this expansion, i.e., linear independence of the elements $\mathrm{A}\mathbf{e}_1,\ldots,\mathrm{A}\mathbf{e}_n$.

If $\alpha_1\mathrm{A}\mathbf{e}_1+\alpha_2\mathrm{A}\mathbf{e}_2+\ldots+\alpha_n\mathrm{A}\mathbf{e}_n=0$ then $\mathrm{A}(\alpha_1\mathbf{e}_1+\alpha_2\mathbf{e}_2+\ldots+\alpha_n\mathbf{e}_n)=0$, owing to linearity; since the kernel of the operator is zero, we have $\alpha_1\mathbf{e}_1+\alpha_2\mathbf{e}_2+\ldots+\alpha_n\mathbf{e}_n=0$; therefore $\alpha_1=\alpha_2=\ldots=\alpha_n=0$ because $\mathbf{e}_1,\ldots,\mathbf{e}_n$ are linearly independent; thus, the statement is proved.

Corollary.

$\mathrm{Im\,A}$ is a finite-dimensional (*n*- dimensional) subspace of C_0.

The linear operator A transforms R^n to $V_n=\mathrm{Im\,A}$ and has zero kernel. Therefore, the operator A has the inverse operator A^{-1} that specifies the transformation $V_n \to R^n$ and is a linear bounded operator.

If $\mathrm{A}\mathbf{f}^{(l)}=\mathbf{s}^{(l)}$, *l=1,2*, then $\mathbf{f}^{(l)}=\mathrm{A}^{-1}\mathbf{s}^{(l)}$, *l=1,2*, therefore, $\mathbf{f}^{(2)}-\mathbf{f}^{(1)}=\mathrm{A}^{-1}(\mathbf{s}^{(2)}-\mathbf{s}^{(1)})$ and the estimate $\|\mathbf{f}^{(2)}-\mathbf{f}^{(1)}\| \le \|\mathrm{A}^{-1}\|\cdot\|\mathbf{s}^{(2)}-\mathbf{s}^{(1)}\|$ is valid for the stability of the inverse problem owing to the boundedness of the operator A^{-1}.

Thus, if solutions of the inverse problem in the form of the segments of the Fourier series exist for two "close" redefinitions, then these solutions are close to each other. Therefore, in that case, stability of surface temperature history reconstruction is proved.

V. Conclusion

In practices, the measured borehole temperature contains continuous set of harmonics. It is due to both errors of measurements and unknown nature of climatic changes. It means that the problem of the past surface temperature reconstruction based on the measured borehole temperature has not the uniqueness and stability properties.

In fact all previous reconstructions of the past surface temperatures implicitly assume that the retrieval surface temperatures can be presented by the finite set of harmonics. In these cases the amplitudes of the harmonics can be found and the solutions are unique and stable.

References

[1] P. Y. Shen and A. E. Beck, Least squares inversion of borehole temperature measurements in functional space, *Journal of Geophysical Research*, 96, pp. 19965–19979, 1991.

[2] P. B. Price, O. V. Nagornov, R. Bay, D. Chirkin, Yu. He, P. Miocinovic, A. Richards, et al., Temperature profile for glacial ice at the South Pole: Implications for life in a nearby subglacial lake, *Proceedings of the National Academy of Sciences*, 12, pp. 7844-7847, 2002.

[3] W. S. B. Paterson, The Physics of Glaciers, 3rd ed., Butterworth–Heinemann, Birlington, 1994.

[4] V. M. Kotlyakov, S. M. Arkhipov, K. A. Henderson, O. V. Nagornov, Deep drilling of glaciers in Eurasian Arctic as a source of paleoclimatic records, *Quaternary Science Reviews*, 23, pp. 1371-1390, 2004.

[5] D. Dahl-Jensen, K. Mosegaard, N. Gundestrup, G. D. Clow, S. J. Johnsen, A. W. Hansen and N. Balling, Past temperatures directly from the Greenland Ice Sheet, *Science*, 282, pp. 268-271, 1998.

[6] D. Mottaghy, G. Schwamborn, and V. Rath, Past climate changes and permafrost depth at the Lake El'gygytgyn site: implications from data and thermal modeling, *Climate of the Past*, 9, pp. 119-133, 2013.

[7] H. Beltrami, L.Z. Cheng, and J.C. Mareschal, Simultaneous inversion of borehole temperature data for past climate determination, *Geophysical Journal International*, 129, pp. 311-318, 1997.
[8] H. Beltrami, G. Matharoo, L. Tarasov, V. Rath, and J.E. Smerdon, Numerical studies on the Impact of the Last Glacial Cycle on recent borehole temperature profiles: implications for terrestrial energy balance. *Climate of the Past*, 10, pp. 1693-1706, 2014.
[9] R. N. Harris and D. S. Chapman, Geothermics and climate change: 1. Analysis of borehole tempratures with emphais on resolving power, *Journal of Geophysical Research*, 103(B4), pp. 7360-7370, 1998.
[10] S. Huang, H. N. Pollack and P.-Y. Shen, Temperature trends over the past five centuries reconstructed from borehole temperatures, *Nature*, 403, pp. 756-758, 2000.
[11] H. Beltrami, Climate from borehole data: Energy fluxes and temperatures since 1500, *Geophysical Research Letters*, 29, 2111, pp. 26-1-26-4, 2002.
[12] H. N. Pollack, D. Yu. Demezhko, A. D. Duchkov, I. V. Golovanova, S. Huang, V. A. Shchapov and J. E. Smerdon, Surface temperature trends in Russia over the past five centuries reconstructed from borehole temperatures, *Journal of Geophysical Research*, 108(B4), pp. 2-1-2-12, 2003.
[13] V. P. Mikhailov, Partial Differential Equations, Mir, Moscow, 1978.
[14] R. Paley and N. Wiener, The Fourier Transforms in the Complex Domain, American Mathematical Society Colloquium Publications, Vol. 19, pp. 116-123, 1934.

The Connection Between Topological Dimension and Some Classes of Operators

Cristina Şerbănescu
Faculty of Applied Sciences
University Politehnica of Bucharest
060042 Bucharest, Romania
Email: mserbanescuc@yahoo.com

Ioan Bacalu
Faculty of Applied Sciences
University Politehnica of Bucharest
060042 Bucharest, Romania
Email: dragosx@yahoo.com

***Abstract*—The purpose of this paper is to explore the role played by the dimension theory in the spectral theory. The work is concentrated on a class of compact sets denoted by C, as well as on examples of sets which do not belong to the class C. The implications between various spectral conditions are emphasized.**

I. Introduction

In 1968, C. Foiaş and I. Colojoară in their paper "Theory of generalized spectral operators" have formulated an open problem in spectral theory (the theory of decomposable operators [10]) namely if any decomposable is hard decomposable (in other words if the operator's restrictions to maximal spectral spaces are decomposable). The answer was partially given in [5] that the decomposable operators with spectrums of a topological dimension ≤ 1 are hard decomposable. On the other hand, E.J. Albrecht has built an example of a decomposable operator that is not hard decomposable. Therefore the answer to the open problem is negative: not all decomposable operators are hard decomposable. The problem of the existence of whole classes of operators (and evidently of certain individual operators) remains open, for which the answer is negative. In [6] the study of the restrictions and quotients of decomposable operators in relation to with an invariant subspace (which maybe isn't maximal spectral) it has been shown that these are S - decomposable (a notion introduced by I. Bacalu in [6]) where S is the intersection of the restriction and quotient spectrums. An interesting observations is that for an S - decomposable operator, $\dim S = 0$ implies $S = \emptyset$ meaning the operator is decomposable. The means by which the dimension theory applies to S - spectral and A_s - scalar (A_2 -spectral) operators will be analyzed in the second section of the article.

In the presented paper, we emphasize, rather systematically and briefly, some geometric and topological aspects and properties of certain sets from the plane and the spaces $\mathbb{R}^n$ and $\mathbb{C}^n$, respectively. We concentrate on the more particular but interesting case of the sets from the class $\mathbf{C}$ (see Definition 1.6), with a long series of significant applications in spectral theory. One of the main purposes of this work is to show the importance of (topological) dimension of the spectrum of an operator or an operator system in the study of their spectral properties, especially the cases of operators or operator systems having their spectra of dimensions 0 or 1 (or the case of restrictions and quotients of operators or systems). It turns out that dimension theory plays a natural role in the study of spectral theory.

The idea of defining the notion of topological dimension (different from that of linear algebra) belongs to Poincaré ([22], [23]), but later, it has been formulated more specifically by Brouwer ([9]). Since 1922, Menger ([20]) and Urysohn ([26]) have been developed the dimension theory in many works, where they were able to establish a larger number of basic properties of the dimension. Moreover, a variety of modern and significant results concerning this theory can be found in Hurewicz and Wallman [16], as well as in [2], [21]. In Romanian, we recommend paper [3].

The paper is organized as follows. Section 2 presents enough details of the topological dimension theory, especially about the spaces and sets (subsets) of dimensions 0 and 1 because they are relevant in the spectral theory applications. Enough examples of properties (including equivalent definitions) of sets of dimensions 0 are given, emphasizing sets of dimensions 1 from class $\mathbf{C}$ and some examples of sets of dimensions 0 which are not in class $\mathbf{C}$. Furthermore, Cantor's discontinuity is emphasized on the line as a model of a subset of dimension 0 as well as some sets homeomorphic with it.

In section 3, the beginning presents a few notions of spectral theory, resolvent, spectrum, maximal spectral space, S-overlay, decomposable operator (S-decomposable), S-spectral measure, S-spectral operator etc. The means in which the theory of the third dimension is used in spectral theory is given by theorems 3.1 and 3.2 in I.Bacalu's PhD thesis. C.Foiaş names theorem 3.2 a fundamental theorem and C. Apostol says that in this theorem the importance of the spectrum's dimension is described in conserving the spectral properties of the operator. The rest of the section reconsider a few of the results obtained in [5] ,[7] ,[8] using the theory of dimension.

II. Topological dimension and the class $\mathbf{C}$

Definition II.1. *Let X be a separable metric space. The notation $\dim_p X$ means the dimension of X at the point $p \in X$. The following three conditions define the notion of dimension at a point through induction (as an inductive concept).*

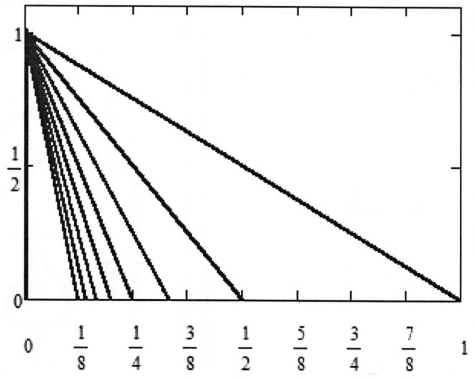
1) $\dim X = -1$ *means* $X = \emptyset$*;*

2) *if* $X \neq \emptyset$, $\dim X$ *is the supremum of* $\dim_p X$ *for any* $p \in X$*;*

ISBN: 978-1-61804-256-9

3) $\dim_p X \leq n+1$ *if there exists a neighborhood of* p *with the boundary of dimension less than or equal to* n.

We recall that a topological space X *is said to be locally connected at a point* $p \in X$ *if for any open subset* $G \subset X$ *which contains* p*, there exists a connected component* V *of* X *such that* $p \in V \subset G$ *(or every neighborhood of* p *contains a connected open neighborhood). The space* X *is locally connected if it is locally connected at each point of it. The "broom set" shown in the figure below is not locally connected. This set is composed by line segments joining the points (0,1) and (0,0), respectively (0,1) and* $\left(\frac{1}{n}, 0\right)$ *for* $n = 1, 2, 3, \ldots$ *; the set is locally connected only at point (0,1).*



We can also provide another definition of the topological dimension formulated in [18]:

Remark 2.1.

(1) The empty set and only it is of dimension -1.

(2) If n is a positive integer, we recall that a topological separable metric space X is of dimension $\leq n$ at the point $p \in X$, and we write $\dim_p X \leq n$, if p has arbitrarily small neighborhoods, each having a boundary of dimension $\leq n-1$.

(3) X is of dimension $\leq n$, and we write $\dim X \leq n$, if X is of dimension $\leq n$ at any point $p \in X$.

(4) X is of dimension n at the point $p \in X$, and we write $\dim_p X = n$, if (2) is true for n, but false for $n-1$.

(5) X is of dimension n, and we write $\dim X = n$, if the condition $\dim X \leq n$ is true, but the condition $\dim X \leq n-1$ is false.

The statement (2) $\dim X \leq n$ is equivalent to the fact that the space X has a basis all of whose elements are open sets with their boundaries of dimension $\leq n-1$.

By definition, a non-empty topological space is *of dimension* 0 if for any point of it there are arbitrarily small neighborhoods whose boundary is empty. Under this definition, for example, the space of rational numbers on the real axis is of dimension 0: each interval with irrational endpoints is a neighborhood for the numbers which are contained within and has an empty boundary (the boundary has irrational numbers). Similarly, the set of irrational numbers, and more generally any boundary set of the real axis are of dimension 0. The space of real numbers $\mathbb{R}$ is of dimension less than or equal to 1, since the boundary of an interval is a two-point set of dimension 0. Analogously, the plane $\mathbb{R}^2$ is of dimension less than or equal to 2 (since the circle is of dimension less than or equal to 1), and generally speaking the Cartesian space $\mathbb{R}^n$ is of dimension less that or equal to n. The proof of the fact that $\mathbb{R}^n = n$ is not elementary. The dimension of a subset of a space is never bigger that the one of the space itself. The set of all points of the plane whose coordinates are one rational and the other one irrational is of dimension 0. The set of all points of the Euclidian n-space $\mathbb{E}^n$ whose coordinates are irrational is of dimension 0. A non-empty set of real numbers is of dimension 0 if it does not contain any nontrivial interval.

Definition II.2. *The set* C *consisting of all real numbers* t *written as*

$$t = \frac{t_1}{3} + \frac{t_2}{3^2} + \frac{t_3}{3^3} + \ldots + \frac{t_n}{3^n} + \ldots$$

where $t_i \in \{0, 2\}$*, for* $i = 1, 2, \ldots$*, is called* the Cantor set (discontinuum)*. Therefore, these are the numbers in the [0,1] interval that have a triadic expansion in which the digit 1 does not occur; for example, the point* $\frac{1}{3}$ *belongs to* C*,*

$$\frac{1}{3} = \frac{0}{3} + \frac{2}{9} + \frac{2}{27} + \ldots + \frac{2}{3^n} + \ldots = (0, 0222)_3,$$

but $\frac{1}{2}$ *does not belong to* C.

We can geometrically describe the Cantor set C as the result of an iterative process: we divide the closed real interval [0,1] into three equal subintervals and first we remove the central open interval $\left(\frac{1}{3}, \frac{2}{3}\right)$, i.e. the middle-third interval; next, from the remaining two closed intervals, $\left[0, \frac{1}{3}\right]$, $\left[\frac{2}{3}, 1\right]$, again one removes their open middle-thirds and we continue in this way infinitely often; consequently, we obtain an infinite sequence of open middle-thirds intervals:

$$\left(\frac{1}{3}, \frac{2}{3}\right), \left(\frac{1}{9}, \frac{2}{9}\right), \left(\frac{7}{9}, \frac{8}{9}\right), \left(\frac{1}{27}, \frac{2}{27}\right), \ldots .$$

What remains of [0,1] at the end of this process is the Cantor set C.

We also remind the following interesting topological properties of the Cantor set:

(i) The Cantor set is of dimension 0 (since the Cantor set has no interval in it) and uncountable.

(ii) The Cantor set C is naturally homeomorphic to the Cartesian product of countable infinite copies of the discrete two-point space $\{0, 2\}$, i.e.

$$C \overset{top}{=} \{0, 2\} \times \{0, 2\} \times \{0, 2\} \times \ldots$$

hence the points of the Cantor set can be identified with the sequences consisting entirely of $0s$ and $2s$, whence

$$C \times C \overset{top}{=} C, C \times C \times \ldots \times C \overset{top}{=} C,\ C \times C \times C \times \ldots C \times \ldots \overset{top}{=} C.$$

(iii) On the other hand, it can be proved that the closed interval [0,1] is a continuous image of the Cantor set and the Hilbert cube $[0, 1] \times [0, 1] \times \ldots \times [0, 1] \times \ldots$ is also a continuous

image of the Cantor set. More generally, any compact metric space is a continuous image of the Cantor set.

(iv) The Cantor set C is regarded as a universal space for the class of all separable metric spaces of dimension 0, meaning that any separable metric space of dimension 0 is topologically embeddable in the Cantor set.

We can also remind certain properties of the topological spaces of dimension 0. Therefore, every non-empty space of dimension 0: has a basis consisting of clopen (closed-open) sets; is topologically embedded in the Cantor set, meaning that it is homeomorphic to a subset of the Cantor set; can be represented as a finite union of closed disjoint sets with diameter $< \varepsilon$, where $\varepsilon > 0$ is arbitrary; has the normalization property: for every two closed disjoint sets A and B, there exists a clopen set G such that $A \subset G$ and $G \cap B = \emptyset$. The union of a countable family of closed sets of dimension 0 is a set of dimension 0.

Theorem II.1. ([3]) *A subset A of $\mathbb{R}^n$ is of dimension n if and only if A contains a non-empty open subset of $\mathbb{R}^n$, i.e.* Int $A \neq \emptyset$ *in* $\mathbb{R}^n$.

A characterization of the space of dimension 0 will be useful to us:

Theorem II.2. ([19]) *A non-empty topological space X is of dimension 0 if for any finite open covering of X, $X = G_0 \cup G_1 \cup \ldots \cup G_m$, there exists a closed covering of X, $X = F_0 \cup F_1 \cup \ldots \cup F_m$ with the property that $F_i \subset G_i$ and $F_i \cap F_j = \emptyset$ $(i \neq j)$ for all $i, j = 0, 1, 2, \ldots, m$. The sets are therefore clopen (simultaneously closed and open).*

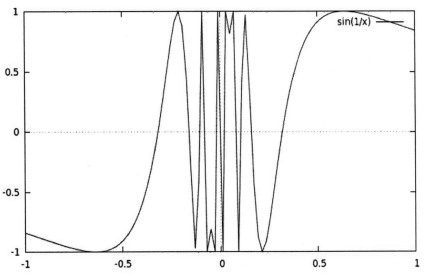
From Theorem 2.2, it follows that the boundary of a subset of the plane is of dimension less than or equal to 1, and the boundary of a set from the real axis is of dimension less than or equal to 0.

If we consider a compact subset L of the plane $\mathbb{R}^2$ such that L is of dimension 1, is it true that the boundary of any compact subset $L_1 \subset L$ (in the relative topology of L) is of dimension 0? There are examples of compact sets of dimension 1 for which the answer of this question is negative.

Example 1. The set Γ is defined as follows:

$$\Gamma : \begin{cases} \left\{ \left(x, \sin \dfrac{1}{x} \right),\ x \in (0, 1] \right\} = M \\ \{(0, y),\ -1 \leq y \leq 1\} = F. \end{cases}$$

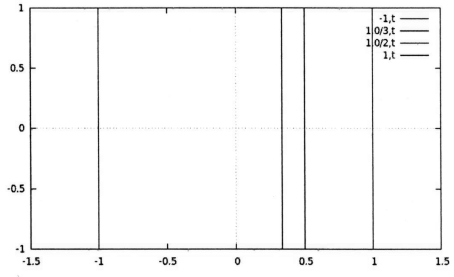
We have $\Gamma = M \cup F$ is compact, $F \subset \Gamma$ is compact and the boundary of F has the property: $\partial F = \partial M = F$, $\dim \partial F = \dim F = 1$.



Example 2. Another example is the "fan set":

$$\Gamma_1 : \begin{cases} \{(x, 1 - nx),\ x > 0, 1 - nx > 0, n = 1, 2, \ldots\} = M_1 \\ \{(0, y),\ 0 \leq y \leq 1\} = F_1 \end{cases}$$

where $\Gamma_1 = M_1 \cup F_1$, $F_1 \subset \Gamma_1$ is closed, $\partial F_1 = \partial M_1 = F_1$, $\dim \partial F_1 = \dim F_1 = 1$.



Example 3. If we consider the set:

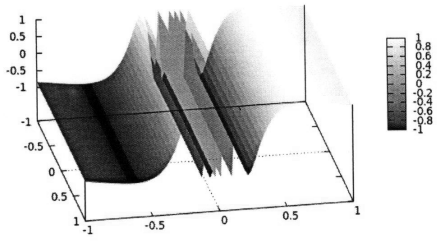
$$\Gamma_2 : \begin{cases} \left\{ \left(\dfrac{1}{n}, y \right),\ -1 \leq y \leq 1, n = 1, 2, \ldots, \right\} = M_2 \\ \{(0, y),\ -1 \leq y \leq 1\} = F_2 \end{cases}$$

we have $\Gamma_2 = M_2 \cup F_2$, $F_2 \subset \Gamma_2$, is compact, $\partial F_2 = F_2$, $\dim \partial F_2 = \dim F_2 = 1$.

Some other examples can also be constructed in the space $\mathbb{R}^3$ (in addition to the three above of $\mathbb{R}^2$) of compact sets of dimension 2 which not have compact subsets with boundary of dimension 1. Each of the previous examples can be viewed as a set of $\mathbb{R}^3$ if we add $-1 \leq x \leq 1$; obviously, the resulting sets are parts of a cylindrical surface; thus, for example:

$$H : \begin{cases} \left\{ \left(x, \sin \dfrac{1}{x} \right),\ x \in (0, 1] \right\} = M \\ \{(0, y),\ -1 \leq y \leq 1\} = F \\ -1 \leq z \leq 1 \end{cases}$$

is a cylindrical surface. All subsets of H of dimension 2 which contain the square $P = \{(y, z), -1 \leq y \leq 1, -1 \leq z \leq 1\}$ have the boundary also of dimension 2, because a part of the boundary is actually P.



For decomposable operators it seems that not only the compact sets of dimensions 1 (spectra of operators or parts of them) that have a good behavior as mentioned above imply a great interest (see Section 2).

ISBN: 978-1-61804-256-9

Our main emphasis is on a class of complex compact sets which plays an important role in what follows:

Remark 2.2. We shall denote by $\mathbf{C}$ the class of all compact sets $\sigma \subset \mathbb{C}$ with $\dim \sigma \leq 1$, and moreover having the property that for any closed subset $\sigma_1 \subset \sigma$ we have $\dim \partial\sigma_1 = 0$ (the boundary $\partial\sigma_1$ being taken in the relative topology of σ).

The family $\mathbf{C}$ is non-empty: every interval or finite union of intervals on the real axis belongs to the class $\mathbf{C}$; every set of the plane that is homeomorphic to [0,1] belongs to the class $\mathbf{C}$; finite unions of sets from $\mathbf{C}$ belong to $\mathbf{C}$; the disk $\{|\lambda| = 1\}$ belongs to $\mathbf{C}$. Let us remark that the countable union of sets of $\mathbf{C}$ may not belong to $\mathbf{C}$.

Example 4. The set

$$L = [0,1] \cup \bigcup_{n=1}^{\infty} \left\{ (x,y) | x \in [0,1],\ y = \frac{1}{n} \right\}$$

does not belong to $\mathbf{C}$. The sets $\Gamma, \Gamma_1, \Gamma_2$, L are not locally connected and do not belong to $\mathbf{C}$; probably there exists a relationship between the sets that do not belong to $\mathbf{C}$ and the sets that are not locally connected.

Definition II.3. *Let A_1, A_2 and B be pairwise disjoint subsets of a topological space X. The set B is called a partition between A_1 and A_2 (or, A_1 and A_2 are separated by B) in X if there are two disjoint open subsets A_1', A_2' in $X \setminus B$ such that $X \setminus B = A_1' \cup A_2'$, $A_1 \subset A_1'$, $A_2 \subset A_2'$. If A_1, A_2 are separated by the empty set $B = \emptyset$, it is said that A_1, A_2 are separated.*

A non-empty topological separable metric space X is of dimension 0 if and only if any point $p \in X$ and any closed set $D \subset X$, which does not contain p, can be separated (i.e. for any point $p \in X$ and for any closed set $D \subset X$, which does not contain p, the empty set is a partition between p and D). Also, a connected space X of dimension 0 is one-point space. Let us observe that for a separable metric space X, with $\dim X = 0$, any two disjoint closed sets in X may be separated.

We recall that a topological space is *totally disconnected* if the connected component of each point of it is the set consisting of the point itself (i.e. if the empty set is a partition between any distinct points x, y of the space X). One proves that a locally compact space X is of dimension 0 if and only if it is totally disconnected.

Next, we present the statement of certain original theorems in which we show the role of the spectral theory of dimension in the spectral theory of operators. The proofs of these theorems can found in [5]-[7] and [8].

Theorem II.3. *(The addition theorem for dimension 0). Let X be a separable metric space and let $(A_i)_{i=1}^{\infty}$ be a countable family of closed subsets of X such that $X = \bigcup_{i=1}^{\infty} A_i$, where $\dim A_i = 0$, $i = 1, 2, \ldots$ Then X is of dimension 0, i.e. $\dim X = 0$.*

Theorem II.4. *(The addition theorem). Let X be a separable metric space and let $(A_i)_{i=1}^{\infty}$ be a countable family of closed subsets of X such that $X = \bigcup_{i=1}^{\infty} A_i$. Then $\dim X = \sup_{i=1}^{\infty} \dim A_i$.*

Theorem II.5. *(The separation theorem for dimension 0). If X is a separable metric space of dimension 0, then for any two disjoint closed subsets A, B of X the empty set is a partition between A and B, i.e. there exists a clopen set $U \subset X$ such that $A \subset U$ and $B \subset X \setminus U$.*

Theorem II.6. *(The decomposition theorem for dimension n). A separable metric space is of dimension $\leq n$ (where n is finite) if and only if it can be represented as the union of $n+1$ subspaces of dimension ≤ 0.*

III. The connection between topological dimension and certain classes of operators

In the monograph "Theory of Generalized Spectral Operators" [10], Colojoară and Foiaş presented seven unsolved problems related to the theory of decomposable operators. One of them has the following statement: if $T \in B(X)$ is decomposable and $Y \subset X$ is a spectral maximal space of T (more general, Y is an invariant subspace to T), are then the operators restriction $T|Y$ and quotient $\dot{T}$ also decomposable? The problem seems to have the correct answer: sometimes yes, sometimes no.

In 1973 in [5], a partial answer was given, namely that for operators with the spectrum in the class $\mathbf{C}$, the answer is affirmative, hence the restriction $T|Y$ and the quotient $\dot{T}$ are decomposable (whence T is strongly decomposable).

In order to illustrate the usefulness of topological dimension in spectral theory, we recall several notations and basic definitions from the specialized literature, that will be employed throughout this paper.

Let $B(X)$ denote the Banach algebra of all linear bounded operators on a given complex Banach space X and let $\mathcal{P}(X)$ denote the set of all projectors on X. If $Y \subset X$ is a linear (closed) subspace invariant of $T \in B(X)$ (i.e. $TY \subset Y$), then $T|Y$ is the restriction of T to Y and $\dot{T}$ is the operator induced by T in the quotient space $\dot{X} = X/Y$. We also denote by $\sigma(T) = \mathbb{C} \setminus \rho(T)$ the spectrum of T, where the *resolvent set* $\rho(T)$ consists of all points $\lambda \in \mathbb{C}$ for which the operator $\lambda I - T$ is bijective on X.

Recall that a linear subspace $T \subset X$ is a *spectral maximal spectral space* of $T \in B(X)$ if $TY \subset Y$ and for any other subspace Z of X with $TZ \subset Z$ and $\sigma(T|Z) \subset \sigma(T|Y)$, we have $Z \subset Y$ ([10]).

A family $\{G_S\} \cup \{G_i\}_{i=1}^n$ of open sets of $\mathbb{C}$ is said to be an *S-covering* of a compact set $\sigma \subset \mathbb{C}$ if:

$$\sigma \cup S \subset G_S \cup \left(\bigcup_{i=1}^{n} G_i \right),\ \bar{G}_i \cap S = \emptyset\ (i = 1, 2, \ldots, n)$$

where S is a compact fixed subset of $\mathbb{C}$ ([24]). If $S = \emptyset$, an S-covering becomes a covering.

An operator $T \in B(X)$ is *S-decomposable* if for any open S-covering $\{G_S\} \cup \{G_i\}_{i=1}^n$ of the spectrum $\sigma(T)$, there exists

a system $\{Y_S\}\cup\{Y_i\}_{i=1}^n$ of spectral maximal spaces of T such that:

(i) $\sigma(T|Y_S)\subset G_S,\ \sigma(T|Y_i)\subset G_i\ (i=1,2,\ldots,n)$.

(ii) $X=Y_S+\sum_{i=1}^{n}Y_i$ ([6]).

If condition (ii) is replaced by (ii')

(ii') $Z=Z\cap Y_S+\sum_{i=1}^{n}(Z\cap Y_i)$, where Z is any spectral maximal space of T, then T is *strongly S-decomposable* ([6]). When $S=\emptyset$, then T is *decomposable* ([10]), respectively *strongly decomposable* ([4]).

Let B_S be the family of all Borelian sets $B\subset\mathbb{C}$ with the property that $B\cap S=\emptyset$ or $B\supset S$, where S is a compact fixed set of $\mathbb{C}$. A mapping $E_S:B_S\to\mathcal{P}(X)$ is said to be an *S-spectral measure* if

1. $E_S=\emptyset,\ E_S(\mathbb{C})=I$

2. $E_S(B_1\cap B_2)=E_S(B_1)E_S(B_2),\ B_1,B_2\in B_S,$

3. $E_S\left(\bigcup_{m=1}^{\infty}B_m\right)x=\sum_{m=1}^{\infty}E_S(B_m)x,\ B_m\in B_S,\ B_m\cap B_p=\emptyset,\ m\neq p,\ x\in X$

4. $\sup_{B\in S_S}||E_S(B)||<\infty$ ([7]).

An operator $T\in B(X)$ is said to be *S-spectral* if there exists an S-spectral measure E_S such that

$TE_S(B)=E_S(B)T$ and $\sigma(T|E_S(B)X)\subset\bar{B},\ B\in B_S$ ([7]).

For $S=\emptyset$, we obtain a spectral measure and a spectral operator ([15]).

The following main results of the paper refer to the implication of the topological dimension in deriving some properties of certain classes of operators (decomposable, with the spectrum in $\mathbf{C}$, S-decomposable, strongly S-decomposable, spectral etc).

Theorem III.1. *Let $T\in B(X)$ be a decomposable operator, let $Y\subset X$ be a spectral maximal space of T and $S=\partial\sigma(T|Y)\cap\sigma(\dot{T})$ of dimension 0. Then both $T|Y$ and $\dot{T}$ are decomposable operators.*

Proof: Let $\{G_i\}_{i=1}^n$ be a finite open covering of $\sigma(T|Y)$ (respectively, of $\sigma(\dot{T})$). Putting $G_i'=G_i\cap\rho(\dot{T})$ (respectively, $G_i'=G_i\cap\rho(T|Y)$), it results that the family $\{G_i'\}_{i=1}^n$ is a covering of $\sigma(T|Y)\setminus S$ (respectively, of $\sigma(\dot{T})\setminus S$). It is clear that $\{G_i\}_{i=1}^n$ is a covering of S as well.

Since $S=\partial\sigma(T|Y)\cap\sigma(\dot{T})$ is of dimension 0 and it is closed, then from Lemma 6.1, [3], one can deduce that for the covering $\{G_i\}_{i=1}^n$ of S, there exists an open covering $\{G_i''\}_{i=1}^n$ of S such that $G_i''\subset G_i,\ G_i''\cap G_j''=\emptyset$ for $i\neq j,\ i,j=1,2,\ldots,n$ and $\bigcup_{i=1}^{n}G_i''\supset S$. Taking $\bigcup_{i=1}^{n}G_i''=G_S$, it is obvious that $\{G_S\}\cup\{G_i'\}_{i=1}^n$ is an S-covering of $\sigma(T|Y)$ (respectively, of $\sigma(\dot{T})$). According to Proposition 2 and Remark 3, [5], it results that

$$Y=(Y_1+Y_2+\ldots+Y_n)+(Y_S^1+Y_S^2+\ldots+Y_S^n)$$

where Y_i and Y_S^i, $i=1,2,\ldots,n$ are spectral maximal spaces of $T|Y$ such that $\sigma(T|Y_i)\subset G_i,\ \sigma(T|Y_S^i)\subset G_i''$.

If we denote

$$X_i=X_T(\sigma(T|Y_i)\cup\sigma(T|Y_S^i))$$

then $X_i,\ i=1,2,\ldots,n$ are spectral maximal spaces of $T|Y$ (Theorem 2.1.5, [10]) and one can obtain $Y=\sum_{i=1}^{n}X_i$, $\sigma(T|X_i)\subset G_i$, hence $T|Y$ is decomposable.

Analogously for $\dot{T}$, according to Remark 3, [5], we have

$$\dot{X}=(\dot{Z}_1+\dot{Z}_2+\ldots+\dot{Z}_n)+(\dot{Z}_S^1+\dot{Z}_S^2+\ldots+\dot{Z}_S^n)$$

where $\dot{Z}_i$ and $\dot{Z}_S^i$, $i=1,2,\ldots,n$ are spectral maximal spaces of $\dot{T}$ such that $\sigma(\dot{T}|\dot{Z}_i)\subset G_i',\ \sigma(\dot{T}|\dot{Z}_S^i)\subset G_i''$, with G_i',G_i'' the same as above. It follows immediately from the proof of Lemma II.2.2, [4] that $\dot{T}$ is decomposable. ■

Theorem III.2. *Let $T\in B(X)$ be a decomposable operator with $\sigma(T)\in\mathbf{C}$. Then both T and T^* are strongly decomposable.*

Proof: The case $\dim\sigma(T)=0$ is contained in [10]. Therefore we have to analyse only the case $\dim\sigma(T)=1$. Let Y be a spectral maximal space of T and $S=\partial\sigma(T|Y)\cap\sigma(\dot{T})$. Since $\sigma(T)\in\mathbf{C}$, it follows that $\dim\partial\sigma(T|Y)=0$ and consequently $\dim S=0$. According to Theorem 3.1, $T|Y$ is decomposable, hence T is strongly decomposable (see Theorem II.3.6, [4]). From Corollary 3.1, [25], it results that T^* is also strongly decomposable. ■

Corollary 3.1. *If $T\in B(X)$ is a decomposable operator with $\sigma(T)\subset\mathbb{R}$ (or $\sigma(T)$ is on a curve), then T is strongly decomposable.*

Remark. The above corollary has already been observed by Foiaş and Apostol.

Corollary 3.2. *Let $T\in B(X)$ be a decomposable operator with $\sigma(T)\in\mathbf{C}$ and let $Y\subset X$ be a spectral maximal space of T. Then $\dot{T}\in B(\dot{X})$ is strongly decomposable.*

Proof: From Theorem 3.2, T is strongly decomposable and then by Theorem II. 3.8, [4], we have that $\dot{T}$ is strongly decomposable. ■

Corollary 3.3. *Let $T\in B(X)$ be a 3-decomposable operator with $\sigma(T)\in\mathbf{C}$. Then the operators $T,T^*,T^{**},\ldots$ are strongly decomposable. If X is reflexive, then T^* is strongly decomposable if and only if T is 3-decomposable.*

Proof: If T is 3-decomposable and $\sigma(T)\in\mathbf{C}$, then by the proof of Theorem 3.1 we deduce that T is strongly 3-decomposable, hence T is strongly decomposable. According to Theorem 3.2, $T^*,T^{**},\ldots$ are strongly decomposable. ■

Corollary 3.4. *If $T\in B(X)$ is a decomposable operator and Y is a spectral maximal space of T, then both $T|Y$ and $\dot{T}$*

are S-decomposable operators, where $S = \partial\sigma(T|Y) \cap \sigma(\dot{T})$, $\operatorname{Int} S = \emptyset$ (i.e. $\dim S \leq 1$) and $S_{\dot{T}} = \emptyset$.

Proposition III.1. *Let $T \in B(X)$ be an S-decomposable operator with $S_T = \emptyset$ and let S_1 be a separated part of S with $\dim S_1 = 0$. Then T is S'-decomposable where $S' = S \setminus S_1$.*

Theorem III.3. *Let $T \in B(X)$ be an S-decomposable operator such that $\dim S = 0$. Then T is decomposable.*

Proof: In the previous proposition we take $S_1 = S$, hence $S' = S \setminus S_1 = \emptyset$ and T is $\emptyset$-decomposable, i.e. T is decomposable. ∎

Corollary 3.5. *If $T \in B(X)$ is decomposable and T is an invariant subspace to T such that $\dim(\sigma(T|Y) \cap \sigma(\dot{T})) = 0$ (in particular, $\dim \sigma(T|Y) = 0$), then $\dot{T}$ is decomposable. When Y is a spectral maximal space of T, both $T|Y$ and $\dot{T}$ are decomposable.*

Proof: The assertions follow easily from the previous theorem and Corollary 3.4. ∎

Theorem III.4. *Let $T \in B(X)$ be a strongly S-decomposable operator such that $\dim S = 0$. Then T is strongly decomposable.*

Proof: Let Y be a spectral maximal space of T. According to Lemma 1.3.15, [7], we deduce that $T|Y$ is S_1-decomposable with $S_1 = S \cap \sigma(T|Y)$. Therefore $\dim S_1 = 0$ and by Theorem 3.3, it follows that $T|Y$ is decomposable, i.e. T is strongly decomposable. ∎

Corollary 3.6. *Let $T \in B(X)$ be strongly decomposable and let Y be an invariant subspace to T such that $\dim(\sigma(T|Y) \cap \sigma(\dot{T})) = 0$ (particularly, $\dim \sigma(T|Y) = 0$ or $\dim \sigma(\dot{T}) = 0$). Then $\dot{T} = B(\dot{X})$ is strongly decomposable.*

Furthermore, a number of important results from the classes of spectral (S-spectral) operators and also of the operator systems are mentioned here only with suitable references, but without proofs.

Theorem III.5. *Let $T \in B(X)$ be a spectral operator and let Y be a subspace invariant of T such that $X_T(\sigma) \subset Y$, where $\sigma = \sigma(T|Y) \setminus \sigma(\dot{T})$ and $S = \sigma(T|Y) \cap \sigma(\dot{T})$. Then both $T|Y$ and $\dot{T}$ are S-spectral operators.*

Corollary 3.7. *Let $T \in B(X)$ be a spectral (scalar) operator and let Y be an invariant subspace to T such that $\dim(\sigma(T|Y) \cap \sigma(\dot{T})) = 0$. Then $T|Y$ and $\dot{T}$ are spectral (scalar).*

Corollary 3.8. *Let H be a Hilbert space and let $T \in B(H)$ be a normal operator. If Y is an invariant subspace to T such that $\dim S = 0$, where $S = \sigma(T|Y) \cap \sigma(\dot{T})$, then $T|Y$ and $T|H - Y$ are normal.*

Proposition III.2. *Let $T \in B(X)$ be a subscalar operator and $\tilde{T} \in B(\tilde{X})$ the minimal scalar extension of T. Then T is S-scalar, where $S = \sigma(T) \cap \sigma(\dot{\tilde{T}})$, $\dot{\tilde{T}}$ being the operator induced by $\tilde{T}$ in the quotient space $\dot{X} = \tilde{X}/X$.*

Proposition III.3. *Let H be a Hilbert space and let $T \in B(X)$ be a subnormal operator. With the same conditions as in the previous proposition, T is S-normal.*

Theorem III.6. *Let $a = (a_1, a_2, \ldots, a_n) \subset B(X)$ be a decomposable operator system of and let Y be a spectral maximal space of a. Then the system $a|Y = (a_1|Y, a_2|Y, \ldots, a_n|Y)$ and $\dot{a} = (\dot{a}_1, \dot{a}_2, \ldots, \dot{a}_n)$, the system induced by a in the quotient space $\dot{X} = X/Y$, are S-decomposable, where $S = \sigma(a, Y) \cap \sigma(\dot{a}, \dot{X})$. If $\dim S = 0$ then $a|Y$ and $\dot{a} = (\dot{a}_1, \dot{a}_2, \ldots, \dot{a}_n)$ are decomposable.*

Theorem III.7. *Let $a = (a_1, a_2, \ldots, a_n) \subset B(X)$ be a spectral system such that $\dim \sigma(a, X) = 0$. Then for any closed subspace $Y \subset X$ invariant of a, the restriction $a|Y = (a_1|Y, a_2|Y, \ldots, a_n|Y)$ is a spectral system.*

Corollary 3.9. *Let $a = (a_1, a_2, \ldots, a_n) \subset B(X)$ be a spectral system and let Y be a closed invariant subspace of a such that $\dim \sigma(a, Y) = 0$. Then $a|Y$ is a spectral system.*

IV. Conclusion

In this paper we observed the efficiency the spectral techniques of the theory of dimension in spectral theory. We underline that the application of the dimension theory in the theory of operators appears for the first time in [5]. Using these ideas in [1] and [25] we obtain new results. In the second section we showed some results obtained in other classes of decomposable (S-decomposable) operators, spectral operators, A-spectral (A_s-scalar) but especially the multidimensional spectral theory (operator systems).

References

[1] E.J. Albrecht and F.H. Vasilescu, *On spectral capacities*, Rev. Roum. Math. Pures et Appl., **18**, 701-705 (1974).

[2] P.S. Aleksandrov, *Dimension theory*, Mat. Ann., **106** (1932).

[3] C. Andreian-Cazacu, A. Deleanu and M. Jurchescu, *Topology. Categories. Riemann Surfaces* (Romanian), Ed. Academiei R.S.R. (1966).

[4] C. Apostol, *Spectral theory and functional calculus* (Romanian), St. Cerc. Mat., **20**, 635-668 (1968).

[5] I. Bacalu, *On restrictions and quotients of decomposable operators*, Rev. Roum. Math. Pures et Appl., **18**, 809-813 (1973).

[6] I. Bacalu, *S-decomposable operators in Banach spaces*, Rev. Roum. Math. Pures et Appl., **20**, 1101-1107 (1975).

[7] I. Bacalu, *Residual spectral decompositions I* (Romanian), St. Cerc. Mat, **32**, 467-504 (1980).

[8] I. Bacalu, *Residual spectral decompositions II* (Romanian), St. Cerc. Mat, **32**, 587-623 (1980).

[9] L.E.J. Brouwer, *Über die natürlichen Dimensionbegriff*, J. Reine. Angew. Math., **142**, 146-152 (1913).

[10] I. Colojoară and C. Foiaş, *Theory of generalized spectral operators*, Gordon Breach, Science Publ., New York-London-Paris (1968).

[11] H.R. Dowson, *Restrictions of spectral operators*, Proc. London Math. Soc., **15**, 437-457 (1965).

[12] H.R. Dowson, *Operators induced on quotient spaces by spectral operators*, J. London Math. Soc., **42**, 666-671 (1967).

[13] H.R. Dowson, *Spectral theory of linear operators*, London Math. Soc. Monographs, **12**, Academic Press, London and New-York (1978).

[14] N. Dunford, *Spectral Operators*, Pacific. J. Math., **4**, 321-354 (1954).

[15] N. Dunford and J.T. Schwartz, *Linear Operators*, Interscience Publishers, New York, part I (1958), part II (1963), part III (1971).

[16] W. Hurewicz and H. Wallman, *Dimension theory*, Princeton University Press (1941).

[17] Şt. Frunză, *An axiomatic theory of spectral decompositions for systems of operators I* (Romanian), St. Cerc. Mat., **27**, 655-711 (1975).

[18] K. Kuratowski, *Introduction to set theory and topology*, Pergamon Press (1961).

[19] K. Kuratowski, *Topology*, Academic Press, New-York, Vol. I (1966), Vol. II (1968).

[20] K. Menger, *Dimensionstheorie*, B.G. Teubner, Leipzig-Berlin (1928).

[21] J. Nagata, *Modern dimension theory*, North-Holland Publishing Co (1965).

[22] H. Poincaré, *Analysis situs*, J. de l'École Polytechniques ser 2, Vol. I, 1-123 (1895).

[23] H. Poincaré, *Complément à l'Analysis situs*, Rendiconti del Circolo Matematico di Palermo **13**, 285-343 (1899).

[24] F.H. Vasilescu, *On the residual decomposability in dual spaces*, Rev. Roum. Math. Pures et Appl., **16**, 1573-1578 (1971).

[25] F.H. Vasilescu, *Analytic functional calculus and spectral decompositions*, Ed. Academiei (Bucharest, Romania), D. Reidel Publishing Company Dordrecht: Holland / Boston: U.S.A. / London: England (1981).

[26] P.S. Urysohn, *Mémoire sur les multiplicités cantoriennes*, Fund. Math. 7-8, 30-157 (1925-1926).

ISBN: 978-1-61804-256-9

Challenge to Create an Estimator for Failure-Detection in safety related systems

O. Krini, A. Krini and J. Börcsök

***Abstract*—** The paper provides an overview of how to create an estimator focusing on failure-defection in safety related systems. Stochastic play a very important role in safety technology. With the help of it, safety systems may be released reliably after an assessment. With the help of the probability theory meaningful statements are achieved and based on them, realistic forecasts may be given. However, in order that reliable forecasts can be conducted, new approaches in thinking need to be developed. This paper serves to give a short synopsis about the actual problem of the probabilistic safety technology on the base of stochastic.

In that, the test methods, however, plays the most important role as the test results are source vectors for probabilistic models. However, this paper tries to describe a suitable, innovative method that will correctly estimate the safety parameters. The first part explains the necessary basic tools. Furthermore, the safety technology is explained with stochastic playing a central role in it. Following this, the approach for the construction of the estimator is introduced. Concluding, the summary and an outlook will be given.

***Keywords*—**Stochastic, mathematical models, safety, Probability, Reliability, Failure/Error, Matrix-calculations

I. Introduction

To To a greater extent than previously, safety related systems have been developed, produced and released to the market. For this reason it is essential, to know and correctly and reasonably apply the current international norms for functional safety as a basis for systems that are used in safety-critical applications.

The functional safety is part of the overall safety in terms of the EUC standards and the EUC control system. It is subjected to the correct function of the E/E/PRE safety-related systems, safety-related systems of other technologies and external devices for risk reduction. In this process it is unimportant whether it refers to a control system or the complete installation.

Concerning the safety of a system, the default rate plays an important role. It describes the amount of default per unit of time and has the unit „FIT“. On principle, when examining errors, it can be differentiated between safe (λS) and dangerous errors (λD). Safe errors, whether they have been found or remain unfound, normally have no influence on the safety-function of a system. However, concerning dangerous mistakes, this is not true. If such errors occur, the system will be transferred into a dangerous state, which under certain circumstances may lead to the massive endangerment of human lives. These errors too are differentiated in dangerous and traceable (λDD) or dangerous and non traceable (λDU) errors.

Concerning dangerous and traceable errors, if accordingly designed, the safety system may bring the overall system or the installation in a safe state. The critical state, however, is given through the non traceable, dangerous errors. If such errors occur in the safety system, there is no possibility to detect it. In the system they may lead to its switch off or, in the worst case, to its dangerous breakdown.

In order to be able to run systems or installations that can be applied in safety related areas, comprehensive measures for development and certification are necessary. These serve to prevent these described dangerous situations from happening and to bring the safety system or the installation into a safe state.

On the base of the default rates the reliability functions and the default probabilities are determined. The distribution of cumulative frequencies plays a central role in this. The challenge is to choose the right density function. Afterwards the model parameters and the safety parameters need to be estimated.

The following chapter will show the new mathematic approach of how an estimator can be constructed in a structured way.

II. Safety technology based on probabilistic approach

According to the norm, the functions of all safety related systems form the functional safety of the overall system. Determining a level of safety integrity (SIL) forms the central element. The SIL is one of four discrete steps towards specification of the requirement for safety integrity of the safety functions related to the E/E/PE safety related system, with level 4 being the highest level of safety integrity, level 1 the lowest. Therefore the IEC-standard 61508 consists of 4 safety levels SIL 1 through 4. Each of these appears in a confidence interval, Fig. 1 showing the distribution of the probability.

ISBN: 978-1-61804-256-9

SIL	Operation with low demand rate PFD_{avg}	Operation with high demand rate PFH [1/h]
4	$10^{-5} \le PFD_{avg} < 10^{-4}$	$10^{-9} \le PFH < 10^{-8}$
3	$10^{-4} \le PFD_{avg} < 10^{-3}$	$10^{-8} \le PFH < 10^{-7}$
2	$10^{-3} \le PFD_{avg} < 10^{-2}$	$10^{-7} \le PFH < 10^{-6}$
1	$10^{-2} \le PFD_{avg} < 10^{-1}$	$10^{-6} \le PFH < 10^{-5}$

Fig. 1 SIL bei niedriger und hoher Anforderungsrate nach IEC

A. *Effective Distribution in safety theory*

The reliability R(t) is the probability that a unit is functional in one view period (0, t). Fig. 2 shows R (t) as function of time [1], [6].

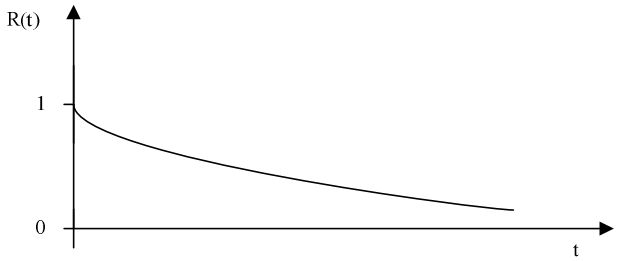


Fig. 2 SIL bei niedriger und hoher Anforderungsrate nach IEC

The probability that the operational time T is within the considered time interval (0...t) is for small t almost equal to one. For larger values of t the probability decreases more and more.

$$R(t) = e^{-\int_0^{\infty} \lambda(t)\,dt} \tag{1}$$

The exponential distribution is useful in many applications in engineering, for example, to describe the lifetime X of a transistor. The most known and most favorite probability model for the reliability analysis of safety systems is the exponential distribution. With this distribution it is possible to represent the time dependent probability $F(t)$ of components for which it is necessary to obtain observed data to determine X.

The failure probability is defined by the exponential distribution as

$$F(t) = 1 - e^{-\lambda \cdot t} \tag{2}$$

where λ is the failure rate. Respectively with failure density

$$f(t) = \begin{cases} \lambda \cdot e^{-\lambda \cdot t} & for\ \ t \ge 0 \\ 0 & otherwise \end{cases} \tag{3}$$

If an exponential distribution for the reliability is valid, then the failure rate is constant:

$$\lambda(t) = \lambda \tag{4}$$

Then the equation can be rewritten as:

$$R(t) = e^{-\lambda \cdot t} \tag{5}$$

An important reliability parameter is the MTTF value (Mean Time To Failure).

$$MTTF = \int_0^{\infty} R(t)dt = \frac{1}{\lambda} \tag{6}$$

If an exponential distribution is suitable equation [8] can be rewritten as:

$$MTTF = \frac{1}{\lambda} \tag{7}$$

Within the interval (0, t] the probability of failure P(t) is calculated applying the reliability function R(t).

$$\begin{aligned} &P(t) = 1 - R(t) \\ &P(t) = 1 - e^{-\lambda \cdot t} \\ &P(t) \approx \lambda \cdot t \qquad for \quad \lambda \cdot t << 1 \end{aligned} \tag{8}$$

Generally, the time t is applied by T1. The time from point in time zero to time T1 is characterized as proof test interval. At time T1 a periodical test or the maintenance of a safety system is taking place. Tests are carried out to allocate undetected, dangerous failures. After a proof test, the system is regarded as new. The calculated PFD-valued depends on the value T1. [1], [6], [9]

In order to be able to make probabilistic statements about possible values of safety parameters, according to the architecture. Different models for analysis can be drawn. In the following these will be introduced.

B. *One out of one Architecture (1oo1)*

The 1oo1 architecture is the simplest safety system around and consists of only one channel. Every dangerous fault can lead towards a failure of the safety function [1], [5], [9]. The 1oo1 architecture is presented in Fig. 3.

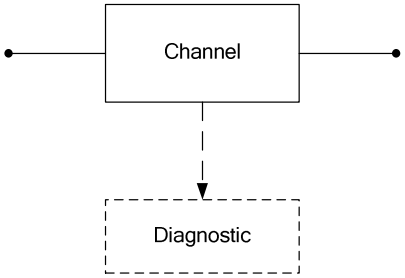


Fig.3 1oo1 architecture

If $\lambda = \lambda_D$ is applied to equation [8] then the result in the

following equation is for the 1oo1 system:

$$P(t) = 1 - e^{-\lambda_D \cdot t} \tag{9}$$

P (t) is developed by the MacLaurin series. For the 1oo1 system the first three terms are needed to be developed. The first three terms plus the remaining term R3 are sufficient for the calculation of the PFDavg values.

$$e^{-\lambda_D \cdot T} = 1 - \lambda_D \cdot T + \frac{\lambda_D{}^2 \cdot T^2}{2!} + R_3 \tag{10}$$

The description of the remaining term R3 is chosen as follows:

R3 is the remaining term to the third order, which belongs to the exponential function with failure rate λ_D.

The remaining term R3 converges for T = 0 to the value 0 and can be neglected compared to the third term when developed towards the limit value at T = 0 [1].

Equation (10) is applied for a 1oo1 system. The PFDavg is:

$$PFD_{avg} = 1 + \frac{1}{\lambda_D \cdot T}\left[1 - \lambda_D \cdot T + \frac{\lambda_D{}^2 \cdot T^2}{2!} - 1\right] = \frac{\lambda_D \cdot T}{2} \tag{11}$$

with,

$$\frac{T}{2} = t_{CE} = \frac{\lambda_{DU}}{\lambda_D}\left(\frac{T_1}{2} + MTTR\right) + \frac{\lambda_{DU}}{\lambda_D} \cdot MTTR \tag{12}$$

Here, t_{CE} is the mean repair time of a channel. The Equation can be presented simplified as follows:

$$PFD_{avg,1oo1} = \lambda_{DU}\left(\frac{T_1}{2} + MTTR\right) + \lambda_{DD} \cdot MTTR = \lambda_D \cdot t_{CE} \tag{13}$$

$$PFH_{1oo1} = \lambda_{DU}$$

C. One out of two Architecture (1oo2)

The 1oo2 architecture, see Figure 4, possesses two channels in parallel, where each channel can execute the safety function by itself.

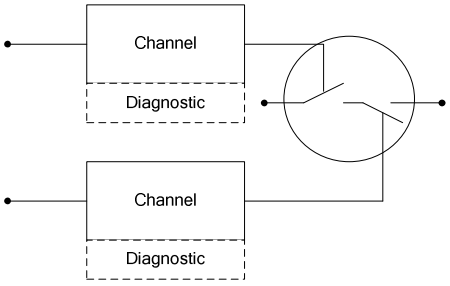


Fig.4 1oo2 architecture

$$PFD_{avg\_1oo2} = 2\left[(1-\beta_D)\lambda_{DD} + (1-\beta)\lambda_{DU}\right]^2 t_{CE} t_{GE} + \beta_D \lambda_{DD} MTTR + \beta \lambda_{DU}\left(\frac{T_1}{2} + MTTR\right) \tag{14}$$

with:

$$t_{CE} = \frac{\lambda_{DU}}{\lambda_D}\left(\frac{T_1}{2} + MTTR\right) + \frac{\lambda_{DD}}{\lambda_D} MTTR \tag{15}$$

and

$$t_{GE} = \frac{\lambda_{DU}}{\lambda_D}\left(\frac{T_1}{3} + MTTR\right) + \frac{\lambda_{DD}}{\lambda_D} MTTR \tag{16}$$

And the PFH-Value is determined by:

$$PFH_{avg\_1oo2} = 2\left[(1-\beta_D)\lambda_{DD} + (1-\beta)\lambda_{DU}\right]^2 t_{CE} + \beta_D \lambda_{DD} + \beta \lambda_{DU} \tag{17}$$

The average time MTTF can be the time estimated between the occurrences of two errors. For this it can be very helpful to develop a Markov-model. Fig 5 shows a possible approach for the One-out-of-two systems 1oo2.

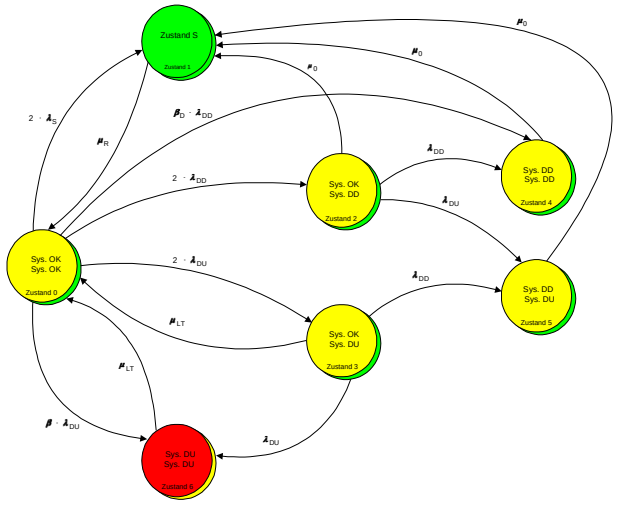


Fig.5 Markov-Chain for 1oo2 Systems

The Markov-Modell for a 1oo2 „Single-Board-System“ is shown in Fig. 5. In the condition 0 both controllers are working error-free. Condition 1 represents the safe condition in which a systems fades after a safe error. The system stands in a condition with no energy. In the condition number 2 one of two channels works incorrect.

The occurred error is dangerous, but is not detected through error diagnostics. Condition 4 is characterized by two dangerous traceable errors, with one of each of them being in one of the two channels. In condition 5, however, there is a dangerous traceable error in one channel, while at the same time there is a dangerous not traceable error occurring in the other channel. In condition 3 one the two channels operated incorrectly.

The occurring errors is dangerous and is not detected in the error analysis. In condition 3, when the error occurs in up until them the error-free channel, there is a fade of the system into the condition 5 or 6. If, however, there is no further error within the whole life span of the system in condition 3, the system may get back to the condition 0, where it is error-free. This practically means: After this the whole system will be exchanged.

If common-cause errors occur in 1oo2 systems, the following two cases are to be distinguished:

1) The joint error source leads to dangerous traceable errors. Then a fade occurs directly from system 0 to condition 4. The transmission rate is $\beta_D \cdot \lambda_{DD}$.
2) The joint error source leads to dangerous traceable errors. Then a fade occurs directly from system 0 to condition 6. The transmission rate is $\beta \cdot \lambda_{DU}$.

In the conditions 0,2 and 3 the system is running. This must be taken into account when calculating the MTTF of the 1oo2 system.

The probability matrix P for the 1oo2 approach is:

$$P_{1oo2} = \begin{bmatrix} P_1 & P_2 \\ P_3 & P_4 \end{bmatrix} \tag{18}$$

Where P_i

$$P_1 = \begin{bmatrix} 1-A_1 \cdot dt & 2 \cdot \lambda_S \cdot dt & 2 \cdot \lambda_{DD} \cdot dt & 2 \cdot \lambda_{DU} \cdot dt \\ \mu_R \cdot dt & 1-\mu_R \cdot dt & 0 & 0 \\ 0 & \mu_0 \cdot dt & 1-A_2 \cdot dt & 0 \end{bmatrix} 19)$$

$$P_2 = \begin{bmatrix} \beta_D \cdot \lambda_{DD} \cdot dt & 0 & \beta \cdot \lambda_{DU} \cdot dt \\ 0 & 0 & 0 \\ \lambda_{DD} \cdot dt & \lambda_{DU} \cdot dt & 0 \end{bmatrix} \tag{20}$$

$$P_3 = \begin{bmatrix} \mu_{LT} \cdot dt & 0 & 0 & 1-A_3 \cdot dt \\ 0 & \mu_0 \cdot dt & 0 & 0 \\ 0 & \mu_0 \cdot dt & 0 & 0 \\ \mu_{LT} \cdot dt & 0 & 0 & 0 \end{bmatrix} \tag{21}$$

$$P_4 = \begin{bmatrix} 0 & \lambda_{DD} \cdot dt & \lambda_{DU} \cdot dt \\ 1-\mu_0 \cdot dt & 0 & 0 \\ 0 & 1-\mu_0 \cdot dt & 0 \\ 0 & 0 & 1-\mu_{LT} \cdot dt \end{bmatrix} \tag{22}$$

From the probability matrix p the Q-matrix is determined. To form the Q-matrix from the P-Matrix one has to mind some criteria. The systems needs to running and the conditions may not be absorbing.

Furthermore it should be ensured that there is no secure condition or conditions showing dangerous untraceable errors. The absorbing conditions means the condition, where there is no further fade except the fade into a secure condition and an error-free condition.

After the Q matrix is formed, the M-Matrix is needed for further estimations. In order to calculate the M-Matrix, the Q-matrix needs to be subtracted from the I-Matrix (unit matrix).

$$I_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \tag{23}$$

$$M = I - Q \tag{24}$$

$$M_{1oo2} = \begin{bmatrix} A_1 \cdot dt & -2 \cdot \lambda_{DD} \cdot dt & -2 \cdot \lambda_{DU} \cdot dt \\ 0 & A_2 \cdot dt & 0 \\ \mu_{LT} \cdot dt & 0 & A_3 \cdot dt \end{bmatrix}_{\tau_{LT}=\infty} \tag{25}$$

if $\tau_{LT}=\infty$, then

$$M_{1oo2} = \begin{bmatrix} A_1 \cdot dt & -2 \cdot \lambda_{DD} \cdot dt & -2 \cdot \lambda_{DU} \cdot dt \\ 0 & A_2 \cdot dt & 0 \\ 0 & 0 & A_3 \cdot dt \end{bmatrix} \tag{26}$$

In order to calculate the MTTF-value, the elements of the first line need to be added to the N-Matrix. The N-Matrix is determined by the inverse of the M-Matrix.

$$N_{1oo2} = M_{1oo2}^{-1} \tag{27}$$

$$N_{1oo2} = \begin{bmatrix} \frac{1}{A_1} & \frac{2\cdot\lambda_{DD}}{A_1\cdot A_2} & \frac{2\cdot\lambda_{DU}}{A_1\cdot A_3} \\ 0 & \frac{1}{A_2} & 0 \\ 0 & 0 & \frac{1}{A_3} \end{bmatrix} \quad (28)$$

The MTTF -value can be determined by adding the elements from the first line to the N-Matrix.

$$MTTF_{1oo2} = \frac{1}{A_1} + \frac{2\cdot\lambda_{DD}}{A_1\cdot A_2} + \frac{2\cdot\lambda_{DU}}{A_1\cdot A_3} \quad (30)$$

Now the safety parameter MTTF can be estimated. With this value, now the reliability and default probability can be determined. The model parameters are of highest importance. In safety technology, certain distribution functions form the basis for the estimation.

In the following chapter an approach will be shown how to determine the model parameters using of estimation-algorithms.

III. Approach for the Construction of Estimators for the Safety Theory

The basis for estimating an unknown parameter ϖ is the assumption that a random variable X belongs to a certain parametric family (f.e. exponential distributed, normally distributed, poisson distribution, …). With the help of this model it is tried to determine this parameter for which the results are the most probable. This is done with events $X_1,...,X_n$ that have already taken place (sample $x_1,...,x_n$ with the values n from the scope n, variables are independent and identically distributed).

Hereby, the idea of the approach is the mapping of the mathematics on the safety technology. This approach is based on the necessary distribution function which is normally applied in safety technology. Only then a realistic statement about the probability functions of the reliability and density can be made.

The expected value $E_H(t)$ of a hardware component has been reached with the following formula:

$$E_H(t) = \lambda_0 \cdot \left(1 - e^{-\int_0^t \lambda(\zeta)\, d\zeta} \right)\Big|_{\lambda(\zeta)=\lambda} \quad (31)$$

where λ_0 is the maximum failure rate. This approach implies that from this point of time $t = 0$ (start of the reliability analysis), a constant failure rate λ_0 exists in the affected hardware system. As a Weibull-distribution has been deemed, the following applies:

$$E_H(t) = \lambda_0 \cdot \left[1 - e^{(-\lambda\cdot t)}\right] \quad (32)$$

If the hypothesis is true that the probability of default of a safety related system is exponentially distributed, the density function will be the following:

$$f(t) = \partial F(t) \Big/ \partial t = \partial\left(1 - e^{\lambda\cdot t}\right) \Big/ \partial t = \lambda \cdot e^{-\lambda\cdot t} \quad (33)$$

With the help of the equation 32, the time sequence of the failure rate $\lambda_H(t)$ can be determined

$$\lambda_H(t) = \frac{\partial E_H(t)}{\partial t} = \lambda_0 \cdot \lambda \cdot e^{(-\lambda\cdot t)}$$

In this connection the new percept is that the maximum failure rate λ_0 is divided into systematic and random hardware errors. This is why the equation needs to be changed into:

$$\lambda_H(t) = \frac{\partial E_H(t)}{\partial t} = \left(\lambda_{0_{SE}} + \lambda_{0_{RE}}\right) \cdot \lambda \cdot e^{(-\lambda\cdot t)}$$

If an optimized algorithm is applied to the new approach, then a forecasting for the hardware system in safety related applications - concerning the hardware error - can be made. This is done so that it can be predicted how many remaining errors $\lambda_{0_{RF}}$ as well as systematic errors $\lambda_{0_{SF}}$ can be found at certain point of time.

The estimations of the default rates λ and λ_0 are necessary. As we do not know of the distribution of the basic population (that is the probability function and the density function and as we have the result of a sampling procedure, we can now look for the parameters $\tilde{\lambda}$, $\tilde{\lambda}_0$, $\tilde{\lambda}_{0_{SF}}$ and $\tilde{\lambda}_{0_{RF}}$ for which the realization of the precise sample is most probable. This, of course, means nothing else than a maximation task. In calculating it the density function f (t) is needed from the equation 33.

$$\begin{aligned} \Delta(\lambda_0, \lambda) &= f(x_1,...,x_n,q) = \prod_{i=1}^{n} f(x_i,q) = L(\lambda_0,\lambda) \\ &= \left[1 - F(t_e)\right]^{\lambda_0 - m_e} \prod_{i=1}^{m_e} (\lambda_0 - i + 1) f(t_i) \end{aligned} \quad (36)$$

Whereby m_e stands for the amount of the overall errors at this point of time t_e. Here $f(t_i)$ is the failure density function and $F(t_e)$ the default probability function. If the

natural logarithm is taken from the equation 36, then the following applies:

$$\ln \Delta(\lambda_0, \lambda) = (\lambda_0 - m_e)\ln[1 - F(t_e)] + \sum_{i=1}^{m_e} \ln(\lambda_0 - i + 1) + \sum_{i=1}^{m_e} \ln f(t_i) \quad (37)$$

Now after that, the Δ - function is to be maximized. This is done with the following approach:

$$\frac{\partial \ln \Delta(\lambda_0, \lambda)}{\partial \lambda_0} = \ln[1 - F(t_e)] + \sum_{i=1}^{m_e} \frac{1}{\lambda_0 - i + 1} = 0 \quad (38)$$

$$\frac{\partial \ln \Delta(\lambda_0, \lambda)}{\partial \lambda} = -\frac{\lambda_0 - m_e}{1 - F(t_e)} \frac{\partial F(t_e)}{\partial \lambda} + \sum_{i=1}^{m_e} \frac{1}{f(t_i)} \frac{\partial f(t_i)}{\partial \lambda} = 0 \quad (39)$$

The partial derivations are replaced by the expected value $E_H(t)$ and the default rate $\lambda_H(t)$. The following applies:

$$E_H(t) = \lambda_0 \cdot F(t)$$
$$\lambda(t) = \lambda_0 \cdot f(t) \quad (40)$$

This leads to:

$$\frac{\partial \ln \Delta(\lambda_0, \lambda)}{\partial \lambda_0} = \ln\left[1 - \frac{E_H(t_e)}{\lambda_0}\right] + \sum_{i=1}^{m_e} \frac{1}{\lambda_0 - i + 1} = 0 \quad (41)$$

$$\frac{\partial \ln \Delta(\lambda_0, \lambda)}{\partial \lambda_0} = -\frac{\lambda_0 - m_e}{1 - \frac{E_H(t_e)}{\lambda_0}} \cdot \frac{\partial \frac{E_H(t_e)}{\lambda_0}}{\partial \lambda} + \sum_{i=1}^{m_e} \frac{1}{\frac{\lambda(t_i)}{\lambda_0}} \cdot \frac{\partial \frac{\lambda(t_i)}{\lambda_0}}{\partial \lambda} = 0 \quad (42), (43)$$

$$\frac{\partial \ln \Delta(\lambda_0, b)}{\partial \lambda} = -\lambda_0 \left[\frac{\lambda_0 - m_e}{\lambda_0 - {}^{s}\mu(t_e)}\right] \cdot \frac{1}{\lambda_0} \cdot \frac{\partial E_H(t_e)}{\partial \lambda} + \lambda_0 \sum_{i=1}^{m_e} \frac{1}{\lambda(t_i)} \cdot \frac{1}{\lambda_0} \cdot \frac{\partial \lambda(t_i)}{\partial \lambda} = 0 \quad (44)$$

$$\frac{\partial \ln \Delta(\lambda_0, \lambda)}{\partial b} = -\left[\frac{\lambda_0 - m_e}{\lambda_0 - E_H(t_e)}\right] \cdot \frac{\partial E_H(t_e)}{\partial \lambda_0} + \sum_{i=1}^{m_e} \frac{1}{\lambda(t_i)} \cdot \frac{\partial \lambda(t_i)}{\partial \lambda} = 0 \quad (45)$$

With the equations 44 and 45 the requested model parameters can be estimated. Therefore the expected value and the default rate may be inserted into the estimated equation and me be written for the summation "$\Sigma = \lambda_{0_{RF}} + \lambda_{0_{SF}}$". Hence, the result is:

$$\frac{\partial \ln \Delta((\Sigma), \lambda)}{\partial \Sigma} = \ln\left[1 - \frac{\Sigma \cdot \left(1 - e^{-(\lambda)t_e}\right)}{\Sigma}\right] + \sum_{i=1}^{m_e} \frac{1}{\lambda_0 - i + 1} = 0$$
$$\frac{\partial \ln L(\Sigma, \lambda)}{\partial \lambda} = -\left[\frac{\Sigma - m_e}{\Sigma - \Sigma \cdot \left(1 - e^{-\lambda t_e}\right)}\right] \cdot \frac{\partial\left(\Sigma \cdot \left(1 - e^{-\lambda t_e}\right)\right)}{\partial \lambda} + \sum_{i=1}^{m_e} \frac{1}{\lambda \cdot \Sigma \cdot e^{-\lambda t_e}} \cdot \frac{\partial\left(\Sigma \cdot e^{-\lambda t_e}\right)}{\partial \lambda} = 0. \quad (46)$$

Consequently. with the substitution of the summation the following estimated equations apply:

$$-\tilde{\lambda} t_e + \sum_{i=1}^{m_e} \frac{1}{\left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) - i + 1} = 0 \quad (47)$$

$$-\left[\frac{\left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) - m_e}{\left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) - \left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) \cdot \left(1 - e^{-(\lambda)t_e}\right)}\right] \cdot \frac{\partial\left(\left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) \cdot \left(1 - e^{-(\lambda)t_e}\right)\right)}{\partial \lambda} + \sum_{i=1}^{m_e} \frac{1}{\lambda \cdot \left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) \cdot e^{-\lambda t_e}} \cdot \frac{\partial\left(\lambda \cdot \left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) \cdot e^{-\lambda t_e}\right)}{\partial \lambda} = 0. \quad (48)$$

If the equation 47 and 48 are solved to the total rate $\tilde{\lambda}$, the equation will be demonstrated as the following:

$$\tilde{\lambda} = \frac{m_e}{\sum_{i=1}^{m_e} t_i + t_e \cdot \left(\left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right) - m_e\right)}. \quad (49)$$

In order to get to the estimated parameter $\left(\sum_{i=0}^{n} \tilde{\lambda}_{0_{RF}} + \sum_{i=0}^{n} \tilde{\lambda}_{0_{SF}}\right)$ the result of the equation 49 needs to be inserted into the estimated equation 47. Hence, the result is:

$$-\left(\frac{m_e}{\sum_{i=1}^{m_e} t_i + t_e \cdot \left(\left(\sum_{i=0}^{n} \tilde{u}_{0_{c_i}} + \sum_{i=0}^{n} \tilde{u}_{0_{nc_i}}\right) - m_e\right)}\right) \cdot t_e + \sum_{i=1}^{m_e} \frac{1}{\left(\sum_{i=0}^{n} \tilde{u}_{0_{c_i}} + \sum_{i=0}^{n} \tilde{u}_{0_{nc_i}}\right) - i + 1} = 0$$

With the help of the Siemens standard SN 295500 the default rates may be taken from the tables of the standard norm. The equations 49 and 50 may be solved with the help of the Cram´sche theory. Therefore both estimated parameters are determined with the Δ -function.

IV. Conclusion

In this paper a new approach has been set up in order to generate a better estimation of the safety parameters. Hereby, the focus was laid on the different default rates. When estimating the probabilities, traditional methods have ignored the differentiation into systematic and random hardware.

It is tremendously important for the safety technology that all error possibilities are taken into account through a stochastic model. Here, too, the work of this paper shows that it may be possible to insert distribution functions other than the exponential distribution.

With this new approach it is possible, too, to minimize or predict systematic hardware errors. Further, a realistic prediction about the probability of reliability as well as the probability of default can be made.

References

[1] Börcsök, Josef, “Functional safety systems”, Heidelberg: Hüthig, 2004.

[2] Börcsök, Josef, “Functional Safety Computer Architecture Part 1 and Part 2”, Kassel: lecture notes, Universit of Kassel, 2001.

[3] Goble, W. M., “Safety of programmable electronic systems – Critical Issues, Diagnostic and Common Cause Strength Proceedings of the IchemE Symposium”, Rugby: UK. Institution of Chemical Engineers, 1995.

[4] Health & Safety Executive (HSE) UK, “The setting of safety standards: A report by an interdepartmental group of external advisors”, London: HM stationery office, 1996.

[5] Health & Safety Executive (HSE) UK, “Programmable electronic systems in safety-related applications, part I”, London: HM stationery office, 1995.

[6] IEC 61508, “International Standard: 61508 Functional safety of electrical electronic programmable electronic safety related systems Part1-Part7”, *Geneva: International Electro technical commission*, 1999-2000.

[7] Lewis, E. E., “Introduction to reliability engineering”, 2nd ed.New York, John Wiley, 1996.

[8] Robert, C.P. & Casella G., “Monte Carlo, statistical methods”, Berlin: Springer Verlag, 1999.

[9] Storey, N., “Safety critical computer systems, Addison Wesley”, 1996.

[10] Velten-Philipp W. &. Houtermans M. J. M., “The effect of diagnostic and periodic testing on the reliability of safety systems”, Köln: TÜV, 2006.

[11] J. Börcsök, P. Holub, M.H.Schwarz, N.T. Dang Pham, “Determine PFD-values for Safety Related Systems Overview”, *ESREL*, Stavanger, 2007

Dr. –Ing. O. Krini Leader for Research and Development/ Functional safety, Bosch GmbH und ZF-Friedrichshafen AG and postdoctoral at the university of Kassel. Certified Safety-Manager for IEC 61508/ISO26262

Prof. Dr. -Ing. habil. J. Börcsök Certified safety-expert of functional safety and leader of the department for computer architecture and system programming at the University of Kassel, Germany.

Dipl. -Ing. A. Krini Ph.D. student (Research and Development) for Functional Safety and department for computer architecture and system programming at the University of Kassel, Germany

V2I-based Velocity Synchronization at Intersection

Xuguang Hao, Abdeljalil Abbas-Turki, Florent Perronnet and Rachid Bouyekhf

Abstract—Recently, some new methods have come into view for intersection management. They take the advantages of the potentials of new technologies that equipped in vehicles and in infrastructure, such as V2I, advanced cruise control, positioning and so on. One of these approaches is the communication between infrastructure and vehicles is used to synchronize vehicles' speed. With the synchronization, many simulations shows the potential of subduing many of current transportation problems. This paper firstly reviews some proposed protocols so as to introduce the synchronization of vehicles' velocity. The synchronization is based on the Sequence-Based Protocol (SBP). So the discussion mainly focus on it. Before proposing the approach, some practical and theoretical problems are highlighted for clarity. Because the problem is complex and need strong contributions on hybrid systems, in this paper the speed synchronization is based on first vehicle arrived first served.

Keywords—Advanced Cruise Control, V2I, Velocity Synchronization

I. Introduction

In modern cities, traditional solutions of transportation congestions, such as traffic lights and road planning, are encountering more and more different kinds of difficulties, especially in big cites. At present, many researchers focus on improving the urban intersection management with the latest technologies, for example wireless communication, positioning, advanced cruise control and so on.

With these informational technologies, people could be able to consider the transportation problems in different visions. The application of these modern tools to the planning and management of intersection is one of the core solutions of urban traffic congestion so as to improve the traffic capacity and efficiency in future. Some works are focusing on Cooperative Intersection Management (CIM) in order to control the passage of vehicles at urban intersection without traffic lights. They are based on the rapid progressing of vehicles equipped with on-board computer, wireless modules, sensors and so on. More precisely, the intelligent vehicle's function includes positioning, wireless communications between vehicles and infrastructure (V2I) as well as controllable motion. The intersection infrastructure and vehicle, as fundamental components of intersection management, all play important roles.

We introduce a new approach for synchronizing vehicles' velocity so as to safely and efficiently traverse the intersection.

This paper is organized as follows. Firstly, it gives an overview of last works on cooperative intersection management. Then, we present the components of Transparent Intersection Management (TIM) and the sequence policy that we adopt. Some important characteristics will be presented and will be used in the fourth section. In the fourth part, basing on sequence-based protocol, we propose a new approach for synchronizing the speed of vehicles. The results of simulation involving at an urban intersection will be the next part. Finally, we conclude on the approach advantages.

II. Literature Overview

A. Reservation-Based Protocol

Based on a multi-agent model, Dresner and Stone have presented the Reservation-Based Protocol in [2]. Vehicles and intersections are implemented as intelligent agents able to communicate together. When the vehicle approaching the intersection, it sends a request for the right-of-way that is kinetic parameters of the vehicle as well as its destination. Accordingly, the intersection manager simulates the journey of the vehicle in the gridded intersection map, in order to reserve space and time of the greeds. It reject the requests of others until the end of reservation. It shows that it is possible to make intersection control much more efficient than traditional control mechanisms.

In [3], the authors implemented a mixed reality platform with a real autonomous vehicle 'Marvin' which could interact with multiple virtual vehicles in a simulation at a real intersection. Its experiment shows that, with several techniques, the Autonomous Intersection Management (AIM) protocol outperform traffic signals. The test result of [4] shows that the protocol has potential to decrease vehicular delay. The more recent work [5] explored the possibility of applying autonomous vehicle auctions at each intersection to determine the order using autonomous reservation protocols with a microscopic simulator performing on city-scale maps.

The mixed reality platform [3] has shown that vehicles are not so controllable as it has been assumed. However, the collision avoidance is strongly dependent on the speed and on the time of traversing the intersection. Thus, there is a high collision risk if there is a tight timing between two vehicles whereas a high idle time between two vehicles will significantly compromises the traffic efficiency without entirely eliminate collision risk.

B. Intersection-based Cooperative Adaptive Cruise Control Protocol

Another team has proposed a heuristic optimization algorithm for controlling the automated vehicles at traditional intersections with a game theory framework entitled CACC-CG [6], [7]. The framework is considered as a decision process that repeats at each time step of simulation to optimize the movement of automated vehicles. The protocol controls trajectories of vehicles which are equipped with Cooperative Adaptive Cruise Control (CACC) to avoid collisions and minimize vehicle's delay and consequently reduce the total delay of intersection.

ISBN: 978-1-61804-256-9

In [7], the authors clearly proposed the Intersection-based Cooperative Adaptive Cruise Control Protocol (iCACC). In order to optimize the movement of autonomous vehicles, three zones are assumed that they fall in the vehicle trajectory. The ”smart” intersection takes into account the physical characteristics that may affect the motion of vehicles to simulate and to optimize the speed of the vehicles. As being fully equipped, the vehicle must adapts itself in Zone II in order to control the point of time that it arrives at the conflict zone. The speed adaptation makes sure the vehicle will reach its maximum velocity when it cross the Conflict Point. And consequently, the vehicle will pass the intersection box at the maximum speed. In iCACC protocol, the fundamental is that, basing on gridded intersection zone, the manager simulates and assumes the journey of vehicle based on the current situation and make a precise reservation that the intelligent vehicle must obey.

As for the reservation protocol, it is hard to guarantee that the vehicle will traverse the intersection at exactly the mentioned high speed. Moreover, the intersection sends messages for slowing down vehicles. However, because of message drop and loss there is a high collision risk.

C. Sequence-Based Protocol

The intersection and vehicle all play important roles in decision process of intersection management. At this point, there are a lot of works could be done to improve current traffic efficiency. In order to define roles in a better way, the Sequence-Based Protocol (SBP) was proposed [1], [8], [9]. As named as Sequence-Based Protocol, one fundamental of this protocol is sequencing all the ”full-automated” vehicles that are waiting to pass the intersection.

In ”this” centralized protocol, basing on the informations that collected from these vehicles, the intersection manager could apply optimization methods to form the passing sequence of vehicles. The sequence means deciding explicitly which vehicle will traverse the intersection first, which is the second vehicle and so on. There is no conditions on times and speeds. Hence, the protocol can be applied for automated vehicles or manned vehicles. For safety reasons, the vehicle cannot traverse the intersection without a consent from the intersection manager. Hence, a vehicle that has not received a message from the intersection automatically stop before the junction box that we will call later conflict zone. Hence, the principle of the default deny is chosen.

For manned vehicles, the intersection manager assigns only ”right-of-way” to vehicles. The right-of-way is a green that allows a vehicle to safely traverse the intersection. The right-of-way can be distributed to several vehicles simultaneously if there is no conflict. Hence the intersection manager permanently broadcasts message with a list of vehicles that have the right-of-way. The vehicle checks if it is in the list.

The unmanned vehicles considers the ”right-of-way” more complex than a simple green. As in the case of manned vehicles, the intersection manager sends a list of allowed vehicles with their movement parameters that are speed, position, movement direction, destination, etc. The list of vehicles is ordered according to the decided sequence. The main difference between the intersections of manned and of unmanned vehicles is that two unmanned vehicles with conflicting movements are included in the same list. This means that both one must synchronizes its speed to do not collide with the other. In general case, an unmanned vehicles should observe all precedent vehicles in the list to avoid collision. So the main raised issue by the sequence based protocol is how to synchronize the speed between all unmanned vehicles. This issue is treated as the core of this paper.

Currently, there are two approaches for synchronizing speeds. The first one assumes that the vehicle immediately slow down until it gets enough space to traverse safely the intersection. The second one assumes that the vehicle slows down to completely stop before the conflict zone but if there is enough space to traverse safely the intersection during the slowing down, the vehicle speed up and traverse the intersection. The main issue with both approaches is that vehicles slow down near the extremity of each side of the lane. This can cause congestions at near intersections if the traffic flow is high. Moreover, both approaches have been developed, simulated and tested for only mini-robots. Since test of both approaches have shown that they are safe and they allow a good performance, then they deserve to be improved in terms of lane occupancy.

III. STUDIED INTERSECTION

Section II-C has given a simple description of the two existing approach for synchronizing speed of vehicles that listed in one sequence. To express the work of this paper more clearly, Fig. 1 shows the two approaches controlling the vehicle coloured yellow .

Controlling by using the first approach, the vehicle will immediately slow down at the beginning of lane to get enough space. Its lower speed even stop will cause the access of the lane that it just has entered be locked. This occupancy of the access of lane will consequently influence the motion of vehicle listed in the sequence but behind it and will also run into the same lane. The sequence is the one maintained by the left one in Fig. 1. The follower will slow down or stop before the conflict zone of the same intersection. As all the vehicles are maintained in one sequence, slowing down of one vehicle may leads to a high congestion in high flow, just as shown in Fig. 1. Another possible consequence of first approach is the lower usage of lane.

The second approach doesn’t have the usage problem, but will influence the fluidness of its downstream intersection shown as right part of Fig. 1. Because of sequence of vehicles, the stop of yellow vehicle before stop line will also finally lead to the stop of the vehicles that listed in the sequence but behind it. Then the congestion will occurs in high traffic flow. Another problem of this approach comes from the speeding up of yellow vehicle when it is traversing the conflict zone. Especially in high flow, the low speed may lead to the slowing down of the vehicles which running on the other lanes and preparing to traverse the same conflict zone.

The most basic reasons of issue of the two approaches is the juncture and strategy of speed adjustment. At extreme

ISBN: 978-1-61804-256-9

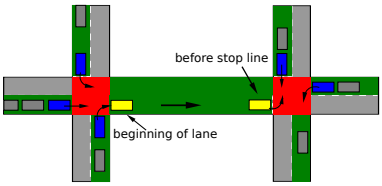


Fig. 1: Current two approaches for synchronizing speeds

situation, the stop of yellow vehicle could not be avoided. So in order to improve the performance of the two approaches, this paper focus on when and where to slow down the vehicle. The target becomes releasing the occupancy of the access of lane to increase the usage of lane and getting velocity as high as possible when the vehicle traversing the conflict zone. This will also have influence as low as possible to the vehicles which are running on other lanes and preparing to pass the same conflict zone.

A. Structure of Intersection

This paper focus on the intersection that the vehicle we will try to control is approaching. The intersection is a 4-leg one which has 4 input lanes and 4 output lanes as shown in Fig. 2. The protocol of intersection management is the Transparent Intersection Management (TIM) [9]. It uses client/server architecture to organize the management of intersection.

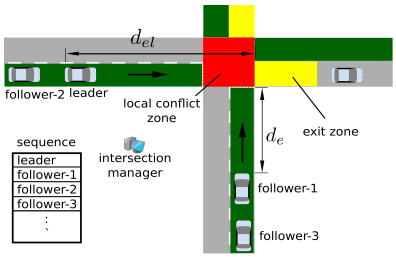


Fig. 2: Structure of Intersection

The necessary components of TIM are as follows:

- Intersection Manger (server) maintains the sequence of vehicles and could optimize it as needed.
- The autonomous vehicle (client) could communicate with the intersection manager. It is equipped with the new control protocol.
- The green ones are the buffer zone. All clients moving on them will communicate with server for negotiating its right-of-way and exchanging other informations which are necessary in the management process.
- The red zone is conflict zone. The manager will make sure that, with right-of-way, at any time, all the vehicles running in it have no conflict of movement route with each other.
- The yellow region is called exit zone. The vehicle that has entered this region will report its exit of conflict zone to manager as quickly as possible.

In the following, we consider that the sampling time of the vehicle is bounded to τ_v. We assume that communication (V2I) is not perfect. That means message drop and loss are possible. Nevertheless, we assume that there exist a stable value of roundtrip time, that is RTT_{v2i}. The positioning system is assumed precise because it is a prerequisite of autonomous vehicle.

B. Obstacles

In TIM there are two kinds of factors will influence the motion of vehicle. The measurement of frontal sensor will be used to avoid collision with precedent vehicle. The information received from manager will be used to adjust the motion in order to achieve the goal of the new protocol. Corresponding to the two kinds of factors, we classify the objects that will influence the motion into Real Obstacle and Virtual Obstacles.

1) Virtual Obstacles: The Virtual Obstacles (VOs) are the objects that do not have direct influence on the motion of vehicle in the new protocol. For example, as listed in the 'sequence' shown in Fig. 2, vehicle 'leader' and 'follower-1' running on different lanes. As to 'follower-1', it does not need to immediately react to the situation of 'leader'. This characteristic plays a fundamental role in this paper. It enable us to do more works than with the real preceding vehicle. This protocol take into account two types of virtual obstacles.

The first type is the Virtual Preceding Vehicle (VPV) that listed in the sequence. Normally, virtual preceding vehicle has higher priority to pass the intersection than its follower. As it is running on other lane, the follower doesn't need to react immediately to its motions, even if the follower is closer to the conflict zone than its virtual preceding. In other words, the follower could smoothly adjust its motion as it desires before it enters the conflict zone. It just needs to make sure that when entered the conflict zone with right-of-way, it has got desired motion. This characteristic is interesting, because we could take the best advantage of it to do some thing we would like.

The other type of virtual obstacle is the conflict zone shown in Fig. 2, the red area. The conflict zone plays a very important role in this paper for the adjustment of motion of vehicle. It's the safe stone that the vehicle will stop before it if the vehicle has not got the right-of-way but has reached the beginning of conflict zone. It has been introduced in one of the control policies of [8]. This paper considers it as a virtual preceding obstacle that has no speed and stay at its position for ever.

2) Real Obstacle: The Real Obstacle is the real preceding vehicle that the traditional cruise control only takes into account. It also be listed in sequence maintained by manager. So the new policy needs to distinguish whether its virtual preceding vehicle is also the real preceding one. Its motion information will comes from the frontal equipped sensors.

Corresponding to the two kinds of obstacles, we could get two accelerations a_r and a_v respectively come from dealing with real obstacle and virtual obstacles. As the safety is the primary standard, the final acceleration a that will be taken by follower is given by:

$$a = \min(a_r,\ a_v) \tag{1}$$

C. Sequence policy

There exist many policies for sequencing the incoming vehicles, for example, First Come First Served (FCFS), Distributed

ISBN: 978-1-61804-256-9

Clearing Policy (DCP) and Autonomous Distributed Clearing Policy (ADCP) [8] etc. As the adjustment of vehicle's motion is concerned in this paper, FCFS is chosen as sequence policy for simplicity.

IV. THE SEQUENCE-BASED COOPERATIVE ADAPTIVE CRUISE CONTROL POLICY

During more than half century, a lot of attentions have been paid on finding some methods that could perfectly reflect driver's behaviours in transportation. Some of their works are Intelligent Driver Model (IDM), enhanced IDM (EIDM) and a special model based on IDM for TIM which named as Cooperative Intelligent Driver Model (CIDM) [9].

In this section, a new Advanced Cruise Control (ACC) policy for real preceding vehicle is firstly proposed. We name it as ExACC policy. We also use it to treat the two kinds of virtual obstacles at same time for safety. After that a simple strategy is introduced for adjusting the speed of vehicle with maximal capacity for traversing the intersection with higher speed.

For clarity, some general symbols are shown in Table I.

TABLE I: Parameters of Advanced Cruise Control

Parameter	Meaning	Value / Unit
v_0	Desired speed of vehicle	15 ms^{-1}
s_0	Minimum distance headway	$2m$
a	Maximum acceleration	4 ms^{-2}
b	Minimum deceleration	-4 ms^{-2}
τ	Sampling time	s
s	Distance headway without s_0	m
v	Speed of vehicle	ms^{-1}

A. Advanced Cruise Control Policy for Real Obstacle

Generally, one of the most important characteristics of traditional adaptive cruise control is to react as quickly as possible to abrupt brake of real preceding vehicle for safety. After getting enough space, it will speed up with an acceptable and comfortable acceleration to achieve an equilibrium situation.

Normally, ACC make action decision for next sampling time basing on the situation of current point of time. They could not give any information about future action of leader vehicle. With this characteristic, they normally could not immediately react to extreme case, for example the leader brakes at maximal deceleration which is bigger too much than that of follower vehicle. So if some assumptions of extreme situation of leader could be introduced into the decision process, the follower may be able to react better so as to improve safety.

We name the the policy that we will introduce as ExACC policy. It originally introduces a prediction of leader's motion into decision process. The prediction assumes that the leader will brake at its maximal capacity until stop from the beginning of next sampling time. Follower will take an acceleration during the next reaction step. And as reaction to leader, the follower will also take its maximal capacity, from the end of next reaction step, to try to stop and make sure that the final distance headway is greater than or is equal to s_0. In other words, follower will react to the assumption with a delay, one sampling time.

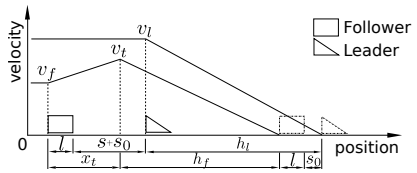


Fig. 3: Strategy of ExACC function

Fig. 3 shows the strategy of ExACC policy. v_l, v_f, v_t are respectively the velocity of current point of time of leader and follower, the velocity of follower at the end of next τ_v. l the length of follower. x_t is the distance the follower will move in the next τ_v. h_l, h_f respectively the movement distance during the brake of leader and follower.

The above figure clearly shows the following relations:

$$v_t = v_f + a_r\tau \tag{2a}$$

$$x_t = \frac{\tau}{2}(v_f + v_t) \tag{2b}$$

$$h_l = -\frac{v_l^2}{2b_l} \tag{2c}$$

$$h_f = -\frac{v_t^2}{2b_f} \tag{2d}$$

$$h = (l + s + s_0 + h_l) - (x_t + h_f + l + s_0) = 0 \tag{2e}$$

where, b_l, b_f are respectively the maximal deceleration of leader and follower, a_r the acceleration that follower will take during the next step τ_v. We draw the reader attention to the fact that for safety reasons, τ is bigger than τ_v and RTT_{v2i}. In the following we consider that $\tau = 1,2\max(\tau_v, RTT_{v2i})$.

Then we have the acceleration, basing on equation (2), that the follower should take during the next τ_v

$$a_r = \frac{b_f\tau - 2v_f \pm 2b_f\sqrt{\frac{b_f b_l\tau^2 + 4b_l v_f\tau + 4v_l^2 - 8b_l s}{4b_f b_l}}}{2\tau} \tag{3}$$

Fig. 3 shows the case that the follower is behind the leader at initial state. Actually, in the scenario that the follower is running before the leader, if we also define the headway $s+s_0$ is from rear bumper of leader to front of follower, we could also have the same result as well as function (3). This scenario could occurs if the leader is a virtual preceding vehicle of follower as shown in Fig. 2.

As we should take into account the extreme situation that the bumper to bumper distance is less than s_0, the final ExACC policy is:

$$a_r(v_f, v_l, s) = \begin{cases} b_f, & \text{for } s < 0 \quad (4a) \\ \dfrac{b_f\tau - 2v_f - 2b_f a^*}{2\tau}, & \text{for } s \geq 0 \quad (4b) \end{cases}$$

where,

$$a^*(v_f, v_l, s) = \sqrt{\frac{b_f b_l\tau^2 + 4b_l v_f\tau + 4v_l^2 - 8b_l s}{4b_f b_l}}$$

The following Fig. 4 shows a extreme scenario that the leader brakes frequently with maximal capacity -15 ms^{-2}. The acceleration capacity of follower is $[-4, +4]ms^{-2}$.

It shows that the follower equipped with ExACC policy could react immediately to abrupt brake of leader without any collision. We take this policy for dealing with the real preceding vehicle in order to get the acceleration a_r in equation (1).

ISBN: 978-1-61804-256-9

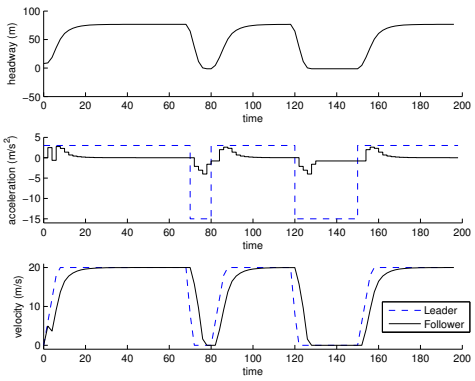


Fig. 4: ACC function: Scenario multi brake

B. Strategy for Virtual Obstacles

As discussed in section III, the most important characteristic of treating virtual obstacles is that the follower only need to make sure that when passing the stop line with right-of-way it has a higher speed for traverse the conflict zone as quickly as possible. So the point of time and position that follower start to adapt its velocity are the key points of the strategy. With the capacity of reacting to extreme situation, we also apply the policy (4), with a adaptation, on dealing with the two types of virtual obstacles.

Table II shows the symbols that will be used in following strategy for virtual obstacles. Some of the symbols could also be found in Fig. 2. Same with previous policy for real preceding vehicle, the headway doesn't include s_0.

TABLE II: Parameters of ACC function for Virtual Obstacles

Parameter	Meaning	Unit
d_{el}	Escape distance of leader	m
d_e	Escape distance of follower	m
s_e	Equilibrium distance headway	m
t_e	Escape time of follower	s
v_t	Target velocity of follower	ms^{-1}

In order to deal with the two types of virtual obstacles, we established a control strategy, function (5). It bases on function (4) and considers the conflict zone as another type of virtual obstacle which has same parameters with virtual preceding vehicle except the velocity and position. As discussed before, this obstacle has speed $0ms^{-1}$ and stay at its position for ever.

$$a_{v1}(v_f, v_l, s_v, s_c) = \begin{cases} a_c, & \text{for } s_v < 0 \text{ or } a_p \leq a_c \quad (5a) \\ a_p, & \text{for other cases} \quad (5b) \end{cases}$$

where,

$$a_c = a_r(v_f, 0, s_c)$$
$$a_p = a_r(v_f, v_l, s_v)$$

s_c, s_v respectively the distance from beginning of local conflict zone or virtual preceding vehicle to follower.

Fig. 5 shows a normal simple scenario with two vehicles running on different lanes at an intersection. They have same initial velocity ($15ms^{-1}$) and same distance ($150m$) to the conflict zone. The virtual preceding leader runs at constant velocity while the follower is controlled by policy (4). The dash point line indicates the conflict zone.

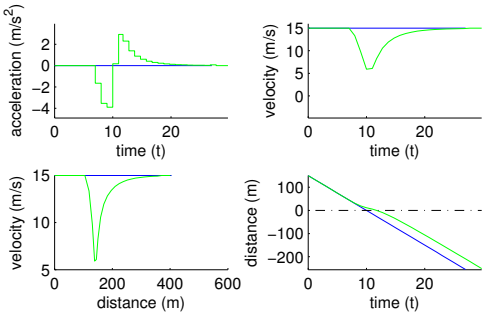


Fig. 5: Application of ExACC on Virtual Obstacles

In above scenario, at the beginning of adjustment, the vehicle brakes sharply. It's not a good experience. In TIM, if the follower does not obtain the right-of-way, it needs to stop before conflict zone. In addition, at some extreme cases, the follower needs to brake sharply even stop for avoiding collision with virtual preceding. One way of getting a smoother motion of vehicle and avoiding sudden velocity adjustment is to make a light deceleration when approaching the stop line.

C. Smoothing Strategy (SS)

For the sake of smoothness, we introduce a simple velocity adjustment strategy. It takes into account a specified distance that from vehicle to conflict zone. During the whole speed adjustment, it will takes a constant deceleration and a constant acceleration. The absolute value of the two acceleration is same. When arrived at the stop line, the vehicle should get the desired velocity v_t, shown in Fig. 6.

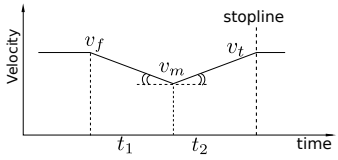


Fig. 6: Smoothing Strategy

Where, v_m, t_1 and t_2 are respectively the minimal speed during the adjustment, the brake time and the speeding up time. Then we have

$$v_m = v_f - at_1^* \quad (6a)$$
$$v_t = v_m + at_2^* \quad (6b)$$
$$t_e = t_1^* + t_2^* \quad (6c)$$
$$(v_f + v_m)t_1^* + (v_m + v_t)t_2^* = 2d_e \quad (6d)$$

For simplicity, we assume that the virtual leader will run at current velocity during its escape journey d_{el} as shown as Fig. 2. So we could have the escape time of leader $t_{el} = \frac{d_{el}}{v_l}$. That's also the escape time t_e that follower will cost when it arrives at conflict zone.

Because that normally the follower will not stop and has a sampling time τ_v, the actual brake time should be divisible by τ_v. With same reason the acceleration time should also be divisible by τ_v, if the target velocity is not the maximal speed of follower. Then we finally have adaptation relation:

$$t_1 = \lceil \frac{t_1^*}{\tau_v} \rceil \tau_v, \; v_m = v_f - bt_1$$

ISBN: 978-1-61804-256-9

and

$$t_2 = \lceil \frac{t_e - t_1}{\tau_v} \rceil \tau_v, \ a = \frac{v_t - v_m}{t_2}$$

Then it is easy to get the acceleration a_{v2} of vehicle with smoothing strategy. Fig. 7 shows the simulation in same scenario as Fig. 5.

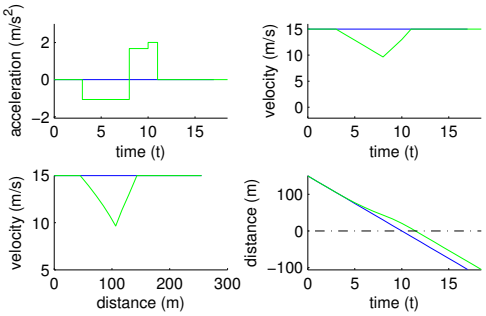


Fig. 7: Simple Scenario of Smoothing Strategy

Combining function (5) and Smoothing Strategy we get the final strategy for dealing with virtual obstacles:

$$a_v = \min(a_{v1}, \ a_{v2}) \tag{7}$$

V. SIMULATION

We make final simulation under MATLAB to test the new policy (1) at the intersection Fig. 2.

Initial escape distance of vehicles from beginning of local conflict zone is $150m$. Initial speed is $15ms^{-1}$. The sampling time of vehicle is $1s$. Every two vehicles are generated at same time but are located at two different lanes. The delay between two successive generations of vehicles is $2s$. The sequence of vehicles is shown as 'sequence' in Fig. 2.

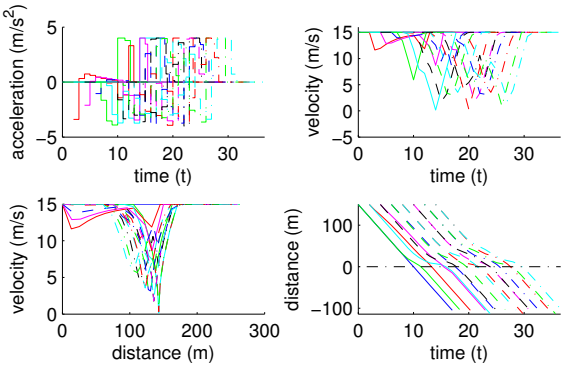


Fig. 8: Combination of function (3) and (5)

Fig. 8 shows that the new approach has successfully synchronized the vehicles' speed before the vehicles enter the conflict zone. And all the vehicles reach their minimal speed during the velocity adjustment. This could help to improve the usage of lane. But the vehicles takes abrupt brake at the beginning of adjustment. After combining smoothing strategy, Fig. 9 shows that the vehicles make a smoother adjustment and get a higher speed when they enter the conflict zone. It is helpful for the efficiency of intersection.

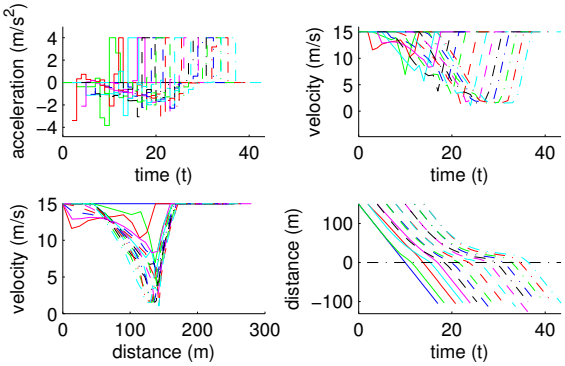


Fig. 9: Combination of the three strategies

VI. CONCLUSIONS

In this paper we have introduced a new approach for controlling the autonomous vehicles that are waiting traverse intersection. The first advantage of this approach is synchronizing the speed of vehicle before it enters the conflict zone soon. This method also helps to improve the efficiency of intersection with a higher speed when they traverse the conflict zone. In addition, because the vehicle adjust its speed when they approach the conflict zone, this strategy could help to improve the use of lane, especially in the case that the vehicle needs to stop during its speed adjustment. With the help of smoothing strategy, the adjustment could be done smoothly. This will be able to have a better ride experience.

REFERENCES

[1] F. Perronnet, A. Abbas-Turki, J. Buisson, A. El Moudni, R. Zeo and M. Ahmane, Cooperative intersection management : Real implementation and feasibility study of a sequence based p rotocol for urban applications. Intelligent Transportation Systems (ITSC), 2012 15 th International IEEE Conference on, pp. 42-47, September 2012.

[2] K. Dresner and P. Stone, Multiagent traffic management: A reservation-based intersection control mechanism, The Third International Joint Conference on Autonomous Agents and Multiagent Systems, pp. 530-537, July 2004.

[3] M. Quinlan, A. Tsz-Chiu, J. Zhu, N. Stiurca and P. Stone, Bringing simulation to life: A mixed reality autonomous intersection, IROS, IEEE/RSJ International Conference on, 2010.

[4] C-L. Fok, M. Hanna, S. Gee, T-C. Au, P. Stone, C. Julien, and S. Vishwanath, A platform for evaluating autonomous intersection management policies, In ACM/IEEE Int. Conf. ICCPS 2012, 2012.

[5] D. Carlino, S. D. Boyles and P. Stone. Auction-based autonomous intersection management. In Proceedings of the 16th IEEE Intelligent Transportation Systems Conference (ITSC), October 2013.

[6] I. H. Zohdy and H. Rakha, Game theory algorithm for intersection-based cooperative adaptive cruise control (CACC) systems. Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on, pp. 1097-1102, September 2012.

[7] I. H. Zohdy, R. K. Kamalanathsharma and H. Rakha, Intersection management for autonomous vehicles using iCACC. Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on, pp. 1109-1114, September 2012.

[8] F. Perronnet, A. Abbas-Turki, A. El-Moudni, J. Buisson and R. Zo, Cooperative Vehicle-Actuator System: A sequence-based optimal solution algorithm as tool for evaluating policies. Advanced Logistics and Transport (ICALT), 2013 International Conference on, pp. 19-24, May 2013.

[9] F. Perronnet, A. Abbas-Turki and A. El Moudni, A Sequenced-Based Protocol to Manage Autonomous Vehicles at Isolated Intersections. Intelligent Transportation Systems - (ITSC), 2013 16th International IEEE Conference on, pp. 1811-1816, Oct. 2013.

ISBN: 978-1-61804-256-9

Designing a Bayer Filter with Smooth Hue Transition Interpolation Using the Xilinx System Generator

Zhiqiang Li, Peter Z. Revesz

***Abstract*—**This paper describes the design of a Bayer filter with smooth hue transition using the System Generator for DSP. We describe and compare experimentally two different designs, one based on a MATLAB implementation and the other based on a modification of the Bayer filter using bilinear interpolation.

***Keywords*—**Bayer array, Demosaicing, FPGA, Interpolation.

I. Introduction

Digital cameras perform a sequence of complicated processing steps while recording color images. A color image usually contains three different color components in each pixel: red (R), green (G) and blue (B). Digital cameras use three separate sensors to capture these three components [1]. In order to reduce the cost, digital cameras capture images using a sensor overlaid with a color filter array (CFA). CFAs allow only one color component for each pixel, which means we need to generate the full color images from the output of the image sensor [2].

Bayer color filter arrays (Bayer CFAs) are currently one of the most common CFAs in digital cameras and can be used together with many different interpolation methods [3, 4]. The *System Generator for DSP,* commonly referred to as just *System Generator* [5, 6], is a MATLAB/Simulink-based simulation tool from Xilinx Inc. [7]. The System Generator is a hardware design package that allows programming on the FPGA and modeling a system using Simulink. The System Generator contains many modules, such as FIR filter, FFT, FIFO, RAM and ROM.

