

MATHMODEL'17

INTERNATIONAL SCIENTIFIC CONFERENCE
13 - 16. DECEMBER, 2017, BOROVETS, BULGARIA

MATHEMATICAL MODELING

**TECHNOLOGICAL AND
SOCIO-ECONOMIC PROCESSES**

PROCEEDINGS

YEAR I ISSUE 1/2017
ISSN (Print) 2535-0978
ISSN (Online) 2603-3003

Published by
SCIENTIFIC-TECHNICAL UNION of MECHANICAL ENGINEERING - INDUSTRY 4.0
Sofia, BULGARIA

INTERNATIONAL SCIENTIFIC CONFERENCE

MATHEMATICAL MODELING

Year I

Volume 1/1

DECEMBER 2017

ISSN 2535-0978 (Print)
ISSN 2603-3003 (Online)

PROCEEDINGS

THEMATIC FIELDS

1. THEORETICAL FOUNDATIONS AND SPECIFICITY OF MATHEMATICAL MODELLING
3. MATHEMATICAL MODELLING OF SOCIO-ECONOMIC PROCESSES AND SYSTEMS
2. MATHEMATICAL MODELLING OF TECHNOLOGICAL PROCESSES AND SYSTEMS
4. MATHEMATICAL MODELLING OF MEDICAL-BIOLOGICAL PROCESSES AND SYSTEMS

**13 – 16 DECEMBER, 2017,
BOROVETS, BULGARIA**

PUBLISHER:

**SCIENTIFIC TECHNICAL UNION OF MECHANICAL
ENGINEERING “INDUSTRY-4.0”**

108, Rakovski Str., 1000 Sofia, Bulgaria

tel. (+359 2) 987 72 90,

tel./fax (+359 2) 986 22 40,

office@mathmodel.eu

www.mathmodel.eu

INTERNATIONAL EDITORIAL BOARD

Chairman:		
Prof. ANDREY FIRSOV Peter the Great St.Petersburg Polytechnic University		RU
Members:		
Abilmazhin Adamov, Prof.	L.N.Gumilyov Eurasian National University	KZ
Alexander Guts, Prof.	Omsk State University	RU
Alexei Zhabko, Prof.	Saint Petersburg State University	RU
Andrey Markov, Prof.	Baltic State Technical University	RU
Andrii Matviichuk, Prof.	Kyiv National Economics University	UA
Andrzej Nowakowski, Prof.	University of Lodz	PL
Anton Makarov, Dr.	Saint Petersburg State University	RU
Armands Gricans, Assoc. Prof.	Daugavpils University	LV
Artūras Dubickas, Prof.	Vilnius University	LT
Avinir Makarov, Prof.	Saint Petersburg State University of Industrial Technologies and Design	RU
Christo Boyadjiev, Prof.	Institute of Chemical Engineering, BAS	BG
Daniela Marinova, Dssoc. Prof.	Technical University of Sofia	BG
Dimitrios Poulakis, Prof.	Aristotle University of Thessaloniki	GR
Evgeniy Smirnov, Assoc. Prof.	Volgograd State Technical University	RU
Giovanni Borgioli, Assoc. Prof.	University of Florence	IT
Haskiz Coskun, Prof.	Karadeniz Technical University of Trabzon	TR
Idilia Bachkova, Prof.	University of Chemical Technology and Metallurgy	BG
Irena Stojkovska, Prof.	Ss. Cyril and Methodius University in Skopje	MK
Ivana Štajner-Papuga, Prof.	University of Novi Sad	RS
Kanagat Aldazharov, Assoc. Prof.	Kazakh Economics University	KZ
Karl Kunisch, Prof.	University of Graz	AT
Mahomed Agamirza ogly Dunyamalyev, Prof.	Azerbaijan Technical University	AZ
Marius Giuclea, Prof.	The Bucharest University of Economics Studies	RO
Mihail Okrepilov, Prof.	D.I. Mendeleyev Institute for Metrology (VNIIM)	RU
Milena Racheva, Assoc. Prof.	Technical University of Gabrovo	BG
Mohamed Kara, Dr.	Ferhat Abbas Sétif 1 University	DZ
Mohamed Taher El-mayah, Prof	MTI University	EG
Neli Dimitrova, Prof.	Institute of Mathematics and Informatics, BAS	BG
Nina Bijedic, Prof.	Dzemat Bijedic University of Mostar	BA
Oleg Obradović, Prof.	University of Montenegro	ME
Olga Pritomanova, Assoc. Prof.	Oles Honchar Dnipropetrovsk National University	UA
Özkan Öcalan, Prof.	Akdeniz University of Antalya	TR
Pașc Găvrută, Prof.	Politehnic University of Timisoara	RO
Pavel Satrapa, Assoc. Prof.	Technical University of Liberec	CZ
Pavel Tvrdík, Prof.	Czech Technical University in Prague	CZ
Pavlina Yordanova, Assoc. Prof.	Shumen University	BG
Petr Trusov, Prof.	Perm State Technical University	RU
Rannveig Björnsdóttir, Prof.	University of Akureyri	IS
Roumen Anguelov, Prof.	University of Pretoria	ZA
Sándor Szabó, Dr. Prof.	University of Pécs	HU
Sashko Martinovski, Assoc. Prof.	St. Kliment Ohridski University of Bitola	MK
Sergey Bosnyakov, Prof.	Moscow Institute of Physics and Technology	RU
Sergey Kshevetskii, Prof.	Immanuel Kant Baltic Federal University	RU
Snejana Hristova, Prof.	University of Plovdiv	BG
Svetlana Lebed, Assoc. Prof.	Brest State Technical University	BY
Tomasz Szarek, Prof.	University of Gdansk	PL
Valeriy Serov, Prof.	University of Oulu	FI
Vasily Maximov, Prof.	Saint Petersburg State University of Industrial Technologies and Design	RU
Ventsi Rumchev, Prof.	Curtin University, Perth	AU
Veronika Stoffová, Prof.	University of Trnava	SK
Veselka Pavlova, Prof.	University of National and World Economy	BG
Viorica Sudacevski, Assoc. Prof.	Technical University of Moldova	MD
Vladimir Janković, Prof.	University of Belgrade	RS
Vladislav Holodnov, Prof.	Saint Petersburg State Institute of Technology	RU
Vyacheslav Demidov, Prof.	Saint Petersburg State University of Industrial Technologies and Design	RU
Yordan Yordanov, Assoc. Prof.	University of Sofia	BG
Yuriy Kuznetsov, Prof.	Nizhny Novgorod State University	RU
Zdenka Kolar - Begović, Prof.	University of Osijek	HR

CONTENTS

THEORETICAL FOUNDATIONS AND SPECIFICITY OF MATHEMATICAL MODELLING

FUNDAMENTALS OF THE KINETIC THEORY OF MULTICOMPONENT EMULSIONS

Prof., Dr. Tech. Sci. Firsov A.N. 5

ABOUT THE PROBLEM OF DATA LOSSES IN REAL-TIME IOT BASED MONITORING SYSTEMS

Prof. Aliksieiev V. PhD., Prof. Gaiduchok O. PhD. 10

COMPARISON OF THE RESULTS OBTAINED BY PSEUDO RANDOM NUMBER GENERATOR BASED ON IRRATIONAL NUMBERS

PhD. Dimitrievska Ristovska V., Prof. PhD. Bakeva V. 12

PARAMETRIC INDUCED INSTABILITIES OF BOSONS IN MAGNETAR'S CRUST

Prof. Dr. Dariescu M.-A., PhD. student Miha D.-A., Prof. Dr. Dariescu C. 16

ON THE USE OF CONFORMING AND NONCONFORMING RECTANGULAR FINITE ELEMENTS FOR EIGENVALUE APPROXIMATIONS

Assoc. Prof Racheva M. Dsc., Prof. Andreev A. Dsc. 20

A MATHEMATICAL MODEL OF VISCOUS LIQUID MIXTURE MOTION THROUGH A VERTICAL CYLINDRICAL PIPE

Asst., MSc Sorokina Natalia 24

MODELING OF LIQUID SPREADING IN RANDOMLY PACKED METAL PALL RINGS

Dr. Petrova T., Stefanova, K., Dr. Dzhonova-Atanasova, D., Prof. Dr. Semkov, K. 26

STATISTICAL METHODS FOR THE ANALYSIS OF THE MONTE CARLO SIMULATION RESULTS IN VISION SYSTEMS

I. Yu. Gendrina, Ass. Prof., Ph.D. 30

ОДНОРАНГОВАЯ АППРОКСИМАЦИЯ ПОЛОЖИТЕЛЬНЫХ МАТРИЦ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ИДЕМПОТЕНТНОЙ МАТЕМАТИКИ

Д-р ф.-м. н. Кривулин Н. К., студент Романова Е. Ю. 33

ИСПОЛЬЗОВАНИЕ РАЗРЕЖЕНИЯ МАТРИЦ ДЛЯ РЕШЕНИЯ МНОГОМЕРНОЙ ЗАДАЧИ ТРОПИЧЕСКОЙ ОПТИМИЗАЦИИ

Д-р ф.-м. н. Кривулин Н. К., аспирант Сорокин В. Н. 36

MATHEMATICAL MODELLING OF TECHNOLOGICAL PROCESSES AND SYSTEMS

ROTATING OF A BALL IN CHAMBER FILLED WITH A FLUID

Prof., Dr. Dementev O. 40

A VARIATIONAL SOLUTION OF THE SCHRÖDINGER EQUATIONS IN AN INHOMOGENEOUS COULOMB FIELD

Freinkman B., Polyakov S., Tolstov I. 44

COMPUTER SYSTEM FOR PREDICTING THE STRUCTURE AND PROPERTIES OF CAST METAL PRODUCTS

Ass. Prof., Dr. Eng. Donii O., Ass. Prof., Dr. Eng. Kulinich A., Ass. Prof., Dr. Eng. Khristenko V. 47

MODELING OF ELECTRONIC STATES OF A SINGLE DONOR IN MIS-STRUCTURE USING THE FINITE DIFFERENCE METHOD

M.Sc. Levchuk E.A., Prof. Lemeshevskii S.V. PhD., Prof. Makarenko L.F. PhD 51

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ МАГНИТНОГО ГИСТЕРЕЗИСА ПРИ ТРЕХОСНОМ НАПРЯЖЕННОМ СОСТОЯНИИ

Mushnikov A.N., PhD. Putilova E.A. 55

MATHEMATICAL MODEL OF SUSTAINABLE INTEGRATED BIOETHANOL SUPPLY CHAINS

Eng. Dzhelil Y., Prof. DSc Ivanov B., Eng. Ganey E., Assoc. Prof. PhD Dobrudzhaliyev D. 59

STABILITY ANALYSIS AND SIMULATIONS OF BIOREACTOR MODEL WITH DELAYED FEEDBACK

Assist. Prof., Ph.D. Borisov M., Prof., Ph.D. Dimitrova N., Prof., D.Sc. Krastanov M.. 65

MODELING AND SIMULATION OF INDUSTRIAL PROCESSES

Christo Boyadjiev 69

APPLICATION OF A HYBRID MODEL: FRACTIONAL EXPERIMENTAL DESIGN + ARTIFICIAL NEURAL NETWORKS+ GREY RELATIONAL ANALYSIS METHOD ON OPTIMIZATION OF PROCESS PARAMETERS OF POWDER METALLURGY	
S. Hartomacıoğlu. PhD., H.O.Gülsoy. PhD., B. Bakırcıoğlu Ms.C., S. Yuksel, Ms.C.	73
SELECTION OF OPTIMAL EXPERIMENTAL CONDITIONS OF TURNING OPERATIONS By USING SATISFACTION FUNCTION AND DISTANCE BASED MULTI-CRITERIA DESISION MAKING METHOD	
B. Bakırcıoğlu Ms.C., S. Hartomacıoğlu. PhD., , S. Yuksel, Ms.C., Ş. Yazman, Ms. C., S. Güvercin, Ms.C., A. Duran, Ph.D.	79
РАЗРАБОТКА ЭФФЕКТИВНОГО И УСТОЙЧИВОГО АЛГОРИТМА ДЕТЕКТИРОВАНИЯ ОБЪЕКТОВ В ПОСЛЕДОВАТЕЛЬНОСТИ КАДРОВ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ВИДЕОНАБЛЮДЕНИЯ	
Хлопин С.В, Белов Н.А.	82
THE EFFECT OF OPTICAL AND RECOMBINATION LOSSES IN CU₂ZNSN(S,SE)₄- BASED THIN-FILM SOLAR CELLS WITH CDS, ZNSE, ZNS WINDOW AND ITO, ZNO CHARGE-COLLECTING LAYERS	
M.Sc. Dobrozhan O.A., St. Danylchenko P.S., Ph.D. Grynenko, V.V., Prof. Dr. Opanasyuk A.S.	84
DECISION OF OPTIMIZATION PROBLEMS USING SYMMETRIC ALGORITHM OF HEAVY BALL METHOD	
Kharlamova Y.N.	88
ENERGY AND EXERGY ANALYSIS OF SEA WATER PUMP FOR THE MAIN CONDENSER COOLING IN THE LNG CARRIER STEAM PROPULSION SYSTEM	
PhD. Mrzljak Vedran, Student Žarković Božica, PhD Student Eng. Poljak Igor	92
THE PECULIARITIES OF THE METALLURGICAL DESIGN DEVELOPED THROUGH PROJECT MANAGEMENT PRINCIPLES	
Prof.Tontchev N., Assoc. Prof. Dimitrov D. Prof. Hai Hao	96
SOFTWARE ASSURANCE OF THE SYNTHESIS AND DESIGN OF HYPERBOLOID GEAR DRIVES	
Assoc. Prof. Abadjieva E. PhD. , Prof. Sc. D. Abadjiev V. PhD.	100
INVESTIGATION OF SAMPLES ACCURACY TO MODEL THE PROCESSES IN 3D PRINTING	
Assoc. Prof. MEng. Minev R. PhD., Assoc. Prof. MEng. Rusev R. PhD., MEng. Antonov S. PhD., MEng. Minev E. PhD.	104
ELECTRIC POWER SYSTEMS MODELING AND EDUCATION: SHORT TERM UNIT COMMITMENT AND ECONOMIC DISPATCH MODELING FRAMEWORK	
M.Sc. Trashlieva V., M.Sc. Radeva T. PhD.	108
ELECTRIC POWER SYSTEMS MODELING AND EDUCATION: HYDRO POWER STATIONS MODELING FRAMEWORK	
M.Sc. Trashlieva V., M.Sc. Radeva T. PhD.	111
ELECTRIC POWER SYSTEMS MODELING AND EDUCATION: THE CONTROLLABLE LOADS IN A SHORT TERM SYSTEM BALANCE	
M.Sc. Trashlieva V., M.Sc. Radeva T. PhD.	114
МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ОПТИМАЛЬНОГО ИСПОЛЬЗОВАНИЯ РИСУБОРОЧНЫХ КОМБАЙНОВ В КЫЗЫЛОРДИНСКОЙ ОБЛАСТИ	
профессор Ж.Садыков	117
КОЛЕБАНИЯ В САМОУСТАНАВЛИВАЮЩИХСЯ МЕХАНИЗМАХ КОНСТРУКЦИИ МАЯТНИК	
Nauryzbaev R., Sadykov Z., Sansyzbayev K, Toilybaev M, Koshanova S.	122
NUMERICAL MODEL FOR SIMULATION OF THE VELOCITY FIELDS FOR THE EXPLOSIVELY FORMED PENETRATOR	
M.Sc. Hutov I. PhD., Prof. M.Sc. Lilov I. PhD.	125
MODELING OF PRODUCTION PARAMETERS OF B₄C + ZrO₂ COMPOSITES VIA ARTIFICIAL NEURAL NETWORKS METHOD	
S. Hartomacıoğlu. PhD., H.O.Gülsoy. PhD., B. Bakırcıoğlu Ms.C.	129
IN SILICO MODELING AND EVALUATION OF BASFIA SUCCINICIPRODUCENS FOR 1,4-BUTANEDIOL PRODUCTION FROM RENEWABLE RESOURCES	
Dr. Zsolt, Bodor Assistant professor,..... Miklóssy I.....	133
МОДЕЛИРОВАНИЕ ПОЛНОЙ МГД-ЗАДАЧИ В СПИРАЛЬНОМ ТОРОИДАЛЬНОМ ПОТОКЕ	
Чупин Антон Викторович кандидат, физ. - мат. наук, Научный сотрудник	134

MATHEMATICAL MODELLING OF SOCIO-ECONOMIC PROCESSES AND SYSTEMS

MODELLING AND EDUCATION: THE ROLE OF MATHEMATICAL MODELLING IN THE REALIZATION OF CONTINUITY OF THE STOCHASTIC LINE IN THE SCHOOL COURSE OF MATHEMATICS

Щербатых Сергей Викторович 135

THE SUBJECTIVE MODEL OF RATIONAL CHOICE IN MULTI-AGENT SYSTEMS

Gennady Pavlovich Vinogradov 138

TRENDS IN DATA ANALYSIS: STATE, DEVELOPMENT PROSPECTS

Doctor in Economic sciences, Prof., I. A. Katsko; PhD in Economic sciences, P. Yu. Velichko; Student, M. Nikogda 142

INTEGRAL ASSESSMENT OF ENVIRONMENTAL QUALITY AND THE QUALITY OF LIFE OF THE POPULATION OF THE ARCTIC REGIONS OF RUSSIA IN THE PERIOD FROM 2003 TO 2015

Prof. Dr. V.V. Dmitriev 147

4. MATHEMATICAL MODELLING OF MEDICAL-BIOLOGICAL PROCESSES AND SYSTEMS

CONCEPTUAL CYBERNETIC MODEL OF TEACHING AND LEARNING

Prof. Ing. Veronika Stoffová, CSc. 152

SIMULATION MODELING OF AUDITORY FUNCTION

Ass. Prof., Dr. Eng. Donii O., Dr. Med. Pisanko V., Ass. Prof., Dr. Eng. Kulinich A., Ass. Kotliar S. 156

GENERATION OF AN ATLAS-BASED FINITE ELEMENT MODEL OF THE HEART FOR CARDIAC SIMULATION

M.Sc. Vasiliev Evgeny 160

ILLUSTRATION OF MODEL CREATION ON EXAMPLE OF APPROXIMATIONS TO THE STEADY STATE CURRENT OF CHEMICAL CYCLIC PROCESSES

Assoc. Prof. Dimitrov A.G. PhD 163

APPLICATION OF FUZZY MODELING TO PREDICT THE DISEASE OF STAFF FROM EXPOSURE TO WORKING CONDITIONS

Klimova I. V., Smirnov Yu. G. 167

EXACT RECONSTRUCTION VERSION OF RADON TRANSFORMATION IN TOMOSYNTHESIS

Morgun O., Nemchenko K., Vaisburd A., Viktynska T. 171

STOCHASTIC COMPUTER SIMULATION OF THE IONIC DIFFUSION THROUGH BIOLOGICAL TISSUES UNDER THE EFFECT OF DIRECT ELECTRIC FIELD

MSc. Fawzy HM, Associate Prof. Dr. Salem NM, Prof. Dr. El Sheikh SM and Prof. Dr. El-Messierly MA 174

FUNDAMENTALS OF THE KINETIC THEORY OF MULTICOMPONENT EMULSIONS

ОСНОВЫ КИНЕТИЧЕСКОЙ ТЕОРИИ МНОГОКОМПОНЕНТНЫХ ЭМУЛЬСИЙ

Prof., Dr. Tech. Sci. Firsov A.N.

Peter the Great St. Petersburg Polytechnic University – St. Petersburg, Russia

E-mail: anfirs@yandex.ru

Abstract: *The paper proposes a mathematical model for describing the dynamics of multicomponent emulsions, based on ideas and methods of the kinetic theory of gases. The methodological basis of the proposed theory is the ideas and methods of the theory of integral kinetic equations.*

Keywords: MULTICOMPONENT EMULSIONS, KINETIC THEORY, MATHEMATICAL MODEL, INTEGRAL KINETIC EQUATIONS

1. Introduction

In some practically important situations (such as the movement of oil through a well), one has to deal with the problem of the motion of a viscous fluid, inside which there are small inclusions in the form of gas bubbles, droplets of water, solid particles, etc. In this paper, we shall consider the case of emulsions¹. This means that, on the one hand, the entire "mixture" (i.e., the liquid plus inclusions) can be considered as a continuous medium, and on the other hand, that in a small element of the volume of the medium there are "sufficiently many sufficiently small particles". In this connection, it seems quite natural to apply the statistical approach in the above-mentioned problem in the spirit of the kinetic theory of gases. However, the theoretical results known to the author in this direction [1 - 7] are connected with concrete (and rather simple) physical models, which does not allow us to sufficiently cover the problem of the motion of emulsions and give its closed mathematical formulation.

We see the main goal of the work, just in order to give a sufficiently general and precise mathematical formulation of this problem, which allows, in particular, to understand the place and role of some simplifying assumptions. From a methodological point of view, the author followed the basic ideas developed by the Leningrad school of aerodynamics of rarefied gas [8], founded by S. V. Vallander.

2. Statement of the problem of modeling multicomponent emulsions based on a kinetic approach

So, we will consider a viscous liquid, inside which there is a very large number of small "particles" (gas bubbles, drops of other liquids, etc.). The fact that these inclusions are emulsions means, in particular, that the whole mixture (liquid plus inclusions) - from a macroscopic point of view - can be considered as a continuous medium². In this connection, it becomes necessary to distinguish three scales of smallness of distances (and volumes). The first of them corresponds to the concept of a small (elementary, physically infinitesimal) volume of a mixture in the hydrodynamic sense. This is the volume within which, on the one hand, the hydrodynamic and thermodynamic quantities related to the mixture as a continuous medium can be regarded as identical, and on the other hand, in which there are a sufficient number of inclusion particles, so that

the latter can be applied a statistical approach.

The second scale corresponds to the concept of a small (from the hydrodynamic point of view) volume of the main fluid. This volume, generally speaking, is of a lower order than the previous one, and its linear dimensions are of the order of the average distance between the particles.

The third scale corresponds to the dimensions of the inclusion particles themselves. These dimensions will be assumed to be small of a higher order than the average distances between the particles.

As the basic elementary volume, it is natural to take an elementary volume corresponding to the first of the mentioned scales of smallness. Hydrodynamic and thermodynamic quantities characterizing the state of the main fluid will be understood as averaged over such an elementary volume.

Finally, we assume that the inclusion particles have a spherical shape and form a set for which one can use the assumptions commonly used in the definition of the concept of "rarefied gas" (the pairing of collisions, the negligibly small duration of the collision time in comparison with the time of free motion, etc., see, for example, [9]).

The nature of the interaction of particles with each other and with the main fluid requires special consideration. Here we will focus on those aspects of this interaction that are essential for the purposes of this article.

It is known that the gas bubbles differ significantly from the other particles (droplets of liquid, solid particles, etc.), both in terms of their individual properties and the effect on the dynamics of the mixture as a whole. First, this difference is manifested in the fact that the dimensions of gas bubbles can change during their free movement inside the main liquid. This change is due to a change in the temperature and pressure of the main fluid and, as a result, the temperature and pressure of the gas inside the bubbles. With a great degree of certainty, we can assume that at any time the gas inside the bubble is in thermodynamic equilibrium with the surrounding liquid (this means that the temperature and pressure of the gas and liquid are coincide). Thus, the radius r of the bubble is related to the temperature T and pressure p of the main liquid by formula

$$r = \left(\frac{3 m RT}{4 \pi \mu p} \right)^{\frac{1}{3}} \quad (1)$$

Where m - is the mass of gas in the bubble, μ - the molecular weight of the gas, R - the universal gas constant. In the general case, we can assume that r is a given function of m , T , p and μ .

The change in the size of the bubble also results the fact that the mixture (as a continuous medium) cannot be regarded as an incompressible fluid, even if the base liquid is incompressible.

The next feature of bubbles is the possibility of their emergence from "nothing" (for example, as a result of chemical reactions occurring in the main liquid), and as a result of their spontaneous decay. The latter, in the simplest case, can be

¹ The presence of gas bubbles essentially distinguishes the problem of the motion of emulsions from the problem of the motion of suspensions; the latter was quite well developed from different points of view [1 - 7].

² In order to distinguish the emulsion mixture as a continuous medium from the actual liquid, in which all these inclusions are found, we shall call the latter "basic liquid".

considered reliably occurring as a result of reaching a certain limiting bubble radius r_0 .

Interaction of particles with each other is simply their direct collision. In this case, however, a phenomenon such as the fusion of two bubbles or droplets, as well as their crushing are possible. We assume that when two bubbles (or droplets) collide, either their merging is possible, or because of the collision, the number of particles does not change (that is, it remains equal to two), and in the collision of two particles of different types (a bubble with a droplet, etc.), these particles change only the velocities.

3. Basic properties of models and connection with problems of the theory of transfer

We now introduce the basic functions that we shall deal with below, and indicate some of their properties. We will mark the type of particle (bubble, drop, solid particle, etc.) by the indexes i, j, k , etc., taking integral values. The number of values of these indices is obviously finite.

For brevity, we introduce the following terminology. We say that some particle of the type i there is a particle of type (i, x, u, m) , if this particle is reliably located at the point of space with a radius vector $x(x^1, x^2, x^3)$, has a speed $u(u^1, u^2, u^3)$ and mass m , and that this particle is of type (i, x, dx, u, du, m, dm) , if its spatial coordinates are enclosed in the interval $[x^k, x^k + dx^k]$, $k=1,2,3$, the projection of the velocities in the gaps $[u^k, u^k + du^k]$, $k=1,2,3$, and the mass in the gap $[m, m + dm]$.

By the distribution function of the particles of the variety i we shall call the function $f_i(x, u, m, t)$, having the property that the quantity $f_i(x, u, m, t) dx du dm$ gives (up to small higher order) the mathematical expectation of the number of particles of type (i, x, dx, u, du, m, dm) at the moment t . As in [9], it is easy to see that $f_i dx du dm$ there is also a probability of detecting one particle in volume $dx du dm$.

Let us denote by $P_i^{(k)}(x, u, m, t)$ a function possessing the property that the quantity $P_i^{(k)}(x, u, m, t) dt$, $k=2,3,\dots$, is a probability of decay of a particle into k parts over a period of time from t to $t + dt$, if this particle at the time moment t was authentically of the type (i, x, u, m) .

Let's denote by $\omega_i^k(u, m | u_1, m_1)$ the density of mathematical expectation (in the space of variables (u_1, m_1)) of the number of particles of the type $(i, u_1, du_1, m_1, dm_1)$, obtained as a result of authentically decay into k parts of a particle of the type (i, u, m) .

Let us set

$$\hat{T}_i^k(x, u, m, t | u_1, m_1) \equiv P_i^{(k)}(x, u, m, t) \omega_i^k(u, m | u_1, m_1). \quad (2)$$

The following properties of functions ω_i^k и \hat{T}_i^k are obvious:

$$\omega_i^k(u, m | u_1, m_1) = 0 \text{ при } m_1 \geq m, \quad (3)$$

$$\hat{T}_i^k(x, u, m, t | u_1, m_1) = 0 \text{ при } m_1 \geq m, \quad (4)$$

$$\int_0^\infty \omega_i^k(u, m | u_1, m_1) du dm_1 = k, \quad (5)$$

$$\sum_{k=2}^\infty \frac{1}{k} \int \hat{T}_i^k(x, u, m, t | u_1, m_1) du dm_1 = P_i(x, u, m, t) \quad (6)$$

Here

$$P_i(x, u, m, t) = \sum_{k=2}^\infty P_i^{(k)}(x, u, m, t)$$

gives, obviously, the probability of decay per unit time of a particle of the type (i, x, u, m) at the time t . Here and below, if the region of integration is not indicated, we mean the entire space R^3 .

We note that, since the quantity $P_i^{(k)} / P_i$ is the probability of decay of a particle of the type (i, x, u, m) in k parts provided that some decay has occurred reliably, the value

$$\hat{T}_i(x, u, m, t | u_1, m_1) = \frac{1}{P_i} \sum_{k=2}^\infty P_i^{(k)}(x, u, m, t) \omega_i^k(u, m | u_1, m_1) \quad (7)$$

gives the density of mathematical expectation of the number of particles of type $(i, x, u_1, du_1, m_1, dm_1)$, in the space of variables (u_1, m_1) , as a result of a reliable decay of a particle of the type (i, x, u, m) .

By virtue of (3)

$$\hat{T}_i(x, u, m, t | u_1, m_1) = 0 \quad m_1 \geq m. \quad (8)$$

We denote by $\hat{T}_{ij}^k(x, u_1, u_2, m_1, m_2 | u, m)$ the density of mathematical expectation (in the space of variables (u, m)) of a number of particles of type (k, x, u, du, m, dm) , obtained as a result of a reliable collision of particles of types (i, x, u_1, m_1) and (j, x, u_2, m_2) . We note that under the assumptions made in §2 $\hat{T}_{ij}^k = 0$ in the following cases:

- 1) $i = j, \quad k \neq i$;
- 2) $i \neq j, \quad k \neq i, j$;
- 3) $i \neq j, \quad k = i, \quad m \neq m_1$;
- 4) $i \neq j, \quad k = j, \quad m \neq m_2$;
- 5) $i = j = k, \quad m > m_1 + m_2$.

Thus, \hat{T}_{ij}^k are nontrivial only for:

- a) $i = j = k$;
- b) $i \neq j, \quad k = i$; c) $i \neq j, \quad k = j$.

In the cases b) and c) an expression for \hat{T}_{ij}^k must contain factors of the form $\delta(m - m_1)$ or $\delta(m - m_2)$, so that in these cases

$$\int_0^\infty \int T_{ij}^k(x, u_1, u_2, m_1, m_2 | u, m) du dm = 1 \quad (9)$$

Of particular interest is the case a), since here it is necessary to take into account the possibility of merging two particles of one kind. For brevity, we denote $T_{ii}^i \equiv T_i$. Let $h_i(x, u_1, u_2, m_1, m_2)$ is a probability of merging two reliably colliding particles of types (i, x, u_1, m_1) and (i, x, u_2, m_2) , and $\tilde{P}_i(u, u_2, m_1, m_2 | u, m)$ is the probability density (in space (u, m)) of the fact that a particle, which as a result of a reliable fusion of these particles, will

have the type (i, x, u, du, m, dm) . We note at once that the expression for \tilde{P}_i should contain a multiplier $\delta(m - (m_1 + m_2))$. It's obvious that

$$\int_0^\infty \int \tilde{P}_i du dm = 1 \quad (10)$$

Let's further $\tilde{T}_i(u_1, u_2, m_1, m_2 | u, m)$ is the density of mathematical expectation (in the space of variables (u, m)) number of particles of type (i, x, u, du, m, dm) , resulting from a reliable collision (without merging) of particles of the type (i, x, u_1, m_1) and (i, x, u_2, m_2) . For \tilde{T}_i (by the assumptions of §3) we have

$$\int_0^\infty \int \tilde{T}_i du dm = 2 \quad (11)$$

In this way,

$$T_i(x, u_1, u_2, m_1, m_2 | u, m) = h_i(x, u_1, u_2, m_1, m_2) \tilde{P}_i(u_1, u_2, m_1, m_2 | u, m) = (1 - h_i) \tilde{T}_i(u_1, u_2, m_1, m_2 | u, m) \quad (12)$$

From (10) and (11) we obtain

$$\int_0^\infty \int T_i du dm = 2 - h_i \quad (13)$$

Suppose, finally, that the function $\Pi_i(x, u, m, t)$ is such that the quantity $\Pi_i dx du dm dt$ is a mathematical expectation of the number of particles of type (i, x, dx, u, du, m, dm) , arising during a period of time from t to $t + dt$ as a result of processes not associated with collisions and particle decays (for example, as a result of chemical reactions in the main liquid).

We note that all the quantities introduced above can depend (as on the parameters) on the macroscopic characteristics of the main liquid at the corresponding point and at the corresponding instant of time.

In concluding this section, we introduce several functions that are important for the future, connected with the free motion of an individual particle of sort i .

Let at the moment t the particle under consideration is of the type (i, x, u, m) . Then the equation of its motion in the main fluid (pressure, velocity, density and temperature are $p(x, t)$, $v(x, t)$, $\rho(x, t)$, $T(x, t)$ respectively) in the presence of acceleration of gravity g , will have the form

$$\dot{x}(\tau) = u(\tau),$$

$$\begin{aligned} \dot{u}(\tau) &= g - \frac{1}{m} \rho V_i(m, p_\tau, T_\tau) + \frac{1}{m} F_{comp}^{(i)}(m, V_i, p_\tau, T_\tau, v_\tau, u(\tau)) \equiv \\ &\equiv G_i(m, p_\tau, T_\tau, v_\tau, u(\tau)). \end{aligned} \quad (14)$$

These equations must be solved under conditions

$$x(\tau)|_{\tau=t} = x, \quad u(\tau)|_{\tau=t} = u. \quad (15)$$

Here we introduce the following notation:

$$\dot{u} \equiv \frac{du}{d\tau}; \quad p_\tau = p(x(\tau), \tau); \quad T_\tau = T(x(\tau), \tau); \quad v_\tau = v(x(\tau), \tau); \quad R_i(x, u, m, t, t+s) = \exp \left[- \int_t^{t+s} Q_i(x_\sigma^{(i)}, u_\sigma^{(i)}, m, \sigma) d\sigma \right] \quad (22)$$

V_i - denotes the volume of the particle under consideration, $F_{comp}^{(i)}$ - denotes the force of resistance to movement of a particle in the main fluid. In the simplest case, for $F_{comp}^{(i)}$ one can use the Stokes formula

$$F_{comp}^{(i)} = 6\pi\eta r a_i (v - u), \quad (16)$$

Where η - denotes dynamic viscosity of the main fluid; r - particle radius; $a_i = 1$, if it is a solid particle, and $a_i = \frac{1}{3} \frac{2\eta + 3\eta'_i}{\eta + \eta'_i}$, if it is a drop or a gas bubble (here η'_i - dynamic viscosity of a liquid (gas) forming a droplet (bubble)).

Further, we denote by

$$x_\tau^{(i)} \equiv \varphi_i(\tau; x, u, m, t), \quad u_\tau^{(i)} \equiv \psi(\tau; x, u, m, t) \quad (17)$$

the solution of problem (14) - (15). Clearly,

$$x_\tau^{(i)} = x, \quad u_\tau^{(i)} = u \quad (18)$$

For each fixed τ functions (18) give a diffeomorphism of the phase space (x, u) into itself. We denote by

$$D_\tau^{(i)} = D_\tau^{(i)}(x, u, m, t) \text{ Jacobian} \quad D_\tau^{(i)} = \frac{D(\varphi_i, \psi_i)}{D(x, u)} \quad (19)$$

In particular, an element of the volume of the phase space $dx_\tau^{(i)} du_\tau^{(i)}$ is associated with an element $dx du$ by the relation

$$dx_\tau^{(i)} du_\tau^{(i)} = |D_\tau^{(i)}| dx du \quad (20)$$

We also note that $D_t^{(i)} = 1$.

4. Derivation of integral kinetic equations of the theory of multicomponent emulsions

The principle of the derivation of a system of integral kinetic equations for functions f_i basically does not differ from usual [9, 10]. We, therefore, indicate here only the necessary changes.

4.1. Probability of free movement.

Let $Q_i(x, u, m, t)$ is the probability of collision or decay per unit time of the particle, which at the moment t authentically had the type (i, x, u, m) .

For Q_i we have the expression

$$Q_i(x, u, m, t) = \sum_j \int_0^\infty \int f_j(x, u_3, m_3, t) (r^{(i)} + r_3^{(j)}) |u_3 - u| du_3 dm_3 + P_i(x, u, m, t), \quad (21)$$

where $r^{(i)}$, $r_3^{(j)}$ - denote radii of particles (i, x, u, m) and (j, x, u_3, m_3) respectively. We emphasize that $r^{(i)}$ и $r_3^{(j)}$ depend, generally speaking, on $x, t, m(m_3)$ - see the formula (1).

Let's further $R_i(x, u, m, t, t+s)$ is a probability of a free (without collision and decay) motion during the time from t to $t+s$ of a particle, which at the moment t authentically had the type (i, x, u, m) . Then

$$R_i(x, u, m, t, t+s) = \exp \left[- \int_t^{t+s} Q_i(x_\sigma^{(i)}, u_\sigma^{(i)}, m, \sigma) d\sigma \right] \quad (22)$$

We note that when the expression (21) is substituted into (22), $r^{(i)}$, $r_3^{(j)}$ should be replaced by $r_\sigma^{(i)}$, $r_{3,\sigma}^{(j)}$.

4.2. Birth function.

We denote by $\phi_i(x, u, m, t)$ the birth function of

particles of sort i , that is, a function possessing the property that the quantity

$$dn_0^{(i)} = \phi_i(x, u, m, t) dx du dm dt$$

is a mathematical expectation of the number of particles of type (i, x, dx, u, du, m, dm) , born within a period of time from t to $t + S$. Clearly,

$$dn_0^{(i)} = dn_1^{(i)} + dn_2^{(i)} + dn_3^{(i)},$$

where $dn_1^{(i)}$ is a mathematical expectation of the number of particles of this type, born as a result of collisions, $dn_2^{(i)}$ - born as a result of decay, a $dn_3^{(i)}$ - arising from other causes.

Using the notation of the preceding section, we obtain

$$\begin{aligned} dn_2^{(i)} &= dx du dm dt \int_0^\infty \int f_i(x, u_1, m_1, t) P_i(x, u_1, m_1, t) \hat{T}_i(x, u_1, m_1, t | u, m) du_1 dm_1; \\ dn_3^{(i)} &= \Pi_i(x, u, m, t) dx du dm dt. \end{aligned}$$

The expression for $dn_1^{(i)}$ has the usual form [10]. In this way,

$$\begin{aligned} \phi_i(x, u, m, t) &= \frac{\pi}{2} \sum_{j,k} \int_0^\infty \int_0^\infty |u_1 - u_2| \left(t_1^{(j)} + t_2^{(k)} \right)^2 f_j(x, u_1, m_1, t) f_k(x, u_2, m_2, t) \times \\ &\times T_{jk}^i(x, u_1, u_2, m_1, m_2 | u, m) du_1 du_2 dm_1 dm_2 + \int_0^\infty \int f_i(x, u_1, m_1, t) P_i(x, u_1, m_1, t) \times \\ &\times \hat{T}_i(x, u_1, m_1, t | u, m) du_1 dm_1 + \Pi_i(x, u, m, t) \end{aligned} \quad (23)$$

The system of integral kinetic equations for f_i is output in the standard way [9], and has the form (in the absence of boundaries³)

$$\begin{aligned} f_i(x, u, m, t) &= f_i(x_{t_0}^{(i)}, u_{t_0}^{(i)}, m, t_0) \times \exp \left[- \int_{t_0}^t Q_i(x_q^{(i)}, u_q^{(i)}, m, q) dq \right] \times \\ &\times \left| D_{t_0}^{(i)} \right| + \int_{t_0}^t \phi_i(x_\tau^{(i)}, u_\tau^{(i)}, m, \tau) \times \exp \left[- \int_\tau^t Q_i(x_q^{(i)}, u_q^{(i)}, m, q) dq \right] \left| D_\tau^{(i)} \right| d\tau. \end{aligned} \quad (24)$$

Here $t_0 < t$ - is arbitrary time moment.

A simple technique analogous to that used in [11] and consisting in the passage to the limit $t \rightarrow t_0$, enables us to obtain from (24) the following system of integro-differential equations

$$\frac{\partial f_i}{\partial t} + u \cdot \frac{\partial f_i}{\partial x} + \frac{\partial (G_i f_i)}{\partial u} = \phi_i - f_i Q_i, \quad (25)$$

Where G_i is given by the relation (14) and all functions are taken at the point (x, u, m, t) .

5. Analysis of the relationship between the macroparameters of the mixture and the main liquid

If given $p(x, t)$, $v(x, t)$, $T(x, t)$ then equations (24) together with equalities (21), (23) form a closed system. However, in practice it is difficult to determine the quantities characterizing the main liquid. On the other hand, (in particular, experimentally) values $p_c(x, t)$, $v_c(x, t)$, $T_c(x, t)$, $\rho_c(x, t)$, characterizing the pressure, velocity, temperature, and density of the mixture as a continuous medium, can be found. Therefore, in order to close the system of equations, it is necessary to add the relations connecting

f_i , p_c , v_c , T_c , ρ_c , p , v , T , ρ . Let's do it. In the future, it will be more convenient to use the value of the internal energy density instead of the temperature. So, let $E(x, t)$ - is a density (per unit mass) of the internal energy of the main fluid. $E_c(x, t)$ - density (per unit mass) of the internal energy of the mixture (as a continuous medium) and e_i - density (per unit mass) of the internal energy of the substance (gas, liquid) in the particle of the variety i . With the assumptions made in §3, we can assume that $e_i = e_i(p, T)$. Let dx - a certain elementary volume of the mixture (see §3) adjacent to the point x . Obviously,

$$dx = d_1 x + d_2 x,$$

where $d_1 x$ - is a part of the volume dx , occupied by particles, and $d_2 x$ - part of the volume dx , occupied by basic fluid.

We have

$$d_1 x = dx \sum_j \int_0^\infty \int V_i(m, p, v, T) f_i(x, u, m, t) du dm$$

where V_i - is a volume of a particle of a variety i .

The mass of the particles inside dx will be equal

$$d_1 M = dx \sum_i \rho_i \int_0^\infty \int V_i f_i du dm = dx \sum_i \int_0^\infty \int m f_i du dm,$$

where ρ_i - is a density of matter in the corresponding particle.

The mass of the main liquid inside dx is equal, hence,

$$d_2 M = dx \rho \left(1 - \sum_i \int_0^\infty \int V_i f_i du dm \right).$$

Thus, for the density of the mixture we obtain

$$p_c(x, t) = \rho - \sum_i \left(\frac{\rho}{\rho_i} - 1 \right) \int_0^\infty \int m f_i du dm \quad (26)$$

Let us turn to the expression for the macroscopic velocity of the mixture $v_c(x, t)$. We use the formula

$$dK_c = v_c \rho_c dx,$$

where dK_c - is the amount of motion of the mixture inside dx . Taking into account that dK_c is composed of the amount of motion of the particles and the main fluid, we obtain

$$v_c = \frac{1}{\rho_c} \left[\rho v \left(1 - \sum_i \frac{1}{\rho_i} \int_0^\infty \int m f_i du dm \right) + \sum_i \int_0^\infty \int m u f_i du dm \right] \quad (27)$$

To find the expression for E_c , note that the total energy of the mixture inside dx is made up of the kinetic and internal energies of the particles and the main fluid. Carrying out the corresponding calculations, we obtain

$$E_c = U_c / \rho_c - v_c^2 / 2, \quad (28)$$

Where U_c - is the bulk density of the total energy of the mixture:

$$\begin{aligned} U_c &= E \rho - \sum_i \left(\frac{E \rho}{e_i \rho_i} - 1 \right) \int_0^\infty \int m e_i f_i du dm + \\ &+ \frac{\rho v^2}{2} \left(1 - \sum_i \frac{1}{\rho_i} \int_0^\infty \int m f_i du dm \right) + \sum_i \int_0^\infty \int \frac{m u^2}{2} f_i du dm \end{aligned} \quad (29)$$

Formulas (26) - (29) give the required additional relations for the closure of the system of equations (24).

6. Concluding remarks

Let us now make a few general remarks.

1. Equations (24), (25) are externally similar to the known kinetic equations for gas mixtures [10], but their content is largely different. In particular, under the assumption that there are no collisions between the particles in the problem under consideration,

³ In practically interesting problems, we can assume that there are no solid boundaries in the mixture, or on them the functions f_i are given.

the right-hand side of (25) does not vanish:

$$\begin{aligned} \phi - f_i Q_i = & \int_0^\infty \int f_i(x, u_1, m_1, t) P_i(x, u_1, m_1, t) \times \\ & \times \hat{T}_i(x, u_1, m_1, t | u, m) du_1 dm_1 + \Pi_i(x, u, m, t) - \\ & - f_i(x, u, m, t) P_i(x, u, m, t) \end{aligned} \quad (30)$$

2. «Equilibrium solution» (that is, a solution that does not depend on x, t) plays a somewhat different role in the problem under study than in the kinetic theory of gases. This is because the dependence on the coordinate is laid inside the very essence of the problem: macro parameters $p(x, t)$, $T(x, t)$ etc. *a priori* are changing over x very quickly.

The latter means that solutions close to equilibrium are of little interest from a practical point of view, and therefore the equilibrium solutions themselves are of little interest. They, however, can be considered as limiting (at $|x| \rightarrow \infty$) function values f_i ; so the equilibrium solutions can be used mainly in the formulation of boundary value problems in unbounded domains.

On the other hand, stationary solutions may be the most interesting from a practical point of view.

3. In the problem of the motion of emulsions, the viscosity of the main liquid plays an important role, and, generally speaking, it can not be neglected. Indeed, if we neglect the term $F_{comp}^{(i)}$ in (14), then any freely moving particle, after a sufficiently long time, will have an arbitrarily high velocity. Consequently, it may turn out that

$$|u|, |x|, t \rightarrow \infty \quad f_i \neq 0,$$

but it is unreasonable.

4. In general case, theoretical search of functions P_i , \hat{T}_i , T_i^k etc., defined in §3, is a very difficult task and is currently hardly feasible. In this connection, in our opinion, the role of experimental studies in this direction increases.

We indicate, however, the formula for $P_i(x, u, m, t)$ in the event that the decay of a particle of a variety i occurs if and only if its radius reaches a certain limiting value $r_0^{(i)}$. As already mentioned, we can assume that $r = \theta_i(p(x, t), T(x, t), m)$.

Let

$$m_0^{(i)} = m_0^{(i)}(p, T, r_0^{(i)})$$

is such that

$$r_0^{(i)} = \theta(p, T, m_0^{(i)}).$$

Then

$$P_i(x, u, m, t) = u \cdot \left(\frac{\partial \theta_i}{\partial p} \frac{\partial p}{\partial x} + \frac{\partial \theta_i}{\partial T} \frac{\partial T}{\partial x} \right) \left| \frac{\partial \theta_i}{\partial m} \right|^{-1} \delta(m - m_0^{(i)})$$

5. In the previous sections we did not take into account the possibility of the proper rotation of the inclusion particles. Accounting for this possibility does not cause fundamental difficulties in the sense of deriving the basic equations, but the solution (in particular, numerical) of these equations is significantly complicated because of the increase in the number of independent

variables for functions f_i . On the other hand, in the broad class of practically interesting cases, the assumption of the absence of proper rotation of spherical particles is justified [1, 2].

7. References

1. Fortier A. Mécanique des suspensions. Masson et C-ie Éditeurs, Paris-VI^e, 1967.
2. Reologiya suspenziy. Seriya: Biblioteka Sbornika «Mekhanika»: Sbornik statey. M.: Mir, 1975. [Rheology of suspensions. Series: Library Collection "Mechanics"]. Moscow: Mir, 1975. (In Russian)
3. Bibik Ye.Ye. Reologiya dispersnykh sistem. L.: Izd-vo Leningr. un-ta, 1981. [Bibik E.E. Rheology of disperse systems]. L.: Leningrad University, 1981. (In Russian)
4. Kelbaliyev G. I. i dr. Mekhanika i reologiya neftyanykh dispersnykh sistem. M.: Izd-vo «Maska», 2017. [Kelbaliev G. I. et al., Mechanics and rheology of petroleum disperse systems]. Moscow: "Mask", 2017. (In Russian)
5. Tsibarov V.A. Kineticheskaya model' psevdoozhizhennogo sloya. – *Vestn. Leningr. un-ta*, 1975, № 13, s. 106-111. [Tsibarov V.A. Kinetic model of a fluidized layer]. – *Vestnik of Leningrad University*, 1975, No. 13, p. 106 - 111. (In Russian)
6. Pokrovskiy V.N. Statisticheskaya mekhanika razbavlenykh suspenziy. M.: Izd-vo «Nauka», 1978. [Pokrovsky V.N. Statistical mechanics of dilute suspensions]. Moscow: Nauka, 1978. (In Russian)
7. Protod'yakonov I.O., Tsibarov V.A., Chesnokov YU. G. Kineticheskaya teoriya gazovzvesey. L.: Izd-vo Leningr. un-ta, 1985. [Protod'yakonov I.O., Tsibarov V.A., Chesnokov Yu.G. Kinetic theory of gas-weights]. L.: Leningrad University, 1985. (In Russian)
8. *Aerodinamika razrezhennykh gazov*. Vyp. 1 – 11. Sbornik statey. Pod red. S.V. Vallandera, R.G. Barantseva, L.: Izd-vo Leningr. un-ta, 1963 – 1983. [Vallander S.V., Barantsev R.G. (editors), Aerodynamics of rarefied gases]. Issues 1 – 11, L.: Leningrad University, 1963 - 1983. (In Russian)
9. Vallander S.V. Uravneniya i postanovka zadach v aerodinamike razrezhennogo gaza. – *Aerodinamika razrezhennykh gazov*. Vyp. 1. L.: Izd-vo Leningr. un-ta, 1963, s. 7-37. [Vallander S.V.. Equations and statement of problems in aerodynamics of a rarefied gas]. - In: Aerodynamics of rarefied gases. Issue. 1. L.: Leningrad University, 1963, p. 7-37. (In Russian)
10. Vallander S.V., Belova A.V. Integral'nyye kineticheskiye uravneniya dlya smesi gazov s vnutrennimi stepenyami svobody. - V kn.: *Aerodinamika razrezhennykh gazov*. Vyp. 1. L.: izd-vo Leningr. un-ta, 1963, s. 45-52. [Vallander S.V., Belova A.V. Integral kinetic equations for a mixture of gases with internal degrees of freedom]. - In: Aerodynamics of rarefied gases. Issue 1. L.: Leningrad University, 1963, p. 45-52. (In Russian)
11. Filippov B.V. Variant nestatsionarnykh kineticheskikh uravneniy. - V kn.: *Aerodinamika razrezhennykh gazov*. Vyp. 1. L.: izd-vo Leningr. un-ta, 1963, s. 67-73. [Filippov B.V. Variant of non-stationary kinetic equations]. - In: Aerodynamics of rarefied gases. Issue 1. L.: Leningrad University, 1963, p. 67-73. (In Russian)

ABOUT THE PROBLEM OF DATA LOSSES IN REAL-TIME IOT BASED MONITORING SYSTEMS

Prof. Aliexsieiev V. PhD.¹, Prof. Gaiduchok O. PhD.¹
Lviv Polytechnic National University, Lviv, Ukraine ¹

Vladyslav.I.Aliexsieiev@lpnu.ua

Abstract: Fast growing market of IoT devices revealed a number of complex problems. Among these problems, there is a problem of data losses caused by data package losses or delays while its transition from sensor to server. As anticipated, there are a number of businesses relying on easy opportunity to build real-time monitoring systems using modern software and IoT hardware solutions. Although the growing reliability of contemporary communication networks one can find the problem of making decision about lost or delayed data packages. Current research is dedicated to building an algorithm for compensation of gaps in data series to support real-time monitoring systems with appropriate artificially generated values. Cases of applicability of the algorithm were also studied and discussed.

Keywords: DATA LOSSES, DATA SERIES, TIME SERIES, IOT, COMPENSATION ALGORITHM

1. Introduction

An IoT growth allowed designing some new platforms for supporting business both with surveillance tools and analytics software applications. Three key features of such platforms are:

- Device – a remote computer like Raspberry Pi or any of its alternatives [1] or some custom hardware device that may include a set of sensors (business solution may consist of a whole network of such devices).
- Internet – any kind of Internet wireless connection via Wi-Fi, GPRS, 3G/4G or anything else supplied with mobile network [2] (business solution may combine different types of Internet service providers to establish connection).
- Cloud – any popular IoT platform [3] to store and process big data.

If one adds some software solution to provide data analytics to that IoT platform than it becomes a powerful business tool supporting real-time monitoring. There is a number of companies developing their own platforms or exploiting powerful cloud services to provide their client with such platform as a reliable business solution.

However, practical use of such system reveals some serious problems. One of most important problem is the problem of lost or delayed data packages. It is obvious that real-time monitoring faces vulnerability to data losses. The data losses or at least delays in package deliveries via Internet over mobile networks is an ordinary problem. Such data losses may cause fake alerts about hardware failures or, otherwise, hide a real failure or a critical state. This means the correct functioning of the whole solution requires some reliable analytical decision about current state of remote device.

2. Prerequisites and means for solving the problem

Let us assume that data packages should come to server within a fixed period of time t_p . This can be any value appropriate for particular business logic. For example, it can be a quarter of an hour – receiving packages every 15 minutes. Again, depending on observed business area we can define a critical time for monitoring and accuracy of alerts. Now we will not discuss the accuracy of data and the actual method or frequency of sensors gathering some values. Thus, we assume that each one sensor gives us a single value across the period of t_p . This value is sent to our server within a single data package (there can be a set of values – one for each sensor in the remote device).

As we analyze some value, than we can assume a range of “good” and “bad” values for normal state and failure state. For example, the range $v_{\min} \leq v \leq v_{\max}$ can be defined to indicate normal state and values outside of this range can be considered as failure state. For the purpose of analysis of the need of invoking an alert, we can simplify these values to Boolean value: TRUE for normal state, FALSE for failure state. We also need one more value to

indicate missing value (package loss) – NULL can be an appropriate one for that state.

Next, we are to define the number of lost packages to consider device to go offline and the size of a “window” to display the current state in the real-time monitoring system. The exact values depend on peculiarities of observed processes and can be defined experimentally. For the purpose of our research, let us assume these values (N_{offline} and N_{window}) to be interdependent:

$$N_{\text{offline}} / N_{\text{window}} = k \in (0, 1] \quad (1)$$

The k value can describe system sustainability or vulnerability to data losses (with consider to the “window” width). $k=1$ means sustainable system, and $k \rightarrow 0$ means vulnerable system. To be clear we can consider the k value as a measure of customer requirement of data losses vulnerability.

The easiest conditions to make decision about system state (normal or failure) are in case with no data losses. We can use simple probability calculation to find whether system should indicate normal state (n_{true} and n_{false} are the numbers of TRUE and FALSE values among all N_{window} values in the “window” of observation, so that $N_{\text{window}} = n_{\text{true}} + n_{\text{false}}$):

$$n_{\text{true}} / N_{\text{window}} = 1 - n_{\text{false}} / N_{\text{window}} > 0.5 \quad (2)$$

Condition “> 0.5” is expected to be strict to be sure that most values indicates the normal state. Table 1 lists some common examples for definition of state indicator according to analysis of sensor values within predefined “window” ($N_{\text{window}}=8$).

Table 1: Examples of state definition in case without data losses ($N_{\text{window}}=8$).

State indicator	Set of values							
Failure	F	F	F	F	F	F	F	F
Failure	F	F	T	F	T	F	F	T
Failure	F	F	F	F	T	T	T	T
Normal	T	T	F	F	T	T	T	F
Normal	T	T	T	T	T	T	T	T

In case of data losses, it looks unclear how to make a reliable decision about current state. On one hand, uncertainty can be ignored, but this will yield mistakes in indication the state. On the other hand, replacement of lost data with artificially generated value is possible, but we cannot be definitely sure about accuracy of the result. Table 2 shows examples of situations, when data loss obstructs ability to indicate system state.

Table 2: Examples of uncertainty caused by data losses ($N_{\text{window}}=8$, F – FALSE, T – TRUE, N – NULL).

State indicator	Set of values							
Unknown	F	N	T	N	F	N	T	F
Unknown	N	N	N	N	N	N	F	T
Unknown	N	N	N	N	T	T	T	T
Unknown	N	N	F	F	T	T	N	N
Unknown	T	T	F	F	N	T	T	F
Unknown	T	T	F	T	F	F	N	N

There is a number of algorithms based on calculation of some kind of average values – arithmetic mean or moving average [4, p.153]. However, such approach usually uses only previous data and does not “cover” gaps. Considering trends in data series with moving average “on-the-fly” can be a good method to replace data losses of the last package. Nevertheless, we cannot rely on several artificially generated values to cover new loss. The best way is to store gaps (for example, as a NULL values) and use some other algorithm to cover gaps later.

Another way to fill in the gap is to use some regression model or prediction technique [5, p.279]. For example, one may use simple linear regression or logistic regression model to cover gaps in our time series. Trying to make it better, we can even use polynomial or spline regression models [6, p.166]. Considering these methods, one can mention a high complexity for its implementation. This means that these methods will not fit the requirement of “on-the-fly” data processing, particularly in case of “big data” volumes.

Finally, we can define a couple of requirements for constructing an appropriate algorithm to cover gaps in time series caused by data losses:

- Simplicity – the algorithm should be easy to implement and fast enough to be used “on-the-fly”.
- Reliability – the algorithm should give a reliable approximation for each lost value as the most probable value.

3. Solution of the examined problem

First, we are to determine key concept for the algorithm, to fit all above mentioned requirements and limitations: *a lost package value can be replaced with a dominant value of neighbouring packages*. Both with a common sense of this concept we should assume to keep the quantitative majority concept of eq. (2). In the case of missing values presented with its number n_{null} we can rewrite (2) as

$$n_{true} / (N_{window} - n_{null}) = 1 - n_{false} / (N_{window} - n_{null}) > 0.5 \quad (3)$$

Here, the number of values in the “window” fits the condition $N_{window} = n_{true} + n_{false} + n_{null}$. We should also mention, that $n_{null} < N_{offline}$ is the rule to consider device staying online. Now we can search the way to determine that dominant value to cover gap.

Second, we are to determine known patterns of data losses presented in terms of Boolean values and most appropriate replacement for each missing value. Table 3 lists examples of patterns containing three packages (these cases are obvious according to eq. (3)).

Table 3: Example of 3-package pattern (F – FALSE, T – TRUE, N – NULL).

Replacement	Set of values (X for F or T)					
N→X	N	X	X			...
N→X	...		X	N	X	...
N→X					X	X N

Next, in Table 4, we list examples of patterns containing four packages (these patterns were formed excluding the subset of patterns similar to smaller patterns in Table 3).

Table 4: Example of 4-package pattern (F – FALSE, T – TRUE, N – NULL).

State indicator	Set of values (X for F or T and Y for Not{X})							
N→X	N	X	Y	X				...
N→X	N	Y	X	X				...
N→X	Y	N	X	X				...
N→X	...		X	N	Y	X		...
N→X				X	X	N	Y	
N→X				X	Y	X	N	
N→X				X	X	Y	N	

Here and in further patterns, we use a general extension to the right rule – “first step right, next step left” – and in case of reaching the border of the “window” we can use extension to the free border (left or right accordingly). The rule of extension allows us find easily the dominant value to fit the majority concept. It is important

here to remember, that we place packages like time series – from the left to the right. This assumption means we have latest values closer to the right border of the “window” and this is why we should use extension to the right rule. Therefore, this is how we accomplish not only implementation of majority concept, but also consider the influence of the latest state (unlike to most of moving average algorithms).

In some cases a “collision” can happen – the state of emerging new gap instead of value during extension procedure and reaching both borders of the “window”. The collision state means repetitive need of extension while gap cover is impossible. To avoid such difficulties in case of collision we use propagation rule – “replace all nulls with major value (if can be found) or last value (if there is no major value), when number of meaning values exceeds number of nulls” – and in case of majority of nulls, we can consider device to be offline.

4. Results and discussion

Now let us formalize description of the algorithm for compensation of gaps in data series:

1. Define a “window” width N_{window} with respect to $N_{offline}$ and transition to “offline” state.
2. Use (2) to make decision when there are no gaps.
3. Following extension to the right rule – “first step right, next step left” – use (3) to make decision when there are gaps.
4. Use propagation rule – “replace all nulls with major or last value, when number of meaning values exceeds number of nulls” – in case of a “collision”.

Such kind of algorithm allows covering all the gaps. It is simple enough to be used the same easily both at back-end or front-end solutions to supply appropriate values in customer’s real-time IoT based monitoring system. Certainly, it is more likely to utilize the algorithm at front-end to lower the load of servers.

For the purpose of greater accuracy, one can choose between different kinds of moving averages (as a common solution for considering missing values), regressions models (as a common predictive solution) and various approximation techniques (like polynomial or spline).

5. Conclusion

The described algorithm for compensation of data losses fits all requirements of implementation simplicity and reliability. The level of computation efforts and complexity of the algorithm are relevant to arithmetic mean calculations. Simultaneously, the algorithm considers latest values (unlike the moving average techniques). The influence of latest values is crucial for indicating current system state. Considering the initial conditions for the problem and a specific time series, the algorithm appears to be the most adequate and reasonable solution.

6. Literature

1. 10 Best Raspberry Pi and Pi 2 Alternatives. // Beebom.com. – April 18, 2017. – <https://beebom.com/raspberry-pi-and-pi-2-alternatives/>
2. Global State of Mobile Networks (February, 2017) // OpenSignal, Inc. – February 6, 2017. – <https://opensignal.com/reports/2017/02/global-state-of-the-mobile-network>
3. Top 20 IoT Platforms in 2018./ Santosh Singh // Internet of Things Wiki – March 8, 2016 – <https://internetofthingswiki.com/top-20-iot-platforms/634/>
4. Allen B. Downey. Think Stats. — O’Reily, 2015. – 225 p.
5. Head First Data Analysis. — O’Reily, 2009. – 488 p.
6. Peter Bruce and Andrew Bruce. Practical Statistics for Data Scientists. — O’Reily, 2017. – 317 p.

COMPARISON OF THE RESULTS OBTAINED BY PSEUDO RANDOM NUMBER GENERATOR BASED ON IRRATIONAL NUMBERS

PhD. Dimitrievska Ristovska V., Prof. PhD. Bakeva V.

Faculty of Computer Science and Engineering- Ss. Cyril and Methodius University, Skopje, Macedonia,

vesna.dimitrievska.ristovska@finki.ukim.mk

verica.bakeva@finki.ukim.mk,

Abstract: Pseudo-random number generators (PRNG) based on irrational numbers are proposed elsewhere. They generate random numbers using digits of real numbers which decimal expansions neither terminate nor become periodic and practically their decimal expansion has infinite period. Using that algorithm, we generate sequences of random numbers and then we check their randomness with statistical tests from Diehard battery. Our main idea is to check if there is a difference in the randomness of the generated sequences if digits of any irrational non- transcendental number (like $\sqrt{2}, \sqrt{3}, \sqrt{5}, \dots$) are used versus the case when digits of a transcendental number (like π or e) are used. In our experiments we use about $3 \cdot 10^7$ digits of a given non-periodic irrational or transcendental number. Many experiments were done and all generated sequences by proposed PRNG based on irrational numbers passed the Diehard tests very well. We may conclude that there is not a significant difference in the randomness of the generated sequences in the both cases (irrational non-transcendental versus irrational transcendental number).

Keywords: PRNG, IRRATIONAL NUMBERS, TRANSCENDENTAL NUMBERS, STATISTICAL TESTS, DIEHARD BATTERY, RANDOMNESS

1. Introduction

Pseudo-random number generator (PRNG) is an algorithm which generates a long sequence of numbers r_1, r_2, \dots which are elements of a given set of numbers and the distribution of generated numbers r_1, r_2, \dots is supposed to be uniform.

A sequence of obtained random numbers r_1, r_2, \dots should have two important properties: uniformity (i.e., they are equally probable everywhere) and independence (i.e., the current value of a random variable does not depend on the previous values).

In practice, we cannot construct an ideal PRNG, since the way we are building the mechanism is not a random one, but in fact it is completely determined by an initial value. This affects the uniformity and independence of the produced sequences and r_1, r_2, \dots and that is why the word "pseudo" is used and we have to measure the randomness of the obtained sequences.

A good random number generator should have some additional qualities as large period and small order computational complexity.

This paper is organized as follows. In Section 2 we give a background and overview of related works. In Section 3 we explain basic ideas for construction of our generator of pseudo random numbers [2], basic principles for usage of statistical tests from Diehard battery and then we present the algorithm for the generator. The obtained results are given in Section 4. In Section 5, some conclusions are made.

2. Background and overview of related works

In this section we present some historical facts recall on L'Ecuyer (2017) in [4] about PRNGs which use irrational numbers.

The inspiration for using successive digits of π , e or any other transcendental number in order to generate a random number sequence is an old idea.

For example, Metropolis et al. (1950) in [6] succeeded to compute 2000 decimals of π and e and confirmed that these sequences pass elementary statistical tests. This testing was extended to the first 10000 decimals by Pathria (1962) in [7] and to 100000 decimals by Esmenjaud-Bonnardel (1965) in [3]), and all of these sequences very well pass elementary statistical tests.

Till now, many sequences of digits of π have been obtained and tested and many papers have discussed this idea. The world record in 2016 was 22 459 157 718 361 decimal digits of π , computed in

about four months by Peter Trueb using an algorithm of Chudnovsky and Chudnovsky (1989) (in [1]), Bellard's formula, and the Y-Cruncher multi-threaded software (Yee, 2017) (in [9]). However, to give good reason that the successive digits of π (or any other given irrational number) in a given base b can be taken as random sequence, it should be good to know that this sequence of digits is uniformly distributed in base b , i.e., that each of the b possible digits appears with frequency $1/b$ (on average) in the infinite sequence. For π , practical counting over several digits suggests that this is true, but there is no known proof of it.

However, the property of uniform distribution of the digits of any given irrational number is not sufficient; we need to have the uniform distribution of the pairs, triplets, and so on.

In the latest years, there are some trials to design a PRNG using digits of any irrational number since irrational numbers have decimal expansions that neither terminate nor become periodic, practically their decimal expansion has infinite period. In [8], Rogers and al. (2015) proposed an algorithm for pseudo-random 5-digit numbers using the digits of π and made some visual and statistical analyses for goodness of proposed generator.

In [2], the author proposed a new algorithm for generating pseudo-random numbers using digits of any irrational number. The randomness of obtained sequences of numbers is checked by some statistical tests and the test results are very well.

In this our paper, the main work is: using algorithm proposed in [2], to check if there is a difference in the randomness of the generated sequences, if digits of any irrational non- transcendental number (like $\sqrt{2}, \sqrt{3}, \sqrt{5}, \dots$ or the golden ratio $\phi = \frac{1+\sqrt{5}}{2}$) are used versus the case when digits of a transcendental number (like π or e) are used.

3. PRNG based on digits of irrational numbers

3.1 The main idea in the algorithm– n -tuples

We will explain the ideas for designing a PRNG using digits of an irrational number ([2]), by example digits of π . In this example, we will use sequential n -tuples, for example 10-tuples of digits from the decimal expansion of π . Using some database we will take l

digits of π . In the next step, decimal 10-digit number from every 10-tuple is generated.

If we take the obtained 10-digit *number* (which is obtained directly from the 10-tuple) then its value is in the range $0 \leq \text{number} < 10^{10}$ and we need to check if the *number* is greater than *max* (maximal allowed generated number). If it is true, we have to omit the obtained *number* and continue with checking the next 10-tuple. The 10-tuples are taken without overlapping.

But, we will improve the previous idea if we scale the obtained 10-digit *number* (which is obtained directly from the 10-tuple) from the range $0 \leq \text{number} < 10^{10}$ in the range $0 \leq \text{number} < 2^{32}$. In this way, there is no need to check if the scaled number is greater than *max*. Note that if the sequence of generated number is uniform then the scaled sequence will be uniform on the set $\{0, 1, \dots, \text{max}\}$.

These steps from the last idea will be repeated until we obtain *l* numbers.

3.2 Input parameters and the algorithm

In order to produce different sequences, each time when we started generating of a new sequence, we must initialize the beginning pointer to an arbitrary digit of the chosen irrational number. The position of the beginning pointer will be an input parameter. Also, the input parameter will be the length *n* of digits (*n*-tuples) for generating of each number in the sequence (in the previous example, we choose $n = 10$). Let stress that we compute the digits of any irrational number using package *Mathematica*.

Algorithm

[1] Choose an irrational number which digits will be used for generating random numbers.

[2] Set the length *n* of digits for generating of each number in the sequence, the position *s* of the beginning pointer (the first digit where the generating starts), the length *l* of generated sequence and the maximum *max* of the generated numbers.

[3] Let *counter*=0.

[4] Until *counter* ≤ *l* do

[4.1] Use slice size of *n* digits to generate a number *r*.

[4.2] Scale $r \leftarrow \left[\frac{r}{10^n - 1} \cdot \text{max} \right]$.

[4.3] Put pointer position $s \leftarrow s + n$.

[4.4] *counter* = *counter* + 1.

We will notice that software realization of this algorithm and many experiments were done using package *Mathematica*.

3.5. Diehard tests

Nowadays there are a lot of tests for randomness and all of them measure the difference between the generated pseudo-random sequences and the theoretically supposed ideal random sequence. We say that a PRNG passes a test if the random sequences produced by that PRNG pass the test with a probability near to 1. We can classify PRNGs depending of the tests they have passed. So, for obtaining a better classification we should have many different tests.

Over several years, George Marsaglia [5] has developed Diehard tests as a battery of statistical tests for measuring the quality of a random number generator. This battery was published in 1995. It consists of 15 statistical tests, and it is a comprehensive set of statistical tests for PRNG and serves as some kind of litmus for checking and certification of PRNG. If a PRNG passes Diehard statistical tests, then it can be used in deeper scientific researches.

The Diehard battery consists of Birthday Spacings Test, Overlapping 5-Permutation Test (OPERM-5), Binary rank tests, 31×31 Binary Matrix, 32×32 Binary Matrix, 6×8 Binary Matrix, Bitstream Test, Test OQSO (Overlapping quadruples sparse occupancy test), Test DNA, Count the 1's Test for specific bytes, Parking test, Minimum Distance Test, 3D Spheres Test, Squeeze Test, Overlapping Sums Test, Runs Test and Craps Test.

We will note that the most of the tests in Diehard return a *p*-value, which should be uniform on $[0, 1]$ if the input file contains truly independent random bits. Those *p*-values are obtained by $p = F(X)$, where *F* is the assumed cumulative distribution function of the sample (random variable *X*) – often normal. But that assumed *F* is just an asymptotic approximation, for which the fit will be worst in the tails. Therefore $p < 0.025$ or $p > 0.975$ means that the PRNG has "failed the test at the 0.05 level".

4. Results obtained from Diehard tests and discussion

Diehard tests have requirements with precise format of the numbers whose randomness they test. Explicitly, the file of the numbers should be a binary file of a hexadecimal integer nonnegative numbers with approximately 11 MB size. There should be ten numbers in each row, about 2 870 000 numbers in the file and the maximum number in the file should be $\text{max} = 2^{32} - 1$.

As we mentioned previously, some of Diehard tests give *p*-value, and some of them are performed several times and the result from these tests is the ratio of the number of passed tests and the total number of tests. Therefore, we presented the results in separated tables depends on the kind of test output.

In the next tables we will present some of the obtained results of Diehard tests applied to the sequences generated by our proposed algorithm in [2].

In Table 1, we present the percentage of passed Diehard test for 14 sequences generated by our PRNG using different irrational numbers or same irrational number with different initial pointer *s* or with different initial length *n*. The bold line in the table separates the sequences obtained from the digits of non-transcendent irrational numbers from them obtained from the digits of transcendent irrational numbers. Note that almost all sequences pass more than 90% of the Diehard test. Exception is only the sequence obtained using the digits from the sin 1, where the percentage of passed tests is between 80% and 90%, but it is satisfactory.

Table 1 Success of Diehard tests

Irrational number	Seq. number	<i>n</i>	<i>s</i>	Time for sequence generation (in sec.)	Success of Diehard tests
ϕ	Seq. 1	10	1	630	91 %
$\sqrt{2}$	Seq. 2a	10	2	670	99 %
$\sqrt{2}$	Seq. 2b	10	1	678	98 %
$\sqrt{3}$	Seq. 3a	10	6	636	97 %
$\sqrt{3}$	Seq. 3b	12	2	653	92 %
$\sqrt{7}$	Seq. 4a	10	6	654	99 %
$\sqrt{7}$	Seq. 4b	10	3	662	91 %
sin 1	Seq. 5a	10	2	351	85 %
sin 1	Seq. 5b	10	4	478	87 %
sin 1	Seq. 5c	9	4	529	86 %
π	Seq. 6	10	8	485	94 %
<i>e</i>	Seq. 7a	10	2	586	90 %
<i>e</i>	Seq. 7b	10	1	577	94 %
ln 2	Seq. 8	10	3	627	95 %

In Table 2 and Table 3, we present the results from Diehard tests obtained from sequences generated from the digits of non-

transcendent numbers. In Table 2, we give the results of Diehard test which output is p -value and in Table 3, the results when the output is the ratio of the number of passed tests and the total number of tests. The red (bold) values in the tables mean that the sequence does not pass the corresponding test.

Table 2: Results from Diehard tests applied on the sequences generated by PGNG when irrational number is non-transcendental. Obtained p -values are presented

Seq. name	Seq. 1	Seq. 2a	Seq. 2b	Seq. 3a	Seq. 3b	Seq. 4a	Seq. 4b
Irr. number	φ	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{3}$	$\sqrt{7}$	$\sqrt{7}$
	p -value						
Birthday Spacings test	0.10	0.87	0.37	0.40	0.23	0.15	0.34
OPERM-5	0.15 0.14	0.82 0.12	0.08 0.57	0.91 0.80	0.29 0.70	0.31 0.72	0.93 0.49
Binary-31 test	0.34	0.85	0.78	0.32	0.68	0.32	0.73
Binary-32 test	0.79	0.36	0.32	0.86	0.64	0.72	0.46
Binary-6x8 test	0.89	0.971	0.54	0.82	0.83	0.57	0.09
Count-Stream test	0.55 0.68	0.91 0.28	0.64 0.68	0.24 0.23	0.36 0.28	0.52 0.50	0.23 0.70
Parking test	0.48	0.30	0.73	0.30	0.18	0.38	0.977
Minim. Distance test	0.99	0.09	0.82	0.57	0.004	0.51	0.32
3D Spheres	0.88	0.96	0.74	0.92	0.28	0.47	0.15
Squeeze test	0.004	0.71	0.14	0.45	0.53	0.96	0.17
O-SUM test	0.42	0.31	0.50	0.03	0.30	0.48	0.26
Run test	0.12 0.80 0.39 0.86	0.85 0.69 0.19 0.27	0.96 0.14 0.34 0.61	0.12 0.02 0.82 0.40	0.47 0.68 0.78 0.40	0.80 0.52 0.87 0.91	0.65 0.31 0.82 0.56
Craps test	0.49 0.81	0.92 0.24	0.26 0.41	0.24 0.03	0.39 0.92	0.75 0.30	0.99 0.24

Table 3: Results from Diehard tests applied on the sequences generated by PGNG when irrational number is non-transcendental. No. of passed tests / No. of total tests are presented

Seq. name	Seq. 1	Seq. 2a	Seq. 2b	Seq. 3a	Seq. 3b	Seq. 4a	Seq. 4b
Irr. number	φ	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{3}$	$\sqrt{7}$	$\sqrt{7}$
	No. of passed tests / total tests						
Bit stream test	20/20	19/20	20/20	19/20	20/20	20/20	20/20
OPSO	23/23	22/23	18/23	21/23	17/23	21/23	22/23
OQSO	26/28	28/28	28/28	28/28	25/28	25/28	28/28
DNA test	31/31	28/31	28/31	31/31	30/31	30/31	31/31
Count Bytes test	23/25	23/25	22/25	23/25	24/25	24/25	24/25

From the last two tables, we can conclude that the generated sequences passed almost all Diehard tests.

In Table 4 and Table 5, we present the results from Diehard tests obtained from sequences generated from the digits of transcendent irrational numbers.

Table 4: Results from Diehard tests applied on the sequences generated by PGNG when irrational number is transcendental. Obtained p -values are presented

Seq. name	Seq. 5a	Seq. 5b	Seq. 5c	Seq. 6	Seq. 7a	Seq. 7b	Seq. 8
Irr. number	$\sin 1$	$\sin 1$	$\sin 1$	π	e	e	$\ln 2$
	p -value						
Birthday Spacings test	0.40	0.05	0.17	0.56	0.03	0.08	0.24
OPERM-5	0.98 0.94	0.05 0.73	0.20 0.99	0.68 0.32	0.95 0.27	0.19 0.99	0.57 0.49
Binary-31 test	0.33	0.66	0.41	0.77	0.49	0.81	0.99
Binary-32 test	0.60	0.61	0.62	0.65	0.72	0.32	0.64
Binary-6x8 test	0.18	0.22	0.89	0.55	0.31	0.16	0.14
Count-Stream test	0.40 0.26	0.01 0.34	0.87 0.03	0.43 0.31	0.90 0.45	0.80 0.51	0.80 0.04
Parking test	0.006	0.19	0.98	0.71	0.27	0.34	0.69
Minim. Distance test	0.92	0.03	0.55	0.01	0.35	0.86	0.72
3D Spheres	0.49	0.09	0.99	0.86	0.78	0.46	0.15
Squeeze test	0.14	0.84	0.61	0.27	0.87	0.66	0.85
O-SUM test	0.99	0.003	0.36	0.31	0.61	0.52	0.82
Run test	0.53 0.08 0.97 0.52	0.25 0.93 0.39 0.013	0.76 0.34 0.84 0.21	0.10 0.61 0.13 0.73	0.42 0.74 0.97 0.48	0.36 0.28 0.39 0.71	0.25 0.65 0.44 0.51
Craps test	0.15 0.35	0.05 0.58	0.71 0.60	0.45 0.67	0.99 0.99	0.86 0.83	0.46 0.64

Table 5: Results from tests in Diehard battery, when irrational number is transcendental. Results are given with No. of passed tests / No. of total tests.

Seq. name	Seq. 5a	Seq. 5b	Seq. 5c	Seq. 6	Seq. 7a	Seq. 7b	Seq. 8
Irr. number	$\sin 1$	$\sin 1$	$\sin 1$	π	e	e	$\ln 2$
	No. of passed tests / total tests						
Bit stream test	19/20	19/20	19/20	16/20	18/20	20/20	19/20
OPSO	19/23	21/23	20/23	20/23	21/23	21/23	23/23
OQSO	27/28	28/28	25/28	28/28	27/28	27/28	24/28
DNA test	30/31	30/31	29/31	31/31	30/31	26/31	31/31
Count Bytes test	20/25	24/25	23/25	25/25	23/25	23/25	24/25

Analyzing the results from Table 4 and Table 5, we can conclude that the sequences generated from the digits of transcendent irrational numbers also passed almost all Diehard tests. As we concluded from Table 1, exception is only the sequence obtained using the digits from the $\sin 1$, where the results are a little

bit worse, but these sequences also passed more of the considered tests.

5. Conclusions

In this paper, we generate sequences using PRNG based on digits of different irrational numbers. Our goal is to check if the kind of irrational numbers (non-transcendent or transcendent) has influence to the randomness of generated sequences. Our expectations were that the transcendent numbers will give better results than non-transcendent ones, but the results reject our expectations. They confirm that almost all tested irrational numbers are good for using in our PRNGs and there is not a significant difference in the randomness of the generated sequences in the both cases.

Acknowledgements

This work was partially financed by the Faculty of Computer Science and Engineering at the "Ss. Cyril and Methodius" University in Skopje.

References

1. Chudnovsky D., G. Chudnovsky The computation of classical constants, in: Proceedings of the National Academy of Sciences of the USA 86, 1989, pp. 8178--8182
2. Dimitrievska Ristovska, V. Pseudo random generator based on irrational numbers, in: Trajanov D., V. Bakeva (eds.): ICT-Innovations 2017, Data Driven Innovations, Conference Web Proceedings, 2017, pp. 105--113
3. Esmeijaud-Bonnardel M. Etude Statistique des D'écimales de Pi, Revue Francaise de Techniques Informatiques 8 (4), 1965, pp. 295--306
4. L'Ecuyer P. History of uniform random number generation, Chan W. K. V., D'Ambrogio A., Zacharewicz G., Mustafee N., Wainer G., and Page E., eds.: DIRO, GERAD, and CIRRELT, Proceedings of the 2017 Winter Simulation Conference, 2017
5. Marsaglia G., W. W. Tsang Some difficult-to-pass tests of randomness, Journal of Statistical Software, Volume 7, Issue 3, 2002
6. Metropolis N., G. Reitwiesner, J. von Neumann Statistical Treatment of Values of First 2000 Decimals Digits of E and Pi Calculated on the ENIAC, Mathematical Tables and Other Aids to Computation 4, 1950
7. Pathria R. K. A Statistical study of randomness among the first 10000 digits of Pi, Mathematics of Computation 16, 1962, pp. 188--197
8. Rogers I., G. Harrell, J. Wang Using Pi digits to generate random numbers: A Visual and statistical analysis, Int'l Conf. Scientific computing | CSC'15 |
9. Yee A. J. Y-Cruncher—A Multi-Threaded Pi-Program, 2017

PARAMETRIC INDUCED INSTABILITIES OF BOSONS IN MAGNETAR'S CRUST

Prof. Dr. Dariescu M.-A.¹, PhD. student Mihu D.-A.¹, Prof. Dr. Dariescu C.¹
Alexandru Ioan Cuza University of Iasi, Romania
E-mail: marina@uaic.ro, denisa.mihu0@gmail.com

Abstract: In the present work, we are discussing the Klein-Gordon equation describing relativistic spinless particles evolving in the (stationary) magnetar's crust. With the wave function expressed in terms of Mathieu's functions, we compute first-order transition amplitudes, pointing out the role of the strong magnetic induction in transitions to states which may be characterized by values of the model's parameters in instability bands.