II. The Bayer Color Filter Array

Bayer CFAs greatly reduce the complexity and the cost of digital cameras. Each Bayer CFA contains twice as many green elements than red or blue ones, reflecting the fact that the cone cells in the human retina are most sensitive to green light. The full color image contains of three components (R, G and B) in each pixel, but a Bayer image, which is the output of a Bayer CFA, contains only one component in each pixel. However, from a Bayer image a full color image is generated by *demosaicing* [8], that is, an interpolation that estimates the values of the missing components [9]. For *demosaicing*, Xilinx uses *bilinear* interpolation, which performs the following three steps:

1. Estimate the missing green values in the red and blue pixels by using their four green neighbors. For example, using the Bayer image in Fig 1., the bilinear interpolation finds:

$$G8 = (G3 + G7 + G9 + G13)/4$$
$$G14 = (G9 + G19 + G13 + G15)/4 \quad (1)$$

2. Estimate the missing red or blue values in the green pixels:

$$B7 = (B6 + B8)/2$$
$$R7 = (R2 + R12)/2 \quad (2)$$

3. Estimate the missing red value of the blue pixel and the missing blue values of the red pixels:

$$R8 = (R2 + R4 + R12 + R14)/4$$
$$B12 = (B6 + B8 + B16 + B18)/4 \quad (3)$$

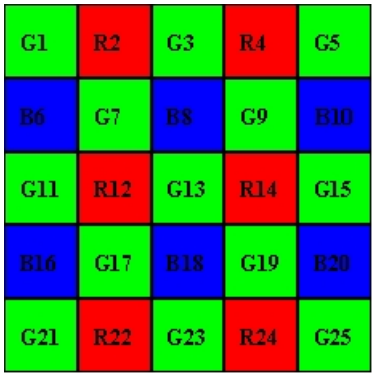


Fig. 1 A Bayer color filter array

Instead of the *bilinear* interpolation, this paper uses *smooth hue transition* interpolation [10]. Before estimating the missing red and blue values, we first estimate the missing green values of the red or blue pixels, the same way as in step (1) of the bilinear interpolation. Let the blue hue be B/G and the red hue R/G. These are used the estimate the missing blue

Zhiqiang Li is with the Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588, USA (zli@cse.unl.edu).

Peter Z. Revesz is with the Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588, USA (revesz@cse.unl.edu).

ISBN: 978-1-61804-256-9

element of the green pixels by:

$$B7 = (G7/2) \times (B6/G6 + B8/G8)$$
$$B13 = (G13/2)(B8/G8 + B18/G18) \quad (4)$$

Then the missing red values of the green pixels are estimated by:

$$R13 = (G13/2) \times (R12/G12 + R14/G14)$$
$$R7 = (G7/2) \times (R2/G2 + R12/G12) \quad (5)$$

The missing red values of the blue pixels are estimated by:

$$R8 = (G8/4) \times (R2/G2 + R4/G4 + R12/G12 + R14/G14) \quad (6)$$

Finally, the missing blue values of the red pixels are estimated by:

$$B12 = (G12/4) \times (B6/G6 + B8/G8 + B16/G16 + B18/G18) \quad (7)$$

III. Development Platforms

Xilinx is a supplier of programmable logic devices. It is famous for inventing the field programmable gate array (FPGA). Xilinx Spartan®-3A DSP [11] FPGA video starter kit (VSK) is a development platform consisting of the Spartan-3A DSP 3400A development platform, the FMC-video daughter card and a VGA camera. This platform enables experimenting with video processing using the Spartan-3A DSP family of FPGAs. VSK also includes a variety of software components, which are the Xilinx ISE® Design Suite 11.1 (includes as well as full versions of EDK and System Generator).

System Generator is a design tool that enables us to use the Mathworks model-based design environment Simulink for FPGA design. Developers do not need to have experience with FPGAs or RTL design when using System Generator. The Simulink modeling environment with a Xilinx specific blockset is used to complete the design. The downstream FPGA implementation steps are automatically performed to generate an FPGA programming file. Over 90 DSP blocks are provided in the Xilinx DSP blockset for Simulink. Common blocks such as adders, multipliers and registers are included. In addition, some complex building blocks, such as FFTs, filters and memories are also provided. The System Generator is based on Simulink from MATLAB.

IV. Implementation in MATLAB of the Smooth Hue Transition Interpolation

We use MATLAB to implement the smooth hue transition interpolation. First, the Bayer image is captured using DH-SV1410, which is widely used in industry. The following algorithm assumes that the size of the input data is a 1040 by 1392 Bayer image. We apply the smooth hue transition interpolation algorithm to the Bayer image to reconstruct the full color image. The function

result_g = shtlin_g_rg (Bayer)

implements Step (1), where result_g stands for the green values. The function

result_r = shtlin_r_rg (Bayer, result_g)

implements Steps (5) and (6) where result_r stands for the red values. The function

result_b = shtlin_b_rg (Bayer, result_g)

implements Steps (4) and (7) where result_b stands for the blue elements. The red, green and blue components constitute the interpolated 1040 by 1392 full color image. Fig. 2 shows an example of a smooth hue transition interpolation.

Fig. 2 Smooth hue transition by MATLAB

V. Reference Design from Xilinx

A. Overall design of the camera

Xilinx provides a reference Bayer filter design using bilinear interpolation, which can be described as follows. One big block called "vsk_camera_vop" is actually the design of a camera. There are three inputs and six outputs. "vsync", "hsync" are the vertical and horizontal synchronization signal of the oscilloscope. "BayerRaw_Raw" contains the one-dimensional Bayer image. "vs_out", "hs_out" are output synchronization signals. "red_out", "green_out" and "blue_out" stand for the color output information. "de_out" is the output enable signal. All of the output signals are connected to the oscilloscope block. The camera pipeline consists of several functional blocks:

- "dyn_range_exp", the dynamic range expansion [12] module, in photography, dynamic range describes the ratio between the maximum and the minimum light intensities. The higher the dynamic range is, the better the image is.
- "spc", the stuck pixel correction [13] module. Some of the pixels cannot be displayed correctly,

we need to correct this.

- "bright_contrast", we also need to control the brightness and contrast of the image.
- "bayer_filter", this is the module we need to focus on, using the bilinear interpolation to complete the reconstruction.
- "color_balance", this module is used to adjust the overall intensity of pixels, making the image look better.
- "stats", this module is used to calculate the maximum and minimum value of the pixels.

B. Bayer filter design

Since the details of the Bayer filter design (see Fig. 3) are complicated, we explain here only the bilinear algorithm using as an example Fig. 1. The block "Delay9" makes sure that the data is synchronized with the vertical and horizontal signals. Numbers from 1 to 9 stand for the location of pixels in the circuit. When the data reach location A, because of the block "delay 7", the data cannot reach location 1 until two clocks later. At the same time, data can reach the block "Single Port RAM", and is written into the RAM according to the address provided. The "Single Port RAM" is set to the mode "Read before write", which means one clock later, the data at location B is the initial value 0, not the data value stored. During the next clock, the address is incremented by 1, and new data is stored into the "Single Port RAM". Since the address is added at this time, the data at location B is still the initial value 0 at the address. From location 4, we can get the value 0 from the last clock. According to the description above, before the pixels in the first row (G1, R2, G3, R4, and G5) reach location A, values from location 4, 5, 6, 7, 8, 9 are all 0.

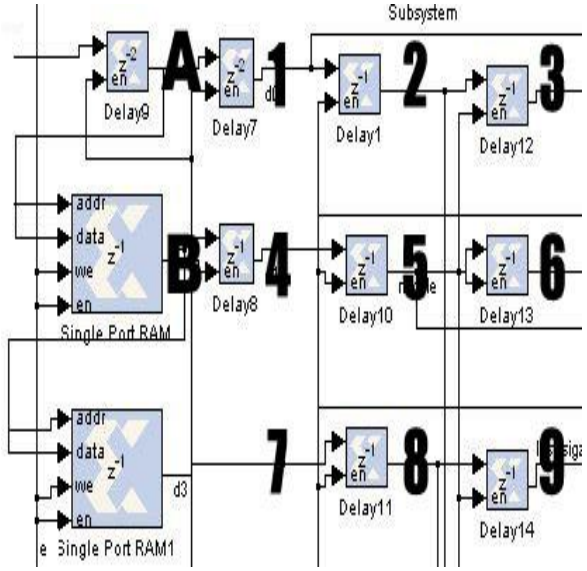


Fig. 3 Part of the design by bilinear

When the pixels in the second row (B6, G7, B8, G9, and B10) start to reach location A, because of the control from horizontal synchronization signal, the address is reset to 0. Similarly to the previous description, "Single Port RAM" is in the "Read before write" mode, the data we get from it is not the data at location A (the data of the second row), it should be the data from the first row (the data is G1). After one clock delay, the G1 appears at location B. Similarly, after getting the initial value 0 from "Single Port RAM1", G1 is stored into it. After all the pixels of the second row (B6, G7, B8, G9, B10) reach location A, the pixels of the first row (G1, R2, G3, R4, G5) are stored into "Single Port RAM1".

When the third row (G11, R12, G13, R14, G15) arrives to A, we can get the pixels of first row from "Single Port RAM1" and store the second row to it. We get the pixels of the second row from "Single Port RAM", and store the third row to it. Finally, locations 1, 2, 3 store the pixels from the third row, locations 4, 5, 6 store the pixels from the second row, and locations 7, 8, 9 store the pixels from the first row.

All of the delay blocks help make the pixels stay in the circuit temporarily, in order to apply the interpolation method to the pixels. For example, at one moment, G1, R2, G3, B6, G7, B8, G11, R12 and G13 can be obtained from location 1 to location 9. Then G7 is the center of the 3 by 3 array (G1, R2, G3, B6, G7, B8, G11, R12, and G13). Now G7 does not have any red and blue values but only has a green value. We directly connect location 5 to "Shift 2", and get the green value through "Mux2". In order to get the blue element, add the values from location 4 and location 6, and get it through "Mux3". The red element is similar, add the values from location 2 and location 8, and get it through "Mux4".

In the design, we get the average value via the slice block. Slice block is used to truncate the binary bits. For example, binary 1110 (decimal 14), and we remove the last bit 0, the result is 111 (decimal 7), which is 14 divided by 2. In this way, we can simplify some of the calculations.

VI. Bayer filter using Smooth hue transition interpolation

Modifying the Xilinx reference design, we designed a Bayer filter that uses a smooth hue transition interpolation. In the reference design, 9 locations (from location 1 to location 9) are needed to store the pixels temporarily, which actually is a 3 by 3 array. Our modified design uses a 5 by 5 array, that is, 25 locations (from location 1 to location 25). Fig. 4 shows part of the design. We again use Fig. 1 to explain the process.

At one moment, all of the pixels in Fig. 1 are corresponding to the locations in Fig. 4. For example, pixel G1 is at location 1, and pixel G13 is at location 13, etc. Further, G13 is the center of the array because it is a green element. We can directly forward its value to the Mux block. We use Step (1) to estimate G12 and G14 and the following to estimate the red values of element 13:

$$R13 = (G13/2) * (R12/G12 + R14/G14) \quad (8)$$

Besides, we need two additional blocks here, Multiplier and Divider. We add values at locations 11, 13, 7, 17, divide the

sum by 4, and get the green value G12 at location 12. We can also estimate G14 through the values at locations 13, 15, 9, 19. Notice that the values at location 12, 14 are red elements. We connect location 12 with G12, location 14 with G14, and calculate the quotients R12/G12 and R14/G14. Then we sum the two quotients by an adder and connect the sum and G13 with a multiplier. Finally, we pass the product to the shift block, right shift 1 bit (divide by 2) and get the red value R14. The estimation of the blue value at location 13 can be done similarly to the estimation of the red value.

At another moment, pixel R14 is the center of the array. Since there is a red value at location 14, we forward it to the "Mux" block. We get the green value using as in the bilinear interpolation. We estimate the blue value of this pixel using Step (9). The elements G8, G10, G18, G20 can be estimated via bilinear interpolation. Then in order to estimate the blue value at location 14, we calculate the quotients B8/G8, B10/G10, B18/G18 and B20/G20, sum the four quotients, multiply the sum by G14, and block shift the results, that is, to divide by four. That can be expressed using the formula:

$$B14 = G14 / 4 \times (B8 / G8 + B10 / G10 + B18 / G18 + B20 / G20) \quad (9)$$

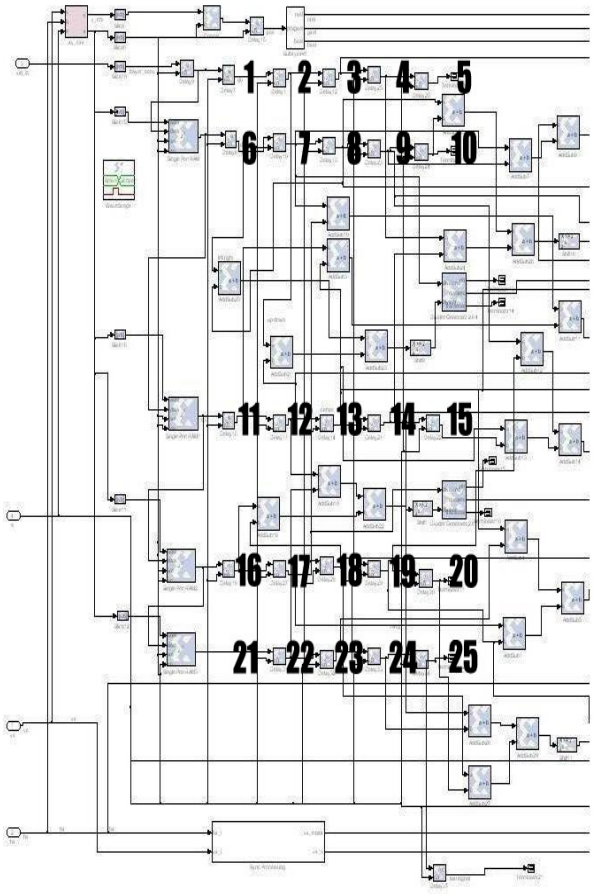


Fig. 4 Design by smooth hue transition

VII. Experiment Results

We use signal-noise ratios to compare the quality of the images reconstructed by our MATLAB implementation described in Section IV and our modified Bayer filter described in Section VI. As an example, Figs. 2 and 5 show a picture of the first author as reconstructed by these two methods, respectively. In addition, Table 1 shows the comparison result for the same image and another image, which is available from the first author's B.S. thesis.

Fig. 5 Smooth hue transition by System Generator

Table. 1 Signal-to-Noise Ratios

	MATLAB Implementation	Modified Bayer
fountain	4.7923	8.1590
person	9.5556	9.6273

The table suggests that modified Bayer filter is generally a better an interpolation method.

References

[1] *Image Sensor Architectures for Digital Cinematography,* DALSA Corp.

[2] X. Lia, B. Gunturkb and L. Zhangc, "Image Demosaicing: A Systematic Survey," *SPIE Proceedings,* vol. 6822, Jan. 2008.

[3] B. Bayer "Color imaging array," U.S. Patent 3 971 065, July 20, 1976.

[4] T. Wittman, "Mathematical Techniques for Image Interpolation," unpublished.

[5] *System Generator for DSP Getting Started Guide,* Xilinx Inc., 2012.

[6] *System Generator for DSP Reference Guide*, Xilinx Inc, Apr. 2008.

[7] Xilinx, Available: http://www.xilinx.com

[8] R. Kimmel, "Demosaicing: Image Reconstruction from Color CCD Samples," *IEEE TRANSACTIONS ON IMAGE PROCESSING*, vol. 8, Sept. 1999, pp. 1221-1228.

[9] R. Lukac, "Color Filter Arrays: Design and Performance Analysis," *IEEE Transactions on Consumer Electronics,* vol. 51, No. 4, Nov. 2005, pp. 1260-1267.

[10] R. Maschal, S. Young, J. Reynolds, K. Krapels, J. Fanning, and T. Corbin, "Review of bayer pattern color filter array (cfa) demosaicing with new quality assessment algorithms," U.S. Army Res. Lab, 2010.

[11] *Spartan 3A-DSP FPGA Video Starter Kit User guide,* Xilinx Inc., Apr. 2008.

[12] S. Battiato, A. Castorina and M. Mancuso, "High dynamic range imaging for digital still camera: an overview," *Journal of Electronic Imaging,* vol. 12(3), Jul. 2003, pp. 459–469.

[13] A. A. Tanbakuchi, A. V. D. Sijde, B. Dillen, A. J. P. Theuwissen and W. d. Haan, "Adaptive pixel defect correction," *SPIE Proceedings,* vol. 5017, May. 2003.

Cost optimization and redundancy allocation of availability constrained heterogeneous series-parallel systems using genetic computing

W. Chaaban, M. Schwarz, and J. Börcsök

***Abstract*—**After analyzing the homogeneous case in a previous work this paper focuses on finding cost optimized heterogeneous redundant structures for series-parallel multi-state systems (SP-MSS) under specified availability constraints. These safety designs identified by non-identical components redundancy are less susceptible against common cause failure and guarantee longer operation times, i.e. longer availability for lower cost. This kind of combinatorial optimization tasks is perfectly solved using biologically inspired genetic algorithms which showed stability, powerfulness, and computing effectiveness. This matter is more complex than the homogeneous case since chromosomes are getting longer and therefore, the search space is getting larger since mixing of components is allowed, the fact that would definitely leads to longer computation times towards convergence. The algorithm has been implemented in Matlab and for validation purposes, tests have been performed using data belonging to already studied models (Levitin, Lisnianski, and Ouzineb).

***Keywords*—**Common Cause Failure(CCF), Genetic Algorithms, heterogeneous Series-Parallel Systems, Redundancy Allocation Problem (RAP), Universal Generating Function (UGF).

I. Introduction

THE Redundancy Allocation Problem (RAP) intends to select best suitable components in addition of determining the redundancy level of each subsystem taking into account predefined system design requirement specifications and constraints like availability, weight, volume, and etc.

Many different deterministic approaches have been previously used in solving such combinatorial optimization matters, e.g. integer programming, dynamic programming, and mixed integer and nonlinear programming.

The Redundancy Allocation Problem (RAP) or Redundancy Optimization Problem (ROP) [1] can be often encountered in many applications areas of the safety engineering world like electrical power systems and in the consumer electronic industry where system designs are mostly assembled using standard certified component types with known characteristics, e.g., reliability, availability, nominal performance, cost, etc.. This matter has been intensively studied over the years and has been classified as a complex nonlinear integer programming combinatorial problem, where deterministic or conventional mathematical optimization approaches become ineffective by means of computational effort and quality of solution while treating complex structurs [1].

The application of heuristic and metaheuristic optimization techniques, e.g., Genetic Algorithms (GAs), Tabu Search (TS), Simulated Annealing, etc., on such kind of combinatorial optimization problems aims to determine an optimal or near optimal solution to the proposed RAP, i.e. to find the best or at least a solution that fulfills the specified or predefined constraints. These approaches have shown instead how powerful and effective they are as means to find high qualitative solutions for the addressed kind of problems, especially when the task becomes more complicated and the conventional solution methods would become ineffective. This kind of problems was first introduced by Ushakov [1] and has been further analyzed by Levitin and Lisnianski et al. [2], [3], and [4], Ouzineb [5], [6] and [7], and many others.

As mentioned at the beginning, this paper deals with heterogeneously structured series assembled systems, where mixing of components is allowed. This feature, compared with homogeneous structure, includes more complexity to the task, because the corresponding search or solution space becomes larger, since every component available on the market has to be taken into account in this case.

The remainder of the paper is organized as follows. Section 2 gives a short introduction into heterogeneous series-parallel multi-states configurations and a brief overview on the advantages obtained through mixing of components. Section 3 discusses shortly genetic algorithms and its different operators while chromosomal encoding and random generating of solution candidates are implemented in section 4. In section 5 a detailed formulation of the optimization problem discussed in this paper which is solved using heuristic traditional GA genetic techniques is presented. Section 6 reports different numerical results, evaluations, and graphical representations obtained by the implemented GA algorithm for different analyzed models which will be compared with previously published evaluations in term of efficiency, solution quality

W.C. Author is with the Department of Computer Architecture and System Programming, University of Kassel, Kassel, D 34121Germany, (e-mail: walid.chaaban @uni-kassel.de).

M.S. Author is with the Department of Computer Architecture and System Programming, University of Kassel, Kassel, D 34121Germany, (e-mail: m.schwarz@uni-kassel.de).

J.B. Author is with the Department of Computer Architecture and System Programming, University of Kassel, Kassel, D 34121Germany, (e-mail: j.boercsoek@uni-kassel.de).

ISBN: 978-1-61804-256-9

and accuracy in addition to algorithm computation speed and convergence time while concluding remarks are resumed in section 7.

II. Heterogeneous Series-Parallel Configurations

In order to improve system reliability and provide longer operation time, safety system designers may introduce different parallel technologies into a system also called redundancy [8]. Including homogeneous components redundancy is a great and effective technique to achieve a desired level of reliability in binary state systems or to increase the availability of multi-state systems. Reliability analysis have shown that the availability of homogeneous redundant structures or systems is extremely affected by common cause failure (CCF), that cannot be ignored since the CCF is the simultaneous failure of all components of the same type due to a common cause (CC), which leads in homogeneous redundant structures definitely to the failure of complete subsystems consisting of identical components causing herewith a total system failure. Common cause events may be caused by environmental loads (humidity, temperature, vibration, shock, etc.), errors in maintenance and system design flaws [9]. In order to partly overcome this kind of facing problems and avoid total system failure subject to CCF heterogeneous redundancy is used.

The main concept of heterogeneous or non- homogeneous structures consists of the mixture of non-identical components within the same subsystem. That means that all non-identical components with the same functionality available on the market and which can be deployed in a redundant manner within the same subsystem have to be taken into account in this case, the fact that would definitely enlarges the size of the search space of feasible solutions and increases the exploration and hence the convergence time towards acceptable solutions.

The main advantages and benefits of components mixing lie in the improvement of the availability of the whole system and reducing the effect of common cause failure in addition of introducing flexibility and diversification into redundant system design through the allowed multiple component choice.

Fig. 1 represents a heterogeneous series-parallel multi-state system consisting of *s* subsystems which are connected serially.

In Fig. 1 r_{ij} represents the reliability of a component of version *j* within the subsystem *i*. In the case of a homogeneous configuration the reliabilities of all components within the same subsystem is the same since subsystems consists of identical components, i.e. all r_{ij}'s are equal for the same *i*.

For a brief explanation in addition to short mathematical computation that shows, why series-parallel configurations are more suitable and studied than parallel-series configurations, the reader is referred to [1] and [10]. This is due to the fact that the overall reliability or availability of a system in a series-parallel configuration is better than the corresponding parallel-series configuration using the same set of components.

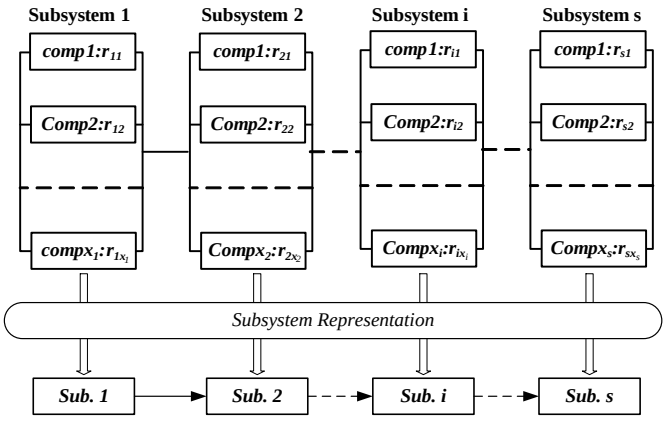


Fig. 1 heterogeneous series-parallel configuration consisting of s-Nodes or subsystems

III. Genetic Algorithms and Corresponding Objective Function

Genetic algorithms (GAs) are biologically inspired metaheuristic search and optimization routines that mimic the act of self-evolution of natural species and are frequently used nowadays in many fields like self- adaptiveness, artificial intelligence and machine learning tasks. As computational efforts and speeds have been increasingly improved over the last decade, GAs have been expanded to cover a wide variety of applications including numerical and combinatorial optimization tasks in engineering like the one discussed in the recent work. Further fundamentals and detailed information on genetic algorithms can be found in [11] and [12].

GAs represent iterative self-adaptive stochastic techniques based on the idea of randomness. They mimic the process of natural evolution of species. GAs became very popular and widely used over the last decade and are very well suited as universal or common techniques for solving combinatorial optimization problems, e.g., the very well-known TSP (Travelling Salesman Problem) and redundancy allocation problems like the one discussed in this paper and many other matters [13].

Genetic algorithms starts the search from a start (initial) population constituting of different randomly generated chromosomes, also called solution candidates that are encoded according to the addressed problem (binary, integer, decimal, etc.) conducting herewith a simultaneous search in many areas of the feasible solution space at once. The encoding of solutions constitutes the most difficult and challenging task of GAs and the evolving procedure from one population to the next is called generation.

After each generation the new generated solutions are decoded and evaluated with the help of the fitness function also called objective function. The fitness value of a chromosome represents a measure for its quality (fitness). A general overview of the genetic cycle is given in Fig. 2.

ISBN: 978-1-61804-256-9

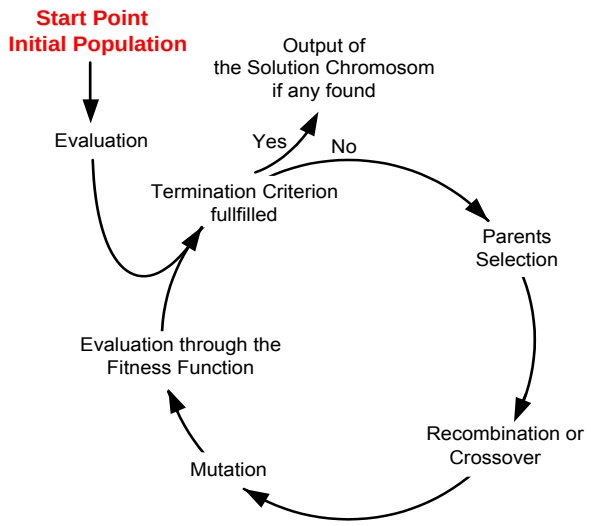


Fig. 2 general overview of the genetic process

The genetic run process terminates when at least one of its predefined termination criterions is met, e.g., when the predefined maximum number of generations or repetitions N_{rep} or a specific number of successive runs without any solution's improvement is reached.

Three main operators, also called genetic operators will be executed during one genetic cycle, hence the selection, crossover or recombination, and the mutation operator. These operators are shortly discussed in the following subsections.

A. *Selection Operator*

Outgoing from a start or an initial population of different solution candidates the selection operator is used to randomly select or choose pair of individuals or chromosomes which will reproduce and help building the next population during the genetic cycle. The chromosomes will be selected randomly according to the Darwinian principle (survival of the fittest-selection probability proportional to relative fitness), which drives the evolution towards optimization. There are many selection methods, some of them are listed in the following [14]

- Roulette Wheel selection,
- Tournament selection,
- Rank selection,
- etc.

B. *Recombination or Crossover*

Whereas the selection operator determines which chromosomes of the recent population are going to reproduce, the crossover operator performs jumps between the different solution subspaces enabling the exploration of new areas of the solution space and avoiding herewith premature convergence in addition of exchanging some basic characteristics and inheriting these properties to the offsprings which will join next populations. The crossover occurs with a predefined crossover rate pc. There are many crossover techniques used in genetic algorithms [1][14][15], like the one-point crossover, two-point crossover, uniform and half uniform crossover and many other crossover techniques.

In the following the one-point crossover operator is shortly discussed. A crossover point depending on the length of the chromosome is randomly selected on both selected parent chromosomes. All data beyond that point in either chromosome is swapped between the two parent organisms so that two new individuals called offsprings or children chromosomes result. The one-point crossover technique is depicted in Fig. 3.

C. *Mutation*

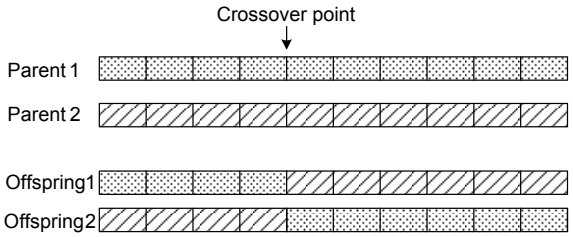


Fig. 3 one-point crossover technique

After crossovering parent chromosomes the resulting offsprings undergo mutation with a low mutation rate pm. The mutation operator introduces diversity into the GA algorithm and inserts small disturbance into the properties (genes) of the proposed solutions avoiding herewith convergence into local maxima. After the mutation process has been accomplished, the new resulting mutated chromosomes can join the next population.

IV. Chromosomal Encoding and Random Generating of Solution Candidates

Genetic algorithms are population based combinatorial optimization approaches where each population consists of a predefined number of solution candidates, also called chromosomes or individuals. These candidates represent vectors of encoded information which will be decoded using the fitness or objective function in order to find the optimal or near optimal solution of the addressed problem, whereas the recent approaches for solving the RAP problem are based on the universal moment generating function for estimating the availability of multi-state systems [2] [5][6][7][16]. The encoding of chromosomes consists in general one of the major challenges faced in the world of genetic computing.

In such problems addressed in this paper, i.e., in the case of heterogeneous redundant structures the chromosome length corresponding to the same system is definitely longer than the homogeneous case since each component version available on the market that may be deployed in a subsystem has to be taken into account, whereas in homogeneous systems, the chromosome length is equal to twice the number of subsystems, since only one component type or version is allowed on each stage. The chromosomes are integer encoded, and each element x_{ij} of the chromosome vector corresponds to

ISBN: 978-1-61804-256-9

the number of components of version j used in subsystem *i*. The chromosome dimension or length l_c is given therefore through

$$l_c = \sum_{i=1}^{s} J_i \quad (1)$$

where s is the total number of subsystems or stages and J_i the maximal number of versions available on the market that can be used on stage *i*. For more clarification, it is important to mention at this point, that two components of different versions connected in parallel are supposed to perform the same task or function. The difference lies in the technical data (reliability/availability, nominal performance and etc.) in addition to the purchasing price.

For example let us consider a system consisting totally of 4 subsystems with the following version vector *m*= [4 6 8 5] representing the number of components available on the market for each subsystem. The chromosome length results in this case as the sum of all elements of the version vector and would give according to (1) a total chromosome length of 4+6+8+5=23 and the chromosome encoding would look like in the following

$$X=([x_{11}..x_{14}],[x_{21}...x_{26}],[x_{31}...x_{38}],[x_{41}...x_{45}]) \quad (2)$$

where x_{ij} denotes as mentioned previously the redundancy number of components type j in subsystem *i*.

For generating chromosomes or solution candidates according to the addressed optimization problem discussed in this paper a pseudo random number generator is needed. Since events should happen at random but some events or numbers within the chromosomal encoding should have a higher probability of occurrence or happening than others, e.g. zero's which means that no components of this kind is used, a weighted random number generator is used.

The upper mentioned step of generating numbers with predefined probabilities of occurrence in addition to the limitation of the maximum number of totally used components within each subsystem should limit the search space, the fact that would increase the computation speed drastically towards convergence making the algorithm more efficient.

V. Cost Optimization and Redundancy Allocation – Formulation

The cost optimization problem of series-parallel redundant systems deals with determining optimal redundant designs and the level of redundancy used in each subsystem which corresponds to the minimal total purchase cost of the system and which fulfils at the same time predefined availability constraint. This kind of optimization gives a rise to safety vs. economics conflicts resumed in the following two points [17]:

Choice of components: choosing high reliable components guarantees high system availability but may be largely non-economic due to high purchase prices; whereas choosing less reliable components for lower costs on one hand may decrease the availability of the system and increase drastically the accident costs on the other hand.

Choice of redundancy configuration: choosing highly redundant configurations increases definitely the reliability and availability of the system and is accompanied at the same time with higher purchase costs caused by additional used equipment units required for improving individual subsystems reliabilities.

The upper described aspects of safety system design call for compromise choices which optimize system operation in view of recommended safety and longer operation time or budget constraint. As mentioned before this paper deals with budgetary optimization and redundancy determination of multi state systems under given availability constraint. This problem can be mathematically formulated as to minimize the cost function $C_{sys}(X)$ (objective function) of the whole system given by [1], [5], [6], [7], and [16]:

$$C_{sys}(X) = \sum_{i=1}^{s} \sum_{j=1}^{m_i} c_{ij} x_{ij} \quad (3)$$

Where c_{ij} being the cost of component of type *j* in subsystem *i* and x_{ij} the number of components of type j used in subsystem *i*. mi is the number of component choices available on the market which may be deployed in subsystem *i*. The (cost) objective function represented in the upper equation results over the sum of the purchasing costs of all components used in system that should fulfil at the same time the system specified availability constraint which implies that the total availability of the system $A_{sys}(X)$ must match or surpass a minimum level of availability required A_0 (inequality or availability constraints)

$$A_{sys}(X) \geq A_0 \quad (4)$$

Based on the *UGF* (Universal Generating Function) or the *Ushakov*- transform, the total availability of the system $A_{sys}(X)$ is estimated as a function of system structure, performance and availability characteristics of its constituting components.

For a detailed overview on the UGF in computing the availability of series-parallel systems the reader is referred to [2], [3], [5], [6], [7], [10], [17], and [18].

VI. Tuning Parameters And experimental Results

The simple genetic algorithm with some simple modifications was used in this work. The algorithm has been implemented in Matlab which provides powerful matrix and vector operations and allows great visualization and graphical representation. The 3 different models analyzed in the homogeneous case in the previous publication [10] and [18] have been treated again in the heterogeneous case, in order to show the effect of mixing of components on investment cost reduction and hence getting safer systems subject to CCFs for lower cost than the homogeneous case for the same given constraints. The used components are assumed to be binary state (perfect working or totally failing). The models and data

as mentioned previously have been taken from [5], [6] and [7], and are listed below:

- Lev4_4_6_3
- Lev5_5_9_4
- Ouz6_4_11_4

For the decoding of the denotation of the individual models it is referred to [5], [6], and [7]. The purchasing price, reliability and nominal performance capacity for the components corresponding to the upper listed systems are supposed to be known and can be retrieved from a list with technical data (excel sheets).

The algorithm starts by retrieving the data of the analyzed problem and by random generating a so called initial population of size Pop_{size} which have been set to 100 chromosomes. The integer encoded chromosomes constituting the initial population have been generated in such a way that solutions that do not fulfil the given availability constraint are rejected and replaced by new acceptable ones in order to get a high qualitative start population. The constituting individuals or chromosomes have been created using a weighted pseudo random number generator which generates numbers between zero and the total number of components allowed in each subsystem which have been set to 10. The probability of occurrence of 0 has been varied between 0.7 and 0.9 during the random chromosome generation process depending on the length of the chromosome corresponding to the analyzed problem.

After each genetic cycle the chromosomes of the new population will be checked for duplicity and multiple occurrence, and the population will be changed in such a way that each chromosome occurs only once within the same population in order to avoid repetition and multiple times evaluation of the same chromosome with the objective function; the fact that would definitely decrease the computation time and hence the algorithm convergence time and would affect the algorithm effectiveness.

The elitism selection technique has been applied, so that the best 10 different chromosome of the preceding population have been always explicitly selected for further crossovering and mutation.

Both crossover techniques the one point and the two point crossover have been used simultaneously in the context of the genetic algorithm analyzed in this work. While crossovering each time 2 random numbers are generated between 1 and the number of subsystems. If the generated numbers match, the one point crossover technique is applied, otherwise the 2 points crossover. The crossover has been performed in such a way that the complete chromosomes corresponding to the subsystems have been exchanged in order to keep the maximum number of allowed components within the subsystem remaining, since taking random points in order to perform crossover in this case would definitely lead to an exceeding in the number of components allowed in each subsystem of the resulting offsprings an may herewith increase the search time. The crossover rate has been set to 0.7 and the mutation rate has been varied between 0.05 and 0.3.

After ranking and evaluating populations chromosomes are selected to mate, recombine, and finally mutate in order to build new offsprings that complete the next population of size Pop_{size}. This genetic procedure repeats until the predefined maximal number of generations N_{rep} is reached. After completing each population through crossover and mutation, the population will be checked for multiplicity and new chromosomes would be generated to replace the chromosomes that have been removed. This procedure of inserting new chromosomes to the population may lead to new search area that has not been explored before and may accelerate the convergence speed. Fig. 6 shows the results of one run of the GA over the Lev4-(4/6)-3 model data by an availability constraint of A_0=0.900. The different plots show the evolution process outgoing from the random initial population up to the predefined maximum number of generations. The best result (Cost and Availability) got after each genetic cycle is depicted.

The time needed to find the best solution depends on the quality of the start population and on how the selected fittest chromosomes evolve throughout crossover and mutation. On the left hand side of Fig. 4 and Fig. 6 the best solution found (Top: cost value, bottom: Availability value for found cost) during each generation is plotted against generation number whereas the same plots are represented on the right hand side against processing time. On the head of each plot, the best chromosome corresponding to the optimal (minimal) found cost subject to the given availability constraint is represented. In the context of the plots the generation number and convergence time are reported for which the best result has been identified.

Fig. 5 and Fig. 7 represent the homogeneous case of the heterogeneous problems analyzed successively in Fig. 4 and Fig. 6. These figures have been included in order to show that through mixing of components lower system costs can be reached in comparison to the homogeneous case subject to the same availability constraint. One additional reason is to show, that with the GA approach analyzed in this paper it was also possible to get same results got with the hybridized GA+TS algorithm implemented in [5] and [7] the fact that shows the effectiveness of the GA approach discussed in this work since with the GA implemented in [6] and [7] different results have been achieved.

The best test results got within 15 successive runs of the genetic algorithm over the different models mentioned previously are represented in Table 1. Computing and convergence time results are also included and show how well the algorithm is performing in term of convergence speed which can also be seen in the results depicted in Fig. 4 and Fig. 6.

ISBN: 978-1-61804-256-9

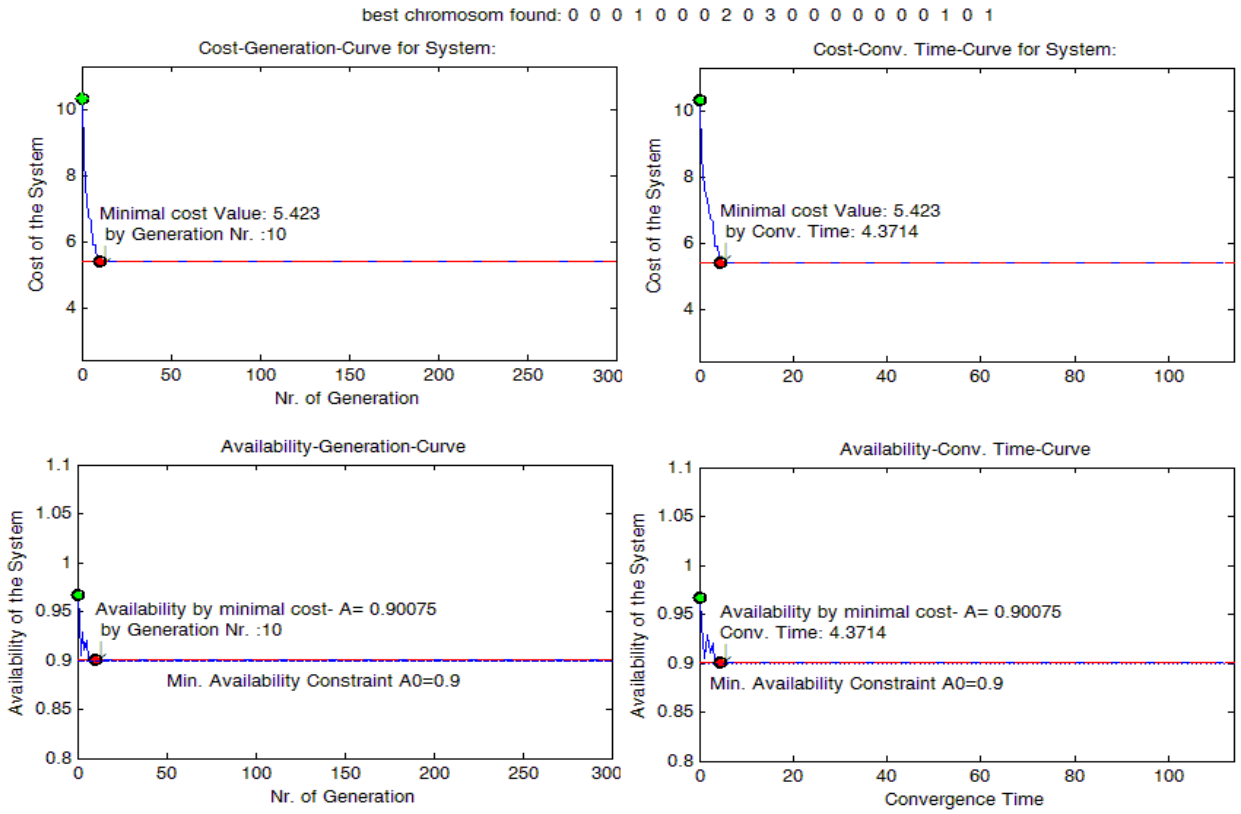


Fig. 4 results of the genetic algorithm run of the heterogeneous case (Levitin- model containing 4 subsystems, availability constraint A0=0.900, and 300 generations). The cost - generation and availability - generation curves dependencies are depicted on the left side, while the cost-time and availability-time dependencies are depicted on the right side.

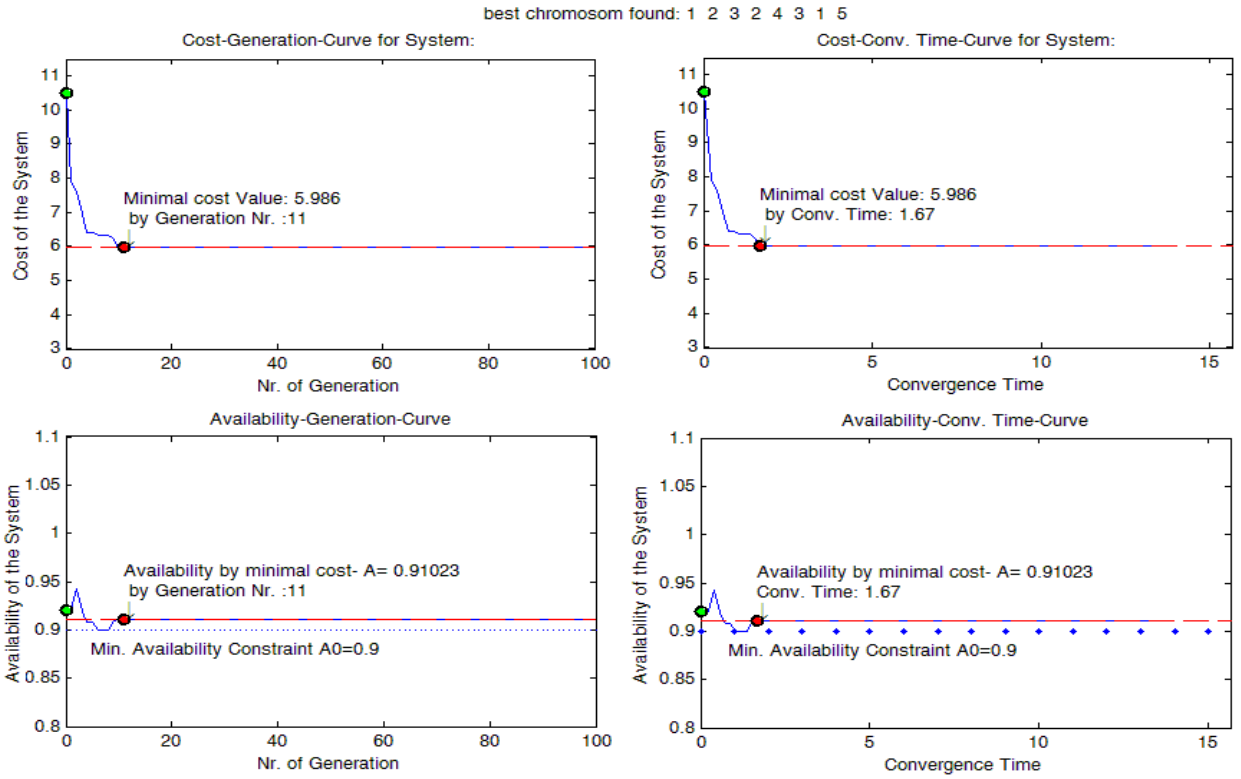


Fig. 5 results of the genetic algorithm run of the homogeneous case for the same upper system (Levitin- model containing 4 subsystems) and subject to the same availability constraint A0=0.900. As one can see in the heterogeneous case a better cost factor can be reached (5.423) than the homogeneous case (5.986) due to the fact of mixing of components.

ISBN: 978-1-61804-256-9

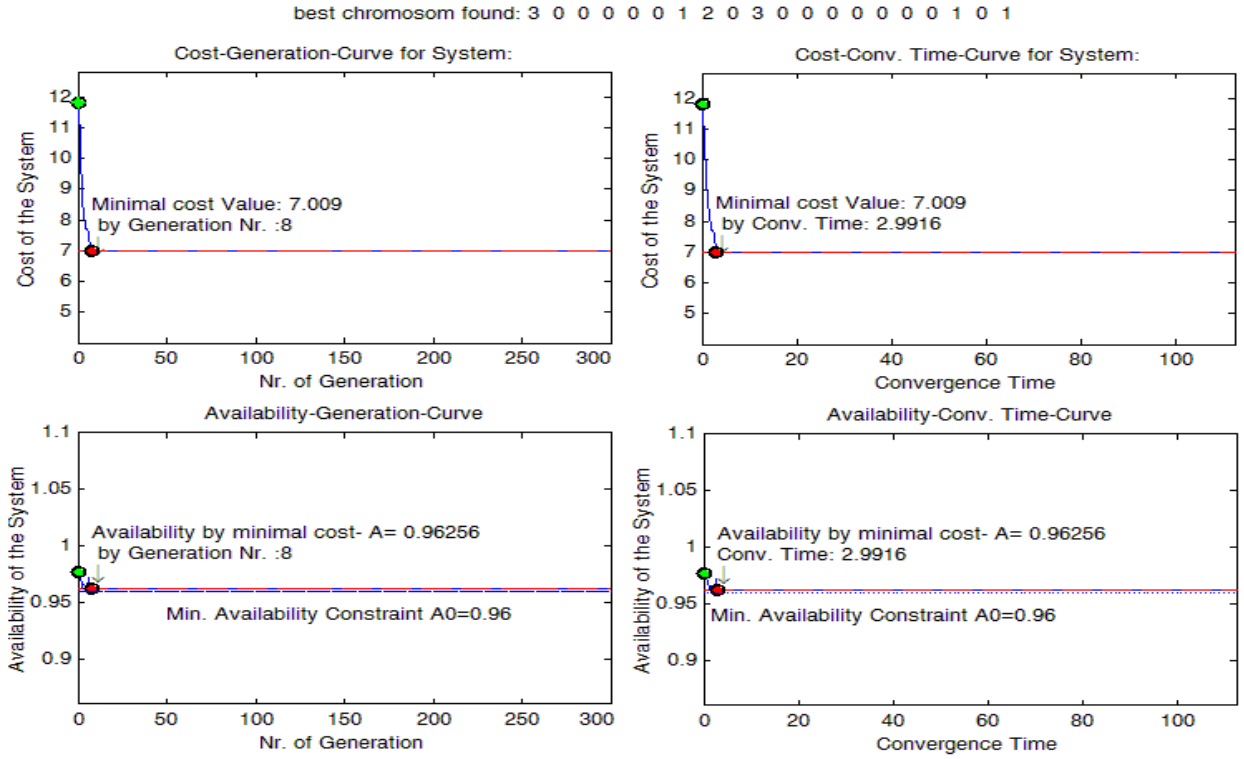


Fig. 6 results of the genetic algorithm run of the heterogeneous case (Levitin- model containing 4 subsystems, availability constraint A0=0.960, and 300 generations). The cost - generation and availability - generation curves dependencies are depicted on the left side, while the cost-time and availability-time dependencies are depicted on the right side.

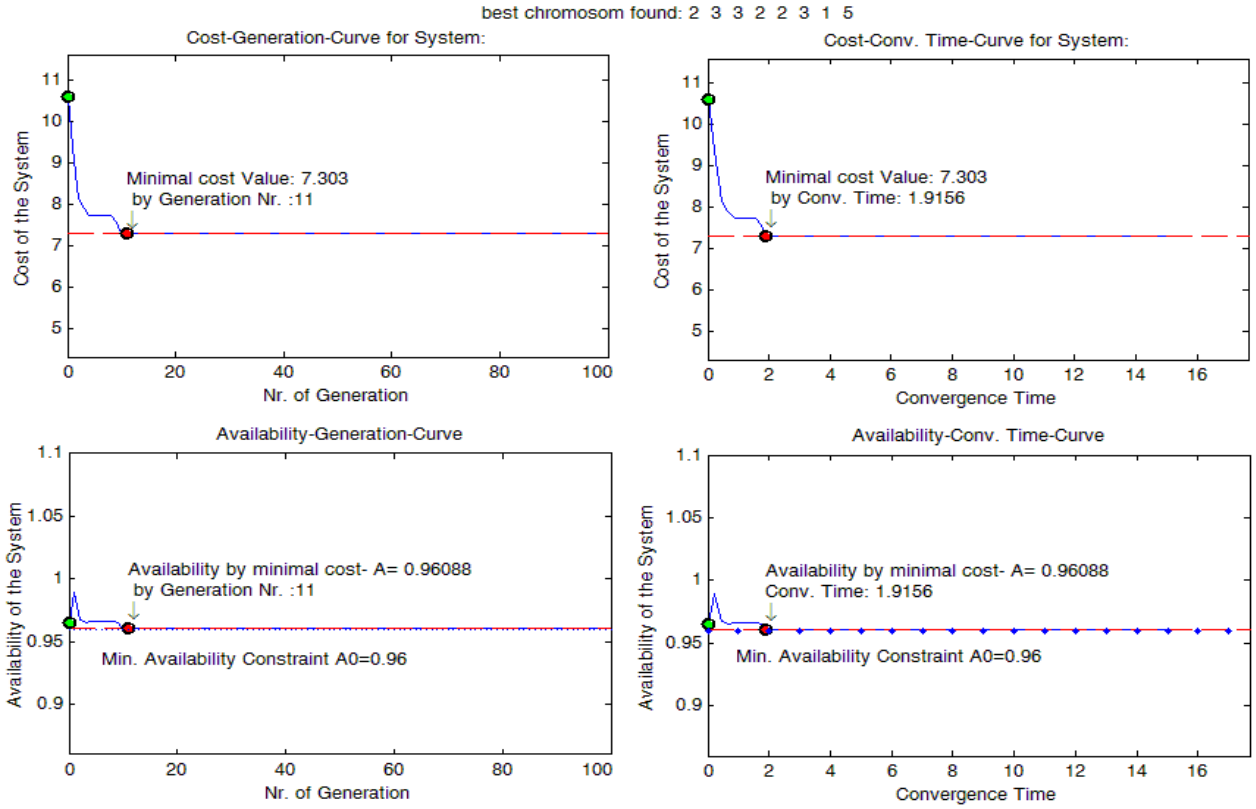


Fig. 7 results of the genetic algorithm run of the homogeneous case for the same upper system (Levitin- model containing 4 subsystems) and subject to the same availability constraint A0=0.960. As one can see in the heterogeneous case a better cost factor can be reached (7.009) than the homogeneous case (7.303) due to the fact of mixing of components.

ISBN: 978-1-61804-256-9

Table 1 computation results of the homogeneous and the heterogeneous case using the GA. It is important to mention at this point that the results got with the GA match either for homogeneous or for the non-homogeneous case the one got by Ouzineb using the hybrid GA+TS (Tabu Search) [5][6][7]. The best values of convergence time of the heterogeneous GA within 15 runs have been also included.

Problem Name	**Av. Constraint** A_0	**Av. Value** ***A(X,J)*** ***Hom.***	**Av. Value** ***A(X)-*** ***Het.***	**Cost** ***C(X,J)(mln $)*** ***Hom.***	**Cost** ***C(J(X))(mln $)*** ***Het.***
lev4-(4/6)-3	0.900	0.9102	0.90075	5.986	5.423
	0.960	0.9609	0.96256	7.303	7.009
	0.990	0.9917	0.99148	8.328	8.180
lev5-(4/9)-4	0.975	0.9774	0.97615	16.450	12.855
	0.980	0.9808	0.98009	16.520	14.770
	0.990	0.9937	0.99211	17.050	15.870
ouz6-(4/11)-4	0.975	0.9790	0.9790	11.241	11.241
	0.980	0.9802	0.9802	11.369	11.369
	0.990	0.9902	0.9902	12.764	12.764

Problem Name	**Best found Chromosome [X,J] Homogeneous case**	**Best found Chromosome [J(X)] Heterogeneous case**						**Convergence time(Het.)**
	Subsystem #	1	2	3	4	5	6	
lev4-(4/6)-3	[1 2 3 2 4 3 1 5]	4(1)	3(2)	1(3)	3(1),5(1)			2.1992
	[2 3 3 2 2 3 1 5]	1(3)	2(1),3(2)	1(3)	3(1),5(1)			1.6700
	[3 3 3 5 1 3 1 2]	1(3)	3(3)	1(3)	3(1),4(2)			2.4501
lev5-(4/9)-4	[2 2 3 3 1 2 3 2 7 2]	4(2),6(1)	5(6)	1(1),4(1)	7(3)	4(3)		16.7701
	[2 6 3 3 1 2 5 2 7 2]	4(2),6(1)	3(2)	2(1),3(2)	7(3)	3(2),4(1)		14.0462
	[2 2 3 3 3 2 3 2 7 4]	4(2),6(1)	3(2)	2(2),3(1)	7(3)	4(3)		11.1849
ouz6-(4/11)-4	[4 4 5 7 2 1 3 1 2 2 3 4]	3(4)	1(4)	2(5)	2(7)	3(2)	4(1)	24.3273
	[4 5 5 8 2 1 3 1 2 2 3 4]	3(4)	1(5)	2(5)	2(8)	3(2)	4(1)	17.4778
	[4 4 4 8 2 2 3 1 2 2 3 4]	3(4)	1(4)	2(4)	2(8)	3(2)	4(1)	16.8972

The algorithm was converging towards optimal values within very promising time factors. It is also important to mention at this point, that all results (availability and cost) got over the handled models match the results obtained by Ouzineb in [5], [6], and [7].

As expected the greater is the number of subsystems and the corresponding available components the longer will be the execution time since the search space and combination possibilities of chromosome combination is getting larger (homogeneous vs. heterogeneous case). The GA algorithm implemented in this paper delivered obviously better results in term of execution and computation speed compared to the results of the GA and GA+TS reported in [7].

VII. Conclusion and Future Works

Based on the facts and experimental results represented in table 1 for the different analyzed models it can be recognized that the GA algorithm implemented in this paper was performing in a great and efficient manner in term of convergence speed towards optimal results being expected and represented by Ouzineb in [5], [6], and [7], and obtained by the hybrid GA+TS metaheuristic approach, in addition to the high accuracy in determining the optimal solution. And since genetic searching seems like searching for a small fish in a big ocean one small disadvantage or drawback is the one known in genetic algorithms and which is resumed in the fact that the best optimal solution is not guaranteed or ensured in each run, due to the limitation of the maximum number of iterations that may result, that some regions of the search or solution space that may include the optimal solution remains unexplored or unreached.

The genetic approach implemented in this paper represents a very effective mean in solving constrained redundancy design problems like the complex heterogeneous one discussed in this work.

One of our future intentions is to tune genetic algorithms with local search algorithms targeting to increase the search accuracy. This kind of tuning is referred to in the literature as hybridization of genetic global searching algorithms.

References

[1] Way Kuo,V. Rajendra Prasad, Frank A.Tillman, and Ching-Lai Hwang, "Optimal Reliability Design, Fundamentals and Applications," Cambridge University Press (2001)

[2] Gregory Levitin, Anatoly Lisnianski, Hanoch Ben Haim, David Elmakis, "Genetic Algorithm and Universal Generating Function Technique for Solving Problems of Power System Reliability

Optimization", The Israel Electric Corporation Ltd., Planning Development & Technology Division (2000)
[3] Gregory Levitin, Anatoly Lisnianski, Hanoch Ben Haim, "Redundancy Optimization for Series-Parallel Multi State Systems" *IEEE Transactions on reliability*, Vol. 47, No. 2 (1998)
[4] Anatoly Lisnianski, Gregory Livitin, Hanoch Ben Haim, David Elmakis, "Power System Optimization subject to Reliability Constraints", *Electric Power Systems Research 39*, 145--152 (1996)
[5] Mohamed Ouzineb, "Heuristiques éfficaces pour l'optimisation de la performance des systèmes séries-parallèles", Département d'informatique et de recherche opérationnelle Faculté des arts et des sciences, Université de Montréal (2009)
[6] Mohamed Ouzineb, Mustapha Nourelfath, Michel Gendreau, "Tabu search for the redundancy allocation problem of homogenous series-parallel multi-state systems", Reliability Engineering and System Safety 93, 1257--1272 (2008)
[7] Mohamed Ouzineb, Mustapha Nourelfath, Michel Gendreau, "A Heuristic Method for Non-Homogeneous Redundancy Optimization of Series-Parallel Multi-State Systems", CIRRELT (2009)
[8] Alice Yalaoui, Chengbin Chu, Eric Châtelet, "Reliability allocation problem in a series–parallel system", University of Technology of Troyes, Institute of Information Sciences and Technologies of Troyes (ISTIT), Reliability Engineering and System Safety 90, 55--61 (2005)
[9] Chun-yang Li, Xun Chen, Xiao-shan Yi, Jun-yong Tao, "Heterogeneous redundancy optimization for multi-state series–parallel systems subject to common cause failures", Reliability Engineering and System Safety 95, 202--207 (2010)
[10] W. Chaaban, M. Schwarz, J. Börcsök, "Budgetary and Redundancy Optimisation of Homogeneous Series-Parallel Systems Subject to Availability Constraints Using Matlab Implemented Genetic Computing", 24th IET Irish, Signals and Systems Conference (ISSC 2013)
[11] Holland J., "Adaptation in Natural and Artificial Systems", The University of Michigan Press, Ann Arbor, Michigan (1975)
[12] Goldberg D., "Genetic Algorithms in Search, Optimization and Machine Learning", Addison Wesley, Reading, MA (1989)
[13] Zhigang Tian, Ming J. Zuo, and Hongzhong Huang, "Reliability-Redundancy Allocation for Multi-State Series-Parallel Systems", IEEE Transactions on Reliability, Vol. 57, No. 2, (2008)
[14] M. Affenzeller, S. Winkler, S. Wagner, and A. Beham, "Genetic Algorithms and Genetic Programming, Modern Concepts and Applications", CRC Press (2009)
[15] Zbigniew Michalewicz, "Genetic Algorithms + Data Structures=Evolution Programs", Third revised and extended Edition. Springer-Verlag (2011)
[16] Frank A. Tillman, Ching-Lai Hwang, Way Kuo, "Optimization Techniques for System Reliability with Redundancy- A Review", *IEEE Transactions on Reliability* (1977)
[17] Gregory Levitin, "The Universal Generating Function in Reliability Analysis and Optimization", Springer (2005)
[18] W. Chaaban, M. Schwarz, J. Börcsök, "Cost and Redundancy Optimization of Homogeneous Series-Parallel Multi-State Systems Subject to Availability Constraints Using a Matlab® Implemented Genetic Algorithm", Recent Advances in Circuits, Systems and Automatic Control. WSEAS 2013, Budapest, Hungary (2013)

2nd Order Differential Equation for Short-Wavelength Defects of the Rail-Head

Konstantinos Giannakos

***Abstract*—** Train circulation is a random dynamic phenomenon. The rail running surface/table, of the rail-head, is a wave in space of completely random form, which imposes a forced oscillation to the wheel. In this paper the second order differential equation of motion is presented for the case of a railway vehicle rolling on a railway track presenting short-wavelength defects on the rail-head/ rail running table.vIts solution is presented and the solution for the dynamic component of the acting load is presented also.

***Keywords*—** Static Stiffness Coefficient, Dynamic Stiffness Coefficient, Suspended Masses, Non Suspended Masses, Dynamic Component of Loads.

I. Introduction

Train circulation is a random dynamic phenomenon and, according to the frequencies of the actions that a railway vehicle imposes, there exists the corresponding response of track superstructure. At the moment when an axle passes from the location of a support point of the rail (sleeper), a totally random dynamic load is applied on the sleeper. The theoretical approach for the most precise identification possible of its probable value, demands the analysis of the total load to individual component loads-actions. The static and semi-static components of the total acting load can be calculated more easily ([1], [2]). Furthermore the rail running surface/table, that is the surface of the rail-head on which the wheel of the railway vehicle is rolling, is a wave in space of completely random form, which imposes a forced oscillation to the wheel and consequently is seriously influencing the dynamic component of the acting load on track. In previous articles the author presented (a) the influence of the track defects and the on the dynamic component loads due to Non-Suspended Masses of the railway vehicles [3], the influence of an isolated defect on the dynamic component of the loads [4] and the influence of a battered or bent joint or welding on the dynamic component of the loads [5]. In the present paper the analysis of the influence of the undulatory wear of the rail running surface/table is presented.

K. G. civil engineer, PhD, F.ASCE, M. TRB AR050 & AR060 Comm., AREMA, fib, was with University of Thessaly, Dpt. of Civil Engineering, Pedion Areos, 38334, Volos, Greece. He is now freelancer consultant 108 Neoreion str., Piraeus 18534, Greece, (e-mail: k.giannakos@on.gr, kyannak @gmail.com).

II. Railway Track: 2nd Order Differential Equation of Motion

The railway vehicle rolls over the railway track and the system "vehicle-track" functions as an ensemble (Fig. 1 depicts the "vehicle-track" system as an ensemble of springs and dashpots). The rail running table imposes to the vehicle a forced vibration. It is not smooth but instead it comprises a lot of faults that give to the rail running table a random surface. Furthermore under the primary suspension of the vehicle there are the Non-Suspended Masses (axle, wheels and a fraction of the semi-suspended motive electromotor in the locomotives), which act without any dumping directly on the track panel. On the contrary the Suspended Masses, cited above the primary suspension of the vehicle, act through a combination of springs and dampers on the track. A part of the track mass is also added to the Non Suspended Masses, which participates in their motion [6]. The rail running table has the shape of a wave that is not completely "rectilinear", that is, it does not form a perfectly straight line but contains faults, varying from a few fractions of a millimeter to a few millimeters, and imposes forced oscillation on the railway vehicles that move on it.

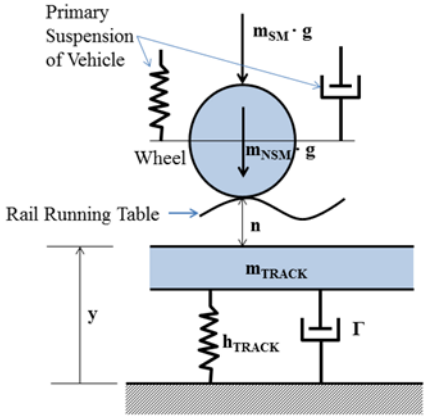


Fig. 1 simulation of the rail running table with its defects with a model of a rolling wheel on it.

The Suspended Masses of the vehicle –masses situated above the primary suspension– create forces with very small influence on the wheel's trajectory and on the system's excitation, particularly in the case of the defects of very short wavelength, like the undulatory wear, with wavelengths fluctuating from some millimeters to a few centimeters. This enables the simulation of the track as an elastic media with

damping as shown in Fig. 1, depicting the rolling wheel on the rail running table [1]. Forced oscillation is caused by the irregularities of the rail running table (like an input random signal) –which are represented by n–, in a gravitational field with acceleration g. There are two suspensions on the vehicle for passenger comfort purposes: primary and secondary suspension. Moreover, a section of the mass of the railway track participates in the motion of the Non-Suspended Masses of the vehicle. These Masses are situated under the primary suspension of the vehicle. If the random excitation (track irregularities) is given, it is difficult to derive the response, unless the system is linear and invariable. In this case the input signal can be defined by its spectral density and from this we can calculate the spectral density of the response. The theoretical results confirm and explain the experimental verifications ([7], p.39, 71). The general form of the equation of a vehicle moving on an infinitely long beam, on elastic ground without damping, is:

$$\left(m_{NSM}+m_{TRACK}\right)\cdot\frac{d^2y}{dt^2}+\Gamma\cdot\frac{dy}{dt}+h_{TRACK}\cdot y= \\ =-m_{NSM}\cdot\frac{d^2n}{dt^2}+\left(m_{NSM}+m_{SM}\right)\cdot g \qquad (1)$$

where: m_{NSM} the Non-Suspended Masses of the vehicle, m_{TRACK} the mass of the track that participates in the motion, m_{SM} the Suspended Masses of the vehicle that are cited above the primary suspension of the vehicle, Γ damping constant of the track, h_{TRACK} the total dynamic stiffness coefficient of the track, n the fault ordinate of the rail running table and y the deflection of the track.

Under the assumption that the case, we analyze, is far from the critical speed of the trains, that is $V_{analyzed} << V_{critical}$=500 km/h: the phenomena of the wheel-rail contact and of the wheel hunting, particularly the equivalent conicity of the wheel and the forces of pseudo-glide, are non-linear. In any case the use of the linear system's approach is valid for speeds lower than the $V_{critical}\approx$500 km/h. The integration for the non-linear model (wheel-rail contact, wheel-hunting and pseudoglide forces) is performed through the Runge Kutta method ([2], p.94-95, 80, [9], p.98, see also [10], p.171, 351).

The total dynamic stiffness coefficient of the track h_{TRACK} is calculated from the total static stiffness coefficient of the track ρ_{total}, which is a quasi-spring constant of the track. The inverse of ρ_{total} is given by the sum of the inverse static stiffness coefficients of the individual layers/ elements of the track's cross-section (Fig. 2):

$$\frac{1}{\rho_{total}}=\frac{1}{\rho_{rail}}+\frac{1}{\rho_{pad}}+\frac{1}{\rho_{sleeper}}+\frac{1}{\rho_{ballast}}+\frac{1}{\rho_{subgrade}} \qquad (2)$$

and:

$$h_{TRACK}=2\sqrt{2}\cdot\sqrt[4]{\frac{EJ\rho_{total}^3}{\ell^3}} \qquad (3)$$

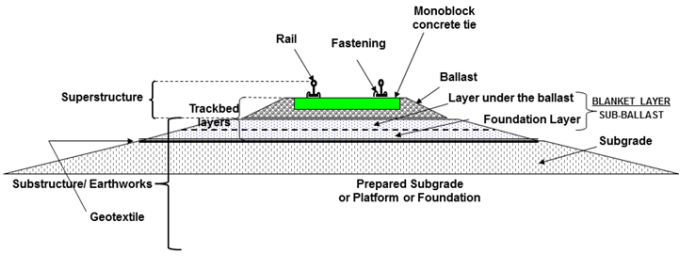


Fig. 2 cross-section of a railway track, with terminology according to the International Union of Railways (U.I.C.[1]).

Where, E, J the modulus of elasticity and the moment of inertia of the rail (steel) and ℓ the distance among the sleepers/ support points of the rail.

III. Transforming the 2nd Order Differential Equation, due to the Track's Defects

In Fig. 1 the rail running table depicts a longitudinal fault/ defect of the rail running table. In the above equation (1), the oscillation of the axle is damped after its passage over the defect. Viscous damping, due to the ballast, enters the above equation under the condition that it is proportional to the variation of the deflection dy/dt. To simplify the investigation, if we ignore the track mass (for its calculation see [6]) in relation to the much larger Vehicle's Non Suspended Mass and bearing in mind that y+n is the total subsidence of the wheel during its motion (since the y and n are added algebraically), we can approach the problem of the random excitation, from cosine defect:

$$\eta=a\cdot\cos\omega t=a\cdot\cos\left(2\pi\cdot\frac{V\cdot t}{\lambda}\right) \qquad (4)$$

where V the speed of the vehicle, $T=2\pi/\omega\rightarrow \omega t=2\pi/(Tt)=2\pi Vt/\lambda$ where λ the length of the defect, run by the wheel in:

$$T=\frac{\lambda}{V}\Rightarrow\lambda=T\cdot V \qquad (5)$$

If we set:

$$y=z+\frac{m_{SM}+m_{NSM}}{h_{TRACK}}\cdot g\Rightarrow\frac{dy}{dt}=\frac{dz}{dt}\quad and\quad \frac{d^2y}{dt^2}=\frac{d^2z}{dt^2} \qquad (6)$$

where the quantity:

$$\frac{m_{SM}+m_{NSM}}{h_{TRACK}}\cdot g \qquad (7)$$

[1] U.I.C. the initials in French: Union Internationalle des Chemins de Fer (International Union of Railways).

represents the deflection due to the static loads only, and z random (see [10]) due to the dynamic loads. Equation (1) becomes:

$$m_{NSM}\frac{d^2z}{dt^2}+\Gamma\cdot\frac{dz}{dt}+h_{TRACK}\cdot z=-m_{NSM}\cdot\frac{d^2n}{dt^2}\Rightarrow \quad (8a)$$

$$\Rightarrow m_{NSM}\left(\frac{d^2z}{dt^2}+\frac{d^2n}{dt^2}\right)+\Gamma\cdot\frac{dz}{dt}+h_{TRACK}\cdot z=0 \quad (8b)$$

Since, in this case, we are examining the dynamic loads only, in order to approach their effect, we could narrow the study of equation (8b), by changing the variable:

$$u=n+z\Rightarrow\frac{d^2u}{dt^2}=\frac{d^2n}{dt^2}+\frac{d^2z}{dt^2} \quad (9)$$

Equations (8) become:

$$m_{NSM}\frac{d^2u}{dt^2}+\Gamma\cdot\frac{dz}{dt}+h_{TRACK}\cdot z=0\Rightarrow \quad (10a)$$

$$\Rightarrow m_{NSM}\frac{d^2u}{dt^2}+\Gamma\cdot\frac{d(u-n)}{dt}+h_{TRACK}\cdot(u-n)=0 \quad (10b)$$

where, u is the trajectory of the wheel over the vertical fault in the longitudinal profile of the rail.

IV. Solving the 2nd Order Differential Equation

If we apply the Fourier transform to equation (8a) (see [11] for solving second order differential equations with the Fourier transform):

$$(i\omega)^2\cdot Z(\omega)+\frac{\Gamma\cdot(i\omega)}{m_{NSM}}\cdot Z(\omega)+\frac{h_{TRACK}}{m_{NSM}}\cdot Z(\omega)= \quad (11)$$

$$=-(i\omega)^2\cdot N(\omega)\Rightarrow$$

$$H(\omega)=\frac{Z(\omega)}{N(\omega)},\qquad |H(\omega)|^2=\frac{m_{NSM}^2\cdot\omega^4}{\left(m_{NSM}\cdot\omega^2-h_{TRACK}\right)^2+\Gamma^2\cdot\omega^2} \quad (12)$$

H(ω) is a complex transfer function, called frequency response function [11], that makes it possible to pass from the fault n to the subsidence Z. If we apply the Fourier transform to equation (10a):

$$(i\omega)^2\cdot U(\omega)+\Gamma\cdot(i\omega)\cdot Z(\omega)+h_{TRACK}\cdot(i\omega)^0\cdot Z(\omega)=0\Rightarrow$$

$$G(\omega)=\frac{U(\omega)}{Z(\omega)},|G(\omega)|^2=\frac{h_{TRACK}^2+\Gamma^2\cdot\omega^2}{m_{NSM}^2\cdot\omega^4} \quad (13)$$

G(ω) is a complex transfer function, the frequency response function, that makes it possible to pass from Z to Z+n.
If we name U the Fourier transform of u, N the Fourier transform of n, p=2πiν=iω the variable of frequency and ΔQ̂ the Fourier transform of ΔQ and apply the Fourier transform at equation (10b):

$$\left(m_{NSM}\cdot p^2+\Gamma\cdot p+h_{TRACK}\right)\cdot U=\left(\Gamma\cdot p+h_{TRACK}\right)\cdot N\Rightarrow$$

$$U(\omega)=\underbrace{\frac{\Gamma\cdot p+h_{TRACK}}{m_{NSM}\cdot p^2+\Gamma\cdot p+h_{TRACK}}}_{B(\omega)}\cdot N(\omega) \quad (14)$$

$$|B(\omega)|^2=\frac{\Gamma^2\cdot\omega^2+h_{TRACK}^2}{\left(m_{NSM}\cdot\omega^2-h\right)^2+\Gamma^2\cdot\omega^2} \quad (15)$$

B(ω) is a complex transfer function, the frequency response function, that makes it possible to pass from the fault n to the u=n+z. Practically it is verified also by the equation:

$$|B(\omega)|^2=|H(\omega)|^2\cdot|G(\omega)|^2=\frac{h_{TRACK}^2+\Gamma^2\cdot\omega^2}{\left(m_{NSM}\cdot\omega^2-h_{TRACK}\right)^2+\Gamma^2\cdot\omega^2} \quad (16)$$

passing from n to Z through H(ω) and afterwards from Z to n+Z through G(ω). This is a formula that characterizes the transfer function between the wheel trajectory and the fault in the longitudinal level and enables, thereafter, the calculation of the transfer function between the dynamic load and the track defect (fault). The transfer function of the second derivative of (Z+n) in relation to time:

$$\frac{d^2(Z+n)}{dt^2} \quad (17)$$

that is the acceleration γ, will be calculated below (and is equal to ω·B(ω)). The increase of the vertical load on the track due to the Non Suspended Masses, according to the principle **force = mass x acceleration**, is given by:

ISBN: 978-1-61804-256-9

$$\Delta Q = m_{NSM} \cdot \frac{d^2 u}{dt^2} = m_{NSM} \cdot \frac{d^2 (n+Z)}{dt^2} \quad (18)$$

If we apply the Fourier transform to equation (18):

$$\hat{\Delta} Q = m_{NSM} \cdot p^2 \cdot U(\omega) = m_{NSM} \cdot p^2 \cdot \hat{f}_{Z+n}(\omega) \Rightarrow \quad (19)$$

$$\left|\hat{\Delta} Q\right| = m_{NSM} \cdot |p|^2 \cdot \left|B(\omega)\right| = m_{NSM} \cdot \beta^2 \cdot \omega_n^2 \cdot \left|B(\omega)\right| \cdot \left|N(\omega)\right| \quad (20)$$

where $p=2\pi i v = i\omega$.

The transfer function B(ω) allows us to calculate the effect of a spectrum of sinusoidal faults, like the undulatory wear on the rail ruuning surface/table. If we replace $\omega/\omega_n=\rho$, where ω_n= the circular eigenfrequency (or natural cyclic frequency) of the oscillation, and:

$$\omega_n^2 = \frac{h_{TRACK}}{m_{NSM}}, \quad \omega = \frac{2\pi V}{\lambda}, \quad 2\zeta\omega_n = \frac{\Gamma}{m_{NSM}}, \quad \beta = \frac{\omega}{\omega_n} \quad (21)$$

where ζ is the damping coefficient. Equation (10b) is transformed:

$$\left|B(\omega)\right|^2 = \left|B_n(\beta)\right|^2 = \frac{1+4\zeta^2 \cdot \beta^2}{\left(1-\beta^2\right)^2 + 4\zeta^2 \cdot \beta^2} \quad (22)$$

V. The Specific Case of Defects of Short-Wavelength on the Rail

The transfer function, let's name it G, of the:

$$\frac{d^2 (Z+\eta)}{dt^2} \quad (23)$$

will be such that $B_1=\omega^2 \cdot B$, *in absolute value*, that is:

$$\left|B_1(\omega)\right| = \omega_n^2 \cdot \beta^2 \cdot \left|B(\omega)\right| \quad (24)$$

and the dynamic component of the load derived will be:

$$\left|\Delta Q(\rho)\right| = m_{NSM} \cdot \omega_n^2 \cdot \beta^2 \cdot \left|B(\omega)\right| = h_{TRACK} \cdot \beta^2 \cdot \left|B(\omega)\right| \quad (25)$$

The equation (25) permits to calculate the influence of one spectrum of consecutive "channels" on the railhead (rail running table), that is defects of short-wavelength, like, for example, the undulatory wear.

VI. Conclusions

In the present paper the dynamic component of the acting load on a railway track is calculated through the solution second order differential equation of the railway track for the case of defects of short-wavelength, like the undulatory wear of the rail running table.

References

[1] Giannakos, K., Actions on the Railway Track, Papazissis publications, Athens, www.papazisi.gr, (2004).

[2] Giannakos, K., Loizos, A., Evaluation of actions on concrete sleepers as design loads – Influence of fastenings, International Journal of Pavement Engineering (IJPE), Vol. 11, Issue 3, June 2010, p. 197 – 213.

[3] Giannakos, K., Track Defects and the Dynamic Loads due to Non-Suspended Masses of Railway Vehicles, International Journal of Mechanics, Issue 3, Volume 7, p. 180-191, (2013).

[4] Giannakos, K., Actions on a Railway Track, due to an Isolated Defect, International Journal of Control and Automation, Vol.7, No.3, p.195-212, 2014.

[5] Giannakos, K., Railway Track: The Transient Solution of the Second Order Differential Equation of Motion and the Acting Loads, in the 2nd International Conference on Mathematical, Computational and Statistical Sciences (MCSS '14), Gdansk, Poland, May 15-17, proceedings, p. 357-366, (2014).

[6] Giannakos, K., Theoretical calculation of the track-mass in the motion of unsprung masses in relation to track dynamic stiffness and damping, International Journal of Pavement Engineering (IJPE) - Special Rail Issue High-Speed Railway Infrastructure: Recent Developments and Performance, volume 11, number 4, p. 319-330, (2010).

[7] Alias, J., La Voie Ferree – Techniques de Construction et Entretien, deuxieme edition, Eyrolles, Paris, (1984).

[8] Fortin, J., La Deformee Dynamique de la Voie Ferree, RGCF, 02/1982.

[9] Thompson, D., Railway Noise and Vibration, Elsevier, (2009).

[10] Gent, I., Janin, G., La Qualite de la Voie Ferree, SNCF's reprint.

[11] Wylie, C. R., Barrett, L.C., Advanced Engineering Mathematics, sixth edition, McGraw-Hill, Inc., USA, (1995).

Technical inspection of remote power supply systems for microgrid development

Stanislav A. Eroshenko, Vladislav O. Samoylenko, Alexander O. Egorov, Pavel. V. Kolobov, Darina A. Firsova, Ekaterina M. Eroshenko

***Abstract*—** the paper provides extended discussion about the procedure of remote areas' electrical supply systems technical inspection, made in an effort to reveal energy quality and reliability problems of power supply systems. Technical inspection is carried out in order to determine prospective alternatives of existing power supply system development. The paper focuses on the possibility of distributed generation implementation for the purpose of power quality and supply reliability improvement in terms of remote territory, located in Far North region.

***Keywords*—** technical inspection, remote areas, power quality, reliability of power supply, distributed generation.

I. Introduction

THE main tasks set before power grid companies are to provide end users with electric energy of required quality, to reduce long line power losses, to optimize loading of 6-10 kV distribution network [1], to improve consumers electricity supply reliability in case of line fault.

The problem dealing with reliable supply of electric energy with required qualitative parameters is of great importance for grid companies, providing power supply for remote rural users [2]-[5]. Taking into account that 6-10 kV feeders can be considerably long (several tens of kilometers), the voltage may reach unacceptably low levels during peak load hours [3]. Moreover, the power of new consumers to be connected (remote from main 110(35) kV substation) is strictly limited owing to insufficient carrying capacity of existing grids.

Consequently, there are several problems, regarding remote rural areas power supply [4], namely:

1) Impermissible voltage reduction at the consumer side.
2) Intolerably high 6-10 kV feeders' technical power losses.
3) Low reliability due to the lack of backup systems.

In order to solve above-mentioned problems the next alternatives are considered: to develop existing network infrastructure or to put into operation distributed generation unit, providing reliability and quality of power supply [5]. The construction of hybrid power supply system is frequently considered as an alternate plan. The very system combines generators, running on fossil fuel, and/or renewable energy sources and storage systems.

However, it is necessary to carry out complex technical inspection in order to choose the plan of power supply system development in the given location [3].

The methodology of technical inspection of remote customers' power supply system is provided in the paper. Much attention is paid to the metering experiment, including data analysis and processing.

The case study provides deep analysis of the results, obtained during technical inspection of remote power supply system, located in northern Ural.

II. Technical inspection procedure

According to the methodology, provided in the paper, complex technical inspection of remote power supply systems is multistage [6]. Firstly, it is necessary to evaluate existing supply system, namely:

1) to evaluate the structure of energy consumption;
2) to analyze external power supply system;
3) to evaluate power supply system modes of operation.

This stage is mostly composed of initial data gathering and analyzing. This information can be found in power utilities' reporting documents. Generally, the initial information includes energy consumption data, electrical supply system data, distribution lines and power equipment data, the information about energy consumption structure and main electrical load units [7].

The analysis of listed above data gives the possibility to estimate energy consumption dynamics and reveal distinctive features of power supply system's operation modes.

At the next stage, it is necessary to make a set of measurements in the given power supply system, including load curves and power quality parameters identification. Special consideration must be given to the number of measuring devices and time of a single measurement interval.

It is necessary to evaluate:

1) power supply center operation mode;
2) operation mode of the backbone network;
3) distribution substations' operation modes;
4) loading conditions of the power equipment;
5) power quality parameters;
6) statistic parameters of power system operation.

To determine prospective plan of power supply system

This work was supported by the Ural Federal university.

S.A. Eroshenko (phone: +7(312)0333335, fax: +7(343)359-16-15, e-mail: stas_ersh@mail.ru), V.O. Samoylenko (e-mail: vvsamoylenko@yandex.ru), A.O. Egorov (e-mail: hiperboreya@yandex.ru), P.V.Kolobov (e-mail: pkolobov1993@gmail.com), D.A. Firsova (e-mail: dashka_firs@list.ru), E.M. Eroshenko (e-mail: trenina_katerina@mail.ru) are with Ural Federal University named after the first President of Russia B. N. Yeltsin, Ekaterinburg, Mira Street, 19, Russian Federation.

development with account of energy quality and reliability requirements, as well as to carry out investment feasibility study for distributed generation implementation, it is essential to perform permanent monitoring of power supply center (main substation) and backbone network modes of operation.

Local (internal) measurements are necessary to develop recommendations for the given power supply system. At the same time the overall clear picture of power supply system mode of operation is not provided with these measurements.

The measurements are to be carried out using power quality analyzing devices and are to be made in accordance with national or international standards [8-9]. For example, in Europe energy quality indices are regulated by EU 50160 "Voltage characteristics of electricity supplied by public" [8].

In this work, the experiments were conducted in conformance with national standard [9]. The voltage, phase currents and power parameters, controlled within minute intervals, included:

- line and phase voltages;
- zero-phase-sequence voltages and negative-sequence voltages;
- voltage unbalance ratio;
- voltage oscillograms, harmonic distortion;
- voltage nonsinusoidality ratio;
- phase currents;
- zero-phase-sequence current and negative-sequence current ;
- phase current unbalance;
- current oscillograms and harmonic distortion;
- current unsinusoidality ratio;
- active and reactive power, electric power;
- power factor.

Additionally, the statistic parameters and dynamic changes of observed values were estimated too.

III. The Experiment Description

The remote territory under consideration is supplied by 10 kV overhead distribution line (Fig. 1). The line extends from 10 kV bus section located at 35/10 kV main substation to sectioning point with the overall length of 56 km. The distribution line is made of AS-50 mm^2 wire (aluminum-steel).

There are six single-transformer distribution substations at the given location: №1232/100 kVA, №1254/250 kVA, №1223/170 kVA, №1244/160 kVA, №1252/160 kVA, №135/100 kVA. The total transformer power is 930 kVA, including consumer substation №135/100 kVA.

Speaking about reliability, there are I-category consumers at the given territory: fire station, rural health post, kindergarten. They are to be supplied with two or more independent power sources. The great deal of electrical load is domestic.

Technical inspection of electrical equipment, located at the very settlement includes:

1. Visual examination of 10 kV distribution feeder at 35/10 kV substation, 10/0.4 kV distribution substations and 0.4-10 kV internal networks of remote territory under consideration in order to control technical construction parameters of the power transmission.

2. Electrical measurements of load curves and power quality parameters.

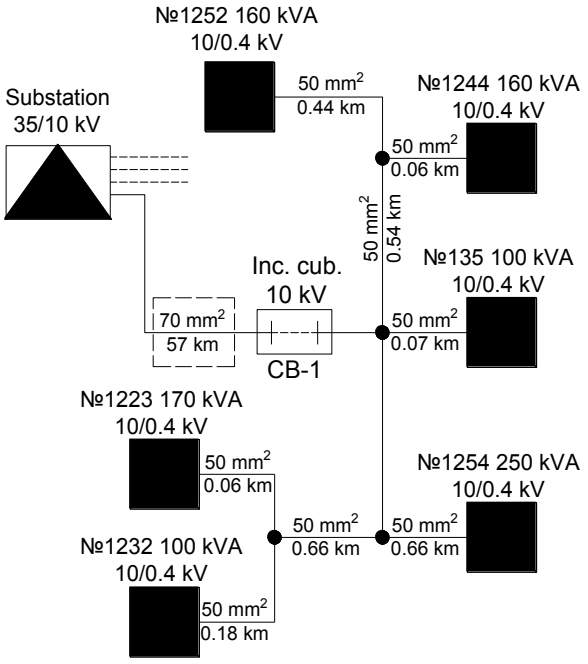


Fig. 1 Case study 35/10 kV network

The measuring points, duration and purposes of the measurements for each point are listed in Table 1.

Within the framework 10/0.4 kV №135 measurements were not carried out because of its weak influence on total load curve and close location to 10 kV sectionalizing point.

TABLE 1
MEASUREMENT ALLOCATION

№	Measurement point	Time	Obtained data
1	35/10 kV substation, 10 kV line cubicle, secondary circuit 0.1 kV	5 days	- 10 kV bus section parameters; - load curve of «line-settlement» system; - parameter variation, curve characteristics.
2	10 kV sectionalizing point, 0.22 kV relay protection circuits	1 day	- 10 kV supply system "input" parameters; - 10 kV line operation parameters; - 10 kV network parameters.
3	10/0.4 kV substation №1254, 0.4 kV transf. bushing	1 day	- load curve; - 0.4 kV network parameters
4	10/0.4 kV substation №1223, 0.4 kV transf. bushing	1 day	- load curve; - 0.4 kV network parameters
5	10/0.4 kV substation №1232, 0.4 kV transf. bushing	1 day	- load curve; - 0.4 kV network parameters
6	10/0.4 kV substation №1244, 0.4 kV transf. bushing	2 days	- load curve; - 0.4 kV network parameters
7	10/0.4 kV substation №1252, 0.4 kV transf. bushing	2 days	- load curve; - 0.4 kV network parameters
8	House, 0.22 kV in-residential network	1 day	- operation mode parameters "at the customers side"

ISBN: 978-1-61804-256-9

IV. Measured results

A. Main substation mode

The active and reactive power curves for 10 kV feeder are illustrated in the Fig. 2. The list of the consumers forming active power curve is the following:

- lighting load;
- household appliances: satellite equipment, computers;
- small heating devices (electric kettles, irons);
- big heating devices (boilers, electric ovens);
- induction motors: machine tools, electric saws, pumps, water-towers, fridges;
- cellular communication stations;
- distribution line losses and contact losses.

The list of the consumers, forming reactive power curve is the following:

- power units and chargers for modern household appliances;
- induction motors: machine tools, electric saws, pumps, water-towers;
- welders;
- cellular communication stations and signal retransmitters;
- distribution line losses and transformer losses.

Lighting and distributed low-rated power devices tend to predominate in the loading structure. They have active consumption characteristics and cause daily peak in the morning on business days. However, the weekly peak loading conditions occur in the evening on weekends. The curve of reactive power consumption is uniform with single surges, caused by electric motors starting.

The daily average active power equals 109.8 kW, reactive power – 39.6 kVAr, total power – 117.0 kVA. The active power ranges within 65.9 – 183.7 kW.

The line voltage profile is presented in Fig. 3. The line voltages vary from 10.23 kV to 10.81 kV. The daily average line voltage equals to 10.53 kV. The voltages are symmetrical.

In general, the loads characteristics are uniform for all phases. The average asymmetry is about 10 %. It is caused by unbalanced distribution of consumers' single-phase loads between different phases in 0.4 kV network.

B. Backbone network operation mode

In order to estimate operation mode of the backbone 10 kV network the additional measurements have been made at the end of 10 kV line in secondary circuits of 10 kV sectionalizing point.

Having no access to the current transformers' and voltage transformers' secondary windings, the current measurements were carried out in control cabinet of sectionalizing point. The clamp meters were connected to internal circuits of two current relays in two phases. The phase voltage was measured at the terminals of the automatic breaker of control cabinet.

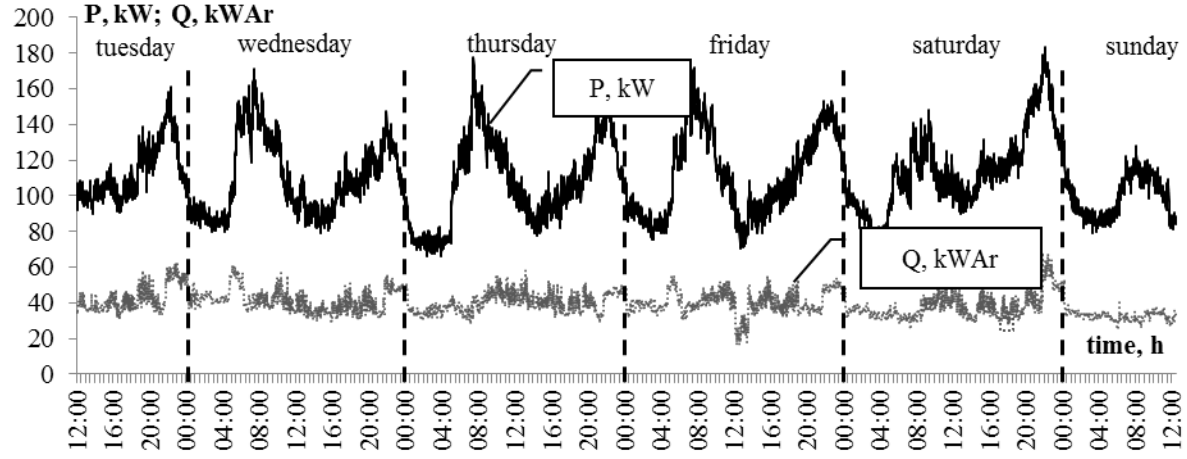


Fig. 2 Active and reactive power flow curves at the beginning of 10 kV feeder

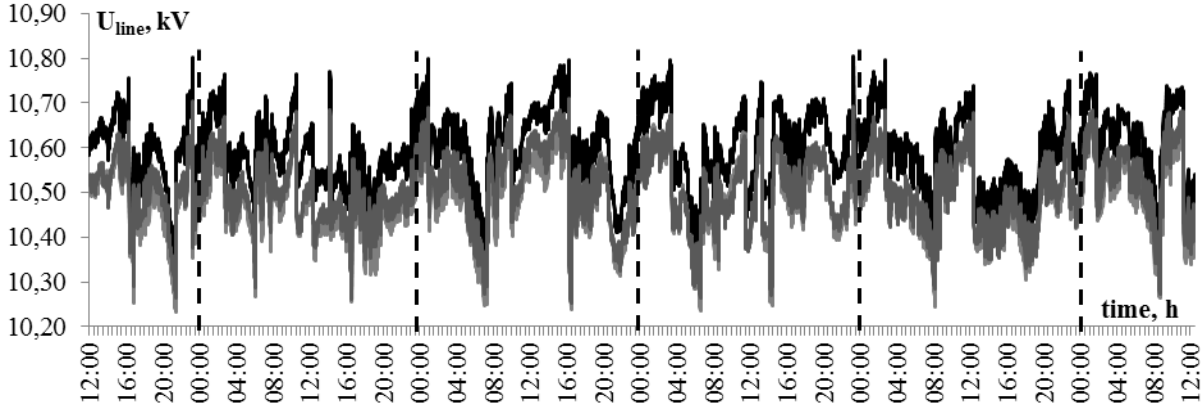


Fig. 3 Line voltages curves at the beginning of 10 kV feeder

ISBN: 978-1-61804-256-9

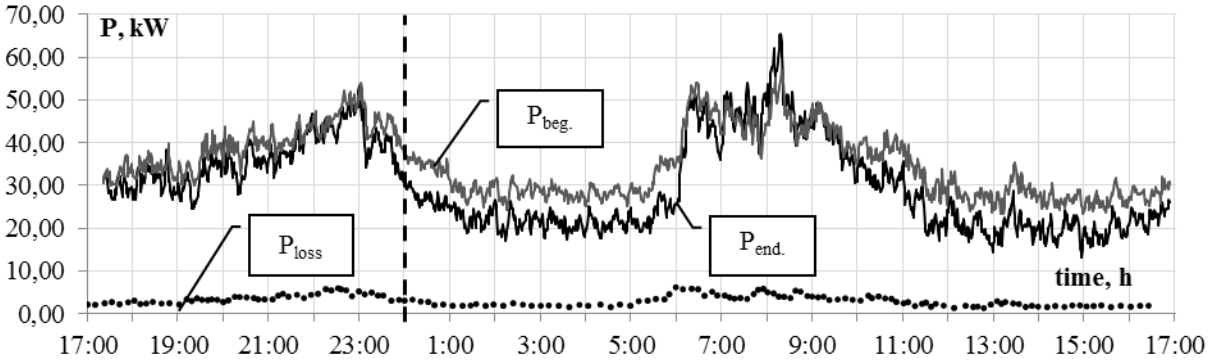


Fig. 4 Active and reactive power flow curves, measured at 10 kV feeder terminals

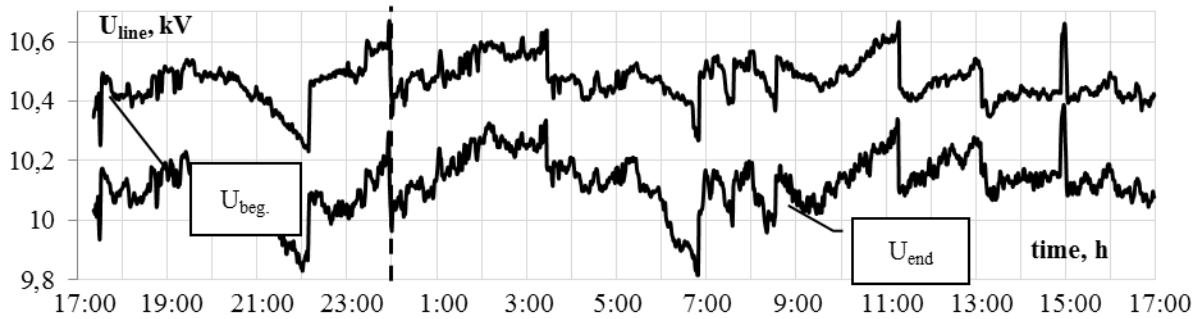


Fig. 5 Line voltages, measured at 10 kV feeder terminals

As a result, it can be stated that the line voltage, measured at 10 kV sectionalizing point, lies in the range of 9.8 – 10.4 kV, that meets voltage magnitude requirements (Fig. 4). The daily average voltage levels are 10.47 kV for distribution line sending end and 10.12 kV – for receiving end.

In this way, the average voltage drop across the given 10 kV line is about 0.35 kV. The length of the line is 56 km and the loading factor of the line is nearly 0.1.

The relation between measured power flows in the beginning and in the ending of the given 10 kV line (Fig. 5) arises from current and voltage transformers inaccuracy. This inaccuracy mainly occurs at the sectionalizing point where there is no energy accounting circuits provided. Current transformer's secondary windings have 10R accuracy rating, which corresponds to 10% error.

Another source of inaccuracy at the sectionalizing point is the configuration of relay protection circuits working on a.c.; such circuits contain a great number of non-linear elements.

The power losses are to be subdivided on three main types, namely: series losses (load losses), shunt losses (leakage currents), contact losses, commercial losses (measurement inaccuracy, caused by current transformers low loading conditions) [10]-[12]. Load losses, which were calculated in accordance with measured data, equal to 8.6%. Leakage currents are about 1%. Moreover, it was estimated that commercial losses and contact losses are about 4.5%. In this way, the total line losses equal to 14.1%.

C. Internal power supply system operation mode

The operation modes of 10/0.4 kV grid of internal power supply system of the remote territory will be further described using measurements made at distribution substation №1254. Final conclusions, made for distribution substation №1254, are suitable for other distribution substations too.

Low-rated household devices predominate in load structure of the given substation. This loads result in morning peak loads. In addition, there is motor load with approximately 12 kW active power consumption. Active and reactive power curves are given in Fig. 6.

Phase voltages range from 226 V to 245 V, that exceeds voltage level standard requirements. The daily average phase voltage is 235 V. The graph illustrating phase voltages is represented in Fig. 7.

The phase loads are not uniform. Phase asymmetry results in rapid deterioration of transformer isolation and worsening of electric power quality in each phase.

V. CONCLUSION

The paper presents methodology of electrical measurements for remote power facilities. The analysis of residential consumers operating conditions and energy quality rates is made to develop recommendations for power quality and reliability improvement. The main statistical rates of remote territory operation are provided in the Table 2.

The 10 kV power supply network of the given remote settlement provides electrical energy quality in compliance with [8-9] and operates in a symmetrical three-phase mode. Taking into account the overall length of the overhead 10 kV distribution line, which equals to 56 km, the level of the power losses in the line is satisfactory.

ISBN: 978-1-61804-256-9

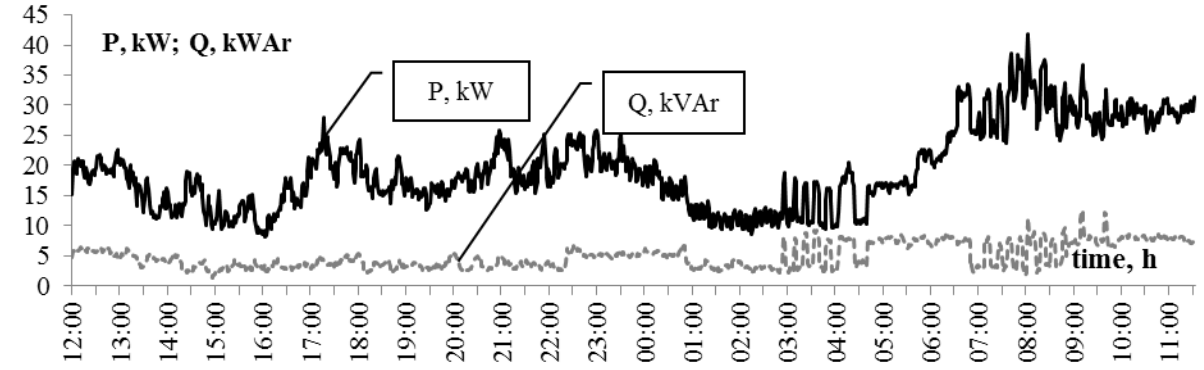


Fig. 6 Distribution substation load curve

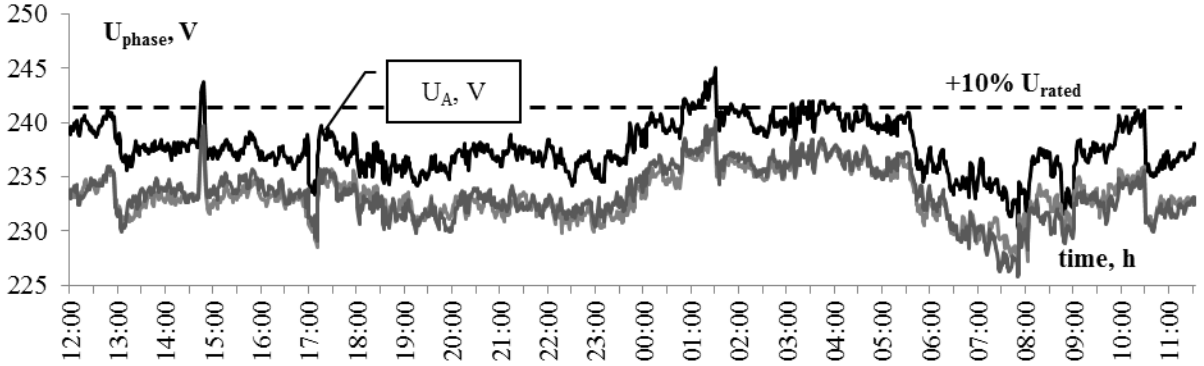


Fig. 7 Distribution substation voltage profile

TABLE 2
POWER SUPPLY PROBLEMS

Parameter	Value	Property
Customer side voltage (0.22 kV network)		
Minimal, V	222,0	High
Maximal, V	251,3	**Inadmissible**
Average, V	236,3	*Admissible*
10 kV line losses		
Load losses, %	8,6	*Admissible*
leakage currents, %	1,0	Low
Contact losses and commercial losses, %	4,4	*Admissible*
Total losses, %	14,0	*Admissible*
Load curve uniformity		
Minimal α, %	18,3	**Low**
Average α, %	20,7	*Admissible*
Load current asymmetry		
Maximal I_2 , %	58,2	**Inadmissible**
Average I_2 , %	35,6	**Inadmissible**

It is important to note that the increased voltage level is observed due to low loading conditions and 10/0.4 kV transformers NLTC systems, switched to winter position. *The latter represents atypical problem of remote consumers power supply.*

It was also observed that load phase balance is inadmissible. Thus, it is recommended to reconnect a share of load to other phases.

The main challenge of remote customers power supply is low reliability. In case of distribution line fault, consumers are not supplied within the repair time period.

In this case, distribution generator implementation becomes the question of growing importance. Basing on the metering data the capacity of the generator, which will be installed at the remote settlement, is assessed to be 500 kVA taking into account winter peak loads.

ACKNOWLEDGMENT

The authors express their gratitude for being supported by Ural Federal University. The authors appreciate the contribution of Open Joint-Stock Company «Interregional Distributive Grid Company of Urals» dealing with assistance in measurements made at 10 kV power facilities within research and development work.

REFERENCES

[1] Kokin, S.; Dmitriev, S.; Khalyasmaa, A., "Assessment of state of urban power supply systems' power transmission lines on the basis of indicative analysis", Applied Mechanics and Materials 291-294,2013, pp. 2143-2148.

[2] Zomers, A., "Remote Access: Context, Challenges, and Obstacles in Rural Electrification" Power and Energy Magazine, IEEE (Volume:12 , Issue: 4), pp. 26–34.

[3] Verwers, J.L.; Sovers, J.R., "Challenges of supplying electric power to a large industrial customer in rural areas" Industry Applications, IEEE Transactions on (Volume:36 , Issue: 4), pp. 972 - 977.

[4] Rudnick, H. ; Mutale, J. ; Chattopadhyay, D. ; Saint, R., "Studies in Empowerment: Approaches to Rural Electrification Worldwide", Power and Energy Magazine, IEEE (Volume:12 , Issue: 4), pp. 35 – 41.
[5] Ijumba, N.M., "Application of distributed generation in optimised design and operation of rural power supply networks", Rural Electric Power Conference, 1999., pp. C3/1 - C3/5.
[6] The methodology of power inspection (energy audit) performance at enterprises and coal utilities (proved by protocol arranged at coordinating council meeting at Russian Energy Department from 29.05.2012 № 6), 2012. (in Russian)
[7] Yong Li ; Jian-Jun Wang ; Tie-Liu Jiang ; Bing-Wen Zhang, "Energy Audit and Its Application in Coal-Fired Power Plant", International Conference on Management and Service Science, 2009. MASS '09, 20-22 Sept. 2009, pp. 1-4.
[8] European Standard EN 50160, Voltage characteristics of electricity supplied by public distribution systems, 1999
[9] International standard GOST 13109-97 "Electric power. Electromagnetic equipment compatibility. Quality standards of electric power in common power supply systems" (introduced by State standard 28.08.1998. N 338), 1998. (in Russian)
[10] Khalyasmaa, A.I.; Dmitriev, S.A.; Kokin, S.E., "Energy information model for power systems monitoring", Advanced Materials Research 732-733, 2013, pp. 841-847.
[11] Pazderin, A.V.; Samoylenko, V.O., "Localization of non-technical energy losses based on the energy flow problem solution", Proceedings of the 6th IASTED Asian Conference on Power and Energy Systems, AsiaPES 2013, pp. 100-103.
[12] Khalyasmaa, A.I.; Dmitriev, S.A., "Power equipment technical state assessment principles". Applied Mechanics and Materials 492, 2014, pp. 531-535.

A Game-Theoretic Analysis of the Nuclear Non-Proliferation Treaty

Peter Z. Revesz
Department of Computer Science and Engineering
University of Nebraska-Lincoln
Lincoln, Nebraska 68588-0115
Email: revesz@cse.unl.edu
http://cse.unl.edu/revesz
Telephone: (1+) 402 472–3488

***Abstract*—Although nuclear non-proliferation is an almost universal human desire, in practice, the negotiated treaties appear unable to prevent the steady growth of the number of states that have nuclear weapons. We propose a computational model for understanding the complex issues behind nuclear arms negotiations, the motivations of various states to enter a nuclear weapons program and the ways to diffuse crisis situations.**

I. Introduction

Numerous international treaties are made with the best of intentions. However, every treaty needs to be examined on its actual affects rather than on its intentions. The *Treaty on the Non-Proliferation of Nuclear Weapons*, commonly referred to as the *Non-Proliferation Treaty (NPT)* aimed to make the world more secure from nuclear weapons. The treaty divided all countries based on their nuclear status as of January 1, 1967, into nuclear weapon states (NWSs), which included China, France, the Soviet Union, the United Kingdom, and the United States, and non-nuclear weapon states (NNWSs), which included all the other states. All the NWSs signed the treaty as well as all the NNWSs except India, Israel and Pakistan. North Korea is the only country that withdrew from the treaty. Hence the NPT enjoyed a great popularity and is often considered a great success.

The essence of the NPT is a bargain between the NWSs and the NNWSs. The NWSs committed themselves to nuclear disarmament and to help the NNWSs to develop civilian use of nuclear technology. In return, the NNWSs committed themselves to foresake ever developing nuclear weapons. Unfortunately, this bargain did not work out as planned. After forty years, the NWSs increased the total number of their nuclear weapons, while many NNWSs engaged in clandestine nuclear weapon development programs. The world does not look safer than it was forty years ago. Nevertheless, NPT defenders claim that the NPT slowed down nuclear proliferation. In other words, without the NPT, nuclear proliferation would have been even worse than it is actually today. In this paper we examine this hypothetical claim using game theory. We start our analysis with some definitions.

Uranium enrichment is the process of dividing any uranium compound into two parts, one part with a higher and another part with a lower concentration of U 235 atoms. Uranium ore has a very low percent of U 235 atoms. Most nuclear reactors can work on *low enriched uranium (LEU)*, where the proportion of U 235 is less than 20 percent. Nuclear bombs require *highly enriched uranium (HEU)*, where the proportion of U 235 is greater than 80 percent. The uranium enrichment technology is the same for LEU and for HEU. To obtain HEU, the uranium enrichment process simply needs to be repeated several times until the desired level is reached.

Plutonium reposessing is the process of separating the plutonium, a byproduct of uranium fission, from the rest of the spent fuel in an uranium atomic reactor. The plutonium can be used either as fuel for plutonium atomic reactors or as material for plutonium atomic bombs.

Dual-use technology is any technology that can be used for both civilian or military purposes. For example, uranium enrichment and plutonium reposessing are both dual-use technologies.

The NPT allows any NNWS to aquire and develop any dual-use nuclear technology. Moreover, citing the NPT, many NNWSs expect the NWSs to provide assistance in aquiring dual-use technologies including uranium enrichment and plutonium reposessing. When a NNWS aquires these technologies, it essentially develops 80 percent of an atomic bomb because civilian and military nuclear technologies largely overlap. Such a NNWS could be tempted to invest the 20 percent extra effort required to develop an atomic bomb. Hence any of its adversaries may become concerned whether it will decide to develop a bomb. Moreover, these adversaries need to be prepared for all eventuality. That means that these adversaries also need to build up their NPT-allowed dual-use nuclear technologies and be ready to activate a nuclear weapons program of their own just in case any of their adversary NNWSs decides to build a nuclear weapon. This leads to a situation, which we define as follows.