KEYWORDS: KLEIN-GORDON EQUATION, MATHIEU'S EQUATION, BESSEL FUNCTIONS, PERTURBATION THEORY, MAGNETARS

1. Introduction

In 1992, Duncan and Thompson introduced the term of magnetar for a type of neutron star with massively boosted magnetic fields whose decay powers the emission of high-energy electromagnetic radiation [1]. The configuration of the background field inside has been extensively worked out and is not unique. Inspired by some previous investigations [2-4], we are assuming a radial magnetic induction, parallel to the surface of a plane slab with finite thickness in the z -direction.

As for the matter content, in the magnetar's crust and core, one may expect the existence of various particles, including boson condensates, which might have a significant influence on the star properties. In this respect, the kaons are seen as best candidates, besides nucleons and leptons [5, 6].

In the present work, we are extending our previous results on boson's wave function, solution to the Klein-Gordon equation [3, 4]. Thus, within a perturbative approach, we are discussing special cases for parameters' ranges which are leading to non-trivial quantization laws and to a favored distribution of the magnetic field inside the crust, corresponding to a high probability of transitions.

With reference to the Mathieu's equation, challenging aspects that surround the interest around it and its solutions are more and more invoked as a solid motivation as these manifest in a rich manner, both theoretically – mostly, from algebraic point of view – and practically, i.e. from the applicative side. Concerning the latter, the range of physical situations in which Mathieu functions appear is extremely extended, from elastic wave equations worked in elliptical boundary geometries as it is the case of resonators or waveguides, the motion of quantum particles in a periodic type potential or the stability of floating vessels for harmonious coherent waves of various frequencies and swings [7], to some modern applications as the physics of a capacitor microphone, ferromagnetic substance manifesting elastic oscillations [8], the particle's behavior in different systems of electromagnetic traps [9] or the quantum dynamics of the electrons within the free electron lasers (FEL) [10].

From algebraic point of view, manipulating Mathieu functions is not at all an easy task. The algebra behind the Mathieu functions is extremely intricate, still admitting a range of unsolved mathematical and computational issues. For instance, if we take into account the convergence of Mathieu functions, some authors report slow convergence or the lack of convergence of specific representations of Mathieu functions [11]. Some codes [12], which led to results

with single-precision accuracy (7 decimal places), proved to respond negatively to the action of extending them to get a double-precision accuracy (15 decimal places).

An ardent area of research on Mathieu functions focuses on deriving expressions for the dependence of the *Characteristic Exponent* γ on the Mathieu equation parameters, α, β . Due to the physical significance of the Characteristic Exponent, which proves to be intimately connected to the wavelike nature of the physical phenomena modelled by the Mathieu equation, finding relations of the form $\gamma = \gamma(\alpha, \beta)$ constitutes an outstanding mathematical approach. In literature, relations of this type written in power series of β , can be found both for integer [13, 14] and noninteger values of γ [15].

We point out that the literature is pretty rich in treating specific algebraic aspects of Mathieu functions and the main difficulties within the computational analysis [16, 17, 18, 19, 20].

To put it concisely, developing packages for calculating Mathieu functions and operating with them represents an instigation for the modern research being an open territory for algebraic and computational explorations.

2. Magnetic field configuration in the crust

Let us consider, besides the vertical induction B_z , a radial component parallel to the surface of the slab and vanishing at $z=0$ and $z=L$, of the form $B_\rho = b(t)\sin(\kappa z)$ [3, 4]. Working in cylindrical coordinates, the potential component

$$(1) \quad A_\phi = \frac{B_0}{\kappa} \exp\left[-\frac{\kappa^2}{\sigma} t\right] \cos(\kappa z)$$

is generating the following electromagnetic field configuration

$$(2) \quad \begin{aligned} E_\phi &= -\frac{\partial A_\phi}{\partial t} = \frac{\kappa B_0}{\sigma} \exp\left[-\frac{\kappa^2}{\sigma} t\right] \cos(\kappa z) \\ B_\rho &= -\frac{\partial A_\phi}{\partial z} = B_0 \exp\left[-\frac{\kappa^2}{\sigma} t\right] \sin(\kappa z) \\ B_z &= \frac{1}{\rho} A_\phi = \frac{B_0}{\kappa \rho} \exp\left[-\frac{\kappa^2}{\sigma} t\right] \cos(\kappa z) \end{aligned}$$

which also satisfy the Maxwell equations.

In the following, we are assuming that the time variable is much less the *characteristic time*, $t \ll \tau = \sigma / \kappa^2 \ll 1 \text{ Myr}$, which is comparable to the average Ohmic timescale, so that the exponential function in (2) can be set to one. However, soon after the crust

forms, the magnetic field is freezing and such objects can be treated as being stationary, with poloidal and toroidal fields of the same order of magnitude [21, 22].

In order to study the semi-classical behavior of the charged scalar field, we start with the Klein-Gordon equation

$$(3) \quad \left[\eta^{ij} D_i D_j - m_0^2 \right] \Phi = 0,$$

where the gauge derivatives are expressed in terms of the four-potential components as $D_i = \partial_i - iqA_i$.

For the essential component

$$(4) \quad A_\phi = \frac{B_0}{\kappa} \cos(\kappa z),$$

the Klein-Gordon equation, in cylindrical coordinates,

$$\left[\Delta - \partial_t^2 - m_0^2 \right] \Phi = \frac{2iq}{\rho} A_\phi \frac{\partial \Phi}{\partial \phi} + q^2 A_\phi^2 \Phi,$$

with the variables separation

$$(5) \quad \Phi = \phi(\rho, z) e^{im_\phi \phi} e^{-i\omega t},$$

leads to the following differential equation for $\phi(\rho, z)$,

$$(6) \quad \frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \frac{\partial \phi}{\partial \rho} \right) + \frac{\partial^2 \phi}{\partial z^2} + \left[\omega^2 - m_0^2 - \frac{m^2}{\rho^2} - \left(\frac{qB_0}{\kappa} \right)^2 \cos^2(\kappa z) \right] \phi = -\frac{2mqB_0}{\kappa \rho} \cos(\kappa z) \phi$$

By identifying the potential operator

$$(7) \quad \hat{V} = -\frac{2mqB_0}{\kappa \rho} \cos(\kappa z),$$

the equation (6) gets the standard form employed in the Perturbation Theory, $\hat{\mathbf{D}}\phi = V\phi$, where $\phi = \psi + \chi$.

The zero-order equation

$$\frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \frac{\partial \psi}{\partial \rho} \right) + \frac{\partial^2 \psi}{\partial z^2} + \left[\omega^2 - m_0^2 - \frac{m^2}{\rho^2} - \left(\frac{qB_0}{\kappa} \right)^2 \cos^2(\kappa z) \right] \psi = 0$$

for $\psi(\rho, z) = F(\rho)T(z)$, splits into the following system

$$\frac{d^2 F}{d\rho^2} + \frac{1}{\rho} \frac{dF}{d\rho} + \left[\omega^2 - m_0^2 - \frac{m^2}{\rho^2} \right] F = 0, \quad (a)$$

$$(8) \quad \frac{d^2 T}{dz^2} + \left[p_z^2 - \frac{1}{2} \left(\frac{qB_0}{\kappa} \right)^2 - \frac{1}{2} \left(\frac{qB_0}{\kappa} \right)^2 \cos(2\kappa z) \right] T = 0. \quad (b)$$

The first equation in (8) is satisfied by the Bessel functions

$$(9) \quad F_m(\rho) = J_m(P\rho), \text{ with } P \equiv \sqrt{\omega^2 - m_0^2 - \frac{m^2}{\rho^2}},$$

while the second one can be identified with the Mathieu-type equation [23]

$$(10) \quad y'' + [\alpha - 2\beta \cos(2\zeta)] y = 0,$$

of parameters

$$(11) \quad \alpha = \frac{p_z^2}{\kappa^2} - \frac{1}{2} \left(\frac{qB_0}{\kappa} \right)^2, \quad \beta = \frac{1}{4} \left(\frac{qB_0}{\kappa} \right)^2,$$

and variable $\zeta = \kappa z$. The general solutions of the equation (10),

$$(12) \quad T(\alpha, \beta, \zeta) = \{ \text{Mathieu}C[\alpha, \beta, \zeta], \text{Mathieu}S[\alpha, \beta, \zeta] \}$$

have been discussed in detail in our previous papers [3, 4], especially with respect to their stability. In general, the Mathieu's

functions are of the form $f(z) \propto e^{i\gamma z} u(z)$, where the Mathieu Characteristic Exponent (MCE), γ , may be real or imaginary, depending on the values of the model parameters.

Nevertheless, we find worth stressing some interesting mathematical properties of these solutions. Thus, the even and odd solutions, can be written as the Fourier series [23]

$$(13.a) \quad \text{Mathieu}C = ce_{2n} = \sum_{j=0}^{\infty} A_{2j}^{2n}(\beta) \cos(2j\zeta)$$

$$(13.b) \quad \text{Mathieu}S = se_{2n+1} = \sum_{j=1}^{\infty} B_{2j+1}^{2n+1}(\beta) \sin((2j+1)\zeta)$$

In explicit calculations, one may employ the periodic expressions of the Mathieu's functions (13.a), valid in the first order in β

$$\begin{aligned} ce_0(z, \beta) &= \sum_{j=0}^{\infty} A_{2j}^0(\beta) \cos(2jz) = A_0^0 + \sum_{j=1}^{\infty} A_{2j}^0(\beta) \cos(2jz) \\ &= \frac{1}{\sqrt{2}} \left[-1 + J_0(\sqrt{\beta} \exp(-iz)) + J_0(\sqrt{\beta} \exp(iz)) \right] \\ &\approx \frac{1}{\sqrt{2}} \left[1 - \frac{\beta}{2} \cos(2z) + \dots \right] \end{aligned}$$

$$\begin{aligned} ce_2(z, \beta) &= \sum_{j=0}^{\infty} A_{2j}^2 \cos(2jz) = A_0^2 + A_2^2 \cos(2z) + \\ &+ \sum_{j=1}^{\infty} A_{2(j+1)}^2(\beta) \cos(2(j+1)z) \\ &= \frac{\beta}{4} + \frac{4}{\beta} \left[J_2(\sqrt{\beta} \exp(-iz)) + J_2(\sqrt{\beta} \exp(iz)) \right] \\ &\approx \cos(2z) + \frac{\beta}{12} (3 - \cos(4z)) + \dots \end{aligned}$$

where $J_n(\sqrt{\beta} \exp(\pm iz))$ are the Bessel functions and we have used the formulas [24]

$$\begin{aligned} A_{2j}^0 &\approx \frac{2(-1)^j}{(j!)^2} \left(\frac{\beta}{4} \right)^j A_0^0; \quad A_{2(j+1)}^2 \approx \frac{2(-1)^j}{j!(j+2)!} \left(\frac{\beta}{4} \right)^j; \\ A_0^0 &= \frac{1}{\sqrt{2}}; \quad A_0^2 = \frac{\beta}{4}. \end{aligned}$$

3. First-order transition amplitude

With the wave function

$$(14) \quad \psi_m = J_m(P\rho) T(\alpha, \beta, \zeta) e^{im_\phi \phi} e^{-i\omega t},$$

one may compute the azimuthal current density component, of quantum origin, which is sustaining the magnetization current,

$$\begin{aligned} j_\phi &= -\frac{iq}{\rho} (\psi_m^* \partial_\phi \psi_m - \psi_m \partial_\phi \psi_m^*) - 2q^2 |\psi_m|^2 A_\phi \\ &= 2q J_m^2 |T|^2 \left[\frac{m}{\rho} - \frac{qB_0}{\kappa} \cos(\kappa z) \right]. \end{aligned}$$

In the transition amplitude

$$(15) \quad \mathcal{A} = \int \psi_m^* V \psi_m \rho d\rho dz dt = -\frac{4\pi mb}{\kappa} I_2$$

where $b \equiv qB_0$, the first integral can be written in terms of the hypergeometric function F_{12} as [23]

$$(16) \quad I_1 = \int_0^\infty J_m(P'\rho) J_m(P\rho) d\rho = \frac{1}{P'} \left(\frac{P}{P'}\right)^m \frac{\Gamma(m+1/2)}{\sqrt{\pi}\Gamma(m+1)} F_{12}\left[\frac{1}{2}, m+\frac{1}{2}, m+1, \left(\frac{P}{P'}\right)^2\right],$$

with $P' > P$. Once we fix the quantum number m , the integral above is expressed as a combination of *EllipticE* and *EllipticK* functions, of variable $(P/P')^2$.

The second integral in (15), i.e.

$$(17) \quad I_2 = \int T^*(\alpha', \beta, \kappa z) \cos(\kappa z) T(\alpha, \beta, \kappa z) dz,$$

should be analyzed for specific ranges of the particle momentum along Oz .

For small values of the parameter β , the theory of the Mathieu's functions is well understood. The characteristic values $\alpha_n(\beta)$ are important since they yield periodic solutions and they separate the regions of stability. The resonance condition $\alpha_n \approx n^2$ leads to the momentum quantization relation

$$(18) \quad p_z^2 \approx n^2 \kappa^2 + \frac{b^2}{2\kappa^2},$$

and to a quantized variable in the Bessel functions (9), i.e.

$$(19) \quad P_n^2 = \omega^2 - m_0^2 - \frac{b^2}{2\kappa^2} - n^2 \kappa^2.$$

For the particle momentum along Oz in the first band of instability

$$(20) \quad \kappa^2 - \frac{b^2}{4\kappa^2} < p_z^2 < \kappa^2 + \frac{3b^2}{4\kappa^2},$$

the imaginary part of the MCE comes into play, $\text{Im}(\gamma) \propto \beta/2$, being responsible for the exponentially growing of the oscillatory wave functions.

By inspecting the stability chart of the Mathieu's functions, one can notice that, for increasing β , the stable regions situated between the characteristic curves $\alpha_n(\beta)$ become more and more narrow and their width is decreasing exponentially to practically discrete eigenvalues. For $\beta > 1$, one has a broad resonance and the solutions (12) are oscillating (along Oz) much faster than the electromagnetic field components, the ratio between the frequencies being $\sqrt{\beta} = b/(2\kappa^2)$.

In the case of a magnetar, whose magnetic field is extremely strong, $B_0 \approx 10^{10} - 10^{12} \text{ (T)}$, the parameter β , proportional to the square of the magnetic induction, gets very large values. Thus, for $0 < \alpha < \beta$, i.e.

$$(21) \quad \frac{b^2}{2\kappa^2} < p_z^2 < \frac{3b^2}{4\kappa^2},$$

one has to use the following asymptotic expansion for the characteristic values [25]

$$(22) \quad \alpha_n \approx -2\beta + 2(2n+1)\sqrt{\beta}.$$

In our case, the above relation turns into

$$(23) \quad \alpha_n = \frac{b}{\kappa^2} \left[(2n+1) - \frac{b}{2\kappa^2} \right],$$

leading to the Landau-type quantization law for the particle momentum along Oz

$$(24) \quad p_z^2 \approx (2n+1)b,$$

and to a quantized variable in the Bessel functions computed with

$$(25) \quad P_n^2 = \omega^2 - m_0^2 - (2n+1)b.$$

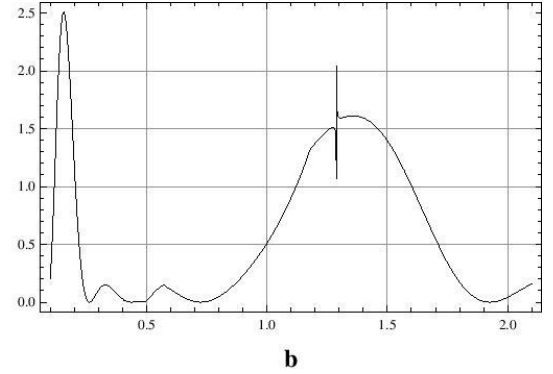


fig. 1 The absolute value of (17) as a function of $b = b/\kappa^2$

In the figure 1, for the transition between $n=2$ and $n'=n+1$, we are giving, in dimensionless units, the graphical representation of the absolute value of (17), as a function of the parameter $b = b/\kappa^2$, characterizing the strength of the magnetic field inside the crust, for a given κ . For both α_n and α_{n+1} positive quantities computed with (23), one may notice the peaks separated by zero minima, corresponding to high probabilities and zero probabilities, respectively.

Let us point out that, for an electron with the momentum along Oz given by (24), the distance between the corresponding energy levels is $W = \sqrt{2\hbar q B_0 c^2} \approx 1.06 \text{ (MeV)}$, for $B_0 = 10^{10} \text{ (T)}$. This ultra-relativistic particle promoted from the level with $n=0$ to the one with $n=1$ can produce, by de-excitation photons which will directly convert to electron-positron pairs.

Thus, one may compute only transitions between $n=0$ and $n'=1$, for which the characteristic values are given by

$$(26) \quad \alpha_0 \approx \frac{b}{\kappa^2} \left[1 - \frac{b}{2\kappa^2} \right]; \quad \alpha_1 \approx \frac{b}{\kappa^2} \left[3 - \frac{b}{2\kappa^2} \right],$$

and one can find similar peaks and minima as in the figure 1.

For $\beta \rightarrow \infty$, the oscillatory behavior of the Mathieu functions happen in a shrinking neighborhood of $\kappa z = \zeta = \pi/2 + O(\beta^{-1/2})$. In explicit calculations, it is very useful to express the Mathieu function $\text{MathieuC}[\alpha_n, \beta, \zeta]$ in terms of parabolic cylinder functions and their derivatives, of variable $t = 2\sqrt{h}\cos\zeta$, with $h = \sqrt{\beta}$, as [25]

$$\text{Mathieu}C \approx C_n \left\{ D_n(t) \sum_s \frac{A_s(t)}{h^s} + D'_n(t) \sum_s \frac{B_s(t)}{h^s} \right\},$$

where

$$D'_n = \frac{n}{2} D_{n-1} - \frac{1}{2} D_{n+1}, \quad C_n \approx \left[\frac{\pi \sqrt{\beta}}{2(n!)^2} \right]^{1/4}$$

and the first terms of the polynomials $A_s(t)$ and $B_s(t)$ are

$$A_0 = 1, \quad A_l(t) = \frac{t^2}{2^6} = \frac{\sqrt{\beta}}{2^4} \cos^2 \zeta,$$

$$B_l(t) = \frac{1}{2^5} [t^3 - (2n+1)t] \approx \frac{\beta^{3/4}}{4} \cos^3 \zeta.$$

4. Conclusions

Within a perturbative approach, in the present paper, we are dealing with the Klein-Gordon equation, describing the bosons evolving in the (stationary) magnetar's crust, endowed with the electromagnetic configuration (2). The solution to the zero-order equation, expressed in terms of Bessel and Mathieu functions, is used for computing first-order transition amplitudes, for physically important ranges of model's parameters.

The width of the resonance bands are solely controlled by the parameter β in (11), which is characterizing the magnitude and the distribution of the magnetic field inside the crust.

A special attention is given to ultra-relativistic particles in ultra-strong magnetic field ($\beta > \alpha > 0$), when the particle momentum along Oz is satisfying the quantization law (24). The probability of the transition between two adjacent n values is discussed for different ranges of the model's parameters $b = qB_0$ and κ , characterizing the magnitude and the distribution of the magnetic field inside the crust. Thus, for $b/\kappa^2 < 2n+1$, a series of peaks, corresponding to high probabilities, can be noticed in the figure 1.

Finally, we mention that for particles moving in an ultra-strong magnetic field so that the α parameter defined in (11) is negative, the amplitude function is exponentially growing along Oz and there is only a narrow interval for the particle momentum p_z , i.e.

$|\alpha| < \beta^2/2$, meaning

$$(27) \quad \frac{1}{2} \left(\frac{b}{\kappa} \right)^2 - \frac{1}{32\kappa^2} \left(\frac{b}{\kappa} \right)^4 < p_z^2 < \frac{1}{2} \left(\frac{b}{\kappa} \right)^2$$

which corresponds to a stable region in the Ince-Strutt diagram [26]. The present work can be extended in several directions, as for example to consider instead of the Klein-Gordon equation, the Dirac equation for relativistic fermions, leading to Mathieu's functions of complex variable and parameters [27].

These situations are very challenging and they are rarely discussed in literature. The existing packages written in MAPLE and MATHEMATICA are insufficient and some routines have been recently developed in [28]. It is worth to mention that the first study which considered a purely imaginary β parameter ($\beta = is$, $s \in \mathbb{R}$) was initiated by Mulholland and Goldstein [29] who deduced an approximative law governing the position of branching points in terms of a quadric growth. The authors detected the first branching point to be $s_0 \approx 1.47$. For a more detailed analysis on the branching points we recommend [30, 31].

Acknowledgement

This work was supported by a grant of Minister of Research and Innovation, CNCS - UEFISCDI, project number PN-III-P4-ID-PCE-2016-0131, within PNCDI III.

References

- [1] Duncan R. C. and Thompson C., *Astrophys. J.* 392 (1992) L9
- [2] Wareing C. J. and Hollerbach R., *Astron. Astrophys.* 508 (2009) L39;
- [3] Dariescu M. A., Dariescu C. and Buhucianu O., *Chinese Phys. Lett.* 28 (2011) 010303;
- [4] Dariescu C. and Dariescu M. A., *Mod. Phys. Lett. A* 26 (2011) 1245;
- [5] Kaplan D. B. and Nelson A. E., *Phys. Lett. B* 175 (1986) 57;
- [6] Nelson A. E. and Kaplan D. B., *Phys. Lett. B* 192 (1987) 193;
- [7] Allievi A. and Soudack A., *Int. J. Control* 51 (1990) 139;
- [8] Phelps III F. M., Hunter J. H., Jr., *Am. J. Phys.* 34 (1966) 533;
- [9] Ruby L., *Am. J. Phys.* 64 (1996) 39;
- [10] Procida L., Lee H. -W., *Opt. Commun.* 49 (1984) 201;
- [11] Erricolo D., *IEEE Antennas Wireless Propagat. Letters* 2 (2003) 58;
- [12] Zhang, S. J., J. M. Jin, *Computation of Special Functions*, New York, Wiley (1996);
- [13] Hill G. W., *Acta Math.* 8, 1 (1886) 1;
- [14] Campbell, R., *Theorie Generale de l'Equation de Mathieu* Paris, Masson and Co. (1955);
- [15] Tamir T., Wang H. C., *National Bureau of Standards - B. Mathematics and Mathematical Physics* 69B (1965) 101;
- [16] Erricolo D., Uslenghi P. L. E., Elnour B., *Electromagnetics*, vol. 26 (2006) 57;
- [17] Frenkel D. and Portugal R., *J.Phys. A: Math. Gen.* 34 (2001) 3541;
- [18] Toyama N. and Shogen K., *IEEE Trans. on Antennas and Propagation* AP-32 (1984) 537;
- [19] Kokkorakis G. C., Roumeliotis J. A., *Mathematics of Computation* 70 (2000) 1221;
- [20] Alhorgan F. A., *ACM Trans. on Math. Software (TOMS)* 26 (2000) 390;
- [21] Rheinhardt M. and Geppert U., *Phys. Rev. Lett.* 88 (2002) 101103;
- [22] Braithwaite J. and Spruit H., *Astron. Astrophys.* 450 (2006) 1097;
- [23] Gradshteyn, I. S., and I. M. Ryzhik, *Table of Integrals, Series and Products*, New York, Academic Press (1965);
- [24] <http://dlmf.nist.gov/>;
- [25] Ogilvie K. and Olde Daalhuis A. B., *Sigma* 11 (2015) 095;
- [26] Simakhina S. V. and Tierb C., *Applied Mathematics and Computation* 162 (2005) 639;
- [27] Dariescu M. A., Dariescu C., *Mod. Phys. Lett. A* 28 (2013) 1350157;
- [28] Blose E. N. et al., *Phys. Rev. A* 91 (2015) 012501;
- [29] Mulholland H. P., Goldstein S., *Phil. Mag.* 8 (1929) 834;
- [30] Hunter C. et al., *Stud. Appl. Math.* 64 (1981) 113.

ON THE USE OF CONFORMING AND NONCONFORMING RECTANGULAR FINITE ELEMENTS FOR EIGENVALUE APPROXIMATIONS

Assoc. Prof. Racheva M. Dsc.¹, Prof. Andreev A. Dsc.^{1,2},
 Technical University of Gabrovo, Bulgaria¹
 Institute of Information and Communication Technologies – BAS, Bulgaria²
 milena@tugab.bg

Abstract: The paper deals with some combinations of conforming and nonconforming rectangular finite elements in order to obtain two-sided bounds of eigenvalues, applied to second-order elliptic operator. The aim is to use the lowest possible order finite elements. Namely, the combination of serendipity conforming and rotated bilinear nonconforming elements is considered in details. This work continues some recent researches of the authors concerning eigenvalue approximations. Computational aspects of the used algorithm are also discussed. Finally, results from numerical experiments are presented.

Keywords: RECTANGULAR FINITE ELEMENTS, CONFORMING/NONCONFORMING ELEMENTS, EIGENVALUE APPROXIMATION, TWO-SIDED BOUND

2010 Mathematical Subject Classification: 65N25, 65N30

1. Introduction and Preliminaries

We consider a bounded polygonal domain Ω in R^2 with boundary $\Gamma = \partial\Omega$. Given an integer $m \geq 0$, we use the Sobolev space $H^m(\Omega)$ with a norm $\|\cdot\|_{m,\Omega}$ and (\cdot, \cdot) denotes the $L_2(\Omega)$ – inner product.

Consider the following weak form eigenvalue model problem: Find a number $\lambda \in R$ and a function $u \in V \equiv H_0^1(\Omega)$, $\|u\|_{0,\Omega} = 1$ such that

$$a(u, v) = \lambda(u, v), \quad \forall v \in V, \quad (1)$$

where

$$a(u, v) = \iint_{\Omega} \nabla u \cdot \nabla v \, dx \, dy \quad \forall u, v \in V.$$

This problem has a countable sequence of real and positive eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \dots$ and the corresponding eigenfunctions u_1, u_2, \dots satisfy $(u_i, u_j) = \delta_{ij}$, $i, j \geq 1$.

Let $\{\tau_h\}$ be a family of a rectangulations of Ω which satisfies the usual regularity conditions (see [1]), i.e. there exists a constant $\sigma > 0$ such that $h_K / \rho_K \leq \sigma$, where h_K is the diameter of the rectangle $K \in \tau_h$ and ρ_K being the diameter of the largest circle contained in K . Then we denote $h = \max_{K \in \tau_h} h_K$.

So, we introduce a conforming finite element space $V_h \subset V$ based on the partition τ_h . Then, the corresponding approximation of (1) is: Find a number $\lambda_h \in R$ and a function $u_h \in V_h$, $\|u_h\|_{0,\Omega} = 1$ such that

$$a(u_h, v_h) = \lambda_h(u_h, v_h), \quad \forall v_h \in V_h. \quad (2)$$

For second- and fourth-order self-adjoint elliptic operator the eigenvalues computed using standard conforming finite element method (FEM) are always above the exact ones. This fact comes from minimum-maximum characterization of the eigenvalues (see for example [2]). So that, it is an important and nontrivial problem to find methods giving lower bounds of the eigenvalues [3]. Herein, we claim a contribution to this specific field. More precisely, this paper is an extension of the authors' research done in [4] and [5].

First, let us consider nine-point conforming rectangular finite element. For any test function v and $K \in \tau_h$ the degrees of freedom

which we will use are (see Fig. 1(a)) $v(a_j)$ and $\frac{1}{|l_j|} \int_{l_j} v(s) \, ds$ and $\frac{1}{|K|} \int_K v(x, y) \, dx \, dy$, where $a_j, j=1,2,3,4$ are the vertices and $l_j, j=1,2,3,4$ are the edges of K .

Then, the finite element space V_h is defined by (see [6]):

$$V_h = \{v \in H^1(\Omega) : v|_K = \text{span}\{1, x, y, xy, x^2, y^2, x^2y, xy^2, x^2y^2\}\}$$

We can use also eight-point serendipity conforming finite element (see Fig. 1(b)). In this case:

$$V_h = \{v \in H^1(\Omega) : v|_K = \text{span}\{1, x, y, xy, x^2, y^2, x^2y, xy^2\}\}$$

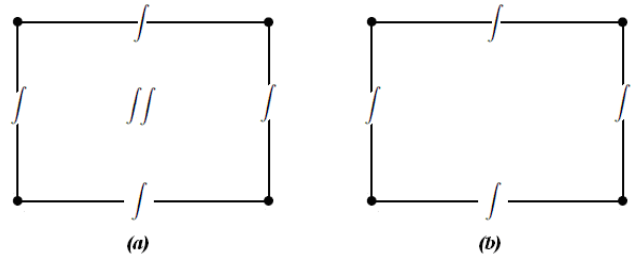


Fig. 1 (a) 9-point rectangle; (b) 8-point serendipity rectangle

Now we introduce non-conforming finite element spaces \tilde{V}_h related to the same partition τ_h . We also introduce mesh-dependent bilinear form

$$a_h(u, v) = \sum_{K \in \tau_h} a_K(u, v), \quad u, v \in V,$$

where

$$a_K(u, v) = \iint_K \nabla u \cdot \nabla v \, dx \, dy.$$

Obviously, in case of conforming FEM, $a(\cdot, \cdot)$ and $a_h(\cdot, \cdot)$ coincide.

The space \tilde{V}_h consists of Rannacher-Tourek (i.e. rotated bilinear) nonconforming finite elements denoted by Q_I^{rot} (Fig. 2(a)). Its degrees of freedom are the integral values at the edges of rectangle. Another case for construction of the space \tilde{V}_h is to use the extension of Rannacher-Tourek element (so-called extended rotated bilinear element) denoted by EQ_I^{rot} . The degrees of freedom for this element are the integral values at the edges of the rectangle and the integral value over the rectangle (see Fig. 2(b)).

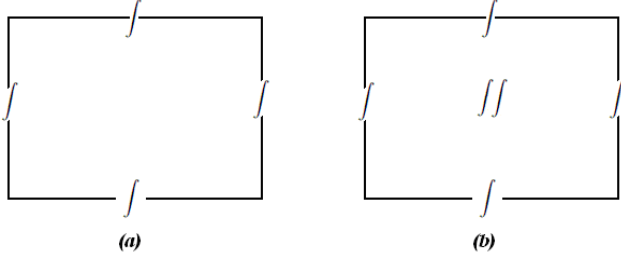


Fig. 2 (a) Q_I^{rot} nonconforming element; (b) EQ_I^{rot} nonconforming element

Let us define a so-called “nonconforming interpolation operator” $\tilde{i}_h: V \rightarrow \tilde{V}_h$ such that

$$a_h(v - \tilde{i}_h v, \tilde{v}_h) = 0, \quad \forall v \in V, \tilde{v}_h \in \tilde{V}_h. \quad (3)$$

2. Main Result

Our result is based on determining a couple of finite element spaces (V_h, \tilde{V}_h) which verify two sided bounds of eigenvalues λ using the following algorithm:

Algorithm

1. Find (λ_h, u_h) from (2) by means of conforming finite element space V_h ;
2. Construct nonconforming space \tilde{V}_h from V_h by eliminating some degrees of freedom in such a way that after obtaining of $\tilde{i}_h u_h \in \tilde{V}_h$ the condition (3) to be fulfilled;
3. Calculate the number $\tilde{\lambda}_h = a_h(\tilde{i}_h u_h, \tilde{i}_h u_h)$ which approximates the exact eigenvalues λ asymptotically from below.

The case when V_h consists of nine-point quadratic conforming finite elements and \tilde{V}_h contains EQ_I^{rot} -elements is proved by the authors in [4] (see Theorem 1). Later on, we will use the same notations as in [4].

Here, our aim is to prove the validity of the algorithm when V_h use the eight-point serendipity elements and $\tilde{V}_h \leftarrow Q_I^{rot}$ -nonconforming elements, respectively.

Thus the nonconforming finite element space is defined by (see [6]):

$$\tilde{V}_h = \{v \in L_2(\Omega) : v|_K = \text{span}\{1, x, y, x^2 - y^2\}\}$$

Also, for any $v \in V$:

$$\int_{l_j} \tilde{i}_h v(s) ds = \int_{l_j} v(s) ds, \quad (4)$$

with edges $l_j, j=1,2,3,4$.

Introducing the notation $|v_h|_h^2 = a_h(v_h, v_h)$ for any $v_h \in V + \tilde{V}_h$, for our considerations, it is enough to assume the following interpolation inequality for $v \in V$:

$$|\tilde{i}_h v - v|_h \geq C h^{3/2}. \quad (5)$$

The following theorem contains a main result of the paper.

Theorem 1. Let (λ_h, u_h) be an approximation of the exact eigenpair (λ, u) obtained from (2) by means of eight-point serendipity elements. If the inequality (5) is valid, the number

$$\tilde{\lambda}_h = a_h(\tilde{i}_h u_h, \tilde{i}_h u_h)$$

approximates λ from below when h is small enough, so that two-sided bounds of λ are obtained:

$$\tilde{\lambda}_h \leq \lambda \leq \lambda_h. \quad (6)$$

Proof. We adopt the following notations for partial derivatives:

$$\partial_x \bullet = \frac{\partial \bullet}{\partial x}, \quad \partial_y \bullet = \frac{\partial \bullet}{\partial y} \quad \text{and so on.}$$

Let $v_h \in \tilde{V}_h$ and $v \in V$. Then

$$\begin{aligned} a_h(\tilde{i}_h v - v, v_h) &= \sum_{K \in \tau_h} \iint_K \nabla(\tilde{i}_h v - v) \cdot \nabla v_h \, dx \, dy \\ &= \sum_{K \in \tau_h} \iint_K (\partial_x(\tilde{i}_h v - v) \partial_x v_h + \partial_y(\tilde{i}_h v - v) \partial_y v_h) \, dx \, dy. \end{aligned} \quad (7)$$

We choose K_0 to be a fixed reference rectangle. Then the result on arbitrary $K \in \tau_h$ will be transformed from K_0 using an affine transformation.

The equations of the edges of K_0 are:

$$l_{1,3} : y - y_0 = \mp \frac{h_2}{2}; \quad l_{2,4} : x - x_0 = \pm \frac{h_1}{2},$$

where (x_0, y_0) is the center of K_0 (Fig. 3) and $h = \sqrt{h_1^2 + h_2^2}$.

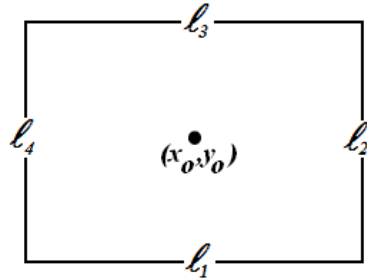


Fig. 3 Rectangular reference element K_0

Having in mind that \tilde{V}_h is an incomplete biquadratic polynomial space on K_0 , we can write

$$v_h(x, y) = v_h(x_0, y_0) + (x - x_0) \partial_x v_h(x_0, y_0) + (y - y_0) \partial_y v_h(x_0, y_0)$$

$$+ \frac{1}{2} (x - x_0)^2 \partial_{xx} v_h(x_0, y_0) + \frac{1}{2} (y - y_0)^2 \partial_{yy} v_h(x_0, y_0).$$

Then obviously

$$\partial_x v_h(x, y) = \partial_x v_h(x_0, y_0) + (x - x_0) \partial_{xx} v_h(x_0, y_0),$$

$$\partial_y v_h(x, y) = \partial_y v_h(x_0, y_0) + (y - y_0) \partial_{yy} v_h(x_0, y_0).$$

Using Q_I^{rot} -element we have

$$\partial_{xx} v_h = -\partial_{yy} v_h = \text{const.} \quad (8)$$

Applying the properties (4), from (7) we obtain

$$\begin{aligned} \iint_{K_0} \partial_x (\tilde{i}_h v - v) \partial_x v_h dx dy &= \partial_x v_h(x_0, y_0) \left(\int_{I_2} - \int_{I_4} \right) (\tilde{i}_h v - v) dy \\ &+ h_I \partial_{xx} v_h \left(\int_{I_2} - \int_{I_4} \right) (\tilde{i}_h v - v) dy - \iint_{K_0} (\tilde{i}_h v - v) \partial_{xx} v_h dx dy \\ &= - \iint_{K_0} (\tilde{i}_h v - v) \partial_{xx} v_h dx dy. \end{aligned}$$

In the same manner,

$$\iint_{K_0} \partial_y (\tilde{i}_h v - v) \partial_y v_h dx dy = - \iint_{K_0} (\tilde{i}_h v - v) \partial_{yy} v_h dx dy.$$

Finally, from (8) it follows that

$$\iint_{K_0} \nabla (\tilde{i}_h v - v) \cdot \nabla v_h dx dy = 0 \text{ for any } v_h \in \tilde{V}_h$$

and the relation (3) is proved for the element Q_I^{rot} .

Now, having in mind that $\|u_h\|_{0,\Omega} = 1$, from (3) it follows

$$\begin{aligned} a_h(\tilde{i}_h u_h - u_h, \tilde{i}_h u_h - u_h) &= a_h(\tilde{i}_h u_h, \tilde{i}_h u_h) - 2a_h(\tilde{i}_h u_h, u_h) \\ &+ a_h(u_h, u_h) = a_h(u_h, u_h) - a_h(\tilde{i}_h u_h, \tilde{i}_h u_h) = \lambda_h - \tilde{\lambda}_h, \end{aligned}$$

consequently

$$\lambda_h - \tilde{\lambda}_h = |\tilde{i}_h u_h - u_h|_h^2 \geq 0.$$

Considering that λ_h is a conforming approximation of λ by means of serendipity rectangular finite element [2], we have

$$0 \leq \lambda_h - \lambda \leq C \|u_h - u\|_{1,\Omega}^2 \leq C_I (h^{2-\delta})^2,$$

where δ is a small positive number.

Moreover, from (5) we obtain asymptotically the inequality

$$\begin{aligned} \lambda - \tilde{\lambda}_h &= \lambda - \lambda_h + \lambda_h - \tilde{\lambda}_h = -(\lambda_h - \lambda) + |\tilde{i}_h u_h - u_h|_h^2 \\ &\geq -C_I h^{4-2\delta} + C_2 h^3 \geq 0. \end{aligned}$$

Thus, (6) is proved. ■

3. Numerical Results

The numerical experiments and the resulting approximations of eigenvalues given in this section verify and confirm the validity, reliability and effectiveness of the proposed algorithm.

For purpose of demonstration of the method we propose, we solve the problem (1) on square domain $\Omega = [0, \pi] \times [0, \pi]$.

This choice we have made because of the fact that in this case the exact eigenvalues are known and are equal to $s_I^2 + s_2^2$, $s_I, s_2 = 1, 2, 3, \dots$. So that, $\lambda_1 = 2; \lambda_2 = 5; \lambda_3 = 5; \dots$

For our numerical implementation we divide the domain Ω into N^2 squares, $N = 4; 8; 12; 16; 20$. Thus the mesh parameter h is equal to $\frac{\pi\sqrt{2}}{N}$.

We solve the variational discrete problem (2) using conforming quadratic rectangular (9-point) finite elements for V_h (see Fig. 1(a)). As a result, we obtain the approximate eigenvalues λ_h , which give upper bounds for the corresponding exact eigenvalues and the approximate eigenfunctions u_h which we interpolate. Interpolation is done by means of EQ_I^{rot} -nonconforming finite element space (Fig. 2(b)). Obtaining the interpolants $\tilde{i}_h u_h \in \tilde{V}_h$ and calculating the numbers $\tilde{\lambda}_h = a_h(\tilde{i}_h u_h, \tilde{i}_h u_h)$ we get lower bounds for the exact eigenvalues. The results from our numerical experiment for the first three eigenvalues are give in Table 1 and Table 2, respectively.