Soft arms race occurs when states develop nuclear-related dual-use technologies with the intent to be strategically prepared to develop nuclear weapons.

ISBN: 978-1-61804-256-9

Several experts are concerned about a soft arms race in the Middle East and North Africa, where many energy rich states insist that they need to develop peaceful nuclear reactors. Developing nuclear technology is expensive, and most of these countries would not have been able to aquire any nuclear technology without direct or indirect assistance from NWSs. Hence the question can be raised whether the NPT contributed to a soft arms race regarding nuclear technology. Further, if there is a soft arms race, how likely it is to lead to an active nuclear weapons program? We try to answer these difficult questions using game theory, and thereby contributing to the theoretical study of nuclear proliferation [2], [4], [8].

This paper is organized as follows. Section II briefly reviews game theory and the history of its use for analyzing nuclear issues. Section III describes a game theoretic analysis of the NPT. Section IV gives a game theoretic analysis of what may happen in a world without the NPT. Finally, Section V gives some conclusions and offers some hope of improving the current nuclear non-proliferation situation.

II. A Review of Game Theory

During the Cold War, game theory was a reasonable approach to arms control negotiations because nuclear tests and total arsenal numbers were hard to verify. Virtually the only thing that could be detected was an already approaching*intercontinental ballistic missile (ICBM)*. There was not enough time and technological sophistication to shield against nuclear ICBM strikes. Therefore, in case of a nuclear attack, each side faced the choice between continued restraint or nuclear retaliation. Table I shows the nuclear options of Russia and the United States during the Cold War expressed in a hypothetical payoff matrix using game theory [9]. The table assumes that it would cost each side 20 points to be destroyed in a nuclear attack. However, if any side is destroyed, at least it can derive a satisfaction of five points by retaliating and destroying the other side too.

Russia ↓ US →	**no strike**	**first strike**	**retaliation**
no strike	*0, 0*	-20, 0*	NA
first strike	*0, -20	-20, -20	*-20, -15*
retaliation	NA	*-15, -20*	NA

TABLE I: A hypothetical payoff matrix during the Cold War.

Clearly, some entries in the table, shown as NA, are not available or logically impossible. For example, it is not possible to retaliate against something that did not happen. Even the case of both countries deciding on a first strike simultaneously would have an extremely small possibility. In this example, game theory gives three *Nash equilibrium points* [3], which are shown as the matrix entries with two stars, that is, one star on the left and another star on the right of the entry. In this case, the rational choice would be *0,0*, which is the best equilibrium point for both sides. This is the game theoretic explanation for how the *mutually assured destruction (MAD)* nuclear posture worked during the Cold War.

The idea behind MAD is that if one side attacks, then it will get destroyed. That is supposed to be the ultimate deterrence. However, for it to work the leaders with access to the nuclear triggers have to be non-delusional and non-suicidal (otherwise, the payoff matrix values could change.) Unfortunately, that cannot be guaranteed. Today there is an increasing danger that not only possible delusional dictators but also terrorist chiefs and suicide bombers may gain access to nuclear weapons.

The success of MAD also depended on maintaining a retaliatory capability because MAD would be impossible if either side could make a first strike that debilitates all the nuclear weapons of the other side. This aspect of MAD tends to lead to an arms race as both sides feel that they need some extra (numerous and/or advanced) weapons to successfully deter the other side.

Russia ↓ US →	**no strike**	**first strike**	**retaliation**
no strike	*0, 0*	-20, 0*	NA
first strike	*0, -20*	-20, -20*	NA
retaliation	NA	*-15, -20*	NA

TABLE II: Modified payoff matrix in case Russia would gain completely debilitating first-strike capability.

To illustrate this last point, Table II shows the changed cost matrix in case Russia could attain such a first strike capability. Here the -20,-15 outcome would no longer be available, and *0,-20* would be a new equilibrium point. Russia would prefer the two equilibria *0,0* and *0,-20* to the third equilibrium *-15,-20*. However, the first two equilibria would be extremely unnerving to the U.S. population. This situation is symmetric. Hence both sides need to maintain a retaliatory capability as a credible deterrent. To maintain a retaliatory capability, both sides kept secret the locations of their nuclear weapons and increased the number of their nuclear warheads to very high levels, leading to a nuclear arms race. Hence Table II is a game theoretic explanation of the nuclear arms race during the Cold War.

In summary, game theory provides insights for cases when there is little or no trust between the participants. Since neither side can trust the other side, they need to play safe first and foremost. Game theory fails to account for trust among the partners in negotiations. Normally, people participate in negotiations because they trust that their partners will keep the agreements, which can be enforced by verification procedures, courts, or the threat of breaking off a relationship. Game theory explains well the purely adversarial strategies but fails to provide a realistic model for negotiations [5], [6], [7].

III. A Game Theoreric Analysis of The NPT

In our analysis, we consider a set of variables shown in the first two columns of Table III. The exact values of these variables can be only estimated, which is something beyond the scope of this paper. However, it is only the relative strength of these variables that is important for our analysis. As shown in the third column, each variable can have either a single

ISBN: 978-1-61804-256-9

	None	Reactor	Bomb
NWS ally	*0, 0	*0, 2*	*−1, 1
NWS adversary	*0, 0	−7, 3	−9, 4*

TABLE V: Payoff matrix.

	None	Reactor	Bomb
NWS ally	*0, 0*	*0, −1	*−1, −4
NWS adversary	*0, 0*	−7, −1	−9, −2

TABLE VII: The revised payoff matrix.

number value, meaning that it is the same for all countries, or it can have two different values for allied and adversary countries, respectively. In the following, we index the variables by 1 for allies and 2 for adversaries if there are differences in values.

For each of the estimates, we provide some explanation in the fourth column of Table III. We assume extra trade benefit *etb* to the NWS states to be zero because the NWS countries were forbidden to sell weapons-related nuclear technology to other countries. This regulation restricted the market and the clandestine transactions that still occurred seem to have been done from political rather than from financial motivations [2].

Table IV shows a matrix where the last three columns describe the three choices of any NNWS: (1) build nothing, (2) build only peaceful nuclear reactors, and (3) build nuclear bombs too. The two rows of the matrix describe the two choices of any NWS. Each NWS could consider the NNWS as either an ally or an adversary. Alliances can shift over long periods of time due to strategic reasons.

Substituting the values in Table III for the variables in Table IV, we obtain Table V. Table V shows that there is a Nash equilibrium, again indicated by two stars, for any NWS and NNWS pair. The Nash equilibrium would mean that the two states would be allies and the NNWS would restrain itself to only a peaceful use of nuclear energy. In practice, this Nash equilibrium may not be reachable because states are locked into various alliances due to other considerations. Hence some countries are bound to remain NWS adversaries. Table V does not show any equilibrium for NWS adversaries. In fact, a NWS would rather have a NNWS adversary with no nuclear technology at all, while the NNWS would rather develop nuclear weapons. The NWS could be naturally suspicious about the peaceful intentions of any adversary NNWS. Hence the current NPT environment encourages peaceful development of nuclear energy among ally NNWSs. This leads to a soft arms race among the ally NNWSs and their adversaries.

IV. Analysis Without the NPT

Imagine a world without the NPT. How the absence of the NPT would effect the values of the variables listed in Table III? First, the development cost for civilian nuclear technology would increase in general. We estimate that for NWS adversaries the development cost may double to 6 as they would have to do essentially everything themselves or pay heavy prices for nuclear technology. NWS allies would also no longer get any free nuclear technology, although they may be able to buy some at a discount. Hence their development cost would increase to about 5.

When the price of civilian nuclear technology increases, the demand decreases. The price increase and demand decrease tend to cancel each other out, hence the trade benefit would not change drastically. We continue to assume that trade benefit is 1. With the increase of civilian nuclear technology, the price of military nuclear technology would also increase. Hence the extra development cost may increase from 2 to 3.

The decreased demand for civilian nuclear technology may prevent the development of the soft arms race in dual-use nuclear technology among the NNWSs. Therefore, the security benefit of civilian nuclear reactors decreases to about 1 for allies and 2 for adversaries. The extra security benefit would decrease to 0 for allies and 2 for adversaries. At the same time, the security cost and the extra security cost to NWSs would remain the same because the NWSs would be still be constrained and lose control over NNWSs that aquire nuclear technology.

To summarize the above discussion, Table VI lists all the variables whose values would be likely different in a world without the NPT.

Repeating now the game theoretic analysis with the new values as shown in Table VII reveals that the no NPT environment has a Nash equilibrium for both NWS allies and NWS adversaries. In both cases the equilibrium implies the choice of developing no nuclear technology.

V. Conclusion

We provided a game theoretic analysis of the choices of NNWSs regarding the use of nuclear technology. According to our estimates of the costs and benefits of certain strategies, it appears that without the NPT, all NNWSs states would choose no nuclear energy. On the other hand, with NPT the NNWS allies of NWSs would choose to develop only civilian nuclear energy, and the NNWS adversaries of NWSs would choose to go all the way to developing nuclear weapons.

Hence according to our analysis, the NPT seems to have made the world less secure by encouraging among the NNWSs a soft arms race of dual-use nuclear technology. Although only a few NNWSs would cross the threshold and later enter an outright nuclear arms race, their entry seems more likely because of the already present soft arms race.

These conclusions depend on the exact values of the costs and the benefits. Each state can have a particular situation which would mean that these values need to be adjusted. In addition, our game theoretic analysis did not include many other cultural, historical and political considerations that influence policy makers' decisions regarding the development of civilian or military nuclear technology. Hence we cannot

Name	Symbol	Value	Explanation
Energy benefit	eb	3	Similar reactors always yield similar amount of energy.
Trade benefit (NWS)	tb	1	Companies equally eager to sell to all.
Extra trade benefit (NWS)	etb	0	Prohibited. Negligible commercial motivation for violations.
Development cost	dc	3	Cost overruns are common in every country.
Extra development cost	edc	2	Construction costs only. Sanctions belong to *esb*.
Security cost (NWS)	sc	1, 8	Both allies and adversaries limit NWS countries' freedom.
Extra security cost (NWS)	esc	1, 2	However, allies are less dangerous.
Security benefit	sb	2, 3	Allies already have security guarantees from NWS.
Extra security benefit	esb	1, 3	Hence allies get a diminished return.

TABLE III: Variables used in the game theoretic analysis.

	None	Reactor	Bomb
NWS ally	0, 0	$tb - sc_1$, $eb + sb_1 - dc$	$tb + etb - sc_1 - esc_1$, $eb + sb_1 - dc + esb_1 - edc$
NWS adversary	0, 0	$tb - sc_2$, $eb + sb_2 - dc$	$tb + etb - sc_2 - esc_2$, $eb + sb_2 - dc + esb_2 - edc$

TABLE IV: The choices of any pair of nuclear weapon state (NWS) and non-nuclear weapon state (NNWS).

Name	Symbol	Value	Explanation
Development cost	dc	5, 6	Cost overruns are common in every country.
Extra development cost	edc	3	Construction costs only. Sanctions belong to *esb*.
Security benefit	sb	1, 2	Allies already have security guarantees from NWS.
Extra security benefit	esb	0, 2	Hence allies get a diminished return.

TABLE VI: Variables with changed values.

draw from our game theoretic analysis any firm conclusion about any particular state. Nevertheless, our game theoretic model suggests that the NPT may have affected the cost and benefit structure of the nuclear technology market, both overt and covert, in a way that encourages instead of discourages non-proliferation. This should raise a concern for the non-proliferation community. The NPT, like any other international treaty, should be evaluated by its actual affects instead of its professed intent. Although the intent of the NPT was to prevent proliferation, its actual affects may have been the opposite.

Our pessimistic analysis of the effects of the NPT, need not be the end of the story. Although it is unlikely that the NPT can be abandoned completely, there are some promising current suggestions by some nuclear non-proliferation experts. One proposal is to offer to replace free the older reactors that produce significant amounts of plutonium with newer *Liquid Fluoride Thorium Reactors (LFTRs)*, which allows for fuel utilization exceeding 99 percent and produces very little weapons grade material. Such a replacement offer may cut down on the temptation to repossess plutonium and use it or sell it to other states. We hope that continued arms control negotiations will lead to a solution that is both well-intentioned and mathematically sound.

ACKNOWLEDGMENT

This work was prompted in part by the author's experience in serving as a *Jefferson Science Fellow* in the U.S. Department of State's Bureau of International Security and Nonproliferation during 2010-2011, while on a leave from the University of Nebraska-Lincoln. The views expressed in this paper are those of the author and do not necessarily reflect the views of the US federal government or its agencies.

REFERENCES

[1] P. C. Kanellakis, G. M. Kuper and P. Z. Revesz, Constraint query languages, *Journal of Computer and System Sciences*, 51 (1), 26-52, 1995.
[2] M. Kroenig, *Exporting the Bomb: Technology Transfer and the Spread of Nuclear Weapons*, Cornell University Press, 2010.
[3] J. F. Nash Jr. The Bargaining problem, *Econometrica: Journal of the Econometric Society*, 18(2):155-162, 1950.
[4] W. Potter and G. Mukhatzhanova, eds., *Forecasting Nuclear Proliferation in the 21st Century: Volume 1 The Role of Theory*, Stanford Security Studies, 2010.
[5] P. Z. Revesz, On the Semantics of Arbitration, *International Journal of Algebra and Computation*, 7(2):133-160, 1997.
[6] P.Z. Revesz, Arbitration solutions to bargaining and game theory problems, *Annales Univ. Sci. Budapest., Sect. Comp.* vol. 43, pp. 2138, 2014.
[7] P. Z. Revesz, *Introduction to Databases: From Biological to Spatio-Temporal*, New York, USA: Springer, 2010.
[8] T. C. Schelling, *Arms and Influence*, 2nd edition, Yale University Press, 2008.
[9] J. Von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, 1947.

Optimal deployment of renewable energy sources considering ancillary services limitation

Andrea Zápotocká, Martin Střelec, and Petr Janeček

Abstract—Nowadays, power system operators (e.g. TSO) face to new phenomena which relate with increasing ratio of intermittent energy sources in power network. Massive installation of renewable energy sources causes significant and unprecedented increase of the uncertainty in power network, which results in new technical issues and challenges. Due to missing historical data and experience, determination of the reasonable ancillary services volume stands for one of the challenge, which has great economical and safety impact to the network operator. Because of limited ancillary services volume, optimal deployment of newly installed intermittent energy sources with respect to network safety criteria can constitute another challenge for power system operator. In this paper, we introduce two methods lead to overcome of these both challenges. First method focuses on the determination of maximal installable power generated by intermittent energy sources which can be absorbed by power network with fixed volume secondary control reserve. Second method computes the optimal deployment of intermittent energy sources in geographical areas and uses statistical properties of secondary control reserve, newly installed energy sources and their coincidence factors. Application of these methods is demonstrated on the case study, which consists in four selected scenarios. Particular aspects of introduced methods are assessed based on results from sensitivity analysis, which is described in the paper as well.

Keywords—intermittent energy sources, transmission and distribution network, ancillary services, constrained optimization

I. Introduction

UNCERTAINTY level increases in power networks mainly due to massive installation of intermittent energy sources (i.e. renewable energy sources), which stands for a new phenomenon faced by transmission and system operators (TSO). One of main TSO's objectives is to ensure safe and reliable operation of the transmission network, which mostly consists in balancing power generation and power consumption. Area control error (ACE) is an important stability performance indicator which is defined as the difference between the actual and the reference value for the power interchange of a Control Area [1]. For grid stability, ACE has to be kept within defined limits by using of ancillary services reserves (AS) which can be split into several categories [2], [3]. The AS categories can differ in particular countries [4]. In this paper, however, we consider a Central Europe categorization, which consists in primary frequency control, secondary power control, and minute reserve.

This work was supported by the European Regional Development Found (ERDF), project NTIS New Technologies for the Information Society, European Centre of Excellence, CZ.1.05/1.1.00/02.0090, Technology Agency of the Czech Republic under project BIOZE (TA01020865)

A. Zápotocká, P. Janeček, M. Střelec are with the NTIS – New Technologies for the Information Society, University of West Bohemia, Pilsen, Czech Republic, e-mail: {fialova, pjanecek}@kky.zcu.cz, strelec@ntis.zcu.cz.

Mostly, fluctuations in ACE caused by RES and power loads are eliminated by secondary control. Grid operator ensures the secondary control reserve (SCR) by long term contracts with ancillary service providers. Therefore, determination of suitable SCR volume has great economical impact to grid operator. In the unprecedented situation of massive installation of RES, historical measurements can be hardly used for the determination of the SCR volume which has been solved in [5].

In this paper, we focus on the situation when operator already contracted fixed SCR volume and has to connect new RES installations into the power network. Two challenges can be considered in this situation, which are described in the following section.

II. Problem Formulation

First challenge relates to the determination of maximal RES production which can be absorbed by the power network under consideration of limited SCR volume. The volume can be restricted by two reasons: (i) physically limits of the installed power equipment (power sources, transmission network) and (ii) fixed contracted SCR volume.

Second challenge stands for optimal deployment of newly installed RES into the controlled area. Global controlled area can be divided into several local areas which can have various coincidence factor of RES production. Coincidence factor of particular areas strongly affect absorption capability of RES production in global area.

In this paper, methods for overcome of these challenges are introduced, where the emphasis is put on second one.

III. Determination of maximal RES production under limited SCR

This section describes a method for determination of maximal RES production which can be absorb by power network under consideration of limited (i.e. fixed) SCR volume.

Suitable SCR volume can be calculated based on statistical properties of ACE_{OL} (i.e. mean value and standard deviation). These parameters include the influence of the volatilities of installed RES and power loads. Installation and integration of new RES into the power network change statistical properties of ACE_{OL}. The influence of the newly installed RES onto ACE_{OL} is assessed by change of prediction error of the RES production estimate in early stage after RES installations.

ISBN: 978-1-61804-256-9

A. Determination of SCR Volume

First, calculation of the SCR volume is described for the situation of undisturbed (normal) operation. As we noted above, the ACE value has to be kept within defined limits. These limits can be defined in the form of "Value at Risk" $VaR\,(P_{bounds}, r_P)$, where $r_P = 3.8\%$ is probability that ACE exceed $P_{bounds} = \pm 100\,\mathrm{MW}$. In this way, SCR can be calculated as

$$SCR^* = q_{ACE} \cdot \sigma^*_{ACE_{OL}} \tag{1}$$

where q_{ACE} is quantile $\frac{r_P}{2}$, $\sigma^*_{ACE_{OL}}$ is standard deviation of ACE_{OL} including newly installed RES. Resulting standard deviation $\sigma^*_{ACE_{OL}}$ can be computed as follows

$$\sigma^*_{ACE_{OL}} = \sqrt{\sigma^2_{ACE_{OL}} + (\sigma^*)^2 + 2\rho_{ACE_{OL}\times RES} \cdot \sigma_{ACE_{OL}} \cdot \sigma^*} \tag{2}$$

where $\sigma_{ACE_{OL}}$ is the original standard deviation of ACE_{OL}, σ^* is overall standard deviation of the prediction error of the RES production estimate, $\rho_{ACE_{OL}\times RES}$ is correlation coefficient between ACE_{OL} and prediction error of RES production. Its value has been empirically set to $\rho_{ACE_{OL}\times RES} = 0,3$ for the Central Europe region.

B. Volatility increase of RES production estimate due to newly installed RES

Newly installed RES bias standard deviation of the prediction error of RES production σ^*. From (1) and (2), maximal tolerable volume of standard deviation σ^* can be derived

$$\begin{aligned}\sigma^* \;=\; & -\rho_{ACE_{OL}\times RES} \cdot \sigma_{ACE_{OL}} + \\ & + \sqrt{\left(\frac{SCR^*}{q_{ACE}}\right)^2 - \sigma^2_{ACE_{OL}}(1-\rho^2_{ACE_{OL}})}\end{aligned}\;, \tag{3}$$

where SCR^* is fixed maximal volume of the reserve.

C. Calculation of overall standard deviation of RES production prediction error

Let us consider the global area $\mathcal{A}$ which is divided into local areas $\mathcal{A} = \{area_1, area_2, \ldots, area_N\}$ where N is number of areas. Particular areas are interconnected with a central (transmission) power network which is operated by a central operator. In order to power network balance, the operator can utilize AS, whose volume is limited. This section describes the computation of the overall standard deviation of RES production prediction error from prediction errors calculated in several local areas. In more details, the method is described in [5]. Here we sketch out only main principles.

We assume that newly installed RES have similar behavioural pattern as already installed RES in local areas. Ratio between already installed power and installed power after installation of new RES (in each local area $a \in \mathcal{A}$) can be defined as an installation factor

$$c_a = \frac{P^*_a}{P_a}\;, \tag{4}$$

where P_a denotes already installed power, P^*_a means power capacity after installation of new energy sources in specified area $a \in \mathcal{A}$.

After installation of new RES, standard deviation of prediction error in area $a \in \mathcal{A}$ is updated according to formula

$$\sigma_a = \sigma_{a,base} \cdot \sqrt{c_a \cdot (1 + (c_a - 1) \cdot \rho_a)} \tag{5}$$

where c_a is relevant installation factor, $\sigma_{a,base}$ is the original standard deviation of the prediction error and ρ_a is correlation coefficient in area $a \in \mathcal{A}$. For all areas, standard deviation vector $\mathcal{E}$ can be defined as

$$\mathcal{E} = (\sigma_a)_{a\in\mathcal{A}} \tag{6}$$

and correlation matrix as

$$R = (\rho_{i,j})_{i,j\in\mathcal{A}}\;, \tag{7}$$

where σ_a means standard deviation of prediction error in area $a \in \mathcal{A}$ after newly installed RES and $\rho_{i,j}$ is correlation coefficient between i-th and j-th areas.

Overall standard deviation of power generation prediction error across all areas is formulated as follows

$$\sigma = \sqrt{\mathcal{E}^T \cdot R \cdot \mathcal{E}}\;, \tag{8}$$

where $\mathcal{E}$ is a standard deviation vector, R is a correlation matrix.

The constraint of the overall standard deviation of RES production prediction error σ implies from (3) and (8) and can be written as follows

$$\sqrt{\mathcal{E}^T(c_1, \ldots c_N) \cdot R \cdot \mathcal{E}(c_1, \ldots c_N)} \le \sigma^*\;, \tag{9}$$

where $\mathcal{E}(c_1, \ldots c_N)$ is standard deviation vector defined by (6), R is correlation matrix given by (7), $c_1, \ldots c_N$ are relevant installation factors specified by (4) and σ^* is maximal tolerable volume of standard deviation for newly installed RES given by (3).

IV. Maximum utilization of the areas

The problem of maximization of RES volume installed in particular local areas under consideration of SCR limitation is solved in this section. Due to RES production coincidence, the solution of the problem above is not straightforward and lead to the constrained maximization problem.

Constrained maximization problem can be rewritten to the minimization one as follows

$$\min_{\forall c_a}\left(-\prod c_a\right),\ \text{for } a = 1, \ldots N \tag{10}$$

subject to equality constraint

$$k_c \cdot \sigma^* = \sqrt{\mathcal{E}^T(c_1, \ldots c_N) \cdot R \cdot \mathcal{E}(c_1, \ldots c_N)} \tag{11}$$

and inequality constraints

$$1 \le c_a \le c_{a,max} \tag{12}$$

where c_a is relevant installation factor in area $a \in \mathcal{A}$, σ^* is maximal value of standard deviation of RES production prediction error, $\mathcal{E}(c_1, \ldots c_N)$ is standard deviation vector, R is correlation matrix. Usability coefficient $k_c \in \langle 0; 1\rangle$ reflects the usage level of maximal potential of SCR volume. The maximal value of $c_{a,max}$ is determined by two ways: 1) theoretical PV production potential of a local area, 2) maximal transfer capability of the power network interconnecting local areas. The value of $c_a = 1$ means no increase energy sources in local area.

ISBN: 978-1-61804-256-9

A. Optimal deployment of newly installed RES into local areas

The solution of the optimization problem (10) can be performed in several ways. In this paper, however, we use the method of Lagrange multipliers.

Consider the optimization problem

$$\min_{c_1,c_2,\dots c_N} J = f(c_1, c_2, \dots c_N) = - \prod_{a\in\{1,2,\dots,N\}} c_a \qquad (13)$$

subject to equality constraint

$$h(c_1, \dots c_N) = k_c \cdot \sigma^* - \sqrt{\mathcal{E}^T(c_1, \dots c_N) \cdot R \cdot \mathcal{E}(c_1, \dots c_N)} \; . \qquad (14)$$

Functions $f(c_1, \dots c_N)$ and $h(c_1, \dots c_N)$ are continuous with continuous first partial derivatives and $\text{grad } h \neq 0$. The Lagrange function is defined as

$$\Lambda(c_1, \dots c_N, \lambda) = f(c_1, \dots c_N) + \lambda \cdot h(c_1, \dots c_N) \qquad (15)$$

where the constant λ is called the Lagrange multiplier. In our case, the Lagrange function (15) is defined as: $\Lambda(c_1, \dots c_N, \lambda) = -\prod_{a=1}^{N} c_a + \lambda \left(k_c \cdot \sigma^* - \sqrt{\mathcal{E}^T(c_1, \dots c_N) \cdot R \cdot \mathcal{E}(c_1, \dots c_N)} \right)$.

Functions Λ and f have same extremes with respect to constraint h. The extreme (c_a^0, λ^0) is found by solving $N+1$ equations (with $N+1$ variables)

$$h(c_1^0, \dots c_N^0) = 0 \qquad (16)$$

$$\text{grad } f(c_1^0, \dots c_N^0) + \lambda \cdot \text{grad } h(c_1^0, \dots c_N^0) = 0 \qquad (17)$$

where (16) and (17) stand for necessary conditions of the optimization problem subject to equality constraints.

Functions f and h have continuous second differentiation which is sufficient condition for finding local minimum (c_a^0, λ^0), $a \in \mathcal{A}$. The second differential is given by

$$d^2\Lambda = d^2 f(c_1^0, \dots c_N^0) + \lambda^0 \cdot d^2 h(c_1^0, \dots c_N^0) > 0 \; . \qquad (18)$$

B. Minimal value of the usability coefficient

Usability coefficient k_c is an cautious coefficient which reflects an usage level of SCR absorption potential[1].

In this part, minimal value of the coefficient k_c is determined which emerges from the equality constraint (14). In the case $\exists c_a > 1$, the usability coefficient is $k_c \leq 1$. For all $c_a = 1$, however, the usability coefficient k_c is given by equation

$$k_c \geq \left. \frac{\sqrt{\mathcal{E}^T(c_1, \dots c_N) \cdot R \cdot \mathcal{E}(c_1, \dots c_N)}}{\sigma^*} \right|_{c_a=1} \qquad (19)$$

where $a \in \mathcal{A}$ $(a = 1, \dots N)$. The condition (19) determines the lower limit for usability coefficient k_c.

[1]Here, SCR absorption potential stands for the capability of the power network to eliminate ACE fluctuation caused by RES production.

V. Cases studies

The application of described optimization approach on selected case study is shown in this section.

Among others, the equality constraint (11) depends on standard deviation of ACE_{OL} without newly installed RES $\sigma_{ACE_{OL}}$ and correlation coefficient $\rho_{ACE_{OL} \times RES} = 0.3$. Usually statistical parameters of ACE_{OL} are computed for selected categories which are characterized by months in year and defined day bands (e.g. working day, working night, non-working day and non-working night). In this example, we consider the case of working day in April where standard deviation of ACE_{OL} is set to $\sigma_{ACE_{OL}} = 113.94\,[\text{MW}]$. Fixed SCR volume is selected to

$$SCR^* = 500\,[\text{MW}] \; . \qquad (20)$$

More details about the method for calculation of minimal volume of SRC can be found in [5]. Based on selected SCR volume and the equation (3), the overall tolerable standard deviation of the prediction error of the RES production estimate σ^* is

$$\sigma^* = 180.89\,[\text{MW}] \qquad (21)$$

In our case study, we consider a global area which consists of two interconnected local areas $a \in \mathcal{A}$ $(N = 2)$ as depicted on the figure 1. Each area is characterized by installed RES power P_a and standard deviation of the RES prediction error $\sigma_{a,base}$ where $a \in \mathcal{A}$. For RES, typically, statistical parameters depend on the time of the day. In this paper, we restrict ourselves to the values at noon.

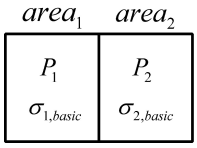


Fig. 1. Structure of the global area

Following table I summarizes the basic parameters of local areas (P_a, $\sigma_{a,base}$).

TABLE I
BASIC PROPERTIES OF LOCAL AREAS

	$area_1$	$area_2$
P_a [MW]	550.00	150.00
$\sigma_{a,base}$ [MW]	65.77	17.33

In the following text, four scenarios are described which consist in application of the optimization method under different conditions. Namely, we change values of correlation coefficients in selected areas ρ_a $(a \in \mathcal{A})$ and correlation coefficient between i-th and j-th area $\rho_{i,j}$.

A. Scenario A

In this scenario, we consider same correlation coefficient in local areas and no correlation between areas. Correlation coefficients for $area_1$ and $area_2$ (i.e. ρ_1 and ρ_2) are included in table II as well as the correlation coefficient between $area_1$ and $area_2$ (i.e. $\rho_{1,2}$).

ISBN: 978-1-61804-256-9

TABLE II
CORRELATION PARAMETERS RELATED TO SCENARIO A

correlation coefficient $area_1$	correlation coefficient $area_2$	correlation coefficient between first and second areas
ρ_1	ρ_2	$\rho_{1,2}$
1.0	1.0	0.0

The correlation matrix between $area_1$ and $area_2$ used in (7) is defined as

$$R = \begin{pmatrix} 1.0 & \rho_{1,2} \\ \rho_{1,2} & 1.0 \end{pmatrix} = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix} \tag{22}$$

For correlation coefficients from table II, the Lagrange function (15) has form

$$\Lambda(c_1, c_2, \lambda) = -c_1 \cdot c_2 + \lambda \cdot \left(\sigma^* - \sqrt{\sigma^2_{1,base} c_1^2 + \sigma^2_{2,base} c_2^2}\right) \tag{23}$$

We use overall tolerable standard deviation of the RES production prediction error given by (21) and values from tables I and II to the (23) for computation of extremes

$$c_1 = 1.94 \tag{24}$$

$$c_2 = 7.38 \tag{25}$$

These values (24) and (25) stand for optimal ratio between already installed power and installed power after installation of new RES (in local $area_1$ and $area_2$) with respect of correlation coefficients of the Scenario A. These values can be used in the (4) and optimal RES production deployment[2] P_a^* in local areas is

$$P_1^* = 1067\,[\mathrm{MW}] \tag{26}$$

$$P_2^* = 1107\,[\mathrm{MW}] \tag{27}$$

Note, these values are computed with the usability coefficient $k_c = 1$ which declares maximal utilization of RES production potential for selected SCR volume.

B. Scenario B

In this scenario, we consider two areas with various correlation inside area. In first area, correlation coefficient is reduced which can represent a large area where RES production is not fully correlated. All correlation coefficients are summarised in table III.

TABLE III
CORRELATION PARAMETERS FOR SCENARIO B

correlation coefficient $area_1$	correlation coefficient $area_2$	correlation coefficient between first and second areas
ρ_1	ρ_2	$\rho_{1,2}$
0.8	1.0	0.0

For selected correlation coefficients (s. Table II), the Lagrange function (15) is

$$\begin{aligned} \Lambda(c_1, c_2, \lambda) &= -c_1 \cdot c_2 + \\ &+\lambda \cdot \left(\sigma^* - \sqrt{\sigma^2_{1,base} \cdot K + \sigma^2_{2,base} c_2^2}\right) \end{aligned} \tag{28}$$

[2]power capacity after installation of new energy sources in certain area

where parameter K substitutes $K = c_1 \cdot (1 + (c_1 - 1) \cdot \rho_1)$. For sake of clarity, parameter K is used throughout following text.

In the (28), we use overall tolerable standard deviation of the RES production prediction error given by (21) and values from tables I and III for the computation of extremes

$$c_1 = 2.08 \tag{29}$$

$$c_2 = 7.28 \tag{30}$$

These values (29) and (30) represent optimal ratio between already installed power and installed power after installation of new RES (in local $area_1$ and $area_2$) with respect of correlation coefficients for Scenario B.

Now these values can be used in the (4) and optimal RES production deployment[3] P_a^* can be calculated as

$$P_1^* = 1144\,[\mathrm{MW}] \tag{31}$$

$$P_2^* = 1092\,[\mathrm{MW}] \tag{32}$$

C. Scenario C

In contrast to previous scenario, here, a correlation between both areas is considered. However, correlation coefficients inside local areas remain same as in previous case. Table IV summarizes the correlation coefficients in $area_1$ and $area_2$ and correlation coefficient between $area_1$ and $area_2$.

TABLE IV
CORRELATION PARAMETERS FOR SCENARIO C

correlation coefficient $area_1$	correlation coefficient $area_2$	correlation coefficient between first and second areas
ρ_1	ρ_2	$\rho_{1,2}$
0.8	1.0	0.73

The correlation coefficients between $area_1$ and $area_2$ are used in the correlation matrix (7)

$$R = \begin{pmatrix} 1.0 & \rho_{1,2} \\ \rho_{1,2} & 1.0 \end{pmatrix} = \begin{pmatrix} 1.00 & 0.73 \\ 0.73 & 1.00 \end{pmatrix} \tag{33}$$

Note, correlation coefficient between local areas is close to real values observed in Central Europe region[4]. In this case, functions f (13) and h (14) are transformed in following form

$$f(c_1, c_2) = -c_1 \cdot c_2 \tag{34}$$

subject to equality constraint

$$\begin{aligned} h(c_1, c_2) &= \sigma^* - \sqrt{\mathcal{E}^T(c_1, c_2) \cdot R \cdot \mathcal{E}(c_1, c_2)} \\ &= \sigma^* - \sqrt{\begin{array}{l} \sigma^2_{1,base} \cdot K^2 + \sigma^2_{2,base} c_2^2 + \\ +2 \cdot \rho_{1,2} \cdot \sigma_{1,base} \cdot K \cdot \sigma_{2,base} \cdot c_2 \end{array}} \end{aligned} . \tag{35}$$

For selected correlation coefficients defined in the Table IV, the Lagrange function (15) is

$$\begin{aligned} &\Lambda(c_1, c_2, \lambda) = -c_1 \cdot c_2 + \\ &+\lambda \cdot \left(\sigma^* - \sqrt{\begin{array}{l} \sigma^2_{1,base} \cdot K^2 + \sigma^2_{2,base} c_2^2 + \\ +2 \cdot \rho_{1,2} \cdot \sigma_{1,base} \cdot K \cdot \sigma_{2,base} \cdot c_2 \end{array}}\right) \end{aligned} \tag{36}$$

[3]power capacity after installation of new energy sources in specified area

[4]Similar values have been empirically observing in operator's data from Central Europe.

ISBN: 978-1-61804-256-9

We use overall tolerable standard deviation of the RES production prediction error given by (21) and values from tables I and IV for computation of extremes according to the (28). Resulting extremes are

$$c_1 = 1.58 \tag{37}$$

$$c_2 = 5.44 \tag{38}$$

Values (37) and (38) mean the optimal ratio between already installed power and installed power after installation of new RES (in local $area_1$ and $area_2$) with respect correlation coefficient of Scenario C.

These values can be consequently used in the (4) and optimal deployment of RES production P_a^* (power capacity after installation of new energy sources in specified area) can be calculated for both areas.

$$P_1^* = 869\,[\mathrm{MW}] \tag{39}$$

$$P_2^* = 816\,[\mathrm{MW}] \tag{40}$$

Based on the results, one can observe decreasing RES installation capacity of local areas with increasing correlation coefficient $\rho_{1,2}$.

D. Scenario D

This scenario stands for the limit case from the operator's perspective, where full correlation of RES production between local areas is considered (i.e. $\rho_{1,2} = 1$). All correlation coefficients used in the calculation are included in the table V.

TABLE V
VERSION D

correlation coefficient $area_1$	correlation coefficient $area_2$	correlation coefficient between first and second areas
ρ_1	ρ_2	$\rho_{1,2}$
0.8	1.0	1

The correlation coefficient between $area_1$ and $area_2$ is used in the correlation matrix (7) which has form

$$R = \begin{pmatrix} 1.0 & \rho_{1,2} \\ \rho_{1,2} & 1.0 \end{pmatrix} = \begin{pmatrix} 1.00 & 1.00 \\ 1.00 & 1.00 \end{pmatrix} \tag{41}$$

In this limit case, functions f (13) and h (14) are transformed in form (34) and (35). For selected correlation coefficients defined in the Table V, the Lagrange function (15) is

$$\begin{aligned} &\Lambda(c_1, c_2, \lambda) = -c_1 \cdot c_2 + \\ &+\lambda \cdot \left(\sigma^* - \sqrt{\begin{array}{c} \sigma_{1,base}^2 \cdot K^2 + \sigma_{2,base}^2 c_2^2 + \\ +2 \cdot \sigma_{1,base} \cdot K \cdot \sigma_{2,base} \cdot c_2 \end{array}} \right) = \\ &= -c_1 \cdot c_2 + \lambda \cdot (\sigma^* - (\sigma_{1,base} \cdot K + \sigma_{2,base} c_2)) \ . \end{aligned} \tag{42}$$

Extremes are computed based on the overall tolerable standard deviation of the RES production prediction error given by (21) and values from tables I and V by use of the (28) as follows

$$c_1 = 1.48 \tag{43}$$

$$c_2 = 5.02 \tag{44}$$

Values of extremes (43) and (44) stand for optimal ratios between already installed power and installed power after installation of new RES (in local $area_1$ and $area_2$) with respect correlation coefficient of the Scenario D. In next step, these values can be used in the (4) and optimal RES production P_a^* (power capacity after installation of new energy sources in specified area) can be expressed.

$$P_1^* = 814\,[\mathrm{MW}] \tag{45}$$

$$P_2^* = 753\,[\mathrm{MW}] \tag{46}$$

The values are computed for extreme situation where both local areas are fully correlated which put highest requirements on SCR volume[5].

E. Scenarios assessment

Results obtained in previous scenarios are assessed and discussed in this section. The assessment focuses on the last three scenarios (i.e. B,C,D), where same correlation coefficients for local areas are used ($\rho_1 = 0.8$, $\rho_2 = 1$). In discussed scenarios, three values of correlation coefficient $\rho_{1,2}$ are tested and the optimal RES production deployment is computed for each scenario. Here, a sensitivity analysis of the maximal installable RES production to the inter-area correlation $\rho_{1,2}$ is performed. Correlation coefficient between $area_1$ and $area_2$ (i.e. $\rho_{1,2}$) is chosen in interval $\rho_{1,2} \in \langle 0, 1 \rangle$. Moreover, we analyse the dependency of the maximal installable RES production on the usability coefficient k_c, which is selected from interval $k_c = \langle 0.6, 1 \rangle$.

The result of the sensitivity analysis is depicted on the figure 2. Aggregated RES power production $P_1^* + P_2^*$ over both

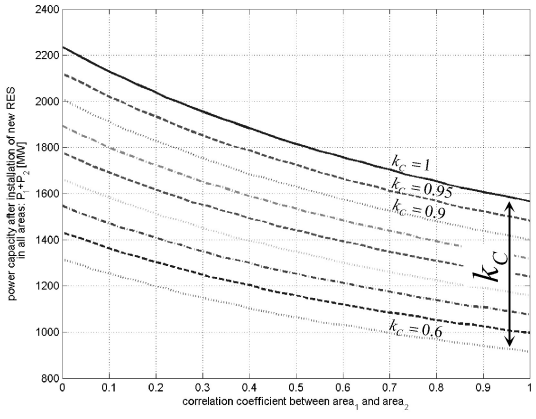


Fig. 2. Dependency of maximal installable RES production $P_1^* + P_2^*$ on the correlation coefficient $\rho_{1,2}$ and usability coefficient k_c

areas is on y-axis while x-axis captures correlation coefficient between local areas. On the figure, several lines for various values of the usability coefficient k_c are depicted. One can see a linear dependence of the power capacity $P_1^* + P_2^*$ on the coefficient k_c. More interesting result can be seen from the dependency of installed power $P_1^* + P_2^*$ on the correlation coefficient between local areas $\rho_{1,2}$. With increasing value of

[5]Because of high coincidence in RES power production.

ISBN: 978-1-61804-256-9

the correlation coefficient, maximal installable power $P_1^* + P_2^*$ decrease. When correlation coefficient is $\rho_{1,2} = 1$, maximal installed power stands for the conservative estimate from the SCR volume point of view. If lower value of correlation is considered, capability of fixed SCR volume to eliminate RES production fluctuations increases and therefore increase maximal installable RES production as well.

Suitable choice of the correlation coefficient is very important for determination of maximal installable RES production. Because estimates of the correlation coefficient $\rho_{1,2}$ are usually influenced by an estimation error, installed RES production has to be lowered by usage of the usability coefficient less than 1. Next figure 3 shows the importance of correct choice of the correlation coefficient $\rho_{1,2}$.

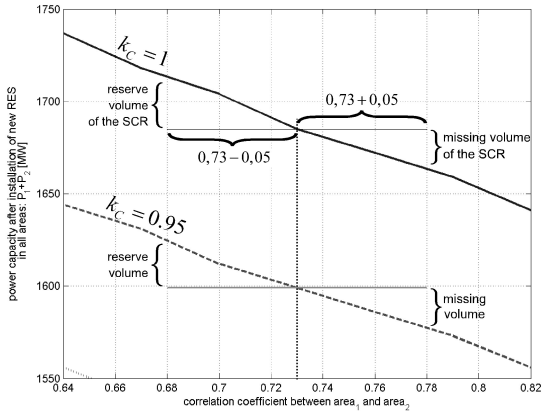


Fig. 3. Influence the correct choice of the correlation coefficient $\rho_{1,2}$

Let us consider the situation, that correlation coefficient is estimated to $\rho_{1,2} = 0.73$ with the absolute error of $\Delta = \pm 0.05$. If the actual value of the correlation coefficient is $\rho_{1,2} < 0.73$, the maximal RES production potential is not fully utilized, but SCR can balance the power network. In opposite case ($\rho_{1,2} > 0.73$), SCR volume can not fully cover power fluctuations caused by RES and higher levels of AS has to be utilized.

VI. Conclusion

In the paper, we describes two methods that can help to overcome current challenges cause by increasing share of intermittent energy sources in power networks. First method focuses on the determination of maximal RES production that can be absorbed by power nerwork with limited SCR volume. Second method stands for constrained optimization problem where deployment of newly installed RES in several local areas is optimized. Applications of both methods are demonstrated on four case studies, which are concluded as a sensitivity analysis of the maximal installable RES production. In the case study, calculations of the optimal RES production deployment under various scenarios are discussed.

Based on sensitivity analysis, several observations can be concluded. Correlation coefficients (i.e. correlation coefficients for local areas and correlation coefficient between local areas) has to be considered, because strongly influences maximal installable RES production. The influence of correlation coefficient for local areas (ρ_1 and ρ_2) is not so important, which is demonstrated on scenarios A and B. On the other hand, the influence of correlation coefficient between local areas $\rho_{1,2}$ is crucial, because strongly affects maximal volume of installable RES. Correlation coefficient between areas is estimated with an estimation error, whose influence to the power network stability can be reduced by suitable choice of the usability coefficient k_c. Scenarios C and D extreme situation for the impact determination of new RES installation.

References

[1] "Operation handbook," ENTSO-E, Tech. Rep., 2010.
[2] "Grid code," ČEPS, a.s., Tech. Rep., January 2012.
[3] A. Zápotocká and J. Fantík, "Proposal for new categories of ancillary services in a transmission system with a massive number of renewable energy sources," in *Proceedings of the 3rd European Conference of Control (ECC 12)*. Paris, France, December 2012, pp. 1–6.
[4] R. Elsen, "Ancillary services unbundling electricity products an emerging market," Union of the electricity industry (Eurelectric), Tech. Rep., 2004.
[5] A. Zápotocká, P. Janeček, and M. Střelec, "A method for assessment of impact of RES integration onto ancillary services," in *Proceedings of the 15th Scientific Conference Electric Power Engineering 2014*. Brno University of Technology, May 2014, pp. 1–5.

Distribution optimization in a single level logistic network

Laila Kechmane, Benayad Nsiri, Azeddine Baalal

Abstract— Optimizing distribution aims at reducing the costs relate[1]d to product transportation while allowing companies to satisfy their customers' needs by supplying the right product, the right quantities at the right time and place. This paper examines a single level logistic network where the only medium between the warehouse and the customer is distribution centers. The objective of this work is to allocate customers to distribution centers and vehicles to travels in order to reduce the transportation costs by cutting down the distance traveled while observing the distribution centers' storage capacities and guaranteeing the satisfaction of the customers' needs. We based on the vehicle routing problem study to propose a mixed integer programming formula that can be solved using Lingo 14.0. A digital example will be given in the end to illustrate the practicability of the model.

Keywords— Mixed integer programming, optimizing distribution, single level logistic network, transportation costs**.**

I. INTRODUCTION

Companies that manage distribution of their goods seek to cut down transport costs and avoid stock shortages at their distribution centers. To ensure a certain quality of customer service, companies have to manage the allocation of customers to distribution centers in a way that minimizes the costs of transport and considers the storage capacity of the vehicles as well as of the various distribution network's nodes.

The problem related to transport and distribution of goods ranges from vehicle routing problems such as the TSP (Travelling Salesman Problem) formulated by the mathematicians WR Hamilton and Thomas Kirkman in 1800 [1], to the construction of whole networks and thus to factories setting up and allocation of various nodes to customers [2]. Vehicle Routing Problem has been an active area of research; Traveling Salesman Problem (TSP) focuses on finding the optimal route to visit a given number of cities while minimizing transportation cost [3]-[4], the Vehicle routing problem (VRP), which is an extension of the TSP, was formulated in 1959 by Danzig and Ramser [5], according to Laporte [6], this problem aims at building the optimal tours of pickup or delivery, from one or several warehouses towards a number of customers or cities that are geographically scattered, while respecting certain constraints. There exist four variants of the VRP [7]: VRP with Time Windows (VRPTW) [8], VRP with Pickup and Delivery (VRPPD) [9], the[2] capacitated VRP (CVRP) [10] and the VRP with Backhauls (VRPB) [11].

The first algorithm to solve the VRP problem, was proposed by Clarke and Wright in 1964 [12], and since then, several methods were proposed and which are either exact methods that allow to find an optimal solution, or approximate methods that allow to obtain a solution to the problem but which is not optimal [13].

Since its introduction, the formulation of several models aiming at the optimization of the transport costs has been based on the VRP. Likewise, the proposed mathematical model is based on the VPR and addresses the minimization of transport costs as well as those of storage, both being parts of logistic distribution.

In the following section, we will begin by presenting our mathematical model, then, we will apply it to a real case and solve it by the Lingo 14.0 software to test its reliability.

We consider a logistic network consisted of one plant, n distribution centers and m customers. The first objective is to allocate customers to distribution centers and vehicles to travels so as to minimize the distances to travel, and ultimately the transport costs. The second objective is to minimize the storage costs at the distribution centers by minimizing the stored quantities while respecting the daily quantities that can be delivered to every center and satisfy the customers' needs.

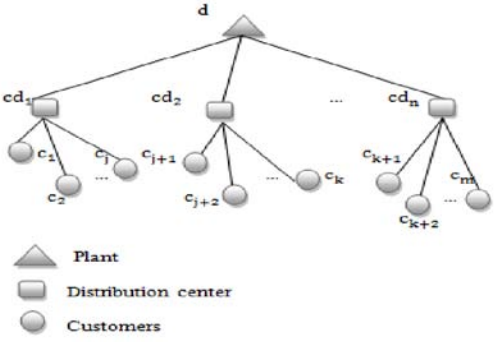


Fig. 1 A single level logistic network

Laila Kechmane, Department Mathematics and Computing, MACS Laboratory, University Hassan II, Faculty of Sciences Casablanca Km 8 Eljadida road, P.B 5366, Maarif 20100, Morocco, (e-mail: kechmanelaila@gmail.com).

Benayad Nsiri, Department Mathematics and Computing, MACS Laboratory, University Hassan II, Faculty of Sciences Casablanca Km 8 Eljadida road, P.B 5366, Maarif 20100, Morocco(e-mail: benayad.nsiri@telecom-bretagne.eu).

Azeddine Baalal, Department Mathematics and Computing, MACS Laboratory, University Hassan II, Faculty of Sciences Casablanca Km 8 Eljadida road, P.B 5366, Maarif 20100, Morocco, (e-mail: abaalal@gmail.com).

II. Mathematical formulation

Sets

I : Collection of distribution centers $i \in I$, $i = 1, 2, \ldots, n$;

M : Set of customers j, $j \in M$, $j = 1, 2 \ldots m$;

T : Set of periods t, $t \in T$, $t = 1, 2 \ldots t'$;

C_h : Set of vehicles vh with a capacity h ;

C : Set of C_h ;

Parameters

d_{ip} : Distance between plant p and center i

d_{ij} : Distance between center i and customer j

c_{vh} : Transportation cost per km for a vehicle vh

c_{si} : Unit cost of storage per day at distribution center i

n_h : Number of vehicles of category C_h

cap_{vh} : Vehicle vh capacity

b_i^t : Center i demand on day t

c_i : Storage capacity at center i

bes_j^t : Customer j demand on day t

stk_i^t : Available stock at center i on day t

Decision variables

x_{ip}^t : Quantity to deliver from the plant p to center i on day t

y_{ij}^t : Quantity to deliver from center i to customer j on day t

$$l_{vh}^{it} = \begin{cases} 1 \text{ if vehicle vh visits center i on day t} \\ 0 \text{ else} \end{cases}$$

$$l_{vh}^{ijt} = \begin{cases} 1 \text{ if vehicle vh travels from center i to customer j on day t} \\ 0 \text{ else} \end{cases}$$

We assume all parameters are nonnegative.

Objective function

$$Min \sum_{v \in Ch} \sum_{t=1}^{t'} \sum_{i \in I} l_{vh}^{it} x_{ip}^t d_{ip} c_{vh} + \sum_{v \in Ch} \sum_{t=1}^{t'} \sum_{i \in I} \sum_{j \in M} l_{vh}^{ijt} y_{ip}^t d_{ij} c_{vh} + \sum_{t=1}^{t'} \sum_{i \in I} c_{si} \left(x_{ip}^t - b_i^t + stk_i^t \right) \quad (1)$$

Subject to

$$x_{ip}^t - b_i^t \le c_i \quad \forall i \in I \ \forall t \in T \quad (2)$$

$$b_i^t \ge x_{ip}^t + stk_i^t \quad \forall i \in I \quad \forall t \in T \quad (3)$$

$$stk_i^t = stk_i^{t-1} + x_{ip}^t - b_i^t \quad \forall i \in I \quad \forall t \in T \quad (4)$$

$$\sum_{i \in I} y_{ij}^t = bes_j^t \quad \forall i \in I \quad \forall t \in T \quad \forall j \in M \quad (5)$$

$$\sum_{vh \in C_h} l_{vh}^{it} \le n_h \quad \forall i \in I \quad \forall t \in T \quad \forall vh \in C_h \quad (6)$$

$$l_{vh}^{it} x_{ip}^t \le cap_{vh} \quad \forall i \in I \quad \forall t \in T \quad \forall vh \in C_h \quad (7)$$

$$l_{vh}^{ijt} y_{ij}^t \le cap_{vh} \quad \forall i \in I \quad \forall j \in M \quad \forall t \in T \quad \forall vh \in C_h \quad (8)$$

$$l_{vh}^{it}, l_{vh}^{ijt} \in \{0,1\} \quad \forall i \in I \quad \forall j \in M \quad \forall t \in T \quad \forall vh \in C_h \quad (9)$$

The objective function (1) expresses the cost to be minimized and which is the sum of:

- Travelling costs from the plant to distribution centers;
- Travelling costs from centers to the customers;
- Storage costs at the distribution centers.

Constraint (2) assures the respect of the storage capacity of every distribution center.

Constraint (3) assures that the daily need for every distribution center is satisfied.

Constraint (4) calculates the quantity available in every distribution center.

Constraint (5) assures that the daily need of every customer is satisfied.

Constraint (6) assures the respect of the number of vehicles available in each category.

Constraints (7) and (8) assure the respect of each vehicle capacity.

III. Illustrative example

To illustrate our model, we apply it to a network consisted of a single plant, 4 distribution centers and 13 customers. Table. I includes storage parameters. There are two categories of vehicles for travels linking plant to centers and two other categories for travels linking centers to customers. Table. II represents the characteristics of different vehicles, we note that when sending a vehicle to a center, it is completely filled, even if the sent quantity exceeds the center need, what explains the existence of stock.

Table. III and table. IV represent respectively distances between plant and various distribution centers, and distances between the latter and customers. Tables. V and table. VI represent respectively the daily needs of distribution centers and of customers over a period of 4 days.

Table. I Parameters values

Parameter	Value
c_{si}	0.20
c_i	500

ISBN: 978-1-61804-256-9

Table. II Characteristics of each type of vehicle vh

vh	A	B	C	D
c_{vh}	0.21	0.21	0.20	020
n_h	2	4	8	10
cap_{vh}	1000	600	350	200

Table. III Distances between the plant and distribution centers

	cd_1	cd_2	cd_3	cd_4
Plant	511	0	291	369

Table. IV Distances between distribution centers and customers

	cd_1	cd_2	cd_3	cd_4
c_1	419	99	390	469
c_2	172	351	642	721
c_3	651	133	166	237
c_4	719	259	204	93
c_5	303	241	485	611
c_6	746	23	60	267
c_7	614	93	198	281
c_8	772	314	29	422
c_9	439	72	361	433
c_{10}	735	217	82	221
c_{11}	87	483	773	818
c_{12}	907	411	120	423
c_{13}	910	390	286	60

Table. V Daily distribution centers' need during period T

	j_1	j_2	j_3	j_4
cd_1	822	832	840	838
cd_2	793	1007	985	1015
cd_3	508	531	516	513
cd_4	476	472	460	481

Table. VI Daily customers' need during period T

	j_1	j_2	j_3	j_4
c_1	300	311	298	288
c_2	302	298	288	278
c_3	200	188	185	186
c_4	150	147	156	148
c_5	340	345	329	330
c_6	188	201	210	198
c_7	347	300	321	311
c_8	150	165	140	160
c_9	250	248	211	200
c_{10}	139	128	148	144
c_{11}	180	189	223	230
c_{12}	170	165	166	155
c_{13}	187	197	156	189

Discussion

We solve this problem using a Mixed Integer Linear Programming solver LINGO 14.0 [14] on an Acer Aspire ONE D255 1.00 GHz machine, running Windows 7 Starter Edition. Results are obtained in 0.56 seconds, and the objective value is 901032.1.

Table. VII represents the optimal quantities to be sent to distribution centers during period T and which meet their needs. Fig. 2, Fig. 3, Fig. 4, and Fig. 5 represent each the affectation of customers to distribution centers, optimal quantities to be sent on every day of period T and which category of vehicle to use.

We notice that obtained results respect the various constraints of our example, which are the storage capacity of distribution centers and the needs of the final customers.

Table.VII Optimal quantities to be sent to the distribution centers during period T

	j_1	j_2	j_3	j_4
cd_1	1000	1000	1000	600
cd_2	1000	1000	1000	1000
cd_3	600	600	600	600
cd_4	600	600	600	600

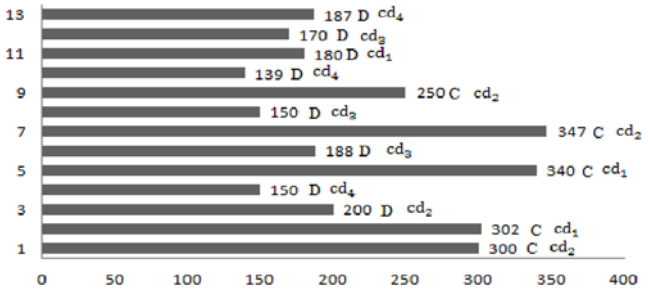


Fig. 2 Affectations on day J_1

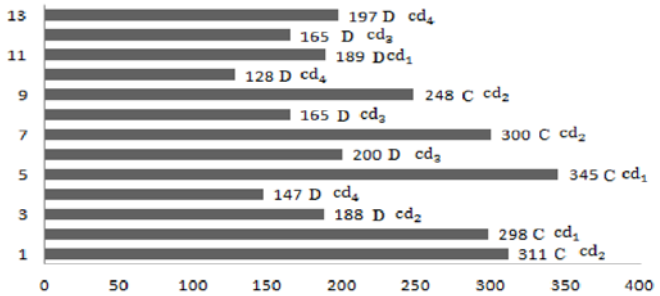


Fig. 3 Affectations on day J_2

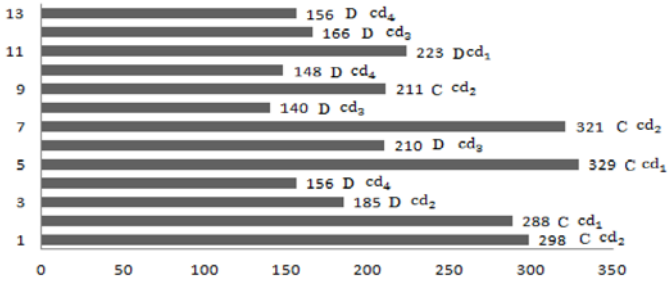


Fig. 4 Affectations on day J_3

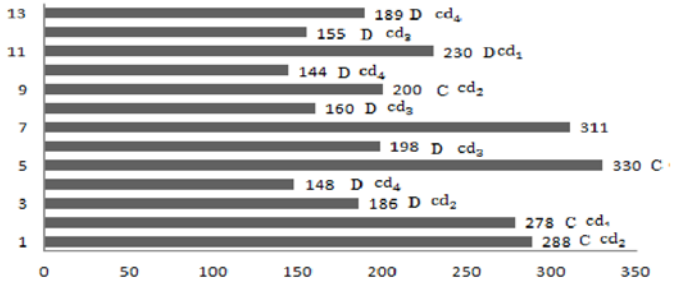


Fig. 5 Affectations on day J_4

IV. Conclusion and perspectives

The number of scientific publications handling transport problems continues to increase, so proving the importance of this function of supply chain. In this paper, we investigate the optimization of the distribution problem, the objective is to minimize both the traveled distances and the storage level, and allocate vehicles to travels. We relied on the vehicle routing problem VRP to develop our mathematical formula.

In this work, a single level logistic network is considered to apply our model. As perspective, we can consider a multi level logistic network, the model can be easily developed and applied in that case.

References

[1] D. Davendra, Traveling salesman problem, theory and applications. India, InTech, 2010

[2] S. Chopra,"Designing the distribution network in a supply chain,"Transportation Research Part E: Logistics and Transportation Review, 39, pp. 123-140, 2003

[3] G. Laporte, "The Traveling Salesman Problem: An overview of exact and approximate algorithms," European Journal of Operational Research 59, pp. 231-247, 1992

[4] G. B. Dantzig, D. R Fulkerson and S. M. Johnson, "The solution of a large-scale traveling salesman problem," Operations Research 2: pp. 393-410, 1954.

[5] G. B. Dantzig and J. H. Ramser, "The truck dispatching problem," Management Science 6: pp. 80-91. 1959.

[6] G. Laporte, "The Vehicle Routing Problem: An overview of exact and approximate algorithms," European Journal of Operational Research 59: pp. 345-358, North-Holland, 1992.

[7] P. Toth, and D. Vigo, (Eds). "The vehicle routing problem," SIAM Monographs on Discrete Mathematics and Applications. Philadelphia, US: Society for Industrial and Applied Mathematics, 2001

[8] M. Solomon, "Algorithms for the vehicle routing problem with time windows," Transportation Science, 29(2):pp. 156-166, 1995.

[9] S. N. Parragh, K. F. Doerner and R. F. Hartl, "A survey on pickup and delivery problems, Part I: Transportation between customers and depot," JfB 58:pp. 21-51, 2008.

[10] T.K Ralphs, L. Kopman, W.R. Pulleyblank, and L. E. Jr. Trotter, "On the capacitated vehicle routing problem," Mathematical Programming 94 (2-3):pp. 343-359, 2003.

[11] P. Toth, and D. Vigo, "An exact algorithm for the vehicle routing problem with backhauls," Transportation science 31(4):pp. 372-385, 1997.

[12] G. Clarke, and J.W. Wright, "Scheduling of vehicles from a central depot to a number of delivery points", Operations Research 12:pp. 568-581, 1964.

[13] G. Laporte, "What you should know about the vehicle routing problem," Naval Research Logistics, 54(8): pp. 811-819, 2007.

[14] LINGO, The modeling Language and optimizer, Chicago, LINDO Systems Inc, 2013.

Multi parameter optimization using Taguchi L_8 (2^7) Array - A case study on additive paper lamination process

S. Karagiannis[a], T. Ispoglou[b], P. Stavropoulos[c] and J. Kechagias[d*]

***Abstract*— Robust design is applied in the current study aiming in the prediction of layer thickness deformation during Laminated Object Manufacturing (LOM) process. Prediction of layer thickness deformation is of importance for quality characterization and build time estimation. Eight different experiments, with seven process parameters each of two levels of detail, have been conducted following an L_8 (2^7) orthogonal array. Results indicate that the layers thickness deformation is affected mainly by the layer thickness and heater speed. A linear regression model has been applied on the experimental results and tested using evaluation experiments giving accurate predictions.**

***Keywords*— Taguchi design, layer thickness deformation; regression modelling**

.

I. Introduction

In the LOM process physical prototypes are built by sequentially laminating, bonding and cutting 2-dimensional cross-sections generated by the horizontal slicing of a CAD model. The material used is ordinary paper with a thin layer of thermoplastic adhesive film on one side. The bonding process is accomplished by applying heat and pressure from a heated cylinder rolling along the sheet. Then, a laser beam is used to cut the area of each layer in three different sections: part perimeter, hatching area and supporting frame perimeter. Finally, any waste material, which is formed into cubes by the laser, is removed once the build process is completed [1]. The LOM process builds large physical prototypes faster than other methods [2, 3] and gives sufficient quality characteristics [4], and tensile strength in the laminates direction [5]. On the other hand the LOM process exhibits sheet-bonding problems [6], which cause process malfunctioning, weak bonding, difficult disengagement between the supporting frame and part, and unequal dimensional accuracy on X, Y, and Z directions [7]. Also, the orientation of the LOM parts, as well as other process variables, affects the final surface quality [1-8]. This article examines the influence of different process parameters onto layer thickness deformation using the Taguchi experimental design and analysis [9]. Matrix experiments were conducted and an analysis of means (ANOM) and an analysis of variances (ANOVA) were carried out in order to investigate the LOM process parameters effect onto the layer thickness deformation. The process parameters tested were the layer thickness (LT), heater temperature (HT), platform retract (PR), heater speed (HS), laser speed (LS), feeder speed (FS) and platform speed (PS). Finally, a linear regression mathematical model is applied on the experimental results and it was tested using evaluation experiments giving accurate predictions.

[a] S. Karagiannis is with the, Department of Mechanical Engineering and Industrial Design, Technological Education Institution of Western Macedonia, Kozani 50100, Greece (email: SKaragiannis@teikoz.gr)

[b] T. Ispoglou is with the, Department of Mechanical Engineering and Industrial Design, Technological Education Institution of Western Macedonia, Kozani 50100, Greece (email: ispoteo@hotmail.com)

[c] P. Stavropoulos is with the Department of Aeronautical Studies, Hellenic Air Force Academy, Dekelia Air-Force Base, 1010 Athens, Greece (email: pstavropoulos.hafa@haf.gr)

[*d] J. Kechagias is with the Department of Mechanical Engineering, Technological Educational Institute of Thessaly, Larissa 41110, Greece (corresponding author: phone: 0030 2410684322, fax: 0030 2410684305, email: jkechag@teilar.gr)

II. Taguchi Design

The Taguchi design method is a simple and robust technique for optimizing the process parameters. In this method, the main parameters, which are assumed to have an influence on the process results, are located in different rows in a designed orthogonal array (orthogonal matrix experiment). With such an arrangement randomized experiments can be conducted. In general, the signal to noise (S/N) ratio (n, dB) represents the quality characteristics of the data observed in the Taguchi design of experiments. In the case of layer thickness deformation, lower values of S/N ratios are desirable (smaller-the-better) and the objective function is defined as follows:

$$\eta = -10\log_{10}\left[\frac{1}{n}\sum_{i=1}^{n} y_i^2\right] \quad (1)$$

where y_i is the observed data at the i^{th} trial and k is the number of trials. From the S/N ratio, the effective parameters having an influence on the process results can be seen and the optimal sets of process parameters can be determined. The set of process parameters and their corresponding levels which were used in the current experimental design are illustrated in Table1. The parameter levels define the experimental area of interest. All the other values of the control parameters and the LOM machine preparation procedure are described in previous work. [3]. The actual height (Z_{max}), and total number of layers (N_{act}) were measured for each experiment. Then the actual layer thickness (ALT) was calculated as follows:

$$ALT=Zmax/Nact \quad (2)$$

The layer thickness deformation (LTD) was measured by the formula:

$$LTD=100*(ALT-LT)/LT\ (\%) \quad (3)$$

where LT is the nominal layer thickness of paper used before for the part production. The matrix experiment selected for this project is given in table 2. It consists of eight individual experiments corresponding to the eight rows. The seven columns of the matrix represent the seven parameters as indicated in the Table 2. The entries in the matrix represent the levels of the parameters.

	Process Parameters	Abb.	Levels	
1	Layer Thickness (mm)	LT	0.1	0.2
2	Heater Temperature (°C)	HT	170	190
3	Platform Retract (mm)	PR	0.3	0.4
4	Heater Speed (mm/sec)	HS	70	140
5	Laser Speed (mm/sec)	LS	150	180
6	Feeder Speed (mm/sec)	FS	50	100
7	Platform Speed (mm/sec)	PS	25	50

Table 1: Process parameters and their levels

No	LT	HT	PR	HS	LS	FS	PS	N_{act}	ALT	LTD
1	0.107	170	0.3	70	150	50	25	60	0.116	8.4%
2	0.107	170	0.3	140	180	100	50	56	0.125	16.8%
3	0.107	190	0.4	70	150	100	50	61	0.115	7.5%
4	0.107	190	0.4	140	180	50	25	57	0.123	15.0%
5	0.203	170	0.4	70	180	50	50	34	0.206	1.5%
6	0.203	170	0.4	140	150	100	25	30	0.233	14.8%
7	0.203	190	0.3	70	180	100	25	35	0.2	-1.5%
8	0.203	190	0.3	140	150	50	50	34	0.206	1.5%
mean (mm)										8.0%

Table 2: Matrix experiment ($L_8(2^7)$) and measurements

III. RESULTS AND ANALYSIS

The mean values of each parameter (S/N ratios of seven parameters according to each level) are shown in Table 3 for the actual layer thickness. The higher the difference between the mean values the higher the effect on the quality characteristic. A primary goal in conducting matrix experiments is to optimise the product or process design – that is, to determine the best or the optimum level for each parameter. The optimum level of a parameter is the level that gives the lower or maximum value of quality characteristic in the experimental area. The effects of process parameters can be seen in fig. 1. The results of the analysis of variance (ANOVA) are shown in table 4, respectively. Based on the statistical analysis of the experimental results, the layer thickness deformation is affected by the layer thickness, hater speed, heater temperature, and platform retract by ((35.5%), (37.2%), (13.1%) and (6.5%), respectively. According to the ANOVA analysis, laser speed, feeder speed and platform speed have a minimum effect on layer thickness deformation. Eliminating the laser speed, feeder speed and platform speed for layer thickness deformation, the error variance is calculated at 0.0009 (Table 4). Thus, the width of the two-standard–deviation confidence interval, which is approximately 95 percent of the confidence interval for each estimated effect, is:

$$e = \pm 2 \cdot \sqrt{\frac{1}{2} \cdot 0.0009} = \pm 4.2\% \quad (4)$$

		LTD	
	Abb.	Lev. 1	Lev. 2
1	LT_i	11.9%	4.1%
2	HT_j	10.4%	5.6%
3	PR_k	6.3%	9.7%
4	HS_l	4.0%	12.0%
5	LS_m	8.0%	7.9%
6	FS_n	6.6%	9.4%
7	PS_k	9.2%	6.8%

.**Table 3:** Mean values of each parameter level

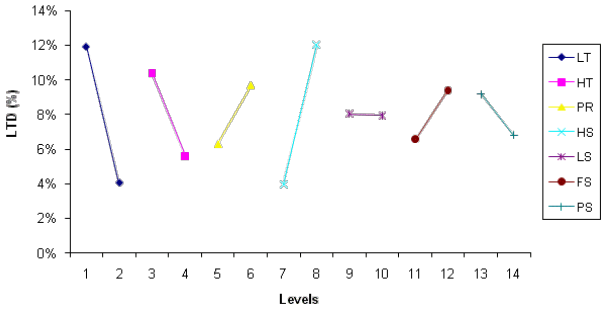


Fig 1. Effect of each parameter on layer thickness deformation

ISBN: 978-1-61804-256-9

	DOF	Sum of squares	Mean square	F	%
LT	1	0.0123	0.0123	13.71	35.5%
HT	1	0.0045	0.0045	5.05	13.1%
PR	1	0.0023	0.0023	2.51	6.5%
HS	1	0.0129	0.0129	14.36	37.2%
LS	1	0.0000	0.0000	0.00	0.0%
FS	1	0.0016	0.0016	1.77	4.6%
PS	1	0.0011	0.0011	1.23	3.2%
Error	0	0.0000			
Total	7	0.0347			
(Error)	**(3)**	**(0.0027)**	**(0.00090)**		

Table 4. ANOVA table for layer thickness deformation

IV. REGRESSION MODELLING AND EVALUATION

Regression modelling uses statistical and mathematical methods to quantify the relationship between the process parameters and the quality indicator obtained. Assuming that the process parameters are continuous and controllable in experiments, the response can be expressed as follows

$$LTD_{pred} = b_1 + b_2 LT + b_3 HT + b_4 PR + b_5 HS + b_6 LS + b_7 FS + b_8 PS \pm e \quad (5)$$

where, LTD_{pred} is the response of the layer thickness deformation, b_i, coefficients, which should be determined, and e is the expected error. In general Eq. (5) can be written in a matrix form.

$$Y = bX + E \quad (6)$$

where, Y is defined to be a matrix of the measured values, X to be a matrix of the process parameters and their products. The matrix b and E consist of coefficients and errors, respectively. The solution of Eq. (6) can be obtained by the matrix approach.

$$b = (X^T X)^{-1} X^T Y \quad (7)$$

where, X^T is the transpose of the matrix X and $(X^TX)^{-1}$ is the inverse of the matrix X^TX. From the observed data listed in Table 2 and Eq. (7), the b_i coefficients of the Eq. (5) are shown in Table 5.

Coefficients	
$\mathbf{b_1}$	0,395369791666683
$\mathbf{b_2}$	-0,817708333333344
$\mathbf{b_3}$	-0,00238250000000004
$\mathbf{b_4}$	0,336499999999989
$\mathbf{b_5}$	0,00114785714285714
$\mathbf{b_6}$	-3,16666666666520e-05
$\mathbf{b_7}$	0,000563999999999994
$\mathbf{b_8}$	-0,000940000000000008

Table 5: b_i coefficients of eq. (5)

Param.	Values
LT	0.203
HT	190
PR	0.3
HS	140
LS	180
FS	100
PS	50
LTD (actual)	**1.9%**
LTD (predicted)	**4.2%**

Table 6: Evaluation Experiments

A confirmation experiment was conducted in order to evaluate the above model (Table 6) and the result shows that the prediction is within the confidence intervals that the methodology gave (less than 4.2% difference between the actual and the predicted value).

V. CONCLUSIONS AND FUTURE APPLICATIONS

The layer thickness deformation was investigated according to an orthogonal array. The following was concluded

- The layer thickness deformation can be predicted accurately using the Taguchi design of experiment methodology
- The analysis of variances shows that the layer thickness deformation is affected mostly by the heater speed, layer thickness, heater temperature and platform retract.
- Using the extracted regression model (eq. 5), predictions of the actual layer thickness can be made.
- Using the extracted model, accurate predictions of total number of layers needed can be made and consequently the prediction of total build time is possible.

Future work will incorporate the above analysis results onto the LOM built time algorithm, which was described in previous work [1] in order to improve the build time estimations.

ISBN: 978-1-61804-256-9

ACKNOWLEDGMENT

This research is implemented through the Operational Program "Education and Lifelong Learning" and is co-financed by the European Union (European Social Fund) and Greek national funds.

REFERENCES

[1] Kechagias, J., Maropoulos, S., Karagiannis, S. (2004) "Process build-time estimator algorithm for laminated object manufacturing", Rapid Prototyping Journal, Vol. 10, No 5, pp. 297-304.

[2] Kechagias, J. (2007a) "An experimental investigation of the surface roughness of parts produced by LOM process" Rapid Prototyping Journal, Vol. 13, No 1, pp 17-22.

[3] Kechagias, J. (2007b) Investigation of LOM process quality using design of experiments approach", Rapid Prototyping Journal, Vol. 13, No. 5, pp. 316-323.

[4] Kruth, J.P. (1991) "Material Incress Manufacturing by Rapid Prototyping Techniques", CIRP Annals, Vol. 40, No 2, pp. 603-614.

[5] Chryssolouris, G., Kechagias, J., Moustakas, P., Koutras, E. (2003) "An Experimental Investigation of the Tensile Strength of Parts Produced by Laminated Object Manufacturing (LOM) Process", CIRP J Manuf Systems, Vol 32(5).

[6] Sonmez, F., Hahn, T. (1998) "Thermomechanical analysis of the Laminated Object Manufacturing Process", Rapid Prototyping Journal, Vol. 4, No 1, pp 26-36.

[7] Park, J.; Tari, M.J.; Hahn, H.T. (2000) "Characterization of the laminated object manufacturing (LOM) process", Rapid Prototyping Journal, Vol. 6, No 1, pp.36-49.

[8] Campbell, R.I., Martorelli, M., Lee, H.S. (2002) "Surface roughness visualisation for rapid prototyping models", CAD Comp Aided Design, Vol 34, No 10, pp. 717-725.

[9] Phadke, M., S. (1989), Quality Engineering using robust design, Prentice hall, EnglewoodCliffs, New Jersey 07632, ISBN 0-13-74-5167-9.

Automated threshold selection for parametric and non-parametric estimates of intensity-duration-frequency curves

Jan Holešovský, Michal Fusek, and Jaroslav Michálek

***Abstract*—**When dealing with extreme values estimation, the threshold models are often used. However, a proper threshold selection is one of the problems which have to be solved. In this paper, we concentrate on this issue in order to compare an automated threshold selection based on multiple-threshold model with double bootstrap technique based on semi-parametric estimation. A case study is carried out to evaluate estimations of intensity-duration-frequency curves which represent commonly used hydrological tool. A special attention is also paid to the assessment of an impact of the series length on the estimation quality.

***Keywords*—**Extreme value, moment estimator, bootstrap, peaks-over-threshold, generalized Pareto distribution.

I. INTRODUCTION

A common objective in the extreme value analysis is to gain information about tails of a probability distribution. This is necessary especially in environmental sciences and engineering where extreme and rare events are of interest. One of the main goals of an extreme value analysis is to estimate the T-year return level z_T, i.e. value which is exceeded on average once every T years. Let $X_1, X_2, \dots, X_n$ be a sequence of independent and identically distributed (iid) random variables and $M_n = \max\{X_1, \dots, X_n\}$. Fisher and Tippett [1] showed that, given a sequences $a_n > 0$ and b_n, the only one non-degenerate distribution of properly normalized block maxima $(M_n - b_n)/a_n$ arises in the form of generalized extreme value (GEV) distribution with cdf

$$G(x) = \exp\left[-\left\{1+\xi\frac{(x-\mu)}{\sigma}\right\}_+^{-1/\xi}\right], \qquad (1)$$

where $\mu \in \Re, \sigma > 0$ and $\xi \in \Re$ are location, scale and shape parameters respectively, and $x_+ = \max(0, x)$. The shape parameter, also referred to as the extreme value index (EVI), plays a crucial role in relation to tail properties, and thus needs to be properly estimated.

In practice, mostly when working with observations over time, a significant dependence structure is often present [2]. Since the estimation procedures, for example the commonly used maximum likelihood (ML) method, are based on iid observations, it is usually necessary to draw samples from a series that can be considered independent. A simple method of drawing samples consists of separating the underlying series into blocks. Considering a block size large enough (usually one year), the obtained block maxima can be modelled by GEV distribution [3], [4].

However, modeling of extremes using block maxima is often unsuitable for practical purposes, because it leads to an extensive reduction of information contained in data, especially if only short time series are available. Therefore, the extreme value analysis is often based on threshold exceedances. Pickands [5] showed that given a sequence u_n of thresholds increasing with n, the limiting distribution of threshold exceedances of a random variable X is the generalized Pareto (GP) distribution. In practice, a high enough threshold u is selected and kept fixed. The variable $Y = X - u$ conditioned by $X > u$ is modelled by the GP distribution with cdf

$$H(y) = 1 - \left(1+\frac{\xi y}{\sigma_u}\right)_+^{-1/\xi}, \xi \neq 0, \qquad (2)$$

where $y = x - u > 0$, ξ and $\sigma_u > 0$ is shape and scale parameter respectively. In correspondence to GEV distribution (1), the relation $\sigma_u = \sigma + \xi(u - \mu)$ holds.

Although the GP-analysis allows us to make use of more observations than the GEV-analysis, selecting a proper threshold can be problematic. A threshold too low leads to desired smaller uncertainty in parameters' estimates, however, it provides a possibly insufficient approximation by the limiting GP distribution, and vice versa.

The purpose of this paper is to compare two lately presented approaches to the automated threshold selection – a multiple-threshold penultimate model based on piecewise constant

This work was supported by the specific research project No. FCH/FSI-J-14-2439 at Brno University of Technology.

J. Holešovský is with the Institute of Mathematics, Faculty of Mechanical Engineering, Brno University of Technology, Technická 2896/2, Brno 61669, Czech Republic (corresponding author to provide phone: 00420-541-142-726; e-mail: honza.holesovsky@gmail.com).

M. Fusek is with the Department of Mathematics, Faculty of Electrical Engineering and Communication, Brno University of Technology, Technická 2848/8, Brno 61600, Czech Republic (e-mail: fusekmi@feec.vutbr.cz).

J. Michálek is with the Department of Econometrics, Faculty of Military Leadership, University of Defence, Šumavská 4, Brno 66210, Czech Republic (e-mail: michalek@fme.vutbr.cz).

shape parameter approximation [6], and a bootstrap-based methodology developed for an optimal sample fraction estimation [7]. Both techniques are applied to estimate intensity-duration-frequency (IDF) curves, which play an important role in hydrological risk assessment.

II. SAMPLE DRAWING AND PROPER SAMPLE FRACTION IDENTIFICATION

Continuous rainfall series with the time step of 1 minute from stations located in the South Moravian Region, Czech Republic, were used for the analysis. Since the observations cannot be considered independent, a pre-processing needs to be applied before carrying on with the analysis. The separating procedure involves identification of the independent rainfall events from which only maxima of mean intensities over different rainfall durations (5, 10, 15, 20, 30, 45, 60, 90, 120, 180, 240 and 360 minutes) are considered. For the purpose of our study, the methodology of Madsen et al. [8] was applied, and the events were separated by ceasing the rainfall for period equal to the considered duration but at least one hour. Together with threshold-based modelling of extremes, this is the well-known peaks-over-threshold (POT) approach [3], [4].

Since some of the stations were established only in the last decades, lengths of the available series vary between 11 and 41 years of records. We aim to study the impact of the chosen threshold estimation methodology on the IDF curves estimates with respect to the series length which can significantly influence the estimates [9], [10].

A. The Multiple-Threshold GP Model

The traditional diagnostic tool for threshold identification consists in graphical inspection of parameter estimates and their stability. Should u_0 be a proper threshold to be selected, then parameter estimates (changing with threshold) $\hat{\xi}(u)$ and $\hat{\sigma}(u)$ would follow a constant and linear trend for $u > u_0$ respectively [3]. Northrop and Coleman [6] proposed an automated method based on the mentioned properties of the shape parameter. In principle, they proposed a discretized version of testing the null hypothesis $H : \xi(u) = \xi(u_0)$, for all $u \geq u_0$.

Let $u_1 < u_2 < \ldots < u_m$ denote an increasing threshold sequence, Y be an excess of u_1, and denote $v_i = u_i - u_1$, for $i = 1, \ldots, m$. The shape parameter is modelled as a piecewise constant function $\xi(y)$ with change-points at $v_i, i = 2, \ldots, m$, i.e.

$$\xi(y) = \begin{cases} \xi_i, & u_i < y < u_{i+1}, i = 1, \ldots, m-1, \\ \xi_m, & y > u_m. \end{cases} \quad (3)$$

In order to avoid discontinuity in density of Y, the scale parameter is considered piecewise linear with σ_1 on interval (u_1, u_2) (for details see [6]). Considering the threshold u_1 first, they test the hypothesis whether a common GP model holds on all intervals $(u_i, u_{i+1}), i = 1, \ldots, m$, that is $H : \xi_1 = \ldots = \xi_m$. Rejection of the null hypothesis attests to requirement of a higher threshold. Let $\hat{\theta}_0$ denote the sequence of ML estimates of shape ξ_1 and scale σ_1 parameters under the null hypothesis, and $\hat{\theta}$ the sequence of ML estimates of σ_1 and $\xi_i, i = 1, \ldots, m$ given by (3). Northrop and Coleman proposed a test based on score test statistic

$$S = U^T(\hat{\theta}_0) \cdot J^{-1}(\hat{\theta}_0) \cdot U(\hat{\theta}_0), \quad (4)$$

where $U(\theta)$ is the score function and $J(\theta)$ is the Fisher information matrix. More details about derivation of the score function, information matrix, and likelihood function can be found in [6]. Provided that $\xi_m > -1/2$ [11] in each case the asymptotic null distribution of the statistic (1) is χ^2_{m-1}. Assume now that the lowest threshold considered is u_k. The null hypothesis to be tested is $H : \xi_k = \ldots = \xi_m$ and the asymptotic null distribution of statistic (4) is χ^2_{m-k}, $k = 1, \ldots, m-1$. The p-values associated with the test indicates whether a threshold higher than u_k is required.

Once a proper value u_0 is selected, a T-year return level z_T is estimated from (2) by $1 - 1/(n_x T)$ quantile of a GP distribution, where n_x denotes the number of observations per year, i.e.

$$\hat{z}_T = u + \frac{\hat{\sigma}_u}{\hat{\xi}} \left[(\hat{\lambda}_u T n_x)^{\hat{\xi}} - 1 \right], \quad (5)$$

$\hat{\lambda}_u := n_u / n \approx P(X > u)$, where n_u denotes the number of observations above the threshold u_0. In (5) all estimated parameter were replaced by their ML estimates, and the asymptotic confidence intervals for $\hat{z}_T$ can be obtained by delta method [3].

B. Bootstrap-Based Optimal Sample Fraction Selection

A different approach to optimal sample fraction identification was studied by Draisma *et al.* [12]. This method is based on k largest order statistics which should in some sense balance the approximation by the GP distribution and properties of the EVI. Here, we focus on moment estimator (MoE) $\hat{\xi}_M$ defined as

$$\hat{\xi}_M(k) = M_n^{(1)}(k) + 1 - \frac{1}{2}\left[1 - (M_n^{(1)}(k))^2 / M_n^{(2)}(k)\right]^{-1}, \quad (6)$$

where

$$M_n^{(j)}(k) = \frac{1}{k} \sum_{i=0}^{k-1} \left(\log X_{(n-i)} - \log X_{(n-k)} \right)^j, \quad (7)$$

is the j-th moment of the GP distribution, $k = 2, \ldots n-1$, and $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ denotes the order statistics. The choice of a proper k is accompanied by difficulties very similar to the choice of an optimal threshold in the POT analysis. When a high value of k is selected, we may expect higher estimation precision but a weak approximation by the GPD which may lead to a significant bias, and vice versa. The optimal value k_0 balancing the bias and variation can be

determined only if the underlying distribution is known. As shown for example in [13], the limiting distribution of $\hat{\xi}_M$ depends on a second-order parameter which is problematic to estimate. Therefore, the double bootstrap methodology proposed in [12] and developed in [7], [14] is used.

Hereby, the optimal value k_0 is chosen to minimize the mean square error (MSE) of MoE

$$k_0 \in \arg\min_k \mathrm{E}(\hat{\xi}_M(k) - \xi)^2, \quad (8)$$

although only in the asymptotic sense. The unknown EVI is replaced by an auxiliary estimator ξ_{AUX} calculated as the third-moment estimator (see [3], page 5).

The bootstrap methodology is used to resample B times independently a sample $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$ of length $m < n$ from the underlying series $\mathbf{X} = (X_1, \ldots, X_n)$ of separated (independent) rainfall observations (see Sect. 2). For each replication and for all $k = 2, \ldots, m$, the estimators $\hat{\xi}_M^*(k)$ and $\hat{\xi}_{AUX}^*(k)$ are evaluated. Finally, the MSE is replaced by its estimator

$$MSE^*(m,k) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\xi}_M^*(k) - \hat{\xi}_{AUX}^*(k)\right)^2, k = 2, \ldots, m, \quad (9)$$

and minimized with respect to k. Denote the optimal value k which minimizes (9) by $k_0^*(m)$. The key step is to apply the bootstrap methodology twice: Firstly, for $m := n_1$; secondly, for $m := n_2$. As derived in [12] or [7], by setting $n_2 := \left\lfloor (n_1)^2 / n \right\rfloor$, the optimal sample fraction $\hat{k}_0^*$ is obtained using formula

$$\hat{k}_0^* = (k_0^*(n_1))^2 / k_0^*(n_2). \quad (10)$$

From the theoretical point of view, n_1 should be smaller than n; however, simulations show better performance if n_1 is set as high as possible [7]. In order to achieve better estimation stability, the whole double bootstrap procedure is applied repeatedly N times leading to N estimates $\hat{k}_{0,i}^*, i = 1, \ldots, N$ of optimal sample fraction (10). The estimator $\hat{k}_0$ of k_0 is computed as median from all the bootstrapped values of $\hat{k}_{0,i}^*$.

Once the optimal value $\hat{k}_0$ is determined, shape parameter is estimated by MoE, and scale parameter is estimated using formula (see Theorem 4.3.3 in [13])

$$\hat{\sigma}_M := X_{(n-k)} M_n^{(1)} \left(1 - \hat{\xi}_M + M_n^{(1)}\right) \quad (11)$$

Standard deviations of the estimated parameters are estimated on the basis of asymptotic normality [13]. The desired return level estimates are obtained from formula (5) where parameter estimates are replaced by the appropriate moment estimators and the threshold is replaced by $X_{(n-\hat{k}_0)}$.

III. RESULTS

In this section, results for stations with the longest (Tuřany, 41 years) and shortest (Jundrov, 11 years) series will be presented. Both stations are located in the highly urbanized area of Brno, the second largest city in the Czech Republic.

In multiple-threshold GP (MT-GP) model (Sect. 2.1), a balance between the level of discretization and the estimation quality has to be found. On one hand, it is good to set m as high as possible in order to ensure a reasonable approximation of the shape parameter by piecewise linear function. On the other hand, too high values imply either high variability or even convergence problems of the ML method. Since the value of m can significantly influence rejection of the tested hypotheses, a mode complex study should be carried out in order to assess this difficulty. In our case, we choose a defensive approach regarding several proposals in [6], and we set m=10, 20, and 40 thresholds between 10% and 99.5% sample quantile for all stations and rainfall durations. For all values of m, the value $u_0(m)$ is selected as the lowest threshold that ensures the validity of the null hypothesis at the significance level of 0.05. Then the optimal threshold is chosen as $u_0 = \max_m(u_0(m))$. In Fig. 1., typical results

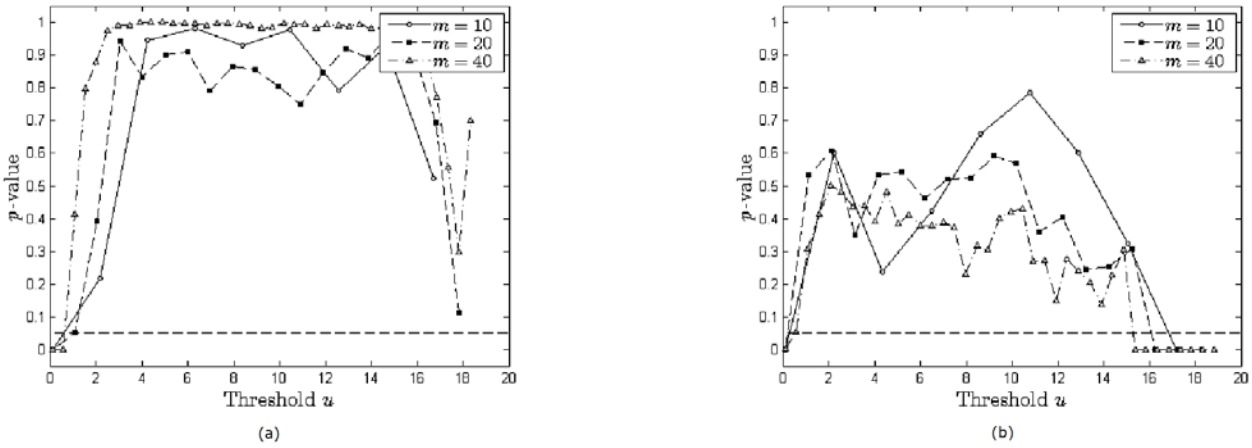


Fig. 1 Obtained p-values of tested hypotheses at Tuřany (a) and Jundrov (b) Station for various values of m. Horizontal dashed line shows significance level of 0.05.

Table I Tuřany Station. Number of observations obtained by the double bootstrap ($\hat{k}_0$) and the MT-GP model (n_u).

Duration [min]	*5*	*10*	*15*	*20*	*30*	*45*	*60*	*90*	*120*	*180*	*240*	*360*
k_0	174	186	243	215	209	111	111	225	245	186	221	278
n_u	423	540	418	556	725	776	804	807	777	740	725	689
k_0 / n_u	0.41	0.34	0.58	0.39	0.29	0.14	0.14	0.29	0.32	0.25	0.30	0.40

obtained from testing the null hypothesis at various thresholds are visualized, specifically the p-values obtained from limiting χ^2 distribution are plotted against the considered thresholds.

Further, ML estimates of parameters of the GP distribution were calculated for all threshold values. Their variances were established on the basis of asymptotic normality, and confidence intervals of return levels were calculated using the delta method [3].

Next, an optimal sample fraction using the double bootstrap methodology (Sect. 2.2) was determined. We set $N = 1000$ and $B = 250$ bootstrap replications, which were considered sufficient (see [14]). As it was mentioned before, it is convenient to set the value of n_1 as high as possible providing $n_1 < n$. In our study, we use $n_1 = \lfloor n^{0.995} \rfloor$ (see [7]). The optimal sample fractions estimated using the MT-GP model and the double bootstrap methodology are summarized in Tables I and II. Notice that use of the MT-GP model leads to lower 'thresholds' implicating larger sample fraction meant for the analysis.

Dependency of the parameter estimates on the selection of k largest order statistics is visualized in Fig. 2. On one hand, setting value k low provides better approximation of the limiting GP distribution. However, it can be seen that it is loaded with higher uncertainty. On the other hand, high k leads to a significant bias. The standard deviations of shape parameter estimates are presented in Table III. It can be seen that the moment estimate is burdened with higher variability than that based on the MT-GP model. A similar behavior was observed in case of the scale parameter.

On the basis of estimated parameters, return levels of 5, 10, 20, 50, and 100-year events were calculated. Similar results were obtained by both approaches in case of the Tuřany station (long series) and lower (< 30 minutes) rainfall durations. For longer rainfall durations, return levels estimated using the MT-GP approach overestimate the bootstrap-based estimates. In case of the Jundrov station (short series), the MT-GP approach gives higher estimates than the bootstrap-based one for all rainfall durations. The bootstrap-based estimates provide narrower confidence intervals for the estimated IDF curves as documented in Fig. 3.

IV. CONCLUSION

In this paper, two automated threshold selection procedures were introduced, compared and used for IDF curves estimation. It was shown that moment estimates of the GP distribution parameters are burdened with higher variability than those based on the MT-GP model. However, the bootstrap-based estimates provide narrower confidence intervals for the IDF curves. Readers can also get an idea about the width of the confidence intervals when dealing with estimation of long-time events in situations where only short rainfall series are available. Although the automation of proper sample fraction identifications through the double bootstrap technique is very comfortable, it is more demanding in terms of computing. To properly compare both threshold selection

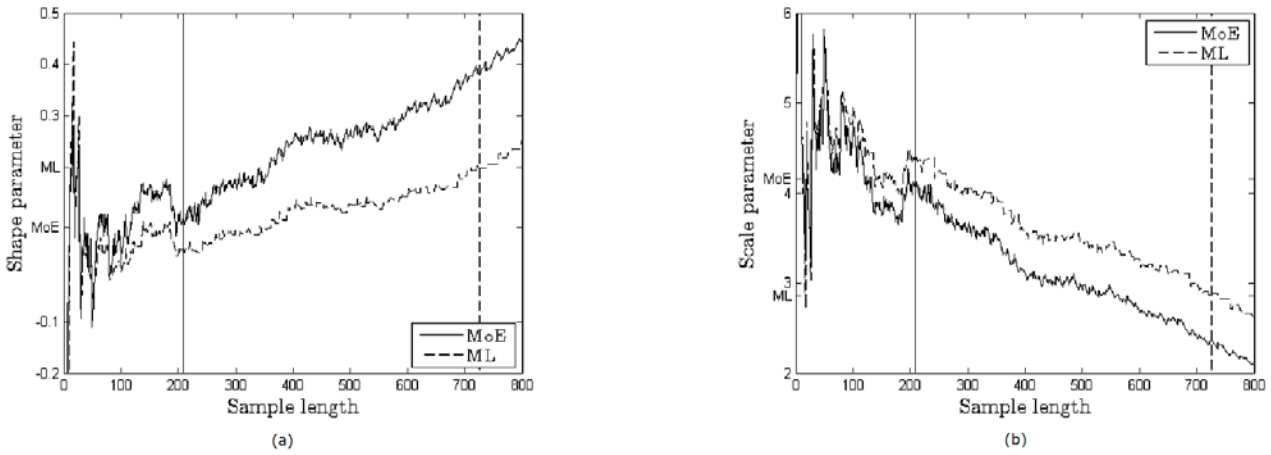


Fig. 2 Dependency of moment and ML estimators of shape (a) and scale (b) parameters on the sample fraction for 30-minutes rainfall at Tuřany station. Vertical lines visualize the optimal values of k (solid) and n_u (dashed).

Table II Jundrov Station. Number of observations obtained by the double bootstrap ($\hat{k}_0$) and the MT-GP model (n_u).

Duration [min]	***5***	***10***	***15***	***20***	***30***	***45***	***60***	***90***	***120***	***180***	***240***	***360***
k_0	56	41	44	50	69	70	83	105	90	103	115	120
n_u	192	205	210	211	207	216	191	194	197	205	206	196
k_0 / n_u	0.29	0.20	0.21	0.24	0.33	0.32	0.43	0.54	0.46	0.50	0.56	0.61

Table III Proportions of estimated standard deviations for shape parameters (MoE/MT-GP) at both stations.

Duration [min]	***5***	***10***	***15***	***20***	***30***	***45***	***60***	***90***	***120***	***180***	***240***	***360***
Tuřany	1.32	1.46	1.24	1.37	1.41	1.94	2.06	1.61	1.62	1.86	1.62	1.50
Jundrov	1.15	1.47	1.45	1.47	1.26	1.36	1.25	1.12	1.31	1.30	1.18	1.17

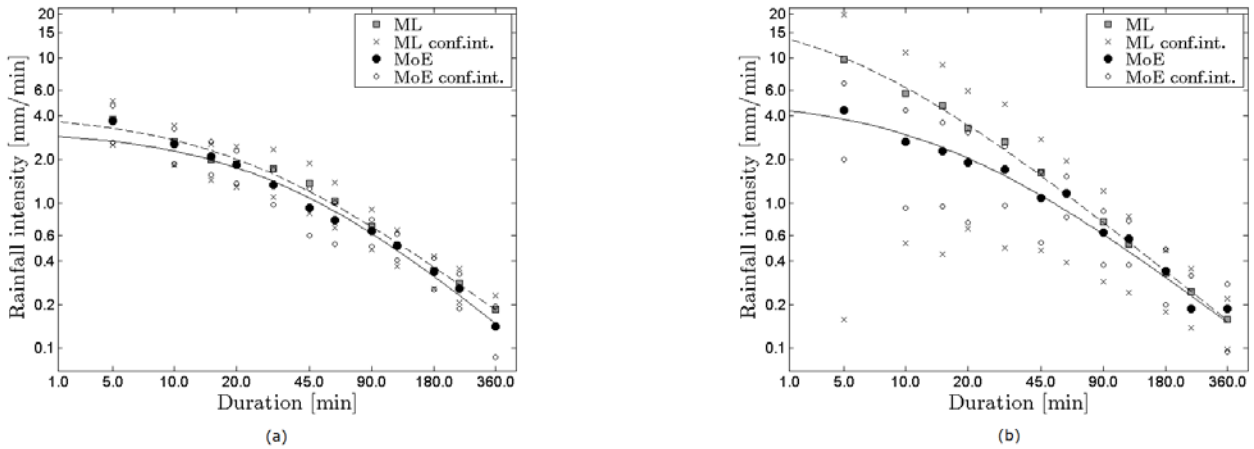


Fig. 3 Estimated 100-year return levels at Tuřany (a) and Jundrov (b) stations with their 95% confidence intervals; logarithmic scale on both axes. IDF curves were obtained using the non-linear regression.

procedures, a thorough simulation study needs to be carried out.

REFERENCES

[1] R. A. Fisher, L. H. C. Tippett, "Limiting forms of the frequency distribution of the largest or smallest members of a sample," in *Proc. of Cambridge Philosophical Society 24*, 1928, pp. 180–190.

[2] J. Holešovský, M. Fusek, J. Michálek, "Extreme value estimation for correlated observations," in *Proc. 20th MENDEL Conf.*, 2014, pp. 359–364.

[3] S. G. Coles, *An introduction to statistical modelling of extreme Values.* London, Springer, 2001.

[4] M. N. Khaliq, T. B. M. J. Ouarda, J.-C. Ondo, P. Gachon, B. Bobèe, "Frequency analysis of a sequence of dependent and/or non-stationary hydro-Meteorological observation: A review," in *J. Hydrol.*, vol 329, no. 3-4 **329**(3-4), 2006, pp. 534–552.

[5] J. Pickands, "Statistical inference using extreme order statistics," in *Ann. Stat.*, vol. 3, 1975, pp. 119–131.

[6] P. J. Northrop, C. L. Coleman, "Improved threshold diagnostic plots for extreme value analyses," in *Extremes*, vol 71, 2014 pp. 289–303.

[7] M. I. Gomes, O. Oliveira, "The Bootstrap methodology in statistics of extremes – Choice of the Optimal Sample Fraction," in *Extremes*, vol. 4, no. 4, 2002, pp. 331–358.

[8] H. Madsen, P. S. Mikkelsen, D. Rosbjerg, P. Harremöes, "Regional estimation of rainfall intensity-duration-frequency curves using generalized least squares regressions of partial duration series," in *Water Resour. Res.*, vol 38, no. 11, 2002, pp. 21-1—21-11.

[9] A. Ben Zvi, "Rainfall intensity-duration-frequency curves relationships derived from large partial duration series," in *J. Hydrol.*, vol 367, no. 1-2, 2009, pp. 104–114.

[10] P. Willems, "Revision of urban drainage design rules after assessment of climate change impacts on precipitation extremes at Uccle, Belgium," in *J. Hydrol.*, vol. 496, 2013, 166–177.

[11] R. L. Smith, "Extreme value theory based on the r largest annual events," in *J. Hydrol.*, vol. 86, no. 1-2, 1986, pp. 27–43.

[12] G. Draisma, L. de Haan, L. Peng, T. T. Pereira, "A Bootstrap-based method to achieve optimality in estimating the extreme-value index," in *Extremes,* vol. 2, no. 4, 1999, pp. 367–404.

[13] L. de Haan, A. Ferreira, *Extreme value theory: An introduction.* New York, Springer, 2006.

[14] F. Caeiro, and M. I. Gomes, "Semi-parametric tail inference through probability-weighted moments," *J. Stat. Plan. Infer.*, vol. 141, pp. 937–950, 2010.

[15] B. Efron, R. Tibshirani, *An Introduction to the bootstrap.* Boca Raton, Chapman, 1994.

Using Random Hypernets for Reliability Analysis of Multilayer Networks

Alexey Rodionov, and Olga Rodionova

***Abstract*—The general approach to constructing structural models of non-stable multi-level networks is proposed. This approach is based on hypernets relatively new mathematical object, which is successively used for modeling different multi-level networks in the Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Russia, for last 30 years. Hypernets allow standard description of neighboring levels interconnection in a mathematically correct way. Using this mathematical object allows easy modifications of data with model changing and/or development and efficiently organize data search for different computational or optimization algorithms. Optimization of mapping of secondary (logical) network onto structure of unreliable primary (physical) network is considered as example.**

***Index Terms*—multilevel networks; modeling; hypernets, reliability analysis**

I. Introduction

RANDOM graphs and hypergraphs are usually used for the reliability analysis of unreliable networks [12], [13], [15]–[18], etc. This model seems quite appropriate as a structural model for analysis of structural and functional reliability of information networks: failures of nodes or links are simulated by removal of correspondent vertexes or edges (arcs) of a random graph with given probabilities, or by reducing their throughputs. At the same time, in many cases modeling network by a random graph is not sufficient, which can be shown by the following example of two-level network. Let us have a cable network that is presented by a graph G_1, and let us have some data-transferring network realized inside it. This network can be presented by some graph G_2, that not necessarily coincides with G_1. We will name G_1 as primary network (PN) while G_2 we will name secondary network (SN). Laying of G_2 into G_1 may be done by different ways (if G_1 is not a tree). Thus, we have some mapping of links of SN onto paths constructed from links of PN (further, we will name these links as branches). Obviously, failure or change of throughput of a branch may lead to failure or change of throughput of several SNs links or may not touch any of them at all. So, for analyzing multi-level networks more complicated models than random graphs are needed.

We can find different descriptions of such models: bigraphs [1], sandwiching graphs [2], graphs with different kinds of edges [3], descriptions on the application level [4]–[6], layered complex networks (LCN) [7], etc. All these models take into account different connections between layers, but all of them (may be excluding LCN) are not universal. Even LCN model consider mapping of neighbor layers only. Yet for more than 30 years the hypernet model is successively used for modeling multi-layer embedded networks of different nature in several Russian, Kirghiz and Kazakh universities. Unfortunately, until now most of papers and books with description of the model and its applications are in Russian (main monograph is [8]), but there are some conference publications in English in which the hypernet models are used also [9]–[11]. Hypernets allow adequately describe multilevel networks with an arbitrary number of levels. In this paper we discuss the simplest case of two-level networks and its usage for reliability analysis and optimization.

This work is supported by the grant of the Program of basic researches of the Presidium of Russian Academy of Science.

Alexey Rodionov is with the Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Lavrentjeva ave. 6, Novosibirsk, 630090, Russia, e-mail: alrod@sscc.ru; Olga Rodionova is with the Higher College of Informatics of the Novosibirsk State University, Russkaja str. 35, Novosibirsk, 630058, Russia, e-mail: rolcon@mail.ru

II. Random Hypernet Model Specification

General description of a hypernet model is given in [8]. Here concept of abstract hypernet is formally presented that describes multi-level network in general case, each layer is presented by a hypergraph. In partial case hypergraphs retrograde to graphs and we have simple multi-level hypernet. For the purpose of this paper the concept of two-level hypernet or simply hypernet is enough.

Definition: Hypernet $H = (X, V, R; P, F, W)$ consists of the following objects (see Fig.1):

$X = (x_1, \ldots, x_n)$ – the set of vertexes;

$V = (v_1, \ldots, v_m)$ – the set of branches (edges of the graph of a primary network);

$R = (r_1, \ldots, r_g)$ – the set of edges (edges of the graph of a secondary network);

$P : V \to X \times X$ – the mapping that defines graph $PN = (X, V)$ named a primary network;

$W : R \to X \times X$ – the mapping that defines graph $SN = (X, R)$ named a secondary network;

F – the mapping that assigns to each element $r \in R$ the set $F(r) \subseteq V$ of its branches (route in graph $PN = (X, V)$).

The incidence and adjacency in PN and SN are defined similar to those for graphs, while mapping F gives these concepts for a hypernet in a whole.

In many cases several secondary networks are embedded in a single primary network, for example one cable network may be used for public phone network, cable TV and Internet; several independent working groups can use the same LAN and so on. For modeling such situations the extension of hypernet named S-hypernet is used (see Fig. 2)

ISBN: 978-1-61804-256-9

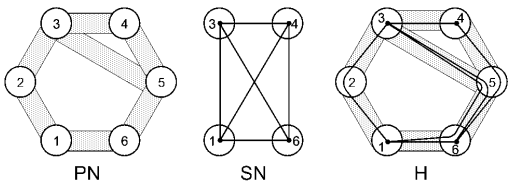


Figure 1. Illustration to the hypernet definition: PN is the primary network; the shadowed vertexes $\{1, 3, 4, 6\}$ form the set of vertexes that belongs to SN; SN is the secondary network (that is a complete graph in our case); H is the hypernet (SN is mapped to the PN)

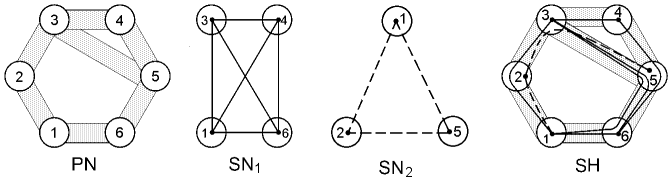


Figure 2. Illustration to the S-hypernet definition: PN is the primary network; vertexes $\{1, 3, 4, 6\}$ form the set of vertexes that belongs to SN_1 and vertexes {1,2,5} form the set of vertexes that belongs to SN_2, both are complete graphs); SH is the S-hypernet (SNs are mapped onto the PN)

Random hypernet is random realization of feasible sextuple H. We consider the following cases:

1) Fixed PN with given probabilities of existence of v_i, that corresponds to unreliable physical network with reliable nodes and unreliable links; SN and F are fixed. This model suits for the case of long-term interconnections.
2) PN and SN are fixed, F is random (randomly chosen feasible mapping). This model better suits case of short-term interconnections.

First model allows analyze probabilistic connectivity of a logical network at possible failures of channels of a physical one, which may occur as because of natural reasons, as anthropogenic ones.

III. Solving Reliability Problems with Hypernets

In [19] we discuss some improvements of well-known factoring method [20] for calculating reliability of a random graph. These improvements could be adopted for calculating reliability of SN. Factoring is executed by states of unreliable elements of PN (branches in our case) while connectivity of SN is checked. Thus we have:

$$R(SN) = p_{ij}R(SN|v_{ij} \text{ works}) + (1 - p_{ij})R(SN|v_{ij} \text{ fails}),$$

where p_{ij} is a reliability of a branch v_ij. When calculating first summand, nodes x_i and x_j are contracted into one new node, while when calculating second summand the branch v_{ij} is simply removed. On the other hand, in some cases the following direct equation may be more convenient:

$$R(SN) = \sum_{i=0}^{g} \sum_{j=1}^{C_g^i} A_{ij} I(SN_{ij}).$$

Here A_{ij} is a probability of an event corresponding to realization of j-th mode of removal exactly i branches from PN, SN_{ij} - corresponding to this event remaining part of SN, and $I(SN)$ - indicator function, 1 if SN is connected and 0 otherwise. For highly reliable branches this equation may be used for lower approximation of SNs reliability by stopping summation at some $g\_ < g$. If reliabilities of all branches are equal to some p, we can obtain a reliability polynomial for SN:

$$R(SN, p) = \sum_{i=0}^{g} B_i p^{g-i}(1 - p)^i.$$

where B_i number of connected realizations of SN when exactly i branches are removed from PN. Examination of this polynomial may help in choice of best mapping of SN onto unreliable PN.

Two possible mappings of a tree-like SN onto cyclic PN are presented in the Fig.3. Mappings differs in placement of SNs nodes into nodes of PN, shortest paths realize edges. Comparison of polynomials shows that second mapping is better for all p (see Fig.4).

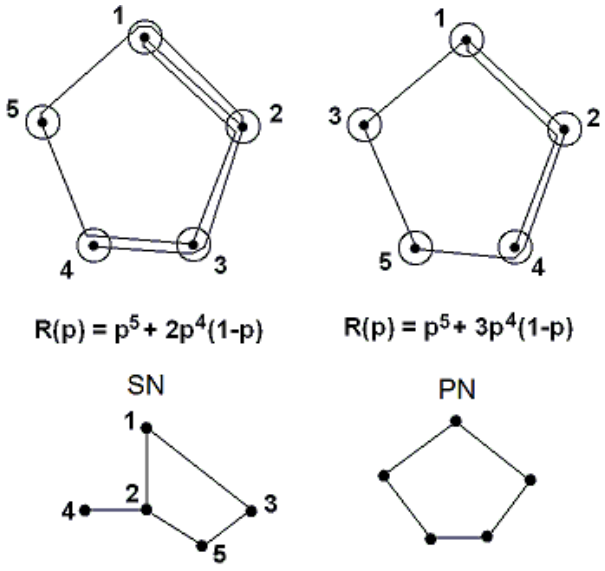


Figure 3. Two possible mappings of SN onto PN

Second model (random mapping) is harder to analyze as it assumes averaging of reliability indices by all possible mappings and so requires exhaustive search for exact calculation or Monte-Carlo method for approximate one. In the last case, using Hypernet model allows unification of data presentation and storage. When modeling one must take into account that if there are some restrictions on PN or SN, then mapping is not always possible. For example, throughput limitations on branches of PN may not allow realize all edges of SN, or there may be limitation on a number of branches realizing an edge and so on.

IV. Discussions

In this paper, we discuss possible usage of the Hypernet model for analyzing multi-layer network reliability. Possible advantages of using this model are greater. First to all it gives general means for describing multi-layer network structure,

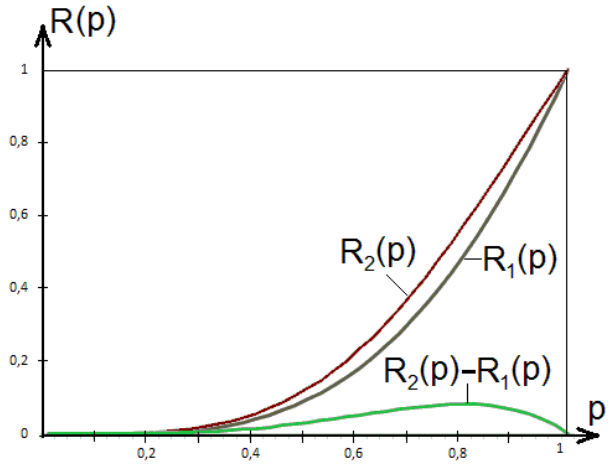


Figure 4. Reliability polynomials for SN in case of different mappings

then it allows designing algorithms that are more effective and solving problems that are impossible or hard model by other means. Our further researches in this direction concerns constructing models of multi-service sensor networks and networks in which several networks of upper level may exist inside one network of lower one.

REFERENCES

[1] R. Milner, *Bigraphs, a Tutorial,* at www.cl.cam.ac.uk/users/rm135, (2005)

[2] J.H. Kim and V. Vu, *Sandwiching random graphs,* Advances in Mathematics, 188 (2004) 444-469

[3] F. Dijkstra, B. Andree, K. Koymans, J. van der Hama, P. Grosso, C. de Laat, *A multi-layer network model based on ITU-T G.805,* Computer Networks, 52 (2008) 19271937

[4] F. He, C. Xin, *Cross-Layer Path Computation for Dynamic Traffic Grooming in Mesh WDM Optical Networks,* Norfolk State University, Technical Report NSUCS-2004-009

[5] S. Orlowski, M.C.A. Koster Arie, C. Raack, R. Wessally, *Two-layer Network Design by Branch-and-Cut featuring MIP-based Heuristics,* Proceedings of INOC 2007, International Network Optimization Conference.

[6] C. Chigan, G. Atkinson, R. Nagarajan, *On the Modeling Issue of Joint Cross-Layer Network Protection/Restoration,* Proceedings of Advanced Simulation Technologies Conference 2004 (ASTC'04), (2004) 57-62

[7] M. Kurant, P. Thiran, *Layered Complex Networks,* Phys. Rev. Lett. 96, (2006) 138701-1 – 138701-4,

[8] V.K. Popkov, *Mathematical Models of Connectivity,* Inst. Of Comp. Math. and Math. Geoph., Novosibirsk, 2006 (in Russian)

[9] V.K. Popkov, O.D. Sokolova, *Application of Hyperneet Theory for the Networks Optimazation Problems,* 17th IMACS World Congress, July 2005, Paper T4-I-42-011.

[10] A.S. Rodionov, O. Sokolova, A. Yurgenson, H. Choo, *On Optimal Placement of the Monitoring Devices on Channels of Communication Network,* ICCSA 2009, Part II, LNCS, Vol. 5593 (2009) 474487

[11] A.S. Rodionov, H. Choo, K.A. Nechunaeva, *Framework for Biologically Inspired Graph Optimization,* Proceedings of ICUIMC 2011, Seoul, Republic of Korea, (2011) paper 2.5, 4 pages.

[12] B.M. Waxman, *Routing of Multipoint Connections,* IEEE JSAC, 9 (1993) 1617162

[13] M. Doar, *Multicast in the ATM environment,* PhD thesis, Cambridge Univ., Computer Lab., (1993)

[14] C.-K. Toh, *Performance Evaluation of Crossover Switch Discovery Algorithms for Wireless ATM LANs,* Proc. INFOCOM'96 (1993) 1380-1387

[15] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal, *Stochastic models for the Web graph,* Proc. 41st Annual Symposium on Foundations of Computer Science (2000) 5765

[16] R. Albert and A.L. Barabasi, *Statistical mechanics of complex networks,* Review Modern Physics, 74 (2002) 4797.

[17] A. Yano and T. Wadayama, *Probabilistic Analysis of the Network Reliability Problem on a Random Graph Ensemble,* arXiv:1105.5903v3, 2011: http://arxiv.org/pdf/1105.5903.pdf

[18] A. Bobbio, R. Terruggia, E. Ciancamerla, and M. Minichino, *Evaluating network reliability versus topology by means of bdd algorithms,* In In: PSAM-9, Hong Kong (2008)

[19] O.K. Rodionova, A.S. Rodionov, H. Choo, Network Probabilistic Connectivity: Exact Calculation with Use of Chains. ICCSA-2004, Springer LNCS, 3046 (2004) 315324

[20] A. Satyanarayana and M.K. Chang, *Network reliability and the factoring theorem,* Networks 13 (1983) 107120.

ISBN: 978-1-61804-256-9

Mathematical Model of Influenza Dynamics Compare the incubation period and Control: in THAILAND

R. Kongnuy*., and E. Naowanich

***Abstract*—**In this paper, we propose the mathematical model to study the endemic of Influenza in Thailand. The aims of the research are to study and analyze the epidemiology of Influenza in Thailand by using the mathematical modeling. The data from the annual number of cases reported to the Division of Epidemiology, Ministry of Public Health during the period 1997-2013 are analyzed. We construct the system of nonlinear differential equations for two models. The first one, we divide the human population into four groups: the susceptible human, infectious human, the recovered human who are totally immune to the strain and the recovered human who are partially immune to that strain classes. For the second model, we enlarge the model by considering the incubation period. The standard dynamical modeling method are applied to determine the behaviors of solutions to each model. The conditions required of the parameters for the disease free and endemic equilibrium states to be local asymptotically stable is obtained. Numerical simulations are seen to support the theoretical predictions. The alternative way to control the outbreak of this disease in Thailand are suggested in our research.

***Keywords*—**Influenza, Mathematical Modeling, Nonlinear Differential Equations, Numerical Simulation.

I. Introduction

INFLUENZA or the flu is a common respiratory disease caused by influenza virus. Influenza is spread from person to person when droplets of moisture from a person with influenza are spread through the air when that person coughs, sneezes, talks and hands touching eyes, mouth or nose [1]. It caused by influenza virus which are of three types A, B and C [2]. Types A and B being clinically important. Different strains dominate from year to year. The symptoms usually start with sudden onset of chills, shakes, headache, muscle aches, fever and dry cough. The respiratory symptoms then become more prominent. People of all ages are susceptible to the flu. Symptoms appear typically 1 to 3 days after exposure to respiratory droplets from an infected person. Usually the diagnosis is based on the appearance of specific signs and symptoms of influenza. Confirmation can be achieved through laboratory testing of throat specimens or blood samples.

Thailand is situated in Southeast Asia, which is bordered to the north by Burma and Laos, to the east by Laos and Cambodia, to the south by the Gulf of Thailand and Malaysia and to the west by the Andaman Sea and the southern extremity of Burma. Thailand is divided into four geographic regions, North, Central, South and Northeast. The country of 513,000 square kilometers. In Thailand, the influenza patients have been reported a total of 809,989 cases between 1997 and 2013. The peak of influenza endemic top in 2009 and the peak in Trang province between 1997 to 2013.

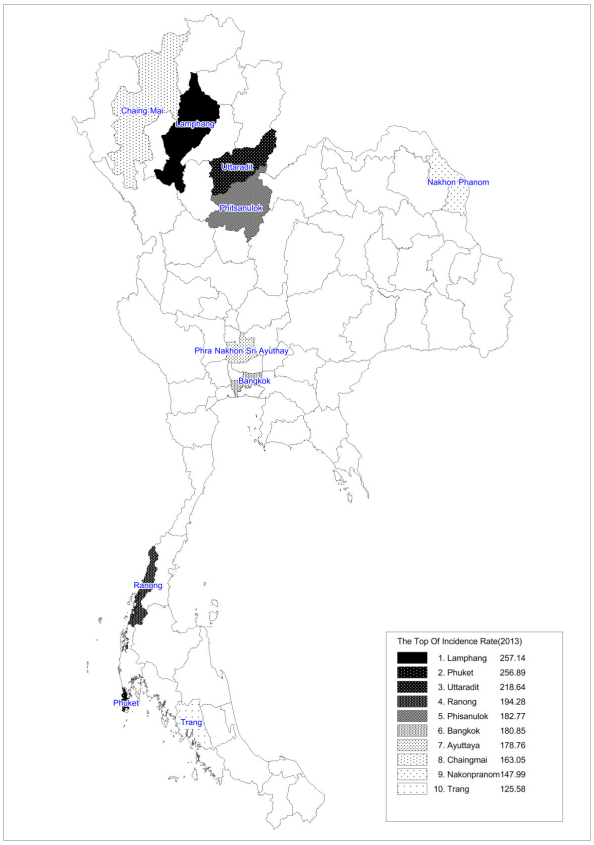


Fig. 1 Geographical distribution of the top ten of provinces by the incidence rate of influenza in 2013, Thailand [18]

This work was supported in part by Rajamangala University of Technology Suvarnabhumi.

R. Kongnuy* is with the Department of Mathematics, Faculty of science and Technology, Rajamangala University of Technology Suvarnabhumi, Nonthaburi Center, Nonthaburi, 11000 THAILAND (corresponding author to provide phone: 668-4682-1922; fax: 660-2525-2682; e-mail: rujirakung@yahoo.co.th).

E. Naowanich., was with Department of Computer Sciences, Faculty of science and Technology, Rajamangala University of Technology Suvarnabhumi, Nonthaburi Center, Nonthaburi, 11000 THAILAND (e-mail: ekachai_n@hotmail.com).

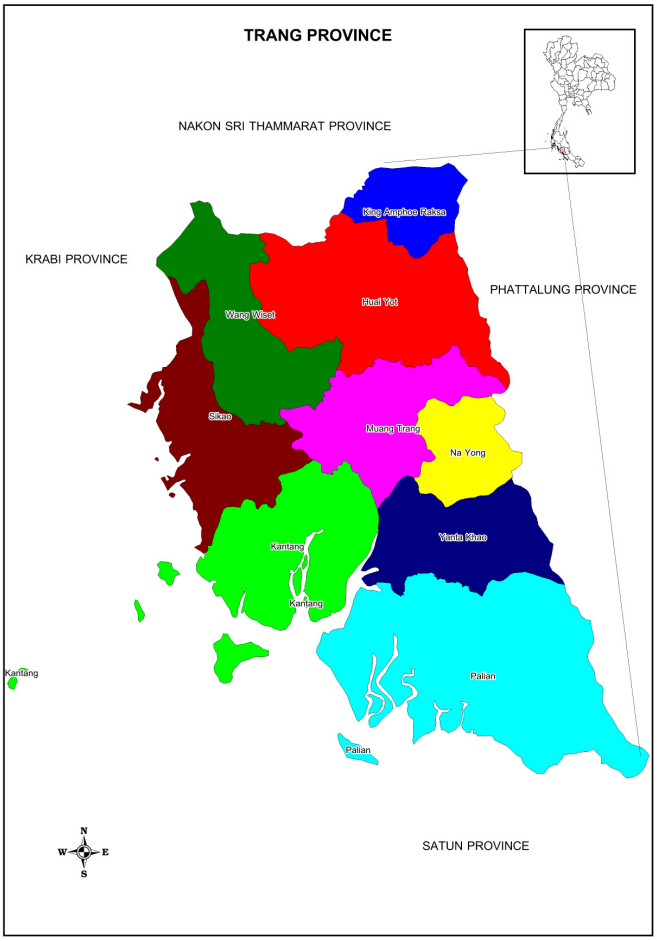


Fig.2 Geographical Trang Province which has the peak of influenza cases by average mean between 1997 and 2013 [3-18]

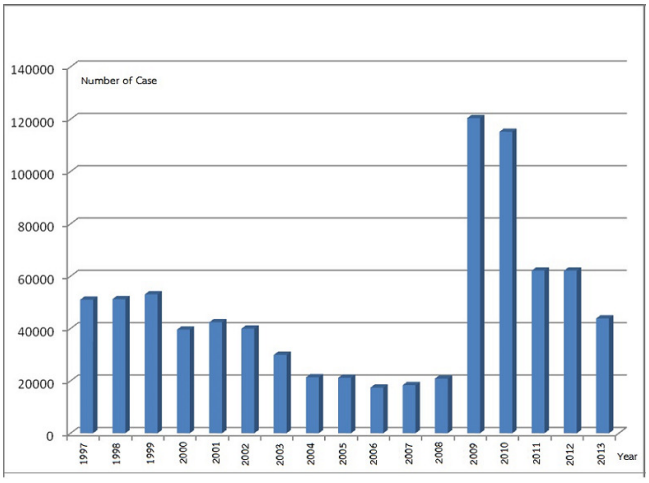


Fig. 3 The number of Influenza cases between 1997 and 2013, in Thailand [3-18]

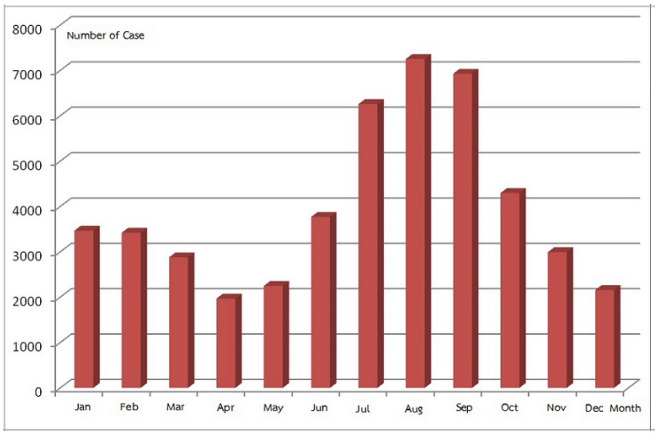


Fig. 4 Time distribution of Influenza outbreaks in Thailand (average mean in period 1997 – 2013) [3-18]

Influenza is a seasonal disease in temperate regions. Most cases in Thailand occur during the rainy months between July and September. Fig. 4 shows the time distribution of the influenza outbreak in humans in Thailand [3-18]. The disease activity starts at the beginning of the rainy season (July), peak in August. Most of the influenza outbreaks is in Trang province (average mean from 1997-2013). Trang province is in the south region of Thailand.

Several researchers such as Andreasenv et al. [19], in 1997, they develop a model that describes the dynamics of a finite number of strains that confer partial cross-protection among strains. The immunity structure of the host population is captured by an index-set notation where the index specifies the set of strains to which the host has been exposed. In 1999, Lin et al. [20], analyze an epidemiological model consisting of a linear chain of three co circulating influenza A strains that provide hosts exposed to a given strain with partial immune cross-protection against other strains. In the extreme case where infection with the middle strain prevents further infections from the other two strains. In 2002, Earn et al. [21], they developed mathematical and computational models that elucidate many properties of multi strain systems. The theoretical insights are also required to simplify model structures and facilitate predictions that can be tested with accessible data. In 2004, Alexender et al. [22], they construct a deterministic mathematical model to study the transmission dynamics of influenza. The model is analyzed qualitatively to determine criteria for control of an influenza epidemic and is used to compute the threshold vaccination rate necessary for community-wide control of influenza.

In this paper, we compare the behavior of the transmission of influenza virus by formulating the mathematical models. There is no incubation period condition in the first model. The second model, we take the incubation period into the model.

II. Mathematical Models and Analysis

A. The First Model

To compare the endemic of Influenza in Thailand for two models, the initial model, we construct a system of nonlinear

ISBN: 978-1-61804-256-9

differential equations and divide the human population (no effect of incubation period) into 4 groups: the susceptible human, infectious human, the recovered human who are totally immune to that strain and the recovered human who are partially immune to that strain classes.

Let

S be the number of susceptible human,

I be the number of infectious human,

R be the number of recovered human who are totally immune to that strain ,

C be the number of recovered human who are partially immune to that strain classes.

The dynamics of the model for the pandemic influenza with no effect of incubation period can be described by the following equations

$$\frac{dS}{dt} = B_n N - (\mu + \beta I)S, \quad (1)$$

$$\frac{dI}{dt} = \beta SI - (\mu + \alpha)I, \quad (2)$$

$$\frac{dR}{dt} = \alpha I - (\mu + \delta)R, \quad (3)$$

$$\frac{dC}{dt} = \delta R - \mu C \quad (4)$$

with the conditions $N = S + I + R + C$

where

N is the total of human population,

B_n is the birth rate of human population,

μ is the natural death rate of human population,

β_1 is the transmission rate which the susceptible human population become to infectious human,

α is the transmission rate which the infectious human population become to the recovered human who are totally immune to that strain,

δ is the transmission rate which the recovered human who are totally immune to that strain become to the recovered human who are partially immune to that strain classes.

The total number of human population is assumed that constant. So the rates of change for the total human population is equals to zero. Then we obtain $B_n = \mu$ and we normalize (1) to (4) by letting $S^* = S/N$, $I^* = I/N$, $R^* = R/N$ and $C^* = C/N$ then our equations become

$$\frac{dS^*}{dt} = \mu - (\mu + \beta_1 I^* N)S^*, \quad (5)$$

$$\frac{dI^*}{dt} = \beta_1 S^* N I^* - (\mu + \alpha)I^*, \quad (6)$$

$$\frac{dC^*}{dt} = \delta R^* - \mu C^* \quad (7)$$

with the condition $R^* = 1 - S^* - I^* - C^*$

B. Disease free equilibrium and endemic equilibrium states of the first model

The model (5) to (7) has exactly one disease free equilibrium state $E^1 = (1,0,0)$ in the region ε when

$$\varepsilon = \{(S^*, I^*, C^*) \,|\, S^*, I^*, C^* \geq 0,\ S^* + I^* + R^* + C^* = 1\}.$$

We use the next generation matrix approach as described in [23-24] to define the basic reproductive number R_{b1}, R_{b1} as the number of secondary infections that one infectious individual would create over the duration of the infectious period in the presence of vaccination, provided that everyone else is susceptible. It occur when $R_{b1} \leq 1$. From our model, we have the endemic equilibrium state E^2 when $E^2 = (\hat{S}, \hat{I}, \hat{C})$ which

$$\hat{S} = 1/(1 + h\hat{I}), \quad (8)$$

$$\hat{I} = (h - A)/Ah, \quad (9)$$

$$\hat{C} = ((\delta\hat{I})(h(1 - \hat{I}) - 1))/((\delta + \mu)(1 + h\hat{I})) \quad (10)$$

where $h = \beta_1 N/\mu$, $A = (\mu + \alpha)/\mu$ and it occur when $R_{b1} > 1$.

C. Stability of the disease free and endemic equilibrium states of the first model

For the local stability analysis of disease free equilibrium state, the linearized systems of (5) to (7) around E^1. The Jacobian of linearized is

$$J_{E^1} = \begin{bmatrix} -\mu & -\mu h & 0 \\ 0 & \mu h - \mu A & 0 \\ -\delta & -\delta & -\delta - \mu \end{bmatrix}_{(1,0,0)}. \quad (11)$$

The characteristic equation of J_{E^1} is as follows:

$$(\mu + \lambda)(\mu A + \lambda - \mu h)(\delta + \mu + \lambda) = 0. \quad (12)$$

Then we have the negative three roots, $\lambda_1 = -\mu$, $\lambda_2 = -\mu - \delta$ and $\lambda_3 = \mu(h - A)$ when $R_{b1} = \sqrt{h/A} < 1$. Hence by Hurwitz's criteria we have established the following result.

For the endemic equilibrium state, we analyze the local stability by linearizing systems (5) to (7) around E^2, then we have the Jacobian

$$J_{E^2} = \begin{bmatrix} -(\mu + \mu h\hat{I}) & -\mu h\hat{S} & 0 \\ \mu h\hat{I} & \mu h\hat{S} - \mu A & 0 \\ -\delta & -\delta & -\delta - \mu \end{bmatrix}_{(\hat{S},\hat{I},\hat{C})}. \quad (13)$$

The characteristic equation for the endemic state is given by

$$(\lambda+\delta+\mu)(\lambda^2+a_1\lambda+a_2)=0 \quad (14)$$

when

$a_1=(1+A+h\hat{I}-h\hat{S})\mu$ and $a_2=(A+Ah\hat{I}-h\hat{S})\mu^2$.

The eigenvalues of (14) are $\lambda_1=-\delta-\mu$,

$\lambda_2=\frac{-a_1-\sqrt{a_1^2-4a_2}}{2}$ and $\lambda_3=\frac{-a_1+\sqrt{a_1^2-4a_2}}{2}$.

$\lambda_3<0$ when $h>A$. So that the stability of the endemic equilibrium point by using Routh-Hurwitz criteria, we found that the endemic equilibrium state is locally stable when $R_{b1}=\sqrt{h/A}>1$.

D. The Second Model

For the second mathematical model, we consider the time of incubation period when the susceptible human become to exposed human population and we consider the re-infection in recovered human population. The dynamic of human population can be described by the following equations

$$\frac{dS}{dt}=B_nN+\gamma C-(\mu+\beta_2 I)S, \quad (15)$$

$$\frac{dE}{dt}=\beta_2 SI-(\mu+\phi)E, \quad (16)$$

$$\frac{dI}{dt}=\phi E-(\mu+\alpha)I, \quad (16)$$

$$\frac{dR}{dt}=\alpha I-(\mu+\delta)R, \quad (17)$$

$$\frac{dC}{dt}=\delta R-(\mu+\gamma)C \quad (18)$$

where

S be the number of susceptible human,

E be the number of infectious human which in the incubation period (exposed human population),

I be the number of infectious human,

R be the number of recovered human who are totally immune to that strain ,

C be the number of recovered human who are partially immune to that strain classes.

N is the total of human population,

B_n is the birth rate of human population,

μ is the natural death rate of human population,

β_2 is the transmission rate which the susceptible human population become to exposed human,

ϕ is the transmission rate which the exposed human population become to infectious human,

α is the transmission rate which the infectious human population become to the recovered human who are totally immune to that strain,

δ is the transmission rate which the recovered human who are totally immune to that strain become to the recovered human who are partially immune to that strain classes.

γ is the transmission rate which the recovered human population become to re-infection again (in susceptible human population).

Introducing the proportion $S^*=S/N$, $E^*=E/N$, $I^*=I/N$, $R^*=R/N$ and $C^*=C/N$ and with the conditions $R^*=1-S^*-E^*-I^*-C^*$. The previous (15)-18) become

$$\frac{dS^*}{dt}=\mu+\gamma C^*-(\mu+\beta_2 I^*N)S^*, \quad (19)$$

$$\frac{dE^*}{dt}=\beta_2 NS^*I^*-(\mu+\phi)E^*, \quad (20)$$

$$\frac{dI^*}{dt}=\phi E^*-(\mu+\alpha)I^*, \quad (21)$$

$$\frac{dC^*}{dt}=\delta(1-S^*-E^*-I^*-C^*)-(\mu+\gamma)C^* . \quad (22)$$

E. Disease free equilibrium and endemic equilibrium states of the second model

The model (19) to (22) has two equilibrium states. This gives the disease free equilibrium state $E^1=(1,0,0,0)$ and the endemic disease equilibrium state $E^2=(\hat{S},\hat{E},\hat{I},\hat{C})$ where

$$\hat{S}=\frac{1+j\hat{C}}{1+h\hat{I}}, \quad (23)$$

$$\hat{E}=\frac{h\hat{I}(1+j\hat{C})}{(1+k)(1+h\hat{I})}, \quad (24)$$

$$\hat{I}=\frac{\phi(1+j\hat{C})}{(\mu+\alpha)(1+k)}-\frac{1}{h}, \quad (25)$$

$$\hat{C}=\frac{\delta}{(\delta+\mu+\gamma)}(1-\hat{S}-\hat{E}-\hat{I}) \quad (26)$$

which $h=\frac{\beta_2 N}{\mu}$, $j=\frac{\gamma}{\mu}$ and $k=\frac{\phi}{\mu}$.

F. Stability of the disease free and endemic equilibrium states of the second model

The local stability of disease free equilibrium solutions can be examined by linearizing (19) to (22) around E^1. This gives the Jacobian matrix as follow

$$J_{E^1}=\begin{bmatrix}-\mu & 0 & -\mu h & \mu j\\ 0 & -(\mu+k\mu) & \mu h & 0\\ 0 & k\mu & -(\mu+\alpha) & 0\\ -\delta & -\delta & -\delta & -\delta-\mu-\mu j\end{bmatrix}_{(1,0,0,0)} \quad (27)$$

The characteristic equation for the disease free is given by

$$(\lambda+\mu+\delta)(\lambda+\mu+j\mu)(\lambda^2+b_1\lambda+b_2)=0 \qquad (28)$$

when $b_1=\alpha+(2+k)\mu$ and

$b_2=\alpha\mu(1+k)+\mu^2(1+k-hk)$.

The eigenvalues of J_{E^1} are $\lambda_1=-\mu-\delta$, $\lambda_2=-\mu-j\mu$,

$\lambda_3=\dfrac{-b_1-\sqrt{b_1^2-4b_2}}{2}$ and $\lambda_4=\dfrac{-b_1+\sqrt{b_1^2-4b_2}}{2}$.

$\lambda_4<0$ when $R_{b2}=\sqrt{(\mu hk)/((\alpha+\mu)(1+k))}<1$. So that the stability of the disease free equilibrium state is locally stable by using Routh-Hurwitz criteria .

For the endemic equilibrium state of second model, we analyze the local stability by linearizing systems (19) to (22) around E^2, then we have the Jacobian

$$J_{E^2}=\begin{bmatrix} -(\mu+\mu h\hat{I}) & 0 & -\mu h\hat{S} & \mu j \\ \mu h\hat{I} & -(\mu+k\mu) & \mu h\hat{S} & 0 \\ 0 & k\mu & -(\mu+\alpha) & 0 \\ -\delta & -\delta & -\delta & -\delta-\mu-\mu j \end{bmatrix}_{(\hat{S},\hat{E},\hat{I},\hat{C})} \qquad (29)$$

The characteristic equation for the endemic state is given by

$$\lambda^4+a_1\lambda^3+a_2\lambda^2+a_3\lambda+a_4=0 \qquad (30)$$

when

$$a_4=\mu^2((1+j+k+jk+h\hat{I}(1+j+k))\alpha\delta+(1+h\hat{I})(1+j)(1+k)\alpha\mu$$
$$+(1+j)(1+k+h(\hat{I}+k\hat{I}-k\hat{S}))\mu(\delta+\mu)),$$
$$a_3=\mu((2+h\hat{I}+j+k))\alpha\delta+(3+2j+92+j)k$$
$$+h\hat{I}(2+j+k))\alpha\mu+\mu((3+2j+(2+j)k$$
$$+h\hat{I}(2+j+k)-hk\hat{S})\delta+(4+3j+3k+2jk$$
$$+h\hat{I}(3+2j+(2+j)k)-h(2+j)k\hat{S})\mu)),$$
$$a_2=\alpha\delta+(3+h\hat{I}+j+k)(\alpha+\delta)\mu+(jk+3(2+j+k)$$
$$+h\hat{I}(3+j+k)-hk\hat{S})\mu^2,$$
$$a_1=\alpha+\delta+(4+h\hat{I}+j+k)\mu.$$

The stability of the endemic equilibrium state can be determined without solving the actual values of eigenvalues by using the Routh-Hurwitz criteria. So the four conditions of Routh-Hurwitz criteria for local asymptotical stability in 4^{th} order characteristic polynomial equation are

i) $a_1>0$, (31)

ii) $a_3>0$, (32)

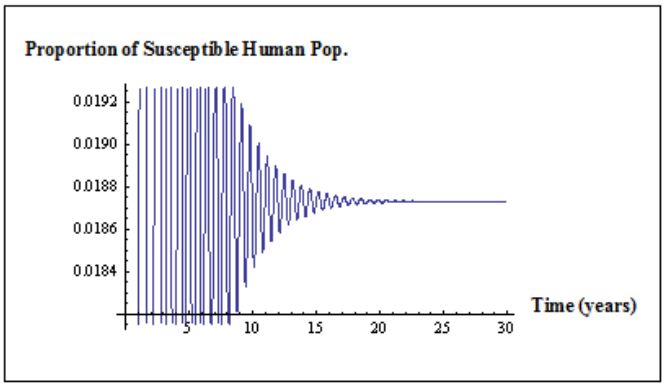
iii) $a_4>0$, (33)

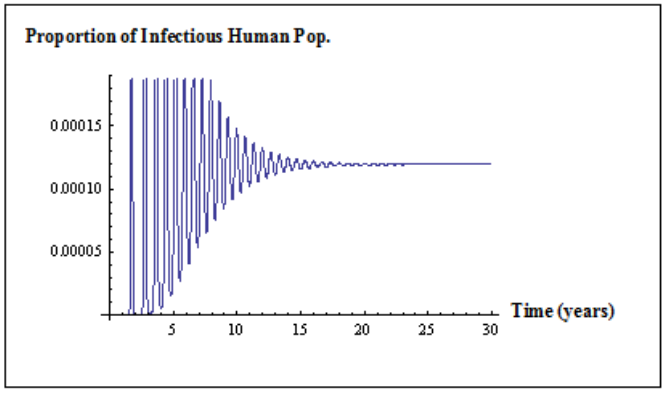
iv) $a_1a_2a_3>a_3^2+a_1^2a_4$. (34)

After we check the stability of the endemic equilibrium state, we found that the endemic equilibrium state is locally stable when $R_{b2}=\sqrt{(\mu hk)/((\alpha+\mu)(1+k))}>1$.

III. Numerical Result

In this study, we are interested in the endemic of influenza virus in Thailand. Numerical solutions are presented comparing used the real data from 1997 to 2013. The values of certain parameters of the model such as the birth rate, the death rate and the total of human population, were obtained from the demographic data of Thailand. They are $\mu=1/70$ per year, which corresponding to a life expectancy of 70 years; $N=62,644,913$, $\phi=0.5$ and $\delta=0.75$ [25]-[26].





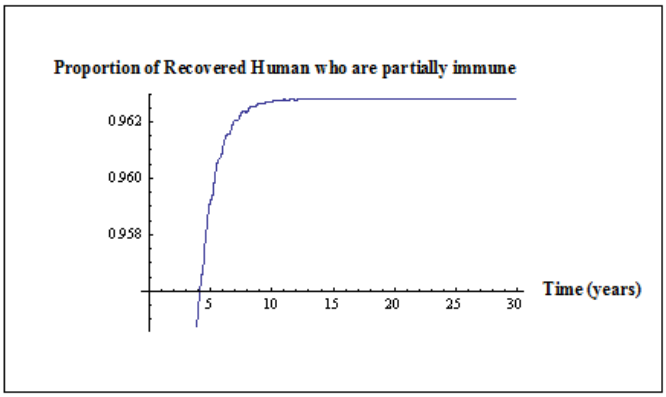


Fig. 5 Numerical solution of (5) to (7) yield the time series solutions of the proportion susceptible , infectious and recovered human populations. Values of parameters are $\beta_1=0.0001$, $\alpha=117.32$ and $\delta=0.75$. $(\hat{S},\hat{I},\hat{C})=(0.018730, 0.000119, 0.962811)$.

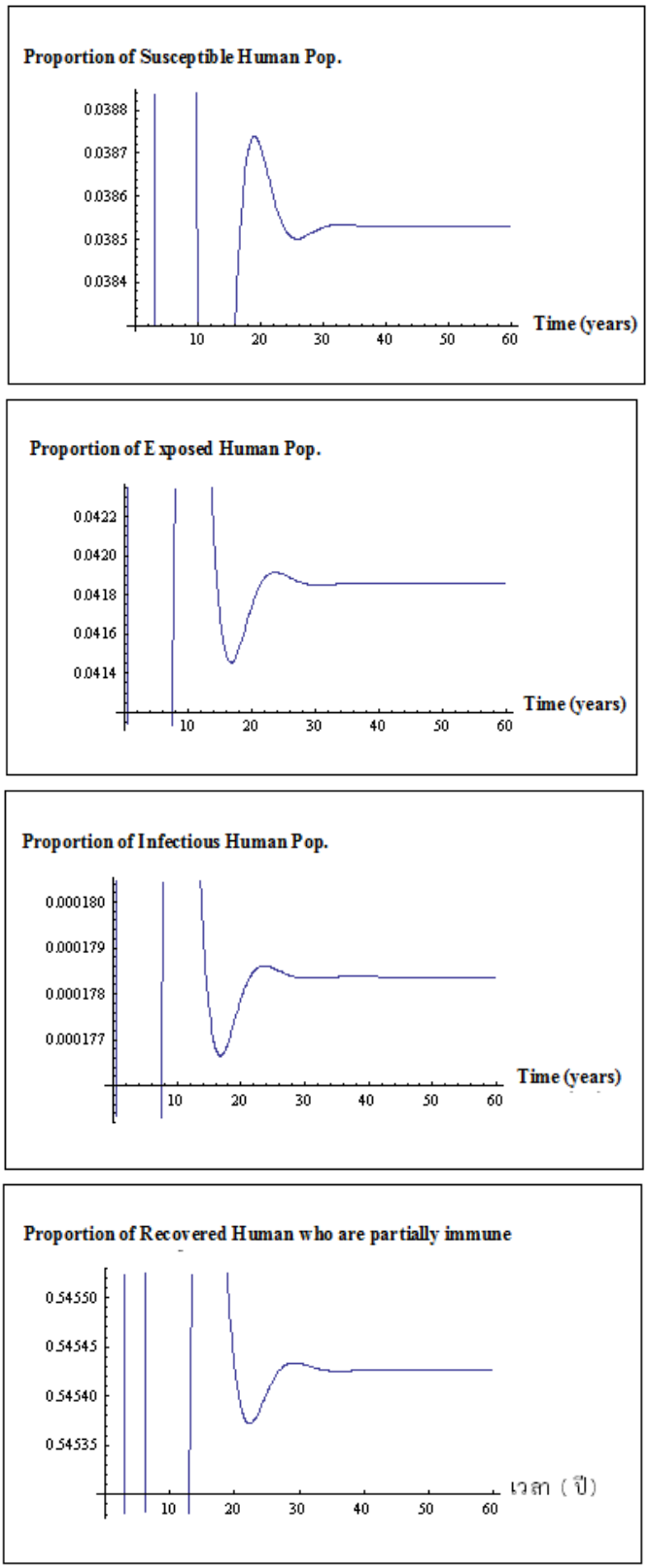


Fig. 6 Numerical solution of (19) to (22) yield the time series solutions of the proportion susceptible, exposed, infectious and recovered human populations. Values of parameters are $\mu = 1/70$, $\beta_2 = 0.00005$, $\gamma = 0.5$, $\phi = 0.5$ $\alpha = 117.32$ and $\delta = 0.75$. $(\hat{S}, \hat{E}, \hat{I}, \hat{C}) = (0.0385304, 0.041905, 0.000178, 0.545425)$

IV. Discussion and Conclusion

In our paper, we analyze the endemic influenza in Thailand by using the real data between 1997 and 2013. The numerical simulations for two mathematical models have different behaviors. Fig. 5 shows time series solution when there is no the effect of incubation period and re-infection in human population. Fig. 6 shows time series solutions when there is the effect of incubation period and re-infection. The values of parameters are satisfied Routh-Hurwitz criterions for the endemic equilibrium states.

Moreover, we consider the basic reproductive number in each model. For the first model which no effect of incubation period

$$R_{b1} = \sqrt{\frac{\beta_1 N}{(\mu + \alpha)}} \tag{35}$$

which shows that the number of secondary case from infectious human with influenza no effect of incubation period depend on the transmission rate which the susceptible human population become to infectious human, the total of human population and the transmission rate which the infectious human population become to the recovered human who are totally immune to that strain. Because of the natural death rate of human population is constant (almost unchanged) . For the second model, we consider the time of incubation period when the susceptible human become to exposed human population and we consider the re-infection in recovered human population. Then we have the basic reproductive number

$$R_{b2} = \sqrt{(\mu hk)/((\alpha + \mu)(1 + k))} \tag{36}$$
$$= \sqrt{\frac{\beta_2 \phi N}{(\mu + \alpha)(\mu + \phi)}}$$

which its value depend on the transmission rate which the susceptible human population become to exposed human, the transmission rate which the exposed human population become to infectious human, the total human population and the transmission rate which the infectious human population become to the recovered human who are totally immune to that strain. In the second model has the proportions of the susceptible human and infectious human are higher than the first model which no effect the incubation period.

The most important thing in mathematical models in this study is to understand how influenza spreads in the real world and how various complexities affect the dynamics for searching the method to prevention and control the endemic of influenza in Thailand.

Acknowledgment

This work is supported by Rajamangala University of Technology Suvarnabhumi, Thailand. We would like to thank Assoc. Prof. Dr. Puntani Pongsumpun at King Mongkut's Institute of Technology Ladkrabang and Prof. Dr. I-Ming Tang at Mahidol University, Thailand.

REFERENCES

[1] British Columbia Ministry of Health, *What is pandemic influenza?*, B C Health Files pandemic influenza series-94a, 2006.

[2] K. Park, *Preventive and Social Medicine*, M/S Banarsi Das Bhanot Publishers, Jabalpur, India, 2005.

[3] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 1997.

[4] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 1998.

[5] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 1999.

[6] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 2000.

[7] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 2001.

[8] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 2002.

[9] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 2004.

[10] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 2005.

[11] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 2006.

[12] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 2007.

[13] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 2008.

[14] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 2009.

[15] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 2010.

[16] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 2011.

[17] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 2012.

[18] Division of Epidemiology, *Annual Epidemiological Surveillance Report*, Ministry of Public Health, Thailand, 2013.

[19] V. Andreasen, J. Lin and S. A. Levin, "The dynamics of co circulating influenza strains conferring partial cross-immunity", *J. Math. Biol.,*vol. 35, pp. 825-842, 1997.

[20] J. Lin, V. Andreason and S. A. Levin, "Dynamics of influenza a drift: the linear three strain model," *Math. Biosci*., vol. 162, pp.35-51, 1999.

[21] D. J. D. Earn, J. Dushoff and S.A. Levin, "Ecology and evolution of the flu," *Trends Ecol. Evol*., vol. 17, pp.334-340, 2002.

[22] M. E. Alexander, C. Bowman, S. M. Moghadas, R. Summers, A. B. Gumel and B. M. Sahai, "A vaccination model for transmission dynamics of influenza," *SIAM J. Apply. Dyn. Syst*., vol 3, pp. 503-524, 2004.

[23] O. Diekmann, J. A.P. Hwwsterbeek and J.A.J Metz, "on the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations," *J. Math. Biol*., vol. 28, pp.365-382, 1990

[24] J. M. Hefferman, R. J. Smith and L. M. Wahl, "Perspectives on the basic reproductive ratio," *J. R. Soc. Interface*., vol. 2, pp.281-293,2005.

[25] J.B. Plotkin, J. Dushoff, S.A. Levin, "Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus," *Proc. Nat. Acad*. Sci. USA vol.96, pp. 6263, 2002.

[26] A. J. Hay, V. Gregory, A. R. Douglass, Y. P . Lin, "The evolution of human influenza viruses," *Proc. Roy. Soc. Lond*., vol. B356, pp.1861,2001.

R. Kongnuy , I was born in Thailand on June, 20, 1976. My educational background: I was earned Ph.D. (Applied Mathematics) 2009 King Monkut's Institute of Technology Ladkrabang, Bangkok, Thailand, M.Sc. (Mathematics) 2005 Ramkhamhaeng University, Bangkok, Thailand, B.Sc. (Mathematics) First Class Honor 1998 Prince of Songkla University, Songkla, Thailand.

I'm a LECTURE at Department of Mathematics, Faculty of Science and Technology, Rajamangala University of Technology Suvarnabhumi, Nonthaburi Campus since 2000. My publications: 'Model for the transmission of dengue disease in pregnant and non-pregnant patients', *International Journal of Mathematical Models and Methods in Applied Sciences*, vol. 1, pp. 127-132, 2007. 'Analysis of a Mathematical Model for Dengue Disease in Pregnant Cases', *International Journal of Biological and Medical Sciences*, vol. 3, pp. 192-199, 2008. 'Mathematical Model for Dengue Disease with Maternal Antibodies', *International Journal of Biological and Medical Sciences*, vol. 1, pp. 5-14, 2010. "Analysis of a dengue transmission model with clinical diagnosis in Thailand", *International Journal of Mathematical Models and Methods in Applied Science*, vol 5, pp. 594-601, 2011. "Contact infection spread in an SEIR model: An analytical approach", *ScienceAsia* vol 39, pp. 410-415, 2013.

E. Naowanich , I was born in Thailand on June, 19, 1977. My educational background: I was earned M.Sc. (Computer Sciences) 2007 King Monkut's University of Technology North Bangkok, Bangkok, Thailand, B.Eng. (Electrical Engineering) 2003 Rajamangala Institute of Technology Thanyaburi, Pathumthani, Thailand, B.I.Ed (Computer Engineering) Second Class Honor 1999 Rajamangala Institute of Technology North Bangkok, Bangkok, Thailand.

I'm a LECTURE at Department of Computer Sciences, Faculty of Science and Technology, Rajamangala University of Technology Suvarnabhumi, Nonthaburi Center since 2000. My publications: Analysis of a dengue transmission model with clinical diagnosis in Thailand", *International Journal of Mathematical Models and Methods in Applied Science*, vol 5, pp. 594-601, 2011.

Partial discretization method for stability analysis of dynamic systems

L.Khajiyeva, Askat Kudaibergenov, Askar Kudaibergenov

Abstract – This paper investigates stability of motion of elastic systems with nonlinear characteristics. Model of steady motion of elastic systems in the absence of oscillatory processes is considered. The model is based on Lyapunov stability criterion. Analysis of the perturbation equation is carried out by use of the partial discretization method. Partial discretization of the Hill equation in a class of generalized functions (Dirac's delta function) is employed to considerably simplify the Hill parametrical equation and obtain its quasi-analytical solutions. Efficiency of the offered method is shown on the example of stability of resonant oscillations of physically nonlinear systems.

Keywords – Nonlinear system, partial discretization method, resonance, stability.

I. INTRODUCTION

This paper considers the issues of stability of motion of nonlinear mechanical systems and methods of their analysis which is of practical interest.

Stability of motion is one of the main problems of modern machinery dynamics. Under the influence of large inertial forces and technological loadings in links of machines as a result of their elastic deformation, difficult oscillatory processes connected with modulation of frequencies and emergence of resonant phenomena occur. These undesirable processes significantly affect strength characteristics of separate elements, and also functionality of machines. Misalignment of links and their deviation from the set trajectories can be observed. Therefore, as well as in the previous works [1]-[2], steady motion of mechanical systems is considered as their movement in the absence of oscillatory processes.

Research of stability of motion of mechanisms and machines depends on a choice of their dynamic model. Widely used model of motion of machinery elements as absolutely rigid, considerably narrows a framework of its application. As a rule, it is research of quasi-static and to resonant modes of motion. However, this model is only the first approximation for the majority of problems.

Nonlinear dynamic models of machines taking into account deformability of links are of the greatest interest. In [3]-[6] models of machine motion were developed assuming all links to be elastic. Their geometrical and physical nonlinearity was considered. Connection between elastic displacements of links was considered through reactions in hinges of adjacent links. Nonlinearity of models can cause resonances on sub- and ultra-frequencies. Therefore, ensuring steady motion of system depends on identification and elimination of frequencies causing resonant vibrations from operating modes.

Most research on stability of periodic oscillations was performed by use of asymptotic methods and methods of small parameter. They are quasi-linear and quasi-Lyapunov systems [7]-[9], etc. By means of Lyapunov's function at rather rigid restrictions on degree of nonlinearity, conditions of asymptotic global stability were obtained.

Among works on research of parametrical instability of nonlinear mechanical systems works of S. Hayashi [10], A.Tondl [11], W. Szemplinska-Stupnicka [12], etc. are well-known. In [13] questions of stability of periodic oscillations of a nonlinear system without restrictions on the size of its nonlinearity and nonautonomous terms were studied.

The objective of this paper is performing stability analysis of nonlinear dynamic deformable systems for elimination of dangerous oscillations from operating modes.

II. DYNAMIC MODELS

One method for solving problems of dynamics of elastic systems is reducing the dimension of equations of motion by applying well-known methods of separation of variables and research of dynamic processes in nonlinear mechanical systems with one degree of freedom in the form:

$$\ddot{f} + \Phi(\dot{f}, f) + \omega_0^2 f = F(\Omega t). \qquad (1)$$

Degree of nonlinearity of the term $\Phi(\dot{f}, f)$ relative to generalized function of displacements $f(t)$ corresponds to assumptions of the model, and characterizes nonlinearity of elastic characteristics (geometric and physical nonlinearity) and dissipative forces.

Considering stability of the periodic solution $f_0(t)$ of (1), we set a small deviation δf from its equilibrium state:

$$f(t) = f_0(t) + \delta f \ . \qquad (2)$$

Stability of the periodic solution $f_0(t)$ depends on the nature of the behavior of its small deviation δf in time, i.e. solution of the equation of the perturbed state of the system:

$$\delta \ddot{f} + \left(\frac{\partial \Phi}{\partial \dot{f}}\right)_0 \delta \dot{f} + \left(\frac{\partial \Phi}{\partial f}\right)_0 \delta f = 0 , \qquad (3)$$

where the symbol $(\ )_0$ means that the solution $f_0(t)$ is taken as argument of functions.

ISBN: 978-1-61804-256-9

If the solution δf of (3) is limited at $t \to \infty$, then motion of the system is considered to be stable. If $\delta f \to \infty$ at $t \to \infty$, by definition, the motion is unstable that is identical to the criterion of Lyapunov stability.

Legitimacy of transition from (1) to (3) is given in work [10] which refers to Treffts's research concerning properties of periodic solutions of equations in the form (1). Limitation of a solution of (1) and its asymptotic stability results in its frequency with the smallest period equaled or multiplied to the period of the external perturbing force. In [10] the Floquet theory was involved to study stability of the periodic solution of (1).

Introduce a new variable η:

$$\delta f = \eta \exp\left(-0,5\left(\frac{\partial \Phi}{\partial \dot{f}} \right)_0 \right). \tag{4}$$

Then (3) reduces to the Hill parametrical equation relative to the variable η.

For the case of basic resonance $f(t) = r_0 + r_1 \cos(\Omega t - \varphi_1)$ the Hill equation is represented as:

$$\frac{d^2\eta}{dt^2} + \eta\left[\theta_0 + \theta_{1s}\sin\Omega t + \theta_{1c}\cos\Omega t + \theta_{2s}\sin 2\Omega t + \theta_{2c}\cos 2\Omega t\right] = 0, \tag{5}$$

where θ_0, θ_{1s}, θ_{1c}, θ_{2s}, θ_{2c} are functions of frequencies, amplitudes, and phases of oscillations of harmonic solutions of (1), Ω, r_1, φ_1 respectively.

Among methods of the dynamic analysis of vibrations of mechanical systems the methods based on creating the characteristic determinants specifying borders of instability zones of the resonant modes are widely known. For this, either the Floquet theory is used, as in work [10], or borders of instability zones are defined directly on amplitude-frequency characteristics, i.e. on resonant curves by means of the Routh-Hurvitz criterion.

Here, in contrast to the mentioned methods, the problem of stability of motion of system (1) based on applying the partial discretization method [14] to the solution of Hill's equation (5) is investigated. This method allows to obtain the analytical solution of the Hill equation characterizing the behavior of small perturbation δf in time *t*.

III. PARTIAL DISCRETIZATION OF THE HILL EQUATION

According to the method of partial discretization [14], the second term of (5) is represented discretely in a class of the generalized functions:

$$\begin{aligned}\frac{d^2\eta}{dt^2} + \frac{1}{2}\sum_{k=1}^{n}(t_k + t_{k+1})[&(\theta_0 + \theta_{1s}\sin\Omega t_k + \theta_{1c}\cos\Omega t_k \\ &+ \theta_{2s}\sin 2\Omega t_k + \theta_{2c}\cos 2\Omega t_k)\cdot\eta(t_k)\delta(t - t_k) \\ &-(\theta_0 + \theta_{1s}\sin\Omega t_{k+1} + \theta_{1c}\cos\Omega t_{k+1} + \theta_{2s}\sin 2\Omega t_{k+1} \\ &+ \theta_{2c}\cos 2\Omega t_{k+1})\cdot\eta(t_{k+1})\delta(t - t_{k+1})] = 0,\end{aligned} \tag{6}$$

where

$\eta(t_k)$ is discrete representation of function $\eta(t)$ for the value of the argument $t = t_k$;

$k = \overline{1,n}$ the number of splitting of the argument t;

$\delta(t - t_k)$ Dirac's delta function.

Taking the following initial conditions: $\eta(0) = \eta_0$, $\dot{\eta}(0) = \dot{\eta}_0$ at $t = 0$, the solution of (6) is expressed as:

$$\begin{aligned}\eta(t) = -\frac{1}{2}\sum_{k=1}^{n}(t_k + t_{k+1})[&(\theta_0 + \theta_{1s}\sin\Omega t_k + \theta_{1c}\cos\Omega t_k \\ &+ \theta_{2s}\sin 2\Omega t_k + \theta_{2c}\cos 2\Omega t_k)\cdot\eta(t_k)H(t - t_k) \\ &-(\theta_0 + \theta_{1s}\sin\Omega t_{k+1} + \theta_{1c}\cos\Omega t_{k+1} + \theta_{2s}\sin 2\Omega t_{k+1} \\ &+ \theta_{2c}\cos 2\Omega t_{k+1})\cdot\eta(t_{k+1})H(t - t_{k+1})] + \dot{\eta}_0 t + \eta_0,\end{aligned} \tag{7}$$

where $H(t - t_k)$ denotes Heaviside step function.

Specifying t discretely, we obtain a recurrent formula for calculation of unknown $\eta(t)$ on *k*-th step of splitting of the argument t:

$$\begin{aligned}\eta(t_k) = \Big[&-(t_1 + t_2)(\theta_0 + \theta_{1s}\sin\Omega t_1 + \theta_{1c}\cos\Omega t_1 + \theta_{2s}\sin 2\Omega t_1 \\ &+ \theta_{2c}\cos 2\Omega t_1)\eta(t_1)\left(\frac{t_k + t_{k+1}}{2} - t_1\right)\Big]\Big/\Big[1 + \frac{1}{2}(t_{k+1} - t_k) \\ &\cdot(t_{k+1} - t_{k-1})(\theta_0 + \theta_{1s}\sin\Omega t_k + \theta_{1c}\cos\Omega t_k + \theta_{2s}\sin 2\Omega t_k \\ &+\theta_{2c}\cos 2\Omega t_k)\Big] - \Big[\sum_{j=2}^{k-1}(t_{j+1} - t_{j-1})(\theta_0 + \theta_{1s}\sin\Omega t_j \\ &+ \theta_{1c}\cos\Omega t_j + \theta_{2s}\sin 2\Omega t_j + \theta_{2c}\cos 2\Omega t_j)\eta(t_j) \\ &\cdot\left(\frac{t_k + t_{k+1}}{2} - t_j\right)\Big]\Big/\Big[1 + \frac{1}{2}(t_{k+1} - t_k)(t_{k+1} - t_{k-1}) \\ &\cdot(\theta_0 + \theta_{1s}\sin\Omega t_k + \theta_{1c}\cos\Omega t_k + \theta_{2c}\sin 2\Omega t_k \\ &+ \theta_{2c}\cos 2\Omega t_k)\Big] + \Big[\dot{\eta}_0\frac{t_k + t_{k+1}}{2} + \eta_0\Big]\Big/\Big[1 + \frac{1}{2}(t_{k+1} - t_k) \\ &\cdot(t_{k+1} - t_{k-1})(\theta_0 + \theta_{1s}\sin\Omega t_k + \theta_{1c}\cos\Omega t_k \\ &+ \theta_{2s}\sin 2\Omega t_k + \theta_{2c}\cos 2\Omega t_k)\Big].\end{aligned} \tag{8}$$

In contrast to [14]-[15], where the method of partial discretization is applied to study of parametrical system oscillations, in this work it is used directly to a solution of the perturbation equation in terms of $\delta(t)$. It is possible to investigate stability of the state by analyzing the nature of behavior of $\delta(t)$, according to Lyapunov stability criterion. If the magnitude decreases with time t (decaying process)

then $\delta f \to 0$, i.e. the state is stable. If the oscillatory process is growing then we have an unstable state.

Efficiency of the offered method will be shown below on the example of stability analysis of resonant oscillations of physically nonlinear systems.

IV. ANALYTICAL SOLUTION OF THE HILL EQUATION IN THE CASE OF PHYSICALLY NONLINEAR SYSTEMS

As an example, consider the motion of physically nonlinear systems. Equations of motion for these systems are taken in the form:

$$\frac{d^2 f}{dt^2} + k_1 \frac{df}{dt} + k_2 \left(\frac{df}{dt}\right)^2 + \alpha_1 f + \alpha_2 f^2 = F_0 + F_1 \cos \Omega t. \qquad (9)$$

In (9) dissipative forces which are supposed to be nonlinear and viscous due to damping properties of physically nonlinear media (rubber and similar materials used as oscillation dampers) are taken into account.

Physical nonlinearity of the system is characterized by an arbitrary angle of rotation of cross elements that corresponds to quadratic nonlinearity of the restoring force.

Stability of a basic resonance is investigated. Solution of (9) is given by:

$$f(t) = r_0 + r_1 \cos(\Omega t - \varphi_1). \qquad (10)$$

The Hill equation in this case is represented as [6]:

$$\frac{d^2 \eta}{dt^2} + \eta \left[\theta_0 + \theta_{1s} \sin \Omega t + \theta_{1c} \cos \Omega t + \theta_{2s} \sin 2\Omega t + \theta_{2c} \cos 2\Omega t\right] = 0, \qquad (11)$$

where

$$\theta_0 = \alpha_1 + 2\,\alpha_2 r_0 - 0.25\, k_1^2 - 0.5\, k_2^2 r_1^2 \Omega^2,$$

$$\theta_{1s} = (2\,\alpha_2 r_1 + k_2 r_1 \Omega^2) \sin \varphi_1 + k_1 k_2 r_1 \Omega \cos \varphi_1,$$

$$\theta_{1c} = (2\,\alpha_2 r_1 + k_2 r_1 \Omega^2) \cos \varphi_1 - k_1 k_2 r_1 \Omega \sin \varphi_1, \qquad (12)$$

$$\theta_{2s} = 0.5\, k_2^2 r_1^2 \Omega^2 \sin 2\varphi_1,$$

$$\theta_{2c} = 0.5\, k_2^2 r_1^2 \Omega^2 \cos 2\varphi_1.$$

According to the above-specified technique, under the given initial conditions, and by the method of partial discretization the analytical solution of (11)-(12) has been obtained:

$$\begin{aligned}
\eta(t_k) = \Big[& -(t_1 + t_2)(\theta_0 + \theta_{1s} \sin \Omega t_1 + \theta_{1c} \cos \Omega t_1 + \theta_{2s} \sin 2\Omega t_1 \\
& + \theta_{2c} \cos 2\Omega t_1) \eta(t_1) \left(\frac{t_k + t_{k+1}}{2} - t_1\right)\Big] \Big/ \Big[1 + \frac{1}{2}(t_{k+1} - t_k) \\
& \cdot (t_{k+1} - t_{k-1})(\theta_0 + \theta_{1s} \sin \Omega t_k + \theta_{1c} \cos \Omega t_k + \theta_{2s} \sin 2\Omega t_k \\
& + \theta_{2c} \cos 2\Omega t_k)\Big] - \Big[\sum_{j=2}^{k-1} (t_{j+1} - t_{j-1})(\theta_0 + \theta_{1s} \sin \Omega t_j \\
& + \theta_{1c} \cos \Omega t_j + \theta_{2s} \sin 2\Omega t_j + \theta_{2c} \cos 2\Omega t_j) \eta(t_j) \\
& \cdot \left(\frac{t_k + t_{k+1}}{2} - t_j\right)\Big] \Big/ \Big[1 + \frac{1}{2}(t_{k+1} - t_k)(t_{k+1} - t_{k-1}) \\
& \cdot (\theta_0 + \theta_{1s} \sin \Omega t_k + \theta_{1c} \cos \Omega t_k + \theta_{2c} \sin 2\Omega t_k \\
& + \theta_{2c} \cos 2\Omega t_k)\Big] + \Big[\dot{\eta}_0 \frac{t_k + t_{k+1}}{2} + \eta_0\Big] \Big/ \Big[1 + \frac{1}{2}(t_{k+1} - t_k) \\
& \cdot (t_{k+1} - t_{k-1})(\theta_0 + \theta_{1s} \sin \Omega t_k + \theta_{1c} \cos \Omega t_k \\
& + \theta_{2s} \sin 2\Omega t_k + \theta_{2c} \cos 2\Omega t_k)\Big].
\end{aligned} \qquad (13)$$

V. NUMERICAL RESULTS

Solution (13) is a recurrent formula for discrete representation of the solution $\eta(t)$ with time on *k*-th step of splitting of the argument t. By analyzing the nature of the behavior of $\eta(t)$, we can judge the stability of the studied state.

In this paper numerical analysis of the behavior of $\eta(t)$ giving a representation of the behavior of a small variation δf with time is realized.

Calculations were done for the parameters of the system $k_1 = 0.2$; $k_2 = 0.1$; $\alpha_1 = 5$; $\alpha_2 = 0.5$; $F_0 = 5$; $F_1 = 50$. Step of discretization was accepted as $\Delta t = 0.05$.

Stability of the solution (13) was studied by putting on amplitude-frequency characteristics of a basic resonance (Fig.1, curve 2) three frequency areas in to-resonant, resonant and post-resonant modes of oscillations.

It is established that both to-resonant and post-resonant modes of oscillations are decaying (Fig. 2, Fig.3) that does not contradict the physical sense of the phenomena investigated. In a zone of resonant frequencies growth of oscillation amplitude is obtained that means the process is instable (Fig.4).

Here, as well as in [16] where research on stability analysis of motion of geometrically nonlinear systems was conducted by method of partial discretization, research results correspond well to graphs of amplitude-frequency characteristics of a system basic resonance (Fig.1).

Thus, application of the partial discretization method to studying stability of oscillations allows to obtain the analytical solution and determine zones of stable and unstable system oscillations. Selection of corresponding geometrical and physical parameters of the elastic system by means of their variation will help to avoid undesirable resonant phenomena in operating system modes.

ISBN: 978-1-61804-256-9

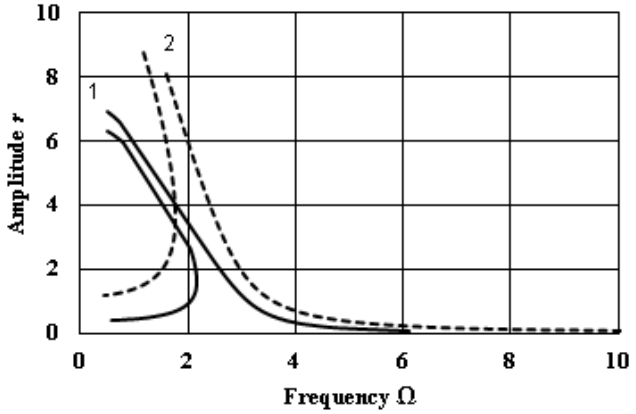


Fig. 1 amplitude-frequency characteristics of a basic resonance at
$\alpha_2 = 1;\ F_1 = 10$ (curve 1), $\alpha_2 = 0.5;\ F_1 = 50$ (curve 2)

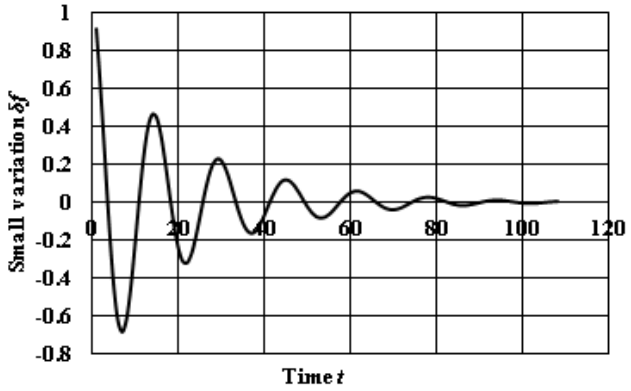


Fig. 2 behavior of the physically nonlinear system in the to-resonant zone of oscillations at $\Omega = 0.5,\ r = 1.5$

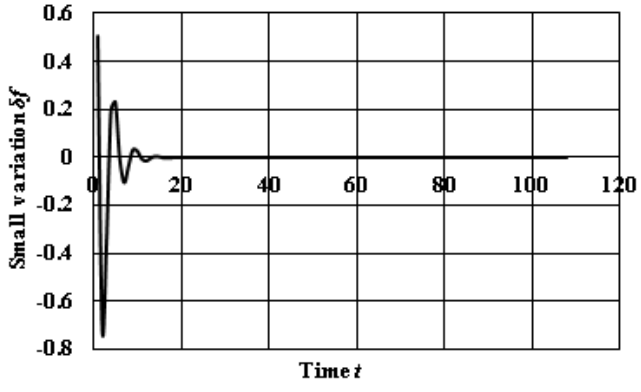


Fig. 3 behavior of the physically nonlinear system in the post-resonant zone of oscillations at $\Omega = 7.26,\ r = 0.15$

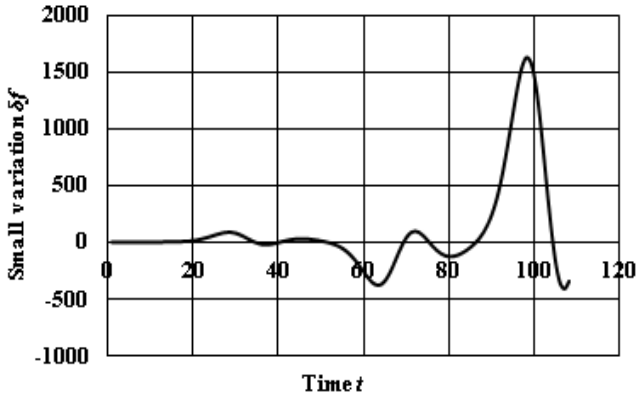


Fig. 4 behavior of the physically nonlinear system in the resonant zone of oscillations at $\Omega = 1.8,\ r = 8$

CONCLUSION

In this work, according to the offered criterion of dynamic stability of elastic systems, stability of motion of nonlinear mechanical systems and methods of their analysis have been considered.

Steady motion of nonlinear systems is considered as their movement in the absence of oscillatory processes. These requirements are identical to determining of Lyapunov stability. Therefore, the technique of stability analysis of nonlinear systems is based on the analysis of solutions for perturbation equations. As a method for solution to the problem, the partial discretization method is offered. The essence of this method consists in discretization of variable coefficients of the Hill equation in a class of the generalized functions. The solution of the Hill equation is considerably simplified by identifying its variable coefficients as constants on each step of discretization. The obtained analytical solution of the perturbation equation is a recurrent formula for calculation of oscillation amplitudes. It allows to predict parametrical instability of resonant modes of motion of nonlinear systems. Efficiency of the used method is shown on the example of physically nonlinear systems. Research results correspond well to the known results obtained by other methods.

REFERENCES

[1] A. Kydyrbekuly, L. Khajiyeva, "The dynamic stability of flat planar linked mechanisms with regard for nonlinear characteristics of elastic links," *in IX Int. Conf. on the Theory of Machines and Mechanisms*, Czech Republic, 2004, pp. 373-379.

[2] Sh. Aitaliev, G. Masanov, L. Khajiyeva, "Dynamic stability of nonlinear and elastic rod systems," *in X Interuniv. Conf. on Math. and Mech.*, Almaty, 2004, vol. 1, pp. 46-52 (in Russian).

[3] L. A. Khajiyeva, "About modeling of machine dynamics with nonlinear characteristics of deformable links," *Reports of NAS RK*, 2005, vol. 4, pp. 38-43 (in Russian).

[4] A. Kydyrbekuly, L. Khajiyeva, "The study of flat mechanism of higher class dynamics with regard for finite deformation of elastic links," *in Proc. X World Congress on the Theory of Machine and Mechanisms*, Finland, Oulu, 1999, vol.4, pp.1453-1458.

ISBN: 978-1-61804-256-9

[5] A. Kydyrbekuly, L. Khajiyeva, “Modeling of physically nonlinear mediums and mediums with initial stresses,” *Proceedings of All-Russian School-Seminar on Modern Problems of Mechanics of Deformable Rigid Body* Novosibirsk, 2003, pp. 119-123 (in Russian).
[6] Sh. Aitaliev, G. Masanov, A. Kydyrbekuly, L. Khajiyeva, “Dynamics of mechanisms with elastic links,” *in Proc.12-th Int. Workshop on Comp. Kinematics*, Italy, Cassino, 2005, pp. 1-11.
[7] I. G. Malkin, *Theory of motion stability.* M: Nauka, 1966, 530 p. (in Russian).
[8] D. A. Merkin, *Inroduction to the theory of motion stability.* M: Nauka, 1976, 320 p. (in Russian).
[9] N. N. Bogolyubov, Yu. A. Mitropolskiy. *Asymptotic methods in the theory of nonlinear oscillations*. M: Nauka, 1974, 504 p. (in Russian).
[10] C. Hayash, *Nonlinear oscillations in physical systems*. M: Mir, 1968, 423 p.
[11] A. Tondl, *Non-linear oscillations in mechanical systems*. M: Mir, 1973, 334 p. (in Russian).
[12] W. Szemplinska-Stupnicka, “Higher harmonic oscillations in heteronomous nonlinear systems with one degree of freedom,” *Internal J. Nonlinear Mech.*, 1968, vol. 3, pp. 17-30.
[13] A. A. Zevin, "Stability of periodic oscillations in systems with soft and rigid nonlinearity," Appl. Math and Mechanics, 1980, vol. 44 (4), pp. 640-649 (in Russian).
[14] A. N. Tyurekhodzhayev, "Some problems of modern engineering," *in Proc. Intern. Conf. Actual Problems of Mechanics and Mechanical Engineering*, Almaty, 2005, vol. 1, pp. 11-28 (in Russian).
[15] A. N. Tyurekhodzhayev, L. Kh. Lukpanova, "Solution of the problem of dynamic stability of an elastic rod by partial discretization method for differential equations” *in Proc. Intern. Conf. Actual Problems of Mechanics and Mechanical Engineering*, Almaty, 2005, vol. 2, pp. 67-70 (in Russian).
[16] L. Khajiyeva, A. Sergaliyev, A. Umbetkulova, “Dynamic Analysis of Steel and Dural Drill Rods,” *Advanced Materials Research,* Switzerland, 2013, Vol. 705, pp. 91-96.

Comparison of Profiling Power Analysis Attacks Using Templates and Multi-Layer Perceptron Network

Zdenek Martinasek and Lukas Malina

***Abstract*—In recent years, the cryptographic community has explored new approaches of power analysis based on machine learning models such as Support Vector Machine (SVM), Multi-Layer Perceptron (MLP) or Random Forest (RF). Realized experiments proved that the method based on MLP can provide almost 100% success rate after optimization. Nevertheless, this description of results is based on the first order success rate that is not enough satisfactory because this value can be deceiving. Moreover, the power analysis method based on MLP has not been compared with other well-known approaches such as template attacks or stochastic attacks yet. In this paper, we introduce the first fair comparison of power analysis attacks based on MLP and templates. The comparison is accomplished by using the identical data set and number of interesting points in power traces. We follow the unified framework for implemented side-channel attacks therefore we use guessing entropy as a metric of comparison.**

***Keywords*—Power Analysis, Neural Network, Template Attack, Comparison.**

I. Introduction

Power analysis (PA) measures and analyzes the power consumption of cryptographic devices depending on their activity. It was introduced by Kocher in [1]. The goal of PA is to determine the sensitive information of cryptographic devices from the measured power consumption and to apply the obtained information in order to abuse the cryptographic device. A detailed description of power analysis including side-channel sources, testbeds, statistical tests and countermeasures is summarized in the book [2].

A. Related Work

Application of neural networks in the field of power analysis was first published in [3]. Naturally, this work was followed by other authors, e.g. [4], [5], who dealt with the classification of individual power prints. These works are mostly oriented towards reverse engineering. Yang et al. [6] proposed MLP in order to create a power consumption model of a cryptographic device in DPA based on correlation coefficient. In recent years, the cryptographic community has explored new approaches based on machine learning models. Lerman et al. [7], [8] compared a template attack (TA) with a binary machine learning approach based on non-parametric methods. Hospodar et al.[9], [10] analysed the SVM on a software implementation of a block cipher. Heuser et al. [11] created the general description of the SVM attack and compared this approach with the template attack. In 2013, Bartkewitz [12] applied a multi-class machine learning model that improves the attack success rate with respect to the binary approach. Recently, Lerman et al. [13] proposed a machine learning approach that takes into account the temporal dependencies between power values. This method improves the success rate of an attack in a low signal-to-noise ratio with respect to classification methods. Lerman et al. [14] presented a machine learning attack against a masking countermeasure, using the dataset of the DPA Contest v4. Interesting method of power analysis based on a multi-layer perceptron was first presented in [15]. In this work, the authors used a neural network directly for the classification of the AES secret key. In [16], this MLP approach was optimized by using the preprocessing of the power traces measured.

B. Contribution

In [15], [16], the authors used the first order success rate for efficiency description of the proposed MLP power analysis method. This is not sufficiently reliable because this value can be deceiving [17]. According the framework, the guessing entropy represents an appropriate metric of two side analysis attack implementation [17]. The metric measures the average number of key candidates to test after the side-channel attack.

Other important fact is that both methods based on MLP (original implementation and optimized one) have not been yet compared with other well-known approaches such as the template attack or the stochastic attack. In this paper, we introduce the first fair comparison of power analysis attacks based on the MLP and templates. The comparison is accomplished by using the identical data set including a number of interesting points. In previous researches described in [15], [16], the adversary uses 1200 interesting points to realize the attack. This large number of interesting points is not practically applicable to TA because of possible numerical problems connected with a covariance matrix. Moreover, we create a general description of the MLP aimed for byte classification including the structure, setting and training algorithm, because this information was also missing in previous research.

This research was funded by project OPVK CZ.1.07/2.2.00/28.0062 "Joint activities of BUT and VSB-TUO while creating the content of accredited technical courses in ICT"

Z. Martinasek is with the Brno University of technology, Brno, Czech republic. He is now with the Department of Telecommunications, Technicka 12, 616 00 Brno, (phone:+420 541 146 960, e-mail: martinasek@feec.vutbr.cz).

L. Malina is with the Brno University of technology, Brno, Czech republic. He is now with the Department of Telecommunications, Technicka 12, 616 00 Brno, (e-mail: malina@feec.vutbr.cz)

ISBN: 978-1-61804-256-9

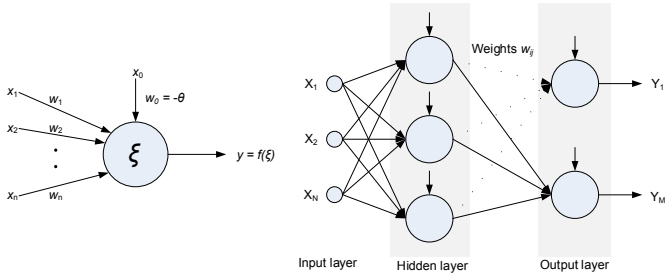


Fig. 1. The general structure of neural network.

II. General Description of the MLP

This section provides only a basic information about the neural networks that we used during the attack (the basic structure and the training algorithm of the MLP). We refer the work [18], [19] for more specific information. The main goal of this section is to show how to use MLP to realize the side channel attack.

The basic element of an artificial neural network is a formal neuron, often called as a perceptron in the literature. The basic model of the neuron is shown on the left side in Fig. 1. The neuron contains x_i inputs that are multiplied by the weights w_i, where $i = 1$ to n. Input x_0 multiplied by the weight $w_0 = -\theta$ determines the threshold of the neuron (bias). During the training of the neuron, weights are updated to achieve a desired output value. Firstly, a post-synaptic potential is calculated. It is defined as the internal function of the neuron:

$$\xi = \sum_{i=1}^{n} x_i w_i - \theta. \tag{1}$$

Subsequently, the output value of the neuron is calculated as $y = f(\xi)$ where f represents a non-linear function, mostly a sigmoid. Naturally, one formal neuron is not able to solve complex problems, therefore we use neurons (perceptrons) connected into a network. The multilayer perceptron consists of two or more layers of neurons that are denoted as an output layer and a hidden layer. Each neuron in one layer is connected with a certain weight w_{ij} to every neuron in the following layer. Frequently, the input layer is not included when one is counting the number of layers because the input layer is not composed of neurons. We follow this notation in this article. An example of the two-layer neural network is shown in Fig. 1 (on the right side).

These networks are modifications of the standard linear perceptron and can distinguish data that are not linearly separable [19]. These networks are widely used for a pattern classification, recognition, prediction and approximation and utilize mostly a supervised learning method called backpropagation [20]. The backpropagation (BPG) algorithm is an iterative gradient learning algorithm which minimizes squares of a cost function using the adaptation of the synaptic weights. This method is described with the following steps (the following equations are valid for the two-layer neural network which is shown in Fig. 1):

- **Step 1:** Weights w_{ij} and thresholds θ of each neuron are initialized with random values.
- **Step 2:** An input vector $\mathbf{X} = [x_1, \ldots, x_N]^T$ and a desired output vector $\mathbf{D} = [d_1, \ldots, d_M]^T$ are applied to the neural network. In other words, one creates a training set containing pairs of $\mathbf{T} = \{[\mathbf{X_1}, \mathbf{D_1}], [\mathbf{X_2}, \mathbf{D_2}], \cdots, [\mathbf{X_n}, \mathbf{D_n}]\}$, where n denotes the number of training set patterns and the training set prepared is applied to the neural network. Provided, that NN represents a ordinary classifier which classifies input data to the desired output groups, the $\mathbf{D}$ represents mostly a classification matrix where the desired outputs are labeled by value 1 and other outputs 0.
- **Step 3:** The current output of each neuron is calculated by the following equations:

$$y_k(t) = f_s\Big(\sum_{k=1}^{N_1} w'_{jk}(t) x'_j(t) - \theta'_k\Big), \tag{2}$$

$$x'_j(t) = f_s\Big(\sum_{i=1}^{N} w_{ij}(t) x_i(t) - \theta_j\Big), \tag{3}$$

  where $1 \leq k \leq M$ denoted output layer and $1 \leq j \leq N_1$ hidden layer.
- **Step 4:** Weights and thresholds are applied according to the following equation:

$$w_{ij}(t+1) = w_{ij}(t) + \eta \delta_j x_i. \tag{4}$$

  Adaptation of weight values starts at the output neurons and proceeds recursively back to the input neurons. In this equation, w_{ij} denotes weights between the i-th hidden or input neuron and the neuron j-th at time t. Output of the i-th neuron is denoted as x'_i, η represents the learning coefficient and δ_j is an error of neuron which is calculated as follows:

$$\delta_j = y_j(1 - y_j)(d_j - y_j), \text{ (output layer)}, \tag{5}$$

$$\delta_j = x'_j(1 - x'_j)\Big(\sum_{1}^{M} \delta_k w_{jk}\Big), \text{ (hidden layer)}, \tag{6}$$

  where k represents all neurons in the output layer.
- **Step 5:** Steps from 3 to 5 are repeated until the error value is less than the predetermined value.

During the training of NN which is based on the BPG algorithm, some problems may occur. These problems are caused by inappropriate setting of training parameters or the improper initialization of weights and thresholds. These difficulties can be reduced by using a modification of the basic algorithm such as Back-Propagation with Momentum or Conjugate Gradient Backpropagation.

III. General description of MLP attack

In this section, we describe the general usage of the MLP in power analysis attack. Machine learning algorithms are mostly used in profiled attacks where an adversary needs a physical access to a pair of identical devices, which we call a profiling device and a target device. Basically, these attacks consist of two phases. In the first phase, the adversary analyzes the

ISBN: 978-1-61804-256-9

profiling device and then, in the second phase, the adversary attacks the target device. Typical examples are template-based attacks [21], [2], [22]. By contrast, non-profiled attacks are one-phase attacks that perform the attack directly on the target device such as DPA based on the correlation coefficient [23].

A. Profiling Phase of MLP Attack

In the attack based on the MLP, we assume that we can characterize the profiling device using a well trained neural network. We assume that desired value by adversary is the secret key stored in cryptographic device. This means that one can create and train a NN for a certain part of a cryptographic algorithm. We execute this sequence of instructions on the profiling device with the same data d and different key values k_j to record the power consumption. After measuring n power traces, it is possible to create the matrix $\mathbf{X}_n$ that contains power traces corresponding to a pair of (d, k_i). These pairs represent a training set $\mathbf{T}$ of the neural network. Input values are power traces measured and values of secret key k_i represent the desired output of the neural network. In this case, secret key values k_i can be easily represented using the $n \times 256$ classification matrix $\mathbf{D}$.

After the measurement phase, an adversary creates a neural network. The number of input neurons has to be equal to the numbers of chosen interesting points. We use only interesting points because memory limitation and time-consuming training process (similar situation like in classical Template attack). Generally, the setting of the hidden layer depends on problem to solve and the training set, therefore the adversary has to set the number of hidden layers and neurons experimentally. The output layer should contain the desired number of neurons corresponding to the aim of the attack (output byte of S-Box, byte of the secret key, Hamming weight etc.). In our example, the NN is aimed on byte classification, therefore the output layer contains 256 neurons. In the last step of the profiling phase, the adversary trains the neural network created by the prepared training set and the chosen training algorithm.

B. Attack Phase of MLP Attack

During the attack phase, the adversary uses a well-trained NN together with a measured power trace from the target device (denoted as $\mathbf{t}$) to determine the secret key value. The adversary puts the $\mathbf{t} = [x_1, \ldots, x_N]^T$ as an input to NN and it classifies the output values using the calculation:

$$y_k = f_s(\sum_{k=1}^{N_1} w_{jk}^{'} x_j^{'} - \theta_k^{'}),\ 1 \leq k \leq M, \tag{7}$$

where w_{ij} denotes weights between i-th hidden neuron (or the input neuron) and the neuron j-th and $x_i^{'}$ denotes the output of hidden neurons:

$$x_j^{'} = f_s(\sum_{i=1}^{N} w_{ij} x_i - \theta_j),\ 1 \leq j \leq N_1. \tag{8}$$

The result of this classification is a vector $\mathbf{g} = [g_1, g_2, \ldots, g_M]$ which contains the probability value 0 to 1 for every output value. The probabilities show how well the measured trace $\mathbf{t}$ corresponds to the training patterns. Intuitively, the highest probability should indicate the correct training pattern in the training set $\mathbf{T}$ and because each training pattern $\mathbf{X_n}$ is associated with a desire value (in our case secret key), the adversary obtains the information about secret key stored in the target device.

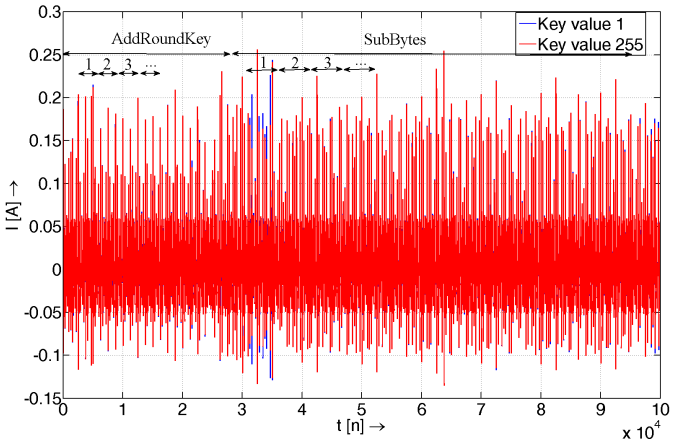


Fig. 2. Measured power traces for different first key values.

IV. Testbed and Implementation Description

This section summarizes the most important facts about the experimental setup and the implementation of the attacks. A complete AES algorithm with a key length of 128 bits was implemented into the cryptographic module and the synchronization was performed only for the `AddRoundKey` and `SubBytes` operations in the initialization phase of the algorithm. The stored secret key can be expressed in bytes as $K_{sec} = \{k_1, k_2, \ldots, k_{16}\}$ where k_i represents individual bytes of the key. The program allowed setting of the secret key and plain text value indicated this operation by sending the respective value via a serial port to a computer. The synchronization signal and the communication with the computer did not affect the power consumption of the cryptographic module. The cryptographic module was represented by PIC 8-bit microcontroller, and for the power consumption measurement we used CT-6 current probe and Tektronix DPO-4032 digital oscilloscope. We used standard operating conditions with 5 V power supply.

Because our implementation was realized in the assembly language and the executed instructions of examined operations (`AddRoundKey` and `SubBytes`) were exactly the same for every key byte k_i, we assume that it is possible to use parts of power traces where first byte is processed (see Fig. 2) to build template and train the neural network to determine the whole secret key byte by byte. In the first step, we determine the value of k_1, and in the second step, byte k_2 and so on. The difference between these steps is in the division of the power traces into parts corresponding to the time intervals in which the cryptographic device works with the respective bytes of the secret key. The division of power traces is indicated in Fig. 2 by numbers and every part of a power trace contained $1,200$ samples. We verify this assumption experimentally and

ISBN: 978-1-61804-256-9

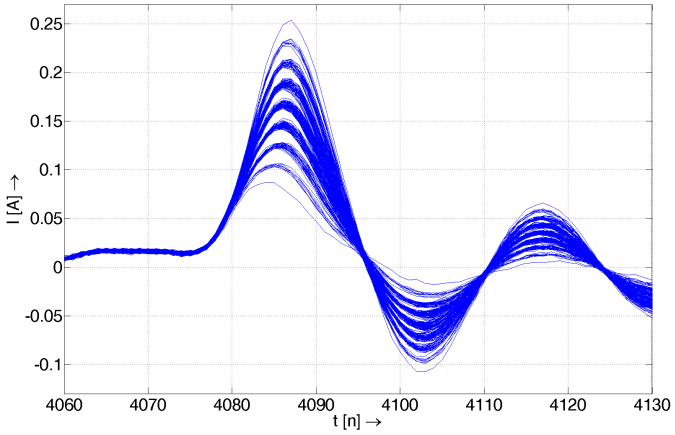


Fig. 3. Detail of measured 256 power traces.

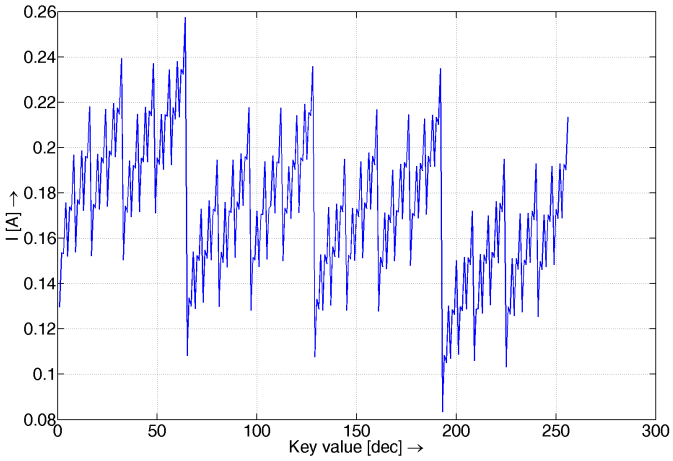


Fig. 4. Measured leaks of Hamming weight for point $4,086$.

it is naturally conditioned by the excellent synchronization of measured power traces.

We measured a set of $2,560$ power traces where ten power traces were independently stored for each value of the first secret key byte. This number of power traces was chosen because we wanted to compare both implementation of the attack (MLP approach and template) using the typical 10-fold cross-validation. In data mining and machine learning, the 10-fold cross-validation is the most common method of model verification. Cross-validation (CV) is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one is used for learning a model and the other one is used for the model validation. In typical cross-validation, the training and validation sets must cross-over in successive rounds that each data point has a chance of being validated against. Therefore, we used 9 power traces in profiling phase of the attack and one power trace in attack phase in every step of validation.

We chose five interesting points according to the information provided in [24]. Our algorithm searched for the maximum differences of an average power consummation and power consumption corresponding to key value 1. The algorithm accepted only the maximums that had a distance of at least one clock cycle from each other. This restriction for having interesting points not too close from each other avoids numerical problems during the covariance matrix inverting. Measured power traces were properly synchronized and our device leaks Hamming weight (HW) of processed data. These facts confirm the plots shown in Fig. 3 and Fig. 4. Figure 3 shows the detail of power traces that correspond to MOV instruction where data values 0 to 255 were processed. Figure 4 shows plot of these measured power traces for one point $t = 4,086$. Each of our chosen interesting points leaked HW of processed data. Same chosen points were used for the template creation and the neural network model.

A well-known fact is that noise always poses the problem during the power consumption measurement. We performed the experimental measurements of a test bed that were made according to the information provided in [2] and we established that the noise level was distributed according to the normal distribution with the parameters $\mu = 0$mA and $\sigma = 5$mA. Every stored power trace was calculated as an average power trace from ten power traces measured using the digital oscilloscope to reduce the electronic noise.

A. Template Attack Implementation

We implemented the classical template attack and reduced template attack to compare the classification results with MLP attack. We were interested in effective template attack based on pooled covariance matrix [22], therefore we calculated the pool covariance matrix as an average value of all covariance matrices and we calculated the probability density function (Eq. 9) with this matrix. Implementations of template attacks were done according to the Eq. 9:

$$p(\mathbf{t}; (\mathbf{m}, \mathbf{C})_{d_i,k_j}) = \frac{exp(-\frac{1}{2} \cdot (\mathbf{t} - \mathbf{m})' \cdot \mathbf{C}^{-1} \cdot (\mathbf{t} - \mathbf{m}))}{\sqrt{(2 \cdot \pi)^{NP} \cdot det(\mathbf{C})}} \tag{9}$$

where $(\mathbf{m}, \mathbf{C})$ represents templates prepared in profiling phase based on multivariate normal distribution that is fully defined by a mean vector and a covariance matrix. Measured power trace from the target device is denoted as $\mathbf{t}$ and NI is the number of interesting points. In following text, classical template, reduced template and template attack based on the pooled covariance matrix are denoted as T_{cls}, T_{red} and T_{pool} sequentially. All template attack implementations were made in the Matlab environment.

B. MLP Attack Implementation

We created and trained the neural network in Matlab using the Netlab neural network toolbox [18]. Ian Nabney and Christopher Bishop from Aston University in Birmingham are the authors of this toolbox and it is available for downloading. We created a typical two layer perception network and we used optimized learning based on the scaled conjugate gradient algorithm (see Sec. II). A standard sigmoid was chosen as an activation function. The created NN is shown in Fig. 5. The input layer contained 5 inputs corresponding with interesting

ISBN: 978-1-61804-256-9

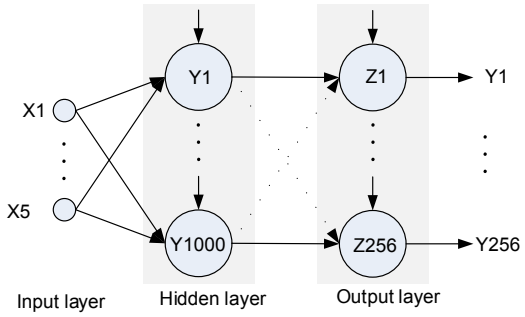


Fig. 5. Created neural network.

points, hidden layer contained $1,000$ neurons and output layer had 256 neurons and we used 200 training cycles. This implementation of NN is denoted as NN_{org} and practically corresponds to the original approach described in [15]. This model differs only in number of inputs. We created the second NN according the optimization based on preprocessing of measured power traces [16]. This implementation is denoted as NN_{opt}.

V. Obtained Results

The measured set of $2,560$ power traces was used for the comparison of implemented methods. We realized a typical 10-fold cross-validation, where nine power traces were used for the template preparation and neural network training in the profiling phase and one power trace was used in the attack phase in every step of the cross-validation. We used the guessing entropy to compare our implemented attacks.

The guessing entropy is defined as follows: let $\mathbf{g} = [p_1, p_2, \ldots, p_N]$ contains the probability such as $p_1 \geq p_2, \geq \ldots, \geq p_N$ of all possible key candidates after N iterations of Eq. 9 or Eq. 7. Indices i correspond with the correct key in $\mathbf{g}$. After the realization of S experiments, one obtains a matrix $\mathbf{G} = [g_1, ..., g_S]$ and a corresponding vector $\mathbf{i} = [i_1, \ldots, i_S]$. Then the guessing entropy determines the average position of the correct key:

$$GE = \frac{1}{S} \sum_{x=1}^{S} i_x. \tag{10}$$

In other words, the guessing entropy describes the average number of guesses, required for recovering the secret key [17], [11].

In the first experiment, we determined the value of one byte of the secret key from one measured power trace. We tried this for all 256 power traces measured corresponding to every key values from 0 to 255. In other words, we determined the value of 256 individual bytes in every step of the cross-validation. After the realization, we calculated the GE according to the Eq. 10. Obtained results are summarized in Tab. I, where ϕ denotes an average value calculated from every cross-validations realized. The template attack based on the pooled covariance matrix T_{pol} achieved the best result in one byte guessing but it is important that the classification based on NN was not much worse. The original implementation of the neural network NN_{org} was the worst of all implemented attacks and achieved $GE = 1.18$ in average. The optimized method achieved $GE = 1.04$ that was almost identical with template attacks.

TABLE I
Guessing entropy for the individual byte determination.

Step of CV	NN_{org}	NN_{opt}	T_{cls}	T_{red}	T_{pol}
1	1.16	1.02	1.07	1.04	1.02
2	1.18	1.04	1.07	1.06	1.02
3	1.32	1.03	1.04	1.04	1.03
4	1.16	1.05	1.04	1.04	1.02
5	1.16	1.05	1.07	1.05	1.02
6	1.23	1.04	1.04	1.04	1.02
7	1.15	1.03	1.08	1.03	1.02
8	1.11	1.05	1.07	1.02	1.02
9	1.18	1.06	1.08	1.02	1.00
10	1.17	1.03	1.03	1.04	1.01
ϕ	1.18	1.04	1.06	1.04	1.02

In the second experiment, we determined the whole 128 bit secret key by using the 16 power traces measured. The secret key stored had value $K = [29, 245, 48, 93, 215, 65, 139, 198, 5, 232, 81, 107, 173, 243, 24, 151]$. Obtained results are written in Tab. II. The second experiment confirmed the previous results. The adversary needs about 18 guesses to determine the correct secret key after the side-channel attack based on the original implementation of neural network NN_{org}. The results of the optimized method were almost identical with template attacks. Potential adversary would need in average about 4 guesses to determine the secret key value after the side-channel attack. Our experiments confirm that success revelation of secret key is comparable for MLP and template based attacks (identical number of interesting points, number of power traces and so on). MLP is able to be trained only for a few interesting points of power traces. In order to complete the comparison of implemented attacks, Tab. III provides the information about the time complexity of attack phase τ and memory complexity m.

TABLE II
Guessing entropy for the whole secret key determination.

Step of CV	NN_{org}	NN_{opt}	T_{cls}	T_{red}	T_{pol}
1	4,00	2,00	4,00	4,00	2,00
2	24,00	4,00	4,00	1,00	2,00
3	32,00	2,00	4,00	4,00	8,00
4	24,00	8,00	2,00	4,00	4,00
5	4,00	2,00	4,00	4,00	4,00
6	30,00	4,00	16,00	4,00	4,00
7	8,00	2,00	4,00	2,00	2,00
8	16,00	6,00	8,00	2,00	2,00
9	32,00	2,00	4,00	4,00	2,00
10	2,00	1,00	2,00	1,00	1,00
ϕ	17,60	3,30	5,20	3,00	3,10

TABLE III
Obtained results

	NN_{org}	NN_{opt}	T_{cls}	T_{red}	T_{pol}
τ $[ms]$	1.59	1.11	174.89	149.85	221.66
m $[kB]$	1,920.00	1,920.00	94.20	22.30	22.60

ISBN: 978-1-61804-256-9

VI. Conclusion

In this paper, we made the first fair comparison of power analysis using the MLP with well-known template attacks. We followed the unified framework for two implementations of the side-channel attack, therefore we used a guessing entropy as a metric of comparison. The comparison was made by using the same data set and same number of interesting points for all implementation. We described the usage of MLP in power analysis attack including the structure, setting and training algorithm because these information were missing in the previous research.

The experiment realized, that determined the whole secrete key of the AES algorithm, confirmed that the efficiency of the power analysis attack based on MLP and the template attack is comparable. By contrast, the adversary needs about 18 guesses to determine the correct secret key using the original implementation of the MLP attack. This result is three times worse in comparison with the classical template attack that needs about 5.2 guesses to reveal the whole secret key. For these reasons, we do not recommend a usage of original implementation of MLP attack. The results of optimized method were almost identical with template attacks. Potential adversary would need in average about 4 guesses after the side-channel attack to determine the secret key value of AES algorithm.

References

[1] P. C. Kocher, J. Jaffe, and B. Jun, "Differential power analysis," in *CRYPTO '99: Proceedings of the 19th Annual International Cryptology Conference on Advances in Cryptology*. London, UK: Springer-Verlag, 1999, pp. 388–397.

[2] S. Mangard, E. Oswald, and T. Popp, *Power Analysis Attacks: Revealing the Secrets of Smart Cards (Advances in Information Security)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[3] J.-J. Quisquater and D. Samyde, "Automatic code recognition for smart cards using a kohonen neural network," in *Proceedings of the 5th conference on Smart Card Research and Advanced Application Conference - Volume 5*, ser. CARDIS'02, Berkeley, CA, USA, 2002, pp. 6–6.

[4] J. Kur, T. Smolka, and P. Svenda, "Improving resiliency of java card code against power analysis," in *Mikulaska kryptobesidka, Sbornik prispevku*, 2009, pp. 29–39.

[5] Z. Martinasek, T. Macha, and V. Zeman, "Classifier of power side channel," in *Proceedings of NIMT2010*, September 2010.

[6] S. Yang, Y. Zhou, J. Liu, and D. Chen, "Back propagation neural network based leakage characterization for practical security analysis of cryptographic implementations," in *Proceedings of the 14th international conference on Information Security and Cryptology*, ser. ICISC'11. Springer-Verlag, 2012, pp. 169–185.

[7] L. Lerman, G. Bontempi, and O. Markowitch, "Side channel attack: an approach based on machine learningn," in *COSADE 2011 - Second International Workshop on Constructive Side-Channel Analysis and Secure Design*, 2011, pp. 29–41.

[8] ——, "Power analysis attack: an approach based on machine learning," *International Journal of Applied Cryptography*, 2013.

[9] G. Hospodar, B. Gierlichs, E. D. Mulder, I. Verbauwhede, and J. Vandewalle, "Machine learning in side-channel analysis: a first study," *J. Cryptographic Engineering*, vol. 1, no. 4, pp. 293–302, 2011.

[10] G. Hospodar, E. Mulder, B. Gierlichs, J. Vandewalle, and I. Verbauwhede, "Least squares support vector machines for side-channel analysis," in *COSADE 2011 - Second International Workshop on Constructive Side-Channel Analysis and Secure Design*, 2011, pp. 293–302.

[11] A. Heuser and M. Zohner, "Intelligent machine homicide - breaking cryptographic devices using support vector machines," in *COSADE*, 2012, pp. 249–264.

[12] T. Bartkewitz and K. Lemke-Rust, "Efficient template attacks based on probabilistic multi-class support vector machines," in *Proceedings of the 11th international conference on Smart Card Research and Advanced Applications*, ser. CARDIS'12. Springer-Verlag, 2013, pp. 263–276.

[13] L. Lerman, G. Bontempi, S. B. Taieb, and O. Markowitch, "A time series approach for profiling attack." in *SPACE*, ser. Lecture Notes in Computer Science, B. Gierlichs, S. Guilley, and D. Mukhopadhyay, Eds., vol. 8204. Springer, 2013, pp. 75–94.

[14] L. Lerman, S. F. Medeiros, G. Bontempi, and O. Markowitch, "A machine learning approach against a masked aes," in *CARDIS*, 2013, in print.

[15] Z. Martinasek and V. Zeman, "Innovative method of the power analysis," *Radioengineering*, vol. 22, no. 2, 2013, iF 0.687.

[16] Z. Martinasek, J. Hajny, and L. Malina, "Optimization of power analysis using neural network," in *CARDIS*, 2013, in print.

[17] F.-X. Standaert, T. Malkin, and M. Yung, "A unified framework for the analysis of side-channel key recovery attacks," in *EUROCRYPT*, 2009, pp. 443–461.

[18] I. T. Nabney, *NETLAB: algorithms for pattern recognition*, ser. Advances in Pattern Recognition. New York, NY, USA: Springer-Verlag New York, Inc., 2002.

[19] N. K. Kasabov, *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*, 1st ed. Cambridge, MA, USA: MIT Press, 1996.

[20] L. C. Jain and N. M. Martin, *Fusion of Neural Networks, Fuzzy Sets, and Genetic Algorithms: Industrial Applications*, 1st ed. Boca Raton, FL, USA: CRC Press, Inc., 1998.

[21] S. Chari, J. R. Rao, and P. Rohatgi, "Template attacks," in *CHES*, 2002, pp. 13–28.

[22] O. Choudary and M. G. Kuhn, "Efficient template attacks," in *CARDIS*, 2013, in print.

[23] E. Brier, C. Clavier, and F. Olivier, "Correlation power analysis with a leakage model," in *CHES*, 2004, pp. 16–29.

[24] M. Bar, H. Drexler, and J. Pulkus, "Improved template attacks," in *COSADE 2010 - First International Workshop on Constructive Side-Channel Analysis and Secure Design*, 2010, pp. 81–89.

ISBN: 978-1-61804-256-9

Mathematical Model for the Home Health Care Routing and Scheduling Problem with Multiple Treatments and Time Windows

Andrés Felipe Torres-Ramos, Edgar Hernán Alfonso-Lizarazo, Lorena Silvana Reyes-Rubiano, Carlos Leonardo Quintero-Araújo

***Abstract*—**Home health care is provided to patients with special conditions in which the assistance is required in their homes. Depending on the pathology, each patient receives specific home care services from specialists, mainly doctors, therapists and nurses. In this context the Home Health Care Routing and Scheduling Problem (HHCRSP) is related with routing and scheduling of the qualified personnel. It integrates the Nurse Rostering Problem (NRP) and the Vehicle Routing Problem (VRP). The HHCRSP considers constraints related with time windows, workload and attention capacity among other limitations associated with patients and staff. Due the cost and quality implications that this kind of services generates in health care companies, this article presents a mixed integer linear programming model for planning the periodic schedule of medical staff and the route planning for to patient visits.

***Keywords*—**Home health care, mixed integer linear programming, qualified staff scheduling, staff routing.

I. Introduction

HOME health care is a service that medical institutions provide to patients who, due to their health conditions, can be treated in their homes, in other cases it's an strategy to increasing the capacity of rooms in hospitals. Taking into account the availability of qualified personnel (doctors, nurses, and therapists), the health sector companies offer a variety of treatments required by patients in which time, cost and quality of the service are crucial; therefore personnel scheduling and the routing of visits has great importance.

The optimization of home health care has long been a field of interest for the operations research, in this context the highlighted areas of study are Home Health Care Routing and Scheduling Problem (HHCRSP), which covers aspects of programming and planning of routes for the medical staff are the Rostering Problem (RP), e.g. Nurse Rostering Problem (NRP), and the planning of routes for visits to patients is considered a Vehicle Routing Problem (VRP). These problems are considered as NP-Hard [1], [2].

In this paper personnel scheduling aims to allocate medical staff members to patients; this scheduling must be performed according to the type of pathology of each patient and the availability of time for patients and staff. In this paper the main treatments studied are related with pathologies of palliative type, chronic care, blood anticoagulation and domiciles for wound care. Depending on the treatment there are three types of specialists who can provides these treatments: doctors, nurses, and therapists. Each patient requires a level of personalized medical care, given that some patients are in severe condition and require fewer intervals between treatments, as opposed to patients presenting better health. This generates a periodic planning of each specialist according to the type of treatment and health conditions of each patient.

In order to schedule and plan the routes of patient visits it is necessary to consider each patient' time windows, which is a time slot in the day which the patient defines or requires the medical care. Another important aspect in the planning of routes is the starting and end point of each staff member's daily, for this paper a multi-depot problem is considered, in which case the home of each specialist (doctors, nurses, and therapists) is the beginning and end of each route, this aspect increase the complexity of the model depending on the number of staff members.

The outline of the article is as follows. Section II presents a literature review for the HHCRSP. The characteristics on staff and patients for home care are presented in section III. A mathematical model for the HHCRSP is presented in section IV. The results of the mathematical model are shown in section V. Conclusions and recommendations for future research are presented in section VI.

II. Literature Review

The home health care routing and scheduling problem (HHCRSP), as mentioned above, make up two problems

This work was supported by the Master in Operations Management and the International School of Economics and Management Sciences (EICEA) of the Universidad de La Sabana, Chía, Colombia.

Andrés Felipe Torres-Ramos is with the International School of Economics and Management Science (EICEA), Universidad de La Sabana, Chía, Colombia (corresponding author to provide phone: 571-8615555 ext.: 25109; e-mail: andrestora@unisabana.edu.co).

Edgar Hernán Alfonso-Lizarazo is with the Engineering Department, Universidad de La Sabana, Chía, Colombia (e-mail: edgar.alfonso@unisabana.edu.co).

Lorena Silvana Reyes-Rubiano is with the International School of Economics and Management Sciences (EICEA), Universidad de La Sabana, Chía, Colombia (e-mail: lorenareru@unisabana.edu.co).

Carlos Leonardo Quintero-Araújo is with the International School of Economics and Management Sciences (EICEA), Universidad de La Sabana, Chía, Colombia (e-mail: carlos.quintero5@unisabana.edu.co).

associated with each operation involved. In terms of staff planning or the allocation of medical staff to patients referred to the Nurse Rostering Problem (NRP) [3], and routes of visits to patients have been considered under different variations of the Vehicle Routing Problem (VRP) [2], [4]. The most studied VRP variation in the HHCRSP is the Vehicle Routing Problem with Time Windows (VRPTW) [1], [5]–[8], which includes the daily time slot that the patient has to receive medical attention. Other variations of the VRP applied to the HHCRSP have been studied independently are the Multi Traveling Salesman Problem with Time Windows (MTSPTW), the Vehicle Routing Problem with Multi-Depot (VRPMD) and the Vehicle Routing Problem with Multi-Period (VRPMP), which intend to characterize multiple staff and multiple points in which the staff start and end each route respectively. [9]–[11].

Different methodologies have been used to solve the HHCRSP within operations research. Within the exact methods is the study of Y. Kergosien, C. Lenté and J-C Billaut [9], which seeks to determine the routes of the medical staff visiting patient's home. In order to do so they determine a whole linear programming model. Another exact method used is the Branch-and-Price algorithm, in the paper [6] the authors use this algorithm to assign staff and determine routes by considering visits per groups of patients. Similarly heuristic methods have been used to solve the HHCRSP, as in the article of D. Mankowska, F. Meisel y C. Bierwirth [10], in which the authors develop a heuristic that determines visits to patients through services interconnected by heterogeneous staff. A. Coppi, P. Detti and J. Raffaelli [12] develop a heuristic based on a local search to determine the personnel planning and routing of visits. Despite the good results that generate the exact and heuristic methods, different authors have used metaheuristics, which allow the problem to be development with more data in a reasonably short time. In the paper [8] the authors present the application of the metaheuristics called Particle Swarm Optimization (PSO) in the programming of the house medical staff. In the paper [13] the authors apply Genetic Algorithm (GA) and Tabu Search (TS) metaheuristics to the delivery of drugs and the collection of biological samples. Simulated Annealing (SA) and Tabu Search (TS) metaheuristics are proposals in the paper [14] to determine the schedule of therapists within the medical treatments of patients home.

On the other hand, the authors have focused their research on various objective functions, the most common is the minimization of the costs of operation, where assignment, overtime and reassignment of staff costs are considered [7], [15], [16], and costs associate to staff and transport [3], [6], [11], [13], [17]. Another objective mainly associated with the routing of the staff is the minimization of time and distance traveled from the operation [8], [10], [18]–[20], which includes minimizing the total journey undertaken by staff to make visits to the patients.

This article focuses on minimizing the total time of operation, as a component of the level of patient satisfaction. Additionally, the study of the scheduling of personal and the planning of routes of patient visits integrating the MTSPTW, the VRPMD and the VRPMP in a single problem: Multi-Traveling Salesman Problem with Time Windows, Multi-Depot and Multi-Period (MTSPTWMDMP).

III. Characteristics of Staff and Patients in Home Health Care

In the most of the revised articles only one type of staff is considered. The home care system studied in this paper, the services can be provided by nurses, doctors and therapists. On the other hand the HHCRSP considers the attention of different services or pathologies of the patients. In this article four types of services are considered (domicile, blood anticoagulation, chronic care, palliative care), and according to the treatment of these pathologies patients require more than one type of staff. The legal and economic aspects related with the working time of the medical staff are considered [5]. A summary of the aspects considered in our model are shown in Figure 1, adapted from Bertels and Fahle [18].