Table 1: Approximations of the exact eigenvalues obtained after finite element implementation by means of 9-point conforming rectangular quadratic finite elements

	$\lambda_{1,h}$	$\lambda_{2,h}$	$\lambda_{3,h}$
$N = 4$	2.001024281	5.030616274	5.032574099
$N = 8$	2.000065532	5.002110541	5.004469411
$N = 12$	2.000013002	5.000450150	5.002890554
$N = 16$	2.000004120	5.000165987	5.0026355170
$N = 20$	2.000001689	5.000087946	5.002571064
Exact eigenvalues	2	5	5

Table 2: Approximations of the exact eigenvalues obtained after nonconforming EQ_I^{rot} -interpolation

	$\tilde{\lambda}_{1,h}$	$\tilde{\lambda}_{2,h}$	$\tilde{\lambda}_{3,h}$
$N = 4$	1.902219920	4.655142682	4.656420455
$N = 8$	1.974624003	4.901823878	4.903948840
$N = 12$	1.988641719	4.955268267	4.957598338
$N = 16$	1.993595054	4.974626877	4.977033035
$N = 20$	1.995896103	4.983705600	4.986147765
Exact eigenvalues	2	5	5

Next, for the same model problem we demonstrate the efficiency of the proposed algorithm for more simple finite elements.

We solve the variational discrete problem (2) using eight-point serendipity conforming finite elements for V_h (Fig. 1(b)). As a result we obtain the approximate eigenvalues λ_h giving upper bounds for the corresponding exact eigenvalues and the approximate eigenfunctions u_h which we interpolate. Interpolation is done by means of Q_1^{rot} -nonconforming finite element space (Fig. 2(a)). Obtaining the interpolants $\tilde{u}_h \in \tilde{V}_h$ and calculating the numbers $\tilde{\lambda}_h = a_h(\tilde{u}_h, u_h, \tilde{u}_h, u_h)$ we get lower bounds for the exact eigenvalues. The results from our numerical experiment for the first three eigenvalues are given in Table 3 and Table 4, respectively.

Table 3: Approximations of the exact eigenvalues obtained after finite element implementation by means of 8-point serendipity conforming rectangular finite elements

	$\lambda_{1,h}$	$\lambda_{2,h}$	$\lambda_{3,h}$
$N = 4$	2.001091866	5.032026684	5.032037751
$N = 8$	2.0000664317	5.002126594	5.004493905
$N = 12$	2.000066432	5.000451611	5.002891734
$N = 16$	2.000004134	5.000166314	5.002635721
$N = 20$	2.000001692	5.000088072	5.002571117
Exact eigenvalues	2	5	5

Table 4: Approximations of the exact eigenvalues obtained after nonconforming Q_1^{rot} – interpolation

	$\tilde{\lambda}_{1,h}$	$\tilde{\lambda}_{2,h}$	$\tilde{\lambda}_{3,h}$
$N = 4$	1.881513965	4.124185537	4.124191017
$N = 8$	1.949450862	4.748511061	4.750459845
$N = 12$	1.977322992	4.885339250	4.887582856
$N = 16$	1.987202566	4.934933147	4.937289058
$N = 20$	1.999994922	4.979244596	4.980117042
Exact eigenvalues	2	5	5

The approximate values in Table 1 and Table 3 are greater than the exact ones and the sequences $\{\lambda_{j,h}\}, j=1,2,3$ obtained when the mesh parameter h decreases are decreasing. This is a reasonable and expected numerical result, because of the fact that conforming finite element methods are used [2].

On the part of the nonconforming interpolation implementation, from the approximations of the eigenvalues given in Table 2 and Table 4, respectively, it is clear that the resulting approximation of the exact eigenvalues is from below. When the mesh parameter h

decreases, the sequences $\{\tilde{\lambda}_{j,h}\}, j=1,2,3$ are increasing and go to the corresponding exact eigenvalue.

As it is well-known from the theoretical point of view, nine-point and eight-point conforming finite elements give similar results concerning the error estimates (Table 1 and Table 3). As a consequence of this, as well as from the fact that Q_1^{rot} and EQ_1^{rot} - nonconforming finite elements give one and the same convergence order, the results in Table 2 and Table 4 are also similar.

3. Conclusions

The proposed algorithm gives lower and upper bounds of the eigenvalues simultaneously – only to solve by conforming FEM the eigenvalue problem once and additionally to apply a nonconforming interpolation to the conforming solution is needed.

In this paper use of more simple and convenient finite elements is proposed, proved theoretically and demonstrated.

Acknowledgements This work was partially supported by the Technical University of Gabrovo under grant D1703E/2017.

References

1. P.G. Ciarlet, The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam, New York, Oxford, 1978.
2. I. Babuška, J. Osborn, Eigenvalue Problems, In Handbook of Numerical Analysis, Vol. II, (Eds. P. G. Lions and Ciarlet P.G.), Finite Element Methods (Part 1) North-Holland, Amsterdam, 641-787, 1991.
3. Y.D. Yang, Z.M. Zhang, F.B. Lin: Eigenvalue approximation from below using nonconforming finite elements, Science China, Mathematics (2010), Vol. 53 No. 1, 137-150.
4. A. B. Andreev, M. R. Racheva, A new algorithm for two-sided eigenvalue approximation, Comp. rend. Acad. bulg. Sci. 70, 1207 – 1214, No 9, 2017.
5. M.R. Racheva, A.B. Andreev: Numerical Aspects for Obtaining Two-sided Bounds of Eigenvalues. International Scientific Journal "Science. Business. Society", 3, Scientific Technical Union of Mechanical Engineering, 2017, ISSN:2367-8380, 104-107
6. S. Brenner, L.R. Scott: The Mathematical Theory for Finite Element Methods, New York, Springer-Verlag, 1994.

A MATHEMATICAL MODEL OF VISCOUS LIQUID MIXTURE MOTION THROUGH A VERTICAL CYLINDRICAL PIPE

Asst., MSc Sorokina Natalia,
Institute of Computer Science and Technology – Peter the Great Saint-Petersburg Polytechnic University, Russia
snv_special@inbox.ru

Abstract: In the paper a mathematical model of the non-stationary motion of a viscous liquid mixture through the vertical straight pipe of the circular cross section is proposed. During the model construction weak compressibility of the mixture is considered. The Navier-Stokes equations system is taken as a basis. Such model can be used in the description of oil motion in a vertical well.

Keywords: NON-STATIONARY HYDRODYNAMICS, LIQUID MIXTURES, WEAK COMPRESSIBILITY, VERTICAL PIPE, MATH MODELING

1. Introduction

The problem of the liquid mixtures (or emulsions, suspensions) motion is not just interesting, but also very important in some areas of the national economy, such as oil and gas industry. It is well known that during the extraction of oil it is not a “clear” product ascending from under the ground, but a mixture, consisting of oil, water and gas. In other words, we deal with the water-oil emulsion.

As emulsion move up through the tubing its properties, such as density and viscosity, are changing. Its change is connected with temperature and pressure. And it is important to know the velocity of mixture’s ascending, of course.

Further in the paper we use the term “liquid” instead of “liquid mixture” since we assume that this mixture is homogeneous – its components are distributed evenly in the main phase, they are well mixed. This is the first assumption. We have to obtain the mathematical model, describing the non-stationary motion of the weakly compressible liquid through the vertical pipe of the circular cross-section.

2. The basic equations

We take the equations of continuum mechanics as the basis [1], consider that there are no sources (or drains on the contrary) of the mass:

$$\frac{d\rho}{dt} + \rho(\vec{\nabla} \cdot \vec{v}) = 0, \quad (1)$$

$$\rho \frac{d\vec{v}}{dt} = \rho \vec{F} + \frac{\partial \vec{\tau}_x}{\partial x} + \frac{\partial \vec{\tau}_y}{\partial y} + \frac{\partial \vec{\tau}_z}{\partial z}, \quad (2)$$

$$\rho \frac{d\vec{M}}{dt} = \rho \vec{\Pi} + \vec{i} \times \vec{\tau}_x + \vec{j} \times \vec{\tau}_y + \vec{k} \times \vec{\tau}_z + \frac{\partial \vec{\pi}_x}{\partial x} + \frac{\partial \vec{\pi}_y}{\partial y} + \frac{\partial \vec{\pi}_z}{\partial z}, \quad (3)$$

$$\rho \frac{dE}{dt} = \varepsilon + \vec{\tau}_x \cdot \frac{\partial \vec{v}}{\partial x} + \vec{\tau}_y \cdot \frac{\partial \vec{v}}{\partial y} + \vec{\tau}_z \cdot \frac{\partial \vec{v}}{\partial z} + \vec{\nabla} \cdot \vec{t}. \quad (4)$$

Since we consider vertical motion, which is obviously due to the head from below, we assume that rotary motion in the liquid is absent hence we will not need the equation (3) further. As the first approximation we may assume that neither density nor pressure depends on temperature and hence energy change may be ignored. That allows us to neglect the equation (4).

3. The construction of the mathematical model

We introduce the Cartesian coordinate system and direct the axis Oz along the pipe axis. We write the system of equations (1)-(2) in projections onto the Cartesian axes:

$$\frac{\partial \rho}{\partial t} + v_x \frac{\partial \rho}{\partial x} + v_y \frac{\partial \rho}{\partial y} + v_z \frac{\partial \rho}{\partial z} + \rho \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} \right) = 0, \quad (5)$$

$$\frac{\partial v_x}{\partial t} + v_x \frac{\partial v_x}{\partial x} + v_y \frac{\partial v_x}{\partial y} + v_z \frac{\partial v_x}{\partial z} = F_x + \frac{1}{\rho} \left(\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{xz}}{\partial z} \right), \quad (6)$$

$$\frac{\partial v_y}{\partial t} + v_x \frac{\partial v_y}{\partial x} + v_y \frac{\partial v_y}{\partial y} + v_z \frac{\partial v_y}{\partial z} = F_y + \frac{1}{\rho} \left(\frac{\partial \tau_{yx}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y} + \frac{\partial \tau_{yz}}{\partial z} \right), \quad (7)$$

$$\frac{\partial v_z}{\partial t} + v_x \frac{\partial v_z}{\partial x} + v_y \frac{\partial v_z}{\partial y} + v_z \frac{\partial v_z}{\partial z} = F_z + \frac{1}{\rho} \left(\frac{\partial \tau_{zx}}{\partial x} + \frac{\partial \tau_{zy}}{\partial y} + \frac{\partial \tau_{zz}}{\partial z} \right). \quad (8)$$

Since it is a vertical pipe, the gravity acts on the liquid (per the elementary volume): $F_x = 0$, $F_y = 0$, $F_z = -\rho g$. Next suppose that velocity has only the vertical component, in other words $v_x = v_y = 0$. In p.2 it is said we consider that density does not depend on temperature. Let us clarify this assumption: let temperature vary with the altitude of the rise, i.e. with the change of the z coordinate, and since the density depends on temperature (also on pressure) we have an implicit density dependence on z coordinate: $\rho = \rho(T(z)) = \rho(z)$.

We write the system (5)-(8) with the assumptions described:

$$v_z \frac{\partial \rho}{\partial z} + \rho \frac{\partial v_z}{\partial z} = 0, \quad (5')$$

$$\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{xz}}{\partial z} = 0, \quad (6')$$

$$\frac{\partial \tau_{yx}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y} + \frac{\partial \tau_{yz}}{\partial z} = 0, \quad (7')$$

$$\frac{\partial v_z}{\partial t} + v_z \frac{\partial v_z}{\partial z} = -\rho g + \frac{1}{\rho} \left(\frac{\partial \tau_{zx}}{\partial x} + \frac{\partial \tau_{zy}}{\partial y} + \frac{\partial \tau_{zz}}{\partial z} \right). \quad (8')$$

In general the stress tensor components have the following form [1]:

$$\tau_{xx} = -p + \lambda \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} \right) + 2\mu \frac{\partial v_x}{\partial x}, \quad (9)$$

$$\tau_{yy} = -p + \lambda \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} \right) + 2\mu \frac{\partial v_y}{\partial y}, \quad (10)$$

$$\tau_{zz} = -p + \lambda \left(\frac{\partial v_x}{\partial x} + \frac{\partial v_y}{\partial y} + \frac{\partial v_z}{\partial z} \right) + 2\mu \frac{\partial v_z}{\partial z} \quad (11)$$

$$\tau_{xy} = \tau_{yx} = \mu \left(\frac{\partial v_x}{\partial y} + \frac{\partial v_y}{\partial x} \right), \tau_{yz} = \tau_{zy} = \mu \left(\frac{\partial v_y}{\partial z} + \frac{\partial v_z}{\partial y} \right), \quad (12)$$

$$\tau_{zx} = \tau_{xz} = \mu \left(\frac{\partial v_x}{\partial z} + \frac{\partial v_z}{\partial x} \right).$$

According to the assumptions made, the stress tensor components take the following form:

$$\tau_{xx} = -p + \lambda \frac{\partial v_z}{\partial z}, \quad (9')$$

$$\tau_{yy} = -p + \lambda \frac{\partial v_z}{\partial z}, \quad (10')$$

$$\tau_{zz} = -p + [\lambda + 2\mu] \frac{\partial v_z}{\partial z} \quad (11')$$

$$\tau_{xy} = \tau_{yx} = 0, \tau_{yz} = \tau_{zy} = \mu \frac{\partial v_z}{\partial y}, \tau_{zx} = \tau_{xz} = \mu \frac{\partial v_z}{\partial x}. \quad (12')$$

Substitute these relations in the system (5')–(8'):

$$v_z \frac{\partial \rho}{\partial z} + \rho \frac{\partial v_z}{\partial z} = 0, \quad (5')$$

$$-\frac{\partial p}{\partial x} + [\lambda + \mu] \frac{\partial^2 v_z}{\partial x \partial z} = 0, \quad (6')$$

$$-\frac{\partial p}{\partial y} + [\lambda + \mu] \frac{\partial^2 v_z}{\partial y \partial z} = 0, \quad (7')$$

$$\frac{\partial v_z}{\partial t} + v_z \frac{\partial v_z}{\partial z} = -\rho g + \frac{1}{\rho} \left(\mu \frac{\partial^2 v_z}{\partial x^2} + \mu \frac{\partial^2 v_z}{\partial y^2} + \lambda \frac{\partial^2 v_z}{\partial z^2} - \frac{\partial p}{\partial z} \right). \quad (8')$$

Let us turn to the cylindrical coordinates in the system (5')–(8'), assuming in advance the cylindrical symmetry of the flow:

$$v_z \frac{\partial \rho}{\partial z} + \rho \frac{\partial v_z}{\partial z} = 0, \quad (5')$$

$$[\lambda + \mu] \cos \varphi \frac{\partial^2 v_z}{\partial r \partial z} - \cos \varphi \frac{\partial p}{\partial r} = 0, \quad (6')$$

$$[\lambda + \mu] \sin \varphi \frac{\partial^2 v_z}{\partial r \partial z} - \sin \varphi \frac{\partial p}{\partial r} = 0, \quad (7')$$

$$\frac{\partial v_z}{\partial t} + v_z \frac{\partial v_z}{\partial z} = -\rho g + \frac{1}{\rho} \left(\mu \left[\frac{\partial^2 v_z}{\partial r^2} + \frac{1}{r} \frac{\partial v_z}{\partial r} \right] + \lambda \frac{\partial^2 v_z}{\partial z^2} - \frac{\partial p}{\partial z} \right). \quad (8')$$

Next we add two equations (6') and (7'):

$$[\lambda + \mu] [\cos \varphi + \sin \varphi] \left(\frac{\partial^2 v_z}{\partial r \partial z} - \frac{\partial p}{\partial r} \right) = 0.$$

Since sum of bulk and dynamic viscosity coefficients cannot be zero, sum of sine and cosine of the same angle cannot be zero, it follows that

$$\frac{\partial^2 v_z}{\partial r \partial z} - \frac{\partial p}{\partial r} = 0,$$

or

$$\frac{\partial^2 v_z}{\partial r \partial z} = \frac{\partial p}{\partial r}. \quad (13)$$

Thus the final system of equations has the following form (we omit the z index since we assume that there is only one velocity component):

$$v \frac{\partial \rho}{\partial z} + \rho \frac{\partial v}{\partial z} = 0, \quad (5')$$

$$\frac{\partial^2 v}{\partial r \partial z} = \frac{\partial p}{\partial r}, \quad (13)$$

$$\frac{\partial v}{\partial t} + v \frac{\partial v}{\partial z} = -\rho g + \frac{1}{\rho} \left(\mu \left[\frac{\partial^2 v}{\partial r^2} + \frac{1}{r} \frac{\partial v}{\partial r} \right] + \lambda \frac{\partial^2 v}{\partial z^2} - \frac{\partial p}{\partial z} \right). \quad (8')$$

Three equations contain three unknown functions: density $\rho(z)$, pressure $p(z)$ and velocity $v(r, z, t)$.

4. The formulation of the corresponding problem and its correctness

As noted earlier, the problem of studying the motion of a viscous liquid (liquid mixture) through the vertical pipe arises in the

field of oil products extraction. It is necessary to predict the pressure, viscosity and velocity of the mixture that rises from the depth along the pipe, to understand whether, for example, pressure changes are so critical that it will lead to partial equipment destruction (the same pipe).

For the mathematical formalization of the corresponding problem, it is necessary to set the initial pressure at the pipe inlet and pipe outlet (at the pipe outlet it is possibly constant), initial velocity distribution over the cross-section and initial density of the liquid. In time, or, in the first approximation, during the ascending along the change in pressure the liquid density changes, since the saturation of the mixture with gas increases [4,5].

In solving partial differential equations (systems of equations), the study of the correctness of the corresponding initial-boundary value problems for equations being solved is especially important. For quasilinear partial differential equations of the second order this analysis is not simple. For the system of equations (5'), (13), (8'), such a study was carried out by OA. Ladyzhenskaya [2]. The results are positive.

5. The solution method

For the initial-boundary value problem, the system of equations obtained in p.3 can be solved by a numerical method – the variable direction method [3]. According to this method, the previously obtained partial differential equations are replaced by difference relations with intermediate computation of functions on the $k + 0.5$ -th time layer: A similar solution has already been applied in [4].

6. Conclusion

In this paper, the construction of a mathematical model describing the nonstationary motion of a viscous compressible mixture over a long vertical rectilinear tube of circular cross-section is considered. Such a model can be used to solve the problem of the movement of oil along a vertical well. A numerical method for solving the corresponding initial-boundary value problem is proposed.

7. References

- [1] Vallander, S.V. Lectures on fluid dynamics. – L.: Izd-vo Leningr. un-ta, 1973, 296 p. (in Russian)
- [2] Ladyzhenskaya, O. A. On the unique solvability in the large of the three-dimensional Cauchy problem for the Navier-Stokes equations in the presence of axial symmetry (in Russian), Zapiski Nauchnikh Seminarov LOMI, 1968, v. 7, p. 155-177
- [3] Filatov, E. Y., Yasinskiy, F. N. Mathematical modeling of liquid and gas flows: Tutorial / State Educational Institution of Higher Professional education "Ivanovo Power Engineering Institute of V. I. Lenin". – Ivanovo, 2007 (in Russian)
- [4] Firsov A., Sorokina N. Mathematical modeling of non-stationary flows of liquid homogeneous viscous mixtures by pipelines // Mathematical modeling. – 2017. – Issue 1/2017. – P.18-22
- [5] Islamov M.K. Borehole oil extraction [Electronic resource]: Theoretical basis for lifting fluid and gas from a well URL: http://doidpo.rusoil.net/storage/Downhole%20oil%20production%20Islamov%29/to/to4_1.html (access date: 21.11.17) (in Russian)
- [6] Everything about oil [Electronic resource]: Oil extraction. Methods of operation of wells. URL: <http://vseonefti.ru/upstream/sposoby-dobychi.html> (access date: 21.11.17) (in Russian)

MODELING OF LIQUID SPREADING IN RANDOMLY PACKED METAL PALL RINGS

МОДЕЛИРАНЕ НА РАЗТИЧАНЕТО В НЕНАРЕДЕНИ МЕТАЛНИ ПРЪСТЕНИ НА ПАЛ

Dr. Petrova T., Stefanova, K., Dr. Dzhonova-Atanasova, D., Prof. Dr. Semkov, K.

Bulgarian Academy of Sciences, Institute of Chemical Engineering
Akad. G. Bonchev Str., bl.103, 1113 Sofia, Bulgaria,

e-mail: tancho66@yahoo.com

Abstract: The present work compares two different approaches, Computational Fluid Dynamics (CFD) and dispersion model, for liquid distribution modeling with experimental data for liquid spreading in randomly packed metal Pall rings. The used experimental data are obtained in a semi-industrial column with a 0.6m diameter for several packing heights and liquid loads. It is shown that the appropriate choice of dispersion model parameters is essential for prediction of liquid distribution. In both models some parameters are determined by fitting with experimental data, the remainder are calculated or taken from literature. Comparison of the two model liquid distributions with experimental data shows that both CFD and dispersion model are in good agreement with the experiment especially for higher packing bed, when the wall flow is fully developed.

Keywords: PACKED-BED COLUMN, LIQUID DISTRIBUTION, RANDOM PACKING, DISPERSION MODEL, PARAMETER IDENTIFICATION

1. Introduction

Packed columns are widely used for separation processes such as rectification and absorption in chemical industry and environmental protection due to their high efficiency at low pressure drop. The recent interest is connected with technologies for flue gas scrubbing, heat recovery and fuel production. The uniform liquid distribution in the packed bed cross-section is crucial for the efficiency of the transfer processes. Several models are proposed to predict the liquid distribution in a packed bed starting from the random walk model of Tour and Lerman [1] for liquid spreading in unconfined packing with no wall effect and the dispersion model of Cihla and Schmidt [2] developed by other authors to account for the wall flow. Lately the techniques of Computational Fluid Dynamics (CFD) are widely employed for prediction of the liquid distribution by treating the packing bed as a porous media with permeability resistance, Yin [3], or by reconstructing the geometry of the packing bed and tracking the gas-liquid interface area.

The present work compares results from a dispersion model and a CFD simulation of liquid spreading in a packed bed of random Pall rings.

2. Model description

In this part a brief concept of dispersion model equations, boundary conditions, and solution is presented. The full description is given in the work [4]. The process of flow distribution in a packed-bed column is described by the following dimensionless equation [2]:

$$(1) \left(\frac{\partial^2 f(r, z)}{\partial r^2} + \frac{1}{r} \frac{\partial f(r, z)}{\partial r} \right) = \frac{\partial f(r, z)}{\partial z},$$

where $r = r'/R$ is dimensionless radial coordinate; r' is radial coordinate, m; R is column radius, m; $z = Dh/R^2$ is dimensionless axial coordinate; D is liquid distribution coefficient, m; h is axial coordinate, m; $f = L/L_0$ is dimensionless superficial velocity; L, L_0 are local and mean liquid superficial velocity, $m^3/m^2 \cdot s$. The boundary conditions are the following [5]:

$$(2) -\frac{\partial f}{\partial r} = B(f - CW), \quad \text{for } r = 1$$

$$(3) \frac{\partial f(r, z)}{\partial r} = 0, \quad \text{for } r = 0$$

where parameter B is a criterion for exchange of liquid between the column wall and the packing; parameter C expresses the equilibrium distribution of entire liquid flow between the wall and the packing when equilibrium state is attained $z \rightarrow \infty$; W is dimensionless wall flow.

The equations defining these parameters are:

$$(4) B = \beta R / D,$$

$$(5) C = \pi R^2 \gamma,$$

where β , $-$, and γ , $1/m^2$, are parameters in the boundary conditions.

At $z = 0$ the uniform initial irrigation is defined as:

$$(6) f(r, z) = 1, \text{ for } 0 \leq r < 1 \text{ and } z = 0$$

There is analytical solution of the above model in the form of infinite series [6]:

$$(7) f^u(r, z) = A_0 + \sum_{n=1}^{\infty} A_n^u J_0(q_n r) \exp(-q_n^2 z),$$

In the above expressions, f^u (dimensionless), denotes the solution for uniform initial irrigation. The coefficients are derived from the expressions:

$$(8) A_0 = \frac{C}{1+C}, \quad A_n^u = \frac{2(q_n^2/B - 2C)}{\left[(q_n^2/B - 2C)^2 + q_n^2 + 4C \right] J_0(q_n)},$$

The dimensionless wall flow W^u is calculated from Eq. (9) and from the material balance:

$$(9) W^u = \frac{1}{1+C} - 2 \sum_{n=1}^{\infty} A_n^u(q_n) \frac{J_1(q_n)}{q_n} \exp(-q_n^2 z),$$

where J_0, J_1 are Bessel functions of the first kind, zero and first order; q_n are the roots of the characteristic equation, following from boundary condition (2):

$$(10) [(2C/q_n) - (q_n/B)] J_1(q_n) + J_0(q_n) = 0$$

3. Experimental data.

Experimental data for the liquid distribution in a column filled with metal Pall rings, of 25 mm, used in this paper are taken from the PhD thesis of Yin [3]. They are obtained in a pilot column with 0.6m diameter, for packing heights 0.9 and 3.0 m, and liquid and gas loads $L = 2.91; 4.78; 6.66, G = 0; 0.75 \text{ kg/m}^2 \cdot \text{s}$. The liquid collecting device consists of 6 segments, the initial irrigation is uniform.

4. Methods of estimation and identification of model parameters

In this paper we propose the following scheme for three parameters' estimation/identification of the mathematical model described above.

1) The value of C can be determined even if there are not available data for wall initial irrigation. In this case the data about wall flow at several packing heights for uniform initial irrigation are needed. We use the relation (9) in the limiting case when $z \rightarrow \infty$; as it is seen, the last term in (9) diminishes and the result is:

$$(11) W^u \Big|_{z \rightarrow \infty} = \frac{1}{1+C},$$

2) In this work the value of the radial spreading coefficient D is taken from the literature [7];

3) Then, only parameter B can be identified by non-linear optimization minimizing the residual variance:

$$(12) S_A^2 = \frac{1}{k-1} \sum_{i=1}^k n_i (f_{ie} - f_{ic})^2,$$

where f_i is the mean dimensionless density of irrigation in i -th annular section of the liquid collecting device, delimited by the radii r_{i-1} and r_i ($r_i > r_{i-1}$) and is determined by the expression

$$(13) f_i = \frac{2}{r_i^2 - r_{i-1}^2} \int_{r_{i-1}}^{r_i} f(r, z) r dr,$$

with indices "e" and "c" denoting experimental and calculated values of liquid density of irrigation.

As explained above, three parameters should to be determined – C, D and B . In the light of experimental data used in this paper, for calculation of C eq. (11) is used. Then, the obtained value for C is 5.95, which is very close to the value ($C = 5.29$), obtained for ceramic Pall rings 25 mm in [4]. Here, the value of C is calculated from the results [3] for the wall flow at uniform initial irrigation for packing heights 0.9, 1.8 and 3.0 m at liquid load $L = 6.66 \text{ kg/m}^2 \cdot \text{s}$ and no gas flow, because at these conditions a fully developed wall flow is achieved. As it is recommended in [8], the value of C is better to obtain from data at higher packing layer, when the equilibrium between bulk zone and wall flow is reached. On the contrary, these authors mentioned, that the parameter B has to be identified/calculated

for lower packing depths, because at higher ones it loses its importance.

In this work the value of D is chosen to be $D = 0.0007 \text{ m}$. This value is the last result [7], reported in the literature about metal Pall rings, 25.4 mm, and, which is more important, the results of Wen et al.[7] are obtained for the same packing, as in [3]. The previous reported data for D is that of Stikkelmann [9], which is about 3 times larger (0.0025m) although the method of calculating of D is the same as in the work [7]. The probable explanation of these differences in the value of D is that a piece of Pall ring packing may have different geometry inside its cylindrical body – see [10]. The Pall rings are created in 1940 and up to now a large number of manufacturers as well as many modifications of this packing type exist (Flexiring, Raflux ring, P-ring, Hy-pak, etc.). The cut "windows" in one packing element may have different width and curvature, which changes the radial spreading inside the packing element, and this result in a large variety of spreading coefficients even for the same packing diameter and material [11]. In [4], the reported value for ceramic Pall rings, 25 mm, for approximate comparison, is $D = 0.0011 \text{ m}$, but the inside geometry of ceramic Pall rings is different from that of metal one.

The only one model parameter to be identified is B . That is done using the minimal residual variance (12) between model and experimental mean dimensionless density of irrigation (13) in each segment of liquid collecting device (detailed procedure is described in [12]. The obtained value is $B = 25$.

4. Comparison of liquid predictions (CFD results of Yin [3] and current model) with experimental data for metal Pall rings [3]

In [3], the comparison between experimental and CFD data for liquid relative velocity is made, for packing heights $H=0.9 \text{ m}$ and $H=3.0 \text{ m}$, for $L=4.78 \text{ kg/m}^2 \cdot \text{s}$ and $G=0.75 \text{ kg/m}^2 \cdot \text{s}$, which is below the reported loading point and the influence of the gas phase is not significant. We compare these results with model solution (7) for already identified values of the parameters. The comparison of both model predictions - CFD and dispersion model from this work are presented in figures 1 and 2.

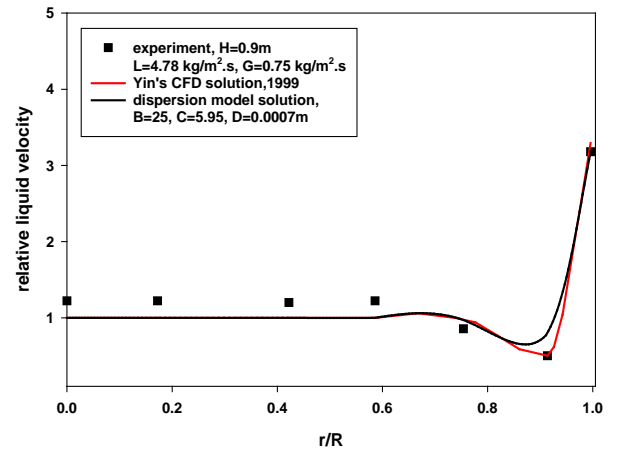


Fig. 1 Comparison of relative liquid velocities between dispersion model predictions, Yin's CFD predictions and Yin's experimental data: $L=4.78 \text{ kg/m}^2 \cdot \text{s}$; $G=0.75 \text{ kg/m}^2 \cdot \text{s}$; for packing height $H=0.9 \text{ m}$.

As can be seen, both models describe very well experimental data. For lower packing height the radial profile of experimental data in the bulk zone are still parabolic and equalize when the packing height is increased. Both predictions are more accurate at $H=3.0 \text{ m}$.

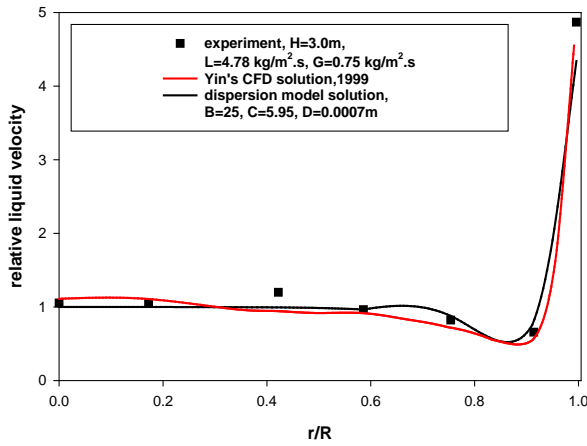


Fig. 2 Comparison of relative liquid velocities between dispersion model predictions, Yin's CFD predictions and Yin's experimental data: $L=4.78 \text{ kg/m}^2\text{s}$; $G=0.75 \text{ kg/m}^2\text{s}$; for packing height $H=3.0 \text{ m}$.

It should be mentioned, that at both figures, model predictions of [3] are performed with the aid of the modern CFD package CFX4.2. Two of the parameters, included in the closure equations and relations, also were obtained by fitting with the experiment. They are connected with 'liquid dispersion coefficient', defined by Yin [3] for volume average concept, and turbulent term, included in it.

In the case of no gas flow, in the next figures 3 and 4, the results of dispersion model predictions for the same heights, but for different liquid loads $L=2.91 \text{ kg/m}^2\text{s}$ and $L=6.66 \text{ kg/m}^2\text{s}$ are also presented. For above identified model parameters the coincidence between the model and experiment is quite well, which is additional verification of dispersion model ability to predict liquid spreading in packed beds.

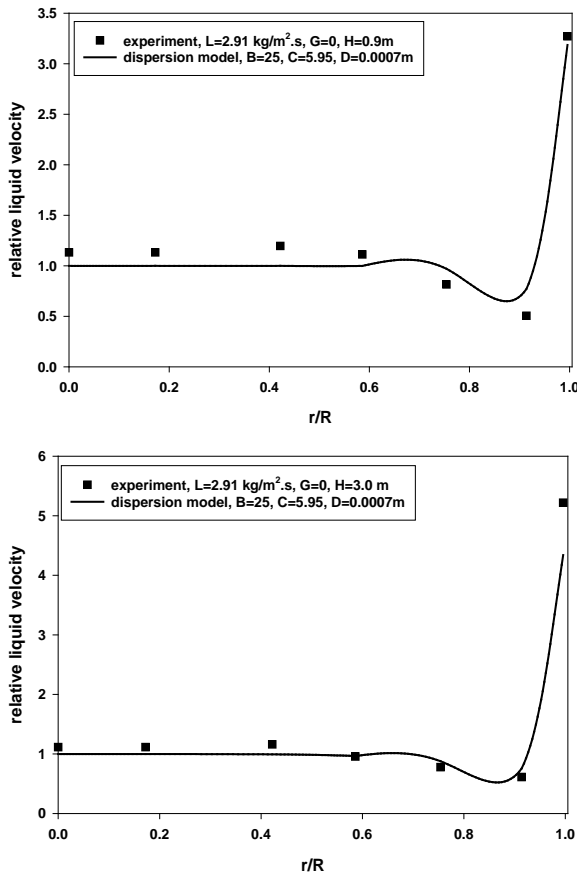


Fig. 3 Comparison of relative liquid velocities between dispersion model predictions and Yin's experiment: $L=2.91 \text{ kg/m}^2\text{s}$; $G=0$; for packing heights $H=0.9 \text{ m}$ (up) and $H=3.0 \text{ m}$ (down).

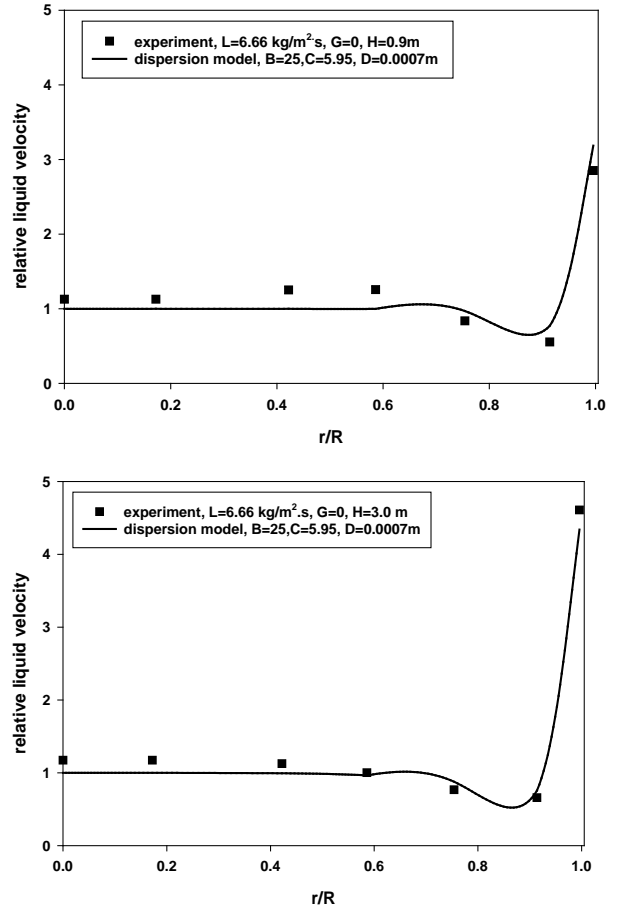


Fig. 4 Comparison of relative liquid velocities between dispersion model predictions and Yin's experiment: $L=6.66 \text{ kg/m}^2\text{s}$; $G=0$; for packing heights $H=0.9 \text{ m}$ (up) and $H=3.0 \text{ m}$ (down).

5. Conclusions

In the current work the results of dispersion model prediction for liquid distribution in random packing and identification procedure for model parameters are presented. A new formula is proposed for estimation of model parameter C in the case of uniform initial irrigation only, and experimental data for liquid distribution at several packing layer heights. Validation of model prediction for identified model parameters with experimental data of [3] for Pall rings, at two different packing heights and three liquid loads shows, that dispersion model analytical solution could be used successfully for fast prediction of liquid distribution in randomly packed beds.

The dispersion model results are also compared with CFD prediction and experimental data of Yin for liquid distribution in a packed bed with metal Pall rings [3]. Comparison of two theoretical liquid distributions with experimental one show, that both describe very well the experimental data.

6. Literature

1. Tour, R. S. , F. Lerman. Unconfined distribution of liquid in tower packing, Trans. AICHE 35, 1939, 709-718
2. Cihla, Z., O. Schmidt. A study of the flow of liquid when freely trickling over the packing in a cylindrical tower. - Collect. Czech. Chem. Commun., 22, 1957, 896-907
3. Yin, F., Liquid maldistribution and mass transfer efficiency in randomly packed distillation columns, PhD thesis, Department of chemical and material engineering, University of Alberta, Edmonton, Alberta, Canada, 1999.

4. Semkov, Kr., T. Petrova, P. Moravec. Parameters Identification of a Mathematical Model for Liquid Distribution in Packed-Bed Columns. - Bulgarian Chemical Communications, 32, 3-4, 2000, 497 – 516
5. Staněk, V., V. Kolář. Distribution of liquid over random packing. - Collect. Czech. Chem. Commun., 30, 1965, 1054-1059
6. Staněk, V., V. Kolář. Distribution of liquid over a random packing . VIII Distribution of the density of wetting in a packing for an arbitrary type of initial conditions. - Collect. Czech. Chem. Commun., 38, 1973, 2865-2873
7. Wen, X., Y. Shu, K. Nandakumar, K. Chuang. Predicting Liquid Flow Profile in Randomly Packed Beds from Computer Simulation. - AIChE Journal, 47, 8, 2001, 1770-177
8. Staněk, V., V. Kolář. Distribution of liquid over a random packing: Verification of the boundary condition of liquid transfer between a packed bed and the wall of a cylindrical column and evaluation of its parameters. - Collect. Czech. Chem. Commun., 33, 4, 1968, 1062-1077
9. Stikkelmann, R. Gas and liquid maldistributions in packed columns. - PhD thesis, Technical University of Delft, Delft, Netherlands, 1989
10. Web brochure : <http://www.rubbersealing.com/TCI/category-34-b0--Random+Tower+packing-pall+ringHiflow+ring.html>
11. Petrova, T., Kr. Semkov, S. Darakchiev, R. Darakchiev. Разтичане на течността в колони с ненаредени пълнежи. - Научни трудове на УХТ –Пловдив, 56, 2, 2009, 245 – 250
12. Dzhonova-Atanasova, D., Kr. Semkov, T. Petrova, S. Darakchiev, K. Stefanova, Sv. Nakov, R. Popov. Liquid distribution in a semi-industrial packed column- experimental and theory. – Scientific Works of University of Food Technology, 64, 1, 2017 (in print).

Acknowledgements

This work was financially supported by the National Science Fund at the Bulgarian Ministry of Education and Science, Contract No DN 07/14/15.12.2016

STATISTICAL METHODS FOR THE ANALYSIS OF THE MONTE CARLO SIMULATION RESULTS IN VISION SYSTEMS

I. Yu. Gendrina, Ass. Prof., Ph.D.
National Research Tomsk State University, Tomsk, Russia
igendrina@bk.ru

Abstract: In this paper, we present some results of statistical processing of the results of numerical simulation of the characteristics of vision systems through the atmosphere obtained with the help of a special software package created by us (Gendrina I.Yu., Kvach A.S.).

Keywords: Vision system, angular brightness distribution, Monte Carlo method, regression.

Introduction

Various methods of statistical research such as correlation-regression analysis, dynamic series, variance analysis, etc. in the study of vision systems are used for studying and subsequent prediction of the patterns of radiation transfer. We have conducted Monte Carlo experiments to calculate the Point Spread Function for linear system "underlying surface – atmosphere". The one is defined as the system response to the input signal, representing a point mass, located at a certain point, and can be determined as the angular brightness distribution of surface-based point source measured with receiving device at the top of the atmosphere. Then we have attempted to apply elements of correlation-regression analysis for the study of the influence of various optical and geometrical parameters on the Point Spread Function.