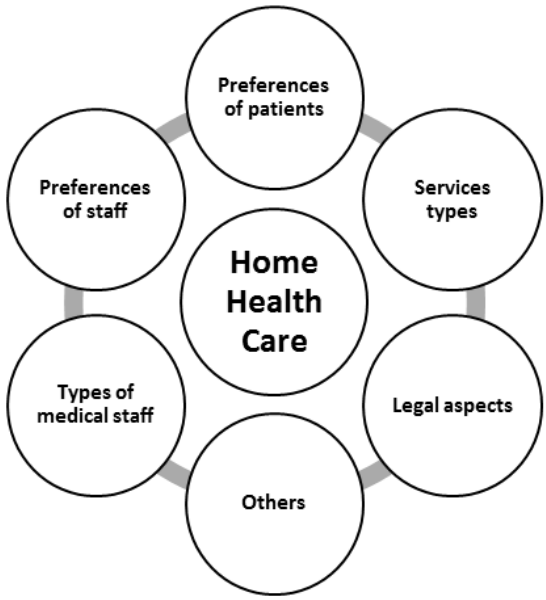


Fig. 1 Characteristics for the HHCRSP

Patients have characteristics that, in addition to the characteristics of the medical staff, delimit the operation; one of the most important is the time window, which represents the time slot that each patient defines or requires for the home visit. There are also features associated with the pathology of each patient. One of them is the demand for personnel, as mentioned above, the type of pathology determines the type of staff required and the frequency between visits, which is dependent on the condition of each patient's health.

IV. Mathematical Formulation for the HHCRSP

This article proposes a mathematical model for the HHCRSP with different types of specialized personnel (doctors, nurses and therapists), which starts and ends every route in their own homes (multi-depot), and is performed in a

horizon of time (multi-period). The problem is defined as a directed graph $G=(V,A)$ with a set $V = CM \cup CE \cup CT \cup PM$ of nodes, which refer to the sets of nodes corresponding to the depot represented by the homes of doctors (*CM*), homes of nurses (*CE*) and homes of therapists (*CT*), and nodes of patients (*PM*). And the set of arcs $A = \{(i, j): i, j \in V, i \neq j\}$.

Every patient $i \in PM$ suffers from a unique pathology, which is classified into four services: domicile, blood anticoagulation, chronic care and palliative care. Each service $s \in S$ is served by the type of personal $p \in P$ required according to the matrix. Additionally, each patient has a demand for visits according to the type of staff and service (DM_i^s, DE_i^s, DT_i^s), these visits are performed within a time horizon in days ($d \in D$) with a periodicity according to the patient and the type of personnel (KM_i, KE_i, KT_i). On the other hand each patient's time window is framed within a length of time per day (e_i^d, l_i^d), in which staff must reach the house of the patient in e_i^d minimum and maximum in l_i^d. As mentioned earlier, each staff member starts and ends its route in their respective home and they have a maximum working time per day *TM*. In addition the travel times (TV_{ij}^p) differ according to the type of personnel, since doctors are mobilized by means of private transport that is faster than public transport by which nurses and therapists are mobilized.

The parameters and decision variables used from modelling the HHCRSP are shown below in Table I.

Table I Notation used from modelling the HHCRSP

Parameters	
TV_{ij}^p	Travel time of personal *p* from the patient *i* to the patient *j*.
TS_i^{ps}	Time of treatment requiring the patient *i* of the personal *p* in service *s*.
DM_i^s	Number of visits required by the patient *i* of doctors in service *s*.
DE_i^s	Number of visits required by the patient *i* of nurses in service *s*.
DT_i^s	Number of visits required by the patient *i* of therapists in service *s*.
SP_p^s	Personal *p* attending the service *s*.
TM	Maximum working time of the day of the staff.
e_i^d	Start time of the time window of patient *i* on the day *d*.
l_i^d	Closing time of the time window of patient *i* on the day *d*.
M	Large number.
N	Index *i* size.
KM_i	Period of time between doctors' visits required by the patient *i*.
KE_i	Period of time between nurses visits required by the patient *i*.
KT_i	Period of time between therapists visits required by the patient *i*.
H	Planning horizon.
Decision variables	
X_{ij}^{pd}	Binary: **1.** If the personal *p* visit the patient *i* and then the patient *j* on the day *d*. **0.** On the contrary.
Y_i^{pd}	Time of arrival of the personal *p* visit the patient *i* on the day *d*.
U_i	Auxiliary variable to avoid subtours to visit each patient *i*.

The proposed mixed integer linear programming model below is to solve the HHCRSP.

Objective function

$$\text{Minimize } Z = \sum_{i\in V}\sum_{j\in V}\sum_{p\in P}\sum_{d\in D}\sum_{s\in S} X_{ij}^{pd}\left(TV_{ij}^{p} + \left(TS_i^{ps} * SP_p^s\right)\right) \quad (1)$$

Subject to

$$\sum_{j\in V} X_{ij}^{pd}\left(\sum_{s\in S} TS_i^{ps}\right) \le TM, \quad \forall i \in V, p \in P, d \in D \quad (2)$$

$$e_i^d \le Y_i^{pd} \le l_i^d, \quad \forall i \in PM, p \in P, d \in D \quad (3)$$

$$Y_i^{pd} + \sum_{s\in S} TS_i^{ps} + TV_{ij}^{p} \le Y_j^{pd} + M\left(1 - X_{ij}^{pd}\right), \quad \forall i, j \in V, i \neq j, p \in P, d \in D \quad (4)$$

$$\sum_{j\in PM} X_{ij}^{pd} = \sum_{j\in PM} X_{ji}^{pd}, \quad \forall i \in PM, p \in P, d \in D \quad (5)$$

$$\sum_{j\in V, j=i} X_{ij}^{pd} = 0, \quad \forall i \in V, p \in P, d \in D \quad (6)$$

$$\sum_{j\in PM} X_{ij}^{pd} \le 1, \quad \forall i \in V / PM, p \in P, d \in D \quad (7)$$

$$\sum_{i\in PM} X_{ij}^{pd} \le 1, \quad \forall j \in V / PM, p \in P, d \in D \quad (8)$$

$$\sum_{j\in V/PM,\, j\neq p} \sum_{p\in P,\, p\neq j} X_{ij}^{pd} = 0, \quad \forall i \in V, d \in D \quad (9)$$

$$\sum_{j\in V,\, j\neq i}\sum_{p\in M}\sum_{f=d}^{d+KM_i} X_{ij}^{pf} \le 1, \quad \forall i \in PM, d \le H - KM_i \quad (10)$$

$$\sum_{j\in V,\, j\neq i}\sum_{p\in E}\sum_{f=d}^{d+KE_i} X_{ij}^{pf} \le 1, \quad \forall i \in PM, d \le H - KE_i \quad (11)$$

$$\sum_{j\in V,\, j\neq i}\sum_{p\in T}\sum_{f=d}^{d+KT_i} X_{ij}^{pf} \le 1, \quad \forall i \in PM, d \le H - KT_i \quad (12)$$

$$\sum_{j\in V}\sum_{d\in D}\sum_{p\in M} X_{ij}^{pd} SP_p^s \ge DM_i^s, \quad \forall i \in PM, s \in S \quad (13)$$

$$\sum_{j\in V}\sum_{d\in D}\sum_{p\in E} X_{ij}^{pd} SP_p^s \ge DE_i^s, \quad \forall i \in PM, s \in S \quad (14)$$

$$\sum_{j\in V}\sum_{d\in D}\sum_{p\in T} X_{ij}^{pd} SP_p^s \ge DT_i^s, \quad \forall i \in PM, s \in S \quad (15)$$

$$\sum_{j\in V}\sum_{p\in M} X_{ij}^{pd} \le 1, \quad \forall i \in PM, d \in D \quad (16)$$

$$\sum_{j\in V}\sum_{p\in E} X_{ij}^{pd} \le 1, \quad \forall i \in PM, d \in D \quad (17)$$

$$\sum_{j\in V}\sum_{p\in T} X_{ij}^{pd} \le 1, \quad \forall i \in PM, d \in D \quad (18)$$

$$U_i - U_j + \left(X_{ij}^{pd} N\right) \le N - 1, \quad \forall i, j \in PM, p \in P, d \in D \quad (19)$$

$$X_{ij}^{pd} \in \{0,1\}, \quad \forall i, j \in V, i \neq j, p \in P, d \in D \quad (20)$$

$$Y_i^{pd} \ge 0, \quad \forall i \in V, p \in P, d \in D \quad (21)$$

$$U_i \ge 0, \quad \forall i \in V \quad (22)$$

The model presents the routing and scheduling of the home medical staff, minimizing the total time of operation (transportation and service) (1). Constraints (2) determines the maximum working load per day for each staff. Constraints (3)

ISBN: 978-1-61804-256-9

and (4) impose the time window per each patient each day according to the staff. Constraints (5) ensures the flow of staff patients every day. Constraints (6) avoid fictitious routes of the staff. Constraints (7), (8) and (9) determine the medical staff every day goes out and returns to their respective home. Constraints (10), (11) and (12) determine the period of time between visits to each patient according to the type of staff and planning horizon. Constraints (13), (14) and (15) guarantee the fulfillment of the demand for visits of each patient according to the type of staff required. Constraints (16), (17) and (18) impose maximum one service with every visit to medical personnel for each patient per day. Constraints (19) eliminates the subtours generated in the programming model. Finally, constraint (20) define the variable X_{ij}^{pd} and constraints (21) and (22) determine non-negativity Y_i^{pd} and U_i variables.

V. RESULTS

As mentioned in section II the model proposed in this article is complex, and is based on a Multi-Traveling Salesman Problem with Time Windows, Multi-Depot and Multi-Period (MTSPTWMDMP). For HHCRSP model validation tests with information from a company's industry in Colombia, test has 16 patients of services: 1. Domicile, 2. Blood anticoagulation, 3. Chronic care and 4. Palliative care. Each service requires attention of different types of staff (doctors, nurses and therapists), according to the service required of one or another personal type as shown in the matrix SP_p^s (see Table II). In addition, Table II shows the number of people for each type of staff, where a total of 19 staff members is determined.

The 16 patients require a total of 101 visits of all medical staff in a two weeks' time horizon, in addition to a periodicity between visits as shown in Table III.

Table II Services handled by each type of personal

Type of staff	Number of specialists	Type of service			
		1	2	3	4
Doctors	3	0	1	1	1
Nurses	8	1	1	1	1
Therapists	8	0	0	1	1
Total	**19**				

Table III Type of service, demand and periodicity of visits required by patients

Patient *i*	Type of service				Periodicity between visits			Demand of visits		
	1	2	3	4	Doctors KM_i	Nurses KE_i	Therapists KT_i	Doctors DM_i^s	Nurses DE_i^s	Therapists DT_i^s
1	1	0	0	0	-	1	-	-	3	-
2	1	0	0	0	-	2	-	-	3	-
3	0	1	0	0	3	1	-	1	3	-
4	0	0	1	0	4	3	4	1	2	2
5	0	0	1	0	4	2	2	2	3	3
6	0	0	1	0	3	1	1	2	3	2
7	0	0	1	0	5	2	1	1	3	2
8	0	0	1	0	4	2	2	2	3	3
9	0	0	1	0	4	1	2	1	3	3
10	0	0	1	0	3	1	2	2	3	2
11	0	0	0	1	3	2	3	2	3	2
12	0	0	0	1	4	3	2	2	2	3
13	0	0	0	1	2	1	2	3	3	3
14	0	0	0	1	4	3	2	1	2	3
15	0	0	0	1	2	1	1	3	3	2
16	0	0	0	1	3	3	2	2	2	2
							Total	**25**	**44**	**32**

The model was implemented using GAMS commercial software version 24.1.3, with a time limit of 4000 seconds in a personal computer Intel(R) Core(TM) i5-4200U CPU with 1.6 GHz with 8 GB of RAM. The solution of the model determines the routes per day needed to meet the demand of visits that patients require. Each route is carried out by a member of the medical staff who begins and ends at home. Table IV shows the routes to perform on day 1, which identifies 3 routes that are performed by the nurse 6, nurse 8 and therapist 5 respectively. For example, the nurse 6 route starts in her home, then visit the patients 6, 5, 8 and 13 in that respective order, and finally part of the last patient (13) to her home as the end point of the route.

The model gives total of 27 routes divided into 11 days (Appendix: Results of the Model of the HHCRSP), routes are carried out by a total of 12 staff members (doctor 1, doctor 2, doctor 3, nurse 2, nurse 6, nurse 8, therapist 1, therapist 3, therapist 4, therapist 5, therapist 6 and therapist 7), as the total number of staff members mentioned above are 19, therefore compliance is evidence with the route with 7 members less, proving the optimization of human resources and the capacity to serve a greater number of patients.

ISBN: 978-1-61804-256-9

Table IV Routs for day 1

Day 1					
Nurse 6		Nurse 8		Therapist 5	
From	To	From	To	From	To
NH6	6	NH8	11	TH5	4
6	5	11	15	4	13
5	8	15	3	13	12
8	13	3	2	12	TH5
13	NH6	2	NH8		

The total operation time of the staff in all the time horizon is 8346 minutes, for which the 64.1% of the time corresponds to the time of travel, and the remaining 35.9% of time corresponds to the time of service.

I. Conclusion

This article proposes a model for the problem of routing and scheduling of medical staff in the home health care systems, which considers characteristics as different types of staff, different services, multi-depot, time windows and multi-period. These features facilitates the application of this model in real conditions related with the home health care services.

The study of the HHCRSP can lead in many directions. First implement heuristics and metaheuristics, which allow the analysis of the problem with more data in less time. On the other hand the integration of other types of services as delivery and pick-up of medicines and biological samples and the emergency services, which involve new constraints and considerations associated with uncertainty in the demand and the availability of staff.

Appendix: Results of the Model of the HHCRSP

The results of the model of the HHCRSP determine a total of 27 routes in 12 days (two weeks) of operation.

Day 1					
Nurse 6		Nurse 8		Therapist 5	
From	To	From	To	From	To
NH6	6	NH8	11	TH5	4
6	5	11	15	4	13
5	8	15	3	13	12
8	13	3	2	12	TH5
13	NH6	2	NH8		

Day 3	
Therapist 3	
From	To
TH3	5
5	TH3

Day 4							
Doctor 3		Nurse 2		Therapist 4		Therapist 6	
From	To	From	To	From	To	From	To
DH3	10	NH2	10	TH4	8	TH6	16
10	6	10	6	8	9	16	14
6	15	6	9	9	13	14	15
15	13	9	15	13	12	15	TH6
13	12	15	1	12	TH4		
12	DH3	1	NH2				

Day 5			
Doctor 1		Nurse 8	
From	To	From	To
DH1	11	NH8	11
11	8	11	5
8	16	5	8
16	7	8	7
7	DH1	7	13
		13	12
		12	NH8

Day 6	
Nurse 2	
From	To
NH2	10
10	9
9	16
16	14
14	4
4	NH2

Day 7							
Doctor 1		Therapist 1		Therapist 5		Therapist 7	
From	To	From	To	From	To	From	To
DH1	5	TH1	7	TH5	11	TH7	10
5	3	7	TH1	11	14	10	6
3	13			14	13	6	5
13	DH1			13	12	5	8
				12	TH5	8	9
						9	TH7

Day 8			
Doctor 1		Nurse 2	
From	To	From	To
DH1	10	NH2	7
10	9	7	3
9	DH1	3	2
		2	1
		1	NH2

ISBN: 978-1-61804-256-9

Day 9	
Doctor 3	
From	To
DH3	11
11	6
6	15
15	DH3

Day 10					
Nurse 2		Therapist 6		Therapist 7	
From	To	From	To	From	To
NH2	10	TH6	8	TH7	10
10	6	8	7	10	6
6	9	7	TH6	6	9
9	15			9	15
15	1			15	TH7
1	NH2				

Day 11			
Nurse 8		Therapist 6	
From	To	From	To
NH8	3	TH6	11
3	2	11	5
2	NH8	5	16
		16	14
		14	4
		4	TH6

Day 12							
Doctor 1		Doctor 2		Nurse 2		Nurse 8	
From	To	From	To	From	To	From	To
DH1	5	DH2	4	NH2	16	NH8	11
5	8	4	12	16	14	11	5
8	16	12	DH2	14	4	5	8
16	14			4	NH2	8	7
14	15					7	13
15	13					13	12
13	DH1					12	NH8

ACKNOWLEDGMENT

The authors thank the sponsorship of this project to the Master in Operations Management and the International School of Economics and Management Sciences (EICEA) of the Universidad de La Sabana.

REFERENCES

[1] S. Nickel, M. Schröder, and J. Steeg, "Mid-term and short-term planning support for home health care services," *Eur. J. Oper. Res.*, vol. 219, no. 3, pp. 574–587, Jun. 2012.

[2] J. Steeg and M. Schröder, "A hybrid approach to solve the periodic home health care problem," *Oper. Res. Proc.*, pp. 297–302, 2007.

[3] W. J. Gutjahr and M. S. Rauner, "An ACO algorithm for a dynamic regional nurse-scheduling problem in Austria," *Comput. Oper. Res.*, vol. 34, no. 3, pp. 642–666, Mar. 2007.

[4] A. Trautsamwieser and P. Hirsch, "Optimization of daily scheduling for home health care services," *J. Appl. Oper. Res.*, vol. 3, no. 3, pp. 124–136, 2011.

[5] E. Cheng and J. Lynn, "A Home Health Care Routing and Scheduling Problem," in *Oakland University, Rice University. Technical Report. USA*, 1998.

[6] M. S. Rasmussen, T. Justesen, A. Dohn, and J. Larsen, "The Home Care Crew Scheduling Problem: Preference-based visit clustering and temporal dependencies," *Eur. J. Oper. Res.*, vol. 219, no. 3, pp. 598–610, Jun. 2012.

[7] P. Eveborn, P. Flisberg, and M. Rönnqvist, "Laps Care—an operational system for staff planning of home care," *Eur. J. Oper. Res.*, vol. 171, no. 3, pp. 962–976, Jun. 2006.

[8] C. Akjiratikarl, P. Yenradee, and P. Drake, "PSO-based algorithm for home care worker scheduling in the UK," *Comput. Ind. Eng.*, vol. 53, no. 4, pp. 559–583, Nov. 2007.

[9] Y. Kergosien, C. Lenté, and J.-C. Billaut, "Home health care problem: An extended multiple traveling salesman problem," in *Proceedings of the 4th Multidisciplinary International Scheduling Conference: Theory and Applications - MISTA*, 2009, pp. 85–92.

[10] D. S. Mankowska, F. Meisel, and C. Bierwirth, "The home health care routing and scheduling problem with interdependent services.," *Health Care Manag. Sci.*, vol. 17, no. 1, pp. 15–30, Jun. 2013.

[11] J. F. Bard, Y. Shao, and H. Wang, "Weekly scheduling models for traveling therapists," *Socioecon. Plann. Sci.*, vol. 47, no. 3, pp. 191–204, Jul. 2013.

[12] A. Coppi, P. Detti, and J. Raffaelli, "A planning and routing model for patient transportation in health care," *Electron. Notes Discret. Math.*, vol. 41, pp. 125–132, Jun. 2013.

[13] R. Liu, X. Xie, V. Augusto, and C. Rodriguez, "Heuristic algorithms for a vehicle routing problem with simultaneous delivery and pickup and time windows in home health care," *Eur. J. Oper. Res.*, vol. 230, no. 3, pp. 475–486, Apr. 2013.

[14] J. D. Griffiths, J. E. Williams, and R. M. Wood, "Scheduling physiotherapy treatment in an inpatient setting," *Oper. Res. Heal. Care*, vol. 1, no. 4, pp. 65–72, Dec. 2012.

[15] E. Lanzarone and A. Matta, "Robust nurse-to-patient assignment in home care services to minimize overtimes under continuity of care," *Oper. Res. Heal. Care*, Jan. 2014.

[16] G. Carello and E. Lanzarone, "A cardinality–constrained robust model for the assignment problem in Home Care services," *Eur. J. Oper. Res.*, Jan. 2014.

[17] P. M. Koeleman, S. Bhulai, and M. van Meersbergen, "Optimal patient and personnel scheduling policies for care-at-home service facilities," *Eur. J. Oper. Res.*, vol. 219, no. 3, pp. 557–563, Jun. 2012.

[18] S. Bertels and T. Fahle, "A hybrid setup for a hybrid scenario: combining heuristics for the home health care problem," *Comput. Oper. Res.*, vol. 33, no. 10, pp. 2866–2890, Oct. 2006.

[19] S. Begur, D. Miller, and J. Weaver, "An integrated spatial DSS for scheduling and routing home-health-care nurses," *Interfaces (Providence).*, vol. 27, no. 4, pp. 35–48, 1997.

[20] E. Alfonso, V. Augusto, and X. Xie, "Mathematical Programming Models for Annual and Weekly Bloodmobile Collection Planning," in *IEEE Transactions on Automation Science and Engineering*, 2014, vol. PP, no. 99, pp. 1–10.

A comparison of random number sequences for image encryption

Antonios S. Andreatos and Apostolos P. Leros

***Abstract*—** The aim of this study is to compare a number of random and pseudorandom number sequences and examine their suitability for image encryption. Five different generators are considered: the pseudorandom generators of three programming languages, a chaotic random number generator, as well as, a truly random number generator. The comparison criteria used are distribution, entropy, statistical tests and encrypted image autocorrelation. Results indicate that all the generators examined provide satisfactory results.

***Keywords*—** Image encryption, random number generator, chaotic number generator, statistical tests, autocorrelation.

I. Introduction

WITH the proliferation of portable devices such as smartphones, tablets and digital cameras, the production as well as the communication of images is an everyday practice. Cyber-crime incidents have raised the issue of confidentiality; hence, data and in particular, image encryption, have gained significant importance today. In recent bibliography several image encryption techniques have been proposed. A common simple encryption practice is to apply the bitwise exclusive OR (XOR) between the image pixels (represented as integers in the range [0, 255]) and random or pseudorandom number sequences [1], [2]. Then the decryption is the XOR between the cipher image and the same random or pseudo random sequence.

It is the purpose of this paper to compare various random and pseudo random number generators for use in image encryption, using the aforementioned XOR method.

A common classification of random number generators based on the source of randomness is the following [1]:

- True Random Number Generators (TRNGs),
- Pseudo-Random Number Generators (PRNGs) and
- Hybrid Random Number Generators (HRNGs).

TRNGs take advantage of unpredictable, nondeterministic sources such as natural processes or physical phenomena which can affect a sensor measuring some physical magnitude and converting the measurement into a sequence of statistically independent data. Physical phenomena commonly exploited in the generation of random numbers are radioactive decay, thermal noise and cosmic microwave background. However, the respective devices are not portable, hence unsuitable for use outside a laboratory. Therefore, good quality random numbers are often obtained by artificial sources such as the rotation of the hard disk in a computer [3], (im)properly connected diodes and transistors [4], noise collected by microphones [5], noise produced by analogue radios [6] and TV sets [7], etc.

Truly random numbers are unpredictable, in the sense that it is impossible to predict the next number, given the previous numbers. For this reason TRNGs are particularly useful in cryptography and especially in key production.

PRNGs are algorithmic generators of numbers which have the appearance of randomness, but nevertheless, their results are predictable. Good random number generators produce very long sequences which look random, in the sense that no efficient algorithm can guess the next number given any prefix of the sequence. Usually, PRNGs use minimal randomness - a randomly chosen initial value called seed. For a specific seed, PRNGs produce a specific, repeatable as well as periodic pattern [8]. This feature is desirable in cryptographic and steganographic telecommunication systems because the pseudorandom sequence used in the transmitter for encryption/ steganography must be faithfully reproduced in the receiver [2].

In practice good TRNGs are hard to find, hard to proof, implemented in hardware, hence often expensive, and, most important for telecom applications, non-reproducible (in the receiver). Therefore, in most applications such as decision making, software testing, simulation and cryptography, PRNGs dominate.

A special category of generators contains the chaotic random number generators (CRNGs) which are based on chaotic phenomena [1]. Physical implementations of chaotic generators (such as those based on electronic circuits) approach TRNGs because real device values have a tolerance and they are also affected by environmental reasons, aging, etc. Software simulations of chaotic phenomena resemble PRNGs, hence they share the same advantages and can be used in cryptography [2] and steganography [9]. The most famous as well as simple chaotic implementation is Chua's circuit [10]. Several cryptographic and steganographic telecommunication systems based on Chua's circuit and its variations have been proposed [1], [2], [9],[11]. Various methods for the production of good quality random number sequences (abbreviated as RNS henceforth) based on Chua's circuit have been proposed [1], [12], [13].

A. S. Andreatos is with the Div. of Computer Engineering & Information Science, Hellenic Air Force Academy, Dekeleia Air Force Base, Dekeleia, Attica, TGA-1010, GREECE (phone: +30-210-819-2360; e-mail: aandreatos.hafa@haf.gr, aandreatos@gmail.com).

A. P. Leros is with the Department of Automation, School of Technological Applications, Technological Educational Institute of Sterea Hellas, 34400 Psachna, Evia, GREECE (e-mail: lerosapostolos@gmail.com).

ISBN: 978-1-61804-256-9

In this paper we compare five sources of random numbers and their suitability in image cryptography. The numbers of these sequences are uniformly distributed, i.e., they have equal probability of appearance.

1/ The PRNG of C (rand() function).

2/ The PRNG of PHP (rand() function).

3/ The PRNG of Matlab (randi function).

4/ A chaotic PRNG based on Chua's circuit, simulated in Matlab, abbreviated henceforth as CRNG [13].

5/ A truly random generator based on the HAVEGE algorithm [14], [15]; see also Haveged manual pages].

The HAVEGE (HArdware Volatile Entropy Gathering and Expansion) algorithm harvests the indirect effects of hardware events on hidden processor state (caches, branch predictors, memory translation tables, etc.) to generate a random sequence. The effects of interrupt service on processor state are perceived as timing variations in program execution speed. Using a branch-rich calculation that fills the processor instruction and data cache, a high resolution timer source such as the processor time stamp counter can generate a random sequence even on an idle system.

The quality of random number sequences (RNSs) is critical in many security applications including cryptography and steganography. Therefore, many kinds of tests have been devised and are used to measure the quality of RN sequences, hence, the corresponding generators. Some of the most common tests are Entropy tests, Statistical Tests, etc. Some of the most common test suites are: the FIPS 140-2 [16], the NIST suite [17], the AIS-31 [18], TestU01 [19], etc.

The tests considered in this paper are the following:

1/ Visual Tests checking distribution;

2/ Entropy tests;

3/ Statistical tests;

4/ Correlation tests.

The FIPS 140-2 Test suite was used for the Statistical Tests. The Federal Information Processing Standard (FIPS) Publication 140-2 (FIPS PUB 140-2) is a U.S. government computer security standard used to accredit cryptographic modules [16]. The official title is Security Requirements for Cryptographic Modules. Initial publication was on May 25, 2001 and was last updated December 3, 2002.

This standard provides increasing, qualitative levels of security intended to cover a wide range of potential applications and environments. The security requirements cover areas related to the secure design and implementation of a cryptographic module.

The test image is a colour jpeg image with dimensions 267x200 pixels and total number of 160,200 bytes or 1,281,600 bits. It is shown in Fig. 1.

Fig. 1 Test image

Poor RNS provide poor (insecure) encryption. In Fig. 2 we can see such a poorly encrypted version of the test image.

Fig. 2 Poorly encrypted image

II. Visual Tests

Visual tests constitute an easy and quick way for humans to check the characteristics of an image and the represented parameter or magnitude.

A. Uniformity Test

Our generators should not be biased, i.e., they should produce random numbers with equal probability, or else, all possible numbers [0—255] should have the same frequency of appearance. This test was performed via a) the Matlab Histogram function; b) via the Binary Viewer software. All generators are producing good results. Figures 3-7 demonstrate the histograms of the number sequences under test, obtained by means of the Binary Viewer.

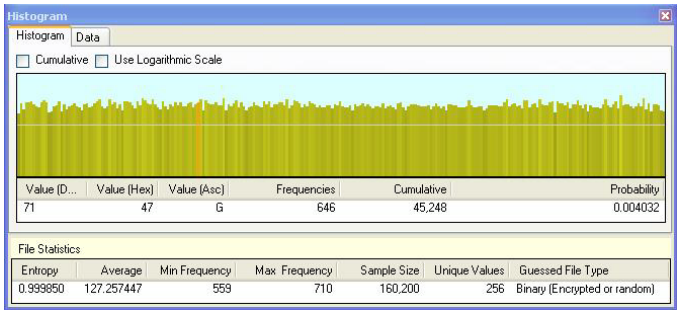


Fig. 3 Histogram of C random numbers

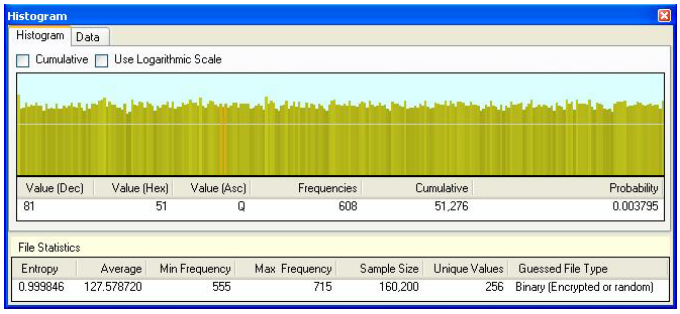


Fig. 4 Histogram of PHP random numbers

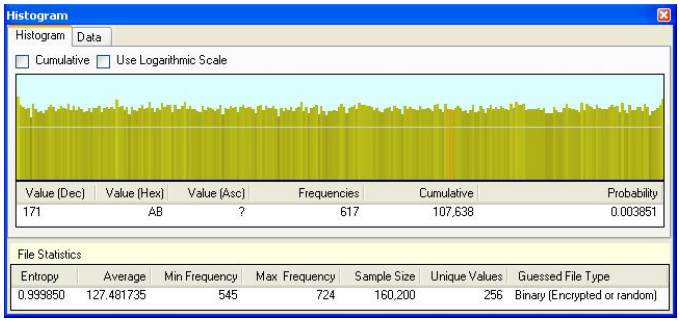


Fig. 5 Histogram of Matlab random numbers

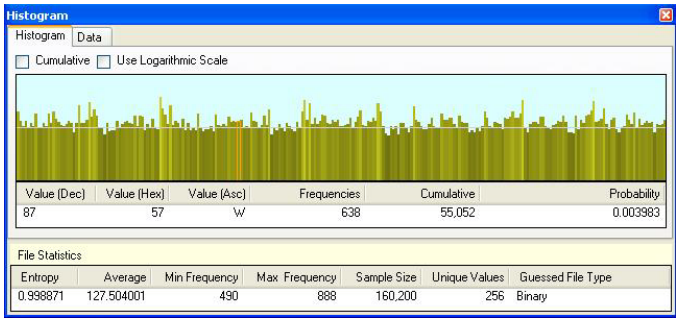


Fig. 6 Histogram of CRNG random numbers

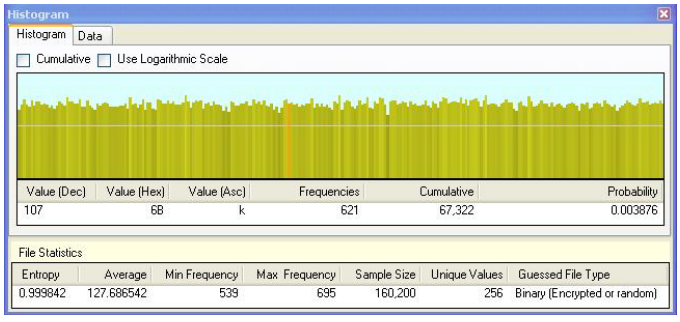


Fig. 7 Histogram of Havege random numbers

Comparing the images we can conclude that all sequences are uniform, but the CRNG produces a coarser histogram. This is also verified by the difference Max Frequency - Min Frequency in the above figures. CRNG random number sequence has the largest difference.

B. Cipher-image inspection

In this test we observe the cipher-images produced by the bitwise XOR between the original (test) image and the random number sequences (Figures 8-12).

Fig. 8 Cipher-images produced by C RNS

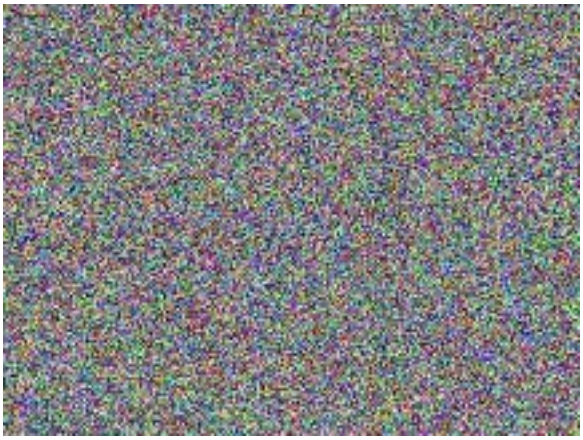


Fig. 9 Cipher-images produced by PHP RNS

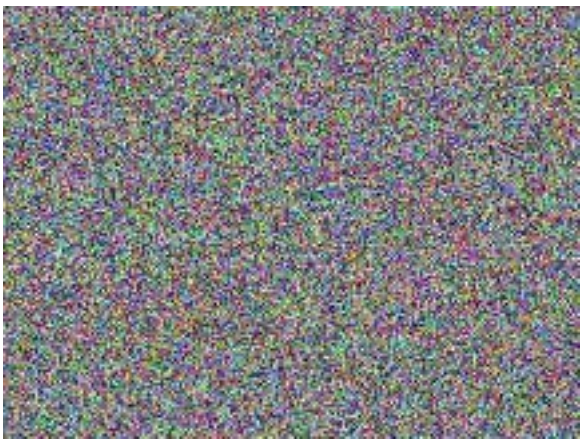


Fig. 10 Cipher-images produced by Matlab RNS

ISBN: 978-1-61804-256-9

Fig. 11 Cipher-images produced by CRNG RNS

Fig. 12 Cipher-images produced by Havege RNS

III. Entropy tests

Entropy is defined as the amount of information contained in a random variable [5]. Entropy of a truly random bit sequence equals its size in bits.

Entropy tests were performed by means of the ***ent*** bash command, as well as, the Binary Viewer software. The higher the Entropy, the better the quality of random numbers. Results as shown in Table 1 below in bits per byte; a result of 8 bits per byte means perfect. In Binary Viewer (BV) the perfect entropy is represented by 1.

Table 1. Entropy tests

	C	PHP	Matlab	CRNG	HAVEGE
Entropy (BV)	0.999850	0.999846	0.999850	0.99871	0.999842
Entropy	7.993240	7.998768	7.998804	7.990964	7.998738
Chi square distribution	266.65	273.95	265.73	2064.94	279.93
Arithmetic mean	127.2574	127.5787	127.4817	127.5040	127.6865
Monte Carlo value for Pi	3.152359551	3.151760300	3.149213483	3.135730337	3.13378277
error %	0.34	0.32	0.24	0.19	0.25
Serial correlation coefficient	-0.003169	0.001998	-0.000521	0.000168	0.002261

IV. Statistical Tests

The results of statistical tests, obtained by means of the *rngtest* bash shell script, are presented in Table 2. Surprisingly enough, the C data set passes all tests, surpassing all other generators.

Table 2. Statistical tests

	C	PHP	Matlab	CRNG	HAVEGE
Successes	64	62	59	59	63
Failures	0	2	5	5	1
Monobit	0	0	0	0	0
Poker	0	0	0	2	0
Runs	0	0	0	0	1
Long run	0	2	5	3	0
Continuous run	0	0	0	0	0

V. Autocorrelation Tests

In this section we compare the autocorrelation of the encrypted images. The autocorrelation of an image is defined here as the similarity of an image with itself, shifted by one pixel horizontally, vertically, diagonally and anti-diagonally. The autocorrelation was calculated in Matlab using the following formulae (1) and (2) [2]. The correlation coefficient γ for a pair of pixels is defined as described in formula (1):

$$\gamma(x,y) = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y} \tag{1}$$

Where:

$$\operatorname{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{N}\left[x_i - E(x)\right]\left[y_i - E(y)\right] \tag{2}$$

In eq. (2) "E" is the expected value operator and σ_x^2 represents the variance of variable x. The values of γ(x,y) lie in the range [−1, 1], with 1 indicating perfect correlation, −1 indicating perfect anti-correlation and 0 indicating no correlation. The evaluation of these formulae in MATLAB was realised by the use of built-in functions.

Table 3. Autocorrelation tests

	Horizontal	Vertical	Diagonal	Anti-diagonal
Original	0.9758	0.9527	0.9278	0.9561
C	0.0025	-0.0028	0.0022	-0.0037
PHP	0.0002	-0.0007	-0.0003	0.0044
Matlab	0.0032	0.0010	0.0031	0.0029
CRNG	0.0035	0.0016	-0.0047	0.0003
HAVEGE	0.0037	-0.0021	-0.0008	-0.0020

ISBN: 978-1-61804-256-9

As expected, the autocorrelation of the original image is close to 1. However, the autocorrelation of all the encrypted images is close to 0. Hence, all generators produce satisfactory results.

VI. Discussion

One might possibly have expected that the Havege TRNG would have the best results by far; however, all other generators produce comparable results, which means that the specific generators are of good quality. In fact, PHP and Matlab use the Mersenne twister generator which is, by far, the best and most widely used PRNG [20]. Mersenne twister is used by many programming languages, including PHP and Matlab; hence the good results. The commonly used version of Mersenne Twister is "MT19937", which has a very long period of 2^19937−1.

In order to check the periodicity of a RNS we have to arrange its numbers (in our case, integers from 0 to 255) as pixels in a matrix. The longer the PRNG period, the longer the RNS in order to observe patterns. This is the reason for not observing patterns in the visual tests performed in the RNS under test, since their size was only 1,281,600 bits. However, if we generate a large RNS, patterns will appear. Fig. 13 shows a visualisation of a PHP pseudo-random sequence of about 40 million numbers, where patterns can be observed.

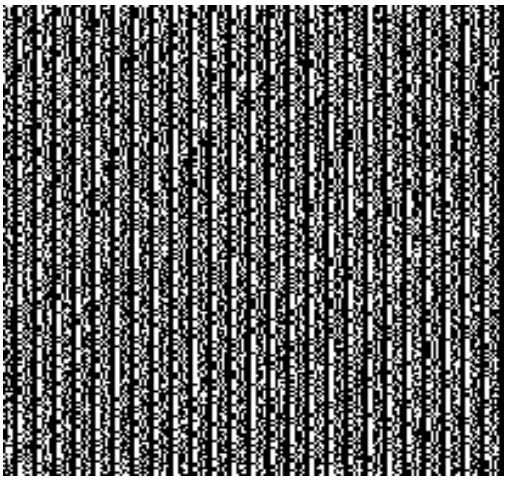


Fig. 13 Patterns in visualisations of pseudo-random sequences indicate periodicity (detail)

Additional tests that could have been performed include the histograms of the cipher-images [2], [21], unpredictability tests, scatter plots [22], the Diehard [23, 24] and the Dieharder [25] test suites, correlation tests [21], etc. This is left for future work.

VII. Conclusion

In this paper we have compared three pseudo-random number sets, a chaotic number set and a truly random number set, for use in image encryption. From the comparison it seems that all sources produce acceptable, high quality results for the selected test image. However, neither pseudo- nor truly random numbers are suitable for cryptography, for different reasons each.

- Pseudo-random numbers are in fact periodic sequences, hence, easy to guess/ break [8]. The reason is that observing a sufficient number of past iterations allows us to predict all future iterations [20]. Hence, they are not suitable for cryptography.
- Truly random numbers are not reproducible; hence, it will be impossible to decode the cipher in the receiver.

The CRNG produces satisfactory results, comparable with the PRNGs and the TRNG; moreover, it has some additional advantages:

- They are periodic but by applying spatiotemporal techniques, their period may become some years long, hence practically unreachable [26];
- Under proper design of the generator, they can use multi-parametric keys rather than a single seed (see for instance [2], [9];
- The CRNG is multi-parametric system and must be fine-tuned in order to improve its performance. A method for tuning Chua-based CRNGs has been proposed in [27].

References

[1] C. K. Volos, "Image Encryption scheme based on coupled Chaotic systems", in *JAMB*, 3, 1, 2013, pp. 123-149.

[2] A. Andreatos and A. Leros, "Secure image encryption based on a Chua chaotic noise generator. Journal of Engineering Science and Technology Review". *JESTR Special Issue on Nonlinear Circuits: Theory and Applications,* 6 (4) (2013) 90-103. Available: http://jestr.org/index.php?option=com_content&view=article&id=31&Itemid=71.

[3] D. Davis, R. Ihaka and P. Fenstermacher, "Cryptographic randomness from air turbulence in disk drives", Proc. CRYPTO' 94, pp. 114 -120. Available: theworld.com/~dtd/random/forward.ps.

[4] Leon's Mini Random Number Generator [Online]. Available: http://www.physics.wisc.edu/~lmaurer/projects/minirng/minirng.html].

[5] S. Theodoulou, "*Improving the Reliability of Cryptographic Services in Personal Computers using Truly Random Numbers and Advanced Coding Techniques*", Diploma Thesis, Hellenic Air Force Academy, June 2010 (in Greek).

[6] "It did not work, it just happened", *Delta Haker magazine*, no. 33, pp. 68-82, June 2014 (in Greek).

[7] "Randomness generators", *Delta Haker magazine*, no. 31, pp. 62-68, Apr. 2014 (in Greek).

[8] A.N. Veneti, G.C. Meletiou and M.N. Vrahatis, "Fractal Dimension As An Assessment Metric for Pseudorandom Number Generators". Presented at the *2nd International Conference on Cryptography, Network Security and Applications in the Armed Forces*. Hellenic Military Academy, April 2, 2014.

[9] Apostolos P. Leros and Antonios S. Andreatos, "A Video Steganography System for Secure Data Communication", Chapter 4.3 in *New Research Trends in Nonlinear Circuits: Design, Chaotic Phenomena and Applications*. Nova Science Publishers, Inc. 2014.

[10] L. O. Chua, "Chua's circuit: ten years later", *IEICE Trans. Fundamentals* E77-A, 11, pp. 1811-1822 (1994).

[11] L. Gámez-Guzmán, C. Cruz-Hernández, R.M. López-Gutiérrez, E.E. García-Guerrero, Synchronization of Chua's circuits with multi-scroll

attractors: Application to communication, *Commun. Nonlinear Sci. Numer. Simulat*. 14, 2765–2775 (2009).
[12] M. E. Yalçın, J. A. K. Suykens and J.Vandewalle, "True Random Bit Generation From a Double-Scroll Attractor", *IEEE Transactions on Circuits and Systems* —I: Regular Papers, 51, 7, pp. 1395-1404 (2004).
[13] A. S. Andreatos and C. K. Volos, "Secure Text Encryption Based on Hardware Chaotic Noise Generator". Presented at the *2nd International Conference on Cryptography, Network Security and Applications in the* Armed Forces, Hellenic Military Academy, April 2, 2014.
[14] A. Seznec and N. Sendrier, "HArdware Volatile Entropy Gathering and Expansion: generating unpredictable random numbers at user level", INRIA Research Report, RR-4592, October 2002.
[15] Haveged - A simple entropy daemon. Available: http://www.issihosts.com/haveged/index.html;
[16] *Security Requirements For Cryptographic Modules*, 2001. Available: http://csrc.nist.gov/publications/fips/fips140-2/fips1402.pdf.
[17] NIST Special Publication 800-22, "A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications". Available: http://csrc.nist.gov/publications/nistpubs/800-22-rev1a/SP800-22rev1a.pdf.
[18] AIS 31, "Functionality Classes and Evaluation Methodology for Physical Random Number Generators", Version 1 (25.09.2001). Available: www.bsi.bund.de/zertifiz/zert/interpr/ais31e.pdf.
[19] TestU01, http://www.iro.umontreal.ca/~simardr/testu01/tu01.html.
[20] Mersenne twister, http://en.wikipedia.org/wiki/Mersenne_twister.
[21] A. Belmeguenaï, K. Mansouri and L. Grouche, "Implementation of Blum Blum Shub Generator for Message Encryption", in *Proceedings of International Conference on Control, Engineering & Information Technology* (CEIT'14), pp. 118-123.
[22] Divyanjali, Ankur and Vikas Pareek, "An Overview of Cryptographically Secure Pseudorandom Number generators and BBS", *International Journal of Computer Applications* (IJCA) (0975 – 8887), pp. 19-28, 2014.
[23] G. Marsaglia, "The Marsaglia random number CDROM including the diehard battery of tests of randomness". Available: http://stat.fsu.edu/pub/diehard, 1996.
[24] Diehard tests, Available: http://en.wikipedia.org/wiki/Diehard_tests
[25] R. G. Brown. "dieharder: A Random Number Test Suite", 2007. Available: http://www.phy.duke.edu/~rgb/General/dieharder.php.
[26] S. Wang, W. Liu, H. Lu, J. Kuang and G. Hu, "Periodicity of chaotic trajectories in realizations of finite computer precisions and its implication in chaos communications", *International Journal of Modern Physics B*, Vol. 18, Issue 17, no. 19, 30 July 2004, pp. 2617-2622.
[27] A. Leros and A. Andreatos, "On the optimisation of Chua chaotic attractors", in *Proceedings of ICACM '13, 2nd WSEAS International Conference on Applied and Computational Mathematics*, Vouliagmeni, Athens, Greece, May 14-16, 2013. Available: http://www.wseas.org/wseas/cms.action?id=2574.

Modeling RIP using Event-B

Bahija Boulamaat. Anas Amamou, Rajaa Filali, Sanae El mimouni, Mohamed Bouhdadi

***Abstract*—**Routing Information Protocol (RIP) is a standard based, distance vector, interior gateway protocol used by routers to exchange routing information among gateways and other hosts. It plays an important role providing the shortest and best path for data to take from node to node. it permits to update periodically the routing information in the RIP network. In this paper, we will use the formal method Event-B to model and prove the Routing information protocol, taking in it step by step by going from an abstract model then refined it and prove the consistency between them. The prover in Event-B assures the correctness of the modeling.

***Keywords*—**Event B - Formal modeling - protocol - refinement.

I. Introduction

In this paper, we will model the RIP using the Event-B method. The first version of the routing information protocol is defined by IETF in RFC 1058,it is considered a classful routing protocol .due to the many limits of the version 1,the version 2 developed in RFC 2453, it is considered a classless routing protocol, and has the ability to carry subnet information. The third version if RIP called RIPng (RIP next generation) is an extension of the RIPv2 to support IPv6.
The informal specification of RIP will be modeled using the formal method Event-B.

The Event B is a formal method for modeling software systems [1]. It has been used in several interdisciplinary such as medicine, transport, aeronautics, trains, space, biology, security, hybrid system, parallel systems, communications protocols. It translates an informal specification to a formal notation using mathematical language (elementary set theory, first order logic...). It goes from developing a discrete system by refinement. This method permits to build a model by successive steps going from an abstract model to a more concrete one. Each version is proved and is consistent with the previous one [2].

Based on back and Dijkstra works, the method Event-B has been developed by Jean Raymond ABRIAL who has developed the B method and the Z language. It works essentially on refinement, composition and generecity [3]. The advantage of the Event -B is to make proofs automatically using the Rodin platform.

The reminder of this paper is as follows. In Section2, we give an overview of the Event-B method. In Section 3, we present the description of the routing information protocol. And, the final section presents the modeling of the protocol).

II. Overview of the Event-B method

Event B is a formal method used to model complex systems. It uses mathematical language to built models step by step going from an abstract one to a refined one and proving the correctness of it. The Event-B models consist of two mean constructs: the contexts and the machines. The context contains the static part of the model like sets and constants, and axioms whereas the machines contain the dynamic part like variables, invariants and events. Between the machines and contexts, there are different relationships. The machines can refine one or several ones. The contexts can be extended by one or several context and can be referenced ‘’see’ by one or several machines.

In Event-B, an event is defined by the syntax: EVENT e WHEN G THEN S END , Where G is the guard, expressed as a first-order logical formula in the state variables, and S is any number of generalized substitutions, defined by the syntax S ::= x := E(v) | x := z : | P(z). The deterministic substitution, x := E(v), assigns to variable x the value of expression E(v), defined over set of state variables v. In a non-deterministic substitution, x := z : | P(z), it is possible to choose non-deterministically local variables, z, that will render the predicate P(z) true. If this is the case, then the substitution,
x := z, can be applied, otherwise nothing happens.

The Event-B consists of three important techniques: refinement, composition, instantiation.

Refinement: the refinement permits to build model gradually by making it more and more accurate. We construct the models by sequence; each one is the refinement of the previous one in the sequence. The refinement uses the concept of the superstition refinement and data refinement.

Decomposition: the decomposition consists on spitting a model to small sub models.

Generic instantiation: the generic instantiation permit to parameterize machines in order to reuse it to refinement. It consists of using a generate theory proved using constants and axioms in a machines to be reused in another machine without proving it.

The proof obligation permits to test and validate the model. The proof obligation rules define what must be. They verify the properties of the machines and ensure the correctness of the modeling and its consistency between the refined and the abstract levels. The proof obligation rules define what must be proven; verify certain properties of the machines. Rodin platform generates automatically this proof with the help of the proof obligation generator. There several proof obligation as INV, FIS, WD...
Invariant preservation proof obligation rule (INV) ensure that each invariant are preserved by each event, the Feasibility proof obligation rule (FIS) ensure that the action are feasible.

The well-definedness proof obligation rule (WD) ensures that a potentially ill-defined axiom, theorem, invariant, guard, action, variant, or witness is indeed well defined

III. Description of RIP

Routing Information Protocol (RIP) is a standard distance-vector protocol used by routers to exchange routing information. it is used to find the best route or path from end-to-end (source to destination) over a network by using a routing metric/hop count algorithm. This protocol is used to determine the shortest path from the source to destination. Hop count is the number of nodes the packet must go through until it reaches its destination [6] - [7] - [8], the routing information is stored in a routing table for future use.

The routing information protocol manage the router information by enable to exchange routing information in network (RFC 1058). RIP allows connecting with destinations that is not directly reachable. It used for a finite number of nodes the longest path supported is 15.if the metric's (the metric is the number of node from the originating node to reach a destination) node value is 16 the destination is considered unreachable. Each node in the network should know all the routes to the other nodes .the routing table is updated periodically to ensure the freshness of the routes.

A node sends a broadcast request to RIP neighbors' interfaces in the network, the request consist of its entire routing table. All the neighbors receiving the request respond with their routing tables, and these tables will also be sent to the receiving neighbors.

The node send a request to its neighbors, each node of the neighbors will send its routing tables to its neighbors until all the nodes in the network have all the routes in their tables.

The messages received, permit to update the receiving 'node table. Each entry in the routing table presents a route, it consist of source, destination and the metric node to reach the destination. The update consist of comparing the received table with the table of the node ,If an entry is in the received node routing table but it does not exist in the node routing table than we add it as a new one in the node table. If it already exists we take the minimum metric, as it considered the best path to the destination.

The routing information table has four important timers, which used in the routing information protocol where:

The Route update timer: each 30 second a router send a copy of its routing table to its neighbors.

Route invalid timer: a time to determine when a route is considered invalid, when there is no update for a route about 180 s then the router sends updates to its neighbors that the route is invalid.

Route hold-down timer: is the time in which a route in unreachable, is about 180s or until a better route is found

Route flush timer: is the time to remote a route from the routing table, it comes after the route is considered invalid

The limits specifications of this protocol are:

IV. Model in event-B

In the initial model we present the exchange between a node and its neighbors. We take the node and one neighbors because the same pattern is repeated with the other neighbors. The source node send messages to its neighbor this messages consist of the source node routing table, and received messages from the neighbors, also consist of the neighbors routing table. in first model we also present the update of the source node table as how it updated. The update happens when we compare the source node routing table and the received node routing table. If there are routes in the received table that are not in the source table we then add those routes as new entry. If the same routes that exist in both tables we consider the route with best path (with the minimum metric)

A. The initial model

We consider the message exchange between the nodes. we model it as an exchange between one node in the RIP network we called the source node and one of its neighbors we called it a neighbor node

In the context:

We define two carrier sets: NODE and MSG.

In the machine:

Sd,and rcv present the messages send or received by a node in the RIP network, we also define the routing table by the variable entrytable, and hopecount who presents the routing table of the resource node (rcvnode)and respectively entrytablercv, and hopecountrcv presents the routing table of the neighbors node (neibornode),

CONTEXT ctx0
SETS NODE,MSG
END

VARIABLES
Sd,rcv,srcnode, neibornode,
entrytable, entrytablercv,
hopecount, hopecountrcv
INVARIANTS
sd ⊆ MSG
rcv ⊆ MSG
srcnode ∈ MSG⇸ NODE
neibornode ∈ MSG⇸ NODE
entrytable ∈ NODE↔NODE
entrytablercv ∈ NODE↔NODE
hopecount ∈ NODE⇸(NODE⇸ℕ)
hopecountrcv∈ NODE⇸(NODE⇸ℕ)
∀i,j· i↦j∈entrytable⇒i≠j
∀i,j· i↦j∈entrytablercv⇒i≠j

We have two events for the exchange messages between a source node and one of it neighbors node. And, three events for the update of the source node routing table.

The event send_request and receiv−request present the messages send or received by a node:

```
send−request
ANY
   Msg,s,n
WHERE
   msg ∈ MSG
   s∈NODE
   n∈NODE
   msg∈dom(srcnode)
   msg∈dom(neibornode)
   srcnode(msg)=s
   neibornode(msg)=n
   s ≠ n
THEN
   sd ≔ sd ∪ {msg}
END
```

```
receiv−request
ANY
   Msg,s,n
WHERE
   msg ∈ MSG,
   s∈NODE,
   n∈NODE
   msg∈dom(srcnode)
   msg∈dom(neibornode)
   srcnode(msg)=s,
   neibornode(msg)=n
   s ≠ n
THEN
   rcv ≔ rcv ∪ {msg}
END
```

After we compare the routes in the route table of the source node and the one received from the neighbor node. We update the source routing table when the route in the received table does not exist in the resource routing table (event updare_table−dif) .or when the route exists in the routing table but the received one has a better path because it has an inferior count metric (update_table_same_entry1). Finally, there is no update when the route in the source routing table has an inferior metric node than the one in the received one, in this case there is no action so it's a SKIP, nothing changing in the resource routing table.

```
update_table−dif
ANY
   s,n,msg1,msg2
WHERE
   msg1 ∈ MSG,msg2∈MSG,              \
   s∈NODE,n∈NODE
   s≠n
   msg1∈dom(srcnode),
   msg2∈dom(neibornode)
   srcnode(msg1)=s,
   neibornode(msg2)=n
   s↦n ∉ entrytable,n↦s ∈ entrytablercv
THEN
   entrytable≔ entrytable ∪ {s↦n}
END
```

```
update_table_same_entry1
ANY
   s,n,msg1,msg2
WHERE
   msg1 ∈ MSG,
   msg2∈MSG
   s∈NODE,n∈NODE
   msg1∈dom(srcnode)
   msg2∈dom(neibornode)
   srcnode(msg1)=s
   neibornode(msg2)=n
   n≠s
   n↦s∈entrytablercv,  s↦n∈entrytable
   s∈dom(hopecount)
   n∈dom(hopecount(s))
   n∈dom(hopecountrcv)
   s∈dom(hopecountrcv(n))
   hopecount(s)(n)>hopecountrcv(n)(s)
THEN
   entrytable ≔ entrytable ∪ {n↦s}
END
```

```
update_table_same_entry2
ANY
   n,msg1,msg2,s
WHERE
   msg1 ∈ MSG
   msg2∈MSG
   s∈NODE,n∈NODE
   msg1∈dom(srcnode)
   n≠s
   msg2∈dom(neibornode)
   srcnode(msg1)=s
   neibornode(msg2)=n
   n↦s∈entrytablercv
   s↦n∈entrytable
   s∈dom(hopecount)
   n∈dom(hopecount(s))
   n∈dom(hopecountrcv)
   s∈dom(hopecountrcv(n))
   hopecount(s)(n)<hopecountrcv(n)(s)
END
```

B. The first refinement

In the first refinement, we add three more events in relation with four timers. And we refine the previous event by adding more guard.

Route update (t1), time route invalid (t2), time, and route flush time (t3), time route hold-down (t4).

The initial context will be extended and the initial machine will be refined.

Firstly we extend the initial context, by defining the states of the route if it is valid or unreachable (the metric is above 15) or invalid and adding the timers.

CONTEXT
ctx01
EXTENDS
ctx0
SETS
ETAT
CONSTANTS
Valid,invalid ,unreachable
t1,t2,t3,t4
AXIOMS
ETAT={valid,invalid,unreachable}
valid ≠ unreachable
unreachable≠invalid
t1 ∈ ℕ ,t2 ∈ ℕ,t3∈ℕ ,t4∈ℕ
END

In this refinement we have two more variable a time temps and a Boolean for the first timer.

VARIABLES
b,temps,etat

INVARIANTS
temps ∈ ℕ, etat ∈ ETAT
b ∈ BOOL = TRUE ⇒ temps < t1

We refine the events send−request by adding two more guards b = TRUE, and etat= valid .
And for the events: update_table_same_entry1, update_table_same_entry2 , updare_table−dif we add the guard etat=valid, the route should be valid so it can be updated.
We add three more event route−invalid, route−remove, route−holddown .the first added event presents when the route is considered invalid. In the second event the invalid routes are removed after a flush time. The third event presents when route is considered unreachable

route−invalid
. ANY
s,n
WHERE
s ∈ NODE
n ∈ NODE
temps∈ℕ
temps≥t2
s ≠ n not
etat=valid
s↦n ∈ entrytable
THEN
etat:=invalid
END

route−remove
ANY
s,n
WHERE
s ∈ NODE
n ∈ NODE
s ≠ n
etat=invalid
temps≥t3
s↦n ∈ entrytable
THEN
entrytable:=entrytable∪(entrytable\{s↦n})
END

route−holddown
ANY
s,n
WHERE
s ∈ NODE
n ∈ NODE
s ≠ n
temps≥t4
s↦n ∈ entrytable
s∈dom(hopecount)
n∈dom(hopecount(s))
hopecount(s)(n)≥16
THEN
act1: etat:=unreachable ›
END

V.Conclusion

In this paper, we model the routing information protocol with the event-B method, going from an initial model and refine it. The Event -B method assure the correctness of the models can prove it using the proof obligation rules in the platform Rodin.

the proof obligation rules who have been failed ,are fixed by adding new invariants or strengthen the guards in the events. In the end we can see that our model in correct and proved using Event-B.

REFERENCES

[1] Jean-Raymond Abrial,Modeling in Event-B - System and Software Engineering. Cambridge University Press 2010, ISBN 978-0-521-89556-9.
[2] Edsger W,Dijkstra, Carel S. Scholten, Predicate calculus and program semantics. Texts and monographs in computer science, Springer 1990, ISBN 978-3-540-96957-0, pp. I-X, 1-220
[3] Jean Raymond Abrial, Stefan Hallerstede, Refinement, Decomposition, and Instantiation of Discrete Models: Application to Event-B Fundamenta Informaticae,Vol.77, No.1-2, 2002, pp. 1-28.
[4] Jean-Raymond Abrial, Michael J. Butler, Stefan Hallerstede, Thai Son Hoang, Farhad Mehta, Laurent Voisin: Rodin: an open toolset for modelling and reasoning in Event-B. STTT, Vol 12, NO.6, 2010, pp. 447-466.
[5] http: //www.event-b.org Rodin Platform

[6] C. Hendrik, RFC 1058, Routing Information Protocol, the Internet Society, June 1988
[7] G. Malkin, RFC 1388, RIP Version 2 - Carrying Additional Information, The Internet Society ,January 1993
[8] G. Malkin ,RFC 2453, RIP Version 2, , The Internet Society ,November 1998

Mathematical Model of Cervical Cancer due to Human Papillomavirus Infection

P. Pongsumpun

***Abstract*—**Cervical cancer is usually found in women from teen-age to older women. Around the world, the cervical cancer cases are increased to 500,000 and 200,000 deaths from this disease per year. About 80% of cervical cases live in developing countries. In Thailand, cervical cancer is the top ten cause of death in Thai women. About 99% of all cervical cancer cases are related to Human Papillomavirus (HPV). In this paper, we formulate the mathematical model to describe the transmission of cervical cancer in women with HPV Infection. The standard dynamical modeling method is used in this study. The numerical results are presented.

***Keywords*—**cervical cancer, local stability, mathematical model , steady state.

I. Introduction

THE body of each person is composed of several living cells. Normal body cells grow, separate into new cells. Early years of each person's life, normal cells divide faster to allow the person to grow. After each person becomes an adult, most cells divide only to replace dying cells or to repair damage cells. Cancer begins when cells in a part of the body start to grow without control. There are many kinds of cancer. They all start because of abnormal growth cells [1]. Cancer is usually named for the part of the body where it starts, even if it spreads to other parts later. The behaviors of different kinds of cancers are difference. The development of each kind of cancer grows at different rates. Cervical cancer cases are occurred around the world. In United Kingdom, there are approximately 2,800 cases of cervical cancer. In each year, there are 1,000 women die from cervical cancer. There were 282 new cases found in 2004 and 127 deaths from this disease in 2005. The survival rate of five-year in Scotland between 1997 and 2001 was 70.6%[2]. In 2008, about 1,300 women were diagnosed with cervical cancer and 380 deaths in Canada. Most cervical cancers are related to Human Papillomavirus (HPV) infection. There are more than 100 different kinds of HPV, several of them are harmless. Nevertheless, some types of HPV can interrupt the normal functioning of the cells of the cervix. HPV consists of many types such as HPV-16, HPV-18, HPV-31, HPV-35, HPV-39, HPV-45, HPV-51, HPV-52, HPV-58. There are about 70% of all cervical cancers are caused by HPV-16 and HPV-18. The symptoms of cervical cancer cases are abnormal vaginal bleeding, pain during sexual intercourse, increased amount of discharge from the vagina and foul-smelling discharge from vagina. Cervical cancer is an abnormal kind of cancer that develops in woman's cervix. It is the entrance to the womb from vagina[3]. The cervix joints the uterus to the vagina. Blood flows from the uterus through
the cervix into the vagina during a menstrual period. Cervical cancer may sometimes be a threat to life. It can attack nearby tissues and organs. It can extend to other parts of the body.
Cervical cancer cells can spread by breaking away from the cervical tumor. They can transmit through lymph vessels to nearby lymph nodes. The spread of cancer cells occurred through the blood vessels to the lungs, liver, or bones [4]. Cervical cancer cases in Thailand are occurred in every year as shown in fig.1. We will see that most cases are occurred in Central region of Thailand.

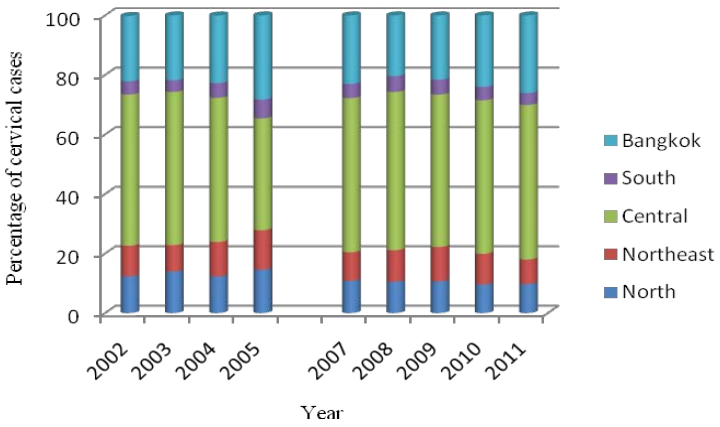


Fig. 1 The data of Thai cervical cases by region of Thailand[5-6]*. * There is no data in year 2006.

In 2010, Muller and Bauch [7] studied the relations of sexual partnerships and the tranmission of Human Papillomavirus(HPV). In 2012, Lee and Tameru [8] constructed the model of Human Papillomavirus(HPV) developed to cervical cancer of African American women in the United States. They gave the method to prevent this disease thought the ideal of modeling. In this study, we construct the mathematical model of cervical cancer in Thailand with the influence of HPV infection.

P. Pongsumpun is with the Department of Mathematics, Faculty of Science, King Mongkut's institute of Technology Ladkrabang,Chalongkrung road, Ladkrabang, Bangkok 10520, Thailand (corresponding author to provide phone: 662-3298400 ext. 320; fax: 02-3298400 ext. 284; e-mail: kppuntan@ kmitl.ac.th).

II. Mathematical Model

The diagram of our dynamical equations can be described by following figure:

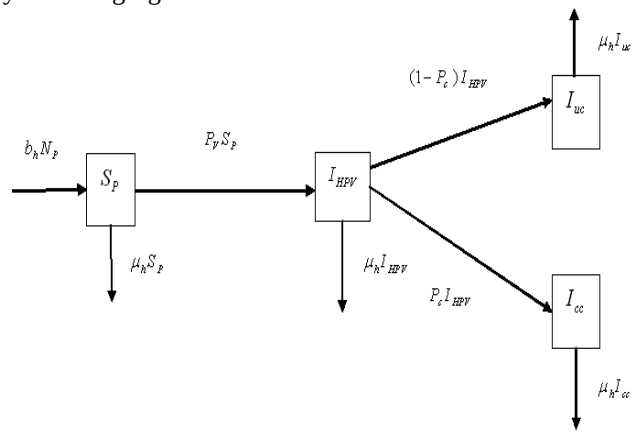


Fig. 2 Diagram of our equations.

The dynamical equations of this model is given by the following systems of differential equations:

$$\frac{dS_p}{dt} = b_h N_p - (P_v + \mu_h)S_p \quad (1)$$

$$\frac{dI_{HPV}}{dt} = P_v S_p - (1-P_c)I_{HPV} - (P_c + \mu_h)I_{HPV} \quad (2)$$

$$\frac{dI_{uc}}{dt} = (1-P_c)I_{HPV} - \mu_h I_{uc} \quad (3)$$

$$\frac{dI_{cc}}{dt} = P_c I_{HPV} - \mu_h I_{cc} \quad (4)$$

with a condition $S_p + I_{HPV} + I_{uc} + I_{cc} = N_p$;

where the variables and parameters in the above equations are given by

S_p is the number of susceptible women,

I_{HPV} is the number of infected women with HPV,

I_{cc} is the number of Infectious HPV women population who be infected with cervical cancer,

I_{uc} is the number of Infectious HPV women population who be uninfected with cervical cancer,

b_h is the birth rate of human population,

μ_h is the death rate of human population,

N_p is the number of women,

P_v is the probability of women who be infected with HPV,

P_c is the probability of women with HPV can be infected with cervical cancer,

We normalize our equations by letting

$$s_p = \frac{S_p}{N_p}, i_{hpv} = \frac{I_{HPV}}{N_p}, i_{uc} = \frac{I_{uc}}{N_p}, i_{cc} = \frac{I_{cc}}{N_p}, \text{ then}$$

the reduced equations become

$$\frac{ds_p}{dt} = \mu_h(1-s_p) - P_v s_p \quad (5)$$

$$\frac{di_{hpv}}{dt} = P_v s_p - i_{hpv}(1+\mu_h) \quad (6)$$

$$\frac{di_{uc}}{dt} = (1-P_c)i_{hpv} - \mu_h i_{uc} \quad (7)$$

where $s_p + i_{hpv} + i_{uc} + i_{cc} = 1$.

III. Analysis of Model

A. Steady states:

Setting our equations(5)-(7) to zero, we obtain a steady state: $(s_p^*, i_{hpv}^*, i_{uc}^*)$

where $s_p^* = \frac{\mu_h}{\mu_h + P_v}$, $i_{hpv}^* = \frac{\mu_h P_v}{(1+\mu_h)(\mu_h + P_v)}$,

$i_{uc}^* = \frac{\mu_h P_v (1-P_c)}{\mu_h (1+\mu_h)(\mu_h + P_v)}$.

To determine the local stability of our steady state, we find the eigenvalues and then check the sign of the real parts. If the sign of the real parts appear negative, we can say that steady state is local stability[9]. The eigenvalues are the solutions of the characteristic equation:

$$\det(J - \lambda I) = 0$$

where $J = \begin{pmatrix} \frac{\partial F_1}{\partial s_p} & \frac{\partial F_1}{\partial i_{hpv}} & \frac{\partial F_1}{\partial i_{uc}} \\ \frac{\partial F_2}{\partial s_p} & \frac{\partial F_2}{\partial i_{hpv}} & \frac{\partial F_2}{\partial i_{uc}} \\ \frac{\partial F_3}{\partial s_p} & \frac{\partial F_3}{\partial i_{hpv}} & \frac{\partial F_3}{\partial i_{uc}} \end{pmatrix}$.

After evaluating our equations(5)-(7), the jacobian matrix(J) is defined by

$$J = \begin{pmatrix} -\mu_h - P_v & 0 & 0 \\ 0 & -\mu_h - 1 & 0 \\ 0 & 0 & -\mu_h \end{pmatrix}.$$

The characteristic equation is

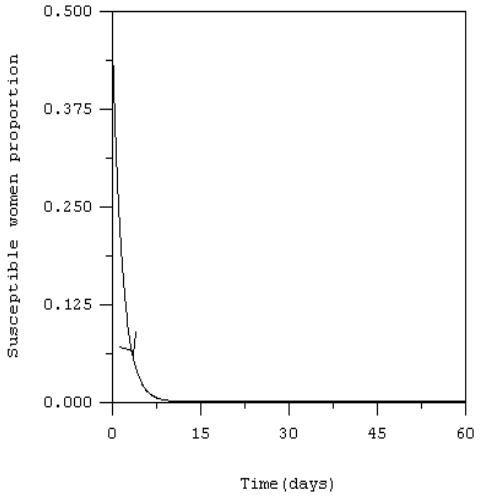
$$(\lambda + \mu_h)(\lambda + \mu_h + 1)(\lambda + \mu_h + P_v) = 0.$$

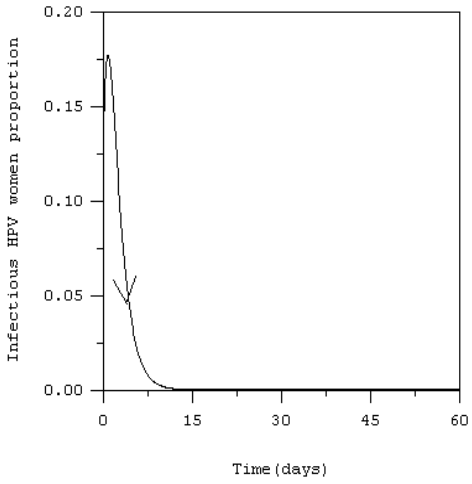
Thus, the eigenvalues are $\lambda_1 = -\mu_h, \lambda_2 = -\mu_h - 1, \lambda_3 = -\mu_h - P_v$. It can be easily seen that the real part of all eigenvalues are negatives. Therefore this steady state is local stability.

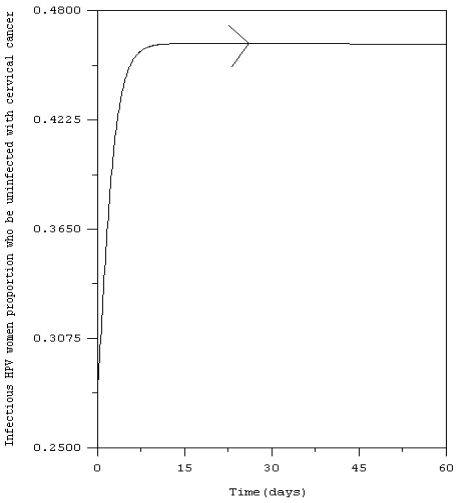
B. Numerical Solutions:

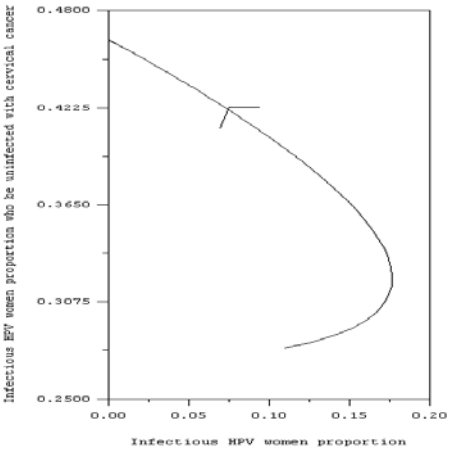
We find the numerical simulations [10] by simulating our equations(5)-(7). The parameters are follows:

$\mu_h = 1/(365\times 65)$ corresponds to the life cycle 65 years of human. $P_v = 0.6$ and $P_c = 0.7$ are arbitrarily chosen.









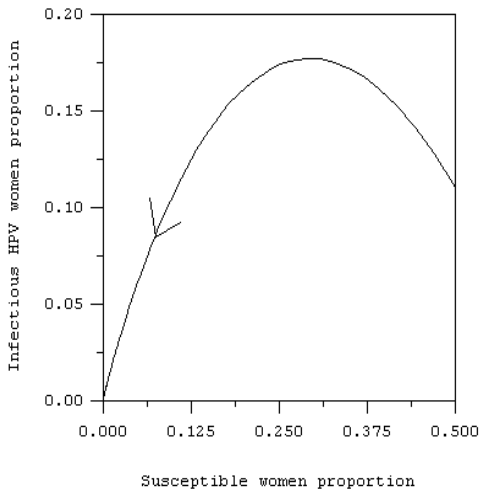


Fig. 3 Numerical solutions of our equations(5)-(7), the parameters are $\mu_h = 1/(365\times 65), P_v = 60\%,\ P_c = 70\%$.

From fig.1, we will see that the solutions converge to the steady state (0.0000702, 0.000042, 0.45298).

C. Analysis of the Parameters P_v and P_c

In this section, we analyze the model given by equation (5)-(7). The trajectories of the solutions, when P_v (the probability of women who be infected with HPV) and P_c (the probability of women with HPV can be infected with cervical cancer) are difference as shown in the following figures.

ISBN: 978-1-61804-256-9

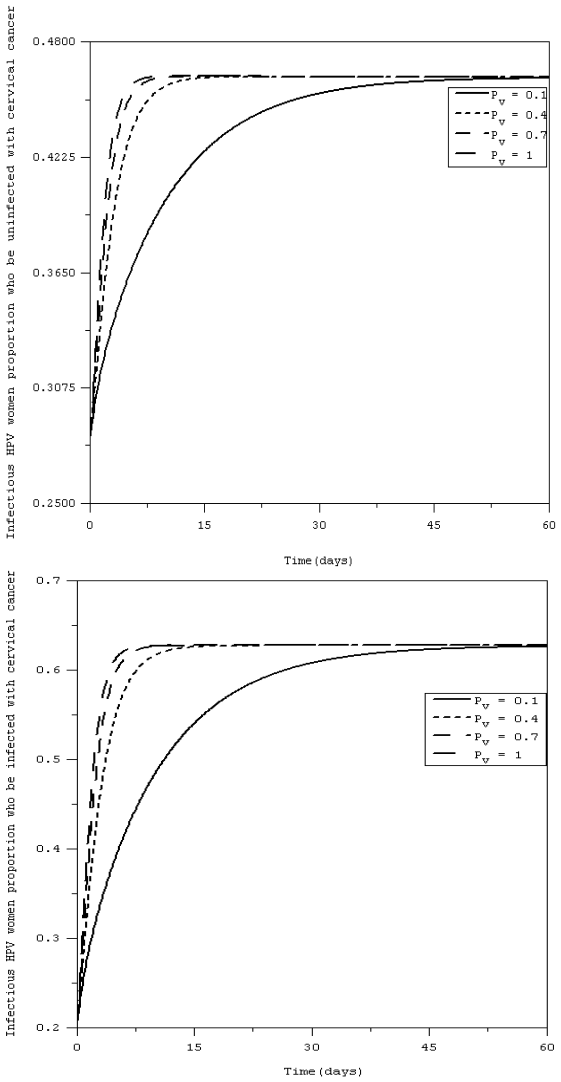
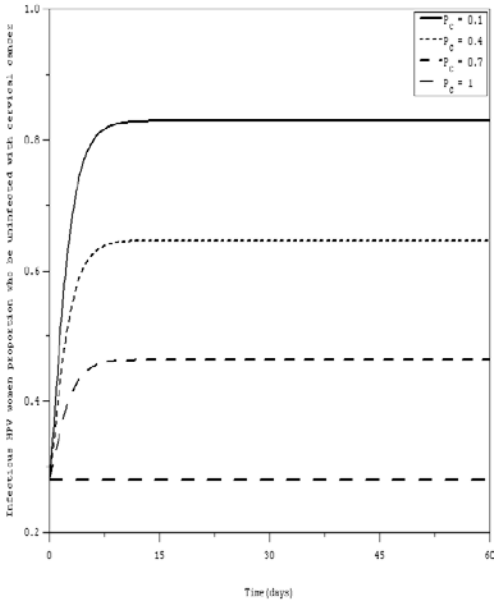


Fig. 4 Time series solutions of our model (5)-(7), when the probability of women who be infected with HPV are difference.



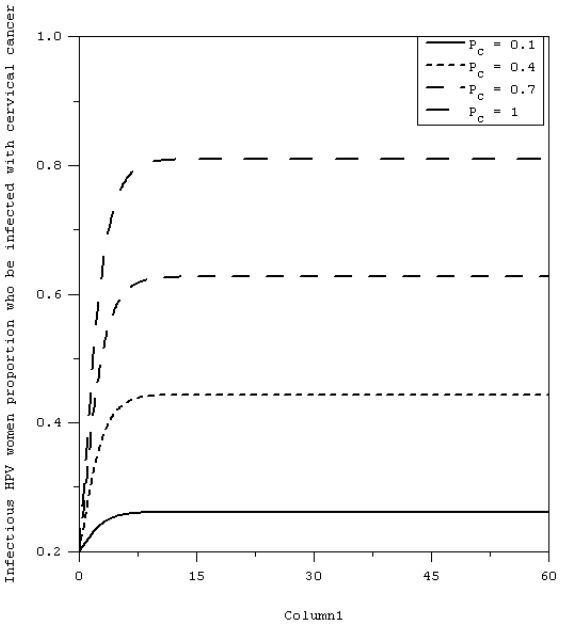


Fig. 5 Time series solutions of our model (5)-(7), when the probability of women with HPV can be infected with cervical cancer are difference.

IV. Discussion

In this study, we analyze a mathematical model of cervical cancer due to HPV. From fig.2 and fig.3, we will see that P_V (the probability of women who be infected with HPV) and P_C (the probability of women with HPV can be infected with cervical cancer) are influence to the behavior of the solutions. When P_V is higher, the time of convergence to the steady states of both infectious HPV women classes(I_{cc} and I_{uc}) are shorter. The steady solution of I_{uc} (Infectious HPV women population who be uninfected with cervical cancer) is smaller when P_C (the probability of women with HPV can be infected with cervical cancer) is higher. But I_{cc} (Infectious HPV women population who be infected with cervical cancer) is higher when P_V (the probability of women who be infected with HPV) is higher. The results are corresponding to the real situations because when the probability of infection with HPV is high then each woman can be infected in a short time. Furthermore, when the probability of women with HPV can be infected with cervical cancer is high, the number of infectious HPV women population who be infected with cervical cancer is also high.

Acknowledgment

This work is supported from faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Thailand.

REFERENCES

[1] What you need to know about Cervical cancer, National Cencer Institute, US.Department of health and human services, National Institutute of Health.
[2] Cervical Cancer[online], Available: http://en.wikipedia.org/wiki/Cervical_cancer
[3] National cancer institute[online], Available: http://www.cancer.gov/cancertopics/types/cervical
[4] American Cancer Society[online], Available: http://www.cancer.org/acs/groups/cid/documents/webcontent/003167 pdf.pdf
[5] Report of cancer cases(in Thai) [online], Available: http://www.nci.go.th/th/cancer_record/cancer_rec1.html
[6] Situation of cervical cancer in Thailand [online], Available: http://www.nci.go.th/th/cancer_record/cancer/cancer_rec1.html, National research Institute.
[7] Muller H. and Bauch C., "When Do Sexual Partnerships Need to Be Accounted for in Transmission Models of Human Papillomavirus? ," *International Journal of Environmental Research and Public Health*. pp.635 – 650, 2010.
[8] Lee SL., "A Mathematical Model of Human Papillomavirus (HPV) in the United States and its Impact on Cervical Cancer," *Journal of Cancer* ,vol. 3, pp.262 – 268, 2012.
[9] Leah EK., *Mathematical models in biology*, Random House, 1988.
[10] Hoffman JD., *Numerical Methods for Engineers and Scientists*, Sigapore: McGraw-Hill

Analysis and forecast of indicators in the industrial production in Slovak Republic

Peter Poór, Gabriela Ižaríková, Jana Halčinová, Michal Šimon

***Abstract*—** The article deals with the indicators in the industrial production. Industry represents an important sector in the world, as well as in Slovak economy. Industry and the related services affect the development of the whole country and also the development of individual regions, is a source of job opportunities. Manufacturing are divided into categories and special aggregates industry classification of economic activities. Classification of economic activities SK NACE is fully harmonized with the European version of NACE that is comparable to the international level. In this paper analyzed indicators of economic activity: number of persons employed, monthly wage labor productivity from revenues from own services and products, receipts for own performances and goods in the industrial production. In the end are prediction of the analyzed parameters of the known values. Relationship between individual variables is explained by using a correlation matrix.

***Keywords*—** industrial production, classification of industrial production, economic indicators, prediction.

I. Introduction

Between the constantly accelerating pace of innovation and technological development the industry must respond flexibly to new requirements. Industry represents an important sector in the world, as well as in the Slovak economy. Its an important part of the industrial production. Manufacturing is part of material production-oriented extraction of minerals and fuels, production and distribution of all kinds of energy, machine processing of extracted materials and derived agricultural products, various repair activities and selected services. Industry and the related services affect the development of the whole country and also the development of individual regions. It is also a source of job opportuni-ties. The impact of economic, social, technical and environmental factors cause in manufacturing various changes. Until 1989 was in every region of at least one supporting industrial plant but after 1989 primarily reflected the transformation of the industry in changing its sector, ownership, size and spatial structure. The regional distribution of industrial production shows that the critical capacity of manufacturing in terms of production, sales and share of employment are concentrated mainly in Western Slovakia.

Table 1 Classification of industrial production by SK NACE

C Manufacturing
CA Manufacture of food products, Manufacture of beverages, Manufacture of tobacco products
CB Manufacture of textiles, Manufacture of wearing apparel, Manufacture of leather and related products
CC Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials, Manufacture of paper and paper products
CD Manufacture of coke and refined petroleum products
CE Manufacture of chemicals and chemical products
CF Manufacture of basic pharmaceutical products and pharmaceutical preparations
CG Manufacture of rubber and plastic products, Manufacture of other non-metallic mineral products
CH Manufacture of fabricated metal products, except machinery and equipment
CI Manufacture of computer, electronic and optical products
CJ Manufacture of electrical equipment
CK Manufacture of machinery and equipment n.e.c.
CL Manufacture of motor vehicles, trailers and semi-trailers, Manufacture of other transport equipment
CM Other manufacturing, Repair and installation of machinery and equipment

Industrial production can be divided into categories and special aggregates industry classification of economic activities. For example breakdown by SK NACE in Tab. 1.

In conclusion, we would like to express thanks for the support of the projects SGS-2012-063 titled "Integrated design of manufacturing system as metaproduct with a multidisciplinary approach and with using elements of virtual reality" and project NEXLIZ – CZ.1.07/2.3.00/30.0038, which is cofinanced by the European Social Fund and the state budget of the Czech Republic and This article was created by implementation of the grant project VEGA no. 1/0102/11. Methods and techniques of experimental modeling of in-house manufacturing and non-manufacturing processes.

Ing. Peter Poór, PhD. is with Department of Industrial Engineering and Management, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic, E-mail: poorpeter@gmail.com, Phone: +420 377 638 401 Fax: +420 377 638 402

Doc. Ing. Michal Šimon, Phd. is with Department of Industrial Engineering and Management, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic, E-mail: simon@kpv.zcu.cz, Phone: +420 377 638 400 Fax: +420 377 638 402

Mgr. Gabriela Ižariková, Ph.D. is with Department of Applied Mathematics and Informatics , Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic, E-mail: gabriela.izarikova@tuke.sk, Phone: + 421 55 602 2227, Fax: +421 55 633 4738

Ing. Jana Halčinová, Ph.D. is Department of Biomedical Engineering and Measurement , Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic, E-mail: jana.halcinova@tuke.sk, Phone: + 421 55 602 2358, Fax: +421 55 633 2380

Classification of economic activities SK NACE is fully harmonized with the European version of NACE. Using this classification is created a statistical binding on all Member States of the European Union. Previous Slovak version of this classification was the Statistical Classification of Economic Activities, the acronym NACE. Reason for revision classification of economic activities was an attempt to take account of a technological and structural changes in the economy and to ensure comparability of economic statistics, not only at European but also at international level. Our statistical office data has been processed according to this classification since 2008, therefore in this paper data are analyzed from 2008 to 2013.

II. Research Method

We use method of least squares (LSM) for the analyz of economic indicators in the industrial production. In describing dynamic phenomena rely on indicators which are grouped into time series. The aim of the analysis time data structure is an appropriate model by which we derived based on data from the past to make predictions for specific periods in the future. Thus created time series model allows us to simulate time series in such a way that the real values and designed a model is not a significant difference. The main task of the analysis of time series is a depiction of the basic tendencies of their development, thus setting the trend.

The principle of least squares method consists in minimizing the sum of squares of empirical values y_i and theoretical values $\hat{y}_i = T$ (i.e. $S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \rightarrow \min$). If we consider a linear trend ($y = a_0 + a_1 t$) respectively linear regression function is for $a_0, a_1 \in R$ valid $\frac{\partial S(a_0, a_1)}{\partial a_0} = 0$, $\frac{\partial S(a_0, a_1)}{\partial a_1} = 0$ where $S(a_0, a_1) = \sum_{i=1}^{n} (y_i - a_0 - a_1 t_i)^2$ the parameters are the solution system of equations:

$$a_0 n + a_1 \sum_{i=1}^{n} t_i = \sum_{i=1}^{n} y_i \quad (1)$$

$$a_0 \sum_{i=1}^{n} t_i + a_1 \sum_{i=1}^{n} (t_i)^2 = \sum_{i=1}^{n} t_i y_i \quad (2)$$

.

For a polynomial of second and third order it is like in the case of a second order polynomial we get three equations with three unknowns.

$$\textbf{A.} \quad a_0 n + a_1 \sum_{i=1}^{n} t_i + a_2 \sum_{i=1}^{n} (t_i)^2 = \sum_{i=1}^{n} y_i \quad (3)$$

$$\textbf{B.} \quad a_0 \sum_{i=1}^{n} t_i + a_1 \sum_{i=1}^{n} (t_i)^2 + a_2 \sum_{i=1}^{n} (t_i)^3 = \sum_{i=1}^{n} t_i y_i, \quad (4)$$

$$\textbf{C.} \quad a_0 \sum_{i=1}^{n} (t_i)^2 + a_1 \sum_{i=1}^{n} (t_i)^3 + a_2 \sum_{i=1}^{n} (t_i)^4 = \sum_{i=1}^{n} (t_i)^2 y_i. \quad (5)$$

In economic practice, but we also meet with functions that that can not be linearized by any transformation. Among them:

- Exponential $T = a\, b^t$ (6)
- Modified exponential trend $T = k + a_0 a_1{}^t$ (7)
- Logistic trend $T = \frac{k}{1 + a_0 a_1{}^t}$, (8)
- Gompers trend $T = k\, a_0^{a_1^t}$, (9)

Selecting the shape of the regression function must respect the logical and factual context of the phenomenon and its laws. Regression function should be as simple and at the same time to guarantee the best possible approximation to the observed values. Selecting the right type of addiction is based on the scatter plot. Choosing the most appropriate model may not always be obvious from the outset, therefore we consider the most appropriate one that is most logical in which the smallest residual variation which has the largest leaks addiction.

The most preferably trend were determined by the value of the correlation coeficient the closer they are to 1, team it is more accurate.

III. Indicators in the Industrial Production

In this paper the underlying data were drawn from the database SLOVSTAT and are processed by statistical methods. In assessing the current state of the industry has been used trend analysis of selected indicators in time series and their comparison. Based on the identified knowledge is made prediction of the analyzed indicators of industrial produc-tion in 2014.

Based on data from the statistical office of the database can be done by analyzing the development of indicators in the industrial production. In this article are analyzed the following variables:

- the average number of persons employed in manufacturing,
- the average nominal monthly wage in manufacturing,
- the labour productivity from revenues from own services and products in manufacturing,
- receipts for own performances and goods in manufacturing.

For analysis was used data for the period 2008-2013, because since 2008 are known values of individual indicators by SK NACE classification. This period is referred as the 2009

crisis year, some indicators show a decline or stagnation, but some of this impact is not noticeable.

A. The average number of persons employed

The average number of persons employed includes the average number of employees and self-employed persons. Average number of employees includes permanent and temporary employees who are at work, employment, public servant or a member of an organization, regardless whether they are actually present at work or not. [4] Developments of the average number of persons employed during the review period is shown in Fig.1.

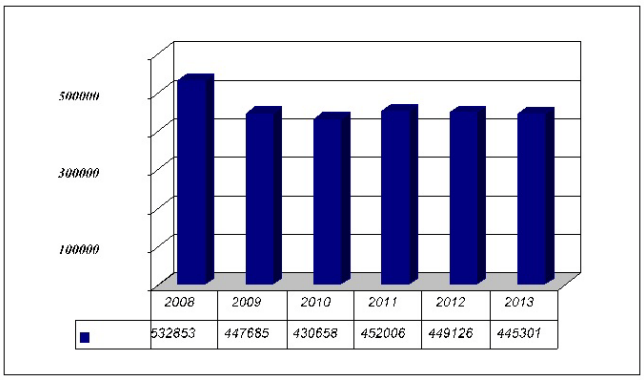


Figure 1 The average number of persons employed in the industrial production

The average number of the employed persons in the industrial production in the crisis year of 2009 decreased compared to 2008, and since then, with minor differences, the trend of nearly constant (fluctuates around the number 450 000 thousand). Trend function parameters are determined by least squares and its best feature is determined by the value of the correlation coefficient. Evolution of the average number of persons employed in industrial production can be described as a third-order polynomial: (year 2008 - t = 1)

$$y = 702453{,}3 - 226202{,}4\,t + 60116\,t^2 - 4937{,}4\,t^3 \qquad (10)$$

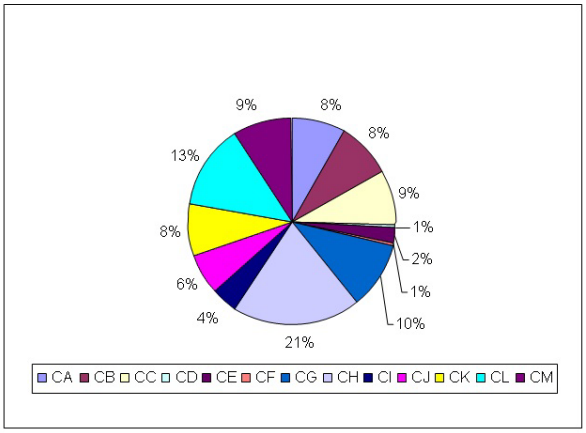


Figure 2 Number of persons employed in the industrial production

The highest number of persons employed during the period was in 2012 year (452006 thousand) and the lowest in 2013 year (445 301 thousand). Proportion of employees under each category shown graphically in Fig. 2. From the all employed in manufacturing is the most people employed in manufacture of basic metals and fabricated metal products except machinery and equipment (21%) and the least in the manufacture of coke and refined petroleum products and in the manufacture of basic pharmaceutical products and pharmaceutical preparations (1%).

B. The average nominal monthly wage

The average nominal monthly wage labor costs shall include the amount paid to its own employees as compensation for work or a replacement on the basis of the legal relationship with the employer (work, service, civil service or membership relation). Its gross wage lowered by legal or agreed with the employee deductions. [4] Developments of the average nominal monthly wage during the review period is shown in Fig.3.

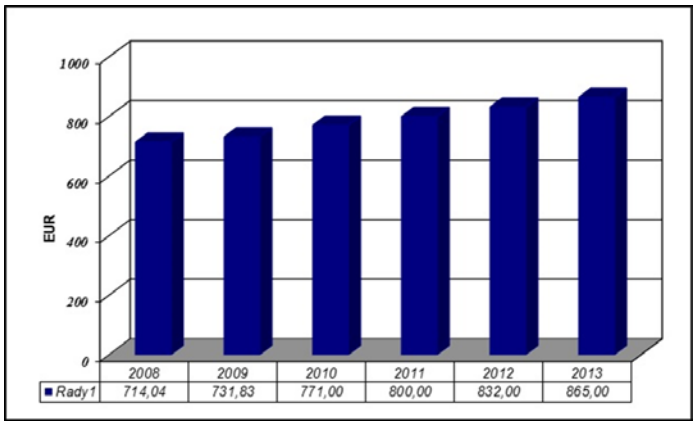


Figure 3 The average nominal monthly wage in the industrial production

The average nominal monthly wage in manufacturing during the period of growth didn't decline even during the cri-sis. The development trend of the average monthly salary can be described as trend function: (year 2008 - t = 1)

$$y = 30{,}98t + 677{,}21 \quad (r = 0{,}9972). \qquad (11)$$

Tab. 2 reflects a comparison of the average nominal monthly salary in manufacturing. The monthly wage is higher than the average across manufacturing employees achieved in eight of the thirteen monitored categories. Maximum wage employees are in manufacture of coke and refined petroleum products (224.4%) and the lowest employees in manufacture of textiles, apparel, leather and related products (66.44%). The difference between the highest (CD) and the lowest (CB) average monthly wage is higher than the average monthly wage in the whole manufacturing. The graph in Fig. 4 presents the evolution of the average monthly wage in manufacturing, by categories. It is evident that the trend of development of all categories is almost identical.

Table 2 Comparison of the average monthly salary by category

Category	C	CA	CB	CC	CD	CE	CF
Wage (%)	100	89	66,4	83,5	224	117	135
Category	CG	CH	CI	CJ	CK	CL	CM
Wage (%)	104,5	102,2	100,2	99,8	112	118	96

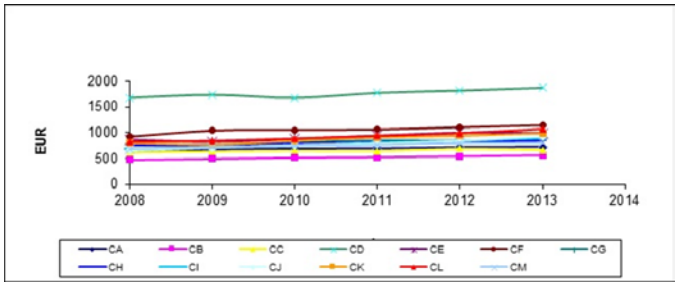


Figure 4 The average nominal monthly wage by category

C. Receipts for own performances and goods in the industrial production

Receipts for own performances and goods sold includes the value of goods and services from their own production and commercial goods destined for domestic and foreign customers. The data are exclusive of value added tax and excise duties. [4] Developments of receipts for own performances and goods during the review period is shown in Fig.5.

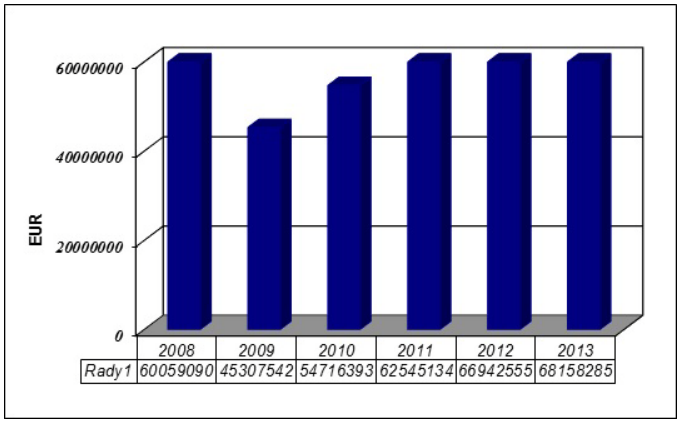


Figure 5 Receipts for own performances and goods in the industrial production

Receipts for own performances and goods in the industrial production in the crisis of 2009 year decreased compared to 2008 year (25%) even in 2010 year were lower than in 2008 year, but higher than in 2009. Development of revenues from own services and products in the industrial production function can be approximated by: (year 2008 – t=1)

$$y = -1317259{,}98 + 14898920{,}3t - 45995839{,}16t^2 + 91458587{,}37t^3 \quad (r = 0{,}9636). \qquad (12)$$

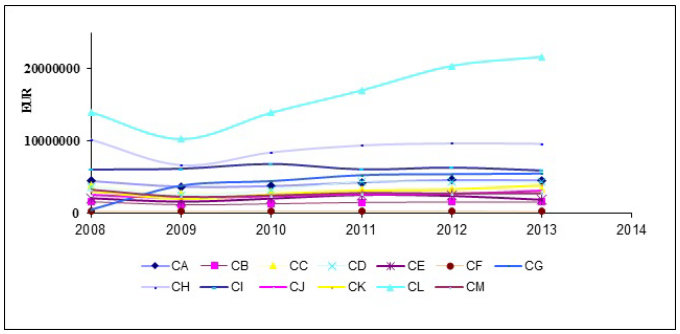


Figure 6 Receipts for own performances and goods by category

Receipts for own performances and goods in the industrial production in each category increased mostly in the manufacture of transport equipment, which is connected especially with the advent of Kia Motors, Volkswagen and others in Slovakia. Nearly 30% of total sales accounted by sales in the manufacture of transport equipment. The second area, which represents 15% of total sales are metals and metal products, except machinery and equipment. Overview of the percentage rate of sales to total sales for each category is shown in Fig.6.

D. The labour productivity from revenues from own products and goods

Developments of the labour productivity from revenues from own products and goods during the review period is shown in Fig.7.,

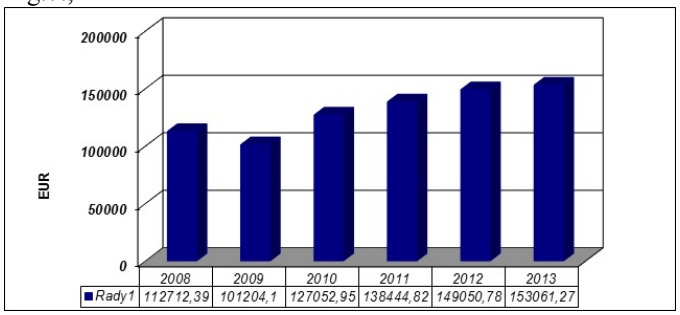


Figure 7 The labour productivity from revenues from own products and goods in the industrial production

The labour productivity from revenues from own services and products in the industrial production in the crisis of 2009 year decreased compared to 2008 year (about 10%) and since then has upward trend (growth factor), trend function has the form: (rok 2008 – t=1)

$$y = 139065{,}4 + 44353{,}43t + 17675{,}28t^2 - 1655{,}1t^3 \quad (r = 0{,}9749). \qquad (13)$$

Category	C	CA	CB	CC	CD	CE	CF
Wage (%)	10	72	29	63	114	167	83
Category	CG	CH	CI	CJ	CK	CL	CM
Wage (%)	67	63	217	61	53	175	41

Tab. 3 shows a comparison of the labor productivity from revenues from own services and products in the in-dustrial production. The highest labor productivity has manufacture of coke and refined petroleum products (11 times higher than average). Values higher than the overall average is only in three categories (CE, CI, CL). Low labor produc-tivity from revenues from own services and products in the manufacture of textiles, apparel, leather and leather products (CB - 29%), which is the lowest average nominal monthly wage. The graphic display (Fig.8) reflects that labor productivity is different in each category, somewhere stagnant, growing somewhere in the categories CA and CD has a variable character.

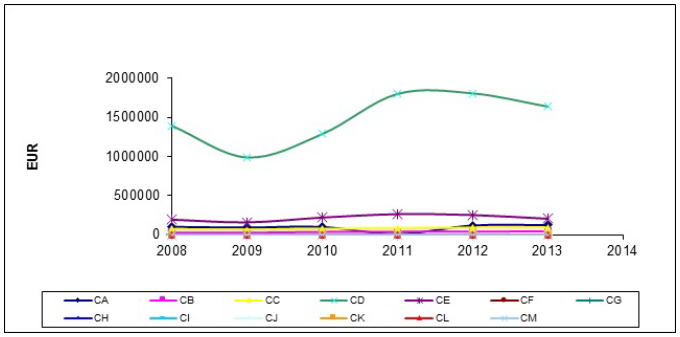


Figure 8 The labour productivity from revenues from own products and goods by category

IV. Prediction Indicators in the Industrial Production for 2014

Prediction of the analyzed parameters established by the estimated approximation of functions, whereas for year 2008 is t = 1, this means that for year 2014 is t = 7. Calculated values of the investigated parameters in the industrial production for 2014 are:

- the average number of persons employed: 371191 thousand persons,
- the average nominal monthly wage: 894.07 €,
- the labour productivity from revenues from own services and products: 126 984.1 €,
- receipts for own performances and goods: 47 714 636 €.

Table 3 Correlation matrix

Indicators	*Number of employees*	*Wage*	*Receipts*	*Labour productivity*
Number of employees	1	-0,55	0,08	-0,38
Wage	-0,55	1	0,75	**0,95**
Receipts	0,08	0,75	1	**0,89**
Labour productivity	-0,38	**0,95**	**0,89**	1

One of monitored indicators in the industrial production and the average nominal monthly wage has a higher value than in the previous period. The other three indicators (the average number of persons employed and the labor produc-tivity from revenues from own services and products and sales of own products and goods) entered values below, which was achieved in 2013. To construct the mathematical model - approximation of functions affected the values that these indicators reached in 2009, so these features have not only increasing character, but also declined values, although preliminary data for 2014 point showed their growth. Individual variables interact the growth of one of them will increase the value of another parameter. The most significant correlations are indicated in the correlation matrix in Tab. 4. It is clear that the wage employee turnover and productivity from revenues from own services and products is a significant correlation.

V. Conclusion

Industry and related services are the core of the Slovak economy, a source of job creation, the driving force of productivity and innovation. Performance of industry and manufacturing is shaping the level of productivity of the Slovak economy in relation to the European Union. The analysis showed that the individual sectors in the industrial produc-tion development are highly differentiated. In the category of coke and refined petroleum products are em-ployed but the least people reach there highest salary Receipts for own performances and goods and hence productivity is highest in this category. Recently, there is also an increase in production indicators in the categories of vehicles. The analysis above shows that in order to increase the competitiveness of industrial production and its individual sectors is necessity to look for the optimal way of industry development.

Acknowledgment

In conclusion, we would like to express thanks for the support of the projects SGS-2012-063 titled “Integrated design of manufacturing system as metaproduct with a multidisciplinary approach and with using elements of virtual reality“ and project NEXLIZ – CZ.1.07/2.3.00/30.0038, which is cofinanced by the European Social Fund and the state budget of the Czech Republic and This article was created by implementation of the grant project VEGA no. 1/0102/11. Methods and techniques of experimental modeling of in-house manufacturing and non-manufacturing processes.

References

[1] HUDEC O a kol.(2007): Štatistické metódy v ekonomických vedách. Košice:Elfa, pp. 195.
[2] RUBLÍKOVÁ E.(2007): Analýza časových radov. EKONÓMIA Bratislava, pp. 207.
[3] AMIRD D ACZEL(1989): Complete Business Statistcs. Irwin, Boston
[4] http://www .statistics.sk
[5] JANUŠKA, M., ŠŤASTNÁ, L. Industrial Engineering in the Non-Manufacturing Processes. In Proceedings of The 22nd International Business Information Management Association Conference. neuveden:

International Business Information Management Association (IBIMA), 2013. s. 747-766. ISBN: 978-0-9860419-1-4
[6] HASSANI, Hossein; HERAVI, Saeed; ZHIGLJAVSKY, Anatoly. Forecasting European industrial production with singular spectrum analysis. International journal of forecasting, 2009, 25.1: 103-118.
[7] JACOBS, Jan; STURM, Jan-Egbert. Do ifo indicators help explain revisions in German industrial production?. Physica-Verlag HD, 2005.
[8] BODO, Giorgio; GOLINELLI, Roberto; PARIGI, Giuseppe. Forecasting industrial production in the euro area. Empirical economics, 2000, 25.4: 541-561.
[9] BOX, George EP; JENKINS, Gwilym M.; REINSEL, Gregory C. Time series analysis: forecasting and control. John Wiley & Sons, 2013.
[10] LASTER, David; BENNETT, Paul; GEOUM, In Sun. Rational bias in macroeconomic forecasts. Quarterly Journal of Economics, 1999, 293-318.
[11] HERAVI, Saeed; OSBORN, Denise R.; BIRCHENHALL, C. R. Linear versus neural network forecasts for European industrial production series. International Journal of Forecasting, 2004, 20.3: 435-446.
[12] SWANSON, Norman R.; GHYSELS, Eric; CALLAN, Myles. A multivariate time series analysis of the data revision process for industrial production and the composite leading indicator. Cointegration, Causality and Forecasting: Festschrift in Honor of Clive WJ Granger, 1999.
[13] KENNEDY, James. An analysis of revisions to the industrial production index. Applied Economics, 1993, 25.2: 213-219.
[14] BULLIGAN, Guido; GOLINELLI, Roberto; PARIGI, Giuseppe. Forecasting monthly industrial production in real-time: from single equations to factor-based models. Empirical Economics, 2010, 39.2: 303-336.
[15] HAUSTEIN, Heinz-Dieter; NEUWIRTH, Erich. Long waves in world industrial production, energy consumption, innovations, inventions, and patents and their identification by spectral analysis. Technological forecasting and social change, 1982, 22.1: 53-89.
[16] BRUNO, Giancarlo; LUPI, Claudio. Forecasting industrial production and the early detection of turning points. Empirical economics, 2004, 29.3: 647-671.
[17] BULLIGAN, Guido; GOLINELLI, Roberto; PARIGI, Giuseppe. Forecasting monthly industrial production in real-time: from single equations to factor-based models. Empirical Economics, 2010, 39.2: 303-336.

The Local Histogram Equalization And Adaptive Thresholding for Hand-Based Biometric Systems

Haryati Jaafar, Salwani Ibrahim and Dzati Athiar Ramli

***Abstract*—** Hand-based biometric systems are the emerging type of biometrics that attracts researchers in biometrics area. As compared to the other biometric traits such as face and iris, the image quality of a hand-based system are robust with more information can be employed even though it is in low resolution. A new approach image enhancement and segmentation called the local histogram equalization and adaptive thresholding (LHEAT) was proposed to improved the quality of image taken. It was firstly obtained to ensure an equal distribution of the brightness levels. The useful information of the image was then extracted and the foreground from the nonuniform illumination background was separated. The sliding neighborhood operation was also applied such that the computation is much faster. Three hand-based biometric databases i.e. the fingerprint, finger vein and palm print databases were employed and evaluated based on the quality of image and classification accuracy (CA). Experimental evaluation based on quality of image shows that the proposed LHEAT has better performance than local histogram equalization (LHE) and local adaptive thresholding (LAT) with more than 45 of peak-signal-to-noise ratio (PSNR). The results also shows that the proposed LHEAT is able to achieved more than 90% in term of CA. This shows that the proposed LHEAT is able to enhance and segmented the images effectively.

***Keywords*—** Hand-based biometric system, LHEAT, LHE, LAT, sliding neighborhood

I. Introduction

Today's complex demands for reliable authentication and identification methods are increasing rapidly. Initially, the traditional technologies such as personal identification number (PIN), smart cards and passwords were introduced [1]. However, they had a number of inherent disadvantages such as duplication, misplacing and hacking. Therefore, biometrics were introduced in the late 90s to recognize a person based on the physiological or biological characteristics [2]. The biometric technology is inherently more reliable. It is capable to provide a level of assurance for the preventions of duplication, stealing and hacking. Due to the specific physiological or behavioral characteristics that are possessed by the users, this technology is able to be implemented in various fields such as door access controls, criminal investigations, logical access points and surveillance applications [3].

There are various kinds of modalities of the biometric systems that are either widely used or developed such as the fingerprint, iris, face, hand geometry, palm print, gait, voice and signature [1]. Among the available biometrics, hand-based systems such as the finger vein, fingerprint and palm print are found to be the most popular due to their high user acceptance and excellent advantages in their application [4].

The features of the finger vein are inside the skin surface, which makes it difficult to be duplicated. Thus, it is more secure compared to other modalities and leads to the high recognition accuracy. In addition, as the veins are located inside the body; it is less likely to be influenced by changes in the weather or physical condition of the individual. Moreover, the rushes, cracked and rough skin does not affect the result of recognition [11]. On the other hands, the images quality of a fingerprint and palm print are robust because of its multiple lines, wrinkles and ridges while the palm print covers even more information and the ridge structures remain unchanged throughout the life, except for a change in size [12]. Currently, they are offering low costs for data acquisition and the possibility of acquiring the data easily. The image can be collected in the real environment where the acquisition devices had no pegs holding the finger or palm.

However, the main problem with hand-based images is that they are of low quality due to several reasons such as the movement of hands, use of low resolution capturing devices and environmental factors. These factors obscure image details and create noise which badly effect object detection and recognition. Commonly, the problem of suppression of noise in these images is solved by a smoothing technique [5]. However, this process has the potential to blur all sharp edges containing an important information about the image [6]. In order to overcome this problem, a combination of image enhancement and segmentation techniques is found to be more appropriate in such ways. Hence, this paper proposed a contrast image enhancement and image segmentation by

This work was supported in part by Research University Grant 814161 and Research University-Post Graduate Grant Scheme 8046019.

H. Jaafar, is with Intelligent Biometric Group, School of Electrical and Electronic, Universiti Sains Malaysia, Engineering Campus, 14300 Nibong Tebal, Pulau Pinang, Malaysia (e-mail: haryati.jaafar@yahoo.com).

S. Ibrahim, is with Intelligent Biometric Group, School of Electrical and Electronic, Universiti Sains Malaysia, Engineering Campus, 14300 Nibong Tebal, Pulau Pinang, Malaysia (e-mail: salwani.ibrahim@gmail.com).

D.A. Ramli is with Intelligent Biometric Group, School of Electrical and Electronic, Universiti Sains Malaysia, Engineering Campus, 14300 Nibong Tebal, Pulau Pinang, Malaysia (corresponding author to provide phone: +604-5996028; e-mail: dzati@usm.my).

introducing the local histogram equalization and adaptive thresholding (LHEAT) technique. This technique is an improved version of the local histogram equalization (LHE) and local adaptive thresholding (LAT) techniques [7, 8]. In the LHEAT, the LHE was firstly obtained to ensure an equal distribution of the brightness levels. The LAT was employed to extract the useful information of the image that had been enhanced by the LHE and separated the foreground from the nonuniform illumination background. In addition, the sliding neighborhood operation was applied such that the computation is much faster. This is an advantage of the LHEAT on reducing the time processing of the image enhancement stage compared to the baseline LHE and LAT techniques. The rest of this paper is organized as follows: The proposed LHEAT technique is described in Section II. The experimental set up and results are explained in Section III, and this paper is concluded in Section IV.

II. The Proposed Local Histogram Equalization And Adaptive Thresholding

The flowchart of the LHEAT techniques is shown in Fig. 1.

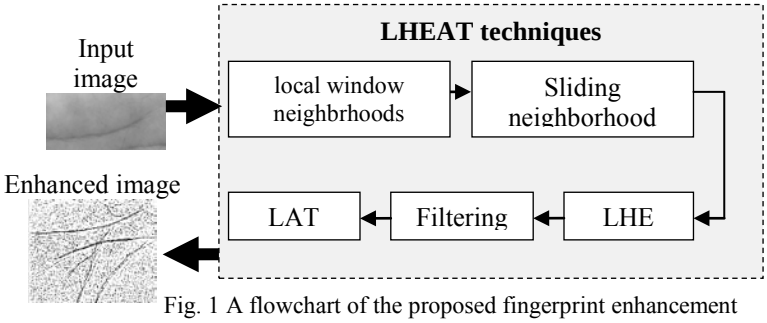


Fig. 1 A flowchart of the proposed fingerprint enhancement algorithm.

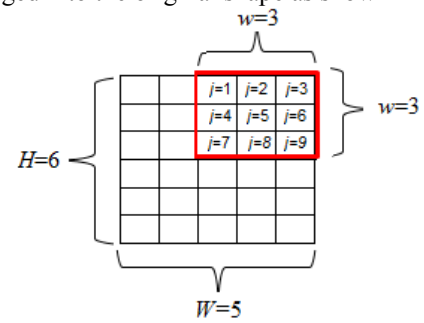
An input image was first broken into small blocks or local window neighborhoods containing a pixel. This was similar in the LHE, LAT and LHEAT. Each block was surrounded by a larger block. The input image was defined as $X \in R^{H \times W}$, with dimensions of $H \times W$ pixels, and the enhanced image was defined as $Y \in R^{H \times W}$, with $H \times W$ pixels. The input image was then divided into the block $T_i = 1,...,n$ of window neighborhoods with the size $W \times W$, where $w < W, w < H$ and $n = \left[\frac{H \times W}{w \times w} \right]$.

Each pixel in the small block was calculated using a mapping function and threshold. The size of w should be sufficient to calculate the local illumination level, both objects and the background [9]. However, it led to a complex computation which can be reduced by employing the sliding neighborhood. This operation can also decrease the acceleration of the computation. Fig. 2 shows an example of the sliding neighborhood operation. An image with a size of 6×5 pixels was divided into blocks of window neighborhoods with a size of 3×3 pixels. It is shown in Fig. 2(a). The 6×5 image matrix was first rearranged into a 30 column (6×5=30) of temporary matrix, as shown in Fig. 2(b). Each column contained the value of the pixels in its nine rows (3×3=9) window. The temporary matrix was then reduced by using the local mean (M_i):

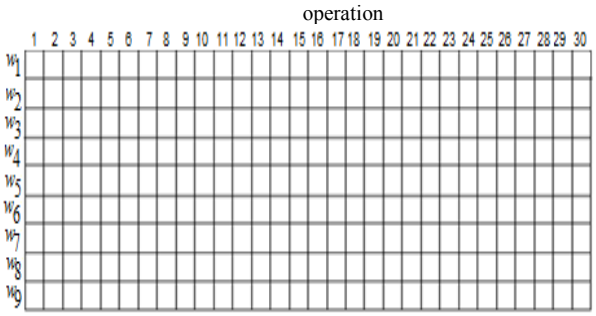
$$M_i = \frac{1}{N} \sum_{j=1}^{n} w_j \quad (1)$$

where w was size of window neighborhoods, j was the number of pixels contained in each neighbourhood,i was the number of column in temporary matrix and N was the total number of pixels in the block.

After determining the local mean in Equation (1), there was only one row left as shown in Fig. 2(c). Subsequently, this row was rearranged into the original shape as shown in Fig. 2(d).

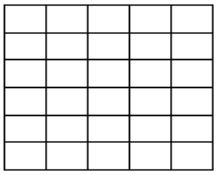


(a) Original image with window neighborhoods



(b) Temporary matrix

(c) One row matrix



(d) Rearranged row into the original shape

Fig. 2 The sliding neighborhood

The LHE was then obtained to ensure an equal distribution of the brightness levels. There are three major steps in the LHE technique. There are the probability density (PD), the cumulative distribution function (CDF) and the

mapping function. The probability distribution of image PD for each block can be expressed as:

$$P(i) = \frac{n_i}{N} \text{ for } i = 0,1,...,L-1 \quad (2)$$

where n_i is the input pixel number of level, i is the input luminance gray level and L is gray level, which was 256 in the investigated case.

The LHE uses an input-output mapping that is derived from CDF of the input histogram as defined in below:

$$C(i) = \sum_{i=0}^{n} P(i) \quad (3)$$

Although the image has been enhanced, it remains mildly degraded because of the background noise and variation in contrast and illumination. Hence, the 2D median filter, containing a 3×3 mask was applied over the grayscale image to reduce the effect of salt and pepper noise and the blur of the edge of the image. Given an input vector is $x(n)$ and $y(n)$ is the output median filter of length l where l defines the number of samples over which median filtering takes place. When l is odd, the median filter can be defined as stated in Equation (4).

$$y(n) = median\{x(n-k:n+k), k = (l-1)/2\} \quad (4)$$

When l is even, the mean of the two values at the center of the sorted samples list is used.

Once it was filtered, the image was segmented using the LAT technique to separate the foreground from the background by converting the grayscale image into binary form. The Sauvola's technique was applied here due to its promising effects on the degraded images. By using the Sauvola's technique, the following formula for the threshold is:

$$T_h(i) = M\left[1 + k\left(\frac{Z}{R} - 1\right)\right] \quad (5)$$

where T_h is the threshold, k is a positive value parameter with k= 0.5, R is the maximum value of the standard deviation, which was set at 128 for grayscale image and Z is the standard deviation which can be found as:

$$Z = \sqrt{\frac{1}{N-1}\sum_{j=1}^{n}\left(w_j - M\right)} \quad (6)$$

The binarization results can be denoted as follows $y(i)$ as in Equation (7)

$$y(i) = \begin{cases} 1 & \text{if } q(i) > T_h(i) \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

As mentioned before, a suitable value for window size, w was greatly affected the image. If the window size is too small, the image resulted segmented regions appear was less visible. Meanwhile the large window size had caused the important details of the image was disappeared. In this study, the best value of w is obtained by the w=11 for the finger vein and fingerprint databases and w=9 for the palm print database.

III. Experimental Results

In this section, a comparative study on the performance of LHEAT technique on the hand-based biometric database had been investigated and compared with the LHE and LAT techniques. The experiments were implemented using Matlab R2010 (b) and were tested in Intel Core i5, 2.1GHz CPU, 6G RAM and Windows 7 operating system.

A. Data Acquisition

The finger vein database was provided by the IBG, USM. It is available for downloading from the following website: http://blog.eng.usm.my/fendi/. The capturing device was comprised of three units of Near-Infrared-light emitted diode (NIR-LED) of wavelength = 850 nm and a Sony PSEye camera with an IR passing filter. The NIR-LEDs were placed in a row on the top section while the camera was attached to the bottom side of the capturing device. To reduce the user's discomfort, the users were simply asked to place their fingers on the acquisition devices and there had no pegs holding the finger. The spatial and depth resolutions of the images were set at 640×480 pixels and 256 grey levels, respectively. The images were then segmented into the region of interest (ROI). A few examples of the ROI of the finger vein images are shown in Fig. 3.

The database was obtained from 123 volunteers who were staffs and students (83 males and 40 females) from University Sains Malaysia (USM). The range of age of the users was from 20 to 52 years old. Each user contributed four of their fingers which were the left index, left middle, right index and right middle fingers resulting in 492 finger classes for this investigation. The images were acquired in two sessions with a time gap of by more than two weeks. Each finger was captured six times in every session. There were 2952 samples extracted from the first and second sessions were used as the training and testing samples, respectively.

The fingerprint database was obtained from the Fingerprint Verification Competition 2006 (2006FVC). The image was collected by using an optical sensor with the resolution of the sensor is 569 dpi in the image format of BMP, 256 gray-levels size of 400x560 pixels. [10]. This database was collected from 150 volunteers who were randomly selected including the manual workers and elderly people. They were simply asked to place their fingers on the acquisition device. There was no constraint was enforced to guarantee the highest quality of the captured images. The final databases were selected from a larger database by choosing the fingers that were more difficult to be evaluated according to a quality index. This was done to make the benchmark sufficiently difficult for a technology evaluation. Each user had provided 12 samples per finger. Thus, the final databases collected were 1800 fingerprint images 1800 samples from 150 users and they were then partitioned in half (900 as the training samples and the other 900 for the testing samples). Some examples of the fingerprint images are shown in Fig. 4

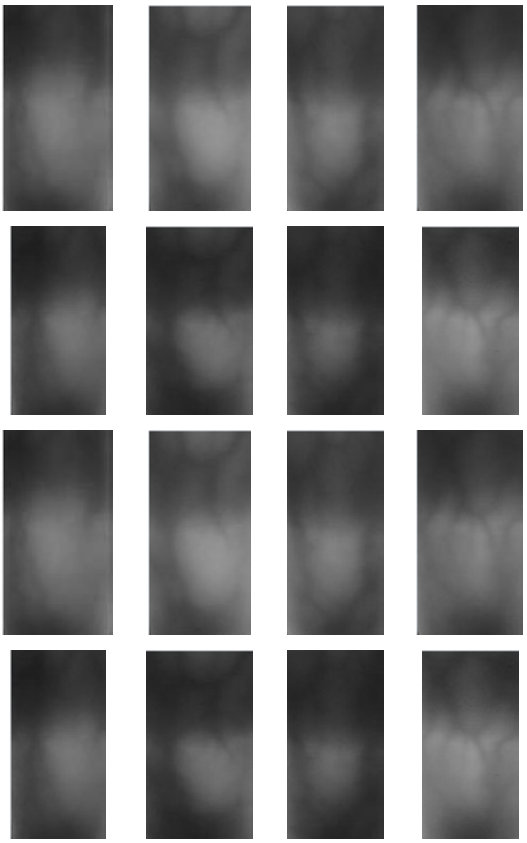


Fig. 3 The example of extracting ROI finger vein image collection

.

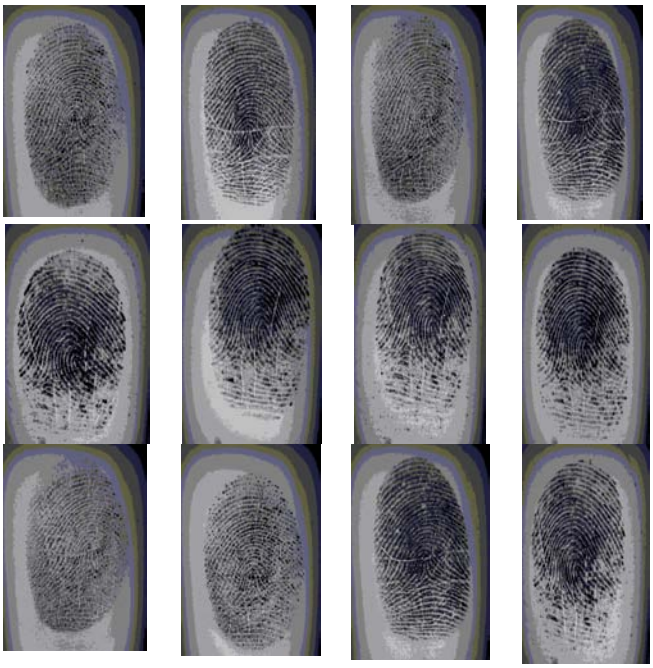


Fig. 4 The examples of fingerprint image collection

The palm print database was provided by the IBG, USM. The image was taken using a HTC One X android mobile phone with the image resolution of 8 megapixels of image resolution at a fixed background the files were saved in JPEG format. It was obtained by collecting the palm print images from 40 users, who were the students from School of Electrical and Electronic, USM. The age range of the users were from 19 to 23 years old. The database were comprised of 60 palm print images from every user in which 20 of them were used as the training samples and the other 40 images were applied as the testing samples. The original image was then transformed into a gray scale image and extracted into the ROI. A few examples of the ROI of the palm print images are shown in Fig. 5.

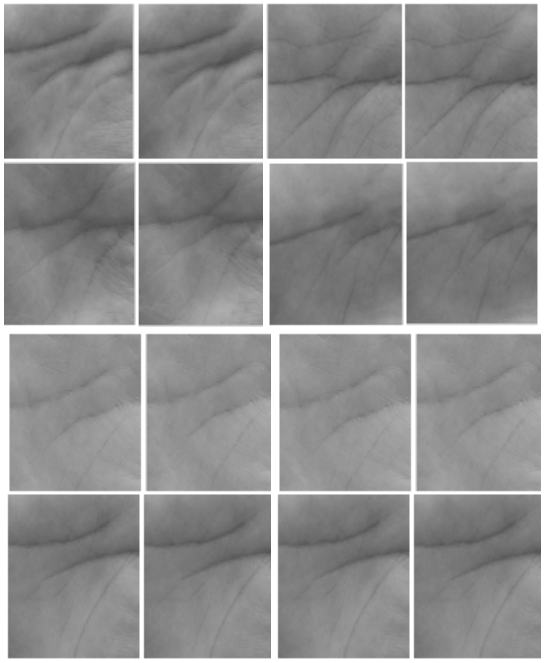


Fig. 5 The examples of extracted ROI palm print image collection

B. Performance Evaluation

In this study, the evaluation of the performance of the hand-based biometric system is based on quality of image and the CA. In the quality of image, the performance of the proposed technique was evaluated in two evaluations subjectively and objectively. The perception of an image quality improvement in the human visual system is a subjective evaluation while the perception of quantitative measures is an objective evaluation. The objective evaluation is determined based on the PSNR computation. The higher value of the PSNR the more improved is an image.
PSNR is calculated using:

$$PSNR = \frac{10\log_{10}(L-1)^2}{MSE} \tag{8}$$

where MSE can be calculated as:

$$MSE = \sqrt{\frac{\sum_{i=1}^{Y}\sum_{j=1}^{X}(m-Y)^2}{R\times C}} \tag{9}$$

where X is the original image, Y is an enhanced image, m is the intensity of the pixel at position (i,j), R and C are the row and column of the image size.

The duration of the processing time of each method is also compared to investigate the complexity of the enhancement

approaches. The enhancement process is expected to be computed with a minimum period of run time.

In order to investigate the effectiveness of proposed technique based on the CA, the k nearest neighbor (kNN) classifier with k=5 was employed to calculate the score of the pattern matching between training and testing data of the databases. The experiments were evaluated in terms of CA such that:

$$C_A = \frac{N_C}{N_A} \times 100\% \quad (10)$$

where N_c was the correct identified number of samples and N_A was the total number of test samples.

Table I, II and III show the comparison of output results based on the quality of images for the finger vein, fingerprint and palm print, respectively. It was observed the LHEAT technique attains the best result in all conditions and exhibits the highest quality results according to visual inspection, PSNR and processing time.

For the subjective evaluation, the details in the enhanced images using LHEAT were clearer and sharper, especially in the fine details like the ridges in which they were became more visible. For the objective evaluation, the LHEAT obtained the highest value of PSNR with more than 45 compared to LHE and LAT techniques. The LHEAT gives another advantage over other methods in term of its simplicity in computation. In the proposed LHEAT technique, the time complexity is $O(n^2)$ because the sliding neighborhood is only used to obtain local mean (M) and local standard deviation (Z). Hence, the time required for LHEAT is much closer to global techniques.

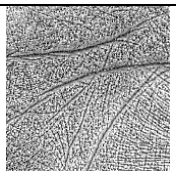
TABLE I. COMPARISON OF THE LHE, LAT AND LHEAT FOR THE FINGER VEIN IMAGE

	LHE	LAT	LHEAT
Image			
Time (s)	0.436	0.976	0.141
PSNR	33.81	38.89	49.55

TABLE II. COMPARISON OF THE LHE, LAT AND LHEAT FOR THE FINGERPRINT IMAGE

	LHE	LAT	LHEAT
Image			
Time (s)	0.551	4.766	0.133
PSNR	42.81	40.79	49.93

TABLE III. COMPARISON OF THE LHE, LAT AND LHEAT FOR THE PALM PRINT IMAGE

	LHE	LAT	LHEAT
Image			
Time (s)	2.847	3.596	0.531
PSNR	40.98	41.44	45.37

To further investigate the superiority of the proposed LHEAT, the analytical results of LHE.LAT and LHEAT in term of CA are also presented in Table IV. It was observed that the LHEAT achieves the highest CA compares to the LHE and LAT, yielding a CA of 90.93%, 93.26% and 92.6% for finger vein, fingerprint and palm print databases, respectively. It can be concluded in the LHEAT yield promising results since the brightness levels has been enhanced by distributing the brightness equally and recovered original images that were over- and under-exposed.

TABLE IV. COMPARISON OF THE LHE, LAT AND LHEAT BASED ON THE CA

	LHE	LAT	LHEAT
Finger vein	78.6%	81.07%	90.93%
Fingerprint	88.31%	83.72%	93.26%
Palm print	87.2%	82.41%	92.66%

IV. CONCLUSION

This paper focused on image enhancement and segmentation of the quality of the hand-based biometric images such as the finger vein, fingerprint and palm print. To enhance images, we propose the LHEAT technique. Because the sliding neighborhood operation is applied in the LHEAT technique, the computation was much faster compared with

previous techniques, such as LHE and LAT. Moreover, this method works well in in the real environments.

Extensive experiments were performed to evaluate the performance of the system in terms of image enhancement and image classification. The proposed system exhibits promising results. In terms of quality of the image, the PSNR with the LHEAT technique was more than 45, and the processing time was three-fold lower than with the LHE and LAT techniques. In addition, the proposed LHEAT was achieved more than 90% in term of CA. The proposed LHEAT technique is convenient and able to manage in the real environment.

ACKNOWLEDGMENT

This work was financially supported by Research University Grant 814161 and Research University-Post Graduate Grant Scheme 8046019.

REFERENCES

[1] H. Jaafar and D. A. Ramli, "A Review of Multibiometric System with Fusion Strategies and Weighting Factor," in International Journal of Computer Science Engineering (IJCSE), vol.2, no.4, 2013, pp. 158-165.

[2] J. P. Campbell, D. A. Reynolds and R. B. Dunn, "Fusing High-And Low-Level Features for Speaker Recognition," in INTERSPEECH, 2003.

[3] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition. Circuits and Systems for Video Technology," IEEE Transactions, vol. 14. No. 1, 2004, pp. 4-20.

[4] L. Zhang, D. Zhang, and H. Zhu, "Online finger-knuckle-print verification for personal authentication," Pattern Recognition, vol. 43, pp. 2560–2571, 2010.

[5] D. Luo, Pattern recognition and image processing, Horwood Publishing Limited, West Sussex, England, 1998.

[6] M. Sonka, V. Hlavac, and R. Boyle, Image processing, analysis, and machine vision. Third Ed. Thompson Corporation, USA, 2008.

[7] H. Zhu, F. H. K. Chan and F. K. Lam, "Image Contrast Enhancement by Constrained Local Histogram Equalization," Comput Vis Image Underst, vol. 73, 1999, pp. 281–290.

[8] T. R. Singh, S. Roy, O. I. Singh, T. Sinam and K. Singh, "A new local adaptive thresholding technique in binarization," International Journal of Computer Science Issues, vol. 8, no 2, 2011, pp. 271-277.

[9] Y. T. Pai, Y. F. Chang and S. J. Ruan, "Adaptive thresholding algorithm: Efficient computation technique based on intelligent block detection for degraded document images," Pattern Recognition, vol. 43, 2010, pp. 3177–3187.

[10] J. Fierrez, J. Ortega-Garcia, D. Torre Toledano and J. Gonzalez-Rodriguez, "Biosec baseline corpus: A multimodal biometric database," Pattern Recognition, vol. 40, no. 4, 2007, pp. 1389-1392.

[11] G. K. O. Michael, C. Tee, A. T. Jin, "Touch-less palm print biometrics: Novel design and implementation," Image Vis Comput. Vol. 26, pp. 1551–1560, 2008.

[12] M. S. M. Asaari, S. A. Suandi and B. A. Rosdi, "Fusion of Band Limited Phase Only Correlation and Width Centroid Contour Distance for finger based biometrics" in *Expert Systems with Applications*, **41**, 2014, pp. 3367–3382.

Well-Timed pattern recognition in Go gaming automation

Arturo Yee and Matías Alvarado

***Abstract*—**Nowadays, the formal analysis of the board game of Go is paradigmatic for computer science, particularly for learning automation. Proficiency of neural networks algorithms for pattern recognition of Go tactics and basic strategies is quantified in this paper. Neural networks pattern recognition of Go *eyes, ladders* and *nets* is the best by the early and middle stages of a Go match. As well, pattern recognition is on the base for *a-priori-knowledge*-based Go gaming strategies for *reduction* and *invasion*, in efficient manner, during the mentioned match steps. Complementary, by the end of a match the use of Monte-Carlo Tree Search for Go gaming automation can be opportune. Test simulations and the quantitative analysis of our claims are given.

***Keywords*—**Go Gaming, Tactics Pattern Recognition, Strategies Building, and Neural Network learning.

I. Introduction

THE formal analysis of the board game called Go is in the core of computer science in progress, likewise the Chess analysis was during the 20th century [1, 2]. Go is a top complex board game and currently, the deployment of learning algorithms for Go gaming automation is a central challenge for computational intelligence to demonstrate sufficient skill to beat the top human Go masters. The Go game official board (*goban*) is a 19 × 19 grid for two players using black-stones versus white-stones with zero-sum, deterministic, and perfect information [3]. By turn, each player places one black/white stone on one empty intersection or point of the board. Black plays first and white receives a compensation *komi*, by playing the second turn [4]. The goal of Go gaming is to control as much of the board area as possible.

To determine the available moves during an automated Go match is a tough problem because of the huge search space to assess. The complexity of computer Go gaming is measured by the game tree size and the state space is quite descriptive. The game tree is the cardinality of the set of all possible manners for a legal sequence of moves, through all Go matches. A Go gaming state (node) is a particular board arrangement, i.e., the positions of the stones on the board at a specific moment in a match. The size of the Go game tree is around 10^{360}: 10^{172} for the state space and up to 361 legal moves [5]. Therefore, the search space on Go gaming solutions is huger –very much– than that for Chess [3]. The moves of a Go match are depicted graphically by a *decision tree* that records the moves and is an element of the game tree. The root state is the match beginning. Any node children are those positions reachable in one move.

In this paper, we quantify the efficiency of pattern recognition by neural networks (NNs) during an automated Go match. The best pattern recognition of tactics and strategies for Go gaming is achieved during the first and middle steps of a match. In these steps, the *a-priori-knowledge*-based moves are very efficient for Go gaming. In a broader computer science perspective, the correct solution of Go gaming automation may enlighten solutions in diverse fields, such as complex network analysis, fractal formation, and bioinformatics. The entire relationship on this issue is beyond the scope of the present paper, but given its relevance, it is outlined briefly in the Discussion section.

A. Go Gaming

In the Go board, a *liberty* of a stone is any vertical or horizontal unfilled adjacent point to the stone, sometimes shared with other stones. Once a stone is placed on the board it can be removed just when it is captured, that happen if it is surrounded by adversarial stones and have lost all its liberties. Black player captures white stones and conversely. Two or more same color stones joint by the horizontal or vertical points form a *chain stone* that cannot be late divided; the diagonal adjacent stones are not in a chain. From now on the term stone refers to both single and chain but explicit difference is made if required. Any *stone is alive* if cannot be captured, and is *dead* if cannot avoid capture. When a player places a stone that will result in an immediate capture is a suicide, what is not allowed. The game ends when both players pass, then the score is computed based on both occupied territory by the player and the captured adversarial stones, so wins who adds the largest result [6].

The Go **basic tactics** of eyes, ladders and nets are used to dominate a local area [7], see Fig. 1. An *eye* is an empty single point being enclosed by same color stones, and cannot be occupied by an adversary's stone due to suicide rule. Two eyes inside a stone make very difficult its capture. A stone having only one liberty is in *Atari*. A *ladder* results from a sequence of moves that forces an adversary's stone into atari. A *net* is a set of stones (not always a chain) that surrounds an adversary's

Arturo Yee is Ph. D. students in Computer Science Department at CINVESTAV-IPN. México City, México. Phone +52 55 5747 3756 Ext. 6555; e-mail: ayee@computacion.cs.cinvestav.mx.

Matías Alvarado is a research scientist in Computer Science Department at CINVESTAV-IPN. México City, México. Phone +52 55 5747 3756 Ext. 6555, e-mail: matias@cs.cinvestav.mx.

ISBN: 978-1-61804-256-9

stone such that could eventually be captured [8]. All these *a-priori*-k*nown* Go basic tactics patterns should be recognized for a fair Go gaming, and are used for training the neural network for learning on their recognition.

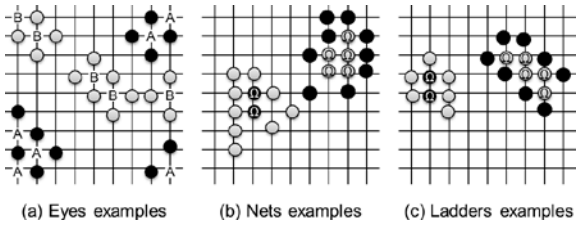


Fig. 1 Go basic tactics: (a) eyes, unavailable points – A is black and *B* is white; (b) Ω stones are surrounded by a net and may soon be captured; (c) Ω stones are in atari by ladders

For broad territory control, Go **strategies** follow a set of planned actions, deployed partly by using the aforementioned tactics as elements. Basic strategies are invasion, reduction, connection, and capture [7] (see Fig. 2). An *invasion* strategy places a stone near friendly stones, in an area where the adversary's stones look likely to dominate. A *reduction* strategy places a stone near friendly stones, to connect them if needed, in an area likely to be occupied eventually by the adversary. *Capture* reduces the liberties of an adversary's stone to zero and removes it from the board.

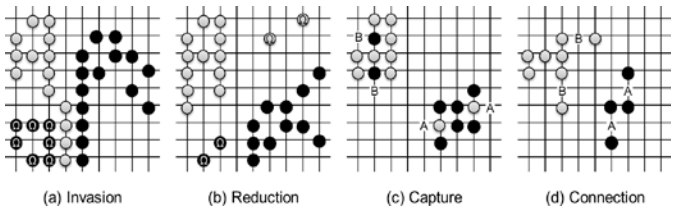


Fig. 2 Go basic strategies: (a) Ω stones perform invasion in territory dominated by white; (b) Ω black/white stones perform reduction in territory of white/black dominance; (c) black/white playing in positions *A/B* capture white/back stones; (d) black/white playing in positions *A/B* perform connections with friendly stone

A Go gaming strategy move, from the root node to the leaves nodes, is aiming to win a match efficiently. Despite the disarming simplicity of the Go rules, Go gaming conceals a huge combinatorial complexity [5, 9] (see Table I) and therefore, the big complexity is to set up an efficient strategy for playing Go**.** The **state space complexity** is the number of all the possible arrangements of the game board, which in a 19 × 19 board is about $3^{19\times19} \approx 10^{172.24}$ for Go, whereas it is 10^{50} for Chess and 10^{18} for checkers. The branching factor for Go ranges from 200–300 possible moves at each player's turn; for Chess, the range is 35–40 moves. The **game tree size** is the total number of different matches that can be played and for Go that is $\approx 10^{360}$ (chess $\approx 10^{123}$ and checkers $\approx 10^{54}$). Even on the 9 × 9 board size, the state space and the game tree size is astronomically large [10].

Table I. COMPLEXITY OF GO, CHESS, AND CHECKERS GAMES

Game	*Board size*	*State space*	*Game tree size*
Go	19 × 19	10^{172}	10^{360}
chess	8 × 8	10^{50}	10^{123}
Checkers	8 × 8	10^{18}	10^{54}

B. State of the Art

The gaming level of ongoing Go automated player is not great successful versus top human Go masters yet [11], contrasting with the achieved by chess automated players. Best Go gaming automation can beat middle level human Go players nowadays [12, 13]. However, the deployment of efficient Go gaming automation is being strengthened. To advance this effort, we quantify the performance of automated pattern recognition for Go basic tactics, such as eye and ladder, and based on the tactics recognition, we perform smart reasoning on Go basic strategies, such as reduction and invasion, both at the early and middle stages of a Go match.

The relevant advances by the Monte Carlo Tree Search (MCTS), applied to overcome the huge complexity of Go gaming, should be complemented in order to achieve victory over top level Go masters. Methods focused on simulation-based search algorithms [12, 14, 15] behave very randomly in the early stages of a Go match and produce high search complexity in choosing the next Go moves, because of the huge set of free board positions at this step. In contrast, using pattern recognition and *a-priori*-known movements, sets the basis for efficient Go strategies/tactics gaming. The Go game is based on *long-term influence moves* [12]. Moves made in the beginning of the match affect the outcome of later moves and thus, this highlights the relevance on making the correct decisions early in the Go match. MCTS is particularly free from expert knowledge and from tactical-solving guidance [16], and the memory of previous games' moves is based on a huge number of simulations [12, 14] that is very costly in time during the early stages of a match. Furthermore, in specific situations, it prevents the identification of any correct move because of the lack of a tactical search; it needs too many simulations per move to achieve an appropriate gaming move [16, 17]. Hence, an *a-priori-knowledge*-based method for movement selection is appropriate at this step.

One very difficult task for Go automation is the evaluation of non-final positions for estimating the potential of occupied territory [4, 18, 19]. Prospective methods for programming Go gaming are being deployed in simulation-based search algorithms, heuristic searches, machine learning, and automated knowledge-based decision making [20]. The main challenge in Go gaming automation is to deal with the huge number of forms in the board, which must be classified prior to deciding on the next correct Go move. Thus, the automation of Go strategies is hugely complex. The process of tactics pattern recognition is essential in learning how to make a Go move [21]. For our quantitative analysis we use:

- A neural network for Go tactics pattern recognition and basic strategy construction.
- A MCTS for move automation in the end stages of a Go

ISBN: 978-1-61804-256-9

match.

The rest of this paper is organized as follows. Section 2 reviews Go automation using NNs and MCTS. Section 3 describes pattern recognition for Go tactics and strategies and Section 4 presents experiments. Section 5 provides the analytical comparison of performance. Section 6 is the Discussion, followed by the conclusions in Section 7.

II. Go Gaming Automation Fundaments

The ability of NNs to find hidden relationships among the input-output mapping of pattern data makes them sufficiently powerful to deal with huge amount of combinations of forms, such as the ones emerging in Go gaming automation. Complementarily, the MCTS working on convenient search space is truly efficient for the automation of a match end step.

A. Neural Networks

The classic back-propagation NN for training on pattern recognition uses supervised learning to adjust the connection weights and enable the recognition of complex patterns. The topology of a multilayer NN is shown in Fig. 3. We use NNs for Go tactics pattern recognition as the basis for deciding the next Go actions, based on the given patterns of atari and capture conditions and on the analysis of the current state of a match.

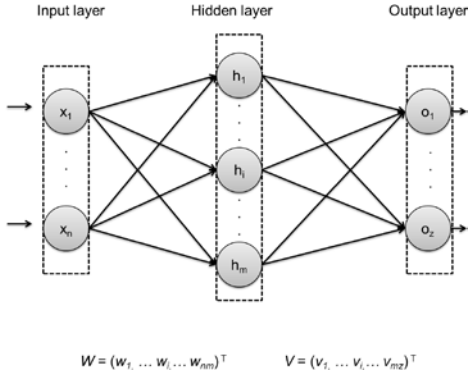


Fig. 3 example of multi-layer NN topology

For Go gaming on small boards, symbiotic adaptive neuro-evolution uses neural networks combined with evolutionary algorithms to determine the good/bad moves for a win/loss [18]. The automated Go player *Honte* [22] uses NNs in conjunction with supervised temporal differences (TD) for learning local shapes, and additional NNs for performing estimates of the value of the territorial potential. In [23], an automated Go player based on a back-propagation NN presents an architecture for learning a model from training data.

TD learning using simulation-based searches has been applied to reinforcement learning for Go gaming automation in two phases: learning and planning [13]. During the learning step, the player improves the defined policy as follows. Each node (state) value is updated from both the MCTS simulations and the value function approximation and bootstrapping from the current node to the match end; the mean outcome of simulated episodes from real Go matches is used to value each node in a search tree and between related nodes. A TD search has been applied on Go matches over 9 × 9 boards by using naive binary features matching simple patterns of stones. Complementarily, during the planning step, the policy is improved by performing iterative simulations that start at each node [24]. A major drawback, however, is that MCTS for Go gaming automation needs too many simulations per move to obtain good results. Therefore, to overcome the lack of efficiency, the TD search integrates Go *a-priori-knowledge* to decide the next moves.

A Go machine learning approach has been developed that focuses on an evaluation function regarding scalability, from the library of local tactical patterns, the integration of patterns across the board, and the size of the board itself [25]. The automated learning is on local patterns from a library of games, by means of a recursive probabilistic Bayesian NN, the outputs of which represent local territory ownership probabilities. A combination of NNs, particle swarm optimization, and evolutionary algorithms has been used to train a board evaluator from zero knowledge [26]; the hybrid algorithm provides an evaluation of the game board through self-play. The authors claim that after experiments against the benchmark game of Capture Go, the hybrid algorithm includes and overcomes the power from the parts.

B. Monte Carlo Tree Search

Monte Carlo methods are rooted in statistical physics for the modeling of stochastic phenomena [15]. In Go gaming automation, MCTS is a best-first search technique, using stochastic simulations to estimate the value of the moves and thus, adjust the policy towards a best-first strategy [15]. MCTS simulation values the nodes in a search tree that is the partial game tree being progressively built. The building of a tree is performed by following the *selection*, *expansion*, *simulation,* and *back-propagation* mechanisms [17].

The well-known Olga and Oleg Go player automation is a pioneer on the application of the Monte Carlo approach [27]. To evaluate a move from a current state *s*, many self-play simulations are performed and the value of *a* is the average value from the outcome of the simulations using a uniform random policy. Progressive pruning and the all-moves-as-first heuristic allows more rapid gaming without decreasing the Go player level. The Rapid Action Value Estimation (RAVE) extension from the MCTS algorithm [14] shares the value of actions across each sub-tree of the tree search, and the heuristic MCTS uses a function to initialize the value of the new positions in the tree search.

III. Tactics and Strategies

In order to create a robust and reliable network, sometimes, random noise is added to training data [7]. *Don´t care* symbols are replaced by 2s, 1s or 0s in each training stage in order to preserve conditions to be a true training set for eyes, ladders,

and nets, respectively. The NN error is obtained from the difference between the output from the training data and the target during iterative steps. The error is fed back repeatedly to the previous layers to modify the connections weight of nodes until a predefined tolerance or number of epochs is achieved. The NN activation of the nodes is by the sigmoid function. The number of neurons in the hidden layer is obtained experimentally and the output layer indicates the recognized patterns.

A. Tactics Pattern Recognition

For the process of pattern recognition analysis, the Go board is segmented into a *window view size* of 3 × 3 in order to identify eyes patterns. A *window view size* of 5 × 5 is used for identifying ladders and nets patterns, and as a result of the neighboring combination of these windows, bigger ladders and nets can be recognized. The NN layer of the input receives a set of board occupied points; 9 for eyes and 25 for a ladder or a net. During the training stage, the training patterns include *don´t care* symbols to represent those points that can be replaced regardless of their value. It is valuable to include *don't care* patterns in Go tactics recognition because of the non-deterministic nature of Go gaming, see example in [2].

To start the process of Go tactics pattern recognition, the positions from the 3 × 3 and 5 × 5 *windows view size* are encoded into a vector that feeds the network. When using pattern recognition to identify Go tactics in a match, the main difficulty concerns verifying that a shape really corresponds to an eye, ladder or net. The NN should check the conditions to authenticate whether the detected shape is a true Go tactic.

For eye pattern recognition, the NN tries to find similarities with the given input to any of the shapes: edge, lateral or normal eyes. If high similarities exist then the conditions of eye must be checked, i.e., there must be an empty point of space surrounded by friendly stones such that no adversary's stone may be set upon it. These conditions are verified outside the NN using verifier conditions. As in the case of eye patterns, the same procedure is applied for ladder and net pattern recognition, but with the proper ladder and net conditions.

B. Building of Strategies from Pattern Recognition

Once the Go tactics patterns such as eyes, ladders or nets, are recognized, as well as the Go gaming strategies of invasion, reduction, connection or capture, offensive/defensive strategies can be employed (see Fig. 4). Hence, based on the tactics pattern recognition, deployment heuristics for suitable defensive/offensive Go *a-priori-knowledge* strategies are available to be applied during the initial and middle steps of an automated Go match. Strategies of reduction and invasion as well as defensive strategies, address saving stones in atari or are devoted to augment the liberties of ally stones. Strategies can be constructed following the next statements, as illustrated in Fig. 5.

Defensive strategies:

- Save a stone in atari by close placement of an ally stone that eventually connects and saves it, or increasing *liberties*.
- For preventing a stone falling within risk of be captured by the adversary's next moves.

Offensive strategies:

- Interrupt the formation of adversary's eye by placing a new stone.
- Reduce *liberties* to adversary's stones, eventually placing it in atari.
- Play close to own stones, sets of stone(s) or close to stone(s) with two or more eyes to ensure high possibilities of making *connections* and *spreading* of stones.
- Capture adversary's stones by placing adversary's stones in or close to atari.

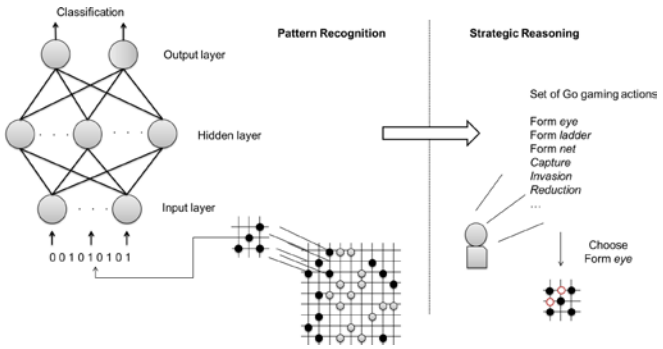


Fig. 4 example pattern recognition of a possible eye

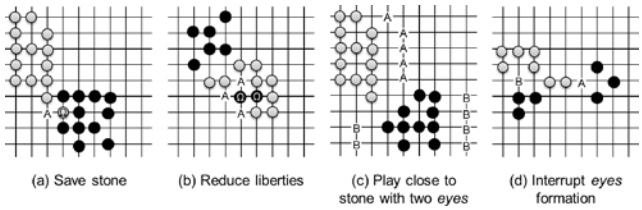


Fig. 5 Go gaming strategies: (a) Ω is in atari, but playing in point A makes a connection for saving Ω; (b) playing in points A reduces Ω liberties; (c) white/black playing to points A/B increases dominance area; (d) white/black playing to points A/B interrupts the formation of adversary's eye

In the latter stages of a Go match, an MCTS-based move becomes a suitable option. This is because the size of the search space has become small and the automated pattern recognition is too difficult to perform over the complex patterns on the board with the few free board positions. Actually, in the latter stages of a Go match, the deployment of *a-priori*-known strategies is hard because the free board points are too restrictive, and few board spaces make it difficult to deploy strategies. Under this circumstance, the gaming method is to perform an MCTS evaluation to play any of the free board positions. Hybrid approaches with Go *a-priori-knowledge*-based strategies are easy to deploy in the initial and middle stages of the game when few board points are occupied. By the end stages, the use of MCTS is better suited to choosing a move on the empty board positions. A description of the hybrid Go players that prove our claim is

ISBN: 978-1-61804-256-9

given in the following.

C. Go Players Simulator

Our gaming simulator compromises a set of automated Go players, using either random, or pattern recognition, or strategic reasoning, or MC-Rave methods. In addition, it has a graphic interface and uses Go Text Protocol for communication with other automated players in KGS [28] or CGOS [29] Go servers (see Fig. 6).

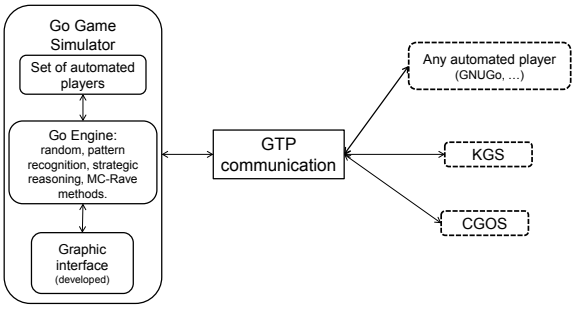


Fig. 6 components of Go gaming simulator

In the present proposal, using MCTS simulations, all actions have the same probability and thus, a uniform distribution is used.

D. The Hybrid Go Players Description

The hybrid Go player SP_1 combines pattern recognition and strategic reasoning with the MC-Rave approach [14]; this SP_1 hybrid use known experience.

GNUGo is a classic and powerful automated open source Go player, the first with huge impact in Go automation [30] and therefore, a comparison versus GNUGo is obligated. GNUGo uses pattern matching algorithms to analyze the stones patterns on the board and then proposes the next moves [24]. In addition:

- GNUGo rates the shape formed by a move and the value of local moves on specific territory,
- GNUGo calculates the max/min value obtained by a move using pattern matching.

Because the GNUGo Go player strength is in the criteria used to evaluate a move, our next proposed hybrid player SP_2 is based on GNUGo. SP_2 uses GNUGo criteria to play during the first 10 moves and then shifts to pattern recognition for the subsequent 10 moves, on average, and so on. The reason for doing 10 moves each is because the match stages are determined by the number of moves played by each player. Statistically, for a 9 × 9 board size, the number of moves is around 40 per player and therefore, on average, each early, middle or late stage in a match is scoped by 13 moves per player. Similar calculus works for a 19 × 19 size board.

IV. Timely Pattern Recognition

The average recognition performance obtained by different numbers of hidden neurons. In order to surpass the major difficulty for Go tactics patterns recognition, given the wide variety of shapes in a Go match; we fix on five for the number of neurons in the hidden layer of the multi-layer NN used. This way, efficient pattern recognition and learning is our best performance. Furthermore, the training time is short enough to avoid over-learning, which produces noise and/or redundancy. The test description to recognize eyes, ladders and nets on the Go board follows.

Eyes *patterns.* 900 tests divided into groups of 150 were performed. 400 eyes positive examples and 100 not-eyes negative examples were used to test the NN. The mean, minimum and maximum accuracy obtained for each group from the eyes/no-eyes examples and the mean accuracy of the NN are reported. The average accuracy is above 70%, as shown in Table II.

Table II. Results of NN eyes (A = **MEAN**, B =**MIN**, C= **MAX**)

Number of tests	Eyes *examples*			*No*-eyes *examples*			*Total classification* Eyes + *No*-eyes
	(Correct classification cases %)						
	A	B	C	A	B	C	A
1-150	83.68	80	90	67.58	58	74	75.77
151-300	83.24	80	88	68.04	60	74	75.64
301-450	83.58	80	91	67.72	56	74	75.65
451-600	83.78	80	91	67.45	60	74	75.62
601-750	83.62	80	90	67.69	60	74	75.65
751-900	83.32	81	88	67.96	60	74	75.64

Ladders and Nets patterns. A similar analysis for ladders and nets was performed by running 900 tests. Each test uses 100 ladders-nets positive examples and 100 no-ladders-nets negative examples. 75% accuracy is obtained (see Table III).

Table III. Results of NN Ladders and Nets (A = **MEAN**, B =**MIN**, C= **MAX**)

Number of tests	Ladders-nets *examples*			*No*-ladders-nets *examples*			*Total classification* Ladders-*net* + *No*-ladders-*nets*
	(Correct classification cases %)						
	A	B	C	A	B	C	Mean
1-150	86.96	60	100	56.06	21	75	71.51
151-300	87.19	61	100	56.16	30	75	71.67
301-450	86.70	48	100	56.89	26	90	71.79
451-600	85.88	29	100	57.58	25	100	71.73
601-750	86.34	43	100	56.83	25	90	71.58
751-900	87.78	63	100	55.09	22	80	71.44

The experimental results proved certain improvements on the NNs effectiveness to recognize complex patterns of Go gaming. Actually, the pattern recognition is through a wide variety of shapes, sometimes too complex and not obviously true Go tactics. The results in Tables II and III show the NNs accuracy. Around 70% efficiency recognition is a good result because of the complexity of these kinds of patterns. The huge amount of forms that occur in a Go gaming match makes the

tactics patterns recognition a difficult task. Even though the recognized patterns correspond to tactics that are significant for the proposed Go strategic analysis, approach [31] tries to determine Go patterns in game records like *edge* and *corner* patterns, but few of them represent proper Go tactics patterns.

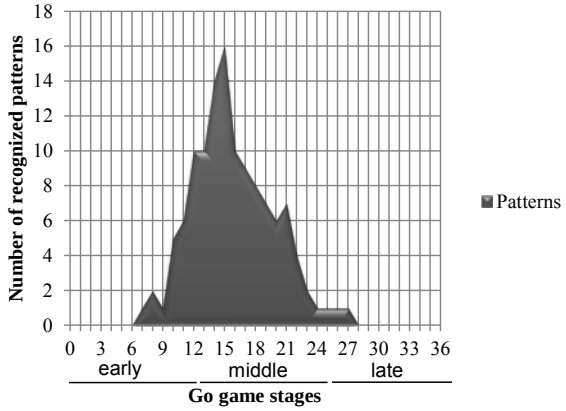


Fig. 7 number of patterns recognized through stages in a Go match

Fig. 7 shows that the highest number of recognized patterns occurs in the match middle stage. In the early and middle stage, based on the Go gaming *a-priori-knowledge* of these states, the pattern recognition and strategic reasoning work, thus a better strategic analysis of offensive/defensive Go actions is available. But when we lack of information or the board free positions arises too restrictive, that correspond to typical circumstance in late stage, the usage of MCTS on the set of free positions do choice the best to play in. Numbers in *x* axis represent Go match turns for both players.

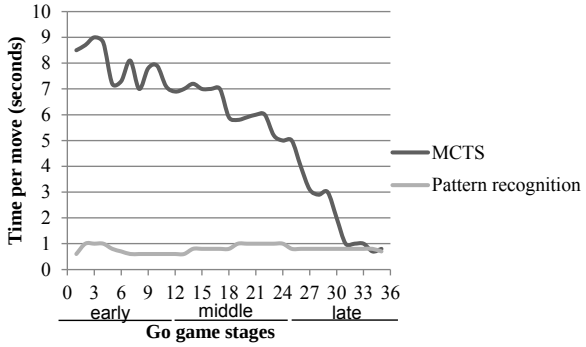


Fig.8 elapsed time per move throughout stages of a Go match

The Fig. 8 shows the elapsed time per move per player throughout the stages of a Go match. Time required per move using pattern recognition is too low and almost constant during the first and middle stage, but strongly increases in the late stage since the difficulty to recognize any pattern on the board in this stage. On the opposite, using MCTS, time spent for doing a move is too high in the early stage, comes down in the middle stage and is truly short in the late stage. Reason is that the size of search space the algorithm works at the early stage is huge, so applying MCTS is expensive and waste a lot of time; in the late stage of a Go match the search space size is small so quick to apply MCTS.

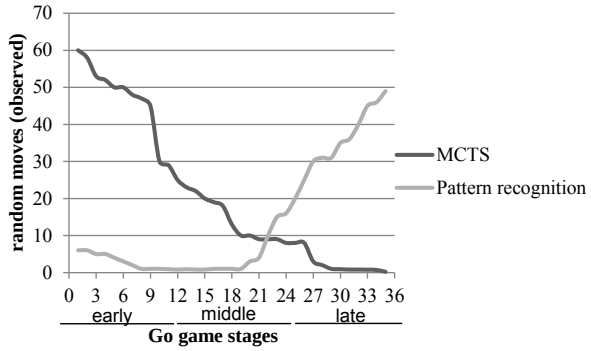


Fig. 9 elapsed time per move throughout stages of a Go match

The Fig. 9 shows the percentage observed of random moves in 100 simulations. In the late stage MCTS movements are suitable since each position is quick to evaluate, so the best one up to the method is selected. In the early and middle stages MCTS is time spending, but not random pattern recognition supporting *a-priori*-known strategic Go movements are ease to apply.

From experiments described, we can conclude that by the early and middle Go match stages is the best time to apply *a-priori-knowledge* for the pattern recognition of tactics and strategies, so efficient gaming is achieved this way on these match steps. On the opposite, because in a Go match early stage the board free points are pretty numerous and any movement by MCTS tends to be random, the huge size of search space for processing is time spending. Hence, at this match early stage MCTS is not an efficient technique but the computer resources are quite waste. Henceforth we propose the systematic usage of hybrid Go automated players for achieving efficient performance during matches.

V. Go Players Performance

The performance comparison of automated Go players and the analytical comparison of approaches that use NNs for Go automation follow. The compared Go automated players are ***SP₁*** that uses pattern recognition of tactics, strategic reasoning and MC-Rave [14], the **Strategic player** that uses pattern recognition of tactics and strategies, and the **MC-Rave player** that uses the method in [14].

A. SP_1 Comparison

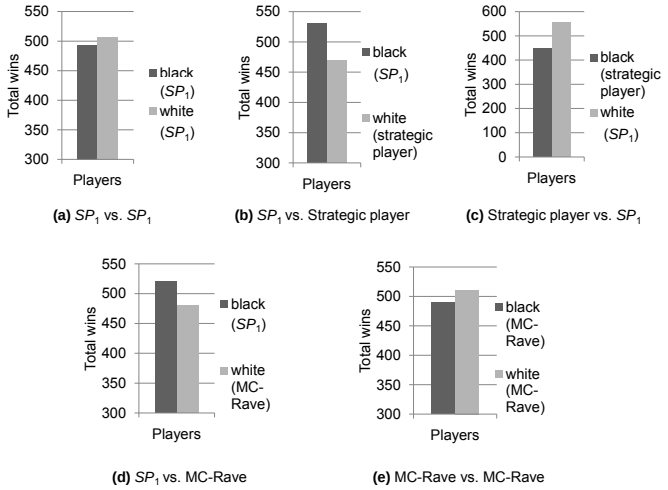
1000 tests on a board size of 9 × 9 of the following combinations were performed to make a comparison among them:

1) SP_1 vs. SP_1,
2) SP_1 vs. Strategic player,
3) Strategic player vs. SP_1,
4) SP_1 vs. MC-Rave,
5) MC-Rave vs. MC-Rave.

In Fig. 10, the performance of the automated players from 1000 simulations is shown. Black SP_1 vs. white SP_1 is 49.3%/50.7% of wins rate (see Fig. 10(A)). Black SP_1 vs.

white Strategic player is 53%/47% of wins rate (see Fig. 10(B)). Black Strategic player vs. white SP_1 is 44.7%/55.3% of wins rate (see Fig. 10(C)). Black SP_1 vs. white MC-Rave player is 52%/48% of wins rate (see Fig. 10(D)). Black MC-Rave player vs. white MC-Rave player is 48.9%/51.1% of wins rate (see Fig. 10(E)).

As shown in the results, SP_1 overcomes the other automated Go player's performances by applying different techniques in the early, middle, and late stages of the Go match.



(a) SP_1 vs. SP_1 **(b)** SP_1 vs. Strategic player **(c)** Strategic player vs. SP_1

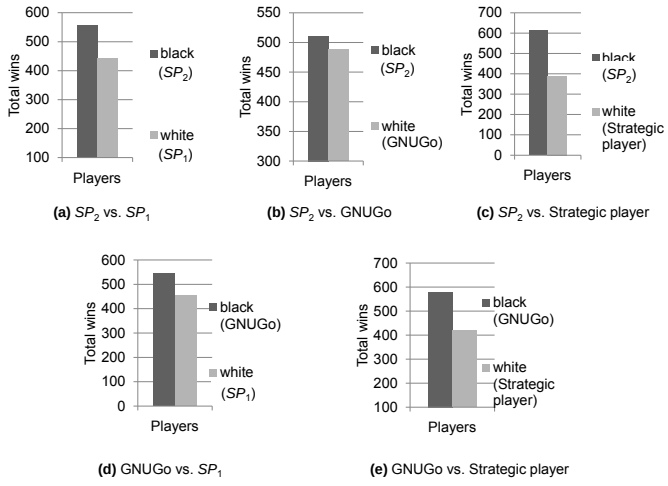
(d) SP_1 vs. MC-Rave **(e)** MC-Rave vs. MC-Rave

Fig. 10 first set of performance results of automated Go players

B. SP_2 Comparison

1000 tests on a board size of 9 × 9 using the following players were performed: ***SP_2 (HybridGNUGo)*** **player** uses pattern recognition, strategic reasoning and GNUGo, **GNUGo player**, **Strategic player** and **Hybrid player** (SP_1) uses pattern recognition, strategic reasoning along with MCTS. The performance comparisons were done as follows:

1) SP_2 vs. SP_1,
2) SP_2 vs. GNUGo,
3) SP_2 vs. Strategic player,
4) GNUGo vs. SP_1,
5) GNUGo vs. Strategic player.



(a) SP_2 vs. SP_1 **(b)** SP_2 vs. GNUGo **(c)** SP_2 vs. Strategic player

(d) GNUGo vs. SP_1 **(e)** GNUGo vs. Strategic player

Fig. 11 second set of performance results of automated Go players

The automated players' performances from 1000 simulations are shown in Fig. 11. Black SP_2 player vs. white SP_1 is 55.7%/44.3% of wins rate (see Fig. 11(A)). Black SP_2 player vs. white GNUGo player is 51.1%/49.9% of wins rate (see Fig. 11(B)). Black SP_2 player vs. white Strategic player is 61.2.7%/38.8% of wins rate (see Fig. 11(C)). Black GNUGo player vs. white SP_1 is 54.5%/45.5% of wins rate (see Fig. 11(D)). Black GNUGo player vs. white Strategic player is 58%/42% of wins rate (Fig. 11(E)).

The results show that SP_2 overcomes the other automated Go player's performances, even the *GNUGo* performance, by applying different techniques in the early, middle, and late stages of a Go match.

VI. Discussion

Analysis on Go gaming automation from the complex network approach, like the one of the World Wide Web, focuses on the non-trivial topology of the network that results from a Go match [32]; the construction of a directed network given a suitable definition of related tactical Go gaming moves. By mining database matches of master level games, this approach discovered the similar patterns arising during the early stages of a Go match. In [33], the proposal for two dynamic randomization techniques is given: one for the parameters and the other for a hierarchical move generator.

Complex pattern recognition is present in Bioinformatics, which is devoted to computer and information analysis and the management of data on biological processes [34-36], particularly in determining or classifying molecular or tissue patterns as equivalent or related to some extent. Pattern recognition for Go gaming and Bioinformatics processes may advance in parallel from now on.

In computer complexity theory [37], the problems pertain to specific complexity classes by regarding certain characteristics: one major is *time*, which refers to the number of execution steps that an algorithm used to solve a problem; the other main complexity character is *space*, which refers to the amount of memory used to deal with a problem. Some complexity classes are **P**, **NP**, **PSPACE**, **and EXPTIME**. As a result of some complexity analyses of Go gaming, *experts* of the area claim that Go gaming belongs to **EXPTIME**-complete game [38], because it is an *unbounded two-player game*. *Unbounded games* are those in which there is no restriction on the number of moves that can be made. However, Go seems to be a bounded game because in each move a stone is placed, but there exist *capturing moves* that reopen spaces on the board. Papadimitriou [39], in one of his analyses, concluded that Go is a **PSPACE**-complete game.

Actually, being aware of what the adversary is doing helps to formulate defensive actions that inhibit her offensive actions. The inspired thinking that humans are capable of, as a result of observing the decisions other people make, applies in Go gaming through performing pattern recognition to decide on the next offensive/defensive move. The proposed NNs recognize forms that are Go tactics patterns and therefore, give

relevant information to strategic decision making during the early and middle stages of a Go match.

VII. Conclusion

In this approach, we proposed the use of NNs for pattern recognition during the early and middle steps of a Go match; the expert's *a-priori-knowledge* for pattern recognition of eyes, ladders and nets is efficiently translated by means of NN for Go gaming automation. Based on this pattern recognition, defensive/offensive movements, such as those involved in complex Go gaming moves, are available to build up and apply during the middle steps of the match. On the other hand, during the latter stages of the game, the use of MCTS is appropriate because of the difficulty of performing *a-priori-knowledge* strategic gaming. A relevant discussion of Go gaming analysis in a perspective of complex networks and fractals was introduced, and a mathematical modeling of a Go game was presented.

Acknowledgment

Thanks are extended to the Mexican National Council of Science and Technology (CONACyT) in relation to Arturo Yee's PhD degree grant, CVU 261089.

References

[1]. J. McCarthy, "AI as sport," Science, vol. 276, no. 5318, pp. 1518–1519, 1997.
[2]. A. Yee, A. and M. Alvarado, "Pattern Recognition and Monte-CarloTree Search for Go Gaming Better Automation," in *Proc. Advances in Artificial Intelligence*, Springer Berlin Heidelberg, 2012, pp. 11–20.
[3]. K. Chen, and Z. Chen, "Static analysis of life and death in the game of Go," Inf. Sci. Inf. Comput. Sci., vol. 121, no. 1–2, pp. 113–134, 1999.
[4]. D. B. Benson, "Life in the game of Go," Inform. Sciences, vol. 10, no. 2, pp. 17–29, 1976.
[5]. L. V. Allis, *Searching for Solutions in Games and Artificial Intelligence*. University of Limburg: The Netherlands, 1994.
[6]. A. Yee, and M. Alvarado, "Formal Language and Reasoning for Playing Go," in *Proc. Seventh Latin American Workshop on Logic / Languages, Algorithms and New Methods of Reasoning*, 2011, pp. 125–132.
[7]. N. Yoshiaki, *Strategic Concepts of Go*. Japan: Ishi Press, 1972.
[8]. J. Kim, *Learn to Play Go*. New York: Good Move Press, 1994.
[9]. E. R. Berlekamp, and D. Wolfe, *Mathematical Go: Chilling Gets the Last Point*. Massachusetts: A K Peters Ltd, 1997.
[10]. P. Drake, and Y.-P. Chen, "Coevolving partial strategies for the game of go," in *Proc the International Conference on Genetic and Evolutionary Methods*, 2008, pp. 312–318.
[11]. M. Müller, "Computer Go," Artif. Intell., vol. 134, no. 1–2, pp. 145–179, 2002.
[12]. S. Gelly, et al., *The Grand Challenge of Computer Go: Monte Carlo Tree Search and Extensions.* Communication of the ACM 55, vol. 3, pp. 106–113, 2012.
[13]. D. Silver, R.S. Sutton, and M. Müller, "Temporal-difference search in computer Go," Mach. Learn., vol. 87, no. 2, pp. 183–219, 2012.
[14]. S. Gelly, and D. Silver, "Monte-Carlo tree search and rapid action value estimation in computer Go," Artif. Intell., vol. 175, no. 11, pp. 1856–1875, 2011.
[15]. C. Browne, et al., "A Survey of Monte Carlo Tree Search Methods," IEEE Transactions on Computational Intelligence and AI in Games, vol. 4, no.1, pp. 1–43, 2012.
[16]. J. B. Hoock, et al., "Intelligent Agents for the Game of Go," IEEE, Computational Intelligence Magazine, vol. 5, no.4, pp. 28–42, 2010.
[17]. G. Chaslot, et al., "Monte-Carlo Tree Search: A New Framework for Game AI," in *Proc. Artif. Intell. Interact. Digital Entert*, 2008.
[18]. N. Richards, D.E. Moriarty, and R. Miikkulainen, "Evolving Neural Networks to Play Go," Appl. Intell., vol. 8, no.1, pp. 85–96, 1998.
[19]. E. D. Werf, H.J. Herik, and J.H.M. Uiterwijk, "Learning to Estimate Potential Territory in the Game of Go," in *Proc. Computers and Games*, H.J. Herik, Y. Björnsson, and N. Netanyahu, Editors., Springer Berlin Heidelberg: Ramat-Gan, Israel, 2006, pp. 81–96.
[20]. B. Bouzy, and T. Cazenave, "Computer Go: an AI Oriented Survey," Artif. Intell., vol. 132, no.1, pp. 39–103, 2001.
[21]. E.H.J. Nijhuis, *Learning Patterns in the Game of Go*, in *Artificial Intelligence*, Universiteit van Amsterdam: Amsterdam, 2006.
[22]. F. A. Dahl, *Honte, a go-playing program using neural nets*. Machines that learn to play games, Nova Science Publishers, 2001 pp. 205–223.
[23]. L. Jianming, Z. Difei, and L. Rui, "Improve Go AI based on BP-Neural Network," in *Cybernetics and Intelligent Systems, IEEE Conference on,* 2008.
[24]. C. Fellows, Y. Malitsky, and G. Wojtaszczyk. "Exploring GnuGo's evaluation function with a SVM," in *Proc. of the 21st national conference on Artificial intelligence*, Boston, Massachusetts: AAAI Press, 2006.
[25]. L. Wu, and P. Baldi, "Learning to play Go using recursive neural networks," Neural Networks, vol. 21, no.9, pp. 1392–1400, 2008.
[26]. X. Cai, G.K. Venayagamoorthy, and D.C.W. II, "Evolutionary swarm neural network game engine for Capture Go", Neural Networks, vol. 23, no. 2, pp. 295–305, 2010.
[27]. B. Bouzy, and B. Helmstetter, *Monte-Carlo Go Developments*. in *Advances in Computer Games*, Springer US, 2004, pp. 159–174.
[28]. *KGS Go Server*. Available from: http://www.gokgs.com/.
[29]. *CGOS*. Available from: http://cgos.boardspace.net/.
[30]. GnuGo. *GnuGo automated player*. Available from: http://www.gnu.org/software/gnugo/.
[31]. Y. Shi-Jim, et al. "Pattern Matching in Go Game Records," presented at *S*econd International Conference on Innovative Computing, Information and Control, 2007.
[32]. B. Georgeot, and O. Giraud, "The game of go as a complex network," Europhysics Letters, vol. 97, no. 6, 2012.
[33]. K.-H. Chen, "Dynamic randomization and domain knowledge in Monte-Carlo Tree Search for Go knowledge-based systems," Knowledge-Based Systems, vol. 34, pp. 21–25, 2012.
[34]. M. Hue, et al., "Large-scale prediction of protein-protein interactions from structures," BMC Bioinformatics, vol. 11, no. 1, pp. 1–9, 2010.
[35]. M. D. Ritchie, et al., "Genetic programming neural networks: A powerful bioinformatics tool for human genetics," Appl. Soft Comput., vol. 7, no. 1, pp. 471–479, 2007.
[36]. O. Wolkenhauer, et al., "SysBioMed report: Advancing systems biology for medical applications," Systems Biology, IET, vol. 3, no. 3, pp. 131–136, 2009.
[37]. A. Sanjeev, and B. Barak, *Computational Complexity: A Modern Approach*. Cambridge: Cambridge University Press, 2009.
[38]. R. A. Hearn, *Games, Puzzles, and Computation*, in *Department of Electrical Engineering and Computer Science*. Massachusetts Institute of Technology: Cambridge, 2006.
[39]. C. H. Papadimitriou, *Computational Complexity*. Reading, Massachusetts Addison Wesley, 1993.

Arturo Yee Rendón received his M. Sc. degree in Computer Science from the Computer Science Department in the Center of Research and Advanced Studies (CINVESTAV-IPN), Mexico, where nowadays is researching on his Ph. D. thesis focus on strategic reasoning for games playing. He received his bachelor degree in Computer Science from the Autonomous University of Sinaloa, Mexico, being distinguished with the Best Student Award in his career promotion.

Matías Alvarado is a research scientist in the Computer Science Department of the Center of Research and Advanced Studies (CINVESTAV-IPN), México. He is a member of the Mexican Academy of Sciences. He is doctor in Science Mathematics from the Technical University of Catalonia. His current major research interest is the mathematical modeling of strategic reasoning and the computer simulation of strategic decision making. He has published around 70 papers, most of them on Decision Making methods, the last ones focusing in selection of strategies in sports and board games.

Integral criterion of the stability of the second order linear D-equations with oscillatory coefficients

A.A. Mukhambetova, Zh.A. Sartabanov

Abstract**—** Established the sufficient conditions of stability and instability of the solutions of the second order linear D-equations with oscillatory coefficients.

Keywords**—** D-equations, oscillating coefficients, function space, stability of solutions.

I. Introduction

FUNDAMENTALS theory of the stability solutions of differential equations systems are incorporated in A. Lyapunov's fundamental paper. One of the fundamental results of this paper is called Lyapunov's integral criterion $[1, p.202]$ concerns the stability of the second order linear ordinary differential equations with oscillatory coefficients. This feature doesn't have a counterpart in the quasi-periodic case. It is generalized to the multiperiodic case for the second order linear D-equations in this note. The previously introduced concepts and research methods use for this purpose [2, 3]. The corresponding results obtained for the second order linear ordinary differential equations with quasi-periodic coefficients by switching to the main diagonal space of the independent variables. Thus, by the theory of the existence the real smooth branches of the logarithm of functional matrices defined on the multidimensional torus [4].

II. Statement of the problem

Consider the following linear equation

$$D^2 x = p(t,\varphi,\psi)\, x \tag{1}$$

with the differential operator $D = \frac{\partial}{\partial t} + \sum_{k=1}^{m} \frac{\partial}{\partial \varphi_k}$, where $D^2 x = D(D x)$, $t \in (-\infty, +\infty) = R$, $\varphi = (\varphi_1, ..., \varphi_m) \in R \times ... \times R^m$, $\psi = \varphi - e\,t -$ is characterization of the operator D, $e = (1, ..., 1) - m -$ vector, $p(t,\varphi,\psi) -$ is the given function with the properties of periodicity and smoothness:

$$p(t+\theta, \varphi + k\,\omega, \psi + k\,\omega) \equiv p(t,\varphi,\psi) \in \in C^{(0,1,1)}_{t,\varphi,\psi}(R \times R^m \times R^m), \forall\, k \in Z^m, \tag{2}$$

$\theta, \omega_1, ..., \omega_m$ - is rationally incommensurable periods, $k\,\omega = (k_1\omega_1, ..., k_m\omega_m)$ - is multiple vector-period.

Note, that the dependence of the coefficient $p(t,\varphi,\psi) = p(t,\varphi,\varphi - et)$ doesn't have the property of periodicity on the variable t, although $p(t,\varphi,\psi)$ becomes to quasi-periodic function along diagonals $\varphi = et$.

Equation (1) can be expressed as an equivalent system

$$D\, z = P(t,\varphi,\psi)\, z, \tag{3}$$

where $z = (z_1, z_2)$, with $z_1 = x, z_2 = D\,x$, $P(t,\varphi,\psi)$ - is matrix of the form

$$P(t,\varphi,\psi) = \begin{bmatrix} 0 & 1 \\ -p(t,\varphi,\psi) & 0 \end{bmatrix},$$

which by (2) has the property

$$P(t+\theta, \varphi + k\,\omega, \psi + k\,\omega) \equiv P(t,\varphi,\psi) \in \in C^{(0,1,1)}_{t,\varphi,\psi}(R \times R^m \times R^m), \ \forall\, k \in Z^m \tag{4}$$

We introduce the function space U *of* $\omega -$ periodic and continuously differentiable vector-functions $u(\varphi) = (u_1(\varphi), u_2(\varphi))$ at $\varphi \in R^m$ with the norm defined by the supremum of the module:

$$U = \left\{ \begin{array}{l} u(\varphi) : u(\varphi + k\,\omega) = u(\varphi) \in C^{(1)}_{\varphi}(R^m), \forall\, k \in Z^m; \\ \|u\| = \sup\limits_{R^m} |u(\varphi)| \end{array} \right\}, \tag{5}$$

where $|u| = \sqrt{u_1^2 + u_2^2}$.

The system (3) admits an unique solution $z = z(t,\varphi,\psi,u(\psi))$ by (4) for each $u(\varphi) \in U$ satisfying the initial condition $z\big|_{t=0} = u(\varphi)$ and this solution belongs to the

space U for each value t. We will consider the solution of the system (3) with the initial data of the functional space (5).

The solution $z^* = z(t,\varphi,\psi,u^*(\psi))$ of the system (3) is called stable if for $\forall \varepsilon > 0$ we can specify $\delta = \delta(\varepsilon) > 0$ such that

$$\left| z(t,\varphi,\psi,u(\psi)) - z(t,\varphi,\psi,u^*(\psi)) \right|_t =$$
$$= \sup_{\psi \in R^m} \left| z(t, et+\psi, \psi, u(\psi)) - z(t, et+\psi, \psi, u^*(\psi) \right| < \varepsilon$$

for $t \geq 0$, as soon as $\left\| u(\varphi) - u^*(\varphi) \right\| < \delta$.

We pose the problem of investigating the stability of the zero solution of system (3) with the condition (4) on the basis of a generalization for the case of Lyapunov's integral criterion of D-equations for the second order linear ordinary differential equations with periodic coefficients.

III. Generalization Lyapunov 's constant on D- system and stability theorem

Let $X(t,\varphi,\psi)$ - is matriciant of system (3). Hence, by (4) have the properties:

$$DX(t,\varphi,\psi) = P(t,\varphi,\psi)X(t,\varphi,\psi),\ X(0,\varphi,\varphi) = E,$$
$$X(t,\varphi+k\omega,\psi+k\omega) = X(t,\varphi,\psi) \in C^{(1,1,1)}_{t.\varphi,\psi}(R \times R^m \times R^m),$$
$$\forall k \in Z^m,$$

$$X(t+\theta,\varphi,\psi) = X(t,\varphi,\psi) \cdot X(\theta,\psi,\psi), \tag{6}$$

where E- is dimensional identity matrix, the matrix $X(\theta,\psi,\psi)$ is called the monodromy matrix system (3), and eigenvalues $\rho = \rho(\psi)$ are called the multiplier of matrix, defined by the equation

$$\det[X(\theta,\psi,\psi) - \rho E] = 0 \tag{7}$$

Obviously, that the matriciant $X(t,\varphi,\psi)$ is represented:

$$X(t,\varphi,\psi) = \begin{bmatrix} \xi(t,\varphi,\psi) & \eta(t,\varphi,\psi) \\ D\xi(t,\varphi,\psi) & D\eta(t,\varphi,\psi) \end{bmatrix}, \tag{8}$$

where $\xi(t,\varphi,\psi)$ and $\eta(t,\varphi,\psi)$ are linear independent solutions of equation (1) satisfying the conditions $\xi(0,\varphi,\varphi) = 1, D\xi(0,\varphi,\varphi) = 0$ and $\eta(0,\varphi,\varphi) = 0, D\eta(0,\varphi,\varphi) = 1.$

Solutions $\xi(t,\varphi,\psi)$ and $\eta(t,\varphi,\psi)$ will obtain in the convergent series form by Lyapunov:

$$\xi(t,\varphi,\psi) = 1 - \int_0^t (t-t_1)\, p(t_1, \psi+et_1, \psi) d\, t_1 +$$
$$+ \int_0^t (t-t_1)\, p(t_1, \psi+et_1, \psi)\, dt_1 \cdot \tag{9}$$
$$\int_0^{t_1} (t_1-t_2)\, p(t_2, \psi+et_2, \psi) dt_2 - \ldots$$

$$\eta(t,\varphi,\psi) = t - \int_0^t (t-t_1)\, p(t_1, \psi+et_1, \psi) t_1 dt_1 +$$
$$+ \int_0^t (t-t_1)\, p(t_1, \psi+et_1, \psi) dt_1 \cdot \tag{10}$$
$$\int_0^{t_1} (t_1-t_2)\, p(t_2, \psi+et_2, \psi) \cdot t_2 dt_2 - \ldots$$

Given that $Sp\, P(t,\varphi,\psi) = 0$ we have

$$\det X(\theta) = \det X(0,\varphi,\varphi) \exp\left[\int_0^\theta SpP(s, \psi+es, \psi) ds \right] = 1. \tag{11}$$

Then, we obtain from equation (7) by (8) - (11)

$$\rho^2 - a\rho + 1 = 0, \tag{12}$$

where $a = a(\psi)$ - is the function defined by

$$a(\psi) = \xi(\theta,\psi,\psi) + D\eta(\theta,\psi,\psi) = S\rho\, X(\theta,\psi,\psi) \tag{13}$$

is the analog of Lyapunov's constant for D-system.

We have from (10)

$$D\eta(t,\varphi,\psi) = 1 - \int_0^t t_1 p(t_1, \psi+et_1, \psi) dt_1 +$$
$$+ \int_0^t p(t_1, \psi+et_1, \psi) dt_1 \int_0^{t_1} (t_1-t_2) t_2\, p(t_2, \psi+et_2, \psi) dt_2 -$$
$$- \int_0^{t_1} p(t_1, \psi+et_1, \psi) dt_1 \int_0^{t_1} (t_1-t_2)\, p(t_2, \psi+et_2, \psi) dt_2 \cdot \tag{14}$$
$$\cdot \int_0^{t_2} (t_2-t_3) t_3\, p(t_3, \psi+et_3, \psi) dt_3 + \ldots$$

Therefore, we obtain the constant Lyapunov's function on the diagonal by (9), (14) from (13):

$$a(\psi) = 2 - \theta \int_0^{\theta} p(t_1, \psi + et_1, \psi) dt_1 + \int_0^{\theta} dt_1 \cdot$$

$$\cdot \int_0^{t_1} dt_2 (\theta - t_1 + t_2)(t_1 - t_2) p(t_1, \psi + et_1, \psi) \cdot$$

$$\cdot p(t_2, \psi + et_2, \psi) - \int_0^{\theta} dt_1 \int_0^{t_1} dt_2 \int_0^{t_2} dt_3 (\theta - t_1 + t_3) \cdot \quad (15)$$

$$\cdot (t_1 - t_2)(t_2 - t_3) p(t_1, \psi + et_1, \psi) p(t_2, \psi + et_2, \psi) \cdot$$

$$\cdot p(t_3, \psi + et_3, \psi) + \ldots$$

From (12)

$$\rho_{1,2} = \frac{1}{2}(a \pm \sqrt{a^2 - 4})$$

Since the function $a(\psi)$ is ω-periodic, then we have three cases by (2):

1) $m = \min|a(\psi)| > 2$, 2) $M = \max|a(\psi)| < 2$ and 3) $m \le 2 \le M$

If $m > 2$, then the equation (12) has two real roots $\rho_1(\psi)$ and $\rho_2(\psi)$, one of them is less by modulo than unity and the other is more. Therefore, the range of the multiplicator is a segment from interval (-1,0) or (0,1) and the range of the other – is a segment from interval $(-\infty, -1)$ or from $(1, +\infty)$ respectively. The spectrum of the monodromy matrix does not include zero in this case, and there is a real smooth multiperiodic branch of the logarithm of the monodromy matrix [4].

If $M < 2$, then multiplicators $\rho_1(\psi)$ and $\rho_2(\psi)$ are complex-valued ω - periodic functions are qual one by module, with $\rho_1(\psi) \neq \rho_2(\psi)$. Therefore their ranges of variations are disjoint closed arcs of the unit circle of the complex space. The spectrum of the monodromy matrix also does not include zero with $M < 2$ and its logarithm has a real branch.

The case $m \le 2 \le M$ requires further study, on which we will not stay here.

It's not hard to prove that the matriciant $X(t, \varphi, \psi)$ of the system (3) for cases 1) and 2) can be written

$$X(t, \varphi, \psi) = \Phi(t, \varphi, \psi) \exp[t \Lambda(\psi)], \quad (16)$$

where $\Lambda(\psi) = Ln\, X(\theta, \psi, \psi)$ - is a real branch of logarithm of the mondromy matrix $X(\theta, \psi, \psi)$, $\Phi(t, \varphi, \psi) = X(t, \varphi, \psi) \exp[-t\Lambda(\psi)] - (\theta, \omega, \omega)$ - periodic by matrix (t, φ, ψ).

Since the solution z of the system (3) can be represented:

$$z(t, \varphi, \psi, u(\psi)) = X(t, \varphi, \psi) \cdot u(\psi), \quad (17)$$

then the system (3) unstable for the case 1) and it's stable for the case 2).

This fact is true for the equation (1) in view of the equivalence of the equation (1) and the system (3).

Theorem 1. When the condition (2) is true, D-equation is stable if the largest value of the module M of function $a(\psi)$ defined by (15) is less than 2, and it's unstable if the smallest value of m its module is greater than 2.

IV. Criterions of Instability and Stability

Let $p(t, \varphi, \psi)$ is not identically equal to zero and $p(t, \varphi, \psi) \le 0$. Then, it is easy to show that $\int_0^{\theta} p(t_1, \psi + et_1, \psi) dt_1 < 0$ and

$$(-1)^n I_n = (-1)^n \int_0^{\theta} dt_1 \int_0^{t_1} dt_2 \ldots \int_0^{t_{n-1}} dt_n (\theta - t_1 + t_n)(t_1 - t_2) \ldots (t_{n-1} - t_n) \cdot$$

$$\cdot p(t_1, \psi + et_1, \psi) \ldots p(t_n, \psi + et_n, \psi) > 0$$

Hence, from (15) we have that $a(\psi) > 2$. We state the following proposition by the Theorem 1

Theorem 2. Let $p(t, \varphi, \psi) \le 0$ and it's not identically equal to zero. Then by the condition (2) D-equation (1) is unstable, with positive multiplicators and one of them larger than one, and the other is less.

Now, assume that $p(t, \varphi, \psi) \ge 0$ and it's not identically equal to zero, moreover, assume that the condition

$$0 < I(\psi) = \theta \int_0^{\theta} p(\tau, \psi + e\tau, \psi) d\tau \le 4. \quad (18)$$

Then it is easy to show that $I_1 = I > 0, I_{k+1} > I_k > 0, (k = 1,2,\ldots)$ and from

$$a = 2 - I_1 + I_2 - ... + (-1)^k I_k + ...$$

we obtain

$$2 - I(\psi) < a(\psi) < 2$$

Therefore, by (18) we have $-2 < a(\varphi) < 2$. Further, by the Theorem 1, the multiplicators $\rho_1(\psi) \neq \rho_2(\psi)$ are complex-conjugate and $|\rho_1(\psi)| = |\rho_2(\psi)| = 1,\ \psi \in R^m$.

Theorem 3. Let $p(t,\varphi,\psi) \geq 0$ and it's not identically equal to zero. Then, by the conditions (2) and (18) D-equation (1) is stable with complex-conjugate multiplicators and their modules are equal to one.

The condition (18) will called the Lyapunov's generalized integral criterion of stability for D-equation (1).

Since that $(\theta,\omega,\omega)-$ are the periodic functions $x(t,\varphi,\psi)$ and functions $D^2 x$ obtained by double application the operator D with $\varphi = et$ refers to quasi-periodic functions $\xi(t)$ and their derivatives $\frac{d^2}{dt^2}\xi(t)$, respectively, we obtain an ordinary differential equation from (1)

$$\frac{d^2\xi}{dt^2} = q(t)\xi \qquad (1^{/})$$

with quasi-periodic coefficient by Bor $q(t) = p(t,et,0)$.

Then we obtain the following consequences of Theorems 1-3 for equation $(1^{/})$:

Consequence 1. The equation $(1^{/})$ is stable for $M < 2$ and unstable for $m > 2$ by the conditions of the Theorem 1, where M, m- are the largest and the smallest values of function- analogue of constant Lyapunov's equation (1).

Consequence 2. The equation $(1^{/})$ is unstable by the conditions of the Theorem 2.

Consequence 3. The equation $(1^{/})$ is stable by the conditions of the Theorem 3.

The proofs of these consequences will obtain from Theorems 1-3 with $\varphi = et$.

REFERENCES

[1] B.P. Demidovich Lectures by mathematical theory of stability M.: Science, 1967. – 472 p.

[2] Vazov V. Asymptotic expansions for ordinary differential equations. M: Mir, 1968-464p.

[3] A.A. Mukhambetova, Zh.A. Sartabanov // Mathematics Journal, 2003. Vol.3 №1 (7) pp. 68-73.

[4] A.M. Samoilenko // Elements of the mathematical theory of multiple oscillations. M.: Science, 1987.-304 p.

[5] A.A. Mukhambetova Stability of the second order linear partial differential equations with oscillatory coefficients. International Journal of Experimental Education, №4, 2013 y., pp. 120-124.

[6] Zh.A. Sartabanov Study of multi frequent periodical oscillations of the system differential equations of the first order in partial derivative//Fen Edebiyat Dergisi. The Journal of ArtsandScienes. –Sakarya, 2007. Volume 9, NoEK.– pp. 18-29

[7] Zh.A. Sartabanov Oscillation solutions of the differential equations determined on the given vector field. Materials of the IV th Congress of Turkish world: The mathematical sciences. Baku, 2011 y., p.226

[8] A. Mukhambetova, Zh. Sartabanov Research of multiperiodic solutions of quasi-linear system in the first order partial derivatives. Bulletind'Eurotalent-Fedjip, 2014, № 4. pp. 33-37.

Computational modelling and simulation analysis of trapezoidal channelled micro heat sinks fabricated using cold spray process

A. Hamweendo, P. A. I. Popoola, and I. Botef

***Abstract*—**Computational modelling and simulation analysis are fundamental in studying the miniaturised microfluidic cooling systems, which are highly recommended for cooling the state-of-the-art high speed microprocessors. The cooling capability of these microfluidic systems depends on the design of microchannels which are incorporated in the micro heat sinks as compact heat extractors. Studies showed that micro heat sinks whose microchannels have trapezoidal cross sectioned shapes outperform other designs in terms of heat extraction capability. However, studies on these devices are limited due to their scarcity and lack of robust method to fabricate them. Consequently, this paper proposes a new method for fabricating microchannels and applies computational modelling and simulation analysis to study the thermo-mechanical performance of these microchannels. The results of this study demonstrate that the fabricated microchannels have geometric ratios which are compatible with high thermal-hydraulic performance of the micro heat sink and the method used to fabricate the microchannels exhibit high process flexibility and capability to produce microchannels with different geometric ratios.

***Keywords*—**Computational Modelling, Simulation Analysis, Micro Heat sinks, Cold Spray, and Microfluidic cooling.

I. Introduction

State-of-the-art computer systems are required to operate at high computational speed. Nonetheless, this requirement puts serious concerns on thermal management of the computers due to large heat fluxes emitted by the high speed microprocessors [1]. Although the anticipated heat flux from these microprocessors are in excess of 100 W/cm^2, their operation temperature is limited to between 55 ^{0}C and 100 ^{0}C [2], [3]. However, the current air convective cooling systems, Fig.1, have diminishing capability to extract larger heat flux from high speed microprocessors. This problem requires innovative cooling solutions, such as the microfluidic cooling systems (MFCSs), Fig.2, whose cooling capability is superior. The MFCS has high heat extraction capability because the micropump delivers the cooling liquid from micro condenser to micro heat sink where the fluid picks up the heat using microchannels.

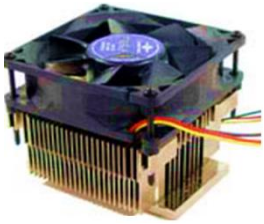


Fig.1: Forced air convection cooling system

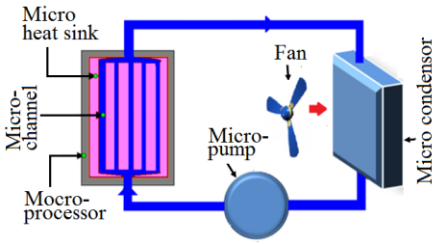


Fig.2: Microfluidic cooling system

The warm liquid then dissipates the heat to the atmosphere through the air cooled micro condenser [4]. Studies showed that trapezoidal shaped microchannels outperform the other designs in terms of heat extraction capabilities [1], [5]. This performance depends on geometric dimensions for trapezoidal shaped microchannels, and on the technology used to fabricate these channels [6], [7]. Currently the major proved technologies for fabricating microchannels are: lithography, chemical etching, and micro-machining. However, these technologies have limitations such as high costs, production of toxic wastes, low productivity, failure to control geometric dimensions of microchannels, and incompatibility for large batch size production [6], [8]. These limitations confirm that there is no robust method for fabricating micro heat sinks [2], [4]. This problem contributes to the scarcity of micro heat sinks (MHSs) with trapezoidal sectioned microchannels, leading to scant availability of information on thermal-mechanical performance of these devices.

Consequently, these shortcomings prompted our research

The support of the DST-NRF Centre of Excellence in Strong Materials (CoE-SM) towards this research is acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the CoE-SM.

A. Hamweendo is finalizing his PhD with the University of the Witwatersrand, School of Mechanical, Industrial and Aeronautical Engineering, Johannesburg, South Africa. (corresponding author: +27-11-717-7438; e-mail agripa.hamweendo@students.wits.ac.za).

P. A. I. Popoola is with Department of Chemical and Metallurgical Engineering, Tshwane University of Technology, Pretoria, South Africa (e-mail: popoolaapi@tut.ac.za).

I. Botef is with the University of the Witwatersrand, School of Mechanical, Industrial and Aeronautical Engineering, Johannesburg, South Africa. (e-mail: ionel.botef@wits.ac.za).

efforts towards developing a new method for fabricating microchannels with trapezoidal cross sections, followed by thermal-mechanical evaluation of these devices. This research was conducted in the cold spray laboratory, at University of the Witwatersrand, Johannesburg.

Subsequently, section II outlines a new method for fabricating microchannels with trapezoidal shaped sections, which has a patent application. Then, section III, applies the computational modelling and simulation analysis to evaluate the thermal-mechanical performance of the microchannels fabricated using the new method. Section IV outlines and discusses the results. Finally, section V presents the conclusions and further work.

II. The New Method to Fabricate Microchannels

This new method is an integration of Cold Spray (CS) with Alloy-de-Alloy concept. CS, shown in Fig.3, coats several metals' powders by exposing a metallic or dielectric substrate to a high velocity (300-1200 m/s) jet of small (1-50 µm) particles accelerated by a supersonic stream of compressed and preheated gas [9]. This coating process lends CS the following advantages: flexibility, capability to depositing coatings with: high thermal conductivity; minimal oxidation; different combinations of metals; high reproducibility; high deposition efficiencies, and lower cost [9].

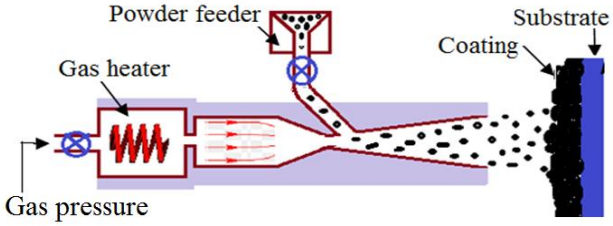


Fig.3: Schematic for low pressure CS process

To fabricate the microchannels, Copper (Cu) powder was deposited as a matrix former, while Aluminium (Al) was coated as a microchannels' forming agent. The powders and the CS equipment were sourced from Centerline, Canada. During fabrication, three layers of Cu powder were directly sprayed on to a grit blasted Cu substrate. Cu deposition was alternated with deposition of one layer of Al. Al was deposited through a mask to make patterned coatings. This alternated deposition was repeated until three layers of Al coatings were imbedded in Cu matrix. After coating, the imbedded Al layers were selectively de-alloyed by repeated dipping of the specimens in dilute acid. Then, the specimens were cross sectioned, metallographically polished and finally characterised by taking images, shown in Fig.4, using the Optical Microscope. From these optical images, 10 geometric dimensions for each side of the microchannels were taken using the Optical Microscope and averaged. These dimensions, indicated in Fig.5, are: a=1091 µm; b = 443 µm; c = 300 µm; d = 400 µm; f = 650 µm, h = 436 µm, w = 4132 µm, L = 4 500 µm; t = 2708 µm; and t_w=870 µm.

To assess the thermal-hydraulic performance and structural rigidity of the Cu MHS, the analytical models (1)-(2) were applied to calculate the geometric ratios, and (3)-(4) to calculate the hydraulic diameter. These models were developed by Bower et al. [10].

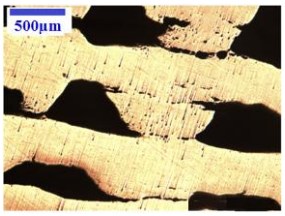


Fig. 4: Trapezoidal shaped microchannels

$$\frac{t}{4c} \approx 1.2 \quad (1)$$

$$\frac{t_w}{f} \le 2 \quad (2)$$

$$d_h = \frac{4A}{P} \quad (3)$$

$$100 \le d_h \le 500\mu m \quad (4)$$

Where t is the channel thickness, c is space between channels, and t_w is the cell width, d_h is the hydraulic diameter of the microchannel, A is cross section area of microchannel and P is the perimeter of the microchannel, and L is the arbitrary length of the MHS.

In this analysis, the geometric ratios were calculated as follows: t/4c = 2.25; and t_w/f =1.34 and the hydraulic diameter = d_h= 4A/P= 336µm.These values signify that the fabricated array of microchannels is rigid and its thermal-hydraulic performance is promising [10].

III. Computational Modelling and Simulation Analysis of Cu MHS

To evaluate the thermal-mechanical performance of the Cu MHS, the computational modelling of trapezoidal cross sectioned microchannels was carried out in order to mimic Cu MHS. In the computational modelling, the computational domain, Fig.5, the geometric dimensions given above, and the Solid Edge 3D modeller were used.

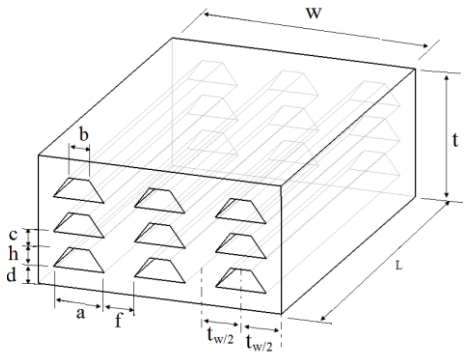


Fig.5: Computational domain

The model which resulted from the computational modelling process, shown in Fig.6, was the object for subsequent simulation analysis of the Cu MHS. In this simulation, de-ionised water was used as the cooling fluid [11]

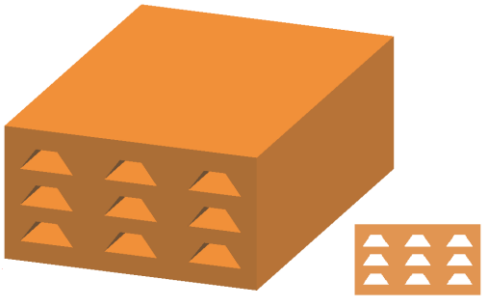


Fig.6: Modelled Cu MHS.

The key parameters required as inputs the simulation window are:

- convective heat transfer coefficient,
- heat fluxes,
- pressure of circulating water and
- ambient temperature.

To calculate the convective heat transfer coefficient, *h*, (5) was applied [12].

$$h = Nu_D * k / d_h \qquad (5)$$

Where Nu_D is the Nusselt number, *k* is thermal conductivity of water, and d_h is the hydraulic diameters of the microchannel. To calculate Nu_D, the fluid laminar flow regime was required, whose Reynolds' number, *Re*, is given by (6) [12].

$$\mathrm{Re} = \frac{u_m d_h \rho}{\mu} \le 2{,}300 \qquad (6)$$

Where u_m is the mean velocity of the fluid, d_h =336 µm, ρ = 1000 kg/m^3 is density of water, and μ = 6.31*10^{-6} kg/s.m is the viscosity of water, and *k* = 0.634 W/mK is thermal conductivity of water.

Additionally, the convective heat transfer coefficient, *h*, relates to u_m as in (7) [13].

$$h = 14300 u_m^{0.3} \qquad (7)$$

From open literature, *h* varies between 50 and 10,000W/m^{2}K, from which u_m was calculated using (8).

$$50 \le h = 14300 * u_m^{0.3} \le 10{,}000 \qquad (8)$$

Equation (8) gave, u_m = 0 to 0.3m/s. Substituting u_m = 0.3m/s in (6), gave R_e = 1,598. This means that the fluid flow regime in the microchannels is laminar. To calculate Nu_D for a laminar flow in trapezoidal shaped microchannels, the shape factor, *b/a* = [(443+1019)/2] / 436 =1.64. Using this shape factor and interpolation in Table 8.1, page 519 [12], Nu_D=3.87. Substituting the values for Nu_D, *k* and d_h in (5), the required forced convective heat transfer coefficient, *h*, was calculated in (9):

$$h = Nu_D * k / D_h = 3.87 * 0.634 / 336 * 10^{-6} = 7{,}245.71 W/m^2.K \qquad (9)$$

Finally, the summary of parameters used in the simulation window is:

- heat transfer coefficient= 7, 242 W/m^2K;
- heat fluxes, *Q*=100, 200, 300, 400, and 500 W/cm^2;
- pressure of water,=1 bar; and
- ambient temperature = 20 ^{0}C.

The simulation results and brief discussion are presented below.

IV. Results and Discussion

Table 1 presents the variation of simulated temperature of Cu MHS with heat flux ranging from 100 to 500 W/cm^2. From this table, it can be seen that the temperature at the top surface of the MHS remains practically constant and it is equal to the ambient temperature, while the temperature at the bottom of the MHS, increase marginally from 22 ^{0}C to 28 ^{0}C, which is below the minimum operation temperature of 55 ^{0}C required for the microprocessor [2]. The surface below the Cu MHS in essence the junction surface between the MHS and the microprocessor, and the accompanying temperature is the junction temperature.

Table I: Variation of temperature with heat flux

Heat flux W/cm^2	*Surface Temp ^{0}C*	*Junction Temp*
100	20	22
200	20	22
300	20	25
400	20	27
500	20	28

Fig.7 is the graph of the variation of the temperature with the heat flux, shown in Table I. This figure shows a rise in junction temperature for the Cu MHS from heat flux of 200 W/cm^2, with the likelihood that this temperature will be remain constant beyond heat flux of 500 W/cm^2.

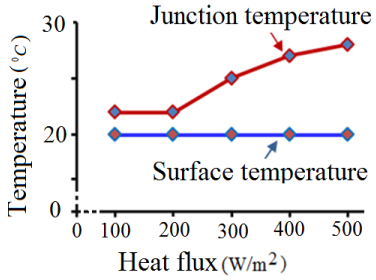


Fig 7: Variation of temperature with heat flux

Fig.8 illustrates the simulated temperature profile of the Cu MHS when the maximum heat flux of 500 W/cm^2 is applied. Even at this heat flux, the temperature at the top surface is 20 ^0C, while the largest temperature of 28 ^0C occur at the bottom of the Cu MHS. This implies that a temperature of the microchannel heat sink rise by only 8 ^0C above the ambient.

Experiments studies by [14] revealed that the micro heat sinks fabricated from Copper and Aluminium can lower the temperature of the microprocessor to 10^0C and 3^0C above ambient temperatures, respectively.

It may be of interest to note that even at 500 W/cm^2, which is the highest anticipated heat flux from microprocessors, the temperature at the bottom of the Cu MHS is far below the minimum operation temperature of 55 ^0C required for the microprocessor. Moreover, it appears that this MHS can attain ambient cooling between the first and second layer of the microchannels. This suggests that the third layer is redundant, which imply that high cooling capability with these microchannels is attainable even when only two layers of microchannels are in use.

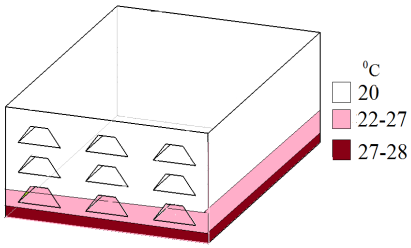


Fig.8: Simulated temperature profile for Cu MHS at 500 W/cm^2

Fig.9 depicts the stress distribution due the fluid pressure of 1 bar considered as the maximum pressure by which the micro pump circulates the cooling fluid [10]. As this figure indicates, the maximum stress induced to the Cu MHS by this pressure is only 250 kPa, which is far below the strength of 259,000 kPa for Cu. This situation signifies that the Cu MHS has high structural rigidity. This agrees very well with the Bowers et al. theory [10].

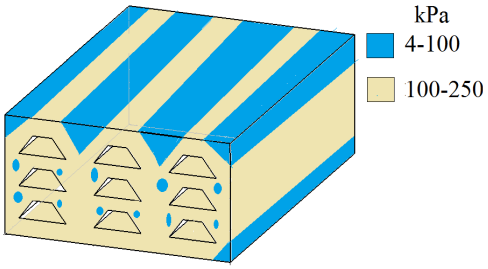


Fig.5: Simulated stress profile for Cu MHS at 1 bar pressure

V. Conclusions and Recommendations

In this paper, computation modelling and simulation analysis were applied to study the thermal-mechanical performance of the Cu MHS with trapezoidal shaped microchannels which were fabricated using a new method. Based on this study, the prospective Cu MHS constructed from these microchannels have demonstrated high thermal extraction capability which can be considered for integration into microfluidic cooling of high speed microprocessors. Integration of this new device in micro cooling of microprocessors will enhance the reliability and service life of computers.

A newly method for fabrication of microchannels with trapezoidal sectioned microchannels was developed and applied to fabricate microchannels with trapezoidal shapes at the University of the Witwatersrand, Johannesburg. This method overcomes several challenges characterising the contemporary microfabrication technologies. Consequently, this new method can be studied for possible integration into advanced manufacturing systems for these microfluidic cooling devices.

Further studies to elucidate the in-situ thermal-mechanical performance of the Cu MHS fabricated by this new method are encouraged. These studies may focus on experimental, analytical, computational and simulation evaluation of thermal-hydraulic performance of microchannel heat sinks with varying geometric aspect ratios fabricated using cold process and Copper as well as Aluminium as matrix formers. Microchannel heat sinks which are alloys of Copper and Chromium could also be considered. It is envisaged that the results of this process can be advanced further and be applied in developing design and optimization strategies for microchannel heat sinks using non-traditional approaches such as multi-objective optimisation algorithms, with a focus of improving the thermal-hydraulic performance of micro heat sinks which are much needed in microfluidic cooling of next generation hyper speed microprocessors.

VI. References

[1] H.Gargi, V.S.Negi, Nidhi, and A. K. Lall, "Numerical Study of Microscale Heat Sinks Using Different Shapes & Fluids", (CSIR-CSIO), *Proc. of the 2013 COMSOL Conf.* India (2013)

[2] M.Ioan, "Heat Transfer in Minichannels and Microchannels CPU Cooling Systems, Heat Transfer Theoretical Analysis", *Expt. Investig. and Indust. Systems*, ISBN: 978-953- 307-226-5 (2011)

[3] Gaikwad, V.P, "Microchannel heat sink fabrication techniques, Second International Conference on Emerging Trends in Engineering", IOSR *J.of Mech. and Civil Engineering*, ISSN: 2278-1684, pp: 51-57 (2014)

[4] G.Upadhya, M.Munch, O. Zhou, J.Hom, D.Werner, D., and M. McMaster, "Micro-Scale Liquid Cooling System for High Heat Flux Processor Cooling Applications", *22nd IEEE Semi-thermal Symposiums*, Mountain View, CA 94043,(2006)

[5] O.Asgari, and M.H. Saidi, "Approximate method of determining the Optimum cross section of microchannel heat sink", *J. of Mech.,Sci.and Tech.*, vol. 23, pp.3448-3458, KSME & Springer (2009)

[6] A.Husain, and K-Y, Kim. "Thermal Optimization of a Microchannel Heat Sink with Trapezoidal Cross Section", *J.of Electronic Pack. Vol.* 131, pp. 021005-1 to 021005-1, ASME (2009)

[7] A. Prakash, and S.Kumar, "Fabrication of microchannels: A review". *Proceedings of the Institution of Mechanical Engineers, Part B*: *J. of Eng. Man.*, pp.1-16, SAGE, (2014)

[8] Zhou, W. Deng, L. Lu, J.Zhang, L, Qin, S. Ma, and Y.Tang, "Laser Micro-milling of Microchannel on copper sheet as catalyst support used in microreactor for hydrogen production", *Int. J. of Hydrogen Energy*, vol.39, pp.4884-4894, ScienceDirect,(2014)

[9] A.Papyrin, V.Kosarev, S. Klinkov, A. Alkhimov, and V.Fomin, "Cold Spray Technology", *Summer Universities* France (2006).

[10] M.B.Bowers, and I.Mudawar, "Two-phase electronic cooling using Mini-channel and micro-channel heat sinks", part 1-design criteria and heat diffusion constraints, *J.of Electronic Packaging, Transactions of the ASME*, Vol. 116, pp.291-297(1994)

[11] R.Gautam, A.K. Sharma, and K.D.Gupta, "Performance analysis of Non-circular microchannels flooded with CuO-water nanofluid", *3rd Micro and Nano Flows Conf.*, Thessaloniki, Greece, (2011)

[12] F.P.Incropera, D.P.Dewitt, T.L.Bergman, and A.S.Lavine, "Fundamentals of Heat and Mass Transfer", S*ixth Edition, John Wiley & Sons*, SBN-13: 978-0-471-457282(2006)

[13] C.Tangwongsan, "Fluid Velocity Measurement Using Convective Heat Transfer Coefficient Measuring System Chanchana", *Conf. Proc., LSSAA Workshop*, pp.81-84, IEEE, Bethesda, MD (2007)

[14] P.Naphon and S.Wiriyasart, "Liquid cooling in the mini-rectangular fin heat sinks with and without thermoelectric for CPU, International Communications in Heat and Mass Transfer, vol. 36, 2009,pp. 166–171

Supply Chain Management for Medical and Psychological Assistance in Post-Disaster Calamities Situation - Case Flood

Lorena Silvana Reyes-Rubiano, Andrés Felipe Torres-Ramos, Carlos Leonardo Quintero-Araújo

Abstract—The humanitarian logistics supply chain can be described as a network of volunteer staff and that interact with a set of goods and services, this with the purpose of satisfying the demand of the population affected by a sudden flood. This article focuses on the formulation of a model's location to a point of distribution and multiple shelters whereas flood risk associated to the area, along with a model of routing of specialized personnel, allowing to relieve the psychological and medical calamities among other present in the population affected in a post-disaster situation; This formulation is made by applying operations research as a solution tool. The aim of this article is to provide a functional model to strategically design network facilities, in addition to coordinate the provision of services required by the population in the shortest possible time. The flood that the town of Santa Lucia, Colombia suffered in 2010 is taken as a case study.

Keywords—Humanitarian logistic, location of shelters and a distribution point, risk area, routing of staff.

I. Introduction

WHEN discussing a flood there are two possible causes: the first is related to natural phenomena such as rain and the rain seasons, the second cause is man-made actions that induce to a large extent natural disasters [1]; These causes include deforestation, damage to watersheds and elimination of wetlands, which contribute to the intensity of floods, landslides and droughts. Additionally a flood may be the side effect of another disaster that occurred in the area. On the other hand, the cutting of mangroves on the coasts contributes to coral erosion, generating storms and hurricanes; the main effect that causes most of these disasters is "human development", which causes the reduction of the ozone, generating climate change along with other problems [2].

This work was supported by the Master in Operations Management and the International School of Economics and Management Sciences (EICEA) of the Universidad de La Sabana, Chía, Colombia.

Lorena Silvana Reyes-Rubiano is with the International School of Economics and Management Sciences (EICEA), Universidad de La Sabana, Chía, Colombia (corresponding author to provide phone: 571-8615555 ext.: 21106; e-mail: lorenareru@unisabana.edu.co).

Andrés Felipe Torres-Ramos is with the International School of Economics and Management Sciences (EICEA), Universidad de La Sabana, Chía, Colombia (e-mail: andrestora@unisabana.edu.co).

Carlos Leonardo Quintero-Araújo is with the the International School of Economics and Management Sciences (EICEA), Universidad de La Sabana, Chía, Colombia (e-mail: carlos.quintero5@unisabana.edu.co).

Floods can be classified into two types: sudden floods and progressive flooding. These have specific characteristics that vary according to the area of occurrence, i.e. the effect of flooding is not the same in a rural area as it is in an urban area. This article will analyze sudden flooding in rural areas, which is caused by rains, storms and hurricanes, among other causes [3].

Sudden floods are characterized by the high level of water and the speed with which it attacks; it is estimated to be within minutes or hours that the area is completely flooded. These can be caused by the rupture of dams, dykes or torrential rains in flat areas, and such events are not possible to detect or predict. Because of this floods are credited with the highest rate of mortality among natural disasters [4]. A flood can drag rocks, knock down trees and destroy infrastructure, generating a large number of deaths and injuries of all kinds. Floods, without exception, generate damage to the population, infrastructure and the damage of a region, giving way to the spread of diseases and the complexity in which the disaster response must perform their operations [3].