Vision system (VS) is understood as an observation scheme including the underlying surface, a "cloudy environment" (atmosphere), and an optical device that captures incoming radiation. To study radiation transfer in such systems, the theory of systems and the theory of radiation transfer are traditionally used (Zuev V.E., Belov V.V., Veretennikov V.V.).

The main system characteristic for VS is the point spread function (PSF); it is defined as the response L of linear system to the input signal, representing a point mass $\delta(x - x_1)\delta(y - y_1)$, located at a certain point (x_1, y_1) : $L[\delta(x - x_1)\delta(y - y_1)] = h(x, y; x_1, y_1)$.

An arbitrary object (function) $f(x, y)$ can be considered as a set of point masses. For instance:

$$f(x, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x_1, y_1) \delta(x - x_1, y - y_1) dx_1 dy_1. \quad \text{Then, a}$$

result of the system impact (image) can be represented in the form:

$$g(x, y) = L[f(x, y)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x_1, y_1) h(x, y; x_1, y_1) dx_1 dy_1.$$

Obviously, regularities of the image distortion due to impact of any system can be studied by analyzing the effect of this system on the point spread function.

The model of the atmosphere includes the following characteristics: the total attenuation coefficient $\sigma(\lambda, \vec{r}) = \sigma_{sc}(\lambda, \vec{r}) + \sigma_a(\lambda, \vec{r})$, where σ_{sc} – the scattering coefficient, σ_a – the absorption coefficient; $g(\lambda, \mu, \vec{r})$ – aerosol phase function. Here $\vec{r} = (x, y, z)$ – radius-vector of the current point in space, $\mu(\vec{\omega}', \vec{\omega})$ – cosine of the scattering angle of radiation coming from direction $\vec{\omega}'$, in the direction $\vec{\omega}$, λ – is the wavelength of incident radiation.

The paper considers two models of the atmosphere:

1. vertically bounded plane-parallel layer-homogeneous aerosol-molecular;

2. vertically bounded plane-parallel aerosol-molecular, including overcast layer. For the cloud layer, different characteristics from those of the first model is assumed: coefficients of attenuation, absorption, scattering, and the aerosol phase function.

1. Problem Statement and methods

The geometric scheme of calculations is as follows: at the lower boundary of the atmosphere (on the underlying surface $z=0$) there is a point source of unit capacity. At the upper boundary of the atmosphere ($z=30\text{ km}$) there is an optical receiver that can receive scattered radiation coming from different directions (observation angles). The brightness of the scattered radiation is solution of the integro-differential transport equation (Marchuk G.I., Mikhailiov G.A., Nazaraliev M.A., Darbinjan, Kargin B.A., Elepov B.S.), which can be practically solved only by approximate or numerical methods.

One of the most universal methods for solving this problem is the simulation method, or the Monte Carlo method. The basis of the Monte Carlo method is the integral transport equation of the second kind with a generalized kernel for the density of particles' collisions (Marchuk G.I., Mikhailiov G.A., Nazaraliev M.A., Darbinjan, Kargin B.A., Elepov B.S.):

$$f(\vec{x}) = \int_{\vec{x}} k(\vec{x}', \vec{x}) f(\vec{x}') d\vec{x}' + \psi(\vec{x}), \quad f = Kf + \psi$$

Here $\vec{x} = (\vec{r}, \vec{\omega})$ – is the point of the phase space of coordinates and directions, $\psi(\vec{x})$ – source function, K – integral operator with kernel $k(\vec{x}', \vec{x})$:

$$k(\vec{x}', \vec{x}) = \frac{\sigma_{sc}(\vec{r}) \cdot g(\mu) \exp(-\tau(\vec{r}', \vec{r})) \sigma(\vec{r})}{\sigma(\vec{r}') 2\pi |\vec{r} - \vec{r}'|^2} \cdot \delta\left(\vec{\omega} - \frac{\vec{r} - \vec{r}'}{|\vec{r} - \vec{r}'|}\right)$$

In this paper, one of the algorithms of the Monte Carlo method was used - the method of local estimation (Marchuk G.I., Mikhailiov G.A., Nazaraliev M.A., Darbinjan, Kargin B.A., Elepov B.S.).

Algorithm for local estimation consists in calculating following functional:

$$J(\Omega_i) = \int_{\Omega_i} \Phi(\vec{r}^*, \vec{\omega}^*) d\vec{\omega}^* = \int_{\vec{x}} l_i(\vec{x}, \vec{x}^*) f(\vec{x}) d\vec{x} = \quad (1)$$

$$= M \sum_{n=0}^N Q_n \cdot l_i(\vec{x}_n, \vec{x}^*)$$

$$l_i(\vec{x}, \vec{x}^*) = \frac{\exp(-\tau(\vec{r}, \vec{r}^*)) \cdot g(\mu^*)}{2\pi |\vec{r} - \vec{r}^*|^2} \Delta_i(\vec{s}^*). \quad (2)$$

$$\text{Here } \vec{s}^* = \frac{\vec{r}^* - \vec{r}}{|\vec{r}^* - \vec{r}|}, \quad \mu^* = (\vec{\omega}, \vec{s}^*), \quad \Delta_i(\vec{s}) - \text{ is}$$

the indicator of region Ω_i . Φ – flux of particles at given point \vec{x}^* . Q_n – weight of the particle, $f(\vec{x})$ – density of collisions.

2. Initial data

We will consider the process of radiative transfer through aerosol-molecular atmosphere, which comprises a layer overcast, by neglecting the reflection from underlying surface. We used the following data (Gendrina I.Yu., Kvach A.S.):

1. Wavelength (mkm) in transparent windows: 0.347; 0.530; 0.694; 0.860; 1.060; 3.390; 10.60.

2. Lower boundary of atmosphere 0 km above Earth's surface, upper boundary H of the atmosphere 30 km above the Earth's surface.

3. Optical thickness for a cloudless atmosphere are presented in Table 1.

Table 1. Optical thickness of the cloudless atmosphere

Wavelength, mkm	Optical thickness
0,347	0,228
0,53	0,158
0,694	0,124
0,86	0,098
1,06	0,092
3,39	0,067
10,6	0,041

4. Lower boundary of the cloud layer - 1 km above the Earth's surface, upper boundary - 2 km above the Earth's surface. The optical models of the cloud layer – “haze H” and “cloud C1” (Deirmendjian D.)

5. In this work, we considered Lambertian model of sources of radiation. In this case the density of the initial areas

looks like: $p(\tilde{\omega}) = \frac{\mu}{\pi}$, where $\mu = \arccos \theta$, θ - zenith angle of initial direction.

3. Simulation results

Quantitative values brightness of scattered radiation for the cloudless atmosphere are presented in our previous publication (Gendrina I.Yu., Alekseenko M. A.). Similar values for various models of the cloud atmosphere are given in Tables 2,3.

Table 2 contains the results for the brightness of scattered radiation for the atmosphere with a cloud layer of the “Haze H” type.

Table 3 contains similar data for the atmosphere with a cloud layer of the “Cloud C1” type.

These types vary in value of the average cosine of scattering phase function:

$$\bar{\mu} = \frac{1}{2} \int_{-1}^1 \mu g(\mu) d\mu.$$

It is known that this parameter characterizes the elongation of aerosol phase function. For example in case $\lambda = 0,694$ mkm the average cosine amounts to 0,745 for type “Haze H” and 0,857 for type “Cloud C1”.

Table 2. Brightness of scattered radiation for the atmosphere with a cloud layer of the “Haze H” type, $W/mkm \cdot m^2$

Angles of reception, grad	Wavelength, mkm			
	$\lambda=0,374$	$\lambda=0,530$	$\lambda=0,694$	$\lambda=0,860$
4,5	2,24E-05	1,67E-05	1,38E-05	1,13E-05
13,5	1,09E-06	7,65E-07	6,60E-07	5,30E-07
22,5	2,23E-07	1,83E-07	1,55E-07	1,39E-07

31,5	8,44E-08	5,63E-08	5,54E-08	4,73E-08
40,5	3,49E-08	2,73E-08	2,81E-08	6,65E-08
49,5	1,86E-08	5,15E-08	1,54E-08	1,18E-08
58,5	1,30E-08	1,10E-08	1,71E-08	7,75E-09
67,5	1,94E-08	1,12E-08	7,95E-09	2,11E-08
76,5	1,58E-07	2,13E-08	4,96E-09	3,68E-09
85,5	1,20E-08	6,37E-09	1,23E-09	1,03E-09

Table 3. Brightness of scattered radiation for the atmosphere with a cloud layer of the “Cloud C1” type, isotropic source, $W/mkm \cdot m^2$

Angles of reception, grad	Wavelength, mkm			
	$\lambda=0,374$	$\lambda=0,530$	$\lambda=0,694$	$\lambda=0,860$
4,5	4,24E-07	5,01E-07	5,45E-07	5,69E-07
13,5	1,32E-07	1,36E-07	1,01E-07	9,47E-08
22,5	5,06E-08	4,30E-08	4,23E-08	3,50E-08
31,5	2,29E-08	1,84E-08	1,20E-08	1,19E-08
40,5	7,00E-09	8,43E-09	6,45E-09	5,16E-09
49,5	3,29E-09	2,44E-09	3,10E-09	2,69E-09
58,5	1,84E-09	1,59E-09	1,36E-09	1,39E-09
67,5	1,17E-09	1,13E-09	7,77E-10	8,25E-10
76,5	8,77E-10	9,73E-10	7,45E-10	5,32E-10
85,5	1,70E-10	5,77E-10	8,95E-11	1,78E-10

4. Statistical analysis of simulation results

To establish functional relationship between the angular distribution of brightness and optical parameters, regression analysis was used, which is widely used to restore aerosol and cloud characteristics, and also to assess their effect on climate (Khayer M. M. et al.). The regression equation for the angular distribution of brightness in the aerosol-molecular and cloud atmosphere relatively wavelength of the incident radiation was

obtained in the form $y = \frac{b_1}{x} + b_0$ for all reception angles.

Regression coefficients for the cloudless atmosphere are given in Table 4. The regression coefficients for the cloudy atmosphere are given in Table 5. The tables also shown coefficients of determination R^2 . This coefficient indicates the proportion of the total variation in the dependent variable y due to variability of x and characterizes the total quality of regression. The statistical significance of determination coefficient can be

confirmed with the help of Fisher statistics: $F = \frac{R^2}{1 - R^2} \cdot (n - 2)$.

Here n is the number of observations. The value F is compared to value $F_{\alpha; k_1, k_2}$ from Fisher table. Here α - the given level of significance, $k_1 = 1$ and $k_2 = n - 2$.

Table 4. Coefficients of the regression equation for the cloudless atmosphere

Angles of reception, grad	b_0	b_1	R^2
4,5	1,19E-05	3,25E-05	0,993
13,5	2,31E-07	5,05E-07	0,976
22,5	6,98E-08	1,19E-07	0,927
31,5	3,92E-08	4,04E-08	0,784
40,5	2,44E-08	1,51E-08	0,570
49,5	1,39E-08	6,51E-09*	0,482
58,5	6,82E-09	3,64E-09	0,609
67,5	2,74E-09	2,35E-09	0,771
76,5	5,21E-10	1,23E-09	0,907
85,5	3,00E-11	1,76E-10	0,929

Table 5 Coefficients of the regression equation for the cloud atmosphere

Angles of reception, grad	b_0	b_1	R^2
4,5	3,52E-06	7,12E-06	0,985
13,5	1,08E-07	3,63E-07	0,997
22,5	3,63E-08	7,53E-08	0,973
31,5	1,26E-08	2,66E-08	0,976
58,5	3,68E-09*	4,34E-09	0,598
76,5	-3,34E-08*	5,04E-09	0,606
85,5	-2,37E-09*	4,50E-09	0,809

We would like to present also the brightness dependence relatively the upper boundary of cloud. The regression equation in this case was obtained in a simple form: $y = b_1x + b_0$ for all reception angles. Regression coefficients and determination coefficients for the case of isotropic source, wavelength of $\lambda = 0,347$ mkm, cloud layer type of the "Haze H" and "Cloud C1" are given in Table 6,7.

Table 6. Coefficients of the regression equation for the cloud atmosphere. Model «Haze H»

Angles of reception, grad	b_0	b_1	R^2
4,5	5,14E-07	-3,07E-08	0,990
13,5	1,53E-07	-7,70E-09	0,975
22,5	5,92E-08	-2,33E-09	0,851
31,5	2,65E-08	-6,54E-10*	0,654
40,5	7,36E-09	-1,75E-10	0,946
49,5	3,60E-09	-6,36E-11	0,625
58,5	2,11E-09	-5,13E-11	0,840
67,5	1,25E-09	-2,96E-11	0,886

Table 7. Coefficients of the regression equation for the cloud atmosphere
Model «Cloud C1»

Angles of reception, grad	b_0	b_1	R^2
4,5	4,98E-07	-2,61E-08	0,966
13,5	1,44E-07	-4,88E-09	0,953
22,5	5,10E-08	-1,48E-10	0,158
31,5	2,14E-08	7,33E-10	0,976
40,5	6,30E-09	1,55E-10*	0,631
49,5	3,15E-09	4,54E-11	0,882
58,5	1,89E-09	-8,70E-12*	0,379
67,5	1,18E-09	-7,32E-12	0,802
76,5	1,11E-09	2,66E-11*	0,151

Note. * - insignificant coefficient.

Conclusions

Statistical estimation of regression equations for significance was carried out on the basis of the F -test and estimation of the determination coefficient. With 90% confidence, it can be argued that the considered dependence is statistically significant.

The analysis shows that between obtained angular distributions of intensity and wavelength) in transparent windows for both cloudless and cloudy for the atmosphere, there is a link that can be with a good degree of accuracy to describe hyperbolic regression equation.

Between the brightness and the upper boundary of cloud, there is a link that can be with a degree of accuracy to describe linear regression equation.

References

1. Gendrina I.Yu., Kvach A.S., "The Monte Carlo method for determining the vision system characteristics", J. of International Scientific Publication: Education Alternatives, 11(1), 236 – 244 (2013).
2. Zuev V.E., Belov V.V., Veretennikov V.V., [Systems with applications in scattering media], Publishing House "Spectrum" Institute of Atmospheric Optics of the Siberian Branch of the Russian Academy of Sciences, 157-168 (1997).
3. Marchuk G.I., Mikhailiov G.A., Nazaraliev M.A., Darbinjan, Kargin B.A., Elepov B.S., [The Monte Carlo method in atmospheric optics], Nauka, Novosibirsk, 4-95 (1976).
4. Deirmendjian D., [Electromagnetic scattering on spherical polydispersions], American Elsevier Pub. Co., 1-290 (1969).
5. Gendrina I.Yu., Alekseenko M. A., "The regression analysis of statistical simulation results in vision systems through the atmosphere", Izvestija-vuzov-fizika, 58 (8/2), 294-296 (2015).
6. Khayer M. M. et al., "Evaluation of a 5-Year Cloud and Radiative Property Dataset Derived from GOES-8 Data Over the Southern Great Plains", Twelfth ARM Science Team Meeting Proceedings, 1-14 (2002).

ОДНОРАНГОВАЯ АППРОКСИМАЦИЯ ПОЛОЖИТЕЛЬНЫХ МАТРИЦ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ИДЕМПОТЕНТНОЙ МАТЕМАТИКИ

RANK-ONE APPROXIMATION OF POSITIVE MATRICES USING METHODS OF IDEMPOTENT MATHEMATICS

Д-р ф.-м. н. Кривулин Н. К.¹, студент Романова Е. Ю.²

Математико-механический факультет – Санкт-Петербургский государственный университет, Россия
Nkk@math.spbu.ru¹, Romanova.ej@gmail.com²

Abstract: Low-rank matrix approximation is widely used in the analysis of big data, in recommendation systems in the Internet, for approximation solution of some equations in mechanics, and other fields. In many applications it makes sense to use matrices of unit rank for approximating since they have the simplest structure. This article provides a method for approximating positive matrices by matrices of unit rank based on the minimization of log-Chebyshev distance. The approximation problem is reduced to the optimization problem, which has a compact representation in terms of an idempotent semifield that taking maximum in the role of addition. Such semifield is often called the max-algebra. The necessary definitions and results of tropical mathematics are given and the solution of the optimization problem is derived from them. Then the solution is represented in terms of the original approximation problem. As a result, all the positive matrices which provide the minimum of approximation error are obtained in explicit form.

KEYWORDS: IDEMPOTENT MATHEMATICS, TROPICAL MATHEMATICS, IDEMPOTENT SEMIFIELD, RANK-ONE MATRIX APPROXIMATION, LOG-CHEBYSHEV DISTANCE

1. Введение

К задаче аппроксимации матриц сводится значительное число прикладных задач из разных областей. Многие вычислительные задачи, например, задачи вычислительной гидродинамики и теории электрических цепей, уравнения балансов и сохранения в механике, требуют решения системы линейных алгебраических уравнений. Применение техники малоранговой аппроксимации повышает эффективность методов, которые находят приближенное решение систем [1, 2].

Потребность в аппроксимации возникает при разработке рекомендательных систем в сети Интернет [3, 4] и при обработке больших массивов данных [5, 6]. Аппроксимация матрицами из выбранного множества матриц дает возможность представить данные в удобной и корректной с математической точки зрения форме. В работах [5, 7, 8] малоранговая аппроксимация матриц используется для решения задач аппроксимации тензоров высоких порядков, возникающих в многомерном анализе данных и методах обработки сигналов в системах телекоммуникации.

Понижение ранга матрицы при помощи аппроксимации существенно упрощает ее структуру и позволяет сократить объем памяти для ее хранения. Особое значение приобретает аппроксимация матрицами единичного ранга, которые устроены наиболее просто. Так, если исходная матрица размера $n \times n$ определяется n^2 элементами, то матрица, полученная в результате одноранговой аппроксимации, задается всего $2n$ элементами. Подобное сжатие информации сопряжено с некоторыми потерями, поэтому такие методы следует применять тогда, когда матрица предположительно имеет малый ранг.

Во многих приложениях предположение о единичном ранге матрицы вполне оправдано. Например, с помощью одноранговой аппроксимации могут решаться задачи, возникающие в области машинного обучения [9], технического зрения [10] и в статистике [11]. Некоторые методы одноранговой аппроксимации описаны в работах [6, 12].

Задача аппроксимации квадратной матрицы $A \in \mathbb{R}^{n \times n}$ матрицами X из некоторого подмножества $S \subset \mathbb{R}^{n \times n}$ формулируется как задача оптимизации

$$\min_{X \in S} d(A, X),$$

где d — функция расстояния на множестве матриц $\mathbb{R}^{n \times n}$, измеряющая величину ошибки аппроксимации.

Подходы к решению задачи аппроксимации могут варьироваться в зависимости от постановки исходной задачи и осо-

бенностей матрицы. Различия между подходами во многом определяются выбором функции расстояния.

Распространенным решением проблемы является применение к аппроксимации матриц разноранговых методов наименьших квадратов, в основе которого лежит минимизация евклидова расстояния. Варианты применения такого подхода описаны в работах [13, 14]. В работе [15] освещается использование расстояния Минковского (l_p -метрики).

Интерес представляет предельный случай l_p -метрики при $p \rightarrow \infty$, называемый метрикой Чебышева. В силу равномерности расстояния Чебышева, его использование в решении задачи аппроксимации должно обеспечивать хорошее приближение исходной матрицы. В работе [15] доказывается существование приближения Чебышева с рангом r для любой матрицы A с большим рангом и результат сравнивается с приближением в метрике Минковского с конечным p . Проблема чебышевской аппроксимации сформулирована в [16] в виде задачи линейного программирования, к решению которой могут применяться соответствующие методы, например, симплексный метод. В работе [17] доказано, что даже в случае аппроксимации матрицами единичного ранга задача чебышевской аппроксимации является NP -полной, однако может быть решена за полиномиальное время, если исходная матрица неотрицательна.

Для аппроксимации положительных матриц иногда целесообразнее перейти к оценке погрешности в логарифмическом масштабе. Задача минимизации лог-чебышевского расстояния может быть сведена к задаче конического программирования второго порядка, как в работе [18], и решена, например, барьерным методом [16]. В работе [19] для лог-чебышевской аппроксимации матриц предлагается применять методы тропической (идемпотентной) математики — области математики, которая изучает теорию и приложения алгебраических систем с идемпотентными операциями. Также тропический подход к аппроксимации матриц затрагивается в [20]. В этой работе получено частное решение задачи аппроксимации матрицами единичного ранга.

Далее в статье предлагается метод аппроксимации положительных матриц матрицами единичного ранга, который находит в явном виде все матрицы, на которых достигается минимум погрешности аппроксимации, путем минимизации лог-чебышевского расстояния между матрицами. Будет показано, что задача минимизации лог-чебышевского расстояния может быть приведена к задаче, записанной в компактной

форме в терминах идемпотентного полуполя с операцией вычисления максимума в роли сложения, которое часто называют \max -алгеброй. Затем для нахождения решения будут использованы результаты [20, 21, 22] из области тропической математики.

2. Предварительные результаты

Опишем некоторые результаты, необходимые для построения решения задачи одноранговой аппроксимации.

2.1. Одноранговая \log -чебышевская аппроксимация

Чебышевская аппроксимация положительной матрицы $A = (a_{ij})$ при помощи положительной матрицы $X = (x_{ij})$ в логарифмической шкале использует функцию расстояния

$$d(A, X) = \max_{i,j} |\log a_{ij} - \log x_{ij}|,$$

где логарифм берется по основанию больше единицы.

Справедливо следующее утверждение.

Утверждение 1. Пусть A, X — положительные матрицы. Минимизация по X величины $d(A, X)$ эквивалентна минимизации

$$d'(A, X) = \max_{i,j} \max(a_{ij}x_{ij}^{-1}, x_{ij}a_{ij}^{-1}).$$

Следовательно, задача \log -чебышевской аппроксимации может быть сведена к задаче

$$\min_X d'(A, X).$$

Принимая во внимание то, что любая матрица X ранга 1 может быть представлена в форме $X = st^T$, где векторы $s = (s_i)$ и $t = (t_i)$ не содержат нулевых элементов, целевую функцию рассматриваемой задачи можно записать так:

$$d'(A, X) = d'(A, st^T) = \max_{i,j} \max(s_i^{-1}a_{ij}t_j^{-1}, s_i a_{ij}^{-1}t_j).$$

В результате задача одноранговой аппроксимации принимает следующий вид:

$$(1) \quad \min_{s,t} \max_{i,j} \max(s_i^{-1}a_{ij}t_j^{-1}, s_i a_{ij}^{-1}t_j),$$

где минимум берется по всем положительным векторам s и t .

2.2. Элементы тропической математики

Приведем основные определения, обозначения и предварительные результаты тропической (идемпотентной) математики [20, 21, 22], на которые будем опираться в дальнейшем. Для более детального изучения могут быть использованы работы [23, 24, 25].

2.2.1. Идемпотентное полуполе

Идемпотентным полуполем $\mathbf{R}_{\max, \oplus}$ называется алгебраическая система $(\mathbf{R}_+, \oplus, \otimes, 0, 1)$, где \mathbf{R}_+ — множество неотрицательных вещественных чисел, операция сложения определена как взятие максимума двух чисел и имеет нейтральный элемент 0, а умножение \otimes определено как арифметическое умножение с нейтральным элементом 1. Заметим, что сложение является идемпотентным, то есть удовлетворяет условию $x \oplus x = x$ для всех $x \in \mathbf{R}_+$, и выполняется свойство дистрибутивности умножения относительно сложения. Понятия обратного элемента и степени имеют обычный смысл. Такое полуполе обычно называют \max -алгеброй.

2.2.2. Матрицы и векторы

Матрица называется разложимой, если перестановкой строк вместе с такой же перестановкой столбцов ее можно привести к блочно-треугольному виду. В противном случае матрица называется неразложимой. Сложение и умножение двух матриц подходящего размера и умножение матрицы на число выполняются по стандартным правилам с заменой обычных арифметических операций на операции \oplus и \otimes . Единичную матрицу будем обозначать символом I .

Для любой ненулевой матрицы $A = (a_{ij}) \in \mathbf{R}_+^{m \times n}$ определена мультипликативно сопряженная матрица $A^- = (a_{ij}^-) \in \mathbf{R}_+^{n \times m}$ с элементами $a_{ij}^- = a_{ji}^{-1}$, если $a_{ji} \neq 0$, и $a_{ij}^- = 0$ в противном случае.

След матрицы $A = (a_{ij}) \in \mathbf{R}_+^{n \times n}$ вычисляется по формуле $\text{tr} A = a_{11} \oplus \dots \oplus a_{nn}$.

Для любой матрицы $A \in \mathbf{R}_+^{n \times n}$ введем в рассмотрение матрицу $A^* = I \oplus A \oplus \dots \oplus A^{n-1}$.

Вектор называется регулярным, если он не имеет нулевых компонент. Для любого ненулевого вектора-столбца $x = (x_i) \in \mathbf{R}_+^n$ определен вектор-строка $x^- = (x_i^-)$, где $x_i^- = x_i^{-1}$, если $x_i \neq 0$, и $x_i^- = 0$ — иначе.

2.2.3. Собственное число и вектор матрицы

Число $\lambda \in \mathbf{R}_+$ и ненулевой вектор $x \in \mathbf{R}_+^n$ называются собственным значением и собственным вектором матрицы $A \in \mathbf{R}_+^{n \times n}$, если они удовлетворяют равенству $Ax = \lambda x$.

Любая матрица A порядка n имеет собственное число, которое вычисляется по формуле

$$\lambda = \bigoplus_{m=1}^n \text{tr}^{1/m}(A^m).$$

Число λ называется спектральным радиусом матрицы.

3. Задача тропической оптимизации

Предположим, что задана матрица $A \in \mathbf{R}_+^{n \times n}$ и требуется решить задачу минимизации

$$(2) \quad \min_{x,y} x^- A y \oplus y^- A^- x,$$

где минимум берется по всем регулярным векторам $x, y \in \mathbf{R}_+^n$.

В работе [20] найдено следующее решение.

Теорема 1. Пусть A — неразложимая матрица, μ — спектральный радиус матрицы AA^- . Тогда минимум в задаче (2) равен $\mu^{1/2}$ и достигается, когда x и $y = \mu^{-1/2} A^- x$ — собственные векторы матриц AA^- и $A^- A$, соответствующие μ .

Теперь представим новый результат, который определяет все векторы x и y , обеспечивающие минимум в задаче (2), и тем самым дает полное решение этой задачи.

Теорема 2. Пусть A — ненулевая матрица, а μ — спектральный радиус матрицы AA^- . Тогда минимум в задаче (2) равен $\mu^{1/2}$ и достигается тогда и только тогда, когда

$$x = (\mu^{-1} AA^-)^* v \oplus \mu^{-1/2} A (\mu^{-1} A^- A)^* w, \\ y = \mu^{-1/2} A^- (\mu^{-1} AA^-)^* v \oplus (\mu^{-1} A^- A)^* w,$$

где v, w — произвольные регулярные векторы.

В частности, минимум достигается, когда x и $y = \mu^{-1/2} A^- x$ — собственные векторы матриц AA^- и $A^- A$, соответствующие μ .

4. Решение задачи аппроксимации

Рассмотрим задачу одноранговой аппроксимации (1). При замене арифметических операций на операции идемпотентного полуполя $\mathbf{R}_{\max, \times}$ получим целевую функцию в виде

$$\bigoplus_{i,j} (s_i^{-1} a_{ij} t_j^{-1} \oplus s_i a_{ij}^{-1} t_j) = s^{-1} A (t^{-1})^T \oplus t^T A^{-1} s.$$

Таким образом, задача (1) принимает вид

$$\min_{s,t} s^{-1} A (t^{-1})^T \oplus t^T A^{-1} s,$$

где минимум берется по всем регулярным векторам s и t .

Положив $x = s$, $y = (t^{-1})^T$, получим задачу тропической оптимизации в форме (2). В силу того, что матрица A — положительная, матрица AA^{-1} тоже является положительной и потому имеет спектральный радиус $\mu > 0$. Таким образом, выполнены условия теоремы 2. Применяя теорему, получим решение задачи аппроксимации в виде следующего утверждения.

Теорема 3. Пусть A — положительная матрица, μ — спектральный радиус матрицы AA^{-1} . Тогда минимальная погрешность аппроксимации в \log -чебышевском смысле матрицы A матрицами единичного ранга, равна $\mu^{1/2}$ и достигается на матрицах вида st^T , где

$$s = (\mu^{-1} AA^{-1})^* v \oplus \mu^{-1/2} A (\mu^{-1} A^{-1} A)^* w,$$

$$t^T = (\mu^{-1/2} A^{-1} (\mu^{-1} AA^{-1})^* v \oplus (\mu^{-1} A^{-1} A)^* w)^{-}$$

и v , w — произвольные регулярные векторы размера n .

В частности, минимальная погрешность достигается, когда s и $t^T = \mu^{1/2} (A^{-1} s)^{-}$ — собственные векторы матриц AA^{-1} и $A^{-1}A$, соответствующие μ .

5. Заключение

В настоящей статье исследована проблема аппроксимации матриц матрицами более низкого ранга и приведено решение задачи одноранговой аппроксимации для случая положительных матриц. Решение основывается на представлении задачи аппроксимации в виде задачи оптимизации и результатах из области тропической математики. Представленный метод находит в явном виде все матрицы, на которых достигается минимум погрешности аппроксимации и имеет трудоемкость порядка $O(n^4)$. В силу того, что все векторы, образующие аппроксимирующую матрицу, определяются явными формулами, становится возможным дальнейшее изучение свойств аппроксимирующих матриц и наложение на них ограничений, соответствующих конкретной задаче.

6. Литература

[1] Соловьев С. А. Решение разреженных систем линейных уравнений методом Гаусса с использованием техники аппроксимации матрицами малого ранга // Выч. мет. программирование. 2014. Т. 15, № 3. С. 441-460.
 [2] Воронин К. В., С. А. Соловьев. Решение уравнения Гельмгольца с использованием метода малоранговой аппроксимации в качестве предобуславливателя // Выч. мет. программирование. 2015. Т. 16, № 2. С. 268-280.
 [3] Lee J., S. Kim, G. Lebanon, Y. Singer. Local low-rank matrix approximation // Proc. 30th Intern. Conf. on Machine Learning, Atlanta, Georgia, USA, 2013. P. 82-90.
 [4] Lee J., S. Bengio, S. Kim et al. Local collaborative ranking // Proc. 23rd Intern. Conf. on World Wide Web (WWW'14), April 7-11, 2014, Seoul, Korea. ACM, 2014. P. 85-96.

[5] Savas B. Algorithms in data mining using matrix and tensor methods // Linköping Studies in Science and Technology: Dissertations, Vol. 1178. Linköping Univ. Tech., 2008. 138 p.
 [6] Ispany M., G. Michaletzky, J. Reiczig et al. Approximation of non-negative integer-valued matrices with application to Hungarian mortality data // Proc. 19th Intern. Symp. on Mathematical Theory of Networks and Systems (MTNS 2010), 5-9 July, 2010, Budapest, Hungary. P. 831-838.
 [7] Тартышников Е. Е. Матрицы, тензоры, вычисления. Учебный курс. М.: МГУ им. М. В. Ломоносова, 2013.
 [8] Wang H., N. Ahuja. A tensor approximation approach to dimensionality reduction // Int. J. Comput. Vis. 2008. Vol. 76, N 3. P. 217-229. DOI:10.1007/s11263-007-0053-0
 [9] Yao Q., J. Kwok. Greedy learning of generalized low-rank models // Proc. 25th Intern. Joint Conf. on Artificial Intelligence (IJCAI'16), New York, 2016. AAAI Press, 2016. P. 2294-2300.
 [10] Шляnnиков А. В. Алгоритм восстановления трехмерной модели лица по фотографии // Научно-технический вестник ИТМО. 2010. Т. 69, № 5. С. 86-90.
 [11] Aissa-El-Bey A., K. Seghouane. Sparse canonical correlation analysis based on rank-1 matrix approximation and its application for fMRI signals // 2016 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China. IEEE, 2016. P. 4678-4682. DOI:10.1109/ICASSP.2016.7472564
 [12] Luss R., M. Teboulle. Conditional gradient algorithms for rank-one matrix approximations with a sparsity constraint // SIAM Review. 2013. Vol. 55, N 1. P. 65-98. DOI:10.1137/110839072
 [13] Саати Т. Принятие решений. Метод анализа иерархий / Пер. с англ. П. Г. Вачнадзе. М.: Радио и связь, 1993. 315 с.
 [14] Shen H., J. Huang. Sparse principal component analysis via regularized low rank matrix approximation // J. Multivariate Anal. 2008. Vol. 99, N 6. P. 1015-1034. DOI:10.1016/j.jmva.2007.06.007
 [15] Zietak K. The Chebyshev approximation of a rectangular matrix by matrices of smaller rank as the limit of l_p -approximation // J. Comput. Appl. Math. 1984. Vol. 11. P. 297-305. DOI:10.1016/0377-0427(84)90004-9
 [16] Boyd S., L. Vandenberghe. Convex optimization. Cambridge: Cambridge Univ. Press, 2004.
 [17] Gillis N., Y. Shitov. Low-rank matrix approximation in the infinity norm // CoRR. 2017. arXiv:1706.00078.
 [18] Lobo M. S., L. Vandenberghe, S. Boyd, H. Lebret. Applications of second-order cone programming // Linear Algebra Appl. 1998. Vol. 284. P. 193--228. DOI:10.1016/S0024-3795(98)10032-0
 [19] Кривулин Н. К., В. А. Агеев, И. В. Гладких. Применение методов тропической оптимизации для оценки альтернатив на основе парных сравнений // Вестник Санкт-Петербургского университета. Прикладная математика. Информатика. Процессы управления. 2017. Т. 13, Вып. 1. С. 27-41. DOI:10.21638/11701/spbu10.2017.103
 [20] Кривулин Н. К. Методы идемпотентной алгебры в задачах моделирования и анализа сложных систем. СПб: С.-Петерб. ун-т, 2009. С. 107-108.
 [21] Krivulin N. K. Eigenvalues and eigenvectors of matrices in idempotent algebra // Vestnik St. Petersburg Univ. Math. 2006. Vol. 39, N 2. P. 72-83.
 [22] Krivulin N. K. Extremal properties of tropical eigenvalues and solutions to tropical optimization problems // Linear Algebra Appl. 2015. Vol. 468. P. 211-232. DOI:10.1016/j.laa.2014.06.044
 [23] Маслов В. П., В. Н. Колокольцев. Идемпотентный анализ и его применение в оптимальном управлении. М.: Физматлит, 1994. 144 с.
 [24] Pin J.-E. Tropical semirings // Idempotency / Ed. by J. Guna-wardena. Cambridge: Cambridge Univ. Press, 1998. P. 50-69.
 [25] Литвинов Г. Л., В. П. Маслов, А. Н. Соболевский. Идемпотентная математика и интервальный анализ // Вычислительные технологии. 2001. Т. 6, № 6. С. 47-70.

ИСПОЛЬЗОВАНИЕ РАЗРЕЖЕНИЯ МАТРИЦ ДЛЯ РЕШЕНИЯ МНОГОМЕРНОЙ ЗАДАЧИ ТРОПИЧЕСКОЙ ОПТИМИЗАЦИИ

SOLUTION OF MULTIDIMENSIONAL TROPICAL OPTIMIZATION PROBLEM WITH THE USE OF MATRIX SPARSIFICATION

Д-р ф.-м. н. Кривулин Н. К.¹, аспирант Сорокин В. Н.²

Математико-механический факультет – Санкт-Петербургский государственный университет, Россия
nkk@math.spbu.ru¹, SovanSB@gmail.com²

Abstract: A complete solution is proposed for a problem of vector-valued function minimization with elements from a tropical (idempotent) semifield. The tropical optimization problem, considered here, arises when one needs, for instance, to find the best, in the sense of Chebyshev metric, approximate solution for tropical vector equations, and occurs in various applications, including scheduling, location and decision-making problems. To solve the problem, first, the minimum value of the objective function is obtained, a characterization of the solution set in the form of a system of inequalities is proposed, and one of the solutions is presented. Then, with introduction of matrix sparsification into the problem, an extended set of solutions, and then a complete solution in the form of a family of subsets are derived. Procedures, allowing to reduce the number of subsets, which one needs to examine when constructing the complete solution, are described in the present paper. It is shown how the complete solution can be represented in parametric way in a compact vector form.

Keywords: idempotent semifield, tropical optimization, Chebyshev approximation, complete solution, matrix sparsification

1. Введение

Тропическая (идемпотентная) алгебра представляет собой раздел математики, который изучает свойства полуколец и полуполей с идемпотентным сложением и их приложения. Исследованиям в этой области посвящено немало работ, включая [1 - 7], а также детальный обзор литературы в [8]. Одним из направлений развития тропической математики является разработка методов решения задач оптимизации, которые могут быть сформулированы и решены в терминах идемпотентной математики (задач тропической оптимизации). Имеется целый ряд практических задач (см., например, [9 - 12]), которые сводятся к наилучшему приближенному решению в смысле метрики Чебышева векторного уравнения $Ax = p$, где A и p обозначают заданные матрицу и вектор, x — неизвестный вектор, а произведение матрицы на вектор понимается в смысле тропической математики.

Проблема чебышевской аппроксимации для решения рассматриваемого уравнения сводится к задаче поиска векторов x , на которых достигается минимум в задаче

$$\min (Ax)^- p \oplus p^- Ax,$$

где матричные и векторные операции понимаются в смысле идемпотентной алгебры.

Исследованию задачи посвящен ряд работ, опубликованных в различное время, включая [3, 10, 12, 13]. Представленные в этих работах результаты обычно сводятся к получению одного из решений и не позволяют найти все множество решений задачи.

Для решения задачи в статьях [13 - 16] предлагается подход, при котором вводится дополнительный параметр для обозначения минимума целевой функции, а затем задача сводится к решению параметризованных неравенств. С помощью такого подхода были получены решения в компактной векторной форме для рассматриваемой задачи, а также некоторых ее вариантов, включая задачи с ограничениями.

Цель настоящей работы состоит в исследовании и решении задачи тропической оптимизации с целевой функцией более общего вида. На основе применения методов решения с использованием разреженных матриц, разработанного в [17], находится полное решение задачи и его представление в компактной векторной форме.

Работа построена следующим образом. Сначала находится минимальное значение целевой функции задачи, предлагается описание множества решений в форме системы неравенств и приводится одно из решений. Далее с помощью разрежения матрицы задачи находится расширенное множество решений, а

затем полное решение в виде некоторого семейства подмножеств. Предлагаются процедуры, позволяющие сократить число подмножеств, которые необходимо исследовать при построении полного решения.

2. Элементы тропической алгебры

Рассмотрим необходимые в дальнейшем основные понятия и предварительные результаты тропической (идемпотентной) алгебры [1 - 7]. Идемпотентное полукольцо определяется как набор $(X, \oplus, \otimes, 0, 1)$, где X обозначает непустое множество с заданными на нем операциями сложения \oplus с нейтральным элементом 0 (ноль) и умножения \otimes с нейтральным элементом 1 (единица).

Сложение идемпотентно: для любого элемента $x \in X$ выполняется равенство $x \oplus x = x$. Умножение обратимо: для любого $x \neq 0$ существует x^{-1} такой, что $x \otimes x^{-1} = 1$.

Идемпотентность сложения задает на множестве X частичный порядок: $x \leq y$ тогда и только тогда, когда $x \oplus y = y$. Отсюда следует, что неравенство $x \oplus y \leq z$ эквивалентно двум неравенствам $x \leq z$ и $y \leq z$. Будем предполагать, что указанный частичный порядок продолжен до линейного на множестве X . Целые степени определяются обычным образом: $x^0 = 1$, $x^p = x \otimes x^{p-1}$, $x^{-p} = (x^{-1})^p$, $0^p = 0$ для всех $x \neq 0$ и натуральных p . В дальнейшем будем считать полуполе алгебраически полным в том смысле, что введенная операция возведения в степень может быть распространена на случай степеней с рациональным показателем. Далее символ \otimes будем опускать для упрощения записи.

В прикладных задачах часто встречаются следующие вещественные идемпотентные полуполя: $R_{\max,+} = (R \cup \{-\infty\}, \max, +, -\infty, 0)$, $R_{\min,+} = (R \cup \{+\infty\}, \min, +, +\infty, 0)$, $R_{\max,\times} = (R_+ \cup \{0\}, \max, \times, 0, 1)$ и $R_{\min,\times} = (R_+ \cup \{+\infty\}, \min, \times, +\infty, 1)$. Здесь $R_+ = \{x \in R | x > 0\}$ — множество положительных вещественных чисел.