To mitigate the impact of the disaster, some organizations such as the Red Cross have developed training programs aimed at the population and institutions impacted; parallel to this a series of protocols and guidelines for the management of such situations have been developed [5]. Despite these measures, floods are natural disaster that present the greatest number of lives lost and affected population [6].

There are two major phases identified in floods and other disasters: the pre-disaster stage and post-disaster stage. Each of these frames a series of operations; the pre-disaster stage focuses on operations preparation and plans of response. In this stage, operations refer to activities related to drills and lines against a flood. The post-disaster phase has two operations: mitigation and recovery. Mitigation operations refers to the design of the logistics network which will allow an optimum operation. Recovery refers to the revival of the population and the area [7].

This article is framed in the post-disaster phase and in the operation of mitigation of the negative impact of the flood. From a strategic point of view determining the optimum location of shelters and from a of operational point of view by defining the optimal distribution of qualified personnel from a distribution point until shelters, with the aim of having a minimum distance between the locations and at the same time ensuring reduced travel times. The article is divided into: II.

ISBN: 978-1-61804-256-9

Literature review. III. Approach of the problem. IV. Case study. V. Result. VI. Discussion. VII. Conclusions.

II. Literature Review

The location of the points of distribution and multiple shelters, is the starting point for the design and management of the supply chain [8]. Location has a great impact on the operation of distribution of humanitarian aid, so the article's [9] defined a criteria to measure the performance of the location, these criteria are related to travel times, transportation costs and equity.

The process of facility locations and transport are operations that are closely related, since the facilities location affects routes and the effectiveness of the assistance required by the population. According to these considerations B. B. Balcik y B. M. Beamon [10], propose a model of temporary location, where the points are re-located according to the requirements of the area or to the requirements of the route, While [11], [12] propose a discrete location models including routing and distribution of food, medicines, water, etc. having as starting point the fixed points of the multiple shelters. The main objective is to improve performance in terms of response times, from the supply chain.

III. Approach of the Problem

This chapter presents two models, the first arises to solve the facility location problem and the second intends to solve the problem of the distribution of personnel. The model is constructed through the application of operations research.

A. Problem Description and Assumptions for Location Model

The proposed model seeks to solve the problem of the installation of a point of distribution and shelters in a neighboring area to the area affected by the flood, to determine potential geographic points considering the probability of risk of flooding in each selected area. The model aims to reduce the time of travel between the distribution point and shelters; thus seeks to ensure that the designed supply chain allows for the assistance of the population affected in minimum response times. The facilities location is based on the assumption that the flood-affected region lacks infrastructure to house and assist the affected population, so it is necessary to evacuate and install this population in safe areas with the purpose of providing the assistance required. In addition the proposed model considers that regions close to the flooded area have an associated risk of flooding.

According to the issues raised by [13] to formulation of the model, the region affected by the flood is divided into sub-zones, and from these sub-zones it was establish that many people must be evacuated and installed in shelters; this model is based on the assumptions that shelters have the same capacity in number of people and the certainty of demand, i.e. the points where the population needs to be evacuated is known with certainty. It is important to clarify that the proposed model is focused on location and not on the evacuation process; however the proposed localization model considers the number of families with some sort of calamity, requiring them to be evacuated and installed in a shelter located in a safe area. Therefore the article focuses on the installation of a point of distribution and multiple shelters.

In addition to the location of the shelters and the installation of a single point of distribution is defined on the basis of a post-disaster situation, which is characterized by damage to access roads, disruption in the information network and the provision of goods and services [14]. According to the experiences of humanitarian aid organizations, operations related to the management of a post-disaster situation are uncontrolled and uncoordinated due to lack of information [3], in addition multiple distribution points causes shelters to be supplied by mistake more than once, leading to inefficient use of limited resources. For these reasons a single point of distribution is considered to ensure the centralization of information, allowing a coordinated and efficient operation.

The facility location is the critical decision in managing the post-disaster of a flood, as that is the operation that defines the supply chain. The definition of points for facility locations are part of the strategy for post-disaster management, therefore the speed with which responding to calamities in a disaster depends on facility locations, as well as the success of supply chain operations. To consider these aspects, a mathematical model is proposed that seeks to reduce the time of travel between the distribution point and shelters.

In the commercial supply chain the main objective is the reduction of logistics costs that are considered fixed costs and variable costs associated with the facility location. In the proposed model considered constraints and goals related to suffering of the affected population and at the same time ensure compliance with human rights, these constraints and goals were not considered since management in post-disaster situation [15].

B. Location Model

For the formulation of the first model datasets and indices were determined, then the variables and parameters.

a) Notation

Sets and indices

V	Set of areas that are potentially unsafe $\{I \cup J\}$
I	Subset of V that describes potentially safe areas to install the distribution point.
J	Subset of V that describes potentially safe areas to install shelters.
D	Set of sub-zones in the flooded region.
M	Set of calamities in the affected population.

Model

$$\text{Mínimize Z} = \sum_{j \in J} \sum_{i \in I} E_{i,j} T_{i,j} RA_j RCD_i \quad (1)$$

$$\sum_{i \in I} a_i = 1 \quad (2)$$

$$\sum_{\forall i \in I} E_{i,j} = b_j, \quad \forall j \in J \quad (3)$$

$$\sum_{\forall j \in J} E_{i,j} = F\,\mathrm{a}_i, \quad \forall i \in I \tag{4}$$

$$b_j(F) \geq \sum_{d \in D} \sum_{m \in M} Q_j^{d,m}, \quad \forall j \in J \tag{5}$$

$$b_j \leq \sum_{d \in D} \sum_{m \in M} Q_j^{d,m}, \quad \forall j \in J \tag{6}$$

$$\sum_{j \in J} b_j \geq \frac{\sum_{d \in D} \sum_{m \in M} O_{d,m} + \Delta}{K} \tag{7}$$

$$\sum_{j \in J} Q_j^{d,m} \geq O_{d,m}\,, \quad \forall d \in D,\ m \in M \tag{8}$$

$$\sum_{d \in D} \sum_{m \in M} Q_j^{d,m} \leq K, \quad \forall j \in J \tag{9}$$

$$a_i + b_j \leq 1, \quad \forall i \in I,\ j \in J,\ i = j \tag{10}$$

$$a_i \in \{0,1\}, \quad \forall i \in I \tag{11}$$

$$b_j \in \{0,1\}, \quad \forall j \in J \tag{12}$$

$$Q_j^{d,m} \geq 0, \quad \forall j \in J,\ d \in D,\ m \in M \tag{13}$$

The objective function of the model (1) is to minimize travel between a distribution point and shelters, considering the risk associated to locate any type of installation in the area *i* or *j*, the risk is a parameters that penalizes journey times. Restriction (2) ensures that the number of points of distribution open is only one. Restriction (3) ensures that shelters will be attended whenever they are opened, (4) restriction guarantees that shelters can only be attended by this open distribution point. Restrictions (5) and (6) ensuring that a shelter can cater to a zone only if it is opened. (7) Restriction ensures that the number of shelters open is able to accommodate all of the affected population. Restriction (8) ensures that the number of people in each shelter must be greater than or equal to the demand of each zone. (9) Restriction ensures that the number of people in each shelter must be less than or equal to the capacity of the shelter. The set *J* is a copy of the set *I* so (10) restriction guarantees that an area may not have more than one use, i.e., the area is used as a shelter or as a distribution point. Restrictions (11), (12) and (13) characterize the variables considered in the problem.

C. Definition Set Zone

To start it is necessary to reference human rights, which defines that all people should have a safe and worthy place like the shelter [1]; Therefore the potential areas to locate the flood-affected population must comply with some considerations, which are related to protection against weather, spread of diseases and health problems. The objective of these considerations is to mitigate the effect of flooding on the affected population.

Potential areas should be characterized by allowing the supply of goods and services, compliance with topographical conditions as the inclination of the land cannot be greater than 6% or less than 1%, and ensure that it has a system for the disposal of solid waste and waste water, in the same way these areas should be located away from stagnant water, wastewater and of all types of waste [1], in order to prevent the emergence and spread of diseases.

Shelters and the single distribution point are installed in potential areas, which must have the shortest possible distance between them, in order to ensure easy access and prompt assistance for the affected population. Usually schools, churches, sports fields, and educational institutions in the neighboring regions to the flood are considered for shelters and distribution points [16], so the whole V refers to these types of places located in potential areas that comply with the characteristics listed above, and in addition must have some kind of infrastructure that will allow the assistance to the population. In the proposed model the capacity of the shelters is homogeneous, and the capacity the distribution point is associated with the number of personnel available to tend to the affected population.

E. Definition of Risk of a Zone

The determination of potential safe zones for the installation of a shelter or a distribution point depends on the region affected by the flood; so it is necessary to assess the risk for each of these areas in order to relocate the population affected by the flood and qualified personnel [1].

In this case the risk associated with each zone is defined as the frequency of occurrence of a flood in this area; these values are an input parameter for the formulation of the model, which are taken as factors that could cause increase or decrease time depending on the risk of each zone. The risk is an external factor that is important to consider, because badly located shelters and the distribution point represent a high risk of facing a second flood, which means double the efforts to assist the affected population and the increase in vulnerability to diseases, psychological disorders and loss of human lives among other problems.

F. Definition of Calamity

Calamities are the kind of health problems that can attack the population as a result of the flood. In the case of sudden floods, there are different types of problems that require a type of specific health care. Injuries that occur in floods are mostly not lethal, however, diseases carried by the flood can be. These calamities cause the floods are the natural disaster that generates the greatest number of lives lost [3], [5].

Sudden floods often cause the fall and the movement of trees, rocks, rubble and bulky items, which are the main cause of physical injuries and drowning of persons. Water contaminated with industrial waste is the main medium for the spread of dengue, malaria, cholera, diarrhea, hepatitis A and E, gastrointestinal, and respiratory diseases among other vector-borne diseases [3]. Other types of calamities are associated with a psychological component, mainly caused by forced displacement, the loss of relatives, and physical and sexual abuse among others [5].

The proposed model considers three groups of calamities by severity which were grouped in the following way (see Table I):

1. Medical calamities.
2. Psychological calamities.
3. Minor calamities.

Table I Grouping of the problems in the calamities groups.

Health problem	Calamity		
	1	2	3
Psychological disorders by flooding (Loss of family).		X	
Psychological disorders by flooding (Economic losses).		X	
Psychological disorders by physical and / or sexual abuse.		X	
Fracture.	X		
Crushing of limbs.	X		
Minor injuries.	X		X
Acute diarrheal diseases.	X		
Respiratory infections.	X		
Vector-borne diseases (Malaria, dengue, cholera, etc.)	X		

Every calamity requires the attention of a specific kind of personnel, however, in a post-disaster situation a coordinated staffing process does not exist and therefore the allocation is d randomly, implying an increase in the suffering of the affected population.

G. Description of the Problem and Assumptions for a Model of Personal Routing

The timely provision of medical and psychological assistance is part of the requirements of the affected population, by which this routing model focuses on the provision of specialized personnel that can alleviate the suffering of the affected population. It is necessary to remember that in post-disaster situations resources are limited, especially the human resource, which is required for optimal coordination to relieve the suffering of the population.

The model is based on the assumption that all kinds of staff are able to handle any calamity, however, the percentage of relief and the time that it takes each staff depends the type of calamity. This is the reason that the prioritization of demand is necessary to minimize and present control over the calamities.

The model proposed considers four types of personnel in the management of the post-disaster situation: doctors, nurses, psychologists and personal first aid; for each type of staff defined a attention time and a relief level by calamity. For the responses to calamities brigades or mobile units are defined for the purpose of providing support in the work of the medical personnel, nurses, psychologists and rescuers, i.e. staff not opera alone, operations are developed together with well-trained medical staff and personal first aid.

The demand for shelters is given by the numbers of families that require attention. It is a value that is known with certainty, since the above location model serves to assess and report the actual status of the affected population. Working in this way ensures that information is concentrated in a single point and from the coordination of personnel is optimal. Humanitarian logistics comprises two levels of coordination: national and international as the Pan American Health Organization, International Federation of Red Cross and Red Crescent, etc., which unite their efforts to provide assistance to the affected population in a timely manner. This highlights the importance of segmenting the type of personnel according to their knowledge and not according to the Organization to which they belong.

H. Routing Model for Staff

For the formulation of the first model datasets and indices were determined, then the variables and parameters.

a) Notation

Sets and indices

I,J	Set of locations $\{W \cup E\}$.
W	Subset Distribution Point.
E	Subset of multiple shelters.
L	Set mobile units.
M	Set of calamities present in the affected population.

Decision variables

$X_i^{j,l}$	Binary. 1. If the mobile unit l goes from the distribution point or shelter i to shelter or point of distribution j. 0. Otherwise.
Y	Days used in the operation.

Parameters

$TV_{i,j}$	Travel time between distribution point or shelter i to shelter or point of distribution j.
$D_{j,m}$	The number of families suffering from the calamity m in the shelter j.
$TA_{l,m}$	Time that it takes the mobile unit l to relief the calamity type m.
C_l	Number of families that can attend the mobile unit l a day.
$A_{l,m}$	Percentage of relief provided by the mobile unit l when it attends the calamity m.
V	Number of nodes to visit.
α	Weight in goal function of the time of travel between the distribution point and shelters
γ	Weight in goal function of the time spent in the care for the affected population.
φ	Conversion factor from minutes to days for operation time.

Model

$$\text{Mínimize } Z = \sum_{i\in V}\sum_{j\in V}\sum_{l\in L}\sum_{m\in M} X_i^{j,l}(\alpha TV_{i,j} + \gamma TA_{l,m} D_{j,m}) \quad (1)$$

$$\sum_{j\in J} X_i^{j,l} = 1, \quad \forall l \in L, i \in W \quad (2)$$

$$\sum_{i\in J} X_i^{j,l} = 1, \quad \forall l \in L, j \in W \quad (3)$$

$$\sum_{\forall i\in I} X_i^{j,l} = \sum_{\forall i\in I} X_j^{i,l}, \quad \forall j \in E, l \in L \quad (4)$$

$$\sum_{\forall i\in I}\sum_{\forall l\in L} X_i^{j,l} C_l A_{l,m} \geq D_{j,m}, \quad \forall j \in E, m \in M \quad (5)$$

$$\sum_{\forall i \in I} \sum_{\forall l \in L} X_i^{j,l} \geq 1, \quad \forall j \in E \tag{6}$$

$$\sum_{\forall j \in J} \sum_{\forall l \in L} X_i^{j,l} \geq 1, \quad \forall i \in E \tag{7}$$

$$Y = \sum_{i \in V} \sum_{j \in V} \sum_{l \in L} \sum_{m \in M} X_i^{j,l} \varphi(\alpha TV_{i,j} + \gamma TA_{l,m} D_{j,m}) \tag{8}$$

$$u_i - u_j + VX_i^{j,l} \leq V - 1, \quad \forall i \in E, j \in E, l \in L \tag{9}$$

$$X_i^{j,l} \in \{0,1\}, \quad \forall j \in J, i \in I, l \in L \tag{10}$$

$$Y \geq 0 \tag{11}$$

Equation (1) describes the objective function of the model, which consists of two parts: the first is related to the travel time between shelters and distribution point and the second refers to the attention time for the affected population; for each part a weight α and γ of importance respectively is defined. Restriction (2) ensures that the staff available at the point of distribution is sent to serve the population. Restriction (3) ensures that all mobile units that came out of the distribution point must return to the same. (4) Restriction is associated with the balance between shelters, i.e., ensures the continuity of the route of the mobile unit. Restriction (5) determines the capacity of each mobile unit and the respective percentage of relief with respect to the attended calamity. It also ensures compliance with the demand. Restrictions (6) and (7) allow each shelter to be attended to by a mobile unit. Equation (8) represents the minimum number of days required to attend to the affected population. Restriction (9) ensures the breaking of the sub-cycles of routes. The restrictions (10) and (11) describe the type of variables considered in the model, where it is a binary variable and it is a positive variable.

I. Definition of Staff

Although all the mobile units are able to serve all the calamities, the percentage of relief varies according to the kind of mobile unit. The mobile units considered in the model differ from the type of personnel, for the model 4 types of mobile units are considered: mobile unit of personal first aid, which are trained to address minor calamities; the medical mobile unit; mobile nurses and psychologists mobile unit, integrated by three people: a doctor, a nurse and a psychologist respectively and two support people.

The percentage of relief provided by each mobile unit was determined according to the Delphi method applied to physicians, nurses, psychologists and personal first aid. Relief capabilities vary due to the type of professional, defining the matrix that describes the variation depending on the type of calamity and staff that attends it; this is because each calamity requires a type of personal specific attention to ensure the maximum possible relief.

The challenge in the management of the post-disaster situation is to provide timely assistance in reduced response times, so the proposed model considered the prioritization of demand through the allocation of appropriate personnel, i.e., supply staff representing a higher percentage of relief to the population concerned according to the type of calamity; with $A_{l,m}$ intends it to prioritize demand, i.e., to assign more trained personnel to every calamity in order to increase the level of survival of the population. The following describes the matrix (see table II):

Table II *Relief for mobile unit capacity* $A_{l,m}$

Mobile Unit	Calamity		
	1	2	3
Doctor	90%	80%	80%
Nurse	80%	10%	80%
Psychologist	20%	50%	10%
Personal first aid	20%	30%	20%

IV. Case Study

The developing countries or with high levels of poverty are those who have a higher risk of disasters [1]. For this article the Flash flood which the town of suffered Santa Lucia in the Department of Atlantic Colombia in 2010 is considered as a case study, (See Figure I), which has the characteristics considered in the model. Flooding affected the region and 1737 families left homeless [17].

Fig. 1 Town of Santa Lucia. Source: Bank of the Republic of Colombia [17].

A. Definition of Areas

To determine the sub-zones of Santa Lucía, the town is divided into north, south, east and west, resulting in four sub-zones. Socio-political issues in the region establishing that the shelters and distribution point must be installed in the town of Sabanalarga, Ponedera, Palmar de Varela, Baranoa and Barranquilla [16], the Table III describes the number of potential areas per Towns.

Table III Potential areas for installing shelters and distribution point

Town	Number of shelters
Sabanalarga	10
Ponedera	10
Palmar de Varela	10
Baranoa	10
Barranquilla	11

B. Definition of Mobile Units

The mobile units of doctors, nurses and psychologists have a capacity for 50 people per day [1], i.e. 10 families per day. The model considers mobile units presented in Table IV.

Table IV Mobile units considered in the model

Mobile unit	Amount
Doctor	2
Nurse	2
Psychologist	4
Personal first aid	6

V. Result

The validation of the models was carried out in commercial GAMS software with a time limit of 0.13 seconds for both an Intel (R) Core TM i7-4500U CPU 1.8 GHz with 8 GB of RAM. The localization model considers a total of 51 possible sites for the location of the shelters and the distribution point, from which 39 for shelters and 1 to the point of distribution are determined (See table V). These shelters and the distribution point are located within the shortest possible time between them. According to the approach of the problem considered, shelters refer to schools and arenas of the neighboring regions. These have the capacity of accommodating 50 families, i.e. on average approximately 250 people. Appendix A shows the number of families with each kind of calamity that are installed in the shelters.

Table V Model output location

Town	Number of shelters	Distribution Point
Sabanalarga	10	0
Ponedera	10	0
Palmar de Varela	10	0
Baranoa	9	1
Barranquilla	0	0

The routing model for the staff determines the minimum time of operation, considering travel time and the time required to assist the affected population. The results of the model determined the assistance of the entire population, however, not all mobile units are assigned. Flood assistance requires a minimum of one medical mobile unit and 2 mobile units of personal first aid. Table VI presents the route of the mobile unit of lifeguard, which is part of the point Baranoa 1, followed by Baranoa 10, Baranoa 4, Ponedera 10 and Sabanalarga 4, in that respective order, and finally ending in Baranoa 10. Each of the routes of the mobile units determined by the routing model is shown in Appendix B.

Table VI Route of mobile Personal first aid unit

Mobile units of personal first aid	
From	**To**
Baranoa1	Baranoa10
Baranoa10	Baranoa4
Baranoa4	Ponedera10
Ponedera10	Sabanalarga4
Sabanalarga4	Baranoa1

VI. Discussion

According to the management of the flood in Santa Lucia report, the flood began November 30, 2010, and on December 3 of that same year the population was evacuated. The performance of the humanitarian work was qualified as regular since it was left to attend a large part of the population, the spread of diseases which gave became Vector-borne diseases and respiratory infections. Finally the families returned to their homes on April 3, 2011 [16]. The cause of this performance was the lack of information and the small number of specialized personnel, according to the standards proposed by [1] it is estimated that the flood was attended by four doctors, four nurses and four psychologists, in addition to the mobile units of personal first aid.

According to the above the actual management of health emergencies was regular, the reason why the proposed model aims to fill the gap of coordination of staff, taking the installation of a single point of distribution in order to unify and to analyze the information and installation of hostels as a strategic decision all while ensuring the shortest distance between them. The proposed model presents a 45.47 operation time where 90% of this time is spent on calamities relief and 10% is part of travel times, for the actual situation the estimated operating time was more than 120 days [3] for which there is no estimates for the time spent in the support of the population and the time spent on travel routes, so that the proposed model represents an approximate improvement of 60% at the time of operation.

VII. Conclusion

The article discusses the problem of the location of facilities and the routing of staff through two models characterized as allocation - location and the travel salesman capacity problem CTSP, respectively, whereas real variables of a State of emergency in the municipality of Santa Lucia, Colombia. In the case of the problem of location, given the socio-political issues in the area, the neighboring regions have a number of facilities available to accommodate the affected population, which have a risk of flooding, in the case of study proposed in the article there are 51 possible places, of which the model determines as optimal 39 shelters opening and a single point of distribution. The routing of the staff in humanitarian aid applied in the case of study warrants assistance to 8681 people, i.e., 1737 families on average, which require assistance according to the calamity that ails them. The routing model considers three types of calamities and four types of mobile units, which are allocated from a point of distribution to multiple shelters, ensuring the fastest response time and the highest percentage of relief for the affected population.

The models proposed in this article are novel since there are no investigations focused on the provision of medical and psychological assistance in post-disaster situations and that consider the location of a single point of distribution and multiple shelters from a strategic point of view. The solution provided by these models is optimal, which ensures the efficient use of human resources and reflects improvement in the response time to a flood and in the level of relief of the

population with regard to the actual situation, as shown in the case study.

Floods affect the world more and more frequently, for which it is necessary to propose new research focused on the level of satisfaction of the population affected. First implement heuristics and metaheuristics, which allow the analysis of the problem with more data. On the other hand the study of evacuation models that consider the stochastic demand behavior, as well as models of rescue in the hours following a disaster. For months after a disaster, it is important to consider socio-economic models that allow the reactivation of the economy in the affected areas and population.

APPENDIX A

Town	Calamity		
	1	2	3
Sabanalarga	1	3	50
	0	2	48
	0	0	50
	50	0	0
	2	3	45
	0	0	35
	0	0	50
	0	0	50
	0	0	50
	0	0	50
Ponedera	0	0	46
	0	0	25
	0	0	46
	0	0	2
	0	0	50
	0	0	50
	0	0	50
	0	0	50
	0	0	50
	50	0	0
Palmar de Varela	0	0	50
	0	0	50
	0	0	50
	0	0	45
	0	0	45
	0	0	28
	0	0	39
	0	0	45
	0	0	45
	0	0	50
Baranoa	0	0	45
	0	0	50
	50	0	0
	0	0	49
	0	0	50
	0	0	50
	1	0	49
	0	0	50
	50	0	0

APPENDIX B

Medical Unit	
From	**To**
Baranoa1	Ponedera4
Ponedera4	Ponedera6
Ponedera6	Ponedera8
Ponedera8	Palmardevarela6
Palmardevarela6	Palmardevarela4
Palmardevarela4	Palmardevarela1
Palmardevarela1	Palmardevarela2
Palmardevarela2	Palmardevarela7
Palmardevarela7	Palmardevarela10
Palmardevarela10	Palmardevarela9
Palmardevarela9	Palmardevarela3
Palmardevarela3	Palmardevarela5
Palmardevarela5	Palmardevarela8
Palmardevarela8	Baranoa2
Baranoa2	Baranoa3
Baranoa3	Baranoa6
Baranoa6	Baranoa5
Baranoa5	Baranoa9
Baranoa9	Baranoa7
Baranoa7	Baranoa8
Baranoa8	Baranoa1

Mobile units of personal first aid	
From	**To**
Baranoa1	Sabanalarga6
Sabanalarga6	Sabanalarga3
Sabanalarga3	Sabanalarga1
Sabanalarga1	Sabanalarga8
Sabanalarga8	Sabanalarga7
Sabanalarga7	Sabanalarga5
Sabanalarga5	Sabanalarga2
Sabanalarga2	Ponedera2
Ponedera2	Ponedera3
Ponedera3	Ponedera9
Ponedera9	Ponedera1
Ponedera1	Ponedera7
Ponedera7	Ponedera5
Ponedera5	Sabanalarga10
Sabanalarga10	Sabanalarga9
Sabanalarga9	Baranoa1

ACKNOWLEDGMENT

The authors thank the sponsorship of this project to the Master in Operations Management and the International School of Economics and Management Sciences (EICEA) of the Universidad de La Sabana.

REFERENCES

[1] SCHR/VOICE/ICVA, "The Sphere Project Humanitarian and Minimum Standards in Disaster", Geneva, Switzerland, 2004. SCHR/VOICE/ICVA, "El Proyecto Esfera Humanitaria y Normas mínimas de respuesta humanitaria en casos de desastre" Ginebra, Suiza, 2004.

[2] USAID/OFDA, "Damage Assessment and Needs Analysis" *www.colombiassh.org*, 2008. USAID/OFDA, "Evaluación de Daños y Análisis de Necesidades", 2008. [Online]. Available: http:// *www.colombiassh.org*. [Accessed: 15-Aug-2014].

[3] T. Street, "Hospitals Safe from floods", Washington, D.C, United States, 2006. T. Street, "Hospitales seguros ante inundaciones", Washington, D.C, Estados Unidos, 2006.

[4] "EM-DAT. The International Disaster Database. Centre for research on the Epidemiology of Disasters. CRED," 2013.

[5] IFRC, "Tecnical report: World Disasters" 2010. IFRC, "Reporte técnico: Mundial sobre Desastres" 2010

[6] V. Gracía G., "International Strategy for Disaster Reduction (ISDR)", *Magazine information ISDR Latin America and the Caribbean*, 2002. [Online]. Available: http://www.eird.org/esp/revista/No6_2002/art13.htm. [Accessed: 15-Aug-2014]. V. Gracía G., "International Strategy for Disaster Reduction (ISDR)", *Revista EIRD informa- America Latina y el Caribe*, 2002. [Online]. Available: http://www.eird.org/esp/revista/No6_2002/art13.htm. [Accessed: 15-Aug-2014].

[7] A. M. Caunhye, X. Nie, and S. Pokharel, "Optimization models in emergency logistics: A literature review", *Socioecon. Plann. Sci.*, vol. 46, no. 1, pp. 4–13, Mar. 2012.

[8] A. J. Pedraza, O. Stapleton, and L. N. Van Wassenhove, "Field vehicle fleet management in humanitarian operations: A case-based approach", *J. Oper. Manag.*, vol. 29, no. 5, pp. 404–421, Jul. 2011.

[9] M. Huang, K. Smilowitz, and B. Balcik, "Models for relief routing: Equity, efficiency and efficacy", *Transp. Res. Part E Logist. Transp. Rev.*, vol. 48, no. 1, pp. 2–18, Jan. 2012.

[10] B. Balcik and B. M. Beamon, "Facility location in humanitarian relief", *Int. J. Logist. Res. Appl.*, vol. 11, no. 2, pp. 101–121, Jan. 2008.

[11] M. M. Dessouky and F. Ordonez, "Rapid Distribution of Medical Supplies", *Int. Ser. Oper. Res. Manag. Sci.*, vol. 91, no. 1, pp. 309–338, 2009.

[12] H. O. Mete and Z. B. Zabinsky, "Stochastic optimization of medical supply location and distribution in disaster management", *Int. J. Prod. Econ.*, vol. 126, no. 1, pp. 76–84, Jul. 2010.

[13] J.-B. Sheu, "An emergency logistics distribution approach for quick response to urgent relief demand in disasters", *Transp. Res. Part E Logist. Transp. Rev.*, vol. 43, no. 6, pp. 687–709, Nov. 2007.

[14] R. Tomasini and L. Van Wassenhove, Humanitarian Logistics, 1st ed. Great Britain: Palgrave Macmillan, 2009, pp. 1–193.

[15] J. Holguín-Veras, N. Pérez, M. Jaller, L. N. Van Wassenhove, and F. Aros-Vera, "On the appropriate objective function for post-disaster humanitarian logistics models", *J. Oper. Manag.*, vol. 31, no. 5, pp. 262–280, Jul. 2013.

[16] Departmental Comptroller of the Atlantic, "State of natural resources and environment", Barranquilla, Colombia. Contraloría Departamental del Atlántico, "Estado de los recursos naturales y del medio ambiente en el departamento del atlántico", Barranquilla, Colombia.

[17] Bank of the Republic of Colombia "Tecnical Report: Working papers on the regional economy", Cartagena, Colombia, 2011. Banco de la República de Colombia, "Reporte Técnico: "Documentos de trabajo sobre la economía regional", Cartagena, Colombia, 2011.

A Mechanically and Incremental Development of the Remote Authentication Dial-In User Service Protocol

Sanae El Mimouni, Rajaa Filali, Anas Amamou, Bahija Boulamaat and Mohamed Bouhdadi

***Abstract*—**The Remote Authentication Dial-In User Service (RADIUS) protocol is a distributed client/server protocol that protects networks against unauthorized access. RADIUS uses User Datagram Protocol (UDP) as transfer protocol and has good capability for real-time applications. It also supports retransmission mechanism and backup server mechanism so that it boasts better reliability. RADIUS is easy to implement, and applicable to the multithreading structure of the server in the time of mass users. While RADIUS is an excellent protocol for it uses, it has never been formally specified. We try to fill this gap by giving a fully formal specification of the protocol using event B method. Event-B is a formal method for system-level modeling and analysis. Event-B is provided with tool support in the form of an Eclipse-based IDE called Rodin. The core of the Rodin tool provides automatic generation of proof obligations that can be analyzed to improve understanding of a model. Our Specification is very general and contains basic message exchange process of RADIUS Client/server.

***Keywords*—**Event-b,Formal specification, RADIUS, Refinement.

I. Introduction

Originally created by Livingston Enterprise which was later acquired by Lucent , and as defined by IETF‘s RFC 2865 (RADIUS authentication and authorization) and RFC 2866 (RADIUS accounting), RADIUS is based on the client-server model and message exchanges takes place over User Datagram Protocol (UDP). The Network Access Server (NAS) acts as a RADIUS client which passes on the user request to the RADIUS server. The other RADIUS clients may be wireless access points, routers, and switches. The RADIUS server performs authentication, authorization, and accounting (AAA) for users after it receives requests from the client.The communication between the client and the server is encrypted using a private key which is never sent over the network. Both the client and server are configured with this secret before communication can take place, and it fails if the secret does not match at both ends.

Even with the practical significance of RADIUS protocol, unfortunately there isn’t a formal specification for it like as done to CSMA/CD Protocol using model checking [8] .So we try to present a formal approach for the protocol. We developed our model specification in Event-B[1].We liberally used refinements, both of machines and of contexts. We give a great deal of attention to proofs. Consequently, we now have a specification of RADIUS protocol where all proof-obligations have been discharged.

The RADIUS protocol was first defined in RFC 2058 [13], in January 1997, this RFC contains proposed standard. Also in January 1997 RADIUS accounting was introduced in RFC 2059 [10], status of which is informational. Later in April 1997 these RFCs were obsolete by RFC 2138 [14] and RFC 2139 [11]. Former of these is proposed standard and latter informational. Then in June 2000 RFC 2865 [15] defined RADIUS draft standard and obsoleted RFC 2138. In same month informational RFC 2866 [12] RADIUS accounting obsoleted RFC 2139.For our article we based on the RFC 2865.

This paper is organized as follows. In Section 2 we will give an informal introduction to the RADIUS protocol, and a brief description of the event B method. The main part of this paper, Section 3 describes our strategy of refinement, Moreover we will specify our protocol using event B. Section 4 summarizes the results and draws a conclusion.

II. Basic Concepts

In this section, we provide some background information on the RADIUS protocol, and the Event-B formal method.

A. RADIUS protocol

The Remote Authentication Dial-in User Service (RADIUS) [5] is an IETF-defined Client/server protocol and software that enables remote access servers to communicate with a central server to authenticate dial-in users and authorize their access to the requested system or service [5]. It is commonly used to provide centralized Authentication, Authorization, and Accounting (AAA) for dial-up, virtual private network, and,

Sanae El Mimouni is with LMPHE laboratory, University of Mohammed V, Faculty of sciences, Rabat, Morocco (e-mail: sanae.elm@ gmail.com).

Rajaa Filali is with LMPHE laboratory, University of Mohammed V, Faculty of sciences ,Rabat ,Morocco(e-mail: rajaafilali@gmail.com).

Anas Amamou is with LMPHE laboratory, University of Mohammed V, Faculty of sciences, Rabat, Morocco (e-mail: amamou.anas@yahoo.fr).

Bahija Boulamaat is with LMPHE laboratory, University of Mohammed V, Faculty of sciences, Rabat, Morocco (e-mail: boulamaatbahija@gmail.com).

Mohamed Bouhdadi is with LMPHE laboratory, University of Mohammed V, Faculty of sciences, Rabat, Morocco (e-mail: bouhdadi@fsr.ac.ma).

wireless network access.

The RADIUS protocol is based on a Client/server model. A Network Access Server (NAS) operates as a client of RADIUS. The client is responsible for passing user information to designated RADIUS servers, and then acting on the response which is returned.

RADIUS servers are responsible for receiving user connection requests, authenticating the user, and then returning all configuration information necessary for the client to deliver service to the user.

A RADIUS server can act as a proxy client to other RADIUS servers or other kinds of authentication servers.

The operation of the RADIUS protocol involves six types of message exchanges between the client and the server, as described in the following sections and a simple procedure of RADIUS communication is shown in the figure 1:

• *Access-Request:* Sent by a RADIUS Client to request authentication and authorization for a network access connection attempt. It determines whether a user is allowed access to a specific NAS, and any other specific service.

• *Access-Accept:* Sent by a RADIUS server in response to an Access-Request message when all conditions are met. The message informs the RADIUS Client that the connection attempt is authenticated and authorized and it contains the list of configuration values for the user.

• *Access-Reject:* Sent by a RADIUS server in response to an Access-Request message if any condition is not met. This message informs the RADIUS Client that the connection attempt is rejected. A RADIUS server sends this message if either the credentials are not authentic or the connection attempt is not authorized.

• *Access-Challenge:* Sent by a RADIUS server in response to an Access-Request message if all conditions are met and RADIUS server wishes to issue a challenge to which the user must respond. The Client in response resubmits its original Access-Request with a new request ID, response (encrypted), and including the Attribute from the Access-challenge.

• *Accounting-Request:* Sent by a RADIUS Client to specify accounting information for a connection that was accepted.

• *Accounting-Response:* Sent by the RADIUS server in response to the Accounting-Request message. This message acknowledges the successful receipt and processing of the Accounting-Request message.

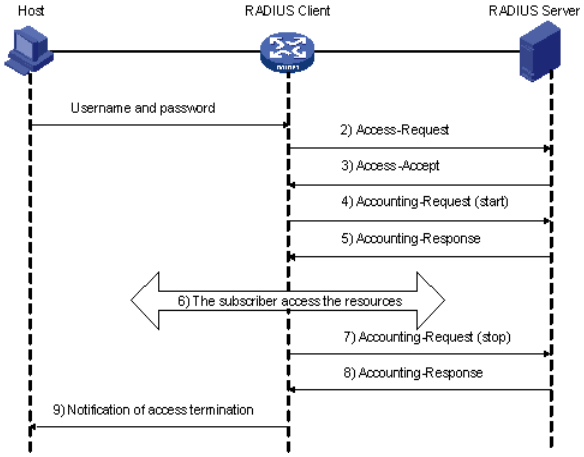


Fig. 1 Basic message exchange process of RADIUS

The following shows how RADIUS operates as shown in the figure above:

1. The user enters the username and password.

2. Having received the username and password, the RADIUS client sends an authentication request (Access-Request) to the RADIUS server.

3. The RADIUS server compares the received user information with that in the Users database. If the authentication succeeds, it sends back an Access-Accept message containing the information of user's right. If the authentication fails, it returns an Access-Reject message.

4. The RADIUS client accepts or denies the user according to the returned authentication result. If it accepts the user, it sends an accounting start request (Accounting-Request) to the RADIUS server, with the value of Status-Type being "start".

5. The RADIUS server returns a start-accounting response (Accounting-Response).

6. The subscriber accesses the network resources.

7. The RADIUS client sends a stop-accounting request (Accounting-Request) to the RADIUS server, with the value of Status-Type being "stop".

8. The RADIUS server returns a stop-accounting response (Accounting-Response).

9. The subscriber stops network resource accessing.

In this paper we model a simple RADIUS procedure of communication without considering accounting messages.

B. Event B method

Event-B is a formal method for specifying, modeling and reasoning about systems. An evolution of the (classical) B-Method developed by Jean-Raymond Abrial [2]. Event-B is now centered on the general notion of events, which also found in other formal methods such as Action Systems [3] [4], TLA [9] and UNITY [5].

Event-B is a formal modeling method for developing systems via step-wise refinement, based on first-order logic. Event-B models are organized in terms of two basic components: contexts and machines. Machines and contexts can be inter-related: a machine can be refined by another one, a context can be extended by another one and a machine can

see one or several contexts as shown in figure 2.

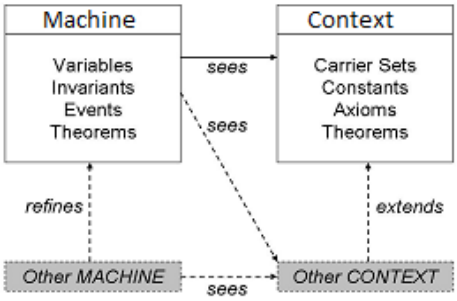


Fig. 2 Event-B Machines and Contexts

- Contexts specify the static part of a model. They may contain carrier sets (similar to types), constants, axioms (containing carrier sets and constants), and theorems (expressing properties derivable from axioms).
-Machines specify behavioral properties of the models. They may contain variables defining the state of a machine, invariants constraining that state, and events (describing possible state changes). Each event is composed of a set of guards and a set of actions. Guard state the necessary conditions under which an event may occur, and actions describe how the state variables evolve when the event occurs.

Contexts/Machines may be refined from more abstract to more concrete contexts/machines. Event-B models are systematically structured in refinement chains.

A key concept in Event-B is proof-obligation (PO) capturing the necessity to prove some internal property of the model such as typing, invariant preservation by events, and correct refinements. Strong tool support is provided in order to support this proof process.

Event-B is not specific to embedded systems design but it is currently being investigated by several industrial from different sectors (automotive, transportation, space) in the context of the DEPLOY project [6].

In Event-B, an event is defined by the syntax: EVENT e WHEN G THEN S END , Where G is the guard, expressed as a first-order logical formula in the state variables, and S is any number of generalized substitutions, defined by the syntax S ::= x := E(v) | x := z : | P(z). The deterministic substitution, x := E(v), assigns to variable x the value of expression E(v), defined over set of state variables v. In a non-deterministic substitution, x := z : | P(z), it is possible to choose non-deterministically local variables, z, that will render the predicate P(z) true. If this is the case, then the substitution, x := z, can be applied, otherwise nothing happens.

It is also important to indicate that the most important feature provided by Event-B is its ability to stepwise refine specifications. Refinement is a process that transforms an abstract and non-deterministic specification into a concrete and deterministic system that preserves the functionality of the original specification. During the refinement, event descriptions are rewritten to take new variables into account. This is performed by strengthening their guards and adding substitutions on the new variables. New events that only assign the new variables may also be introduced. Proof obligations (POs) are generated to ensure the correctness of the refinement with respect to the abstract model. Event-B is supported by several tools, currently in the form a platform called Rodin.

Rodin is an open-source development platform for Event-B. It provides an environment for system modeling and analyses, including support for refinement, i.e. POs are generated automatically between abstraction levels, and support for mathematical proof, i.e. most POs can be discharged automatically or manually. More teaching materials on Event-B and Rodin can be found at [7].

III. Specifying RADIUS protocol using Event B

A. *Refinement strategy*

In this short section, we present our strategy for constructing the RADIUS protocol specially the message type exchanges that take place between the Client and the Server, which is shown in the figure below. This will be done by means of an initial model followed by one refinement.

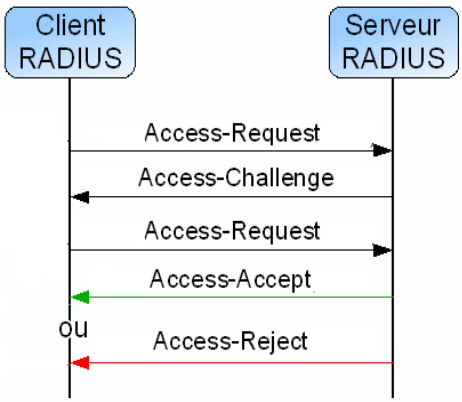


Fig. 3 Simple procedure of RADIUS communication

- The initial model essentially presents message exchange between the client and the server without considering any condition.
- In the first refinement, we introduce the condition that take side the client status and we add a timer.

B. *Initial Model*

The initial model of RADIUS protocol is presented as follow:

The context is made of two sets Requests and the Responses. These sets represent the message type exchanges that take place between the Client and the Server. Which are Access_Request, Access_Accept, Access_Challenge, and Access_Reject.

ISBN: 978-1-61804-256-9

```
SETS
    Requests      ›
    Responses     ›
CONSTANTS
    Access_Request    ›
    Access_Accept     ›
    Access_Reject     ›
    Access_Challenge      ›
AXIOMS
    axm1:   Access_Request ∈  Requests
    axm2:   Access_Accept ∈ Responses
    axm3:   Access_Reject ∈ Responses
    axm4:   Access_Challenge ∈ Responses
END
```

Fig. 4 Carrier Sets

When the client chooses to use RADIUS, it creates an "Access_Request" containing some information and sends it to the server side. We do not discuss in this paper the information that is in the message; we just focus about the operation that happened between the client and the server.

```
clt_access_request:
ANY
    msg   ›
WHERE
    grd1:   msg = Access_Request
    grd2:   msg ∉ paquet_client
THEN
    act1:   paquet_client ≔ paquet_client ∪ {msg}
END

srv_access_accept:
ANY
    msg   ›
WHERE
    grd1:   msg = Access_Accept
    grd2:   msg ∉ paquet_server
THEN
    act1:   paquet_server ≔ paquet_server ∪ {msg}
END

srv_acces_challenge:
ANY
    msg
WHERE
    grd1:   msg = Access_Challenge
    grd2:   msg ∉ paquet_server
THEN
    act1:   paquet_server ≔ paquet_server ∪ {msg}
END

srv_access_reject:
ANY
    msg
WHERE
    grd1:   msg = Access_Reject
    grd2:   msg ∉ paquet_server
THEN
    act1:   paquet_server ≔ paquet_server ∪ {msg}
END
```

Fig. 5 Events of initial model

C. *First refinement*

We are going to refine our abstract model to a more concrete one, by adding new variables and modifying our existing events. For this we introduce the client status and a timer. We define a carrier set named STATUS. It is made of three distinct elements: valid, invalid, moreinfo, which present the RADIUS client status.

```
SETS
    Statut
CONSTANTS
    valid
    invalid
    moreinfo
AXIOMS
    axm1:   Statut ={valid, invalid,moreinfo}
    axm2:   valid≠ invalid
    axm3:   moreinfo ≠ invalid
    axm4:   moreinfo ≠ valid
END
```

Fig. 6 Carrier Set statut

The Access-Request is submitted to the RADIUS server via the network. If no response is returned within a length of time, the request is re-sent a number of times.

```
clt_access_request:
REFINES
    clt_access_request
ANY
    msg   ›
WHERE
    grd1:   msg = Access_Request
    grd2:   msg ∉ paquet_client
    grd3:   Time = FALSE
THEN
    act1:   paquet_client ≔ paquet_client ∪ {msg}
    act2:   Time ≔ TRUE ›
END
```

Fig. 7 Modified event Access_request

If the client is valid then the RADIUS server sends Access-Accept response to the client.

```
srv_access_accept:
REFINES
    srv_access_accept
ANY
    msg   ›
WHERE
    grd1:   msg = Access_Accept
    grd2:   msg ∉ paquet_server
    grd3:   client_st = valid
THEN
    act1:   paquet_server ≔ paquet_server ∪ {msg}
END
```

Fig. 8 Modified event Access_accept

If any condition is not met, the RADIUS server sends an "Access-Reject" response indicating that this user request is invalid.

ISBN: 978-1-61804-256-9

```
srv_access_reject :
REFINES
    srv_access_reject
ANY
    msg  ›
WHERE
    grd1:   msg = Access_Reject
    grd2:   msg ∉ paquet_server
    grd3:   client_st = invalid
THEN
    act1:   paquet_server ≔ paquet_server ∪ {msg}
END
srv_access_challenge :
REFINES
    srv_acces_challenge
ANY
    msg  ›
WHERE
    grd1:   msg = Access_Challenge
    grd2:   msg ∉ paquet_server
    grd3:   client_st = moreinfo
THEN
    act1:   paquet_server ≔ paquet_server ∪ {msg}
END
```

Fig. 9 Modified event Acces_challenge and reject

The server can respond to this new Access- Request with either an Access-Accept, an Access-Reject, or another Access-Challenge.

The last event in our model is the vent of timing.

```
time :
WHERE
    grd1:   Time = TRUE
THEN
    act1:   T ≔ T+1  ›
END
```

Fig. 10 Event time

IV. Conclusion

In this paper we have presented formal modeling of the RADIUS protocol using Event B.

In this approach the modeling process starts with an abstraction of the protocol which specifies the goals of the protocol. In our case study, presents message exchange between the client and the server without considering any condition are the main protocol goals. The abstract level of our Event-B model shows these goals in a very general way, and then during refinement level, features of the protocol are modeled and the goals are achieved in a detailed way.

The use of Event-B and Rodin as a formal modeling environment has several advantages. Firstly, the model can be gradually developed by step-wise refinements, which allows hierarchical design exploration at different abstraction levels. Secondly, the obligation to discharge POs ensures full model consistency throughout all levels.

References

[1] J.-R. Abrial, Modeling in Event-B: System and Software Engineering. Cambridge University Press, 2010.

[2] J.-R. Abrial, The B-Book: Assigning Programs to Meanings, Cambridge University Press, 1996.

[3] R.-J. Back," Decentralization of process nets with centralized control".2nd ACM SIGACT–SIGOPS Symposium on Principles of Distributed Computing, 1983.

[4] R.-J. Back, Refinement Calculus II: Parallel and Reactive Programs. In: de Bakker J. W., de Roever W. P., Rozenberg G. (eds.), Lecture Notes in Computer Science, Springer, vol 430, pp. 67-93, 1990.

[5] K. Chandy, J. Misra, Parallel Program Design: a Foundation, Addison-Wesley, 1989.

[6] DEPLOY FP7 Project, [Online].Available: http://www.deploy-project.eu , January 2014.

[7] Event-B and RODIN. Available: http://wiki.event-b.org, April 2011.

[8] M. Sirjani, M.M. Jaghoori, S. Forghanizadeh, M. Mojdeh, and A. Movaghar. Model Checking CSMA/CD Protocol using an Actor-Based Language, in the Proceedings of the International Conference on Software Engineering, WSEAS, February 2004.

[9] L. Lamport, The temporal logic of actions, Transactions on Programming Languages and Systems (TOPLAS), vol.16 no.3, pp. 872-923, 1994.

[10] C. Rigney, RFC 2059: Radius Accounting [Online].Available: www.ietf.org/rfc/rfc2059.txt , January 1997.

[11] C. Rigney, RFC 2139: Radius Accounting [Online].Available: www.ietf.org/rfc/rfc2139.txt, April 1997.

[12] C. Rigney, RFC 2866: Radius Accounting [Online].Available: www.ietf.org/rfc/rfc2866.txt, June 2000.

[13] C. Rigney, A. Rubens, W. Simpson and S. Willens, RFC 2058: Remote Authentication Dial In User Service (RADIUS). Available: www.ietf.org/rfc/rfc2058.txt, January 1997.

[14] C. Rigney, A. Rubens, W. Simpson and S. Willens, RFC 2138: Remote Authentication Dial In User Service (RADIUS). Available: www.ietf.org/rfc/rfc2138.txt, April 1997.

[15] C. Rigney, A. Rubens, W. Simpson and S. Willens, RFC 2865: Remote Authentication Dial In User Service (RADIUS). Available: www.ietf.org/rfc/rfc2865.txt , June 2000.

Mathematical optimization of powder composition for improved hardness of titanium alloy coated by cold spraying

Damilola I Adebiyi, Patricia A. Popoola, and Ionel Botef

***Abstract*—**The proportion of each of the three powders used in cold spray of titanium alloy was simulated to determine the optimum percentage composition for improved hardness. In cold spray coating, the ability to expeditiously predict the effect of change in powder type or proportion on the properties of the coating is usually advantageous. If the desired property is basically decided by the coating powder composition, an optimization methodology specific to the design of mixture experiments can be successfully used. In the present study, the percentage compositions of titanium, nickel and aluminium were optimized using mixture experimental design and analysis of variance. The determination coefficient, R square, = 96.4%, and the adjusted determination coefficient, adjusted R square = 83.9%, signify a good fit and high statistical significance of the model. The empirical relationship between the hardness and the percentage composition of the powders shows that titanium has the highest coefficient and contributed the most to the overall hardness..

***Keywords*—**Cold spray, optimum percentage, hardness, regression analysis

I. Introduction

IN coating with powders, mixtures of powders with different properties are often used in order to achieve optimum properties. It is critical when working with powder mixtures to use statistical approach to determine the optimum percentage composition for a specific property. According to Muteki [1], there are basically three general degrees of freedom to control the final properties of any product manufactured in a blending operation: (1) the selection of raw materials; (2) the ratios in which to blend them; and (3) the process conditions used to manufacture them. If processing parameters are kept constant, the enhancement in the surface properties, hardness for example, of a material after cold spraying basically depends on the composition of the pre-mixed powder used. This is the idea behind the use of mathematical and statistical techniques to design mixture composition and predict the property response. This means that for a given set of powder, a mathematical relationship could be written for any property response to relate such property with the percentage composition of the powder in the starting mixture

It is a usual practice in the industry to find the optimal percentage quantities of the components which gives the best satisfying characteristics in a mixture [2]. This is called mixture design. The usual practice is to determine the optimum composition via traditional optimization which generally leads to constant reformulation of the percentage powder composition through try and error. This consumes time and materials, and it is quite expensive because it often requires an excessively large number of samples [3]. Moreover, the overall efficiency in traditional optimization is usually quite low because it is virtually impossible to precisely find the optimum point using non-systematically selected samples [4].

Hence the need to apply Design of Experiments (DoE) techniques to determine the optimal composition of a powder mixture for cold spraying. Mixture design makes it possible to identify the synergetic effect of mixing two or more components on a property of interest. This will lead to experiments with high success rates. A method called the simplex design method, which uses theory of statistics and experiments to obtain models that can be used to optimize the composition of mixture for a specified response variable was developed by Scheffe in 1958 [6]. Mixture design has been used in the metallurgy to optimize the composition of alloy and metal matrix composites, and to design and predict material properties [7]-[11]

The purpose of this work is to employ the use of statistical experiment design approach in proportioning the optimum weight percent of the titanium, nickel and aluminum powders for increase in hardness of cold spray coating of titanium alloy (Ti-6Al-4V).

This material is based upon work supported financially by the National Research Foundation. The cold Spray Laboratory of the University of Witwatersrand, Johannesburg is appreciated for Cold Spray facilities, The authors also acknowledge the support from Tshwane University of Technology Pretoria, South Africa which helped to accomplish this work

D. I. Adebiyi is a doctoral candidate of the Department of Chemical, Metallurgical and Materials Engineering, The Tshwane University of Technology, Pretoria, South Africa (e-mail: AdebiyiDI@tut,ac,za

I. Botef is with the University of Witwatersrand, School of Mechanical, Industrial and Aeronautical Engineering, Johannesburg, South Africa (email: Ionel.botef@wits.ac.za)

A. P. I. Popoola is with the Department of Chemical, Metallurgical and Materials Engineering, The Tshwane University of Technology, Pretoria, South Africa (e-mail: PopoolaAPI@tut,ac,za)

II. MATHEMATICAL MODELING

The optimum powder mixture is defined as that mixture which minimizes cost while maximizing the property. Two experiment design approaches can be applied to optimize the pre-mix powder ratio. These are: (1) the classic mixture approach, in which the *q* mixture components are the variables, and (2) the mathematically independent variable (MIV) approach, in which *q* mixture components are transformed into *q-1* independent mixture-related variables. In the classic mixture approach, the sum of the proportions must be 1. The steps involved in mixture experiments can be found in other literature [4], [12].

A. Model Assumptions

In order to conform the mixture experiment to powder optimization for cold spray coating of titanium alloy, the following assumptions are made:

1) The input factors can be controlled by the experimenter.
2) The level of input factors can be determined, controlled and varied. The relevance of this is that the quantity of each powder in the powders mix can be determined and varied.
3) The response depends on the proportions of the individual powder in the premixed ratio and not on the total amount of mixture. Similarly, the improvement in hardness of the titanium alloy after cold spraying depends on the percentage composition of the premixed ratio.
3) This response (improvement in hardness of the titanium alloy) is measurable

.

B. Modeling Procedure

The approach used for determining the optimum percentage composition of the powders is described as follows:

1) The main objective was decided. This is to determine the optimum percentages of each of titanium, nickel and aluminium that will give the best hardness for cold spray coating of titanium alloy.
2) The special cubic Scheffé [6] canonical equation was used to model the hardness data of each of the powder.
3) The design matrix for the powder blend was developed with proportion defined by the simplex-centroid experimental design for mixtures of three components expanded with internal points using D-optimal mixture experiments. 10-run D-optimal design were generated.
4) A statistical analysis tool known as XLSTAT was used to determine the optimal proportion of each of the powders influencing the hardness of the coating. Thus, an empirical relationship between the titanium, nickel and aluminum powders was obtained.

The percentage composition of the powder blend was optimized for maximum hardness.

C. Model Equation

Let the property response (hardness) be denoted by *f*. According to Myers and Montgomery [13] and Cornell [14], the property response, *f*, can be expressed in its canonical form as a typical low degree polynomial of the first or second degree. The polynomial expressed in terms of the weight fractions, *x*, of the three powders which will sum to unity.

Consider a powder mixture consisting of 3 component materials

$$f = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad (1)$$

$$f = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 \quad (2)$$

It is generally preferable to evaluate such these polynomial equations over a number of points, *N* greater than the number of components. This is to afford the representation of the response surface over the entire area using a regular array of uniformly spaced points called *a lattice*. This lattice is referred to as a {*q*, *m*} simplex lattice, *m* being the spacing parameter in the lattice and *q* the number of components. In the standard mixture designs, the proportions of the ingredients can vary between 0 and 1 and must sum up to unity.

Let the number of powder in the premixed ratio be q (q = 3). x_i is the percentage weight of the *ith* powder in the mixture, the constraints in equations (3) and (4) must be satisfied. These constraints keep each mixture component proportion between 0% and 100% (0 and 1), also ensure that at any point in the mixture space, the total sum of the proportions of all the components adds up to unity

$$\sum_{i=1}^{q} x_i = 1 \quad (3)$$

$x_i \geq 0, (i = 1,2,3.....q)$ = the concentration of the component

$$x_i \geq 0, (i = 1,2,3.....q), x_1 + x_2 ..... + x_q = 1 \quad (4)$$

Where

x_1 = Proportion of Titanium powder

x_2 = proportion of Nickel powder

x_3 = Proportion of aluminium powder

The number of points in the design is given by

$$(m + \frac{q}{m} - 1) = \frac{q(q+1).....(q+m-1)}{1.2.....m} \quad (5)$$

According to Sheffe [6], a {q, n} polynomial function can be written for equation (3) as shown in equation (4). Let the degree of the polynomial be n in q variable

$$\eta = \beta_0 + \sum_{1 \leq i \leq q} \beta_i x_i + \sum_{1 \leq i \leq j \leq q} \beta_{ij} x_i x_j + \sum_{1 \leq i \leq j \leq k \leq q} \beta_{jki} x_i x_j x_k ..... + ... \sum_{1 \leq i \leq q} \beta_{i_1 i_2 .....i_n} x_{i_1} x_{i_2 .....x_{i_n}} \quad (4)$$

The β coefficients are constant and the number of coefficients is given by

$$\left(\frac{n+q}{n} \right)$$

But since equation (1) is true, the equation component is thus eliminated and equation (3) is reduced to:

$C^n_q + n - 1$

According to Sheffe [6], if Equation (4) is reduced subject to the normalization condition of Equation (1), the properties of the mixture can be described for a sum of independent variables. The reduced second-degree polynomial for a ternary system is derived as follows

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{12} x_1 x_2 + \beta_{11} {x_1}^2 . + \beta_{22} {x_2}^2 + \beta_{33} {x_3}^2 \quad (5)$$

Since $x_1 + x_2 + x_3 = 1$, then,

$$\beta_0 = \beta_0 (x_1 + x_2 + x_3) \quad (6)$$

i.e.

$$\beta_0 = \beta_0 x_1 + \beta_0 x_2 + \beta_0 x_3 \quad (7)$$

If equation (5) is multiplied by x_1, x_2, x_3 in succession, equations (8), (9) and (10) are obtained

$${x_1}^2 = x_1 - x_1 x_2 - x_1 x_3 \quad (8)$$

$${x_2}^2 = x_2 - x_1 x_2 - x_2 x_3 \quad (9)$$

$${x_3}^2 = x_3 - x_1 x_3 - x_2 x_3 \quad (10)$$

Substituting equation (8)-(10) into (5), equation (11) is obtained

$$\eta = (\beta_0 + \beta_1 + \beta_{11}) x_1 + (\beta_0 + \beta_2 + \beta_{22}) x_2 + (\beta_0 + \beta_3 + \beta_{33}) x_3 + (\beta_{12} - \beta_{11} - \beta_{22}) x_1 x_2 + (\beta_{13} - \beta_{11} - \beta_{33}) x_1 x_3 + (\beta_{23} - \beta_{22} - \beta_{33}) x_2 x_3 \quad (11)$$

A reduced second degree polynomial in three variables can be obtained from equation (11) if the following conditions are indicated:

$$\beta_1 = b_0 + b_i + b_{ii} \quad (12)$$

$$\beta_{ij} = b_{ii} - b_{ij} \quad (13)$$

Thus, if $\beta_i = \beta_1, \beta_2, \beta_3$, $\beta_{ij} = \beta_{12}, \beta_{13}, \beta_{23}$ and η_i and η_{ij} are response property, the reduced polynomial will be obtained as in equation (14)

$$\eta = \beta + \beta_1 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 \quad (14)$$

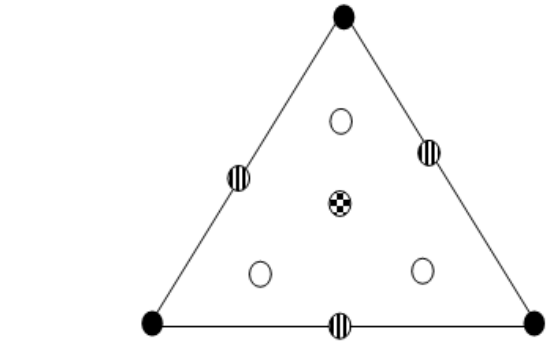
Where, η is the response (i.e. hardness), x_1, x_2 and x_3 are the coded values of the variables and $\beta_1, \beta_2, \beta_3$ etc. are the regression coefficients. The coefficient of the polynomial equation (14) is described as the solution of equation (12) and (13) as:

$$\beta_i = \eta_i \quad (15)$$

and

$$\beta_{ij} = 4\eta_{ij} - 2\eta i \ - 2\eta_j \quad (16)$$

Equation (14) is the governing equation which is the response function for optimization of pre-mixed coating powder consisting of three components (powders). This is represented in Figure 1. The three single components run at the vertices, the three binary blends on the edges, and the ternary blend in the middle. The three additional blend provide checkpoints for evaluation of lack-of-fit



Symbol	Variation in proportion
●	Pure powder {(1, 0, 0), (0,1,0), (0,0,1)
◍	Binary blend {(½, ½, 0), (0, ½, ½), (½, ½, 0)
◈	Ternary blend {1/3, 1/3, 1/3}
○	Check points {(2/3, 1/6, 1/6), (1/6, 2/3, 1/6), (1/6, 1/6, 2/3)}

Fig. 1: Notation for response on the lattice

III. Results and Discussion

A Evaluation of the model coefficients

In order to evaluate the model coefficients and validate the model, the proportions of the components were expressed in percentage form. No constraints or limit (upper/lower) were set for the proportion of any of the powder. The following are the hardness of the powders: titanium is 238BN, nickel, 80 BN and aluminium, 55 BN. The result of the D-optimal design is presented in Table 1

An empirical relationship was developed between resultant hardness and the hardness of each powder as shown in (17)

$$\text{Hardness} = 17091.2 + 17352.6X_1 + 17159.1X_2 + 17134.8X_3 + 1294.8X_1X_2 + 1198.27X_1X_3 - 45.0607X_2X_3 - 5201.59X_1X_2X_3 \quad (17)$$

B Validation of model

The values of the determination coefficient, (R square) and adjusted determination coefficient (adjusted R square) are respectively 96.4% and 89.3%. These results signify a good fit and high statistical significance of the model. Analysis of (17) shows that titanium has the highest coefficient and contributed the highest hardness to the overall hardness.

Table 1: D-optimal design mixture designs with no constraints.

	Ti	Ni	Al					
No.	x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$	**Hardness**
1	1	0	0	0	0	0	0	238
2	0.5	0.5	0	0.25	0	0	0	477
3	0	1	0	0	0	0	0	80
4	0	0.5	0.5	0	0	0.25	0	68
5	0.5	0	0.5	0	0.25	0	0	440
6	0	0	1	0	0	0	0	55
7	0.167	0.167	0.667	0.0279	0.1114	0.1114	0.018602	135
8	0.667	0.167	0.167	0.1114	0.1114	0.0279	0.018602	460
9	0.167	0.667	0.167	0.1114	0.0279	0.1114	0.018602	153
10	0.333	0.333	0.333	0.1109	0.1109	0.1109	0.036926	187

The optimal percentage composition of the powders was found to be 55% titanium, 40% nickel and 5% aluminium. This composition yielded the optimal hardness of 489 BHN. Literatures are hardly available on this specific application of mixture design. Hence, a comprehensive evaluation, and comparison with previous work, could not be carried out.

IV CONCLUSION AND FUTURE WORK

The optimum fractional composition of titanium, nickel and aluminium for the improvement of the hardness of Ti6Al4V has been determined using mixture experiment design. The practical efficiency and usefulness of the proposed method will be verified experimentally. The hardness of 489 BHN was obtained from the optimal percentage composition at 55% titanium, 40% nickel and 5% aluminium. The effect of process parameters could not be established. Particle size distribution, cold wedding during powder mixing and processing conditions and precipitation of secondary phases are all relevant and their effect will be investigated in future works. This model is suitable for the selection of the optimum percentage composition of powder blend for obtaining powder mixture with desired hardness. A more detailed study is however required to gain a deeper insight into the behavior of the powder mixture.

REFERENCES

[1] K. Muteki, J. F. MacGregor, and T. Ueda, "Mixture designs and models for the simultaneous selection of ingredients and their ratios, " *Chem. Intel. Lab. Systems* vol. 86 pp. 17–25, 2007

[2] D. Davidović, D. Letić, V. Petrović, I. Berković, B. Radulović, and D. Z. Zivković, "The designing of the four – component composition of the blend of the polymer fibres on the basis of the numerical simulation," Metabk. 52(2):251-254 (2013)

[3] K. M. Lee, and D. F. Gilmore, "Formulation and process modeling of biopolymer (polyhydroxyalkanoates: PHAs) production from industrial wastes by novel crossed experimental design," *Process Biochemistry*, vol. 40 pp. 229–246, 2005

[4] Z. ZJeirani, B. M. Jan, B. S. Ali, I. M. Noor , S. C. Hwa, and W. Saphanuchart, "The optimal mixture design of experiments: Alternative method in optimizing the aqueous phase composition of a microemulsion," *Chem. Intel. Lab. Systems*, vol. 112 pp. 1–7, 2012

[5] J. V Nardia, W. Accharc, and D. Hotzad, "Enhancing the properties of ceramic products through mixture design and response surface analysis," *J. European Ceramic Soc*. Vol. 24 pp. 375–379, 2004

[6] H. Scheffé, "Experiments with mixtures," *J Royal Statistical Society Series*, vol B 20, pp. 344–366, 1958

[7] S. L. Correia, D. Hotza, and A. M. Segadães, "Optimising mechanical strength and bulk density of dry ceramic bodies through mixture design," *Bol. Soc. Esp. Ceram*. Vol. 44 [1] pp. 53-58, 2005

[8] I. B. Deshmanya, and G. K. Purohit, "Development of mathematical model to predict micro-hardness of Al7075/Al2O3 composites produced by stir-casting*," J Eng. Sci Tech Review* vol. 5 (1), pp. 44-50, 2012

[9] M. Taskın, and U. Çalıgülü, "Modelling of microhardness values by means of artificial neural networks of Al/SiCp metal matrix composite material couples processed with diffusion method," *Mathematical and Computational Applications*, vol. 11(3), pp. 163-172, 2006

[10] A. B. Spierings, K. Wegener, nad G. Levy, "Designing material properties locally with additive manufacturing technology SLM," *Proc. of the Solid Freeform Fabrication Symposium, Austin*, TX, USA. Pp. 447-455, 2012

[11] C. Ramesh, and K. Kumar, "Mathematical and neural network models for prediction of wear of mild steel coated with inconel 718—A comparative study," *Int. Jour Sci Res. Pub*. Vol. 2.7 pp. 1-8, 2012

[12] M. J. Simon, E. S. Lagergren, and K. A. Snyder, "Concrete mixture optimization using statistical mixture design methods," *Proceedings of the PCI/FHWA international symposium on high performance concrete*, 1997

[13] R. H. Myers, and D. C. Montgomery, "Response surface methodology: process and product optimization using designed experiments," John Wiley & Sons, New York, 2002

[14] J. A. Cornell, "Experiments with Mixtures: Designs, Models, and the Analysis of Mixture Data," 2nd ed. Wiley, New York. (1990)

Modeling of SNMP Protocol in Event-B

Rajaa Filali, Sanae El Mimouni, Anas Amamou, Bahija Boulamaat, and Mohamed Bouhdadi

***Abstract*—**This paper presents an incremental formal development of the Simple Network Management Protocol (SNMP) in Event-B. SNMP is an application layer protocol used to manage network resources. This standardization gives network administrators the ability to monitor network performance. To model and verify the protocol, we use the formal technique Event-B which provides an accessible and rigorous development method and enables user to express the problem at abstract level and then add more details in refinement step to obtain concrete specification. This interaction between modelling and proving reduces the complexity and helps in assuring that the SNMP specification is proven in consistency and correctness

***Keywords*—**Simple Network Management Protocol, Formal Modelling, Refinement, Event-B, Rodin

I. Introduction

SIMPLE network Management Protocol is a communication protocol, it is used to administer and manage networked devices. It can be used to manage large networks that span firewalls or embedded devices. The specifications for this protocol can be found in Request For Comments (RFC) 1157 [1].

Increasingly numerous communication protocols are being employed in computer networks of various types. This increases the need of adequate software specification techniques and suitable development methods to make the system more reliable. A number of formal approaches have been applied to model and analyze these protocols, such as Petri Nets [2,3] and State Machine [4]. Recently a new method Event-B [5] has been developed by Jean Raymond ABRIAL who has developed the B method [6] and the Z method [7].

Rajaa Filali, *LMPHE laboratory, University of Mohammed V, Faculty of sciences, Rabat ,Morocco,(* e-mail: rajaa.filali@gmail.com).

Sanae El Mimouni, *LMPHE laboratory, University of Mohammed V, Faculty of sciences, Rabat ,Morocco,(* e-mail: sanae.elm@gmail.com).

Anas Amamou, *LMPHE laboratory, University of Mohammed V, Faculty of sciences , Rabat ,Morocco,(* e-mail: amamou.anas@yahoo.fr).

Bahija Boulamaat, *LMPHE laboratory, University of Mohammed V, Faculty of sciences, Rabat, Morocco (e-mail: boulamaatbahija@gmail.com)*

Mohamed Bouhdadi, *LMPHE laboratory, University of Mohammed V, Faculty of sciences, Rabat ,Morocco,(* e-mail: bouhdadi@fsr.ac.ma).

In this paper, we use Event-B to model and prove the SNMP protocol. The most important benefit of using Event-B is its capability to use abstraction and refinement [8].

Indeed, in this approach the modeling process starts with an abstraction of the system which specifies the goals of the system. The abstract level of our Event-B model shows these goals in a very general way, and then during refinement levels, features of the protocol are modeled and the goals are achieved in a detailed way. Moreover the Rodin tool [9] permits an automated proof of the different models of the system.

The reminder of the paper is organized as follows. Section 2, gives a brief overview of Event-B. Section 3 provides the requirements which are informally defined. In Section 4, the formal development is presented. Finally, a conclusion is presented to summarize the main outcomes of this research

II. Overview of Event-B

Event-B is a formal method for specifying, modeling and reasoning about systems, especially complex systems such as an electronic circuit, an airline seat booking system, a PC operating system, a network routing program, a nuclear plant control system, a Smartcard electronic purse, etc..Event-B has evolved from classical B.

Key features of Event-B are the use of set theory as a modeling notation, the use of refinement to represent systems at different abstraction levels and the use of mathematical proof to verify consistency between refinement levels. From a given model M1, a new model M2 can be built as a refinement of M1. In this case, model M1 is called an abstraction of M2, and model M2 is said to be a concrete version of M1. A concrete model is said to refine its abstraction. Each event of a concrete machine refines an abstract event or refines skip. An event that refines skip is referred to as a new event since it has no counterpart in the abstract model. An Event-B model has two parts, context and machine. Each context specifies the static properties of the system, including sets, axioms, and constants. Each machine specifies the dynamic part of the system, including variables, invariants and events. Variables represent the current state of the system and invariants specify the global specification of the variables and system behaviors.

An event is defined by the syntax: EVENT e WHEN G THEN S END , Where G is the guard, expressed as a first-order logical formula in the state variables, and S is any number of generalized substitutions, defined by the syntax S ::= x := E(v) |x := z : |P(z). The deterministic substitution, x := E (v), assigns to variable x the value of expression E(v), defined over set of state variables v. In a non-deterministic substitution, x := z : |P(z), it is possible to choose non-

deterministically local variables, z, that will render the predicate P(z) true. If this is the case, then the substitution, x := z, can be applied, otherwise nothing happens.

The Rodin is the tool of the Event-B. It allows formal Event-B models to be created with an editor. It generates proof obligations that can be discharged either automatically or interactively. Rodin is modular software and many extensions are available. These include alternative editors, document generators, team support, and extensions (called plugins) some of which include support decomposition and records.

III. Informal Description Of SNMP Protocol

The SNMP is a client/server (agent/manager) protocol.
The **agent** (Server) is a software process that responds to queries using the Simple Network Management Protocol to provide status and statistics about a network node.
The **manager** (Client) is an application that manages SNMP agents on a network by issuing requests, getting responses, and listening for and processing agent- issued traps
SNMP traps enable an agent to notify the management station of significant events by way of an unsolicited SNMP message.

As shown in (Fig. 1), the setup on the left shows a network management system that polls information and gets a response. The setup on the right shows an agent that sends an unsolicited or asynchronous trap to the network management system (NMS).

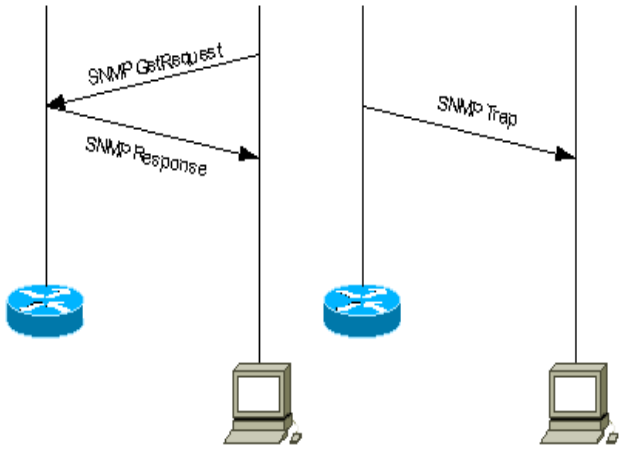


Fig. 1 The two setups of the network management system

Among the SNMP commands are specific protocol operations that facilitate in the requests and responses of managed network devices. The most basic operations include: Get, GetNext, Set, and Trap (see Fig. 2)
GetRequest: Manager requests an update
GetNextRequest: Manager requests the next entry in a table
SetRequest: Manager modifies data on the managed device.
GetResponse: Agent answers a manager request.
Trap: Agent alerts manager of an unusual event.

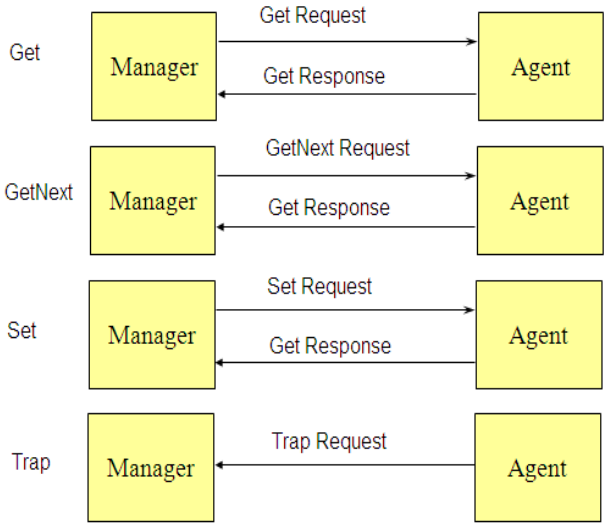


Fig. 2 The permitted operations between managers and agents

IV. Modeling Of SNMP Protocol

A. Initial Model

The first model is the most abstract specification of the system. First, we define three carrier sets:

Requests: set of messages which can be sent by the manager, it contains three constants (GetRequest, GetNextRequest and SetRequest) defined by the axioms (axm1, axm2 and axm3).

Responses: set of responses sent by the Agent, it contains the constant GetResponse which represented by the axiom (axm4).

Notification: set of messages sent by the Agent to inform the Manager. The axiom (axm5) represent that this set contains the constant Trap.

AXIOMS
axm1 : GetRequest ∈ Requests
axm2 : GetNextRequest ∈ Requests
axm3 : SetRequest ∈ Requests
axm4 : GetResponse ∈ Responses
axm5 : Trap ∈ Notification

Then we can use two variables to represent the state of the initial model: *reqt* to denote the number of requests that have been sent, and *resp* to indicate the number of responses that have been given.

We have three invariants: inv1 and inv2 denotes that the two variables *reqt* and *resp* are natural numbers. inv3 specifies that the communication is synchronous: either the number of requests is the same as the number of responses or it is greater than the number of responses by 1 in the case where a response is expected before another request can be created.

VARIABLES
reqt
resp

ISBN: 978-1-61804-256-9

INVARIANTS
inv1 : reqt ∈ ℕ
inv2 : resp ∈ ℕ
inv3 : reqt=resp ∨ reqt=resp+1

Initially, there are no requests or responses hence both variables are initialed by 0.

INITIALISATION
act1 : resp:=0
act2 : reqt:=0

Finally, we define two events in our abstract model. An event **Manager_request** represents the sending request from the manager to the agent, starts when the number of requests and the number of responses are identical and increases the number of requests by 1. An event **Agent_response** represents the response sent from the agent to the manager, guards of this event state that the number of requests and responses are different.

Manager_request
WHEN
grd1 : reqt=resp
THEN
act1 : reqt:=reqt+1
END

Agent_response
WHEN
grd1 : reqt≠resp
THEN
act1 : resp:=resp+1
END

B. First Refinement

In this first refinement, we introduce the channels and the messages sent between the manager and the agent, because in the reality the message needs to be sent via some channel between two parties.

So we add two variables **reqtChan** and **respChan** which represent respectively the channel of messages sent by the manager and the channel of messages sent by the agent

INVARIANTS
inv1 : reqtChan ⊆ Requests
inv2 : respChan ⊆ Responses

We define now our events:
Manager_send_request: refining the abstract event Manager_request: the manager sends a message to the agent.
Agent_receive_request: the agent receives the request sent by the manager.
Agent_send_response refining the abstract event Agent_response: after receiving the request, the agent sends a response to the manager.
Manager_receive_response: the manager receives the response sent by the agent.

Manager_send_request
REFINES
Manager_request
ANY msg WHERE
grd1 : reqt=resp
grd2 : msg ∈ Requests
grd3 : msg ∉ reqtChan
THEN
act1 : reqt:=reqt+1
act2 : reqtChan := reqtChan ∪ {msg}
END

Agent_receive_request
ANY msg WHERE
grd1 : msg ∈ reqtChan
THEN
act1 : reqtChan:= reqtChan \ {msg}
END

Agent_send_response
REFINES
Agent_response
ANY msg WHERE
grd1 : reqt≠resp
grd2 : msg ∈ Responses
grd3 : msg ∉ respChan
THEN
act1 : resp:=resp+1
act2 : respChan := respChan ∪ {msg}
END

Manager_receive_response
ANY msg WHERE
grd1 : msg ∈ respChan
THEN
act1 : respChan := respChan \ {msg}
END

C. Second Refinement

In this refinement, we add a new event "Notify" where the agent can send a trap, or asynchronous notification, to the manager.

Notify
ANY msg WHERE
grd1 : msg ∈ Notification
THEN
act1 : notiChan := notiChan ∪ {msg}
END

ISBN: 978-1-61804-256-9

V. Conclusion

In this paper, we have modeled and proved SNMP protocol using Event-B.

We have explained our approach using refinement, which allows us to achieve a very high degree of automatic proof. The powerful support is provided by the Rodin tool. Rodin proof is used to generate the proof obligations and to discharge those obligations automatically and interactively.

Modeling and analyzing SNMP specification using formal methods can help in assuring correctness, unambiguity, and clarity of the SNMP protocol. Since a well-defined and verified protocol specification can reduce the cost for its implementation and maintenance, modeling and analysis are important steps of the protocol development life-cycle from the point view of protocol engineering.

.

References

[1] Case, J., Fedor, M., Schoffstall, M., and Davin, J., "RFC 1157: Simple network management protocol (SNMP)," *IETF, April*, 1990

[2] Woodside, C.M., "Performance Petri net analysis of communications protocol software by delay-equivalent aggregation," *In Petri Nets and Performance Models,* pp. 64-73, 1991.

[3] Antonidakis, E. "Conferencing protocols and petri net analysis", *WSEAS Transactions on Computers*, vol. 5, no 12, pp. 3112-3118, 2006

[4] Bochmann, G. "Formal Methods in Communication Protocol Design," *IEEE Transactions on Communication*, vol.28, pp. 624-631, 1980.

[5] Abrial, J.R., *Modeling in Event-B: system and software engineering*, Cambridge University Press, 2010.

[6] Abrial, J.R., *The B-book: assigning programs to meaning,* Cambridge University Press. 2005.

[7] Abrial, J.R., "B#: Toward a synthesis between Z and B," In: ZB 2003*: Formal Specification and Development in Z and B,* Springer Berlin Heidelberg, pp. 168-177, 2003.

[8] Back, R.J., *On the correctness of refinement steps in program development*, Department of Computer Science, University of Helsinki, 1978.

[9] Jones, C., Oliver, I., Romanovsky, A., and Troubitsyna, E., *RODIN (rigorous open development environment for complex systems)*, University of Newcastle upon Tyne, Computing Science, 2005.

A Study Of Exergy Analysis for Combustion in Direct Fired Heater (Part I)

Seif Al Nasr Ahmed Abd Al ghany*, Bahgat Kameis Morsy** Ahmed Ali Abd El-Rahman Ali***
*Mechanical Engineering Dept. - Faculty of Engineering – Beni Suef University, Egypt
dr.sife2011@yahoo.com
** Mechanical Engineering Dept. - Faculty of Engineering – Minia University, Egypt
Bahgat52@yahoo.com
***Mechanical Engineering – Egyptalum Company, ahmed_3a2000@yahoo.com

Abstract: Heat transfer plants with organic media have often been able to replace or improve the classic steam–water operation. The possibility of transferring and closely controlling temperature up to > 300 ^{o}C has provided the heat transfer media technology with many new fields of application. This growing application of heat transfer plants with liquid heat transfer media other than water has made it necessary to produce complete and accurate engineering database for combustion and its devices to continuous improvement of industrial heating. Heating is an important operation in almost all industrial fields. A large variety of heating techniques is available at the market. Some examples are fuel burning, electrical heating, and so on. The analysis of related combustion process and estimation of the effective coefficients is the first step toward a successful design. The process of combustion fuel and their combustion and combustion devices are considered in this study. Direct fired heater exergy and energy analysis are performed taking into account precise calculation of chemical exergy for products of combustion.

Keywords: Exergy, Energy, Combustion, Thermal System.

Nomenclature

A	Area [m^2]	S	Distance [m]
A_{Abs}	The absolute availability of a system [-]	S_o	Entropy of a system at environmental state [J/kg]
C_p , c	Specific heat capacity [J/kgK] or heat capacity [J/kg]	S_i	Specific entropy, of substance I, [J/kgK]
E, E	Specific exergy [J/kg] or available work [J]	T	Temperature [K]
E_f	Fuel exergy [kJ/kg]	u, U	Specific internal energy [J/kg] or internal energy [J]
$e^{ch} = A_{ch}$	Chemical exergy [kJ/kg]	v, V	Specific volume [m^3/kg] or volume [m^3]
E/Q	Exergy factor [no unit, %]	σ	Stefan-Boltzmann constant [W/m^2K^4]
H, H	Specific enthalpy [J/kg] or enthalpy [J]	Є	The emissivity [-]
m_a	Mass flow rate of air [kg/s]	ώ	Exergetic efficiency [-]
m_f	Mass flow rate of fuel [kg/s]		
m_e	Mass flow rate of exhaust flue gas [kg/s]		
P	Power [kW]		
P_o	Environment pressure [bar]		
Q, Q	Specific heat [J/kg] or heat [J]		
Q, Q	Specific heat [J/kg] or heat [J]		
$\bar{R}$	Molar gas constant [J/mol K]		

I. Introduction:

Traditional methods of thermal system analysis are based on the first law of thermodynamics. These methods use an energy balance on the system to determine heat transfer between the system and environment. The first law of thermodynamics introduces the concept of energy conservation, which states that energy entering a thermal system with fuel, electricity, flowing streams of matter, and so on is conserved and can not be

I.1 Energy and Exergy:

Exergy is a measure of the quality or grade of energy and it can be destroyed in the thermal system. The second law states that a part of the exergy entering a thermal system with fuel, electricity, flowing streams of matter, and so on is destroyed. In general, energy balances provide no information on the quality or grades of energy crossing the concept of exergy in the analysis of thermal systems. destroyed within the system due to irreversibilities. Thermal system boundary and no information about internal losses. By contrast, the second law of thermodynamics introduces the useful. In recent years, exergy analysis has played a key role in order to evaluate processes by taking into account not only the quantity of energy but also both the quantity and quality of energy.Various definitions have been used to describe the term of exergy. Exergy is defined as the maximum amount of work which can be produced by a system or a flow of matter or energy. Exergy is a measure of the potential of the system or flow to cause change, as a consequence of not being completely stable equilibrium relative to the reference environment. Unlike energy, exergy is not a subject to a conservation law (except for ideal or reversible processes). Rather exergy is consumed or destroyed due to irreversibilities in any real process. The exergy consumption during a process is proportional to the entropy created due to irreversibilities associated with the process.

I.2 Theoretical Analysis:

Exergy analysis is a method using the conservation of mass and conversion of energy principles together with the second law of thermodynamic for the analysis, design and improvement of energy and other systems. An exergy balance applied to a process or a whole plant tells us how much of the usable work potential or exergy supplied as the input to the system under consideration, has been consumed by the process.

The loss of exergy or irreversibility provides a generally applicable quantitative measure of process inefficiency. In other words, an Exergy analysis is similar to an energy analysis, but it takes into account the quality of the energy as well as the quantity. Since it includes a consideration of entropy, Exergy analysis allows a system to be analyzed more comprehensively by determining where in the system the Exergy is destroyed by internal irreversibilities, and the causes of those irreversibilities.

2. Combustion and Exergy:

The purpose of combustion in industrial applications, for the most part, is to transform chemical energy available in various types of fuels to thermal energy or heat to be used in the processing of gas or liquid streams or solid objects. Typical examples involve the heating of air, water, and steam for use in heating of other processes or equipment, the heating of metals and nonmetallic minerals during production and processing, the heating of organic streams for use in refining and processing, as well as heating of air for space comfort conditioning. For all of these, it is necessary to have a workable method for evaluating the heat that is available from a combustion process. Available heat is the heat accessible for the load (useful output) and to balance all losses other than stack losses.

Exergy is a measure of the energy available for useful work in a system. This property is also referred to as Availability. Exergy is a better measure of the work that may be extracted from a system rather than properties such as the internal energy or enthalpy of the system. No device or process can extract a quantity of work greater than the availability of the system without violating the second law of thermodynamics. Thus, the availability of a system also helps to define the upper limit on the efficiency of the device/process. cause and true magnitude of energy resource waste and to determine losses. Such information can be used in the design of new energy-efficient systems and for improving the performance of existing systems

ISBN: 978-1-61804-256-9

The method of Exergy analysis presented in this .

3. Exergy Balance:

Exergy can be transferred by three means:
1- Exergy transfer associated with work.
2- Exergy transfer associated with heat transfer.
3- Exergy transfer associated with the matter entering and exiting a control volume.
Exergy is also destroyed by irreversibility within the system or control volume. Fig.1. shows exergy flow diagram. In this Figure E_{tr} represent exergy transit and E_{pr} represent exergy used by process.

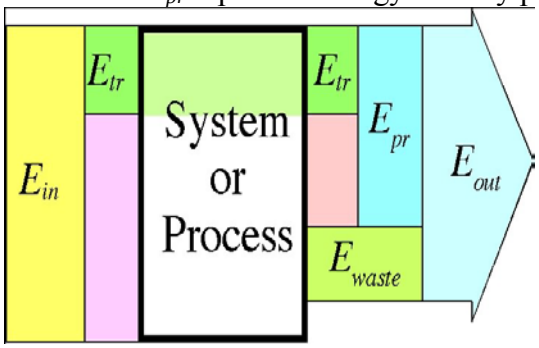


Fig.1. Exergy flow diagram

4. The Objective of Investigation:

Direct fired heaters are the most common industrial heating facilitates. They are readily employed for a wide range of applications and can be powered on various fuels depending on the local availability. Their many advantages and relative lack of competition has made direct fired heaters one of the preferred devices for converting the chemical energy of fuels into thermal energy. The current work attempts to understand the destruction of exergy (availability) in combustion processes and compare it with the losses of energy, with specific application to direct fired heaters. However, the analysis is not restricted to direct fired heaters and is applicable to all combustion processes.
From the literature review, it is evident that a comprehensive second law examination of combustion processes is lacking. Such a study would provide a more fundamental understanding of combustion processes and help in identifying strategies to reduce the destruction of exergy during combustion processes. Some work has been done towards applying the second law to combustion by Dunbar and Lior [1] (constant pressure combustion) and Daw et al. [2] (constant pressure combustion) and Caton [3] (constant investigation enables us to identify the location, volume). These studies however, were restricted to a particular combustion process and did not strictly quantify the contribution of the various exergy terms. The current study wishes to apply the second law to the combustion process, while relaxing most of the approximations and simplifications made in the past. It is hoped that an inclusive examination of the various parameters will provide a more fundamental and complete understanding of the combustion processes. The current work also aims to incorporate excess air ratio into the study to allow for comparison of the combustion of different equivalent ratios. For more accurate analyses chemical exergy will be calculated in this study. The combustion process analyzed will be in the (direct fired heater), used in purpose of heating of heat transfer medium (Mineral Oil MOBILTHERM 605), in plant of heating of thermal oil in EGYPTALUM company in Nag-Hammady, Egypt.

5. Thermal Oil Plant Operation:

Thermal oil system is shows in Fig. 2. The Figure provides an efficient means of supplying indirect heat to one or more process systems. Such systems offer both high temperature and low pressure, making them ideal for a wide variety of process heating application. The heat transfer fluid firstly heated by means of a direct fired heater then circulated through a closed loop systems to the users. Heat from the fluid is transferred to the user and then re-circulated for reheating and the cycle repeated. However, organic media has become more common and often replaces a classic steam-water operation. The heaters are made with coils made of seamless tubes. The thermal fluid is heated during the flow through the tubes. The heat is transferred to the fluid as radiant heat in the combustion chamber, where the inner cylindrical tube coil and a flat tube coil form the chamber wall and the bottom respectively

ISBN: 978-1-61804-256-9

. Consequently refractory concrete is avoided. The combustion gasses are hereafter cooled in the outer convection part, as the gasses pass the space between the two tube coils. The thermal design ensures a modest volume of the thermal fluid relative to the size of the heater, and allows unlimited thermal expansion due to the high fluid temperature.

The fuel that will be used in this test will be fuel oil No.6 that is named in Egypt and Arab countries as (Mazout); this fuel must be heated before using because it's high viscosity at ambient temperature. Table (1) shows fuel chemical analysis for typical heavy oil No.6 by weight.

Table (1): Fuel chemical analysis for typical heavy oil no.6 by weight

Fuel Components	C	H_2	N_2	O_2	S
% By Weight	87.87	10.33	0.14	0.50	1.16

Almost all industrial liquid fuel burners use atomization to aid vaporization by exposing the large surface area (relative to volume) of millions of droplets in the size range of 100-400 μm. Evaporation then occurs at a rapid rate even if the droplets are not exposed to furnace radiation or hot air due to enhanced mass transfer rates.

Rotary-cup atomization delivers the liquid fuel to the center of a fast-spinning cup surrounded by an air stream. Rotational speed and air pressure determine the spray angle. This is still used in some large boilers, but the moving parts near the furnace heat have proved to be too much of a maintenance problem in higher temperature process furnaces and on smaller installations where a strict preventive maintenance program could not be affected.

6. Flue Gas Analysis:

The major constituents in flue gas are CO_2, O_2, N_2 and H_2O. Excess air is determined by measuring the O_2 in the flue gas. Before proceeding with measuring techniques, consider the form of the sample. A flue gas sample may be obtained on a wet or dry basis. When a sample is extracted from the gas stream, the water vapor normally condenses and the sample

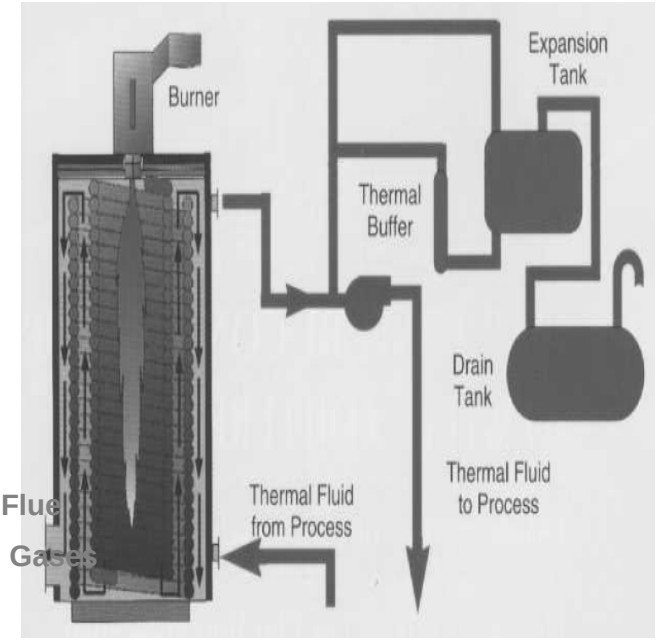


Fig. 2. Simple schematic diagram for thermal oil plant

is considered to be on a dry basis. The sample is usually drawn through water near ambient temperature to ensure that it is dry. The major constituents of a dry sample do not include the water vapor in the flue gas.

When the gas is measured with an in situ analyzer or when precautions are taken to keep the moisture in the sample from condensing, the sample is on a wet basis. The amount of O_2 in the flue gas is significant in defining the status of the combustion process. Its presence always means that more oxygen (excess air) is being introduced than is being used. Assuming complete combustion, low values of O_2 reflect moderate excess air and normal heat losses to the stack, while higher values of O_2 mean needlessly higher stack losses. The quantity of excess O_2 is very significant since it is a nearly exact indication of excess air.

The O_2 is an equally constant indication of excess air when the gas is sampled on a wet or in situ basis because the calculated excess air result is insensitive to variations in moisture for specific types/sources of fuel. The current industry

ISBN: 978-1-61804-256-9

standard for heaters operation is continuous monitoring of O_2 in the flue gas with in site analyzers that measure oxygen on a wet basis. For testing, the preferred instrument is an electronic oxygen analyzer.

The flue gas analyzer unit, which measures (CO_2 SO_2) and O_2 on a dry volumetric basis, remains a trusted standard for verifying the performance of electronic equipment. The flue gas analyzer uses chemicals to absorb the (CO_2 SO_2) and O_2, and the amount of each are determined by the reduction in volume from the original flue gas sample

7. Temperature Measurements:

Thermal oil inlet, thermal oil outlet, ambient temperature, fuel oil temperature, flue gas temperature and outer surface temperature, all are measured in this study for more accurate calculation. All measurements of temperature included in this investigation will be executed by PT-100 thermocouple (Type K) Fig. 3. Temperature of thermal oil inlet, thermal oil outlet, fuel oil inlet, and flue gas were measured by PT-100 thermocouple. Temperature of outer surface of heater was measured by using device which uses Infrared sensor technology to measure temperature of surfaces without contact.

8. Method of Calculation:

The most commonly used indicator for the efficiency of energy conversion process is the ratio of the output of useful energy to the total energy input. This ratio is called first law efficiency. It is based on a quantitative accounting of energy, which reflects recognition of the first law of thermodynamics and the law of conservation of energy.

It is well known that the second law of thermodynamics defines the availability of energy more restrictively than the first law. Principally, first law is silent on the effectiveness with which availability is concerned. Analysis in terms of the second law of thermodynamics more closely describes the effectiveness with which systems or processes use available energy.

Each calculation of exergy and thus each exergetic analysis imply reference state called 'dead state'. If a system is in thermal and mechanical equilibrium with the reference environment that is at the environmental temperature T_O and Pressure P_O, it is said to be in a thermodynamically dead state or restricted dead state. In general it is taken as T_O = 298 K and P_O = 1 atm.

Exergy losses are calculated by making exergy balance for each component of the system. Unlike energy balance where the inflow is equal to outflow (when there is no internal energy generation or consumption), in exergy balance due to reasons of irreversibility, exergy inflow is always greater than the exergy outflow and their difference gives the exergy loss or exergy destruction. Ratio of exergy output to exergy input gives the exergetic efficiency of a system [4].

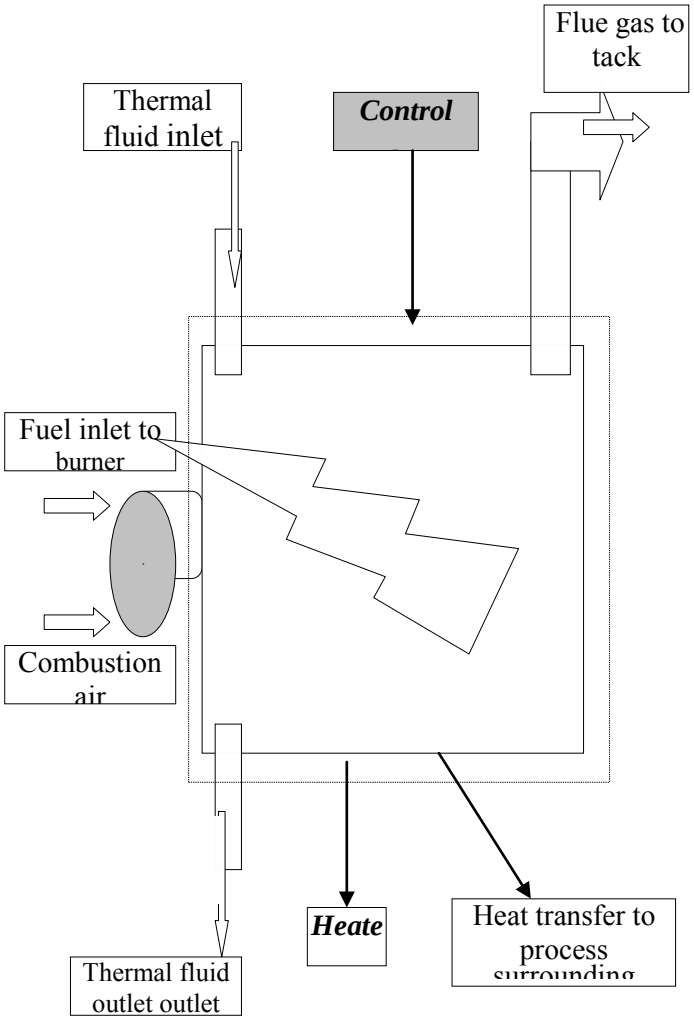


Fig. 4. Schematic diagram for the control volume of a test rig.

$$\omega = \text{Exergetic efficiency} = \frac{\text{Exergy output}}{\text{Exergy input}} \quad (1)$$

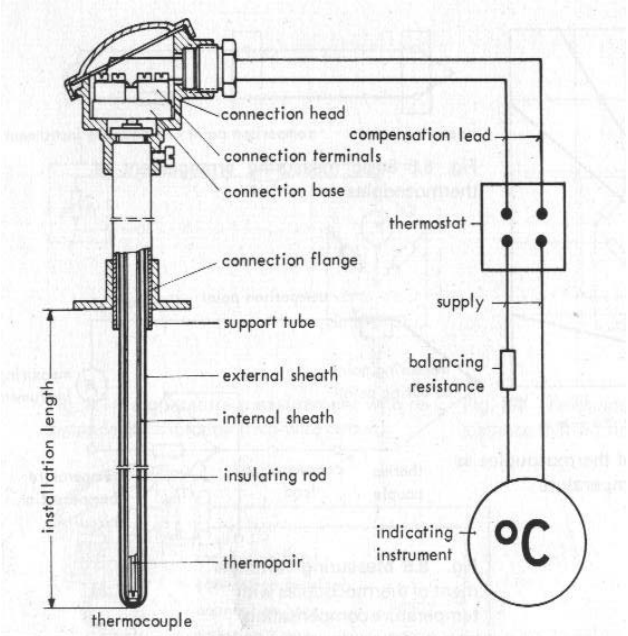


Fig. 8.7 Protective sheath for thermocouple

PT- 100 Thermocouple (type K)
Fig. 3

Exergy calculation (second law of thermodynamics):

The objective of this section is to introduce exergy analysis, a method that uses the conservation of mass and conservation of energy principles together with the second law of thermodynamics for the investigate and analysis of combustion process and thermal systems. Another term frequently used to identify exergy analysis is availability analysis.

Following usual conventions [5-7], the absolute availability, $\mathbf{A_{Abs}}$ of a system is defined as:

$$A_{Abs} = U - T_0 S + P_0 V \qquad (2)$$

Where U, S and V are the internal energy, entropy and volume of the system respectively, while T_0 and P_0 are the reference temperature and pressure. The work that may be extracted from a system is also limited by the reference conditions. The work that may be extracted from the system is then given by the (thermo-mechanical) availability, A_{TM}, of the system, which is defined as:

$$A_{TM} = (U - U^o) - T_0 (S - S^o) + P_0 (V - V^o) \quad (3)$$

Where U^o, S^o and V^o are the internal energy, entropy and volume of the restricted dead state respectively.

The restricted dead state is achieved by allowing the system to come to thermo-mechanical equilibrium with the environment, typically the atmosphere. The restricted dead state has the same pressure and temperature as the environment, however, the composition of the restricted dead state is the same as that of the original system and is not necessarily the same as that of the environment. The current work uses this definition of the restricted dead state, in conformance with the standard literature [5-7], with a temperature of 298.15 K and pressure 101.325 kPa for the restricted dead state and the reference conditions. This difference in composition between the restricted dead state and the environment can be exploited to further obtain work from the system. This work, obtained by allowing the restricted dead state to come to chemical equilibrium with the environment, is referred to as the chemical exergy, A_{Ch}, of the system.

$$A_{Ch} = \sum_{k=1}^{n} N_k \left(\mu_k^0 - \mu_{k,0} \right) \qquad (4)$$

Where N_k is the number of moles of the respective species (k) and $\mu_{k,o}$ & μ_k^O are the chemical potentials of the respective species in the restricted dead state and the environment, respectively. The chemical potentials may further be expressed as:

$$\mu_k = g_k(T_0, P_0) + \bar{R} T_0 \ln\left(\frac{p_k}{p_0} \right) \qquad (5)$$

Where g_k is the Gibbs energy of the k^{th} species in the mixture $\bar{R}$ is the Universal gas constant and p_k is the partial pressure of the k^{th} species in the mixture. If the restricted dead state and the environment, both had the same constituent species, differing only in their respective compositions, the Gibbs energy term would cancel out, leaving a simpler expression for the chemical availability of the system:

$$A_{Diff} = \bar{R} T_0 \sum_{k=1}^{n} N_k \ln\left(\frac{p_k^0}{p_{k,0}} \right) \qquad (1.5)$$

The difference in concentrations of the various species in the system and the atmosphere may be exploited by first separating the various components in the mixture (using devices such as

semi-permeable membranes) and then allowing them to expand or compress to the atmospheric partial pressures, as the case may be. Work may be gained or lost during this process and this creates an additional potential for work. Since this term may be attributed to the work obtained by allowing the species in the system to diffuse to the atmospheric concentrations, it would be appropriate to refer to this as the "diffusion availability". It may be noted that diffusion availability of a system can be positive or negative, depending on the concentrations of the various species in the system.

The diffusion availability of a system is largely ignored since its contribution is often small relative to the thermo-mechanical availability A_{TM} of the system. Also, it is not easy to extract the diffusion availability component of the availability since it would require the use of semi-permeable membranes to extract the various species in the mixture before allowing them to diffuse to atmospheric concentrations. It is also evident from the expression for the diffusion availability of a system that it depends on the composition of the environment. The assumed composition of the atmosphere therefore, makes a difference on the diffusion availability of the system. The current work uses a standard wet atmospheric unless otherwise stated.

The availability of a system, A_{Total}, incorporating the various components would then be

$$A_{Total} = (U - U_0) - T_0(S - S_0) + P_0(V - V_0) + \sum_{k=1}^{n} N_k\left(\mu_k^0 - \mu_{k,0}\right) \qquad (1.6)$$

The above expression for availability is valid for closed systems. For open systems, the flow availability, $A_{Total,\,f}$ needs to be considered. This is defined

$$A_{Total,f} = (H - H^0) - T_0(S - S^0) + \sum_{k=1}^{n} N_k\left(\mu_k^0 - \mu_{k,0}\right) \qquad (1.7)$$

Where H and H^0 are enthalpies of the system and the restricted dead state respectively.

In general, then, the availability of a system, A_{Total}, may be expressed as a sum of the thermo-mechanical availability and chemical availability.

$$A_{Total} = A_{TM} + A_{Ch} \qquad (8)$$

The chemical availability term may further be split into constituents, the reactive availability and diffusive availability as:

$$A_{Total} = A_{TM} + A_{Reactive} + A_{Diff} \qquad (9)$$

The importance of developing thermal systems which uses fossil fuel in the process of combustion that make effective use of nonrenewable resources such as oil, natural gas, and coal is apparent. The method of Exergy analysis is particularly suited for furthering the goal of more efficient resource use, since it enables the locations, types, and true magnitudes of waste and loss to be determined. This information can be used to design thermal systems, guide efforts to reduce sources of inefficiency in existing systems, and evaluate system economics.

8.2 Exergy Analysis Formulas

The start point in the Exergy analysis is Exergy balance for a system, Exergy balance in this system can be symbolized as:

$$E_{in} = E_{oil} + E_{stack} + E_{s,loss} + E_{Destruction} \qquad (11)$$

Where, E_{in} represent chemical exergy involved in fuel oil entering to combustion chamber, also exergy of fuel oil = LHV * 1.04, [5] that means E_{in} = 41033*1.04 = 42674.32 kJ/kg fuel.

E_{oil} in above equation represent Exergy flow to thermal fluid, and is calculated here from:

$$E_{oil} = \text{Exergy flow to thermal fluid} = \dot{m} c_p \left(T_{out} - T_{in} - T_o \ln \frac{T_{out}}{T_{(8)}}\right) \qquad (12)$$

Where, the value 1.04 is factor multiply in lower heating value of fuel to get exergy value contained in fuel. [5].

$\dot{m}$= Thermal fluid mass flow rate kg/hr, Cp = average specific heat capacity kJ/ (kg.k)

T_{out} = Thermal fluid outlet temperature Kelvin, T_{in}=Thermal fluid inlet temperature Kelvin.

E_{stack} in above equation represents exergy flow to surrounding with flue gas and calculated here from equation: (9)

$$E_{stack} = \underline{h - h_0 - T_0 (s - s_0)} + e^{ch}$$

Where, in above eq., h and s represent the specific enthalpy and entropy, respectively, at the inlet or

exit under consideration; h_0 and s_0 represent the respective values of these properties when evaluated at the dead state. Values of h, s, h_0 and s_0 are from standard tables of thermodynamics. Where the underlined term is the thermo-mechanical contribution of exergy in combustion products, $\boldsymbol{e^{ch}}$ is the chemical contribution evaluated as following::

$$\bar{e}^{ch} = \bar{R}T_0 \sum_i y_i \ln\left(\frac{y_i}{y_i^e}\right) \quad (14)$$

Where, $\bar{R}$= Universal Gas Constant=8.314 kJ/kmol. K and y_i and y_i^e denote, respectively, the mole fraction of component i in the mixture of combustion products at T_0, P_0 and in the environment, with assumption that products of combustion are modeled as an ideal gas mixture at all states considered.

$E_{s.loss}$ in main equation represents exergy flow to surrounding by radiation from the surface of heater and calculated here from equation:

$$\boldsymbol{E_{s.loss} = (1 - [T_o / T_{surf}])* Q_e} \quad (15)$$

Where Q_e calculate (surface losses) which is quantified macroscopically by a modified form of the Stefan–Boltzmann law equation relation:

$$\mathbf{Q_e = \varepsilon\sigma A (T_s^4 - T_o^4)} \quad (16)$$

$\boldsymbol{E_{Destruction}}$ in main equation represents Exergy Destruction inside furnace because irreversibility and are calculated by making Exergy balance for control volume in this study.

9. Results Presentation and Analysis

In this study, exergy analysis was carried out for combustion in direct fired heater. A flue gas sample for 40 runs of restricted heater was taking and input and output streams for each run were studied. Exergy and energy balance for each run was evaluated and theoretical analysis was carried out using these results. These results include a complete energy and Exergy analysis for direct fired heater, therefore energy efficiency, exergetic efficiency; exergy losses, energy losses, irreversibility and exergy destruction within the system (control volume of test rig) are calculated.

Excess air ratios at variation of fuel oil flow rate, the following operating condition are tested:

1-Fuel oil flow rate=120 kg/hr, 2-Fuel oil flow rate=144 kg/hr, 3-Fuel oil flow rate=192 kg/hr, 4-Fuel oil flow rate=240 kg/hr, 5-Fuel oil flow rate=279 kg/hr, 6-Fuel oil flow rate=298 kg/hr 7-Fuel oil flow rate=318 kg/hr, 8-Fuel oil flow rate=336 kg/hr

Figure 5 represents the variation of energy efficiency with excess air at different levels of fuel flow rate. The figure shows that the energy efficiency, for all curves, tends to decrease with the increase of excess air level. A closer look in the figure would show that the high values of energy efficiency are achieved in the range of 8% to 20% of excess air. For all levels of fuel flow rate, the energy efficiency values are limited in the range of 60% to 82%.

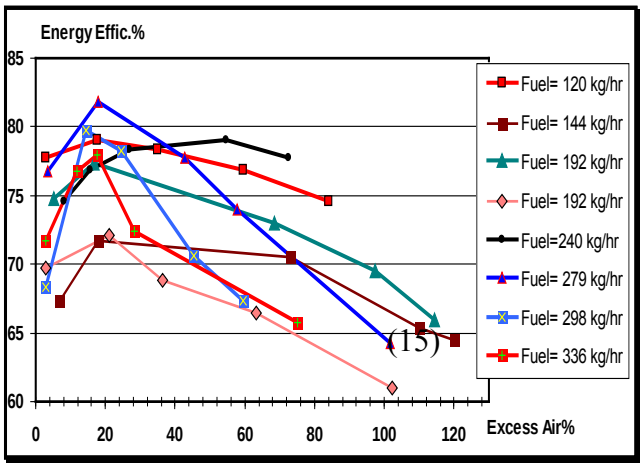
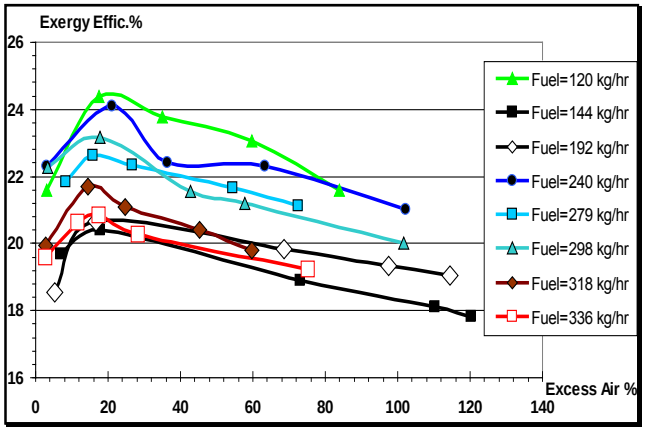


Fig. 5. Variation of the energy efficiency with excess air.

Figure 6 shows the variation of exergy efficiency with excess air at different levels of fuel flow rate. It is clear that the trend is the same as it appears in Fig. (5) but with change in values of exergy efficiency and energy efficiency. The Figure shows that the energy efficiency reached a range of 18% to 24% instead of 60% to 82% at the same excess air of the range of 8% to 20%.



ISBN: 978-1-61804-256-9

Fig. 6. Variation of the exergy efficiency with excess air.

Figure 7 shows the variation of energy stack losses with the excess air at different level of fuel flow rate. In the Figure, it is clear that general trend for the lines to go to high level of energy which goes to surrounding with flue gases with increase of excess air level, but in rang of 8% to 20% excess air stack losses decrease and return to increase again with excess air. Energy stack loss values are limited by the range 10% to 33% for all runs.

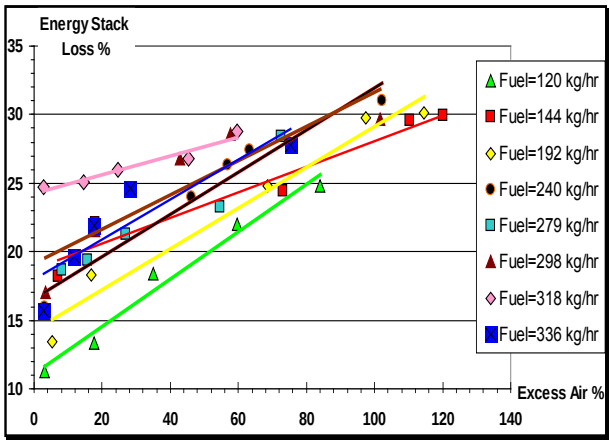


Fig. 7. Variation of the energy stacks losses with excess air.

Figure 8 shows the variation of exergy stack losses with the excess air at different level of fuel flow rate, it is obvious that the trends of the curves are the same as they appear in Fig. 7; but with changes in value of exergy stack losses than energy stack losses, since energy stack losses vary in range of 10% to 33%, while exergy stack losses vary in range 32% to 73% for all runs.

Figure 9 shows the variation of exergy destruction with the excess air at different level of fuel flow rate. In this Figure, it appears that the exergy destruction within the system under investigation decreases while the excess air level increases for all runs that mean the irreversibility within the system decreases with the increase of excess air levels.

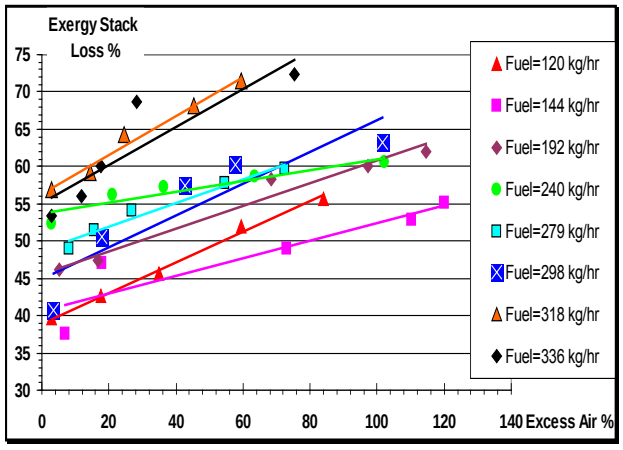


Fig. 8 Variation of the exergy stack losses with the excess air.

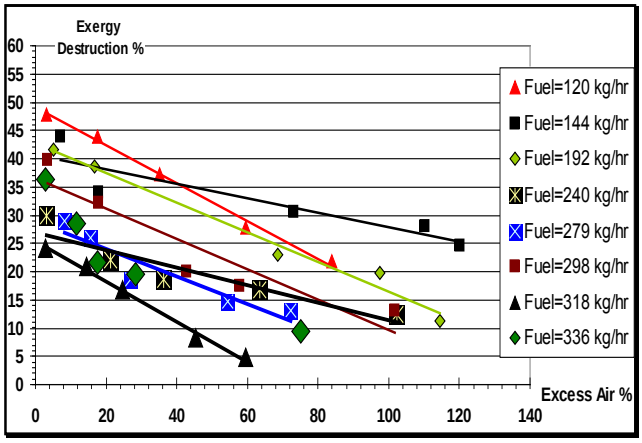


Fig. 9. The variation of exergy destruction with variation of excess air.

10. Conclusions

1-The exergitec efficiency is very low comparing the energy efficiency in direct fired heater. The Exergy destruction is in the range of 12% to 60% in direct fired heater.

2- The exergy loss through the flue gas is in the range of 20% to 65% in direct fired heater, and it is in the range of 0.004% to 0.008% for surface emission.

3- The average exergetic efficiency becomes 22% in direct fired heater. According to this analysis the minimum possible exergy losses in a direct fired heater should be within the following limits: Percentage exergy loss through flue gas: 20%, percentage exergy loss through surface emission: 0.005%. With reference to these limits the percentage exergy destruction is 60% and the exergetic efficiency is 22%. This is the maximum

ISBN: 978-1-61804-256-9

possible exergetic efficiency that can be taken by maintaining the optimum running condition.

4 - As increasing of exergetic efficiency, as the energy efficiency also increased

1. References

[1] Dunbar, W.R., and Lior, N., "Sources of Combustion Irreversibility", Combustion Science and Technology; 103, (1994), 41-61.

[2] Daw, S., Chakravarthy, K., Conklin, J., and Refining, G. R., "Understanding of Combustion Irreversibility", Proceedings of the Technical Meeting of the Central States, Section of the Combustion Institute, Austin, Texas, USA, March 21 – 23, (2004).

[3] Caton, J.A., "On the Destruction of Availability (Exergy) Due to Combustion Processes – with Specific Application to Internal Combustion Engines", Energy; 25, (2002). 1097-1117.

[4] Kotas, T.J., "The Exergy Method of Thermal Plant Analysis. Essex: Butterworth's", John Wiley and Sons Inc., New York, 1985.

[5] Moran, M.J., and Shapiro, H.N. Fundamentals of Engineering Thermodynamics, Fourth Edition, John Wiley and Sons Inc., New York, (2004).

[6] Moran, M.J., Availability Analysis: A Guide to Efficient Energy Use, Prentice Hall Inc., Englewood Cliffs, NJ, (1982).

[7] Cengel, Y.A., and Boles, M.A., Thermodynamics: An Engineering Approach, Fourth Edition, McGraw Hill Publications: New York, 2002.

[8] Kutz, M., Mechanical Engineers' Handbook: Energy and Power, Third EditionJohn Wiley, Vol. 4, & Sons Inc., (2006).

[9] Caton, J.A., A Review of Investigations Using the Second Law of Thermodynamics To Study Internal Combustion Engines, SAE Technical Paper Series, Society of Automotive Engineers, (2000).

[10] Gerpen, Van J.H., and Shapiro, H.N., "Second Law Analysis of Diesel Engine Combustion", Journal of Engineering for Gas Turbines and Power; .112, (1990), 129-137.

[11] Richter H.J., and Knoche, K.F., "Reversibility of Combustion Processes", Efficiency and Costing: Second Law Analysis of Processes, ACS Symposium series, 235, and (1983), 71-85...

ISBN: 978-1-61804-256-9

On the component-based reliability in open multi-server queueing networks

Edvinas Greičius, Saulius Minkevičius

Abstract – This paper is motivated by performance in terms of reliability of multi-server computer networks. Limit theorems on the queue length and virtual waiting time in an open multi-server queueing network in heavy traffic are derived and applied to a reliability model for a multi-server computer network, where the time of failure of a multi-server computer network is related to the parameters of the system.

Keywords – Heavy traffic, performance evaluation, queueing theory, probability limit theorem.

I Introduction

Probabilistic models and queueing networks have long been used to study the performance and reliability of computer systems [1, 2] and to analyse the performance and reliability of computer networks and of distributed information systems [3, 4]. In this paper, we will first briefly review the works related to using the queueing theory of computer systems reliability, and then present some new results on the estimation of the time of failure of a computer network.

In one of the first papers of this kind [6], the reliability of execution of programs in a distributed computing system is considered, showing that a program, which runs on multiple processing elements that have to communicate with other processing elements for remote data files, can be successfully executed despite that certain system components may be unreliable. In order to analyse the performance of multimedia service systems which have unreliable resources and to estimate their capacity requirements, a capacity planning model using an open queueing network is presented in [9], and in [5] a novel model for a reliable system composed of N unreliable systems, which can hinder or enhance each other's reliability, is discussed. In [10], the management policy of an $M/G/1$ queue with a single removable and non-reliable server is discussed and analytic results are explored, using an efficient Matlab program to calculate the optimal threshold of the management policy and to evaluate the system performance. In [11], the authors consider a single machine subject to break down and employ a fluid queue model with repair. In [13], the behaviour of a heterogeneous finite-source system with a single server is considered and applications in the field of telecommunications and reliability theory are treated.

In this paper, first we present the proability limit theorem on the queue length and virtual waiting time of the customer in heavy traffic for open multi-server queueing networks.

II The network model

Consider a network of j stations, indexed by $j = 1, 2, \ldots, J$, and the station j has c_j servers, indexed by $(j, 1), \ldots, (j, c_j)$. A description of the primitive data and construction of processes of interest are the focus of this section. No probability space will be mentioned in this section, and certainly, one can always think that all the variables and processes are defined on the same probability space.

First, $\{u_j(e), e \geq 1\}, j = 1, 2, \ldots, J$, are J sequences of exogenous interarrival times, where $u_j(e) \geq 0$ is the interarrival time between the $e-1$ job and the e-th job which arrive at the station j exogenously (from the outside of the network). Define $U_j(0) = 0, \quad U_j(n) = \sum_{e=1}^{n} u_j(e), \ n \geq 1$ and $A_j(t) = \sup\{n \geq 0 : U_j(n) \leq t\}$, where $A_j = \{A_j(t), t \geq 0\}$ is called the exogenous arrival process of the station j, i.e., $A_j(t)$ counts the number of jobs that arrived at the station j from the outside of the network.

Second, $\{v_{jk_j}(e), e \geq 1\}, \ j = 1, 2, \ldots, J, \ k_j = 1, 2, \ldots, c_j$, are $c_1 + \ldots + c_J$ sequences of service times, where $v_{jk_j}(e) \geq 0$ is the service time for the e-th customer served by the server k_j of the station j. Assume that $V_{jk_j}(0) = 0, \quad V_{jk_j}(n) = \sum_{e=1}^{n} v_{jk_j}(e), \ n \geq 1$ and $x_{jk_j}(t) = \sup\{n \geq 0 : V_{jk_j}(n) \leq t\}$, where $x_{jk_j} = \{x_{jk_j}(t), t \geq 0\}$ is called the service process for the server k_j at the station j, i.e., $x_{jk_j}(t)$ counts the number of services completed by server k_j at the station j during the server's busy time. We define $\mu_{jk_j} = \left(M\left[v_{jk_j}(e)\right]\right)^{-1} > 0, \ \sigma_{jk_j} = D\left(v_{jk_j}(e)\right) > 0$ and $\lambda_j = \left(M\left[u_j(e)\right]\right)^{-1} > 0, \ a_j = D\left(u_j(e)\right) > 0, \ j = 1, 2, ..., k$, with all of these terms assumed finite. Let p_{ij} be probability of the job after service at the ith station of the network are arrived to the jth station of the network, $i, j = 1, 2, \ldots J$.

Now we introduce the following process $Q_{jk_j} = \{Q_{jk_j}(t), \ t \geq 0\}$, where $Q_{jk_j}(t)$ indicates the number of customers waiting to be served by server k_j of the station j at time t; $j =$

This research was supported in part by the National Complex Programme "Theoretical and engineering aspects of e-service technology creation and application in high-performing calculation platforms".

E. Greičius is with Vilnius University, Naugarduko 24, 03225 Vilnius, Lithuania (e-mail: edvinas.greicius@gmail.com).

S. Minkevičius is with VU Institute of Mathematics and Informatics, Akademijos 4, 08663 Vilnius, Lithuania and Vilnius University, Naugarduko 24, 03225 Vilnius, Lithuania (e-mail: minkevicius.saulius@gmail.com).

ISBN: 978-1-61804-256-9

$1, 2, \ldots, J$ $k_j = 1, 2, \ldots, c_j$. Thus, we introduce the following process $V_{jk_j} = \{V_{jk_j}(t),\ t \geq 0\}$, where $V_{jk_j}(t)$ indicates the virtual waiting time of customer (workload process) in k_j server of the station j at time t; $j = 1, 2, \ldots, J,\ k_j = 1, \ldots, c_j$.

The dynamics of the queueing system (to be specified) depends on the service discipline at each service station. To be more precise, "first come, first served" (FCFS) service discipline is assumed for all J stations. When a customer arrives at a station and finds more than one server available, it will join one of the servers with the smallest index. We assume that the service station is work-conserving; namely, not all servers at a station can be idle when there are customers waiting for service at that station. In particular, we assume that a station must serve at its full capacity when the number of jobs waiting is equal to or exceeds the number of servers at that station.

III The main results

Let the number of servers k_i in j-th station of the network divide into parts: $k_j = 1, 2, \ldots, p_j$ (where the probability limit theorem is valid for queue length of customers) and $k_j = 1, 2, \ldots, r_j$ (where the probability limit theorem is valid for the virtual waiting time of the customer), $p_j + r_j = c_j$.

Let us denote $\hat{p}_{ij} = \frac{1}{c_i} \cdot \frac{1}{c_j} \cdot p_{ij}$,

$$p_j = 1 - \sum_{j=1}^{J} \sum_{k_i=1}^{c_i} \hat{p}_{ij},\ \tilde{\beta}_{jk_j} = \frac{\lambda_j}{c_j \cdot \mu_{jk_j} \cdot p_j} - 1 > 0,$$

$$\tilde{\sigma}^2_{jk_j} = \frac{\lambda_j^3}{\mu_{jk_j}} \cdot \frac{a_j}{\sigma_{jk_j}} \cdot \frac{1}{c_j^2 \cdot p_j^2} + 1 > 0,\ j = 1, 2, \ldots, J,$$

$$k_j = 1, 2, \ldots, c_j,\ t \geq 0.$$

We also define

$$\hat{\beta}_{jk_j} = \sum_{k_i=1}^{J} \mu_{ik_i} \cdot p_{ij} + \lambda_j - \mu_{jk_j} > 0,$$

$$\hat{\sigma}^2_{jk_j} = \sum_{k_i=1}^{J} \mu^3_{ik_i} \cdot \sigma_{ik_i} \cdot p^2_{ij} + \lambda_j^3 \cdot a_j + \mu^3_{jk_j} \cdot \sigma_{jk_j} > 0,$$

$$j = 1, 2, \ldots, J,\ k_j = 1, 2, \ldots, c_j.$$

We also assume that the following "overload conditions" are fulfilled

$$\sum_{i=1}^{J} \sum_{k_i=1}^{c_i} \mu_{ik_i} \cdot p_{ij} + \lambda_j > \sum_{k_i=1}^{c_j} \mu_{ik_i}, \quad (1)$$

$$j = 1, 2, \ldots, J.$$

Note that these conditions guarantee that the length of all the queues will grow indefinitely with probability one. The results of the present paper are based on the following theorems.

Theorem 1. *If conditions (1) are fulfilled, then*

$$\lim_{n\to\infty} P\left(\frac{Q_{jk_j}(nt) - \hat{\beta}_{jk_j} \cdot n \cdot t}{\hat{\sigma}_{jk_j} \cdot \sqrt{n}} < x\right) = \int_{-\infty}^{x} \exp\left(-\frac{y^2}{2t}\right) dy,$$

$$0 \leq t \leq 1,\ k_j = 1, 2, \ldots, p_j,\ j = 1, 2, \ldots, J$$

and

Theorem 2. *If conditions (1) are fulfilled, then*

$$\lim_{n\to\infty} P\left(\frac{V_{jk_j}(nt) - \tilde{\beta}_{jk_j} \cdot n \cdot t}{\tilde{\sigma}_{jk_j} \cdot \sqrt{n}} < x\right) = \int_{-\infty}^{x} \exp\left(-\frac{y^2}{2t}\right) dy,$$

$$0 \leq t \leq 1, k_j = 1, 2, \ldots, r_j,\ j = 1, 2, \ldots, J.$$

Proof. These theorems are proved in [7], and the proof is therefore omitted here so as not to lengthen this short paper.

IV The reliability of a multi-server computer network

In this section, we prove the following theorem on the probability that a computer network fails due to overload.

If $t \geq \max\left(\max\limits_{1\leq j\leq p_j} \frac{m_{jk_j}}{\hat{\beta}_{jk_j}}, \max\limits_{1\leq j\leq r_j} \frac{\gamma_{jk_j}}{\tilde{\beta}_{jk_j}}\right)$ *and conditions (1) are fulfilled, the computer network becomes unreliable (all computers fail).*

Proof. At first, using Theorem 1 and Theorem 2, we get that for $x > 0$

$$\lim_{n\to\infty} P\left(\frac{Q_{jk_j}(nt) - \hat{\beta}_{jk_j} \cdot n \cdot t}{\hat{\sigma}_{jk_j} \cdot \sqrt{n}} < x\right) = \int_{-\infty}^{x} \exp\left(-\frac{y^2}{2t}\right) dy,$$

$$k_j = 1, 2, \ldots, p_j \quad (2)$$

and

$$\lim_{n\to\infty} P\left(\frac{V_{jk_j}(nt) - \tilde{\beta}_{jk_j} \cdot n \cdot t}{\tilde{\sigma}_{jk_j} \cdot \sqrt{n}} < x\right) = \int_{-\infty}^{x} \exp\left(-\frac{y^2}{2t}\right) dy,$$

$$k_j = 1, 2, \ldots, r_j,\ j = 1, 2, \ldots, J. \quad (3)$$

Let us investigate a computer network which consists of the elements (computers) α_j that are indicators of stations X_j, $j = 1, 2, \ldots, p_j$ and the elements (computers) γ_i that are indicators of stations Y_i, $i = 1, 2, \ldots, r_j$

Denote

$$X_j = \begin{cases} 1, & \text{if the element } \alpha_j \text{ is reliable} \\ 0, & \text{if the element } \alpha_j \text{ is not reliable,} \end{cases}$$

$j = 1, 2, \ldots, p_j$ and

$$Y_i = \begin{cases} 1, & \text{if the element } \beta_i \text{ is reliable} \\ 0, & \text{if the element } \beta_i \text{ is not reliable,} \end{cases}$$

$i = 1, 2, \ldots, r_j$.

Note that $\{X_j = 1\} = \{Q_j(nt) < k_j\}$, $j = 1, 2, \ldots, p_j$ and $\{Y_i = 1\} = \{V_i(nt) < \gamma_i\}$, $i = 1, 2, \ldots, r_j$. Denote the structural function of the system of elements, connected by scheme 1 from $p_j + r_j$ (see, for example, [8]), as follows:

$$\phi(X_1, X_2, \ldots, X_p, Y_1, Y_2, \ldots, Y_r, t) =$$

$$\begin{cases} 1, & \sum_{j=1}^{p_j} X_j + \sum_{i=1}^{r_j} Y_i \geq 1 \\ 0, & \sum_{j=1}^{p_j} X_j + \sum_{i=1}^{r_j} Y_i < 1. \end{cases}$$

Assume $y = \sum_{j=2}^{p_j} X_i + \sum_{i=1}^{r_j} Y_i$. Estimate the reliability function of the system (computer network) using the formula of conditional probability

$$h(X_1, X_2, \ldots, X_p, Y_1, Y_2, \ldots, Y_r, t) =$$

$$E\phi(X_1, X_2, \ldots, X_p, Y_1, Y_2, \ldots, Y_r, t) =$$

$$P(\phi(X_1, X_2, \ldots, X_p, Y_1, Y_2, \ldots, Y_r, t) = 1) =$$

$$P(\textstyle\sum_{j=1}^{p_j} X_j + \sum_{i=1}^{r_j} X_i \geq 1) =$$

$$P(X_1 + y \geq 1) = P(X_1 + y \geq 1 | y = 1) \cdot$$

ISBN: 978-1-61804-256-9

$P(y=1)+P(X_1+y\geq 1|y=0)\cdot P(y=0)=$

$P(X_1\geq 0)\cdot P(y=1)+P(X_1\geq 1)\cdot P(y=0)\leq$

$P(y=1)+P(X_1\geq 1)=P(y=1)+P(X_1=1)\leq$

$P(y\geq 1)+P(X_1=1)=$

$P(\sum_{j=2}^{p_j}X_j+\sum_{i=1}^{r_j}X_i\geq 1)+P(X_1=1)\leq\cdots\leq$

$\sum\limits_{i=1}^{m}\sum_{k_i=1}^{p_j}P(Q_{ik_i}(nt)\leq m_{jk_j})+$

$\sum\limits_{i=m+1}^{J}\sum_{k_i=1}^{r_j}P(V_{ik_i}(nt)\leq \gamma_{jkj}).$

Assuming that $k_j=p_j+r_j$

$0\leq h(X_1,X_2,\ldots,X_p,Y_1,Y_2,\ldots,Y_r,t)\leq$

$\sum\limits_{i=1}^{m}\sum_{k_i=1}^{p_j}P(Q_{ik_i}(nt)\leq m_{jk_j})+$

$\sum\limits_{i=m+1}^{J}\sum_{k_i=1}^{r_j}P(V_{ik_i}(nt)\leq \gamma_{jkj}).$ (4)

Applying Theorem 1, we obtain that for $m_{jkj}<\infty$

$0\leq \lim\limits_{n\to\infty}P(Q_{jk_j}(nt)<m_{jk_j})=$

$\lim\limits_{n\to\infty}P\left(\frac{Q_{jk_j}(nt)-\hat{\beta}_j\cdot n\cdot t}{\hat{\sigma}_j\cdot\sqrt{n}}<\frac{m_{jk_j}-\hat{\beta}_j\cdot n\cdot t}{\hat{\sigma}_j\cdot\sqrt{n}}\right)=$

$\int_{-\infty}^{-\infty}\exp\left(-\frac{y^2}{2t}\right)dy=0,$ (5)

where $k_j=1,2,\ldots,p_j$ and $j=1,2,\ldots,J.$

It follows from (5), that, for $m_{jk_j}<\infty$,

$\lim\limits_{n\to\infty}P\left(Q_{jk_j}(nt)<m_{jk_j}\right)=0,$ (6)

where $k_j=1,2,\ldots,p_j$ and $j=1,2,\ldots,J.$

Similarly as in (5) - (6), we prove that for $\gamma_{jk_j}<\infty$

$\lim\limits_{n\to\infty}P\left(V_{jk_j}(nt)<\gamma_{jk_j}\right)=0,$ (7)

where $k_j=1,2,\ldots,r_j$ and $j=1,2,\ldots,J.$

Consequently,

$\lim\limits_{n\to\infty}h(X_1,X_2,\ldots,X_p,Y_1,Y_2,\ldots,Y_r,t)=0$

(see (4), (6) and (7)), which completes the proof. □

V Concluding remarks and future research

1. Conditions (1) are fundamental, - the behaviour of the whole network and its evolution is not clear, if conditions (1) are not satisfied. Therefore, this fact is the object of further research and discussion.
2. Note that a computer with Windows operating system functions steadily if the number of jobs does not exceed 5 (therefore, $m_{jk_j}>=5$). In other cases, the computer fails (see paragraph 1).

References

[1] Gelenbe E. (1973). Unified approach to the evaluation of a class of replacement algorithms, IEEE Transaction on Computers, C22 (6): 611-618.

[2] Gelenbe E. (1979). Probabilistic models of computer systems, Part II, Acta Informatica, Vol. 12: 285-303, 1979.

[3] Gelenbe E., Finkel D., Tripathi S.K. (1986). Availability of a distributed computer system with failures, Acta Informatica, 23 (6): 643 - 655.

[4] Gelenbe E, Wang X. W., Onvural R. (1996). Diffusion based statistical call admission control in ATM', Performance Evaluation, 27-8: 411-436.

[5] Gelenbe E., Fourneau J. M. (2002). G-networks with resets, Performance Evaluation, 49(1-4): 179-191.

[6] Lin M. S., Chen D. J. (1997). The computational complexity of reliability problem on distributed systems, Information Processing Letters, 64(3): 143-147.

[7] Minkevičius S., Laws of the iterated logarithm in open, closed and mixed queueing networks, to be published, 2015.

[8] Morder J.J., Elmaghraby S. E. (1978), Handbook of operational research models and applications, Van Nostrand Reinhold, New York.

[9] Park K., Kim S. (2002), A capacity planning model of unreliable multimedia service, Journal of Systems and Software, 63(1): 69 -76.

[10] Pearn W. L, Ke J. C., Chang Y. (2004). C., Sensitivity analysis of the optimal management policy for a queueing system with a removable and non-reliable server, Computers and Industrial Engineering, 46 (1): 87-99.

[11] Perry D., Posner M. J. M. (2000). A correlated M/G/1-type queue with randomized server repair and maintenance modes, Operations Research Letters, 26(3): 137-148.

[12] Sakalauskas L. L., Minkevičius S. (2000), "On the law of the iterated logarithm in open queueing networks, European Journal of Operational Research, 120, 632-640.

[13] Sztrik J., Kim C. S. (2003). Markov-modulated finite-source queueing models in evaluation of computer and communication systems, Mathematical and Computer Modelling, 38(7-9): 961-968.

[14] Zubov V., Mathematical theory of reliability of queueing systems, Radio i Sviaz, Moscow, 1964.

Mathematical model for predicting process parameters in cold spray of porous Ti coatings

A. Hamweendo, P. A. I. Popoola, and I. Botef

***Abstract*—**Studies showed that there is no mathematical model to describe the correlation between the process parameters and the porosity of Titanium (Ti) coatings fabricated by cold spray process. Consequently, this paper proposes such a new mathematical model. This model was built using the second order polynomial regression modelling and MATLAB software. Central composite design of experiments was applied to generate and collate process parameters and porosity of Ti coatings. To verify and validate the new model, sensitivity and least square regression analyses were applied, and proved that the model is rigorous and therefore could be used to predict process parameters and/or porosity of Ti coating with high accuracy.

***Keywords*—**Cold Spray, Mathematical Model, Process Parameters, Titanium, Porosity.

I. Introduction

Mathematical methods are fundamental in modelling of phenomena in, for example, science, engineering, and economics [1],[2] and so, facilitate the description, optimization, analysis, forecasting, design and prediction of the final results of a process [3],[4],[5],[6],[8]. For this reason, Cold Spray (CS), which is the newest surface coating innovation, has received a lot of mathematical modelling attention [9], [10]. CS process, schematically presented in Fig.1, applies coatings by exposing a substrate to a high velocity (300-1200 m/s) jet of small (1-50 µm) particles accelerated by a supersonic jet of compressed gas [11]. The coating process takes place at a temperature always lower than the melting point of the powder's material, resulting in coating formation in the solid state. As a consequence, the deleterious effects of the high-temperature oxidation, evaporation, melting, crystallization, residual stresses, gas release, and other common problems arising from traditional thermal spray processes are minimized or eliminated [11]. These advantages made CS process an attractive deposition method for temperature and phase sensitive materials such as Titanium (Ti) that received renewed research attention due to its applications in aerospace and biomedical implants [12]. In these applications, Ti coatings with varying percentages of porosity are required. However, there is no mathematical model relating the input and output process parameters of Ti coatings fabricated using CS process. As a result, researchers and operators conduct costly test trials or extensive literature survey to find the best process parameters for their applications.

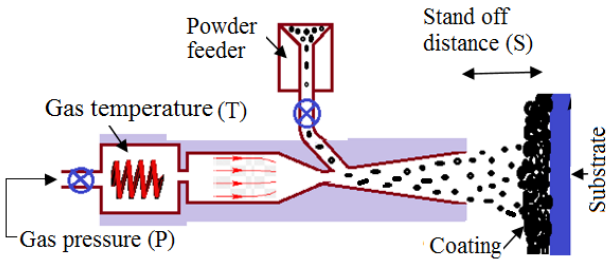


Fig.1 Schematic of the low pressure CS process

Consequently, this paper aims to develop and validate a new mathematical model for predicting process parameters in CS of Ti with various percentage of porosity. Subsequently, section 2 reviews relevant mathematical modelling techniques for process parameters. Section 3 introduces the methodology used for the new mathematical model. Then, in section 4, a summary of the mathematical modelling process is presented. Finally, section 5 draws conclusions about the research problem; highlights the theoretical and practical implications of the new proposed mathematical model, and indicates further research directions.

II. Short Literature Review

In mathematical modelling of process parameters, Thirumalaikumarasamy et al [13] recommended the application of the second-order regression polynomial modelling technique because of its ability to establish relationships and influences between the input and output parameters of a process. Also, Balasubramanian et al [14] used the Central Composite Design of Experiment (CCDOE) method to generate empirical data for their mathematical model and so estimate the grain size and hardness in welding of Ti-6Al-4V alloy. Finally, Demirel and Kayan [15] applied a

The support of the DST-NRF Centre of Excellence in Strong Materials (CoE-SM) towards this research is acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the CoE-SM.

A. Hamweendo is finalizing his PhD with the University of the Witwatersrand, School of Mechanical, Industrial and Aeronautical Engineering, Johannesburg, South Africa. (corresponding author: +27-11-717-7438; e-mail agripa.hamweendo@students.wits.ac.za).

P. A. I. Popoola is with Department of Chemical and Metallurgical Engineering, Tshwane University of Technology, Pretoria, South Africa (e-mail: popoolaapi@tut.ac.za).

I. Botef is with the University of the Witwatersrand, School of Mechanical, Industrial and Aeronautical Engineering, Johannesburg, South Africa. (e-mail: ionel.botef@wits.ac.za).

similar approach to optimize the interaction of oxygen pressure, temperature and time on the textile dye degradation by wet air oxidation. This was the basis for the following mathematical development model.

III. MATHEMATICAL MODEL DEVELOPMENT

The development of the mathematical model for predicting process parameters in cold spray of porous Ti coatings described here is a modification of the steps followed in [13] with new key steps such as: the selection of process parameters; development of experimental design matrix; conduction of experiments; and the building and verification of the model. The selection of process parameters was based on the literature survey and the working limits of the CS equipment, supplied by Centreline, Canada. Consequently, for this study, three process parameters which have predominant influence on the porosity of Ti coatings [12] were selected, namely: the stagnation temperature (T), stagnation pressure (P), and the standoff distance (S). The working limits for these parameters, which were determined from the deposition trials conducted in the CS laboratory at the University of the Witwatersrand, Johannesburg were as follows: stagnation temperature between 623 and 873 K; stagnation pressure between 0.7 and 9.5 MPa; and standoff distance between 10 and 30 mm.

To develop the design matrix, five coded levels, namely 1.682; -1; 0; +1; and 1.682 were used. These coded levels are coordinates of the location points where possible values of process parameters could be found [16]. The process parameters were calculated using (1), which relates the given coded level, X_i, with respective the process parameters X, and X_{max} and X_{min} are the upper and lower limits of process parameters [17]. The results of these calculations are presents in Table 1.

$$X_i = 1.682[2X - (X_{\max} + X_{\min})]/[X_{\max} - X_{\min}] \quad (1)$$

The process parameters and the coded levels in Table 1 were used to develop the experimental design matrix according to CCDOE [13], after calculations and collation, not included here due to space constraints, are presented in Table 2.

Table 1. Selected CS process parameters and their levels.

Parameters	*Notation*	*Units*	*Level*				
			-1.682	-1	0	1	1.682
Temperature	T	K	623	664	723	782	823
Pressure	P	MPa	0.7	0.75	0.825	0.9	0.95
Standoff Distance	S	mm	10	14	20	26	30

Table 2. Design matrix and experimental results.

Expt	*Design matrix*			*Parameters*			*Porosity*	
				T	P	S	Measured	Predicted
1	1	-1	-1	782	0.75	14	4.6	6.8
2	-1	1	-1	664	0.9	14	9.6	10.2
3	1	1	-1	782	0.9	14	2.8	4.3
4	-1	0	-1	664	0.825	14	25.8	27.7
5	1	-1	1	782	0.75	26	19.3	19.5
6	-1	1	1	664	0.9	26	31.5	29.2
7	1	1	1	782	0.9	26	9.2	9.1
8	1.652	0	0	823	0.825	20	0.8	0.8
9	0	1.682	0	723	0.95	20	1.4	0.0

IV. CS EXPERIMENTS AND MODEL VALIDATION

The CS experiments were conducted to acquire the requisite data for building and validating the model. To collect this data, Ti was deposited on a grit blasted steel substrate. Titanium powder, brand code SST5001 supplied by Centerline, Canada, was of a size range between 10-50 µm. The nozzle used was a de-Laval converging/diverging type with 120 mm diverging length, 2 mm throat diameter, and 6.5 mm exit diameter. Air was used as both process and powder carrier gas. The robot manipulated the spray gun at traversing speed of 10 mm/s throughout the experiments. After deposition, the coatings were sectioned and metallographically polished using the Struers automatic polishing machining. The porosity of the coatings was measured by taking optical images of the polished sections, followed by the analysis of the images using ImageJ

software. The results of the measured porosity are presented in Table 2.Furthermore, the following section is an example of the use of the mathematical model introduced in section 3 but also not completely presented here due to the space constraints.

To build the mathematical model, a generalised second order polynomial regression model presented in (2) was used, and where: Y is the response; b_0 is the average of the responses; n is the number of variables; x_i and x_j are the variables; b_i, b_j, and b_{ij} are the coefficients for the first order, second order and the interacting variables, respectively. Then, (2) was expanded as a generic polynomial regression model with (3) as result.

By collating the measured porosity, and T, P, S, (from Table 2) the simultaneous equations (not shown here) were developed and re-written into a matrix (4).

To calculate the 'b' coefficients in (3), (4) was rearranged into (5) which was solved using matrix algebra and MATLAB software. Then, by replacing the values of these coefficients in (3), the mathematical model was built and shown as (6).

Furthermore, the verification of the model was accomplished through sensitivity analysis (SA)[13], while regression analysis (RA)[18] was applied to validate this model. SA identifies critical parameters and their affect to the model output. The resultant partial differential equations in (7), (8) and (9) were used to analyse the change of porosity of Ti coating with respect to individual process parameters.

$$Y = b_0 + \sum_{i=1}^{n} b_i x_i + \sum_{i=1}^{n} b_{ii} x_i^2 + \sum_{i=1}^{n-i} \sum_{j=i+1}^{n} b_{ij} x_i x_j \tag{2}$$

$$Porosity = b_0 + b_1 T + b_2 P + b_3 S + b_{11} T^2 + b_{22} P^2 + b_{33} S^2 + b_{12} TP + b_{13} TS + b_{23} PS \tag{3}$$

$$[Porosity] \approx [x][b] \tag{4}$$

where:

$$[Porosity] = \begin{bmatrix} 4.6 \\ 9.6 \\ 2.8 \\ 25.8 \\ 19.3 \\ 31.5 \\ 9.2 \\ 0.8 \\ 1.4 \end{bmatrix} \quad [x] = \begin{bmatrix} T & P & S & T^2 & P^2 & S^2 & TP & TS & PS \\ 782 & 0.75 & 14 & 611524 & 0.5625 & 196 & 586.5 & 10948 & 10.5 \\ 664 & 0.9 & 14 & 440896 & 0.81 & 196 & 597.6 & 9296 & 12.6 \\ 782 & 0.9 & 14 & 611524 & 0.81 & 196 & 703.8 & 10948 & 12.6 \\ 664 & 0.825 & 14 & 440896 & 0.6806 & 196 & 547.8 & 9296 & 11.6 \\ 782 & 0.75 & 26 & 611524 & 0.5625 & 676 & 586.5 & 20332 & 19.5 \\ 664 & 0.9 & 26 & 440896 & 0.81 & 676 & 597.6 & 17264 & 23.4 \\ 782 & 0.9 & 26 & 611524 & 0.81 & 676 & 703.8 & 20332 & 23.4 \\ 823 & 0.825 & 20 & 677329 & 0.6806 & 400 & 679.0 & 16460 & 16.5 \\ 723 & 0.9 & 20 & 522729 & 0.9025 & 400 & 686.8 & 14460 & 19.0 \end{bmatrix} \quad [b] = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_{11} \\ b_{22} \\ b_{33} \\ b_{12} \\ b_{13} \\ b_{23} \end{bmatrix}$$

$$[b] \approx [x]^{-1}[Porosity] \tag{5}$$

$$\begin{aligned} Porosity = {} & 11.6656 - 0.2866T + 299.8704P + 11.0097S - 0.00057T^2 - \\ & 788.452P^2 + 0.0333S^2 + 1.3389TP - 0.01014TS - 4.4700PS \end{aligned} \tag{6}$$

$$\partial(Porosity)/\partial T = -0.2866 - 0.00114T + 1.3389P - 0.01014S \tag{7}$$

$$\partial(Porosity)/\partial P = 299.8704 + 1576.9P + 1.3389T - 0.4700S \tag{8}$$

$$\partial(Porosity)/\partial S = 11.0097 + 0.0666S - 0.010014T - 4.4700P \tag{9}$$

In (7), the negative signs of the first two terms imply that an increase in temperature results in a decrease in predicted porosity of Ti coatings. This is in total agreement with the experimental results reported by Zahiri et al [12]. In this case a higher gas temperature in CS process tends to reduce the porosity of Ti coatings [12]. In (9), the positive signs of the first two terms denote that the predicted porosity of the coatings increase with the standoff distance (S), which is also in total agreement with the experimental results by [12]. However, the positive pressure terms in (8) gives inconclusive interpretation as they contradict the existing theory [12]. Therefore, to bring more understand to the

influence of the change in gas pressure on porosity of Ti coatings, (6) was used to predict the porosity for Ti for the range of process parameters. The predicted results were plotted in graphs as, shown in Fig.2 and Fig.3. These figures illustrate that an increase in gas temperature and pressure leads to a reduction of predicted porosity, while the opposite is true for the standoff distance. This variation of predicted porosity for Ti coatings coincides well with what Zahiri et al. observed [12].

Fig.2 shows that at a low temperature (T=400 °C and S=30 mm), if pressure is increased from 7 to 9 bars (an increase of 28.5 %), this results in the decrease in porosity of Ti coating from 78 to 30 % (a decrease of 61.5 %), while at S=10 mm, porosity decrease from 29 to 7 % (a decrease of 75 %). This suggests that the change in stagnation pressure has more influence on porosity, with this influence being predominant at smaller standoff distance than at larger standoff distance.

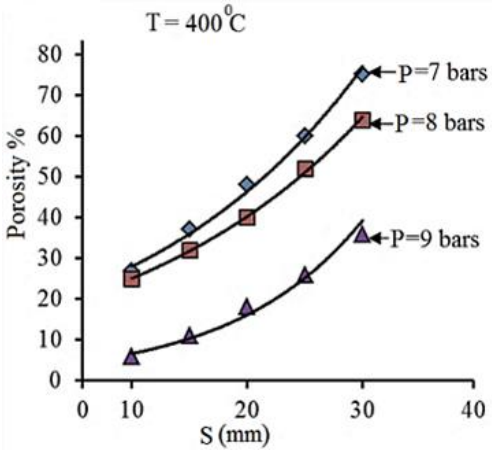


Fig.2. Variation of predicted porosity with S, P and T for Ti coatings for T = 400 °C.

Fig.3 shows that at 500 °C the porosity of Ti coating increase steeply with the standoff distance at lower pressure of 7 bars. At S of 30mm, an increase in pressure from 7 to 9 bars (an increase of 28.5 %) results in the decrease of porosity of Ti from 27 to 15 % (a decrease of 44.4 %), while at S of 10 mm, porosity decrease from 5 to 0.1 (a decrease of nearly 100%) for the same increase of pressure. Similarly, this suggests that stagnation pressure has predominant influence on porosity of Ti coating at lower standoff distance and also clarifies that an increase in stagnation pressure has negative influence on porosity of the coatings.

Furthermore, the Regression Analysis (RA) validated the mathematical model through defining the relationship between predicted and measured values [13]. In this study, the predicted and measure porosity values for Ti coatings were calculated using (6) and plotted as shown in Fig.4. In this graph, the resultant least square regression line showed a good fit with the scatter plot and the coefficient of determination for the graph of 98.3 %. This means that the variation of predicted porosity with the measured porosity is 98.3 % collinear.

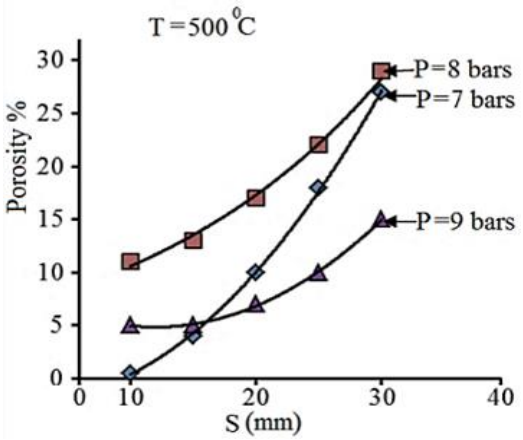


Fig.3. Variation of predicted porosity with S, P and T for Ti coatings for T = 500 °C.

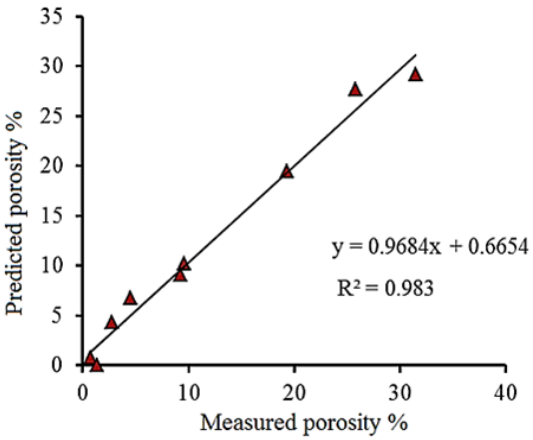


Fig.4. Variation of predicted and measured porosity for Ti coatings.

V. CONCLUSIONS

In this paper, a new mathematical model for predicting process parameters in CS of Ti was developed through the use of CCDOE, polynomial regression modelling, matrix algebra and MATLAB software. The CCDOE generated the design matrix for the process parameters in CS experiments.

The model was verified and it showed a high accuracy in predicting the process parameters and/or porosity of Ti coatings. Consequently, it can be concluded that the new mathematical model could eliminate the actual costly trial experiments used to establish the desired process parameters to be used for a specific required porosity.

Further research recommended may include the application of this new model and model development process to the cold spraying of other materials such as Aluminium, copper, zinc, and nickel. In addition, the selected limits for the stagnation temperature (T), stagnation pressure (P), and the standoff distance (S) can be extend to

ISBN: 978-1-61804-256-9

include the limits of other Low Pressure and High Pressure Cold Spray systems as well as using nitrogen and helium as operating gases.

REFERENCES

[1] A. Ochoche, "On Error Estimation In General Linear Methods: Runge Kutta (Rk) and Almost Runge-Kutta (Ark) Methods, *Proc. of the 2013 Inter'l Conf. on Applied Math. and Comp. Methods in Engineering* , Recent advances in AMCME, Rhodes Island, Greece, 2013, pp.126-130.

[2] V. M. Kovenya, "Problems and trends in mathematical modelling", *J. of Applied Mechanics and Technical Physics*, vol. 43, Inst.of Comp.Tech., Siberian Div., Russia, Plenum Publishers, 2002, pp. 345–353.

[3] G. Mutanov, and Zh. Yessengalieva, "Qualitative information method of an assessment of scientific and innovative projects during implementation of the industrial and innovative program in Kazakhstan, *Proc. of the 2014 Inter'l Conf. on AMCME*, Prague, Czech, Republic, pp132-136.

[4] V. Passannante, M. Sibillo, and V.D'Amato "The prediction of mortality by causes of death in Critical Illness, *Proc. of the 2014 Inter' Conf. on AMCME*, Prague, Czech Republic, pp.75-80.

[5] Ibrahim, M. A.A., "Forecasting Trend of Traffic Fatalities in the United Arab Emirates", *Proc. of the 2014 Inter'l Conf. on AMCME*, Prague, Czech Republic, pp.81-86

[6] M.Ehrgott, C. Güler, H.W. Hamacher, L. Shao, "Mathematical optimization in intensity modulated radiation Therapy", *Annals of Operations Research,* vol.175, no.1, Springer-Verlag, 2008, pp.309-365

[7] F. A. Maksimov, D. A. Churakov, and Yu. D. Shevelev, "Development of Mathematical Models and Numerical Methods for Aerodynamic Design on Multiprocessor Computers, *Computational Mathematics and Mathematical Physics*, Vol. 51, No. 2, Inst. for CAD, Moscow, Pleiades Publishing, Ltd, 2011, pp. 284–307

[8] V. V. Penenko and E. A. Tsvetova, "Mathematical models for studying environment pollution risks", *Journal of Applied Mechanics and Technical Physics*, Vol. 45, No. 2, Institute of Comput. Math. and Math. Geophysics, Siberian Div., Russia, Plenum Publishing Corp., 2004, pp. 260–268.

[9] A. P. Alkhimov, V. F. Kosarev, and S. V. Klinkov, "The Features of Cold Spray Nozzle Design", *J. of Thermal Spray Tech.*, vol. 10, pp. 375-381, 2001.

[10] W. Wong, P. Vo, E. Irissou, A. N. Ryabinin, J.-G. Legoux, and S. Yue, "Effect of Particle Morphology and Size Distribution on Cold-Sprayed Pure Titanium Coatings", *J of Thermal Spray Tech.*, vol. 22, pp. 1140–1153, 2013.

[11] A. Papyrin, V. Kosarev, S. Klinkov, A. Alkimov, and V. Fomin, *Cold Spray Technology*, Elsevier, 2007.

[12] S. H. Zahiri, C. I. Antonio and M. Jahedi, "Elimination of porosity in directly fabricated titanium via cold gas dynamic spraying", *J of Thermal Spray Tech.*, vol. 209, pp. 922-929, 2009.

[13] D. Thirumalaikumarasamy, K. Shanmugam and, V. Balasubramanian, V, "Establishing empirical relationships to predict porosity level and corrosion rate of atmospheric plasma-sprayed alumina coatings on AZ31B magnesium alloy", *J. of Magnesium and Alloys*, pp.1-14, 2014.

[14] M. Balasubramanian, V. Jayabalan, V. Balasubramanian, "Developing mathematical models to predict grain size and hardness of argon tungsten pulse current arc welded titanium alloy", *J of Mat. Proc. Tech.*, vol. 196, pp. 222-229, 2008.

[15] M. Demirel and B. Kayan, "Application of response surface methodology and central composite design for the optimization of textile dye degradation by wet air oxidation", *I J of Ind Chem*, Springer Open, 2012.

[16] D. C. Montgomery, "Design and Analysis of Experiments", 6th Edition, John Wiley and Sons, 2005.

[17] R. G. Miller, J. E. Freund, D. E. Johnson, "Probability and Statistics for Engineers", *Prentice Hall of India Pt Ltd*., New Delhi, 1999, p. 75.

[18] S. Karthikeyan, V. Balasubramanian, and R. Rajendran, "Developing empirical relationships to estimate porosity and microhardness of plasma-sprayed YSZ coatings", *Ceramics International*, vol. 40, 2014, pp. 3171 – 3183,.

Medical Images Understanding based on Computational Intelligent Techniques

Abdalslam AL-Romimah
ACC and AL Saeed University
Department of IT
Yemen
lapromr@gmail.com

Amr Badr
Cairo University
Faculty of Computers and Information
Egypt
a.badr@fci-cu.edu.eg

Ibrahim Farag
Cairo University
Faculty of Computers and Information
Egypt
I.Farag@fci-cu.edu.eg

Abstract: A computational intelligent system for regions of interest (ROIs) understanding is presented. It constitutes of fuzzy pulse-couple neural networks (FPCNNs) for ROIs and automatic understanding based on integer-CHC genetic algorithm (ICHCGA) with fuzzy artmap neural networks (FAMNNs). The system is applied on mammogram images, the mammogram understanding method consisting essentially of, automatic segmentation method based Fuzzy-PCNNs, and classification method based on ICHCGA feature selection and receiver operating characteristic (ROC) is generated by FAMNN for performance evaluation. The distinction between normal and abnormal cases by FAMNN is carried out by generated areas under ROC curve ranging from 0.88000 to 0.98604, whereas distinction by MLPNNs is carried out by generated areas under ROC curve ranging from 0.72000 to 0. 86936. FAMNN is used the distinction between benign and malignant mass is with fitness degree of 98.00 ranging from 0.87000 to 0.97845 under ROC curve, whereas distinction by MLPNNs with fitness degree of 92.00 ranging from 0.87000 to 0.95702 under ROC curve.

Key–Words: Digital Mammography, Fuzzy-PCNNs, FAMNN, Feature Extraction,Wavelet, integer-CHC genetic algorithm, ROC.

1 Introduction

This field of bioinformatics is a crossway of numerous academic fields simultaneously; it provides an interface between medical sciences and information technology, using the most recent computerized means in analysis of medical images and thus diagnosing diseases and disorders on basis of establishing more analysis and understanding models for medical images. Recently, in the last 30 years, there has been massive increase in the field of medical equipment and information technology in diagnostic imaging, the researchers have developed many different aid methods and systems in a field of biomedical informatics, whether traditional or computational intelligence techniques to improve tools of diagnostic effectiveness. For mammogram segmentation techniques [1] unsupervised and supervised approaches also known as model-based segmentation. Supervised approaches depend on prior information about image components if only objects or background. In unsupervised segmentation methods, image is partition into the set of regions dependent on specific features, intensity value, shape, texture and color. S. Fu and J. K. Mui [2] divided unsupervised segmentation into three major groups: region-based methods, contour-based methods and clustering methods. Of course, all categorizations types for segmentation techniques are based upon color, intensity, or texture characteristics. Like Fu and Mu [3] considered the threshold methods as a special case of partitional clustering methods; where only two clusters are considered, threshold methods have been widely used for mass segmentation. There are two types of thresholding value which are used for image segmentation, hard and soft thresholding techniques, the hard thresholding techniques categorize in six groups as follows: histogram shape-based methods, clustering-based methods, entropy-based methods, object attribute-based methods, the spatial methods use higher-order probability distribution and local methods adapt the threshold value [4]. More recently, many studies for mammogram classification are presented; an automated mass detection method is presented by Timp et al [5] to detect temporal changes in mammographic masses between two consecutive screening rounds. Two kinds of temporal features, difference features and similarity features are designed to realize the interval change analysis. A SVM is employed as a classifier to detect the temporal changes in mammographic masses. The classification performance is evaluated with and without the use of temporal features. In experimental results, the database consisted of 465 temporal mammogram pairs contain-

ISBN: 978-1-61804-256-9

ing 238 benign and 227 malignant cases. The Az = 0.74 without temporal features and 0.77 with the use of temporal features. Lcio et al [6] proposed an independent component analysis a feature extraction method and classification of mammograms with benign, malignant and normal tissues using three neural networks: MLPNN, probabilistic NN and radial basis function NN. The best performance is obtained with probabilistic NN, resulting in 97.3 success rate, with 100 of specificity and 96 of sensitivity. Retico et al [7] used the 16 features based on size and shape of the lesion are extracted: (area, perimeter, circularity, mean and the standard deviation of the normalized radial length, radial length entropy, zero crossing, maximum and the minimum axis of the lesion, mean and the standard deviation of the variation ratio, convexity, the mean, the standard deviation, the skewness and the kurtosis of the mass grey-level intensity values). For classification a standard three-layer feed-forward NN classifier merges the features into an estimated likelihood of malignancy. A data set of 226 massive lesions (109 malignant and 117 benign) is used. The system performances are evaluated in terms of the ROC analysis, obtaining Az ranging 0.80 0.04 as the estimated Az. Pasquale et al [8] used the same 16 features extracted by [7] and the same dataset, but here feature selection procedure that are carried out on the basis of the feature discriminating power and of the linear correlations interplaying among them. 12 selected features out of the 16 computed use the Az of ROC evaluation with MLPNN. A MLPNN classifier is trained by error back-propagation algorithm. The masses dataset divided to 3 different categories: correctly, acceptably and non-acceptably segmented masses, Az=0.8050.030, 0.7870.024 and 0.7800.023, respectively.

This paper is structured in foure sections. In Section 2 automatic mammogram understanding is presented, it consisting essentially of, automatic segmentation based on Fuzzy-PCNNs, and automatic mammogram classification based on ICHCGA feature selection is performed FAMNNs categories classification. In section 3 FAMNN and MLPNNs evaluation results presented, and in section 4 the conclusions and futre work.

2 Automatic mammogram understanding

An automatic mammogram understanding method relates to improvements in image understanding methods, it consisting essentially of, automatic segmentation method based on fuzzy-pulse-couple neural networks (Fuzzy-PCNNs), and classification method based on integer-CHC genetic algorithm (ICHCGA) feature selection is performed with fuzzy artmap neural networks (FAMNNs) categories classification method.

2.1 An Automatic Segmentation

An automatic segmentation method based on Fuzzy-PCNNs method relates to improvements in image segmentation methods and systems. Fuzzy rule inference and fuzzy entropic threshold are adapted to improve a performance of PCNNs for image segmentation. Fuzzy entropic threshold is computed according to fuzzy max entropic that is depended on the image normalization, 2D image histogram and fuzzy partition, thus fuzzy entropic thresholding is adapted for PCNNs thresholding matrix ij [n]. The fuzzification and fuzzy rule are adapted to compute the coefficient matrix(i,j) (n) of a linking modulation layer of Fuzzy-PCNNs based fuzzy rule inference between the pixel and surround pixels in the image matrix, thus Fuzzy-PCNNs method consisting essentially of, feeding and linking layers, Fuzzy-PCNNs filter based on the inverse of 2D Laplacian of Gaussian filter method of the resulted images after remove non-information regions, Fuzzy-PCNNs thresholding and Fuzzy-PCNNs pulse generator. Fuzzy-PCNNs filter is adapted as sharpening or high-pass filter, allow high frequencies pass and reduce the lower frequencies, and consequence is extremely sensitive to shut noise. To construct a high-pass filter, the kernel coefficients should be set positive near a center of kernel and in the outer periphery negative, for more details about Fuzzy-PCNNs see the equations from 2 to 6. The sequences of binary resulted images are filled by a polygon mask.

2.1.1 Fuzzy entropic threshold

It is computed according to fuzzy max entropic [9] that is depended on the image normalization, 2D image histogram and fuzzy partition, thus it is adapted to get Fuzzy-PCNNs thresholding matrix ij [n].

2.1.2 Image normalization

Normalization of the image resulted based on min-max normalization formula NZ_S see eq.[1], this image having gray levels ranging from lmin to lmax can be modeling as an array of fuzzy number; each element in the array is the value representing the degree of brightness of gray level between 0 and 1.

$$NZ_S = \frac{\sum_{i=0}^{N} x(i,:) - l_{min}(i,:)}{l_{max} - l_{min}} \quad (1)$$

ISBN: 978-1-61804-256-9

2.1.3 2D histogram

2D histogram method proposed by Kirby and Rosenfeld [10] that is computed of resulted images, where each the bin of the 2D histogram represent a frequency of occurrence of each (level, local average gray level) pair. The bins form a surface with ideally two peaks corresponding to background and object regions. Thus, the pixels interior to the object or background are found mainly to the near-diagonal bins of a 2D histogram and off-diagonal bins being contributed by edges and noise in the region. For an n grey-level region there are obviously x^2 bins. By means of two thresholds S and T a 2D histogram is divided into 4 quadrants. Since the shaded quadrants of 2D histogram will in general contain information only about edges and noise that are ignored in the calculation. The quadrants 0 and 1 contain the distributions corresponding to the background and object classes.

2.1.4 Fuzzy partition

In this section see [9], fuzzy entropy is adapted for image thresholding based on both intensity distribution and local information among pixels. The purpose of this method is to automatically determine the fuzzy region and optimal decay threshold parameter, therefore matrix of threshold $_ij$, which is based on fuzzy entropy principle, given 2D histogram array of L_ij region NM and K gray levels. The 2D histogram is divided to three regions: background region, fuzzy region and bright region. The background region is defined as the region with left top point (0, 0) and right button point (c, c). The overlapping region denoted by fuzzy-region, which starts at point (a, a) and ends at point (c, c). For more details see the steps (3, 4, 5, and 6) in Fuzzy-PCNN method as it show below.

2.1.5 Fuzzy-PCNNs Model

Firstly, fuzzy pulse-coupled neural networks (Fuzzy-PCNNs) as developed model of PCNNs [11] shown as follows:

$$F_{ij}\,[n] = e^{-\alpha F \delta n}\,.F_{ij}[n-1] + s_{ij} + VF \sum_{kl} M_{ijkl} Y_{kl}[n-1] \tag{2}$$

$$L_{ij}\,[n] = \sum_{kl} M_{ijkl} Y_{kl}[n-1] \tag{3}$$

$$U_{ij}[n] = F_{ij}[n].*(1+\beta_{ij}(n).*L_{ij}\,[n] \tag{4}$$

Where$_ij$=defuzzification (centroid method of $_ij$ as shown in fuzzy rules.

$$Y_{ij}[n] = \begin{cases} 1 & if\ U_{ij}\ [n]\ > \theta_{ij}\ [n-1] \\ 0 & Othewise \end{cases} \tag{5}$$

$$\theta_{ij}\,[n] = e^{-\alpha\theta\delta n}\,.\theta_{ij}\,[n-1] + V\theta Y_{ij}\,[n] \tag{6}$$

Where n=eq.11 in (see Fuzzy-PCNNs thresholding).

According to equations from 2 to 6 and 11, Fuzzy-PCNNs consisting essentially of, feeding and linking layers, Fuzzy-PCNNs filter based on the inverse of 2D Laplacian of Gaussian filter of the resulted images, Fuzzy-PCNNs thresholding and Fuzzy-PCNNs pulse generator. The main purpose of this method is separation the mammogram image regions well and full robotic. Fuzzy-PCNNs system consisting essentially of, 11 components shown as follows: Filter component is an inverse of 2D Laplacian of Gaussian filter of the resulted images, which is adapted as sharpening or high-pass filter, let high frequencies pass and reduce the lower frequencies, and consequence is extremely sensitive to shut noise. To construct a high-pass filter, the kernel coefficients should be set positive near a center of kernel and in the outer periphery negative. Feeding and linking component, fuzzification component of linking coefficient (), fuzzy rule inference component of linking coefficient, defuzzification component, linking modulation component of Fuzzy-PCNNs, Fuzzy-PCNNs thresholding component, which is computed according to fuzzy max entropic and Fuzzy-PCNNs pulse generator component, sequence of resulted images and polygon mask is adapted to fill the ROIs that are resulted from Fuzzy- PCNNs.

1. Fuzzy entropic thresholding

As it shown in above.

2. Fuzzy-PCNNs filters

A Fuzzy-PCNNs filter is an inverse of 2D Laplacian of Gaussian filter of the resulted images. Which is adapted as sharpening or high-pass filter, let high frequencies pass and reduce the lower frequencies, and consequence is extremely sensitive to shut noise. To construct a high-pass filter, the kernel coefficients should be set positive near a center of kernel and in the outer periphery negative.

3. Fuzzy-PCNNs feeding and linking

This component represents the Fuzzy-PCNNs feeding and linking see eqs. (2, 3).

4. Fuzzification of linking coefficient β

In this component a fuzzification of linking coefficient β is presented. The pixels in nn nieghbarhood region X surrounding each pixel (i, j) from feeding inputs,$x_(i,j)$ is gray level of (i, j) pixel in X. Let $_X(x_(i,j))$ denote the membership value represents the degree of coefficient between (i, j) pixel (F_ij in PCNNs) with (L_ij in PCNNs) in X. A fuzzy membership of region set X is mapping from X into interval

[0, 1]. For membership function, the homogeneity and edgeness measures are computed as fuzzy rules [12].

5. Fuzzy rule of linking coefficient

The degree of membership for linking coefficient parameterβ(i,j) is calculated as a matrix of values to knowing which the pixel Fij with surrounding neighborhood region Lij belongs to the four types $(very_high, high, low, very_low)$. The input space of the linguistic variable H(i,j) is comprised of the three fuzzy sets (low, med, high), and E(i,j) is comprised of two fuzzy sets labeled (low, high). Fuzzy rules can be defined as a conditional statement in the form:

$$if\ (H_{i,j}\ is\ low\ )\quad then\quad beta_{i,j}\quad is\, very\_high \tag{7}$$

if
$(H_{i,j}\ is\ med)\ AND\ (E_{i,j}\, is\ high)\ OR\ (H_{i,j}\ is\ high)$
$AND\ (E_{i,j}\, is\ high)\quad then\, \beta_{i,j}\ is\, high$

(8)

if
$H_{i,j}\, is\ med\quad AND\ E_{i,j}\, is\ low\ then\, \beta_{i,j}\, is\, low$

(9)

if
$H_{i,j}\, is\ low\, AND\ E_{i,j}\, is\ low\, then\, \beta_{i,j}\, is\, very\_low$

(10)

Where H(i,j), E(i,j) and β(i,j) are linguistic variables and (low, med, high), (low, high) and (veryhigh, high, low, verylow) are linguistic values determined by fuzzy sets on the universes of discourse X and Y respectively. And fuzzy OR is defined as max (a, b) and fuzzy AND is defined as min (a, b).

6. Defuzzification of linking coefficient

In this component, a defuzzification of fuzzy rule of linking [12, 13].

7. Fuzzy-PCNNs linking modulation

In this component, the linking modulation of Fuzzy-PCNNs is represented see eq.3.

8. Fuzzy-PCNNs thresholding In this component, in Fuzzy-PCNNs, an optimal decay thresholding parameter $\alpha\theta\delta n$ is calculated as follows:

$$\alpha\theta\delta n = \max\left(\left(t_{tiss} + \mu_{maxfn}\right), \max\left(\mu_{maxmax}\right)\right) \tag{11}$$

Where μ_{maxfn} is the global maximum fuzzy entropy [14] of a normalize image (maximum entropy of fuzzy number) in fuzzy partition unit 36, t_{tiss} is the coefficient based on the type of mammogram tissue, which is determined based on the experiments. And the $max(\mu_{max}\ ))$ is the maximum fuzzy entropy of matrix. Where is a matrix of membership for optimal thresholding (maximum fuzzy entropy) for feeding with its surrounding neighborhood.

9. Fuzzy-PCNNs pulse generator

See eq.5.

10. Resulting images

In this component, a sequence of binary resulted images. See eq.5.

11. Filling of ROIs

In component polygon mask is adapted to fill the ROIs that are resulted from Fuzzy- PCNNs system. In binary image ,the regions of interest are detected by polygon mask (tracing boundary contours) and fill it using filtering the ROI from original image, which returns an image that consists intensity values for pixels in locations where ROI image contains 1's, and unfiltered values for pixels in locations where ROI image contains 0's. Then save the each region of interest as image.

2.2 Fuzzy-PCNNs (Pseudo-Code)

Fuzzy-PCNNs system steps for mammogram mass segmentation and micro-calcification detection passes through various components as shown in follows:

1. Before use Fuzzy-PCNNs for mammogram mass segmentation and micro-calcification detection, a system of automatic tissue types identification is worked, a MLPNNs classifier is adapted to know the tissue type. If tissue type is not a dense tissue, in the other hand, one from first four types, a mammogram image is enhanced by AHE method on a specific range of gray levels.
2. Image normalization.
3. 2D histogram is calculated one only.
4. Set initial values of all Fuzzy-PCNNs parameters and matrixes.
5. $\beta i, j(n)$ is calculated at each iteration. The coefficient parameter $\beta_{i,j}(n)$ is different from $F_{i,j}$ with its surrounding pixels in the same region to other. Thus the coefficient degree is determined based on the strong relationship between the pixels with its surrounding pixels.
6. Given an eq. 11 an optimal decay thresholding parameter $\alpha\theta\delta n$ is obtained base on a maximum of local maximum fuzzy entropy matrix μ_{max}, the global maximum fuzzy entropy of the normalize image (maximum entropy of fuzzy number), and a value of parameter, which is based on MLPNNs results.

ISBN: 978-1-61804-256-9

7. The binary images are resulted using the Fuzzy-PCNNs method. These images are included the ROIs (exactly a first binary image). All the ROIs are detected by polygon mask to draw each ROI separately. Therefore a new binary image for each ROI is created separately based on values of its boundary (by tracking boundary is aforementioned). Each boundary of a ROI has same the location in original image. The ROI is filled using a filtering it with original image.

2.3 Automatic mammogram classification

Automatic mammogram classification relates to improvements in computational intelligent methods for classification of medical images. This method consisting essentially of, feature extraction, feature selection and classification.

2.3.1 Feature extraction

In this section, various special methods are adapted to extract the ROIs features and generate a features matrix. Textural features such first order statistics, second order statistics features of well-known gray level co-occurrence matrixes (GLCMs), gray level run length matrixes (GLRLMs) features, fractional dimension features and multi-level wavelet decomposition features. Shape features and density features also are extracted.

2.3.2 Feature selection

Mammogram feature selection relates to improvements in computational intelligent methods for the medical images classification. Integer-CHC Genetic Algorithm (ICHCGA) is proposed to attain a best balance between the exploration and exploitation [15], CHC (cross-generational elitist selection, heterogeneous recombination, and cataclysmic mutation). This is accomplished by maintaining diversity in the population and allowing the algorithm to focus in several areas of search space simultaneously, and it is used to force diversity onto a population. A CHC algorithm is developed to solve the problems of premature convergence that genetic algorithm frequently suffers, and it uses a conservative strategy of selection. In ICHCGA, integer-coded is adapted in lieu of binary coded, because the last one require a decodification step to apply the fitness function and also does not fit well when the number of features is fixed.

1. Integer coded :
   For feature subset selection integer coded is not require a decodification step to apply the fitness function and does fit well when the number of features is fixed.

2. Fitness function :
   $$Fitness(subset) = |(accuracy/FAMNNerrorrate)| \quad (12)$$

3. CHC method:
   ICHCGA based on CHC (cross-generational elitist selection, heterogeneous recombination, and cataclysmic mutation) is adapted to force diversity onto a population, when it may have become trapped around a sub-optimal solution
   (a) Elitist selection:
   This method is one of the elitist steady-state selection algorithms, which explicitly borrow from the (+) evolutionary strategies [16, 17, 18]. It is based on survival of fittest instead of reproduction with emphasis; the survivors are chosen from the old parent population to the next generations parent population and select the remaining members from the offspring population. And the survivors are the elite chromosomes having the best criterion value determined by the fitness function.
   (b) Incest Prevention:
   To avoid premature convergence, ICHCGA employs the incest prevention mechanism, which can be used to promote exploration at the start of the search. If the minimum difference between parents is relatively large, the offspring will be sufficiently different to promote exploration. As this required difference decreases in later generations, the similarity of the parents and therefore the offspring increases and this focuses the search into a particular region of search space. In other words, two parents are only mated, if their Hamming distance (in binary coded) is above a threshold and an increase in the mutation probability is not required. Therefore, before applying HUX to two parents, the dissimilarly between them is measured by a Hamming distance of the gene strings, which is a count of number of the differing bits. In case, the integer or real coding, the dissimilarly is obtained by sum of the absolute differences between the values across all loci (their Manhattan distance or Euclidean distance). The individuals are able (or allowed to them) to mate and produce offspring,

ISBN: 978-1-61804-256-9

only if, the average of Euclidean distance is above of a certain (mating threshold) is achieved. This mean the elite chromosomes are rank-ordered from top down only chose points that have a decoded, Euclidean distance greater than a threshold from all previously selected points. Only these points are used in mating and the parents and offspring are used to cast out several new offspring.

(c) HUX crossover:
Using HUX, the substrings are switched between offspring with a probabilityP_cross; and this probability decreases with each generation. In essence this is a biased uniform crossover between integer-coded strings (where the fitness of the parent determines the probability that its gene will be expressed), and the bias increases with each generation. Although the elite chromosome was paired with another parent chromosome it remained unchanged after crossover was performed.

(d) cataclysmic mutation:
To keep on the production of offspring with maintains diversity and slows population convergence a cataclysmic mutation (called re-start mechanism) is applied. According to CHC adaptive algorithm, the value of this cataclysm threshold is decreased as the population converges and individuals become more similar. This threshold is calculated as follows: the initial threshold is set at$L/4$, where L is the length of the chromosomes, or threshold:= $MP * (1.0 - MP) * L$, where MP is mutation probability (0.35). According to Eshelman scheme [15], the cataclysmic mutation can also be used to construct families if it is extended to use multiple parents with a given threshold separation instead of simply using just the elite chromosomes. It must be emphasized that this process generates multiple offspring from a single parent by only using a mutation operator. If no offspring are inserted into the new population at the next generation, then the threshold is reduced by one. In other words, In order to avoid very slow convergence, threshold will be also decremented by one, when no improvement is achieved respect to the best chromosome of the previous generation. On the other hand, whenever the population converges towards a certain points, a cataclysm occurs (If the threshold¡0).

(e) restart (can be included in cataclysmic mutation):
The new population includes one copy of the best individuals, while the rest of the population is generated by mutating some percentage of genes of such best individuals. In other way, the elite chromosomes are used as a template to re-seed the population. Randomly, the rate changing of bits is 35 in the template chromosome to form each of the other chromosomes in the population. The Euclidean threshold is reset and the algorithm resumes in the usual manner.

2.3.3 Classification

Fuzzy artmap neural network (FAMNN) with receiver operating characteristics (ROC), FAMNN for training and testing, and ROC for evaluate the performance of FAMNN. FAMNN is one of the incremental learning algorithms are presented by Carpenter et al [19, 20, 21], in response to stability-plasticity dilemma (the catastrophic forgetting phenomenon through neural network learning). This technique is characterized by the following:

1. The FAMNN (nonlinear separability): able to build decision boundaries that separate classes of any shape and size.

2. The FAMNN (overlapping classes): creates decision boundaries to minimize the misclassification for all overlapping classes. In other words, there is no overlap between hyperboxes of different classes.

3. The FAMNN (training time): needs only one pass to learn and refine its decision boundaries.

3 FAMNN and MLPNNs Evaluation Results

In this work, the FAMNN and MLPNNs performance are evaluated by fitness (an accuracy or error rate) of ICHCGA and AUC of ROC.

1. The discernment results between two classes using FAMNN and MLPNNs shown as follows:

   (a) For normal or abnormal see(Table I, Table II) and $(Figure1)$. And the AUC of ROC in the best population using FAMNN is higher than MLPNNs.

ISBN: 978-1-61804-256-9

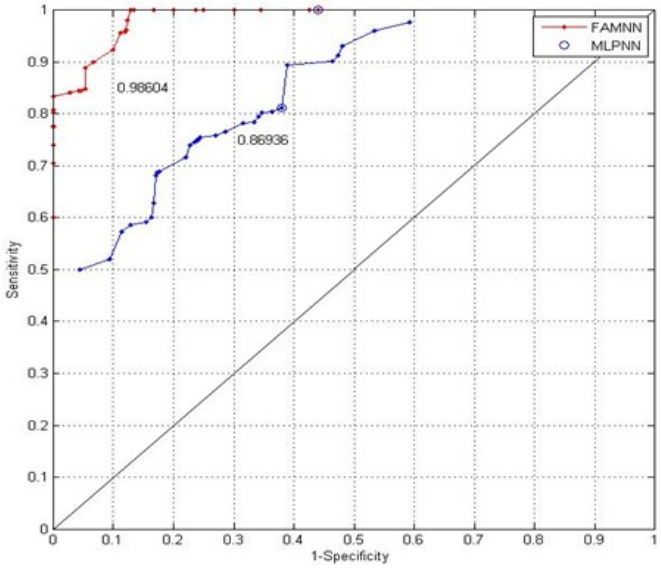


Figure 1: : ROC curves for normal and abnormal tissues for comparison between FAMNN- ICHCGA and MLPNNs performance.

(c) For benign or malignant see(Table I, Table II) and $(Figure2)$. And the AUC of ROC in the best population using FAMNN is higher than MLPNNs.

2. The discernment results between $multi-class$ using FAMNN and MLPNNs shown as follows:

(a) For normal or benign or malignant see(Table I,Table II) and (Figure 3).And the AUC of ROC in the best population using FAMNN is higher than MLPNNs.

Finally, we note that best results are at using $GA-FAMNN$ for stepwise GA-MLPNNs see Table II ,Table II .

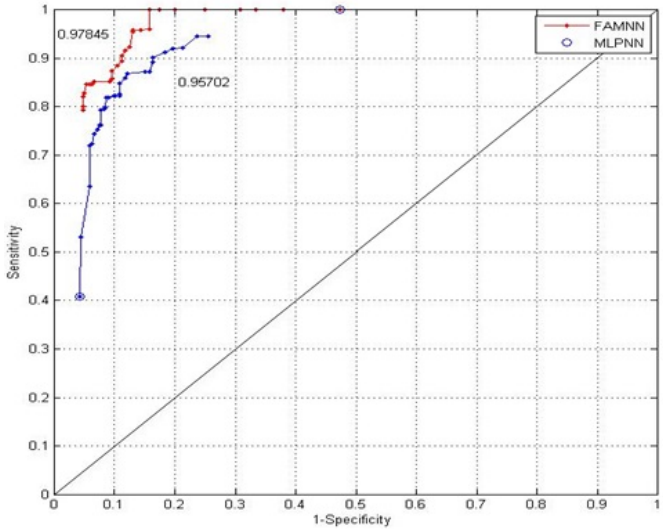


Figure 2: : ROC curves for benign and malignant tissues for comparison between FAMNN- ICHCGA and MLPNNs performance.

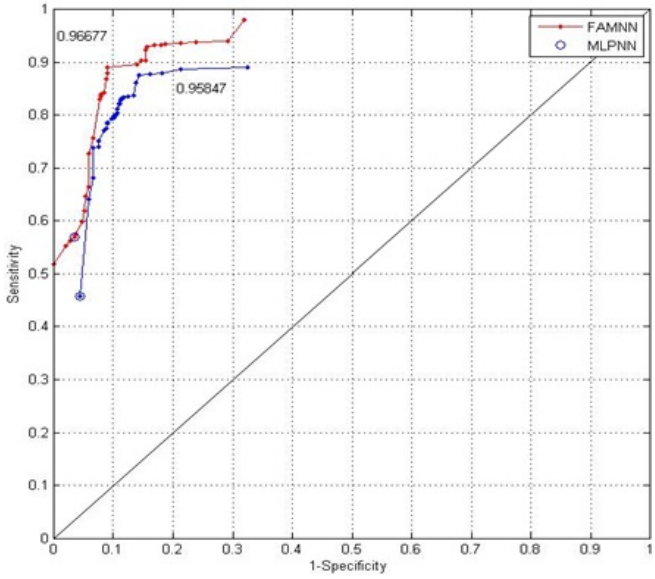


Figure 3: : ROC curves for normal, benign and malignant tissues for comparison between FAMNN-ICHCGA and MLPNNs performance.

4 Conclusions and Future Work

The main goal of this thesis has addressed the investigation of computational intelligence techniques and their applications especially in medical images understanding. Using MIAS dataset, 200 mammograms are used for mass segmentation and classification, 96 mammograms have a normal case with all tissue types (fatty, glandular, density), 53 mammograms have a benign mass with all tissue types and

ISBN: 978-1-61804-256-9

Table 1: Input and result of breast cancer classification by FAMNNs and evaluated it performance using ICHCGA and AUC of ROC curve.

Classification according to breast cancer tissue	Training set	Tesing set	Fitness	Az ROC	No. features selection	Generation
Normal abnormal	50-50	46-44	1.0000	0.98604	6	240
Benign-malignant	28-21	25-20	0.9800	0.97845	6	270
Normal-benign-malignant	50-28-21	46-25-20	0.9889	0.96677	6	290

Table 2: Input and result of breast cancer classification by MLPNNs and evaluated it performance using ICHCGA and AUC of ROC curve.

Classification according to breast cancer tissue	Training set	Tesing set	Fitness	Az ROC	No. features selection	Generation
Normal abnormal	50-50	46-44	0.9578	0.86936	6	300
Benign-malignant	28-21	25-20	0.9200	0.95702	6	300
Normal-benign-malignant	50-28-21	46-25-20	0.9100	0.95847	6	300

41 mammograms have a malignant mass with all tissue types. Other 18 mammograms are used for microcalcification detection, 11 mammograms have a benign case from all tissue types (fatty, glandular, density) and 7 mammograms have a malignant case from all tissue types. The potential of a novel segmentation technique based on Fuzzy-PCNNs is investigated. This computational intelligence model is unsupervised, context sensitive, robotically and invariant to a tissue type. Therefore, Fuzzy-PCNNs own rather interesting properties for the automatic processing of most applications. The Fuzzy-PCNNs approach aiming at separate the ROIs in the image (high spots) based on fuzzy membership degree of coefficient between pixel and it neighbors and fuzzy membership degree of different between them by maximum fuzzy entropy (soft thresholding) rather than segment the ROIs based on initial value (hard thresholding). For feature extraction 188 features are used, whether statistical or geometric, as well as Wavelet technique is used in order to deal with the ROI with multi-scale. Also for feature selection, the ICHCGA is used to select the best available features. For discernment between normal and abnormal, 50 rows from normal data and 50 rows from abnormal data are used as training set. And 46 rows from normal data and 44 rows from abnormal data are used as testing set. The best results when FAMNN is used compare with MLPNNs are as follows: error rate = 0.0000, AUC of ROC =0.98604, features selection number=6. For discernment between benign and malignant, 28 rows from benign data and 21 rows from malignant data are used as training set. And 25 rows from benign data and 20 rows from malignant data are used as testing set. The best results when FAMNN is used compare with MLPNNs are as follows: error rate = 0.0200, AUC of ROC =0.97845, features selection number=6. For discernment between normal, benign and malignant, 50 rows from normal data, 28 rows from benign data and 21 rows from malignant data are used as training set. And 46 rows from normal data, 25 rows from benign data and 20 rows from malignant data are used as testing set. The best results when FAMNN is used compare with MLPNNs are as follows: error rate = 0.0111, AUC of ROC =0.96677, a features selection number=6.

- Future Work : The work in Future will be to develop a medical images understanding for diseases prognosis. This model will use to help discover possible cancers before its occurring in the future based on a time series of mammogram images for women, who come early to screen up.With other view, a Fuzzy-

PCNNs model can be used to develop other applications such as automatic change detection in very high resolution images (satellite images analysis).

References:

[1] Arnau Oliver, Jordi Freixenet, Joan Marti, Elsa Perez, Josep Pont, Erika R.E. Denton,Reyer Zwiggelaar," A review of automatic mass detection and segmentation in mammographic images", *Medical Image Analysis.* 14, 2010, pp. 87–110.

[2] K. S. Fu and J. K. Mui., *A survey on image segmentation,* Pattern Recognition,13:3–16, 1981–1986

[3] Radhika Sivaramakrishna, Nancy A. Obuchowski, William A. Chilcote, Kimerly A. Powell.," Automatic Segmentation of Mammographic Density", *academic radiology.* Volume 8, Issue 3, Pages 250–256 (March 2001).

[4] Mehmet Sezgin.,Survey over image thresholding techniques and quantitative performance evaluation, *Journal of Electronic Imaging.* 13(1), 146–165 (January 2004).

[5] S. Timp and N. Karssemeijer.,"A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography", *IEEE Transactions on Medical Imaging.* 31(5):958–971, 2004.

[6] Lcio F.A. Campos, Aristfanes C. Silva, and Allan Kardec Barros.,"Diagnosis of Breast Cancer in Digital Mammograms Using Independent Component Analysis and Neural Networks", *CIARP 2005.* LNCS 3773, pp. 460 –469, 2005.

[7] Retico, P. Delogu, M.E. Fantacci, P. Kasaec.An automatic system to discriminate malignant from benign massive lesions on mammograms,*Nuclear Instruments and Methods in Physics Research* A 569 (2006) 596–600.

[8] Pasquale Delogu, Maria Evelina Fantacci, Parnian Kasae, Alessandra Retico,Characterization of mammographic masses using a gradient-based segmentation algorithm and a neural classifier,*Computers in Biology and Medicin* e 37 (2007) 1479 – 1491.

[9] H. D. Cheng, Yen-Hung Chen, Fuzzy partition of two-dimensional histogram and its application to thresholding,*Pattern Recognition* , Volume 32, Issue 5, May 1999.

[10] R.L. Kirby and A. Rosenfeld, A note on the use of (gray-level, local average gray-level) space as an aid in threshold selection,*IEEE Trans. Syst. Man Cybernet.* SMC-9 12 (1979), pp. 860–866.

[11] Lindblad, Th.; and Kinser, J.M. (1998). Image Processing using Pulse- Coupled Neural Networks,*Perspectives In Neural Computing. Springer-Verlag Limited.* ISBN 3-540-76264.

[12] Zadeh, L.A. Outline of a New Approach to the Analysis of Complex Systems and Decision Processes ,*Information science* , Vol.9, pp.43–80. (1973).

[13] B. Riecan, D. Markechova, The entropy of fuzzy dynamical systems, general scheme and generators,*Fuzzy Sets and Systems,* Volume 96, Issue 2, 1 June 1998, Pages 191–199.

[14] H. D. Cheng, Jim-Rong Chen, Automatically determine the membership function based on the maximum entropy principle,*Information Sciences,* Volume 96, Issues 3-4, February 1997, Pages 163-182.

[15] L. Eshelman, The CHC Adaptive Search Algorithm, How to Have Safe Search When Engaging in Non-traditional Genetic Recombination,*Morgan Kaufman, S.* 265 283-1991.

[16] Rechenberg, Evolutions strategies, *From mann-Holzboog* , 1973.

[17] H. Schwefel, Numerical optimization of computer models (M. Finnis, trans.), *Chichester: John Wiley* , 1981 (Original work published 1977).

[18] Larry J. Eshelman, James D. Schaffer, Method for optimizing the configuration of a pick and place machine, *United States Patent* No. 5,390,283, 14 February 1995.

[19] R. Polikar, L. Udpa, S. Udpa, V. Honavar, Learn++: An incremental learning algorithm for multilayer perceptrons. *Proceedings of 25th. IEEE International Conference on Acoustics, Speech and Signal Processin* , Vol. 6, pp: 3414-3417, Istanbul, Turkey,2000.

[20] R. Polikar, L. Udpa, S. Udpa, V. Honavar. Learn++: An incremental learning algorithm for supervised neural networks. *IEEE Transactions on Systems, Man, and Cybernetics.Part C: Applications and Review* , Vol. 31, No. 4, pp: 497-508, 2001.

[21] S. Grossberg, Adaptive pattern recognition and universal encoding II:Feedback, expectation, olfaction, and illusions, *Biol. Cybern.* , vol. 23,pp. 187202, 1976.

Computational technique for optimization of the process parameter for cold spray coating of titanium

Damilola I. Adebiyi, Ione Botef, and Patricia A. Popoola

***Abstract*—**Cold spray coating is a solid-state coating process that uses a high-speed gas jet to accelerate powder particles toward a substrate causing plastic deformation and consolidation of the particles upon impact. The process involves many parameters, thus making it very complex, and highly dependent and sensitive to small changes in process parameters. This results in a small operational window of these parameters. Consequently, optimization of the process parameters by means of detailed mathematical study of the process is a key to improve the coating quality and reduce the spraying costs. In this study, a mathematical model is employed to optimize the outlet gas velocity, applied gas pressure and deposition temperature of the powder particles at the exit of the nozzle. One important implication of this is that the deposition characteristic of the coating can be analyzed and the possibility of the particles achieving the much desired critical velocity can be established

***Keywords*—**Cold spray, critical velocity, mathematical model, process parameters.

.

I. Introduction

COLD spray (CS) is a material deposition process in which relatively small particles (ranging in size from approximately 5 μm to 100 μm in diameter) in solid state are accelerated to a critical high velocity (typically 300-1400 m/s), and are subsequently plastically deformed to develop a deposit on a metallic substrate [1, 2]. CS is a relatively recent spray technology and there are different approaches known by different names such as: Cold Gas Dynamic Spraying, Kinetic Spraying, High Velocity Particle Consolidation (HVPC), High Velocity Powder Deposition and Supersonic Particle/Powder Deposition (SPD) [3]. In cold spray, the powders do not melt before impacting the substrate making the process commendable for different applications which involves various materials such as metals, polymers, composites, etc. Attachment of powder to substrate otherwise known as bonding is achieved by the kinetic energy of the powder particles rather than the thermal energy as the case in most of the thermal spray processes [4]. Bonding takes place when the velocity of the powder particles exceeds a certain value called the critical velocity (CV). Hence, the CV is defined as the velocity the particle must attain before deposition can take place after impacting the substrate [5]. Many process parameters affect the CV [6]. Thus critical velocity is a major parameter in cold spray process [7]. This is because CV determines which of particle deposition or substrate erosion will occur upon the impact of spray particles [6]. Typically, the CV is the velocity at which the transition from erosion of the substrate to deposition of the particle takes place [6]. The critical velocity depends on the type of spray material, the powder quality, the particle size and the particle impact temperature. Below the critical velocity, plastic deformation is too low to cause bonding, above the critical velocity, hydrodynamic penetration leads to strong erosion. Therefore, the optimum conditions for deposition lie between these two characteristic velocities [8]. According to Assadi [9], the value of CV is determined by the temperature, thermo-mechanical properties of the sprayed material [10] and the characteristics of the substrate [11-13]. In this work, an attempt is made to use numerical model to optimize the velocity, temperature and pressure of the particles exiting the de lava nozzle used in the cold spray process.

II. Mathematical Modelling Procedure

Cold spray is carried out in a de Laval nozzle called cold gas-dynamic spray (CGDS) system. Currently, two variants of the commercially CGDS system are available. These are the Low Pressure Cold Gas-Dynamic Spray (LPCGDS) system and the High Pressure Cold Gas-Dynamic Spray (HPCGDS) system.

This material is based upon work supported financially by the National Research Foundation. The cold Spray Laboratory of the University of Witwatersrand, Johannesburg is appreciated for Cold Spray facilities, The authors also acknowledge the support from Tshwane University of Technology Pretoria, South Africa which helped to accomplish this work

D. I. Adebiyi is a doctoral candidate of the Department of Chemical, Metallurgical and Materials Engineering, The Tshwane University of Technology, Pretoria, South Africa (e-mail: AdebiyiDI@tut,ac,za

I. Botef is with the University of Witwatersrand, School of Mechanical, Industrial and Aeronautical Engineering, Johannesburg, South Africa (email: Ionel.botef@wits.ac.za)

A. P. I. Popoola is with the Department of Chemical, Metallurgical and Materials Engineering, The Tshwane University of Technology, Pretoria, South Africa (e-mail: PopoolaAPI@tut,ac,za)

The basic principle of the cold spray (CS) involves a high-velocity flow of the particles made possible by high-pressure and high-velocity of the carrier gas. The high pressure jet is preheated to compensate for the adiabatic cooling due to expansion. The powder particles are transported by the energy of the preheated, high-pressure, high velocity supersonic gas jet.

A. Modeling Assumptions

In formulating this model, it is assumed that:

1) Ideal gas law is obeyed by the gas.
2) The gas flow is one-dimensional, frictionless and adiabatic.
3) Steady-state conditions exist.
4) Gas expansion is uniform; no shocks or discontinuities.
5) Particles effect on gas conditions is negligible.
6) Inter-particle collision is negligible.
7) Particles effect on space charge is negligible

B. Model equations

According to Papyrin [14], Lee et al. [15] and Janzhong et al. [16], the flow through the nozzle in the LPGDS process is governed by the continuity equation, momentum equation and the energy equation. A constitutive equation is however required in order to close the system.

C. The continuity Equation

A continuity equation is an equation that describes the transport of a conserved quantity. According to the continuity equation (otherwise known as the law of conservation of mass), the rate at which mass enters a system is equal to the rate at which mass leaves the system in any steady state process, i. e. the total time rate of change of mass in a fixed region is zero. Therefore, the mass, energy and momentum of the powder in the gas stream are conserved. The continuity equation for the LPCGDS can be written as:

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho v_i)}{\partial x_i} = 0 \qquad (1)$$

Where ρ is the density of the gas and v is the velocity

D The momentum Equation

In the LPCGDS, the principle of conservation of translational momentum also applies. The principle states that the total momentum of a system of colliding objects remains constant provided no resultant external force acts on the system. In other words, when external forces are acting on them, the time rate of change of the momentum is equal to the net force acting on the particle. Taking internal stress and the gravitational acceleration into account, the application of principle of conservation of translational momentum for the LPCGDS can be written as:

$$\frac{\partial \rho v_i}{\partial t} + \frac{\partial (\rho v_j v_i)}{\partial x_j} = \frac{\partial \tau_{ij}}{\partial x_j} - \frac{\partial p}{\partial x_i} \qquad (2)$$

τ is the internal stress and the g is the gravitational acceleration

E The Energy Equation

The kinetic energy of the particles on impact is important for plastic deformation of the particles to take place and form splats, which bond together to produce coatings. The energy equation is given in equation (3)

$$\frac{\partial (\rho E)}{\partial t} + \frac{\partial (\rho v_j E)}{\partial x_j} = \frac{\partial}{\partial x_j}\left(k \frac{\partial T}{\partial x_j} \right) + \frac{\partial}{\partial x_j}(\tau_{ij} v_i) \qquad (3)$$

F The Constitutive Equation

Constitutive equations relate thermo-mechanical parameters, i.e. strain (ε), strain rate ($\dot{\varepsilon}$) and temperature (T), with flow stress (σ). Although the conserved quantitative (mass, momentum and energy) are the basic quantities describing the flow through the LPCGDS) system, in order to close the system, a constitutive equation is required for stress and flow/flux, otherwise stress and flow must be added to the list of variable. Equations (1)-(3) are solved in conjunction with an appropriate equation of state and the constitutive equation. The stress equation for Newtonian fluid is given by:

$$\tau_{ij} = \mu \left[\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) - \frac{2 \partial v_k}{3 \partial x_k} \right] \qquad (4)$$

Where $i, j = (x, y, z)$

The total stress tensor σ_{ij} in the fluid is given by the sum of internal stresses due to the fluid pressure p and the stress due to viscous forces as shown in equation (5)

$$\sigma_{ij} = -p\delta_{ij} + \tau_{ij} \qquad (5)$$

Where δ_{ij} is the *Kronecker delta*, defined such that $\delta_{ij} = 1$ if $i = j$, otherwise $\delta_{ij} = 0$

G Discretization of the Equations

The discrete approximation to the momentum, energy and continuity equations can be written in a form shown in equations (6) and (7):

$$\boldsymbol{M}\dot{\boldsymbol{u}} + \boldsymbol{A}(\boldsymbol{u})\boldsymbol{u} + \boldsymbol{G}\boldsymbol{P} = \boldsymbol{K}\boldsymbol{u} + \boldsymbol{f}(\boldsymbol{t}) \qquad (6)$$

$$\boldsymbol{D}\boldsymbol{u} = \boldsymbol{g}(\boldsymbol{t}) \qquad (7)$$

M – Mass matrix, (for equidistant discretization of the unit matrix), A – advection matrix, G – gradient matrix, K – diffusion matrix, D – divergence matrix, f (t) and g(t) represents the effects of the boundary conditions on velocity. Using the following notations and equalities:

$$b \equiv Ku + f(t) - A(u)u$$

$$D = G^T$$

And denoting G with C, the following is obtained:

$$\begin{bmatrix} M & C \\ C^T & 0 \end{bmatrix} \begin{bmatrix} P \\ \dot{u} \end{bmatrix} = \begin{bmatrix} b \\ \dot{g} \end{bmatrix}$$

Consequently, the discrete system for the constitutive equation is obtained as shown in equation (8) while that for the

momentum, energy and continuity equations are obtained as shown in equation (9)[4]

$$\begin{bmatrix} N & 0 & 0 \\ 0 & N & 0 \\ 0 & 0 & N \end{bmatrix}\begin{bmatrix} \tau_{11} \\ \tau_{22} \\ \tau_{12} \end{bmatrix} - \begin{bmatrix} 2L_1 & 0 & 0 \\ 0 & 2L_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} Vi \\ V_j \\ P \end{bmatrix} + \begin{bmatrix} D_1 & 0 & 0 \\ 0 & D_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} Vi \\ Vj \\ P \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad (8)$$

$$\begin{bmatrix} c_i u_i & 0 & 0 \\ 0 & c_i u_i & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} V_i \\ V_j \\ P \end{bmatrix} + \begin{bmatrix} 2K_{11}+2K_{22} & K_{21} & Q_1 \\ K_{22} & 2K_{11}+2K_{22} & Q_2 \\ Q_1^T & Q_2^T & 0 \end{bmatrix}\begin{bmatrix} V_i \\ V_j \\ P \end{bmatrix} = \begin{bmatrix} F_1 \\ F_2 \\ 0 \end{bmatrix} \quad (9)$$

Using the equation (8) and (9), the finite element equation for the flow process can be written by defining the coefficient matrix as given below:

$$K_{11} = \left(2S^{11} + S^{22}\right)\left(\mu + \mu_1\right)$$

$$K_{22} = \left(S^{11} + 2S^{22}\right)\left(\mu + \mu_1\right)$$

$$K_{21} = \left(K_{12}\right)^T$$

$$C_i U_i = \int \alpha \rho_g V_j . \left(\frac{\partial \alpha^T}{\partial x_j}\right) dx$$

$$Q = \int \gamma \frac{\partial \alpha^T}{\partial x_j} . dx$$

$$S_{ij}^{11} = \int \left[\frac{\partial \alpha}{\partial x_j} . (\mu + \mu_i) \frac{\partial x^T}{\partial X_i} . dx \right]$$

$$L_1 = \int \left[\psi \left[(\mu + \mu t) \frac{\partial \alpha^T}{\partial X_j} \right] . dx \right]$$

$$D_1 = \int \psi \frac{2}{3} \mu_i \frac{\partial \alpha^T}{\partial X_i} . \delta i_j . dx$$

Thus, the finite element equations (8) and (9) become:

$$C_{(u)} + KU = F \quad (10)$$

$$N_\tau - LU + DU = 0 \quad (11)$$

Equations (10) and (11) are respectively the typical form of the Newtonian equation and the extra stress finite element analogue of the constitutive equation. These equations were solved by substituting the boundary conditions, and were used to calculate the exit velocity, pressure and temperature of the nozzle by using a CFD software called Solidworks

III. Results and Discussion

The geometry of the LPCGDS used for the modeling was obtained from Goyal et all [4].The velocity, temperature and pressure distribution were analyzed and calculated using the meshing tool of Solidworks and substituting the thermo-physical properties of the carrier gas (air) and the metal powder, and the boundary conditions.

A Thermo-Physical Properties and Boundary Conditions

The boundary conditions are determined by the properties of the carrier gas and the powder used. The carrier gas is air and the powder is irregular shaped titanium powder with -325 mesh particle size and metal base purity of over 99.5%. The properties of the air and titanium are given in Table 1

Table i: Thermo-physical properties of the carrier gas (air) and the titanium powder

Carrier gas (air)		Titanium powder	
Density,	1.205 kg/m^3	Density	4.5 kg/m^3
Specific heat capacity, *Cp*	1.005 J/kg K	Heat capacity, *C*	523 J kg^{-1} K^{-1}
Thermal conductivity, *h*	0.0257 W/m.K	Thermal Conductivity	6 W m^{-1} K^{-1}
Kinematic viscosity, v	15.11 x 10^{-6} m^2/s	Thermal conductivity,	27.5 W/m-K
Expansion coefficient	3.43 x 10^{-3} 1/K	Elastic modulus, *E*	10.3x10^4 MPa
Prandtl's number- P_r -	0.713	Poisson ratio, μ	0.32
		Melting Point	1933 K
		Fusion Heat:	18.8 kJ/mol

B Boundary Conditions

The air into the nozzle is at temperature 723 K and pressure of 1 M Pa. The titanium powder is at room temperature at inlet and its mass flow rate at room temperature is 10 g/min. At the outlet, the pressure of the nozzle, $u = v = w = 0$, T = Constant (room temperature).

C Distribution of Velocity, Pressure and Temperature in the Nozzle

The distribution of the velocity, temperature and pressure in the nozzle is shown in Fig 1. As shown in the Fig, the

ISBN: 978-1-61804-256-9

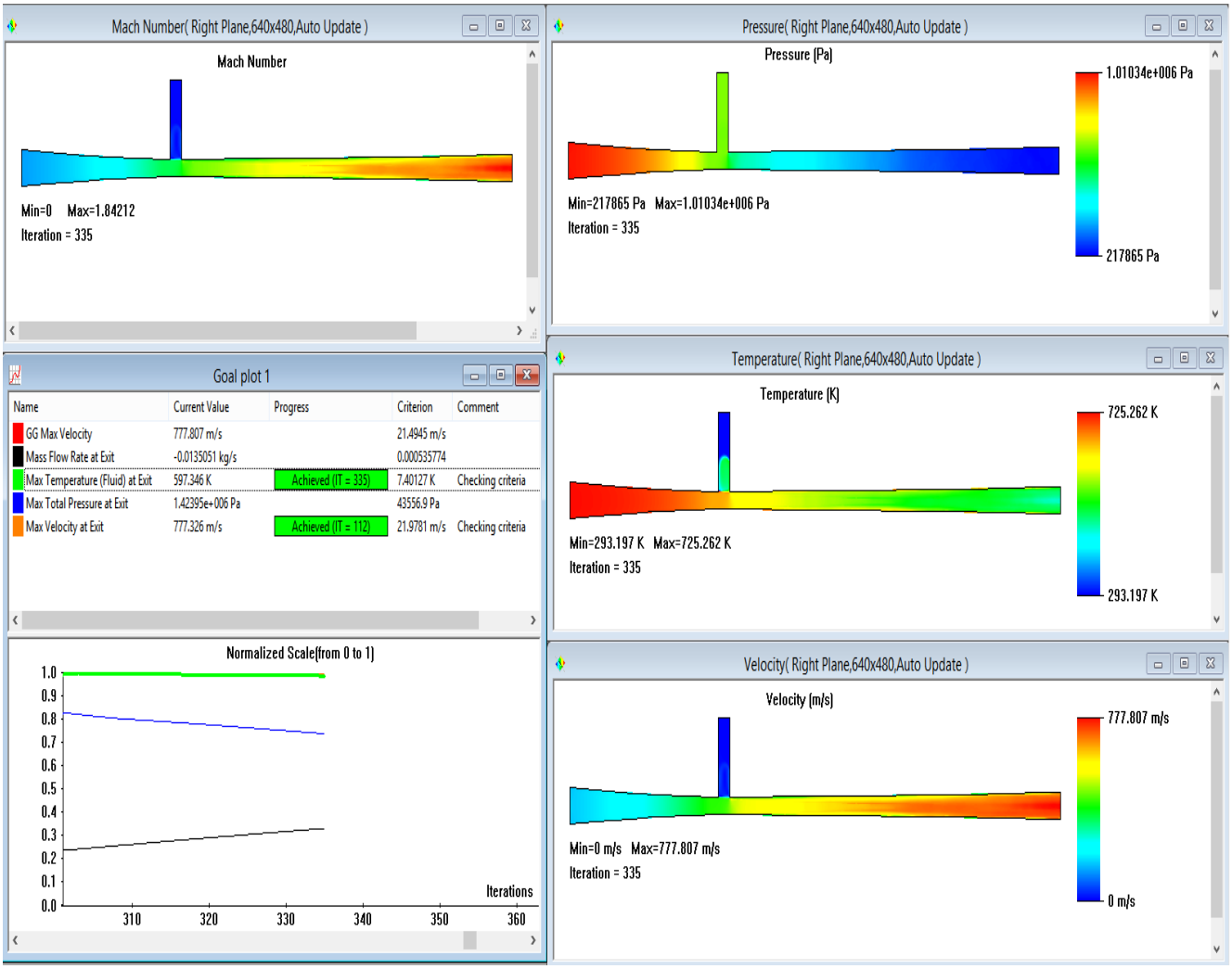


Figure 1: The distribution of the velocity, temperature and pressure through the nozzle

outlet velocity is about 777.326 ms^{-1} whereas the outlet pressure is about 1.42595 bars. The temperature distribution through the LPGDS nozzle as indicated by the contours of the temperature shows that the outlet temperature is less than the inlet temperature, and the temperature is maximum at inlet. This is because the powder was at room temperature when it enters the gas stream. Although the maximum temperature is 725.262 K, the exit temperature obtained is 597.346 K (324 ^{O}C)

C Validation

In order to validate the computational model, the values obtained were compared to previously published experimental results. Triantou et all. [2] obtained titanium coatings on titanium alloy (Ti-6Al-4V) at a gas temperature 550 ^{o}C and pressure of 3.0 MPa. Lima et all [17], deposited titanium particles on aluminium by cold spray. According to the authors, the particles achieved critical velocity at temperature between 370 to 480 ◦C. The results of these published works agree with that obtained from mathematical model.

IV CONCLUSIONS AND FUTURE WORK

The continuity momentum and energy equations have been solved by transforming the partial differential equations into a single ordinary differential equation; and introducing a constitutive equation to close the system and also account for stress and flow viscosity. An estimate of velocity, pressure and temperature at the exit of the nozzle has been provided by this model. The values obtained in the estimate compare favourably with those of the open literature. The values of the estimate will however be validated using experimental techniques. This indicates that the cold spray coating process can be successfully simulated, and that the pressure, velocity and temperature can be predicted.

APPENDIX

ρ_g gas (air) density
δ_g *Kronecker delta*, which is a component of the identity tensor defined such that δ_{ij}=1if *i*=*j*, otherwise δ_{ij} =0
ν velocity vector
τ Stress tensor
μ molecular viscosity
ω weight function
χ, α, ψ vector of interpolation function

ISBN: 978-1-61804-256-9

ACKNOWLEDGMENT

This material is based upon work supported financially by the National Research Foundation. The Cold Spray Laboratory of the University of Witwatersrand, Johannesburg is appreciated for Cold Spray facilities; the authors also acknowledge the support from Tshwane University of Technology, Pretoria, South Africa, which helped to accomplish this work

REFERENCES

[1] T. Stoltenhoff, H. Kreye, and H. J. Richter, "An Analysis of the Cold Spray Process and Its Coatings," *J. Ther. Spray Tech*, vol. 11(4), pp. 542—550, 2002

[2] K. I. Triantou, C. I. "Sarafoglou, T. Tsiourva, D. I. Pantelis, D. K. Christoulis, and V. Guipont, "Case studies of cold sprayed coatings. In: 7th International Conference on Coatings in Manufacturing Engineering, Chalkidiki, Greece. (2008)

[3] H. Singh, T. S. Sidhu, and S. B. S. Kalsi, "Cold spray technology: future of coating deposition processes," *Frattura ed Integrità Strutturale*, vol. 22, pp. 69-84, 2012

[4] T. Goyal, S. Prince, R. S. Walia, and T. S Sidhu, "Effect of nozzle geometry on exit velocity, temperature and pressure for cold spray process," *Int. J. Mat. Sci. Eng*., Vol. 2, pp. 65-72, 2011

[5] V. Champagne, "The cold spray materials deposition process, Fundamentals and Application, Woodhead, (2007)

[6] R. Ghelichi, S. Bagherifard, M. Guagliano, and M. Verani, "Numerical simulation of cold spray coating," *Surf. & Coat. Tech*. vol. 205, pp. 5294–5301, 2011

[7] A. Papyrin, "The development of the cold spray process," *Cold Spray Technol*., (CST), USA 2006

[8] T. Schmidt, F. Gaurtner, H. Assadi, H. Kreye, and A. Materialia, "Development of a generalized parameter window for cold spray deposition," *Acta Materialia, vol.* 54(3) pp. 729–742, 2006

[9] H. Assadi, F. Gaurtner, T. Stoltenho, and H. Kreye, "Bonding mechanism in cold gas spraying," *Acta Materialia*, vol. 51, pp. 4379-4394, 2003

[10] A. P. Alkimov, V. E. Kosarev, and A. N. Papyrin, "A method of cold gas-dynamic deposition," *Dokl. Akad. Nauk, SSSR* vol. 318, pp. 1062–1065, 1990

[11] J. G. Legoux, E. Irissou, and C. Moreau, "Effect of substrate temperature on the formation mechanism of cold-sprayed aluminum, zinc and tin coatings," *J. Therm. Spray Tech*. vol. 16 (5–6), pp. 619-626, 2007

[12] P. Gao, C. Li, C. Yang, Y. Li, and C. Li, "Influence of substrate hardness on deposition behavior of single porous WC-12Co particle in cold spraying" *Sur & Coatings Tech.* vol. 203, pp. 384 –390, 2008

[13] T. Schmidt, F. Gärtner, H. Assadi, and H. Kreye, "Development of a generalized parameter window for cold spray deposition" *Acta Mater.* Vol. 54(3), pp. 729–742, 2006

[14] A. N. Papyrin, "Cold spray: State of the art and applications; Cold spray technology. Albuquerque, NM, USA, pp. 1-21, 2006

[15] J. C. Lee, H. G. Kang, W. S. Chu, and W. S. Ahn, "Repair of damaged mold surface by cold-spray method," *CIRP Annals – Man. Tech*, vol. 56, pp. 577—580, 2007

[16] L. Janzhong, and C. Guobang, "Fluid Mechanics. Beijing: Tsinghua University Press, 72-91 2005

[17] R. S. Lima, A. Kucuk, and C. C. Berndt, "Deposition efficiency, mechanical properties and coating roughness in cold-sprayed titanium," *J. Mat. Sci. Let.* Vol. 21, pp. 1687 – 1689, 2002

Authors Index

ISBN: 978-1-61804-256-9