Рассмотрим полуполе $R_{\max,+}$, которое обычно называют $(\max, +)$ -алгеброй. В нем множество X определено как множество действительных чисел R , расширенное путем добавления числа $-\infty$. Роль тропического сложения играет операция взятия максимума, а в качестве умножения берется арифметическое сложение. Для любого $x \neq 0$ существует обратный по умножению элемент x^{-1} , равный противоположному числу $-x$ в обычной арифметике. Степень x^y определена для всех $x, y \in R$ и соответствует арифметическому произведению $x \cdot y$. Индуцированный

идемпотентным сложением порядок совпадает с обычным линейным порядком на R .

Обозначим через $X^{m \times n}$ множество матриц размера $m \times n$ над полуполем X .

В нулевой матрице все элементы равны 0 . Матрицу, в которой отсутствуют нулевые строки (столбцы), назовем регулярной по строкам (столбцам). Матрица без нулевых строк и нулевых столбцов называется регулярной.

Операции над матрицами выполняются по стандартным правилам с заменой обычных скалярных операций сложения и умножения на тропические. Будем называть мультипликативно сопряженным транспонированием ненулевой матрицы $A \in X^{m \times n}$ операцию преобразования в матрицу $A^- \in X^{n \times m}$, элементы которой определяются по правилу: $a_{ij}^- = a_{ji}^{-1}$, если $a_{ji} \neq 0$, и $a_{ij}^- = 0$ в противном случае.

Рассмотрим множество $X^{n \times n}$ квадратных матриц порядка n . Матрица является диагональной, если все ее недиагональные элементы равны нулю. Единичной называется диагональная матрица I , у которой все элементы на диагонали равны единице.

Для регулярных по строкам матриц A выполняется неравенство $AA^- \geq I$.

Матрица, состоящая из одного столбца (строки), образует вектор-столбец (вектор-строку). Далее все векторы, если не указано иначе, считаются вектор-столбцами. Множество вектор-столбцов размерности n обозначим через X^n . Нулевой вектор имеет все компоненты равными 0 . Вектор -- регулярный, если у него нет нулевых компонент. Вектор, все компоненты которого равны 1 , обозначается как $e = (1, \dots, 1)^T$.

Для ненулевого вектора $x \in X^n$ справедливо равенство $x^-x = 1$. Если вектор x является регулярным, то выполняется неравенство $xx^- \geq I$.

Выпуклой линейной комбинацией векторов x_1, \dots, x_k называется выражение вида $u_1x_1 \oplus \dots \oplus u_kx_k$, где числа u_1, \dots, u_k удовлетворяют условию $u_1 \oplus \dots \oplus u_k = 1$. Это условие означает, что $u_i \leq 1$ для всех индексов $i = 1, \dots, k$, и по крайней мере для одного i выполняется равенство $u_i = 1$.

Пусть заданы матрица $A \in X^{m \times n}$, а также вектор $d \in X^m$, и требуется найти все векторы $x \in X^n$, удовлетворяющие неравенству

$$(1) \quad Ax \leq d.$$

Решение этой задачи обеспечивается следующим утверждением, полное доказательство которого приводится, например, в работах [7, 18].

Лемма 1.1 Для любой регулярной по столбцам матрицы A и регулярного вектора d решение неравенства (1) имеет вид $x \leq (d^-A)^-$.

3. Задачи тропической оптимизации

Задачи тропической оптимизации обычно состоят в минимизации или максимизации некоторой целевой функции, заданной на векторах над идемпотентным полуполем. Такие задачи возникают, например, при исследовании уравнения $Ax = p$, для которого требуется найти точное или приближенное решение [3, 10 - 12].

Сначала предположим, что заданы матрица $A \in X^{n \times n}$ и вектор $p \in X^n$. Пусть требуется найти регулярные векторы $x \in X^n$, которые решают задачу

$$(1) \quad \min (Ax)^-p \oplus p^-Ax.$$

Решение этой задачи обеспечивает наилучшее приближенное решение уравнения $Ax = p$ в смысле чебышевской метрики. Исследование задачи было проведено, например, в работах [13, 15, 16], где приводится частичное решение в следующем виде.

Лемма 2 Для любых регулярных матрицы A и вектора p минимум в задаче (2) равен $\Delta = ((A(p^-A)^-)^-p)^{1/2}$ и достигается при $x = \Delta(p^-A)^-$.

Обобщением задачи (2) с дополнительным вектором $q \in X^n$ является задача нахождения регулярных векторов x , которые обеспечивают

$$(2) \quad \min (Ax)^-p \oplus q^-Ax.$$

В настоящей работе задача (3) рассматривается в следующей форме. Пусть заданы матрица $A \in X^{n \times n}$ и векторы $p, q \in X^n$. Требуется найти все регулярные векторы $x \in X^n$, на которых достигается

$$(3) \quad \min (Ax)^-p \oplus q^-x.$$

В работах [14, 16] было получено частичное решение этой задачи.

Ниже для решения задачи (4) сначала также, как в работах [14, 16], будет определен минимум целевой функции и получено одно из решений. Затем для полного решения задачи строится система неравенств, которая определяет множество всех решений. На основе использования разреженных матриц находится более широкое множество решений, а затем полное решение задачи в форме семейства подмножеств решений.

4. Предварительный анализ

Цель этого раздела состоит в том, чтобы найти минимум целевой функции, охарактеризовать множество решений и описать некоторые свойства этого множества. Для этого будем использовать подход [14 - 16, 18], при котором вводится параметр для обозначения минимума целевой функции, а затем задача оптимизации сводится к решению параметризованного неравенства. Справедливо следующее утверждение.

Лемма 3 Пусть A -- регулярная по строкам матрица, а p и q -- регулярные векторы. Тогда минимум в задаче (4) равен

$$(4) \quad \Delta = ((Aq)^-p)^{1/2},$$

а все регулярные решения x определяются системой неравенств

$$(5) \quad Ax \geq \Delta^{-1}p, \quad x \leq \Delta q.$$

В частности, минимум достигается при $x = \Delta q$.

Путем прямой подстановки и проверки на соответствие системе (6) доказывается следующий факт.

Следствие 4 Множество решений задачи (4) вместе с любыми решениями содержит их всевозможные выпуклые линейные комбинации.

Для расширения множества решений задачи (4) применим метод с использованием разрежения матриц, предложенный в [17]. Сначала преобразуем матрицу задачи $A = (a_{ij})$ в разреженную матрицу $\hat{A} = (\hat{a}_{ij})$, приравнявая к нулю все элементы, которые строго меньше порогового значения по следующему правилу:

$$\hat{a}_{ij} = \begin{cases} a_{ij}, & \text{если } a_{ij} \geq \Delta^2 p_i q_j^{-1}; \\ 0, & \text{в противном случае.} \end{cases}$$

Далее матрицу \hat{A} , полученную при помощи такого преобразования, будем называть разреженной матрицей задачи. Ее свойства отражает следующий результат.

Лемма 5 Замена матрицы A на \hat{A} не меняет множество решений задачи (4).

Заметим, что ненулевые элементы матрицы \hat{A} отвечают условию $\hat{a}_{ij} \geq \Delta^{-2} p_i q_j^{-1}$. Тогда для элементов матрицы $\hat{A}^- = (\hat{a}_{ij}^-)$ справедливо соотношение $\hat{a}_{ij}^- \leq \Delta^2 q_i p_j^{-1}$. В матричной форме имеем неравенство $\hat{A}^- \leq \Delta^2 q p^-$, которое будет использовано ниже.

5. Полное решение задачи

В этом разделе будет представлено полное решение задачи (4) в виде семейства решений, которое задается множеством матриц, полученных из разреженной матрицы задачи путем дальнейшего обнуления ее элементов.

Начнем с того, что расширим решение $x = \Delta q$, полученное в лемме 3, до некоторого множества с помощью следующего утверждения.

Лемма 6 Пусть A -- регулярная по строкам разреженная матрица задачи (4), где p и q -- регулярные векторы, и $\Delta = ((Aq)^{-}p)^{1/2}$. Тогда минимум в задаче (4) равен Δ и достигается на любом векторе x , который удовлетворяет условию

$$(6) \quad \Delta^{-1}A^{-}p \leq x \leq \Delta q.$$

Теперь можно сформулировать лемму, которая обеспечивает полное решение задачи (4).

Лемма 7 Пусть A -- регулярная по строкам разреженная матрица задачи (4), где p и q -- регулярные векторы, и $\Delta = ((Aq)^{-}p)^{1/2}$. Обозначим через \mathcal{A} множество матриц, полученных из A путем сохранения по одному ненулевому элементу в каждой строке и обнулению остальных.

Тогда минимум в задаче (4) равен Δ , а все регулярные решения x образуют семейство решений, каждое из которых определяется условием

$$(7) \quad \Delta^{-1}A_1^{-}p \leq x \leq \Delta q, \quad A_1 \in \mathcal{A}.$$

Отметим, что у различных множеств решений из семейства, описанного в лемме 7, могут быть пересекающиеся подмножества.

Решение в параметрическом виде предоставляет следующая теорема.

Теорема 8 Пусть A -- регулярная по строкам разреженная матрица задачи (4), где p и q -- регулярные векторы, и $\Delta = ((Aq)^{-}p)^{1/2}$. Обозначим через \mathcal{A} множество матриц, полученных из A путем сохранения по одному ненулевому элементу в каждой строке и обнулению остальных, а через B -- матрицу, столбцами которой являются векторы $b_1 = \Delta^{-1}A_1^{-}p$ для всех матриц $A_1 \in \mathcal{A}$.

Тогда минимум в задаче (4) равен Δ , а все регулярные решения x имеют вид

$$(8) \quad x = Bu \oplus v, \quad v \leq \Delta q, \quad e^T u = 1.$$

В качестве следствия полученного результата представим решение задачи (2), которая является частным случаем задачи (4).

Следствие 9 Пусть A -- регулярная по строкам разреженная матрица задачи (2), где p -- регулярный вектор и $\Delta = ((A(p^{-}A)^{-})^{-}p)^{1/2}$. Обозначим через \mathcal{A} множество матриц, полученных из A путем сохранения по одному ненулевому элементу в каждой строке и обнулению остальных, а через B -- матрицу, столбцами которой являются векторы $b_1 = \Delta^{-1}A_1^{-}p$ для всех матриц $A_1 \in \mathcal{A}$.

Тогда минимум в задаче (4) равен Δ , а все регулярные решения x имеют вид

$$x = Bu \oplus v, \quad v \leq \Delta(p^{-}A)^{-}, \quad e^T u = 1.$$

6. Процедуры построения решения

Перебор всевозможных матриц $A_1 \in \mathcal{A}$ для построения подмножеств семейства решений задачи в соответствии с результатом теоремы 8 может представлять определенные трудности. Ниже описываются процедуры, позволяющие во многих случаях сократить число подмножеств, которые необходимо учесть при построении общего решения.

Рассмотрим следующую процедуру построения нижних границ в неравенстве (8) для семейства решений, которая

заключается в последовательным выборе по одному ненулевому элементу в строках разреженной матрицы \hat{A} .

Предположим, что в матрице \hat{A} зафиксированы элементы в некоторых строках, в результате чего получена матрица $\tilde{A} = (\tilde{a}_{ij})$. Пусть выбран ненулевой элемент в одной из оставшихся строк, например элемент \tilde{a}_{rj} в строке r и столбце j , значение которого фиксируется, а остальные элементы в этой строке замещаются нулями.

Всякий вектор $x = (x_i)$, который является решением задачи (4), удовлетворяет системе (6), в частности, ее первому неравенству в форме $\tilde{A}x \geq \Delta^{-1}p$. Скалярное неравенство для строки r , где все элементы кроме \tilde{a}_{rj} равны 0, записывается в виде $\tilde{a}_{rj}x_j \geq \Delta^{-1}p_r$, что эквивалентно неравенству $x_j \geq \Delta^{-1}\tilde{a}_{rj}^{-1}p_r$.

В столбце j возьмем элемент \tilde{a}_{sj} , который расположен в одной из еще не рассмотренных строк s . При выполнении условия $\tilde{a}_{sj}p_s^{-1} \geq \tilde{a}_{rj}p_r^{-1}$, которое эквивалентно условию $\tilde{a}_{sj} \geq \tilde{a}_{rj}p_r^{-1}p_s$, имеем $\tilde{a}_{sj}x_j \geq \tilde{a}_{rj}p_r^{-1}p_s\Delta^{-1}\tilde{a}_{rj}^{-1}p_r \geq \Delta^{-1}p_s$. Тогда неравенство $a_{s1}x_1 \oplus \dots \oplus a_{sn}x_n \geq \Delta^{-1}p_s$ в системе (6) выполняется вне зависимости от значений x_l для всех $l \neq j$. В этом случае дальнейшее исследование ненулевых элементов \tilde{a}_{sl} в строке s не может дать новых решений. Эти элементы можно заменить нулями, не нарушая неравенство, что уменьшает число альтернатив, которые требуется рассмотреть.

Заметим, что для проверки условия $\tilde{a}_{sj}p_s^{-1} \geq \tilde{a}_{rj}p_r^{-1}$ удобно предварительно умножить каждую строку i матрицы \hat{A} на элемент p_i^{-1} , а затем исследовать элементы полученной матрицы, которую обозначим через $\hat{A}' = (\hat{a}'_{ij})$.

Если в какой-то строке матрицы \hat{A}' преобладают элементы, являющиеся минимальными ненулевыми элементами в своих столбцах, то, вероятно, стоит начать перебор с подобной строки. Это позволит поочередно фиксировать такие элементы и не рассматривать строки, содержащие остальные ненулевые элементы соответствующих столбцов, а при отсутствии в таком столбце нулевых элементов сразу получать одну из границ.

Отсюда видно, что выбор строки в матрице может сыграть существенную роль для уменьшения количества рассматриваемых подмножеств семейства решений.

Некоторые нижние границы подмножеств, полученные с помощью процедуры построения полного семейства решений, могут быть несущественными в том смысле, что их удаление не повлияет на полученное множество решений. Нетрудно видеть, например, что если для каких-то двух границ b_i и b_j выполняется неравенство $b_i \leq b_j$, то граница b_j может быть удалена из списка без потери решений.

Сформулируем критерий для отбрасывания несущественных границ.

Предложение 10 Пусть B -- матрица, столбцы которой определяют нижние границы для некоторого набора подмножеств семейства решений (8), а $b_1 = \Delta^{-1}A_1^{-}p$, где $A_1 \in \mathcal{A}$, -- еще одна граница. Тогда граница b_1 является несущественной, если выполняется неравенство

$$(9) \quad e^T(b_1^{-}B)^{-} \geq 1.$$

7. Заключение

В работе было получено полное решение задачи минимизации функции, заданной на векторах с элементами из тропического (идемпотентного) полуполя. Решение этой задачи, в частности, обеспечивает наилучшее приближенное решение уравнения $Ax = p$ в смысле чебышевской метрики. Ее исследование уже проводилось в работах [13 - 16], где, однако, было найдено только частичное решение.

Для вывода полного решения вначале представлено частичное решение, которое впоследствии расширяется благодаря использованию техники разрежения матриц, разработанной в исследовании [17]. Общее решение задачи представляется в виде семейства подмножеств решений,

которое также может быть записано в компактной векторной форме в параметрическом виде.

Некоторые семейства решений могут содержаться в других семействах и поэтому при записи общего решения могут быть отброшены. Был разработан критерий отбрасывания подобных несущественных семейств, а также описаны процедуры, которые позволяют сократить число подмножеств, которые необходимо учесть при построении общего решения.

8. Литература

- [1] Baccelli F., G. Cohen, G. J. Olsder, J.-P. Quadrat. Synchronization and Linearity. Chichester : Wiley Series in Probability and Statistics. Wiley, 1993. 514 p.
- [2] Маслов В. П., В. Н. Колокольцов. Идемпотентный анализ и его применение в оптимальном управлении. М.: Физматлит, 1994. 144 с.
- [3] Cuninghame-Green R. A. Minimax algebra and applications - Advances in Imaging and Electron Physics. Vol. 90 / ed. by P. W. Hawkes. San Diego: Academic Press, 1994. P. 1–121.
- [4] Golan J. S. Semirings and Affine Equations over Them. Mathematics and Its Applications. Vol. 556 of Springer, 2003. 241 p.
- [5] Heidergott B, G. J. Olsder, J. van der Woude. Max Plus at Work. Princeton Series in Applied Mathematics: Princeton Univ. Press, 2006. 226 p.
- [6] Gondran M., M. Minoux. Graphs, Dioids and Semirings. Computer Science. of Operations Research Vol. 41, Springer, 2008. 383 p.
- [7] Кривулин Н. К. Методы идемпотентной алгебры в задачах моделирования и анализа сложных систем. СПб.: Изд-во С.-Петербург. ун-та, 2009. 256 с.
- [8] Glazek K. A Guide to the Literature on Semirings and Their Applications in Mathematics and Information Sciences. Springer, 2002. 392 p.
- [9] Воробьев Н. Н. Экстремальная алгебра положительных матриц. Elektron. Informationsverarb. Kybernet. 1967. Bd. 3, N 1. S. 39-72.
- [10] Cuninghame-Green R. A. Projections in minimax algebra. Math. Program. Vol. 10. 1976. P. 111–123.
- [11] Zimmermann K. Some optimization problems with extremal operations. Mathematical Programming at Oberwolfach II. Vol. 22 of Mathematical Programming Studies / ed. by B. Korte, K. Ritter. Berlin: Springer, 1984. P. 237–251.
- [12] Butkovič P., K. P. Tam. On some properties of the image set of a max-linear mapping. Tropical and Idempotent Mathematics. Vol. 495 of Contemporary Mathematics. Ed. by G. L. Litvinov, S. N. Sergeev. Providence: AMS, 2009. P. 115-126.
- [13] Кривулин Н. К. О решении линейных векторных уравнений в идемпотентной алгебре. Математические модели. Теория и приложения. Вып. 5: Сб. науч. статей под ред. М. К. Чиркова. СПб.: BBM, 2004. С. 105–113.
- [14] Krivulin N. A new algebraic solution to multidimensional minimax location problems with Chebyshev distance. WSEAS Trans. Math. Vol. 11, N 7. 2012. P. 605–614.
- [15] Krivulin N. Solution of linear equations and inequalities in idempotent vector spaces. Int. J. Appl. Math. Inform. Vol. 7, N 1. 2013. P. 14–23.
- [16] Krivulin N., K. Zimmermann. Direct solutions to tropical optimization problems with nonlinear objective functions and boundary constraints. Mathematical Methods and Optimization Techniques in Engineering. Ed. by D. Bielek, and H. Walter, I. Utu, C. von Lucken. WSEAS Press, 2013. P. 86–91.
- [17] Krivulin N. Algebraic solution of tropical optimization problems via matrix sparsification with application to scheduling. J. Log. Algebr. Methods Program. 2017. Vol. 89. P. 150–170.
- [18] Krivulin N. Extremal properties of tropical eigenvalues and solutions to tropical optimization problems. Linear Algebra Appl. Vol. 468. 2015. P. 211-232.

ROTATING OF A BALL IN CHAMBER FILLED WITH A FLUID

ВРАЩЕНИЕ ШАРА В КАМЕРЕ, ЗАПОЛНЕННОЙ ЖИДКОСТЬЮ

Prof., Dr. Dementev O.
Chelyabinsk State University, Chelyabinsk, Russia
e-mail: dement@csu.ru

Abstract: Influence of form errors of a chamber filled with a liquid on the movement and stability of a ball, rotating in the chamber ([1-3]), is studied. Two cases of the influence of a chamber form errors on the forces, acting on the ball, are defined. The first case describes the situation when limitations on the rotor shift are not imposed and disturbances of the chamber form are set by spherical harmonics not above the first order. Then the chamber of a disturbed form, from the point of view of the reaction forces of the liquid and their moments, does not differ from a similar spherical chamber. In the second case disturbance of a chamber form are arbitrary and the rotor shift is supposed small. Then the force, acting on the rotor, depends on its displacement only, and the momentum does not depend on shift. A chamber of any form is equivalent to an ellipsoid. A rising here diffractive moment tends to direct the angular speed vector along the small semiaxis of the ellipsoid, i.e., a stable position of the rotor appears.

KEYWORDS: INFLUENCE OF FORM, STABILITY, FLUID, ROTATION.

1. Let us consider a Cartesian system of coordinates, the origin of which is located in the center of a rotating ball - a rotor. Otherwise, the system is arbitrary. Direction of angular speed vector $\vec{\Omega}$ is defined with the angle α (between the axis OZ and $\vec{\Omega}$ and β (between the axis OX and the projection of $\vec{\Omega}$ on the plane x, y). Also we will consider spherical coordinates (r, θ, φ) , related to Cartesian formulas:

$$x = r \sin \theta \cos \varphi, y = r \sin \theta \sin \varphi, z = r \cos \theta.$$

Let us introduce characteristic thickness of the gap $\delta = (R_2 - R_1)/R_1$, where R_2 - is the chamber radius. Then we will move on to dimensionless variables, choosing the rotor R_1 radius as a unit of distance and $1/\Omega$ is a unit of time. The equation for the rotor surface is $r = 1$ and the equation for the chamber inner surface is $r = r(\theta, \varphi)$. The problem of viscous incompressible fluid flow in the gap between the rotor and the chamber within the Stokes approximation outside the field of mass forces is the following [4]:

$$(1) \quad \Delta \vec{\omega} = 0, \quad \vec{\omega} = \text{rot} \vec{v},$$

$$\text{div} \vec{v} = 0.$$

Boundary conditions are:

$$\vec{v}|_{r=1} = -\sin \alpha \sin (\varphi - \beta) \vec{e}_\theta - [\sin \alpha \cos \theta \cos (\varphi - \beta) - \cos \alpha \sin \theta] \vec{e}_\varphi,$$

$$\vec{v}|_{r=r(\theta, \varphi)} = 0.$$

Now let us use the assumption of small thickness of the gap between the layer. To do that, we will introduce a new radial variable ξ , that is, when the inner surface of the chamber is a sphere of radius R_2 , we suppose that

$$\xi = (r - 1)/\delta.$$

Equation for the rotor surface now is $\xi = 0$, and equation for the inner surface of the chamber is:

$$\xi = h(\theta, \varphi), \quad h(\theta, \varphi) = (r(\theta, \varphi) - 1)/\delta.$$

As for the equations of motion (1) and the boundary conditions we will retain principal terms of their asymptotics only at $\delta \rightarrow 0$. Then equations (1) can be rewritten with a precision of terms of order δ .

$$(2) \quad \frac{\partial^2 \omega_r}{\partial \xi^2} = 0, \quad \frac{\partial^2 \omega_\theta}{\partial \xi^2} = 0, \quad \frac{\partial^2 \omega_\varphi}{\partial \xi^2} = 0,$$

$$(3) \quad \frac{1}{\delta} \frac{\partial v_r}{\partial \xi} + \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} (\sin \theta v_\theta) + \frac{1}{\sin \theta} \frac{\partial v_\varphi}{\partial \varphi} = 0,$$

$$(\Delta = \frac{1}{\delta^2} [\partial^2 / \partial \xi^2 + O(\delta)]),$$

where

$$(4) \quad \omega_r = \frac{1}{\sin \theta} \left[\frac{\partial}{\partial \theta} (\sin \theta v_\varphi) - \frac{\partial v_\theta}{\partial \varphi} \right],$$

$$(5) \quad \omega_\theta = -\frac{1}{\delta} \frac{\partial v_\varphi}{\partial \xi}, \quad \omega_\varphi = \frac{1}{\delta} \frac{\partial v_\theta}{\partial \xi}.$$

Now the boundary conditions are as follows:

$$(6) \quad v_r|_{\xi=0} = 0, \quad v_\theta|_{\xi=0} = -\sin \alpha \sin (\varphi - \beta),$$

$$v_\varphi|_{\xi=0} = -\sin \alpha \cos \theta \cos (\varphi - \beta) + \cos \alpha \sin \theta,$$

$$v_r|_{\xi=h} = v_\theta|_{\xi=h} = v_\varphi|_{\xi=h} = 0.$$

The solution of equations (2) looks like

$$\omega_\theta = A(\theta, \varphi) \xi + B(\theta, \varphi),$$

$$\omega_\varphi = C(\theta, \varphi) \xi + D(\theta, \varphi).$$

Differentiating equation (4) twice in respect to ξ and using the first equation (2) and the resultant solutions for ω_θ and ω_φ , we will obtain the following equation

$$(7) \quad \frac{\partial}{\partial \theta} (A \sin \theta) + \frac{\partial C}{\partial \varphi} = 0.$$

Now it is possible to express the components of the field of velocities v_θ and v_φ in terms of the introduced coefficients A, B, C, D , using equations (5) and the boundary equations

$$(8) \quad v_\theta = -\sin \alpha \sin (\varphi - \beta) + \delta (C \xi^2 / 2 + D \xi),$$

$$(9) \quad v_\varphi = -\sin \alpha \cos \theta \cos (\varphi - \beta) + \cos \alpha \sin \theta - \delta (A \xi^2 / 2 + B \xi).$$

Coefficients B and D can be obtained with the found above components of velocity and conditions on the chamber surface

$$(10) \quad D = \frac{1}{\delta h} \sin \alpha \sin (\varphi - \beta) - \frac{Ch}{2},$$

$$B = -\frac{1}{\delta h} [\sin \alpha \cos \theta \cos (\varphi - \beta) - \cos \alpha \sin \theta] - \frac{Ah}{2}.$$

So, angular velocity components of the fluid can be expressed in terms of coefficients A and C

$$(11) \quad v_\theta = \left(\frac{\xi}{h} - 1 \right) [\sin \alpha \sin (\varphi - \beta) + \frac{\delta}{2} Ch \xi],$$

$$(12) \quad v_\varphi = \left(\frac{\xi}{h} - 1 \right) [\sin \alpha \cos \theta \cos (\varphi - \beta) - \cos \alpha \sin \theta - \frac{\delta}{2} Ah \xi]$$

and basing on the equation of continuity and condition $v_r|_{\xi=0} = 0$

$$v_r = \frac{\delta}{\sin \theta} \left[\frac{\xi^2}{2} \left(\frac{\sin \alpha}{h^2} \frac{\partial h}{\partial \theta} \sin \theta \sin (\varphi - \beta) + \frac{\partial h}{\partial \varphi} \cos \theta \cos (\varphi - \beta) \right) - \frac{\cos \alpha}{h^2} \frac{\partial h}{\partial \varphi} \sin \theta + \right.$$

$$(13) \quad + \frac{\delta}{2} \left(\frac{\partial}{\partial \theta} (\sin \theta h C) - \frac{\partial}{\partial \varphi} (h A) \right) - \frac{\xi^3 \delta}{6} \left(\frac{\partial}{\partial \theta} (\sin \theta C) - \frac{\partial A}{\partial \varphi} \right)].$$

Using condition $v_r|_{\xi=h} = 0$, we obtain one more equation for coefficients A and C

$$(14) \quad \frac{\partial}{\partial \theta} (\sin \theta h^3 C) - \frac{\partial}{\partial \varphi} (h^3 A) = - \frac{6 \sin \alpha}{\delta} \left[\frac{\partial h}{\partial \theta} \sin \theta \sin (\varphi - \beta) + \frac{\partial h}{\partial \varphi} \cos \theta \cos (\varphi - \beta) \right] + \frac{6 \cos \alpha}{\delta} \frac{\partial h}{\partial \varphi} \sin \theta.$$

We can define coefficients A and C , solving equation (14) together with equation (7), then it is possible to find the whole field of velocities.

Let us consider, first, a special case when function h does not depend on φ , that is the chamber has axial symmetry as related to axis oz . It is possible when the chamber is spherical and axis oz is directed along the line of the rotor and the chamber centres. In this case it is sufficient to differentiate the latter equation with respect to φ and substitute expression $\frac{\partial C}{\partial \varphi}$ from (7) into it. Then we will get the equation for function A :

$$\begin{aligned} \frac{\partial}{\partial \theta} [\sin \theta h^3 \frac{\partial}{\partial \theta} (\sin \theta A)] + h^3 \frac{\partial^2 A}{\partial \varphi^2} \\ = \frac{6 \sin \alpha}{\delta} \frac{dh}{d\theta} \sin \theta \cos (\varphi - \beta), \end{aligned}$$

the solution of which should be in the following form:

$$(15) \quad A(\theta, \varphi) = \frac{6 \sin \alpha}{\delta} f(\theta) \cos (\varphi - \beta).$$

As a result we obtain an ordinary differential equation for $f(\theta)$:

$$(16) \quad [\sin \theta h^3 (\sin \theta f)']' - h^3 f = h' \sin \theta$$

(here prime means a derivative with respect to θ). It is required to find solution for this equation which is continuous at $0 \leq \theta \leq \pi$. Actually, possibility to find such a solution depends on function $h = h(\theta)$, that is, on the chamber form. For a spherical chamber $h = 1 + \lambda \cos \theta$, where λ is relative displacement of centers (the distance between the centers is $|\lambda|\delta$) and the solution for equation (16) is as follows (compare [9]):

$$(17) \quad f(\theta) = \frac{\lambda}{\lambda^2 + 4} \left(\frac{1}{h} + \frac{1}{h^2} \right), \quad h = 1 + \lambda \cos \theta.$$

In the general case when h depends on φ it is possible to "integrate" equation (7) first, representing it as condition of equality of mixed second derivatives of a new function $E = E(\theta, \varphi)$:

$$(18) \quad \sin \theta A = \frac{\partial E}{\partial \varphi}, \quad C = - \frac{\partial E}{\partial \theta}.$$

The following expression can be chosen as E

$$E(\theta, \varphi) = - \int_0^\theta C(\bar{\theta}, \varphi) d\bar{\theta},$$

for which the conditions written above can be verified directly if A and C are connected with equation (7). Let us substitute this function $E(\theta, \varphi)$ into equation (14):

$$(19) \quad \begin{aligned} \frac{\partial}{\partial \theta} (\sin \theta h^3 \frac{\partial E}{\partial \theta}) + \frac{\partial}{\partial \varphi} \left(\frac{h^3}{\sin \theta} \frac{\partial E}{\partial \varphi} \right) = \\ = \frac{6 \sin \alpha}{\delta} \left[\frac{\partial h}{\partial \theta} \sin \theta \sin (\varphi - \beta) + \frac{\partial h}{\partial \varphi} \cos \theta \cos (\varphi - \beta) \right] - \frac{6 \cos \alpha}{\delta} \frac{\partial h}{\partial \varphi} \sin \theta. \end{aligned}$$

2. Let us consider the case when the chamber shape differs little from spherical. Let us set spherical form of the chamber in the form of $h_0 = 1 + \lambda \cos \theta$ (axis oz along the line of centers). Then we will set the form, which differs little from spherical, by function

$$h = h_0 + \mu h_1, \quad |\mu| \ll 1$$

and we will look for a solution of equation (19) in the form of

$$E(\theta, \varphi) = E_0(\theta, \varphi) + \mu E_1(\theta, \varphi) + O(\mu^2).$$

Substituting h and E into equation (19) and equating the coefficients at the same degrees μ . We obtain equations for definition E_0 and E_1 , where E_0 , which satisfies the condition of norming, is already known from (15), (17) and (18):

$$(20) \quad E_0(\theta, \varphi) = \frac{6 \sin \alpha}{\delta} \frac{\lambda}{\lambda^2 + 4} \left(\frac{1}{h_0} + \frac{1}{h_0^2} \right) \sin \theta \sin (\varphi - \beta).$$

Now the right part of the equation E_1

$$(21) \quad \begin{aligned} \frac{\partial}{\partial \theta} (\sin \theta h_0^3 \frac{\partial E_1}{\partial \theta}) + \frac{\partial}{\partial \varphi} \left(\frac{h_0^3}{\sin \theta} \frac{\partial E_1}{\partial \varphi} \right) = \\ = -3 \frac{\partial}{\partial \theta} (\sin \theta h_0^2 h_1 \frac{\partial E_0}{\partial \theta}) - 3 \frac{\partial}{\partial \varphi} \left(\frac{h_0^2 h_1}{\sin \theta} \frac{\partial E_0}{\partial \varphi} \right) + \\ + \frac{6 \sin \alpha}{\delta} \left[\frac{\partial h_1}{\partial \theta} \sin \theta \sin (\varphi - \beta) + \frac{\partial h_1}{\partial \varphi} \cos \theta \cos (\varphi - \beta) \right] - \frac{6 \cos \alpha}{\delta} \frac{\partial h_1}{\partial \varphi} \sin \theta \end{aligned}$$

is completely known and because it is lineary dependent on function h_1 , which can be expanded into series with respect to spherical function, it is sufficient to consider spherical functions themselves as h_1 ([5])

$$(22) \quad h_0 = P_n^m(\cos \theta) \cos m \varphi, \quad h_1 = P_n^m(\cos \theta) \sin m \varphi,$$

$$0 \leq m \leq n, \quad n = 0, 1, 2, \dots$$

Here P_n^m are adjoint Legendre functions of first kind. Let us substitute function h_1 and its derivatives into the right part of (21) and expand this part of the equation into Fourier series with respect to variable φ . If

$$E_1(\theta, \varphi) = \frac{a_0(\theta)}{2} + \sum_{m=1}^{\infty} [a_m(\theta) \cos m \varphi + b_m(\theta) \sin m \varphi],$$

then equation (21) decomposes into a finite system of ordinary differential equation. The system is finite because for functions (20), (22) in the right part of (21) a finite number of nonzero Fourier components is retained. From the condition of norming ($E(0, \varphi) = 0$) $a_m(0) = 0$. Let us look for continuous $0 \leq \theta \leq \pi$ solution for equations in the form of Fourier series with respect to the orthogonal system of adjoint Legendre functions of appropriate weight with a fixed superscript. Let us clarify what we get at small values of n , that is in the first harmonics with respect to angle φ . If $m = 0$, then $h_1 = \text{const}$ and the equation for the chamber surface takes the following form:

$$h = 1 + \lambda \cos \theta + \mu h_1 = (1 + \mu h_1) \left[1 + \frac{\lambda}{1 + \mu h_1} \cos \theta \right],$$

If $n = 1$, then for m two values are possible: $m = 0, 1$. At $m = 0$ we get

$h_1 = \cos \theta$, $h = 1 + \lambda \cos \theta + \mu \cos \theta = 1 + (\lambda + \mu) \cos \theta$ and function h is reduced to h_0 at $\lambda \rightarrow \lambda + \mu$. Let us take an arbitrary spherical function of order $n = 1$ as h_1 :

$$h = 1 + \lambda \cos \theta + \mu_1 \sin \theta \cos \varphi + \mu_2 \sin \theta \sin \varphi + \mu_3 \cos \theta =$$

$= 1 + (\lambda + \mu_3) \cos \theta + \mu_1 \sin \theta \cos \varphi + \mu_2 \sin \theta \sin \varphi$. Performing a turn, converting axis oz into axis oz' with the directing vector

$$\vec{e} = \frac{1}{\sqrt{(\lambda + \mu_3)^2 + \mu_1^2 + \mu_2^2}} \{ \mu_1, \mu_2, \lambda + \mu_3 \},$$

we will convert function h into function $h' = 1 + \lambda' \cos \theta'$, where

$$\lambda' = \sqrt{(\lambda + \mu_3)^2 + \mu_1^2 + \mu_2^2}, \quad \cos \theta' = \frac{z}{r} =$$

$$= \frac{1}{\sqrt{(\lambda + \mu_3)^2 + \mu_1^2 + \mu_2^2}} [\mu_1 \sin \theta \cos \varphi + \mu_2 \sin \theta \sin \varphi + (\lambda + \mu_3) \cos \theta].$$

So, for disturbance of the shape of chamber inner surface, which is set by spherical function h_1 of zero or first order, we have exact

solution for equation (19), deriving from (20) with simple substitution of parametres. Geometrically zero order h_1 means small extension or compression of the chamber, and the first order means a small shift with a small turn, so that the spherical shape of the chamber does not change, though the chamber undergoes some deformation and shift. Therefore for any disturbance of the chamber surface with spherical function h_1 of the order not more than one there exists an effective spherical chamber with close meanings of relative gap δ and relative shift $\lambda\delta$, giving the same values of the main forces of the fluid response to the rotor and the main vector of moment of these forces and hence leading to the same equations of the rotor motion and the same perturbing moment. Specifically, central position of the rotor equilibrium at small disturbances of the chamber shape of kind ϵh_1 remains unstable, as in the problem for the spherical chamber.

3. Let us consider a chamber of arbitrary shape which differs little from spherical. Taking into consideration that the equation for the arbitrary chamber inner surface $r = r(\theta, \varphi) > 1$ (absence of contact of rotating ball and the chamber is supposed). As a measure of the relative gap δ value, we choose average thickness of the layer with respect to the sphere

$$\delta = \frac{1}{4\pi} \int_0^{2\pi} d\varphi \int_0^\pi [r(\theta, \varphi) - 1] \sin \theta d\theta.$$

Radial variable ξ is determined as before $\xi = (r - 1)/\delta$. Then the chamber inner surface is set with equation $\xi = (r(\theta, \varphi) - 1)/\delta = h(\theta, \varphi)$. Let us suppose that function $r(\theta, \varphi)$, together with $h(\theta, \varphi)$ are twice continuously differentiated and decompose ξ into uniformly convergent series with respect to spherical functions:

$$(23) \quad h(\theta, \varphi) = \sum_{n=0}^{\infty} \sum_{m=0}^n P_n^m(\cos \theta) (a_n^m \cos m\varphi + b_n^m \sin m\varphi),$$

where $P_n^m(\cos \theta)$ are adjoint Legendre functions [5,6]

$$P_n^m(x) = \frac{(-1)^m}{2^n n!} (1 - x^2)^{\frac{m}{2}} \frac{d^{m+n}(x^2 - 1)^n}{dx^{m+n}}, \quad 0 \leq m \leq n.$$

It is obvious from orthogonality correlation of Legendre functions and trigonometric functions the average value of function $h(\theta, \varphi)$ with respect to sphere

$$\begin{aligned} \bar{h} &= \frac{1}{4\pi} \int_0^{2\pi} d\varphi \int_0^\pi h(\theta, \varphi) P_0^0(\cos \theta) \sin \theta d\theta = \\ &= \frac{1}{4\pi} \int_0^{2\pi} d\varphi \int_0^\pi a_0^0 [P_0^0(\cos \theta)]^2 \sin \theta d\theta = a_0^0, \end{aligned}$$

but it is obvious from definition ξ and δ that $\bar{h} = 1$, that is $a_0^0 = 1$ and

$$(24) \quad h(\theta, \varphi) = 1 + \sum_{n=1}^{\infty} \sum_{m=0}^n P_n^m(\cos \theta) (a_n^m \cos m\varphi + b_n^m \sin m\varphi).$$

Now, this function should be substituted into equation (19). Let us solve it, supposing that the chamber differs little from a sphere, concentric regarding the rotor. As for geometry it means not only that the chamber shape is close to spherical, but that the rotor center is close to the chamber center, that is coefficients a_n^m and b_n^m are small. As a measure of the chamber deviation from the sphere, which is concentric regarding the rotor, we will choose function $h - 1$, equal to zero if and only if the chamber is spheric and its centre coincides with the rotor centre. We will consider as the value of this function its norm in Hilbert space L_2 of functions, which are square-integrable on the sphere

$$\|h - 1\| = \sqrt{\int_0^{2\pi} d\varphi \int_0^\pi [h(\theta, \varphi) - 1]^2 \sin \theta d\theta}.$$

Because

$$a_n^m = \frac{(2n+1)(n-m)!}{2\pi(n+m)!} \int_0^{2\pi} d\varphi \int_0^\pi [h(\theta, \varphi) - 1] P_n^m(\cos \theta) \cos m\varphi \sin \theta d\theta,$$

evaluating the latter integral with the help of Cauchy-Bunyakovskii inequality, we will obtain

$$|a_n^m| \leq \sqrt{\frac{(2n+1)(n-m)!}{2\pi(n+m)!}} \|h - 1\|,$$

that is, every coefficient a_n^m is a small value, not less than the first order regarding $h - 1$. At $a_n^m = 1$

$$\begin{aligned} &\frac{1}{\sin \theta} \frac{\partial}{\partial \theta} (\sin \theta \frac{\partial E_n^m}{\partial \theta}) + \frac{1}{\sin^2 \theta} \frac{\partial^2 E_n^m}{\partial \varphi^2} = \\ &= \frac{3 \sin \alpha}{\delta} [(m+n)(n-m+1)P_n^{m-1}(\cos \theta) \sin((m-1)\varphi + \beta) + \\ &+ P_n^{m+1}(\cos \theta) \sin((m+1)\varphi - \beta)] \\ &+ \frac{6 \cos \alpha}{\delta} m P_n^m(\cos \theta) \sin m\varphi. \end{aligned}$$

(25)

Solution of equation (25), which is continuous on the sphere, is the following:

$$\begin{aligned} E_n^m &= -\frac{3 \sin \alpha}{\delta n(n+1)} [(n+m)(n-m+1)P_n^{m-1}(\cos \theta) \sin((m-1)\varphi + \beta) + \\ &+ P_n^{m+1}(\cos \theta) \sin((m+1)\varphi - \beta)] - \\ &- \frac{6 \cos \alpha}{\delta n(n+1)} m P_n^m(\cos \theta) \sin m\varphi. \end{aligned}$$

Solution for the equation similar to (25), but at $b_n^m = 1$, is obtained from (26) with the turn for $\pi/2$.

In terms of main force \vec{F} vector definition, acting on the rotor, the chamber of arbitrary shape, which differs little from the sphere with the centre in the centre of the rotor in the first assumption regarding $h - 1$ does not differ from a spherical chamber with the same value of the gap δ , the centre of which is displaced regarding the rotor centre for the appropriate value in the appropriate direction. In particular, it is impossible to make the position of the rotor equilibrium stable, using selection of the chamber shape, if this position was unstable for the spherical chamber.

4. Let us place the beginning of Cartesian system of coordinates in the centre of spherical rotor. Let the chamber have the shape of an ellipsoid close to a sphere with semiaxis $1 + \delta_1, 1 + \delta_2, 1 + \delta_3$. Let us direct the axis of coordinates along the main axis of the chamber. We will define coordinates of the chamber centre as (x_0, y_0, z_0) . Then the equation for the chamber inner surface is the following:

$$(27) \quad \frac{(x - x_0)^2}{(1 + \delta_1)^2} + \frac{(y - y_0)^2}{(1 + \delta_2)^2} + \frac{(z - z_0)^2}{(1 + \delta_3)^2} = 1.$$

Let us consider $\delta_1, \delta_2, \delta_3, x_0, y_0, z_0$ to be so small that their squares and pair products can be neglected in comparison with themselves. Then equation (27) can be transformed into this equation $x^2 + y^2 + z^2 - 2(\delta_1 x^2 + \delta_2 y^2 + \delta_3 z^2) - 2(xx_0 + yy_0 + zz_0) = 1 + O(\delta_1^2 + \dots + z_0^2)$.

In spherical coordinates this equation has the following form:

$$\begin{aligned} &r^2(1 - 2\delta_1 \sin^2 \theta \cos^2 \varphi - 2\delta_2 \sin^2 \theta \sin^2 \varphi - 2\delta_3 \cos^2 \theta) - \\ &- 2r(x_0 \sin \theta \cos \varphi + y_0 \sin \theta \sin \varphi + z_0 \cos \theta) = 1 + O(\delta_1^2 + \dots + z_0^2). \end{aligned}$$

Then we will define its approximate solution, linear with respect to $\delta_1, \delta_2, \delta_3, x_0, y_0, z_0$:

$$r = 1 + \delta_1 r_1 + \delta_2 r_2 + \delta_3 r_3 + x_0 s_1 + y_0 s_2 + z_0 s_3 + O(\delta_1^2 + \dots + z_0^2).$$

Substituting r into equation (28) and equating coefficients at δ_1, \dots, z_0 , to zero, we will find r_1, \dots, s_3 . So, with the accuracy up to small values of the second order:

$$\begin{aligned} r(\theta, \varphi) &= 1 + \delta_1 \sin^2 \theta \cos^2 \varphi + \delta_2 \sin^2 \theta \sin^2 \varphi + \delta_3 \cos^2 \theta + \\ &+ x_0 \sin \theta \cos \varphi + y_0 \sin \theta \sin \varphi + z_0 \cos \theta. \end{aligned}$$

Let us introduce average thickness of the gap δ according the formula: $\delta = (\delta_1 + \delta_2 + \delta_3)/3$, therefore:

$$h = \frac{r-1}{\delta} = \frac{1}{2} \left[\frac{\delta_1 + \delta_2}{2} \sin^2 \theta + \delta_3 \cos^2 \theta + x_0 \sin \theta \cos \varphi + y_0 \sin \theta \sin \varphi + z_0 \cos \theta + \frac{\delta_1 - \delta_2}{2} \sin^2 \theta \cos 2\varphi \right].$$

Decomposition of function h with respect to spherical harmonics is the following:

$$h = 1 + \frac{1}{\delta} [z_0 P_1^0(\cos \theta) - x_0 P_1^1(\cos \theta) \cos \varphi - y_0 P_1^1(\cos \theta) \sin \varphi + (\delta_3 - \delta) P_2^0(\cos \theta) + \frac{\delta_1 - \delta_2}{6} P_2^2(\cos \theta) \cos 2\varphi],$$

where $P_1^0(\cos \theta) = \cos \theta$, $P_1^1(\cos \theta) = -\sin \theta$, $P_2^0(\cos \theta) = (3 \cos^2 \theta - 1)/2$, $P_2^2(\cos \theta) = 3 \sin^2 \theta$; that is the chamber centre shift regarding the rotor centre, assigned with parametres x_0, y_0, z_0 , contributes to h in the form of harmonics with $n = 1$ only, though does not influence the momentum. Let us define the deviating moment \vec{M}^p , that is the component of \vec{M} , orthogonal to angle vector velocity $\vec{\Omega}$:

$$\begin{aligned} \vec{M}^p &= \vec{M} - \frac{1}{\Omega^2} (\vec{M}, \vec{\Omega}) \vec{\Omega}, \\ M_x^p &= -\frac{8\pi\mu' \Omega R_1^3}{15\delta^2} [(\delta_1 - \delta_2) \sin^2 \alpha \sin^2 \beta + (\delta_1 - \delta_3) \cos^2 \alpha] \sin \alpha \cos \beta, \\ M_y^p &= -\frac{8\pi\mu' \Omega R_1^3}{15\delta^2} [(\delta_2 - \delta_3) \cos^2 \alpha + (\delta_2 - \delta_1) \sin^2 \alpha \cos^2 \beta] \sin \alpha \sin \beta, \\ M_z^p &= -\frac{8\pi\mu' \Omega R_1^3}{15\delta^2} [(\delta_3 - \delta_1) \sin^2 \alpha \cos^2 \beta + (\delta_3 - \delta_2) \sin^2 \alpha \sin^2 \beta] \cos \alpha. \end{aligned}$$

It is obvious that in the ellipsoidal chamber deviation moments occur in the first approximation with regard to small parameters

$$\frac{\delta_1 - \delta}{\delta}, \frac{\delta_2 - \delta}{\delta}, \frac{\delta_3 - \delta}{\delta},$$

but for spherical chamber ($\delta_1 = \delta_2 = \delta_3 = \delta$) they are absent in the first approximation with regard to λ ($h = 1 + \lambda \cos \theta$) and occur in the second approximation with regard to λ only. It is known from the exact solution [4]. Moreover, in the case of sphere, deviating moment depends on the rotor shift. If the rotor follows the circular path, the averaged with regard to period deviating moment is equal to zero. In the case of ellipsoidal chamber deviating moment in the first approximation does not depend on the rotor position.

Let us clarify the evolution of the vector angular velocity $\vec{\Omega}$ in the simplest case of isotropic spherical rotor (density from the centre only). In this case inertia tensor is spherical and is set with a scalar I , equal to inertia moment regarding any axis, crossing the centre. Vector $\vec{\Omega}$ evolution is defined with the equation of moments $I \frac{d\vec{\Omega}}{dt} = \vec{M}$. Let us substitute components of the moment into it and obtain the system of equations

$$\begin{aligned} \dot{\Omega}_x &= -\frac{8\pi\mu' R_1^3}{3\delta I} \left(1 + \frac{\delta_1 - \delta}{5\delta}\right) \Omega_x, \\ \dot{\Omega}_y &= -\frac{8\pi\mu' R_1^3}{3\delta I} \left(1 + \frac{\delta_2 - \delta}{5\delta}\right) \Omega_y, \\ \dot{\Omega}_z &= -\frac{8\pi\mu' R_1^3}{3\delta I} \left(1 + \frac{\delta_3 - \delta}{5\delta}\right) \Omega_z. \end{aligned}$$

Its solution is the following:

$$\Omega_x = \Omega_x^0 \exp \left[-\frac{8\pi\mu' R_1^3}{3\delta I} \left(1 + \frac{\delta_1 - \delta}{5\delta}\right) t \right],$$

$$\Omega_y = \Omega_y^0 \exp \left[-\frac{8\pi\mu' R_1^3}{3\delta I} \left(1 + \frac{\delta_2 - \delta}{5\delta}\right) t \right],$$

$$\Omega_z = \Omega_z^0 \exp \left[-\frac{8\pi\mu' R_1^3}{3\delta I} \left(1 + \frac{\delta_3 - \delta}{5\delta}\right) t \right].$$

We can always suppose (probably, remaining the axes) that

$$(32) \quad 0 < \delta_1 \leq \delta_2 \leq \delta_3.$$

Then $\delta_3 = 3\delta - \delta_1 - \delta_2 < 3\delta$, whence it follows that:

$$0 < \frac{\delta_1}{\delta} \leq \frac{\delta_2}{\delta} \leq \frac{\delta_3}{\delta} < 3,$$

or

$$-1 < \frac{\delta_1 - \delta}{\delta} \leq \frac{\delta_2 - \delta}{\delta} \leq \frac{\delta_3 - \delta}{\delta} < 2,$$

that is addition to vector damping decrement $\vec{\Omega}$ from (31), dependent on non-spherical type of the chamber is not more than 20% in the direction of decrease and 40% in the direction of increase in comparison with a spherical chamber with the same relative thickness of the gap.

To study evolution of the vector direction $\vec{\Omega}$ we will rewrite the system of equations (30) in spherical coordinates. Therefore it is sufficient to find solution for this equation, satisfying initial condition $\beta(0) = \beta_0, 0 < \beta_0 < \frac{\pi}{2}$.

Such a solution is set by the formula

$$(33) \quad \beta(t) = \arctg[\tg \beta_0 \exp(-(A_2 - A_1)t)],$$

$$\alpha(t) = \arccotg[\cotg \alpha_0 \sqrt{\frac{1 + tg^2 \beta_0}{\exp(2(A_2 - A_1)t) + tg^2 \beta_0}} \cdot \exp(-(A_3 - A_2)t)],$$

$$\alpha(t) \rightarrow \frac{\pi}{2} \text{ for } t \rightarrow \infty.$$

So, if at initial time the end of vector $\vec{\Omega}$ is in semisphere $-\frac{\pi}{2} < \beta_0 < \frac{\pi}{2}$ with the centre at the ellipsoidal chamber, then as time passes it is attracted to the end of this semiaxis, except the case of elongated ellipsoid of rotation when it is attracted to the equator along its meridian. In the opposite semisphere the same situation takes place: the end of vector $\vec{\Omega}$ evolves to the end of the small semiaxis (or, in case of elongated ellipsoid of rotation - to the equator. In this case every point of which is the end of the small semiaxis).

As a result we can conclude that in case of the ellipsoidal chamber direction of the small semiaxis is stable for vector $\vec{\Omega}$, which, under influence of deviating moment, is attracted to this direction from any initial position.

REFERENCES

1. Semenyuk, N.F., Bachinskaya N.K. "Mechanics of frictional contact of rough surfaces. The area of contact". Friction and deterioration, 1993; No. 6. 984-990.
2. Taylor, S. "Measurement and calculation of influence of a non-uniform roughness of a surface on factor of friction at turbulent current". Modern mechanical engineering, No. 7. 1989, P. 17-33.
3. Dementiev, O. "Stability of a rotor motion in a chamber filled with viscous gas". Computers & Mathematics with Application, T. 39. No. 7-8. 2000, P. 183-191.
4. Loitzanskii, L. "Mechanics of fluid and gas", Moscow, Nauka, 1978, P. 736.
5. Bateman H., Erdelyi A. "Higher transcendental functions", New York, Toronto, London, MC Graw-Hill Book comp. 1973. P.295.
6. Abramowitz, M., Stegun, I.A. Handbook of mathematical function, National Bureau of Standards Appl. Math. ser., Moscow, Nauka, 1964, P. 830.

A VARIATIONAL SOLUTION OF THE SCHRÖDINGER EQUATIONS IN AN INHOMOGENEOUS COULOMB FIELD

Freinkman B., Polyakov S., Tolstov I.

Keldysh Institute of Applied Mathematics of RAS – Moscow, Russia

E-mail: freink@newmail.ru, polyakov@imamod.ru, tolstov_ilya@inbox.ru

Abstract: The present work is devoted to computer modeling of the emission processes from the surface of graphene. The pivotal obstacle for emission is a model of the unperturbed emission surface. The hydrogen-like atom model is one of the useful approaches describing the states of the emission surface. In [1] this model was used considering ion screening in the Brandt model [2]. To calculate the ground state of the electron, we used the variational solution of the Schrodinger equation, based on the minimization of the potential energy of an electron in the field of a homogeneous ion. However, the field of the screened singly ionized carbon atom in the Brandt model is not homogeneous. Therefore, it was shown in [3] that it is possible to obtain a binding energy error of up to 40% when using only the external screening parameter without taking into account the inhomogeneity. In this paper, we consider the effect of the ion screening parameter in the Brandt model λ and the algorithm for determining it by minimizing the total energy of the electron interaction in s state in two parameters: the effective ion charge and the ion screening parameter. The obtained solution of the Schrödinger equation is used to calculate the ground state of a hydrogen-like carbon atom in a graphene lattice at zero temperature and is compared with the results of [2, 4].

KEYWORDS: GRAPHENE, PSEUDOPOTENTIAL OF A HYDROGEN-LIKE ATOM, COMPUTER SIMULATION FOR GRAPHENE LATTICE

1. Introduction

The present work is devoted to computer modeling of the emission processes from the surface of graphene - a perspective material for micro- and nanoelectronic devices. The pivotal obstacle for emission is a model of the unperturbed emission surface, determining the theoretical spectrum and the emission threshold current. One of the approaches to describe the states of the emission surface is the hydrogen-like atom model, which is used for light elements of the periodical table. In [1], a lattice of hydrogen-like atoms with a screened ion in the Brandt model was used to calculate the ground state of atoms of the graphene surface [2]. To calculate the ground state of an electron, the Schrödinger equation was solved by minimization of the potential energy of the electron in the ion field, which assumes homogeneity of the ion field. However, the field of the shielded singly ionized carbon atom in the Brant model is not homogeneous. Therefore, it was shown in [3] that using only the external screening parameter without taking into account the inhomogeneity, it is possible to obtain a binding energy error of up to 40%, which is unacceptable even for qualitative theory. In this paper, we consider the influence of the ion screening parameter in the Brant model λ and the algorithm for determining it by minimizing the total energy of the electron interaction in s state in terms of two parameters: the effective ion charge and the ion screening parameter. The obtained solution of the Schrodinger equation is used to calculate the ground state of a hydrogen-like carbon atom in a graphene lattice at zero temperature. The obtained solution is compared with the results of [2, 4].

2. Prerequisites and means for solving the problem

In [1], for the variational solution of the Schrödinger equation for an electron in a hydrogen-like atom, the method of determining the wave function in a given field was used [6]. This technique is based on minimizing the discrepancy of the difference between the total energy of an electron in a known ion field and the effective field

$$\tilde{U}_i(r) = \frac{q}{r} + A \quad (1)$$

for which a self-similar solution of the Schrodinger equation is known, depending on the parameters of this field. For the ground state of a hydrogen-like carbon atom, this leads to a minimization of the functional

$$J_s(q, \lambda) = \int_0^\infty \psi_{2s}^2(x) \left[U_i(r, \lambda) - \frac{q}{r} \right] x^2 dx + \frac{q^2}{2n^2} + A, \quad (2)$$

$$\psi_{2s}^2(r) = \left(1 - \frac{x}{2} \right)^2 \exp(-x), \quad x = \frac{2q}{n} r.$$

In this case, it is implicitly assumed that the field is uniform over r . However, the field of the screened ion in the Brandt-Kitagawa model is not uniform:

$$U_i(r) = \frac{1}{r} + \frac{Z-1}{r} \exp\left(-\frac{r}{\lambda_i}\right), \quad \frac{\partial}{\partial r} U_i(r) \neq k U_i(r). \quad (3)$$

According to the virial theorem, the average total energy of the finite motion of an electron in the Coulomb field is related to the average kinetic and mean potential energy by the equation

$$\bar{E} = -\bar{T}, \quad \bar{U} = 2\bar{E}. \quad (4)$$

Using the expression for the mean kinetic energy in a nonuniform field [5], we obtain the distribution of the total energy of the bound electron in an inhomogeneous field ion:

$$E(r) = \frac{r}{2} \frac{\partial}{\partial r} U_i(r) + U_i(r). \quad (5)$$

Fig. 1 shows the distribution of the total energy of an electron with and without an accounting of the inhomogeneity of this field. As can be seen from the figure, the inhomogeneity of the field appears mostly near the nucleus. Therefore, taking into account the distribution of the probability density in s and p states, the influence of the field inhomogeneity on the binding energy of the electron with the ion will be larger for the electron in the s state than in the p state.

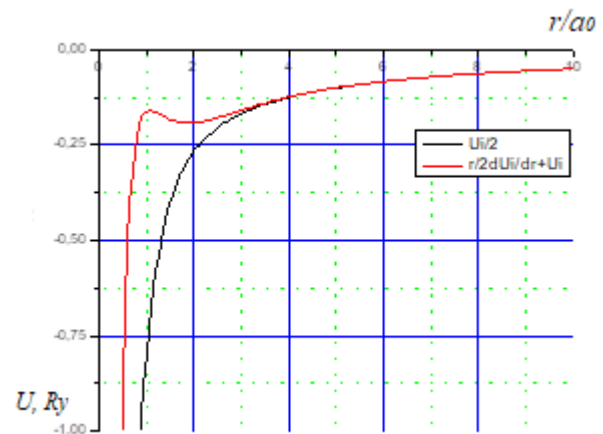


Fig. 1. Comparison of the distribution of the total energy along the radius for a finite motion of an electron

in the field of the singly charged carbon ion in the Brandt model without and taking into account the inhomogeneity of the field. Therefore, in an arbitrary field ion at solving the problem of variations on the characterization of the effective uniform field should be based on minimizing the total energy of the finite electron motion in a particular quantum state. The effective charge of the ion - q , and the external screening parameter - λ , are determined by the following system of nonlinear equations:

$$\frac{\partial}{\partial q} \left\{ \int_0^\infty \psi_{2s}^2(x) \left[U_i(r, \lambda) - \frac{q}{r} \right] x^2 dx + \frac{q^2}{2n^2} \right\} \bigg|_{q=q_m} = 0 \quad (6)$$

$$A = - \int_0^\infty \psi_{2s}^2(x) \left[U_i(r, \lambda) - \frac{q_m}{r} \right] x^2 dx - \frac{q_m^2}{2n^2}$$

The hypothesis adopted by us was realized with the help of a numerical solution of the problem (6).

3. Solution of the examined problem

Figures 2a and 2b show numerical solutions of this problem for an electron in the s state with and without consideration for the inhomogeneity of the ion field for different values of the internal screening parameter λ of the ion in the Brandt model.

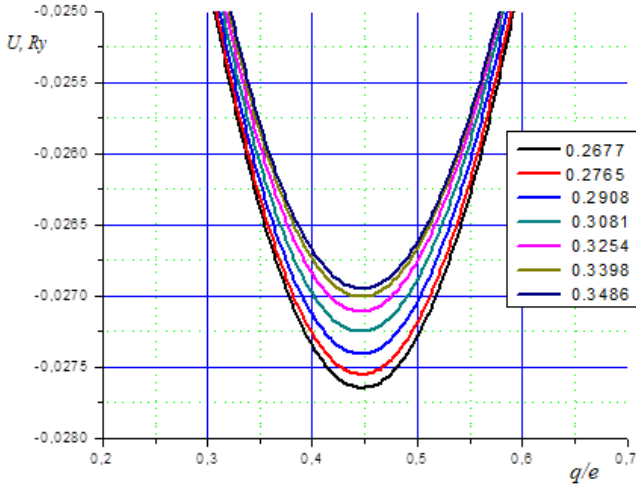


Fig. 2a Numerical solution of the variational task for an electron in the s state taking into account the inhomogeneity of the ion field.

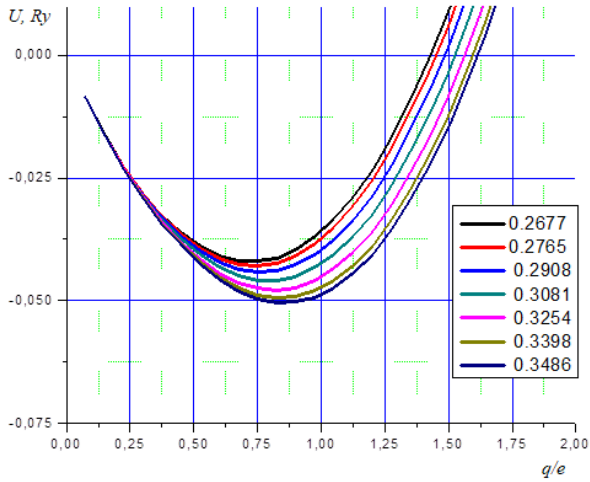


Fig. 2b Numerical solution of the variational task for an electron in the s state without taking into account the inhomogeneity of the ion field. Different lines explain different values of screening parameter λ

Comparison of these solutions shows that the lowest value of the effective charge of the ion q and the external screening potential of the atom A will be when the heterogeneity of the ion field is considered.

4. Results and discussion

For comparison, Fig. 3 shows the solution of this task using the previously used method of the variational solution of the Schrödinger equation [1, 6].

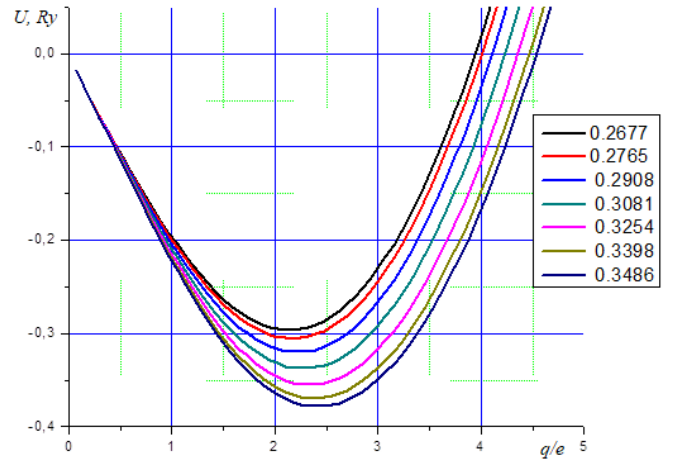


Fig. 3 The variational solution of the Schrödinger equation by the method [6].

Comparison of this solution with previous ones shows that the effective charge of the ion and the screening parameter of the atom are much higher and the binding energy exceeds the ionization potential of carbon. It can be assumed that these discrepancies go from the fact that the method of the variational solution of the Schrödinger equation [6] is more suitable for atoms with a large number of valence electrons.

An analysis of the dependence of the solution of the variational task on the parameter of the internal screening λ of an ion in the Brandt model is shown in Fig. 4.

It is interesting to note that the optimal value of λ is close to its value for a neutral atom in the Brandt model.

Figure 5 compares the obtained distribution of the field of a hydrogen-like atom along the radius with analogous distributions in [2,4].

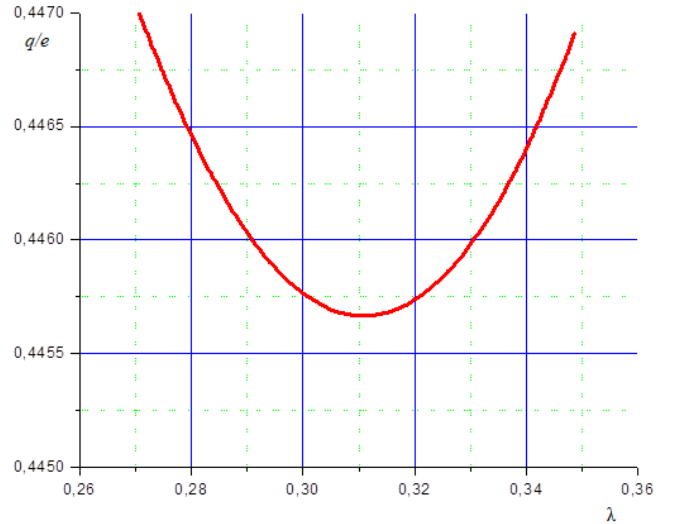


Fig. 4 Dependence of the effective charge of an ion on the parameter of the internal screening of the ion λ in the Brandt model.

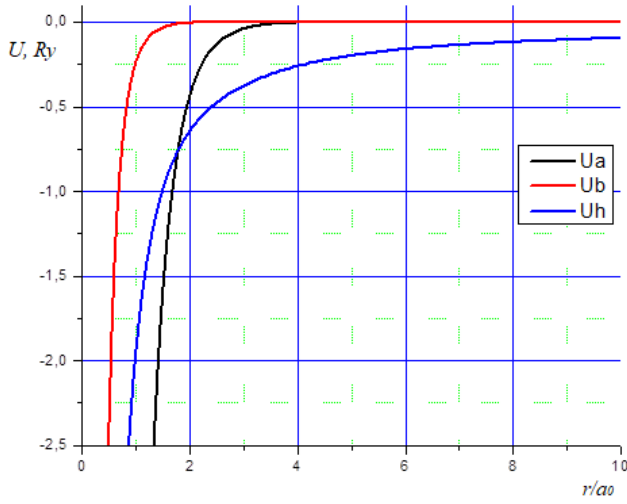


Fig. 5 Comparison of the distribution of the field along the radius of a hydrogen-like atom with similar distributions in [2,4].

It can be seen that the distribution of the field of the hydrogen-like atom is allocated between the distributions near the nucleus in other models and it decreases much more slowly in the distance. When considering these results, it should be noted that the model Brandt atom is surrounded by an electron gas, but only the effect of neighboring atoms is taken into account in the model Abrahamson. Because in the future we will be interested in the ground state of a hydrogen atom in the lattice of graphene, then the profiles of the potential distribution between two lattice sites in these models were calculated.

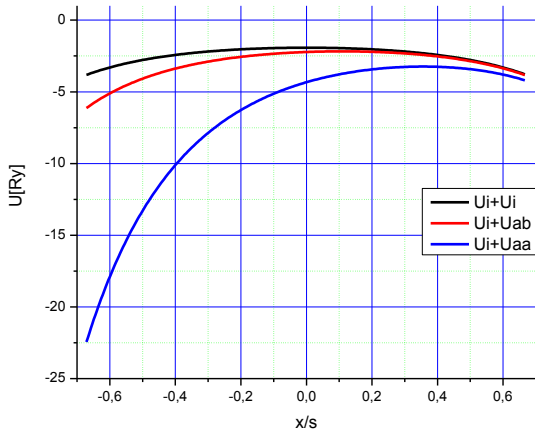


Fig. 6 Potential well for a weakly bound electron accounting the field of an ion or atom at an adjacent graphene lattice site.

Figure 6 shows that in the Abrahamson model the potential well is unnaturally deep while in the Brandt model it is close to the atom and the surrounding ion.

5. Conclusion

The problems of modeling the emission processes from the graphene surface are considered. To describe the states of the emission surface at zero temperature, we used a hydrogen-like atom model, taking into account the screening of the ion in the Brandt model. To calculate the ground state of an unbound electron located in a graphene lattice site, we used the variational solution of the Schrödinger equation. It corresponds to a minimum of the total interaction energy of an unbound electron in the s state with the corresponding lattice ion. In numerical experiments, minimization was carried out in two parameters: the effective ion charge and the ion screening parameter. However, we restricted ourselves to the first iteration of the minimization of the energy functional, since it is assumed hereinafter account of the influence of the field in the immediate environment of the atoms of the electron energy with the ion in the lattice site. Our calculations showed that in the considered model it is important to take into account the inhomogeneity of the field produced by the screened lattice ion.

The work was supported by Russian Fund for Basic Research (Project No. 17-01-00973-a)

6. References

1. B.G. Freinkman. Pseudo-potential atom model for carbon one in graphen lattice // *Matematicheskoe modelirovanie*, 2015, v. 27, No. 7, p. 122-128 (in Russian).
2. W. Brandt, and M. Kitagawa. Effective stopping-power charges of swift ions in condensed matter // *Phys. Rev.*, 1982, v. B25, n. 9, p. 5631-5637.
3. Tolstov I., Freinkman B., Polyakov S. Variation solution of the Shrodinger eqiation in an inhomogeneous central field as applied to emission problems / *Mathematical modeling and computational physics (MMCP'2017): Abstracts of International Conference (2017 July 3-7, Dubna, Russia)*. - Dubna: JINR, 2017. - p. 54.
4. A.A. Abrahamson. Born-Mayer-type interatomic potential for neutral ground-state atoms with $Z = 2$ to 105 // *Phys. Rev.*, 1969, v. 178, p. 76-79.
5. Fok V.A., Nachala kvantovoi mekhaniki (Introduction to quantum mechanics) (Moscow, 1974)
6. A.F. Nikiforov, I.G. Novikov, V.B. Uvarov. *Kvantovostatisticheskie modeli vysokotemperaturnoi plasmy*. - M.: Fizmatlit, 2000 (in Russian).

COMPUTER SYSTEM FOR PREDICTING THE STRUCTURE AND PROPERTIES OF CAST METAL PRODUCTS

Ass. Prof., Dr. Eng. Donii O., Ass. Prof., Dr. Eng. Kulinich A., Ass. Prof., Dr. Eng. Khristenko V.
National Technical University of Ukraine "Kyiv Polytechnic Institute named after Igor Sikorsky" - Kyiv, Ukraine
Email: dosha@iff.kpi.ua

Abstract: The basic principles of functioning of subsystems of the information-technological complex, which is intended for the forecast of structure and properties of the cast metal, are presented. The basis of subsystems are mathematical and simulation models of crystallization. The subsystem of thermal analysis is based on mathematical models, which were developed within the framework of the dynamic theory of metallic castings. The subsystem of modeling of crystallization is based on a combination of mathematical models of heat conduction and diffusion with cellular automata. The model makes it possible to investigate in computational experiment the effect of various cooling conditions on the process of formation of the structure during crystallization. The results of computer experiment are shown.

KEYWORDS: COMPUTER THERMAL ANALYSIS, SIMULATION MODELING, CRYSTALLIZATION, STRUCTURE OF METALS AND ALLOYS, COMPUTER MODELING, CELLULAR AUTOMATON

1. Introduction

One of the factors to solve the problem of improving the quality of castings and increasing the efficiency of foundry production is the availability of methods for the on-line monitoring of the melting process and evaluation of the melt's state immediately before casting. Although the process of forming the quality of metal products is complex and multistage, crystallization plays a special role in it since it is at this stage the "primary" structure of solid metal is formed, which significantly influences the formation of the properties, which the final product must have. That is, it is necessary to be able to on-line check of the melt's state and control the crystallization process during the technological process of smelting. As one of the variants of realization of this task it is possible to use information-technological complex consisting of subsystems of computer thermal analysis [1] and of simulation modeling of metals' and alloys' crystallization [2]. Thermal analysis provides information on thermal effects during hardening of sample, which reflect the processes of formation of the cast metal's structure. This makes it possible to analyze the kinetics of the crystallization process using data of the cooling curve and to predict the service properties of metals and alloys in the solid state. Simulation modeling makes it possible in the mode of computer experiment to optimize known and develop new technological modes for creating foundry products, which significantly reduces time of development and material costs, gives an opportunity to obtain important scientific information about the process of crystallization of metals and alloys under both equilibrium and nonequilibrium conditions (even those ones, which is difficult or impossible to obtain under in laboratory conditions). Simulation modeling makes it possible in the mode of computer experiment to optimize known and develop new technological modes for creating foundry products, which significantly reduces time of development and material costs, gives an opportunity to obtain important scientific information about the process of crystallization of metals and alloys under both equilibrium and nonequilibrium conditions.

2. Preconditions and means for resolving the problem

The thermal analysis subsystem should record the change of the temperature of metallic material, which is in the process of crystallization (the cooling curve). Using the data of the cooling curve with the help of mathematical modeling it is possible to single out information on the dynamics of the crystallization process, which helps to predict the properties of metals or alloys in the solid state. Mathematical models in this case should be simple, since the calculations based on them must be performed in the computer for the minimum possible time [3]. Similar models are proposed in the framework of the dynamic theory of hardening of metallic castings [4]. Based on these models, a technique for computer thermal analysis is proposed, which simulates differential thermal analysis using the hardware of simple thermal analysis. The application of this technique significantly increases the sensitivity of the method, as well as the reliability of the forecast of the structure and properties of the metal in the solid state.

In the process of developing of a subsystem of simulation modeling of the formation of the structure of metallic materials during their crystallization, problems arise due to the complexity of the mathematical formulation of the problem. At the same time, studies of cellular automata have shown the principal possibility of their use for modeling of similar processes [5]. However, the rules for cellular automata' operation of are usually given without connection with real physical processes. In this paper we propose a simulation model of crystallization, which is based on combination of mathematical models of heat conduction and diffusion with cellular automata. At that, the external cooling conditions and composition of the liquid metal or melt at a certain point determine the possible state (liquid or solid) of the corresponding cell of the cellular automaton.

3. Structure and capabilities of the models

3.1. The subsystem of computer thermal analysis

Mathematical difficulties which arise while creating mathematical models of crystallization can be significantly reduced if we consider the case of solidification of a body whose temperature gradient across its volume can be neglected. In the method of thermal analysis, which is used in this work, portions of metal or metal alloys which have small dimensions and cylindrical shape with a diameter of 20 ... 40 mm are examined. Since metallic materials are good conductors of heat, in this case it is possible to proceed to an analysis of the heat balance equation for the whole sample as a whole [4]. Then the heat balance equation is represented as:

$$\frac{dQ}{dt} = \frac{dQ_c}{dt} + \frac{dQ_L}{dt}, \quad (1)$$

where $\frac{dQ}{dt}$ - the change in the amount of heat in the sample due to its release into the environment; $\frac{dQ_L}{dt}$ - change in the amount of heat, which is released as a result of formation of a solid phase; $\frac{dQ_c}{dt}$ - change in the amount of heat in the sample due to changes of the material's temperature.

Substituting the known relations in (1):

$$dQ_c = -cmd(t), \quad (2)$$

$$\frac{dQ}{dt} = fS[T(t) - T_{cp}] + \sigma \varepsilon S[T^4(t) - T_{cp}^4], \quad (3)$$

we obtain the differential equation in the form:

$$\frac{dT(t)}{dt} + k_1[T(t) - T_{cp}] + k_2[T^4(t) - T_{cp}^4] = Z(t), \quad (4)$$

$$\text{where } Z(t) = \begin{cases} 0, & \text{при } t \leq t_{kp}, t \geq t_{ms} \\ \frac{L}{c} \frac{m(t)}{m_0}, & \text{при } t_{kp} \leq t \leq t_{ms} \end{cases} \quad (5)$$

$$V(t) = \frac{m(t)}{m_0}, \quad (6)$$

$$k_1 = \frac{fS}{cm_0}, \quad (7)$$

$$k_2 = \frac{\sigma \varepsilon S}{cm_0}, \quad (8)$$

where $T(t)$, T_{cp} - are the temperatures of the metal (of cooling curve) and of the environment; t - is the total time of the process; t_{kp} - is the time of onset of crystallization; t_{me} - time of the end of crystallization; c , L , m_0 - are, respectively, specific heat, specific latent heat of crystallization and portion's mass of the material that is being investigated; $m(t)$ - is the mass of the solidified part of the sample to the instant t ; ε - is the degree of blackness of the cooling surface; $\sigma = 5,56 \cdot 10^{-8} \text{ Дж/(м}^2\text{K}^4\text{)}$ - is the Stefan-Boltzmann constant.

Equation (4) simulates the technique of differential thermal analysis, which has a high sensitivity of the beginning and end of phase transformations. During the experiment, the cooling curve $T(t)$ is fixed in the computer memory. Then, its fourth degree and derivative are calculated. Knowing the values of the coefficients k_1 , k_2 , we can calculate the right-hand side of $Z(t)$ equation (3). Before the beginning of crystallization and after its termination, the value of Z is zero. During crystallization, heat is released and the balance is disrupted. These changes at the points t_{kp} and t_{me} are easily recorded in the experiment.

To determine the values of the coefficients k_1 and k_2 , the initial and final sections of the cooling curve are used, where there is no release of crystallization heat and the right-hand side of equation (4) is zero. These sections of the cooling curve are linearized by dividing of both parts (4) by $[T(t)-T_{cp}]$ and introducing of new variables:

$$x = \frac{T^4(t) - T_{cp}^4}{T(t) - T_{cp}}, \quad (9)$$

$$y = -\frac{T'(t)}{T(t) - T_{cp}}. \quad (10)$$

Then equation (4) in these areas becomes the equation of a straight line:

$$y = k_1 + k_2 x. \quad (11)$$

The values of the coefficients are determined by the method of least squares [6].

Thus, this technique of computer thermal analysis uses the hardware part of simple thermal analysis and simulates differential thermal analysis. Its use in the subsystem of the computer thermal analysis makes it possible to analyze the kinetics of the crystallization process using the data of the cooling curve and to predict the service properties of metals in the solid state.

3.2. The subsystem of simulation modeling

The two-dimensional heat equation is used in the development of the subsystem of simulation modeling of formation of the structure of metallic materials during their crystallization. It assumes that cooling of the body comes from all four sides of the plane, and thermophysical characteristics of the melt (specific heat, density and thermal conductivity) do not depend on temperature (ie, c , ρ , λ = const):

$$\frac{\partial T(x, y, t)}{\partial t} = a \left(\frac{\partial^2 T(x, y, t)}{\partial x^2} + \frac{\partial^2 T(x, y, t)}{\partial y^2} \right) \pm \frac{L}{c} \frac{\partial \varepsilon(x, y, t)}{\partial t}, \quad (12)$$

$$T(x, y, 0) = T_0 = \text{const}, \quad (13)$$

$$\frac{\partial T(0, y, t)}{\partial x} = -\alpha_1 [T(0, y, t) - T_0], \quad (14)$$

$$\frac{\partial T(x_0, y, t)}{\partial x} = -\alpha_2 [T(x_0, y, t) - T_0], \quad (15)$$

$$\frac{\partial T(x, 0, t)}{\partial y} = -\alpha_3 [T(x, 0, t) - T_0], \quad (16)$$

$$\frac{\partial T(x, y_0, t)}{\partial y} = -\alpha_4 [T(x, y_0, t) - T_0], \quad (17)$$

where $a = \frac{\lambda}{c\rho}$ - is the coefficient of thermal diffusivity; $\varepsilon(x, y, t)$ - is

the fraction of the solid phase in some elementary volume; T_0 - is the initial temperature; L - is the specific latent heat of crystallization; α_1 , α_2 , α_3 , α_4 - heat transfer coefficients from four sides; x_0 , y_0 - are the dimensions of the system. The last term in the heat equation.

The last term in the heat equation (12) takes into account the release (or absorption) of heat during the phase transformation. Since a cellular automaton is used in the simulation model of crystallization, it is convenient to solve this problem numerically. This solution is realized by the splitting method using an implicit numerical scheme.

Taking into account the need to enter the system time in the model, the order of calculations is organized as follows. First and foremost, the heat conduction problem is solved numerically, in which the time step determines the system time of the entire model. At that, one step of the system time is the total time of observation of the diffusion problem. Therefore, the calculation of this task is carried out anew at each step of the system time. At that, the initial condition changes each time: the previous solution becomes a new initial condition. The boundary conditions assume the absence of exchange of matters at the boundaries of the system:

$$\frac{\partial K(x, y, t)}{\partial t} = D \left(\frac{\partial^2 K(x, y, t)}{\partial x^2} + \frac{\partial^2 K(x, y, t)}{\partial y^2} \right), \quad (18)$$

$$K(x, y, 0) = K_0(x, y), \quad (19)$$

$$\frac{\partial K(0, y, t)}{\partial x} = 0, \quad (20)$$

$$\frac{\partial K(x_0, y, t)}{\partial x} = 0, \quad (21)$$

$$\frac{\partial K(x, 0, t)}{\partial y} = 0, \quad (22)$$

$$\frac{\partial K(x, y_0, t)}{\partial y} = 0, \quad (23)$$

where $K(x, y, t)$ - is concentration of the second component in the melt; $K_0(x, y)$ - is an initial concentration, which is updated at each step of the system time; D - is the diffusion coefficient of the second component of the alloy in the melt. The solution of this problem is completely analogous to the solution of the thermal problem.

The temperature and concentration of the second component at each point of melt determine the size of the local supercooling, which is calculated as the difference of the liquidus temperature for this point and its temperature at a given time. Subcooling in the liquid state is the main driving force of crystallization. In its presence, there are conditions for formation of a crystal or growth of an already existing embryo. To calculate the presence and magnitude of supercooling, the linearization of the liquidus lines and of the solidus of the double-alloy state diagram with the eutectic is used.

Thus, in the subsystem of simulation modeling, mathematical models of heat conduction and diffusion are organized in such a way that it is possible to simulate various conditions of cooling which take place during crystallization in the technological processes of foundry. The simulation model itself has the possibility of changing the cooling conditions (reducing or increasing the intensity of the heat removal) directly in the course of the computer experiment.

4. Results and discussion

The basis for predicting of the mechanical properties of the alloys of the AL-Si and Al-Si-Mg systems in the cast state is that the temporary tear resistance σ_e and relative elongation δ are mainly determined by the chemical composition of alloys' main alloying components (Si, Mg) and impurities, which significantly affect the mechanical properties of these alloys (for example, Fe). In addition,

inherited metal's properties and technology's factors that are difficult to control (for example, overheating or modifying) can also change the average level of mechanical properties. The total effect of these factors is manifested in the thermogram in the form of a change in the values of recalcence during the formation of the α -phase and the eutectic. As shown in [3], these parameters, in particular eutectic recalcence, correlate with the level of mechanical properties. Therefore, it is advisable to introduce its value as an additional parameter into equation of forecasting of mechanical properties. To create appropriate regression equations, experiments were performed in which content of the silicon for Al-Si and Al-Si-Mg alloys varied from 6.1% to 11.3% and of magnesium - in the range from 0.15% to 0.40%. Equations are obtained for the alloy of the Al-Si system, which bind the temporary tear resistance and relative elongation with the calculated content of silicon, iron, and with the value of eutectic recalcence. They have the following form:

$$\sigma_s = 14,37145 + 0,31076 \cdot Si - 13,1125 \cdot Fe - 9,3304 \cdot Fe^2 + 0,2431 \cdot \Delta T_{sp}, \quad (24)$$

$$\delta = 4,4195 - 0,16663 \cdot Si - 1,7786 \cdot Fe - 1,06714 \cdot Fe^2 + 0,0283 \cdot \Delta T_{sp}, \quad (25)$$

а для сплава системы Al-Si-Mg:

$$\sigma_s = 51,01231 + 4,13723 \cdot Si + 8,05964 \cdot Mg - 4,6346 \cdot Fe + 0,01782 \cdot \Delta T_{sp}, \quad (26)$$

$$\delta = -2,1175 - 1,3332 \cdot Si - 0,2143 \cdot Mg - 5,60762 \cdot Fe + 0,0171 \cdot \Delta T_{sp}, \quad (27)$$

where ΔT_{sp} - the value of the recalcence of temperature, determined by the cooling curve in the region of formation of the eutectic. Composition of alloys is also determined by the cooling curve by means of the method, which is described in [3]. Fisher criterion was used to prove the adequacy of the models, and the error in forecasting of the properties was 5 ... 7%. When testing the subsystem of computer thermal analysis under industrial conditions, the equations described above showed the convergence of results and the accuracy of the prediction of properties at the level of standard methods of mechanical testing.

The use of cellular automata greatly simplifies mathematical calculations while usage of the model and gives the principal possibility of modeling of the emerging polycrystalline structure of metallic materials. The influence of the external cooling conditions of the system under study on the functioning of the cellular automaton makes it possible to establish technological methods for

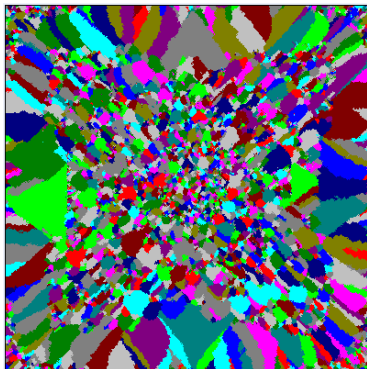


Fig. 1. Macrostructure of pure aluminum, which is simulated at speed of cooling 5 deg/s

obtaining various structures of solid metal. Model structures were obtained as a result of the computer experiment of homogeneous crystallization of aluminum with usage of various cooling speed during the process.

With equal speed of heat removal from all sides and a cooling rate of 5 deg/s, a diverse granular structure of pure aluminum is formed (Fig. 1). There are small grains in the center, and on the sides grains'

sizes are much larger. It can be also observed that a solid metal's structure is interrelated with the shape of the temperature's field inside the system. Figure 2 shows the structure of aluminum, which

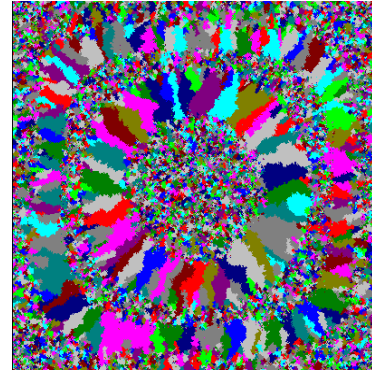


Fig. 2. Macrostructure of pure aluminum, simulated by an alternate change in the coefficients of heat transfer in the process of crystallization

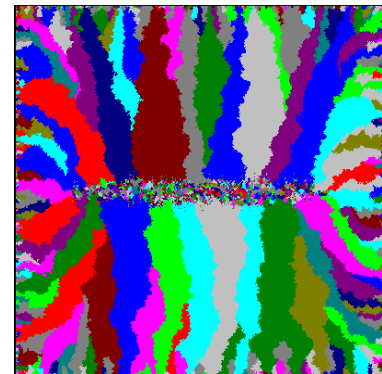


Fig. 3. Macrostructure of pure aluminum, simulated with asymmetric variation of heat transfer coefficients

was obtained by repeatedly changing intensity of heat removal. At first, fine grains were formed at the boundaries of the system in conditions of high cooling rate. Then the intensity of heat removal from all sides was reduced symmetrically and the size of the growing grains increased. After some time, the cooling rate was increased, then reduced and then increased again. As a result, the final structure consists of several alternating areas with different shapes and different average sizes of grains (Fig. 2). Complex aluminum structure has been obtained in which grains of different shapes, sizes, and orientations are observed (Fig. 3) as a result of varying of the cooling rate by setting asymmetrical coefficients of heat removal from different sides and by varying of their magnitude. It can be assumed that in this case the properties of the real metal will have anisotropy. Such regimes of crystallization are not easy to be realized in practice, but these experiments demonstrate the capabilities of this modeling subsystem.

5. Conclusion

1. Usage of mathematical models of solidification of a small portion of metal allows to increase the sensitivity of thermal analysis to the level of differential thermal analysis due to calculating of the sample's heat balance before, after and during crystallization.
2. The resulting regression equations allow to predict the level of breaking point and relative elongation of alloys of Al-Si and Al-Si-Mg systems using data from the cooling curve of the thermal analysis with an error of 5 ... 7%.
3. Simulation model of crystallization makes it possible to visualize the process of formation of the macro structure of metals under different conditions of crystallization, that makes it possible, within the framework of a computer experiment, to develop techniques for management of this process to obtain the necessary properties of cast products.

4. Combination of the subsystem of simulation modeling of crystallization and subsystem of computer thermal analysis as part of the informational and technological complex significantly expands possibilities of controlling the cast products' quality and allows to obtain information by engineers-technologists both in applied and in theoretical aspects.

6. Literature

1. Donii O. System of control and management of the quality of foundry melts on the basis of computer thermal analysis. Acta Universitatis Pontica Euxinus. Dnipropetrovsk, Varna, Volume III, 2011, P. 84 - 87. (Ukrainian)

2. Donii O. Modeling of metal crystallization by means of cellular automation. Materials science. Non-equilibrium phase transformations, 4, 2017, P. 150 - 153.

3. Bialik O., Donii O., Golub L. Forecast of properties of metals and alloys by the method of computer thermal analysis. Preprint, Kyiv, "Politechnika". (Ukrainian)

4. Bialik O., Mentkovskyi Yu. Questions of the dynamic theory of hardening of metal castings. Kiev, Higher school, 1983. (Russian)

5. Hayes B. The cellular automaton creates a model of the world and the world around itself. In the world of science. N5, 1984, p. 97 - 104. (Russian)

6. Lvovskyi E. Statistical methods of constructing of empirical formulas. M, Higher school, 1988. (Russian)

MODELING OF ELECTRONIC STATES OF A SINGLE DONOR IN MIS-STRUCTURE USING THE FINITE DIFFERENCE METHOD

M.Sc. Levchuk E.A.¹, Prof. Lemeshevskii S.V. PhD.², Prof. Makarenko L.F. PhD¹
 Faculty of Applied Mathematics and Computer Science – Belarusian State University, Belarus¹
 Institute of Mathematics, National Academy of Sciences of Belarus, Belarus²

liauchuk.alena@gmail.com

Abstract: Numerical modeling of electronic state evolution due to external electric field in the structure metal-insulator-semiconductor with solitary donor center is carried out. Considering a nanometer disc-shaped gate as a source of the electric field, the problem for the Laplace equation in infinite multilayered medium is solved to determine the gate potential. The energy spectrum of a bound electron is calculated from the problem for the stationary Schrödinger equation. Finite difference schemes are constructed to solve both the problems. Difference scheme for the Schrödinger equation takes into account cusp condition for the wave function at the donor location. To solve the problem for the Laplace equation, asymptotic boundary conditions for approximating the potential at large distances from the gate are proposed. On the basis of calculation results, a controlling parameter is suggested, which allows to determine the localization of electron wave function regardless of insulator thickness and permittivity.

Keywords: FINITE DIFFERENCE METHOD, MIS-STRUCTURE, NANOGATE, SCHRÖDINGER EQUATION, ENERGY LEVEL, NUMERICAL MODELING

1. Introduction

The sensitivity of the electrical, optical, and magnetic properties of semiconductors to doping impurities is widely used to create various semiconductor devices. Reducing the physical size of the devices led to the need for taking into account quantum effects in the design and optimization of modern semiconductor structures [1]. Moreover, the achievements in technology of manufacturing nanoscale structures have made it possible to construct devices with only one impurity atom in the working region [2]. Designing of such devices requires additional fundamental studies devoted to the quantum mechanical modeling of their physical properties.

There are several proposals of using single impurity atoms to physically realize qubits, which can be based on the nuclear spin of a phosphorus impurity in silicon [3], the electron spin of a bound electron [4] or its charge [5]. These proposals stimulated the appearance of a number of papers devoted to modeling the electronic states of a donor near semiconductor surface (see [6-13]). The main attention was paid to the control of the donor electron density with an external electric field when the critical characteristic of the system under study was the critical field [9-10] or the critical potential at the gate [6], which corresponds to the relocation of the wave function of the donor electron into the gate area.

Taking into account all the features of the system (the finite size of the gate, the presence of a dielectric layer of nonzero thickness, the difference between the dielectric permittivities of semiconductor and insulator layers) requires solving the problem for the Laplace equation in a multilayered medium. As a result, it becomes necessary to use numerical methods, in particular, the finite difference method (FDM), which had shown its effectiveness in solving similar problems [14]. Attempts of applying FDM for modeling quantum mechanical effects in nanoscale structures with single donors were made in Refs. [15-16]. However, in these papers, the structures with confining potential of a simple form were considered, that does not require solving the problem for the Laplace equation.

The aim of this paper is to develop an algorithm for the numerical simulation of electronic states in a metal-insulator-semiconductor (MIS) structure with a single donor in the gate region based on sequential solution of the problems for the Laplace equation and the stationary Schrödinger equation. To calculate the gate potential, wave functions, and energy levels, FDM is used. The results of FDM calculations are compared with the results obtained with the finite element method (FEM).

2. Formulation of the problem

We consider a singly charged donor located in semiconductor at a distance z_0 from the semiconductor-insulator interface. An external electric field is created by an infinitely thin disc-shaped gate of diameter d and potential Φ_0 . The insulator layer, separating the gate from the semiconductor, is located in the area $-t_{ox} < z < 0$. The donor and the center of the gate are positioned on the Oz axis.

According to the effective mass approach, the energy E and the electron wave function Ψ for states with zero projection of the orbital angular momentum on the Oz axis are described with the problem for the stationary Schrödinger equation (in cylindrical coordinates):

$$(\hat{T} + \hat{V})\Psi = E\Psi, \quad \rho > 0, \quad z > 0, \quad (1)$$

where \hat{T} is a kinetic energy operator, \hat{V} is a potential energy operator.

In Eq. (1), we use effective Bohr radius as a unit of length

$$a^* = \frac{4\pi\epsilon_0\epsilon_2\hbar^2}{m^*e^2} \quad (2)$$

and effective Rydberg as a unit of energy

$$Ry^* = \frac{\hbar^2}{2m^*(a^*)^2} = \frac{1}{4\pi\epsilon_0\epsilon_1} \frac{e^2}{2a^*}, \quad (3)$$

where m^* is electron effective mass in semiconductor, ϵ_1 is dielectric permittivity of semiconductor. The potential is measured in units of Ry^*/e , respectively.

In Eq. (1), the kinetic energy operator \hat{T} is defined as

$$\hat{T} = -\frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \frac{\partial}{\partial \rho} \right) - \frac{\partial^2}{\partial z^2}, \quad (4)$$

The potential energy operator is the sum

$$\hat{V} = \hat{V}_D + \hat{V}_{D'} + \hat{V}_{sim} + \hat{V}_G,$$

where \hat{V}_D is the interaction between the electron and the donor

$$\hat{V}_D = -\frac{2}{\sqrt{(z-z_0)^2 + \rho^2}}, \quad (5)$$

$\hat{V}_{D'}$, \hat{V}_{sim} are the interactions of the electron with the donor image and electron image respectively [12]:

$$\hat{V}_D = -\frac{2Q^*}{\sqrt{(z+z_0)^2 + \rho^2}} + \frac{8\varepsilon_1\varepsilon_2}{(\varepsilon_1 + \varepsilon_2)^2} \frac{\varepsilon_3 - \varepsilon_2}{\varepsilon_3 + \varepsilon_2} \sum_{k=1}^{\infty} \frac{(P^*)^{k-1}}{\sqrt{(z+z_0+2kt_{ox})^2 + \rho^2}}, \quad (6)$$

$$\hat{V}_{sim} = \frac{Q^*}{2z} - \frac{2\varepsilon_1\varepsilon_2}{(\varepsilon_1 + \varepsilon_2)^2} \frac{\varepsilon_3 - \varepsilon_2}{\varepsilon_3 + \varepsilon_2} \sum_{k=1}^{\infty} \frac{(P^*)^{k-1}}{z + kt_{ox}}, \quad (7)$$

where dielectric permittivity of the medium $\varepsilon(\rho, z)$ is

$$\varepsilon(\rho, z) = \begin{cases} \varepsilon_1, & (\rho, z) \in (0, +\infty) \times (0, +\infty) = D_1; \\ \varepsilon_2, & (\rho, z) \in (0, +\infty) \times (-t_{ox}, 0) = D_2; \\ \varepsilon_3, & (\rho, z) \in (0, +\infty) \times (-\infty, -t_{ox}) = D_3; \end{cases} \quad (8)$$

$$Q^* = \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1 + \varepsilon_2}; \quad P^* = Q^* \frac{\varepsilon_3 - \varepsilon_2}{\varepsilon_3 + \varepsilon_2}. \quad (9)$$

Operator $\hat{V}_G = -u(\rho, z)$ describes the electron potential energy in external field, where gate potential $u(\rho, z)$ is solution of the following problem:

$$\frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \varepsilon \frac{\partial u}{\partial \rho} \right) + \frac{\partial}{\partial z} \left(\varepsilon \frac{\partial u}{\partial z} \right) = 0, \quad \rho > 0, \quad -\infty < z < +\infty, \quad (10)$$

$$\frac{\partial u}{\partial \rho} \Big|_{\rho=0} = 0, \quad (11)$$

$$u \rightarrow 0 \quad \text{as} \quad z \rightarrow \pm\infty, \quad \rho > 0, \quad (12)$$

$$u \rightarrow 0 \quad \text{as} \quad \rho \rightarrow +\infty, \quad -\infty < z < +\infty. \quad (13)$$

At the semiconductor boundary, an infinitely high potential barrier is assumed, which allows to set the boundary condition for the wave function in the form

$$\Psi|_{z=0} = 0. \quad (14)$$

For bound states:

$$\Psi \rightarrow 0 \quad \text{as} \quad \rho \rightarrow \infty, \quad z \rightarrow \infty. \quad (15)$$

At points on the axis $\rho=0$ outside the donor location, the wave function must be continuously differentiable, so

$$\frac{\partial \Psi}{\partial \rho} \Big|_{\rho=0} = 0, \quad z \neq z_0. \quad (16)$$

At the donor location, condition (16) is not satisfied, since the potential becomes infinite at this point. Using a method described in [11], one can obtain the following condition

$$\left(\frac{\partial \Psi}{\partial \rho} + \Psi \right) \Big|_{\rho=0} = 0, \quad z = z_0. \quad (17)$$

3. Calculation of the gate potential

For numerical calculation of the gate potential, the unbounded domain $\rho > 0$, $-\infty < z < +\infty$, on which the problem is posed, was replaced by a bounded domain $0 < \rho < L_\rho$, $-L_{z3} < z < L_{z1}$. The conditions on the boundary of this domain were set using the assumption that at large distances the gate potential can be approximately assumed to be equal to the potential of the point charge q located in the center of the gate. This potential has been found using Fourier-Bessel transform [18]:

$$v_1(\rho, z) = \frac{q}{\pi\varepsilon_0} \frac{\varepsilon_2}{(\varepsilon_2 + \varepsilon_3)(\varepsilon_2 + \varepsilon_1)} \sum_{k=0}^{\infty} \frac{(P^*)^k}{\sqrt{(z+t_{ox}(2k+1))^2 + \rho^2}},$$

$$v_2(\rho, z) = \frac{q}{2\pi\varepsilon_0(\varepsilon_2 + \varepsilon_3)} \left(\sum_{k=0}^{\infty} \frac{(P^*)^k}{\sqrt{(z+t_{ox}(2k+1))^2 + \rho^2}} - \sum_{k=0}^{\infty} \frac{Q^*(P^*)^k}{\sqrt{(z-t_{ox}(2k+1))^2 + \rho^2}} \right),$$

$$v_3(\rho, z) = \frac{q}{4\pi\varepsilon_0\varepsilon_3} \left(\frac{2\varepsilon_3}{\varepsilon_2 + \varepsilon_3} \frac{1}{\sqrt{(z+t_{ox})^2 + \rho^2}} - \frac{4\varepsilon_2\varepsilon_3}{(\varepsilon_2 + \varepsilon_3)^2} \sum_{k=1}^{\infty} \frac{Q^*(P^*)^{k-1}}{\sqrt{(z-t_{ox}(2k-1))^2 + \rho^2}} \right).$$

After differentiation we get the boundary conditions:

$$\left(\frac{\partial u_1}{\partial z} - \frac{1}{v_1} \frac{\partial v_1}{\partial z} u_1 \right) \Big|_{z=L_{z1}} = 0, \quad \rho > 0, \quad (18)$$

$$\left(\frac{\partial u_3}{\partial z} - \frac{1}{v_3} \frac{\partial v_3}{\partial z} u_3 \right) \Big|_{z=-L_{z3}} = 0, \quad \rho > 0, \quad (19)$$

$$\left(\frac{\partial u_k}{\partial \rho} - \frac{1}{v_k} \frac{\partial v_k}{\partial \rho} u_k \right) \Big|_{\rho=L_\rho} = 0, \quad -\infty < z < +\infty, \quad k = 1, 2, 3. \quad (20)$$

Difference scheme for the problem (10) – (11), (18) – (20) was constructed on the grid Ω_L , which is defined as follows:

$$\omega_\rho = \left\{ \rho_i = ih_\rho : h_\rho = \frac{L_\rho}{N_\rho}, \quad i = \overline{0, N_\rho} \right\},$$

where h_ρ is a grid step, N_ρ is a number of nodes on ρ direction,

$$\omega_z^{(1)} = \left\{ z_j^{(1)} = jh_z^{(1)} : h_z^{(1)} = \frac{L_{z1}}{N_z^{(1)}}, \quad j = \overline{0, N_z^{(1)}} \right\},$$

$$\omega_z^{(2)} = \left\{ z_j^{(2)} = jh_z^{(2)} - t_{ox} : h_z^{(2)} = \frac{t_{ox}}{N_z^{(2)}}, \quad j = \overline{0, N_z^{(2)} - 1} \right\},$$

$$\omega_z^{(3)} = \left\{ z_j^{(3)} = -L_{z3} + jh_z^{(3)} : h_z^{(3)} = \frac{L_{z3} - t_{ox}}{N_z^{(3)}}, \quad j = \overline{0, N_z^{(3)} - 1} \right\},$$

$h_z^{(k)}$ is a grid step, $N_z^{(k)}$ is a number of nodes on z direction. Then

$$\Omega_L = \omega_\rho \times (\omega_z^{(1)} \cup \omega_z^{(2)} \cup \omega_z^{(3)}).$$

The values of the potential in grid nodes are denoted as

$$y_{ij}^{(k)} = u_k(\rho_i, z_j^{(k)}), \quad k = 1, 2, 3.$$

As a result, we get the following difference scheme [14]:

$$\frac{1}{\rho_i} \left(\rho_{i+1/2} y_{\rho}^{(k)} \right)_{\bar{\rho}} + y_{zz}^{(k)} = 0, \quad k = 1, 2, 3. \quad (21)$$

$$\varepsilon_k \frac{y_{i,1}^{(k)} - y_{i,0}^{(k)}}{h_z^{(k)}} - \varepsilon_{k+1} \frac{y_{i,0}^{(k)} - y_{i, N_z^{(k+1)}-1}^{(k)}}{h_z^{(k+1)}} + \frac{\varepsilon_k h_z^{(k)} + \varepsilon_{k+1} h_z^{(k+1)}}{2} \times \frac{\rho_{i+1/2} (y_{i+1,0}^{(k)} - y_{i,0}^{(k)}) - \rho_{i-1/2} (y_{i,0}^{(k)} - y_{i-1,0}^{(k)})}{\rho_i h_\rho^2} = 0, \quad (22)$$

$$k = 1: \quad i = \overline{1, N_\rho - 1}; \quad k = 2: \quad i = \overline{N_d + 1, N_\rho - 1}.$$

$$\frac{y_{1,j}^{(k)} - y_{0,j}^{(k)}}{h_\rho} + \frac{h_\rho}{4} \frac{y_{0,j+1}^{(k)} - 2y_{0,j}^{(k)} + y_{0,j-1}^{(k)}}{(h_z^{(k)})^2} = 0, \quad (23)$$

$$j = \overline{1, N_z^{(k)} - 1}, \quad k = 1, 2, 3.$$

$$y_{i,0}^{(2)} = \Phi_0, \quad i = \overline{0, N_d}, \quad (24)$$

where $N_d = \lceil d/2h_p \rceil$. Conditions (18) – (20) are approximated with the second order using additional points outside the computational domain and Eqs. (21) – (22).

4. Numerical solving of the Schrödinger equation

As in the case of potential computation, the problem for the Schrödinger equation has been solved on the finite domain $0 < \rho < L_p$, $0 < z < L_{z1}$. Since there is an exponential decay of the wave function at infinity, zero boundary conditions for the wave function were set on the boundary of the computational domain, i.e. the boundary condition (15) has been replaced with the conditions:

$$\Psi|_{z=L_{z1}} = 0, \quad \Psi|_{\rho=L_p} = 0. \quad (25)$$

The solution of the Schrödinger equation has been constructed on a grid:

$$\Omega_S = \omega_\rho \times \omega_z^{(1)}.$$

Denoting the values of the wave function at the grid nodes as

$$w_{ij} = \Psi(\rho_i, z_j^{(1)})$$

we get the following difference scheme:

$$-\frac{1}{\rho_i} \left(\rho_{i+1/2} w_\rho \right)_\rho - w_{zz} + V_{ij} w = E_h w, \quad (26)$$

$$w_{i,0} = 0, \quad i = \overline{0, N_\rho}, \quad (27)$$

$$w_{i, N_z^{(1)}} = 0, \quad i = \overline{0, N_\rho}, \quad (28)$$

$$w_{N_\rho, j} = 0, \quad j = \overline{1, N_z^{(1)} - 1}, \quad (29)$$

where E_h is approximate energy value, $V_{ij} = \hat{V}(\rho_i, z_j^{(1)})$.

To increase the order of approximation of the boundary condition (16), the difference approximation of the Schrödinger equation (1) on the axis $\rho = 0$ is used. As a result, we get

$$w_\rho + \frac{h_p}{4} (w_{zz} - V_{0j} w + E_h w) = 0, \quad j \neq j_0, i = 0, \quad (30)$$

where $(0, j_0)$ is the node corresponding to the donor location. The value of the wave function at the donor location can be found from the second order approximation for the condition (17):

$$3w_{0,j_0} = 4w_{1,j_0} - w_{2,j_0} + 2h_p w_{0,j_0}. \quad (31)$$

5. Error estimation

The error of calculating the energy will be different when the electron is in the gate region and when it is located in the region near the donor. In the first case, both the error in calculating the potential and the error of the solution of the Schrödinger equation affect the accuracy of energy calculation.

To estimate the error in calculating the potential, arising from setting conditions (18) – (20) on the boundaries of computational domain, we consider a test problem in which an infinitely thin disc is in a homogeneous medium or on the boundary of two homogeneous media. In this case, the exact analytic expression for the potential is known [18]. Calculations of the gate potential using FDM have been carried out for three model problems with different conditions on the boundaries of the computational domain. In the first problem, the value of the potential at the boundary is assumed to be zero, in the second problem, the potential on the boundary is exact [18] and, in the third problem, conditions (18) – (20) are set on the boundary. The solutions obtained in all three cases were compared with the exact value of the potential in the integral norm on the domain $0 < \rho < 10a^*$, $-10a^* < z < 10a^*$. Calculations for gate diameter $d = 6a^*$, $L = L_\rho = L_{z1} = L_{z3}$ in range from $10a^*$ to $40a^*$

and grid step $h = h_p = h_z^{(k)} = 0.1$ have shown that the use of the conditions (18) – (20) instead of exact boundary conditions does not increase relative energy error even for small computational domains. On the contrary, when using zero boundary conditions on the boundary of the computational domain, the computational error increases sufficiently: the relative error at $L = 10a^*$ with zero boundary conditions is more than twenty much greater than the corresponding error for the problem with exact boundary conditions.

We consider the case, when the electron is localized on the donor, in the absence of external gate field and image charges. Then, at the donor location, the potential in the Schrödinger equation has a singularity, which can lead to additional errors. In order to verify the effect of condition (17) on the accuracy of the calculations, the dependences of the donor ground state energy on the grid step in the problem with condition (17) and with condition (16) at the donor location were calculated (Fig. 1). As one can see from Fig. 1, the use of the condition (16) at the location of the donor leads to the appearance of additional fluctuations in the error caused by the relative position of the donor location ($\rho = 0$, $z = z_0$) with respect to the grid. Thus, the use of conditions (17) has a significant effect on the accuracy of the computations.

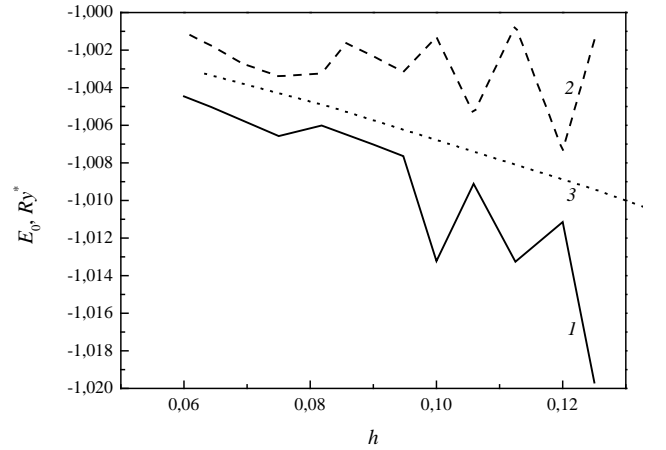


Fig. 1 The effect of non-analyticity of the potential on calculation results of the energies of the ground state (E_0), for the distance from the donor to the semiconductor surface $z_0 = 8a^*$, at zero gate potential and neglecting image charges; h is the grid step. The curves 1, 2 are calculated using FDM and FEM, respectively, with conditions (16) at the donor location, the curve 3 is calculated using FDM with condition (17).

6. Results and discussion

One of the main characteristics of the structure under study is the value of the external electric field at which relocation of electron wave function from the donor to the near-gate region takes place. We denote corresponding critical gate potential as Φ_{0C} . The critical potential depends on a number of parameters of the system: the size of the gate, the position of the donor, the dielectric constant of the media, and the thickness of the dielectric layer.

It had been shown in [19] that the critical field corresponds to a minimum of the difference between the energies of the first excited and ground states ($E_1 - E_0$). Therefore, to determine Φ_{0C} it is sufficient to calculate the energy values of two states, which essentially simplifies the problem.

Dependences of the critical potential on the thickness of the dielectric layer are shown in Fig. 2. As one can see from the figure, the critical potential increases with increasing thickness of the dielectric faster for smaller values of ϵ_2 , e.g., for $t_{ox} = 2a^*$ (≈ 6.3 nm in silicon) $\Phi_{0C} \approx 231$ mV for $\epsilon_2 = 3.8$ and $\Phi_{0C} \approx 95$ mV for $\epsilon_2 = 34.2$. This means that when an insulator with a lower dielectric permittivity is used, the device requires a larger potential at the gate, which additionally increases the probability of dielectric breakdown.

It had been shown in [19] that at zero insulator thickness the redistribution of the ground state wave function depends only on the potential difference between the point of the donor location and semiconductor surface and does not depend on the position of the donor. An analogous dependence also takes place for a nonzero dielectric thickness. We consider the value $\Delta\Phi$ which is the potential difference between the donor and the semiconductor surface:

$$\Delta\Phi = \Phi(0, z_0) - \Phi(0, 0),$$

where

$$\Phi(\rho, z) = \hat{V}_G(\rho, z) + \hat{V}_D(\rho, z) + \hat{V}_{sim}(\rho, z).$$

Correspondingly, we introduce the value of $\Delta\Phi_C$ which is the critical potential difference. The calculations show (Fig. 2) that the value of $\Delta\Phi_C$ is practically independent of the thickness and material of the dielectric. This allows us to use the results of calculations of the critical potential difference for a zero thickness of the dielectric in modeling of structures with an arbitrary thickness of the insulating layer.

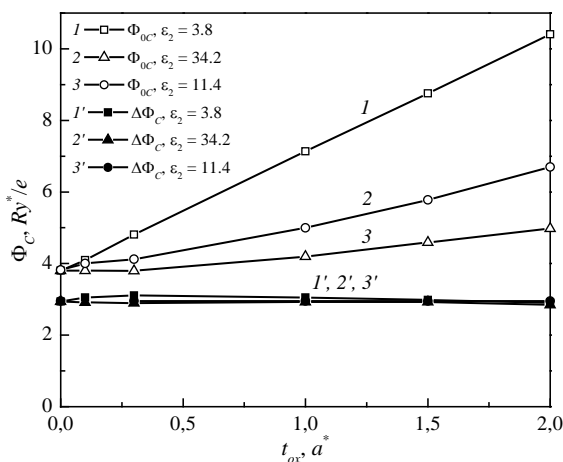


Fig. 2 The dependences of the critical potential (Φ_{OC}) and critical potential difference ($\Delta\Phi_C$) on dielectric thickness for different dielectric constants of the insulating layer; $d = 6a^*$, $z_0 = 8a^*$, $\epsilon_1 = 11.4$, $\epsilon_3 = 1$.

7. Conclusion

Numerical simulation of the electronic structure of a metal-insulator-semiconductor in the presence of a near-surface donor has been carried out. To calculate the gate potential and solve the problem for the stationary Schrödinger equation, a difference scheme has been proposed. A method to approximate the boundary conditions at infinity at calculations of the gate potential has been proposed. It has been found that the use of the proposed boundary conditions makes it possible to significantly reduce errors in calculated potential as compared with the use of zero conditions at the boundaries of calculation domain.

Using the suggested difference schemes, the dependences of the ground and first excited state energies of the donor on the field of disc-shaped gate in a three-layered medium have been obtained. On the basis of these dependences, the critical potential of the gate has been calculated, at which the electron density of the ground states is relocated from the donor to the gate region. It has been found that the critical potential difference between the donor and the semiconductor surface is the controlling parameter when describing the functioning of the device, since it practically does not depend on the thickness and permittivity of the dielectric layer.

References

1. Asenov, A., A.R. Brown, J.H. Davies, S. Kaya, G. Slavcheva. Simulation of intrinsic parameter fluctuations in decanometer and nanometer-scale MOSFETs. – IEEE transactions on electron devices, vol. 50, no. 9, 2003, pp. 1837-1853.

2. Koenraad, P.M., M.E. Flatte. Single dopants in semiconductors. – Nature materials, vol. 10, no. 2, 2011, p. 91.
3. Kane, B.E. A silicon-based nuclear spin quantum computer. – Nature (London), vol. 393, no. 6681, 1998, pp. 133-137.
4. Vrijen, R., E. Yablonovitch, K. Wang, H.W. Jiang, A. Balandin, V. Roychowdhury, T. Mor, D. DiVincenzo. Electron-spin-resonance transistors for quantum computing in silicon-germanium heterostructures. – Physical Review A, vol. 62, no. 1, 2000, p. 012306.
5. Hollenberg, L.C.L., A.S. Dzurak, C. Wellard, A.R. Hamilton, D.J. Reilly, G.J. Milburn, R.G. Clark. Charge-based quantum computing using single donors in semiconductors. – Physical Review B, vol. 69, no. 11, 2004, p. 113301.
6. Smit, G.D.J., S. Rogge, J. Caro, T.M. Klapwijk. Gate-induced ionization of single dopant atoms. – Physical Review B, vol. 68, no. 19, 2003, p. 193302.
7. Kettle, L.M., H.S. Goan, S.C. Smith, C.J. Wellard, L.C. Hollenberg, C.I. Pakes. Numerical study of hydrogenic effective mass theory for an impurity P donor in Si in the presence of an electric field and interfaces. – Physical Review B, vol. 68, no. 7, 2003, p. 075317.
8. MacMillen, D.B., U. Landman. Variational solutions of simple quantum systems subject to variable boundary conditions. II. Shallow donor impurities near semiconductor interfaces: Si, Ge. – J. Chem. Phys, vol. 80, no. 2, 1984, pp. 1691-1702.
9. Calderon, M.J., B. Koiller, S. Das Sarma. Quantum control of donor electrons at the Si-SiO₂ interface. – Physical Review Letters, vol. 96, no. 9, 2006, p. 096802.
10. Calderon, M.J., B. Koiller, S. Das Sarma. External field control of donor electron exchange at the Si/SiO₂ interface. – Physical Review B, vol. 75, no. 12, 2007, p. 125311.
11. Slachmuylders, A.F., B. Partoens, F.M. Peeters, W. Magnus. Effect of a metallic gate on the energy levels of a shallow donor. – Appl. Phys. Lett., vol. 92, no. 8, 2008, p. 083104.
12. Hao, Y.L., A.P. Djotyan, A.A. Avetisyan, F.M. Peeters. Shallow donor states near a semiconductor-insulator-metal interface. – Physical Review B, vol. 80, no. 3, 2009, p. 035329.
13. Nikolyuk, V.A., I.V. Ignatiev. The energy structure of quantum dots induced in quantum wells by a nonuniform electric field. – Semiconductors, vol. 41, no. 12, 2007, pp. 1422-1429.
14. Samarskii, A.A. The theory of difference schemes, New York, CRC Press, 2001.
15. Souza, G.V.B., A. Bruno-Alfonso. Finite-difference calculation of donor energy levels in a spherical quantum dot subject to a magnetic field. – Physica E, vol. 66, 2015, pp. 128-132.
16. Galeriu, C., L.L.Y. Voon, R. Melnik, M. Willatzen. Modeling a nanowire superlattice using the finite difference method in cylindrical polar coordinates. – Comp. Phys. Commun., vol. 157, no. 2, 2004, pp. 147-159.
17. Bingel, W.A. A physical interpretation of the cusp conditions for molecular wave functions. – Theoretica Chimica Acta, vol. 8, no. 1, 1967, pp. 54-61.
18. Smythe, S. Static and dynamic electricity, New York, Hemisphere Publishing, 1988.
19. Levchuk, E.A., L.F. Makarenko. On controlling the electronic states of shallow donors using a finite-size metal gate. – Semiconductors, vol. 50, no. 1, 2016, pp. 89-96.

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ МАГНИТНОГО ГИСТЕРЕЗИСА ПРИ ТРЕХОСНОМ НАПРЯЖЕННОМ СОСТОЯНИИ

MATHEMATICAL MODEL OF MAGNETIC HYSTERESIS UNDER TRIAXIAL STRESS STATE

Mushnikov A.N., PhD. Putilova E.A.

Institute of Engineering Science, Ural Branch of the Russian Academy of Sciences, Ekaterinburg, Russian Federation.

mushnikov@imach.uran.ru

Abstract: The magnetomechanical hysteresis models of Jiles and Sablik is extended to treat magnetic properties in the case of triaxial stress state. Simulation results are compared with experimental results. Magnetic measurements were made under applied loading. To obtain a triaxial stress state, hollow cylindrical samples were subjected to elastic deformation by uniaxial tension (compression) and torsion under internal hydrostatic pressure.

Keywords: MAGNETIC HYSTERESIS, ELASTIC STRESS, TRIAXIAL STRESS

1. Введение

Проблема теоретического описания изменений намагниченности ферромагнетика при воздействии на него внешнего магнитного поля и механических напряжений связана с необходимостью учёта его полной свободной энергии, сложным образом зависящей от внутренних факторов (внутренних и приложенных напряжений, магнитной анизотропии зерен, дислокации различных типов, неравновесных точечных дефектов и включений, фазового состава). Одна из наиболее популярных математических моделей магнитного гистерезиса ферромагнитных материалов была разработана в 1984 году Джайлсом и Атертоном [1] и в дальнейшем была дополнена усилиями других ученых. Модель Джайлса-Атертона имеет более высокую вычислительную производительность, по сравнению с феноменологической моделью Прейзаха и моделью гистерезиса Стонера-Вольфарта [2]. А с учетом модификаций, важными преимуществами модели Джайлса-Атертона являются возможность установить связь с реальными физическими параметрами ферромагнитного материала, расчет предельных и частных циклов гистерезиса на одних и тех же параметрах модели, моделирование материалов с изотропными и анизотропными свойствами, возможность учета действующих механических напряжений.

Для учета механических напряжений модель гистерезиса получила развитие в ряде работ Саблика, Квуна и Буркхардта [3-7]. Однако модель применялась для случаев одноосного и некоторых видов двухосного нагружения. Целью данной работы является расширение модели на случай трехосного напряженного состояния и сравнении расчетных результатов с экспериментальными данными.

2. Базовая модель

Модель Джайлса-Атертона разрабатывалась как попытка создания количественной модели гистерезиса на основе макромагнитной формулировки. Модель описывала изотропные поликристаллические материалы (многодоменные зерна) с движением доменной стенки в качестве основного процесса намагничивания. Модель строится в два шага: на первом вычисляется безгистерезисная намагниченность, затем при помощи системы дифференциальных уравнений моделируется гистерезис, с учетом изменений внешнего магнитного поля. Под безгистерезисной намагниченностью понимается намагниченность, полученная при одновременном действии постоянного поля и переменного поля с убывающей до нуля амплитудой.

Рассмотрим энергию единицы объема домена с магнитным моментом m в внешнем поле напряженностью H :

$$(1) E = \mu_0 m (H + \alpha M),$$

где α – параметр, определяющий междоменную связь. Сумму $H_e = H + \alpha M$ называют действующим (или эффективным) полем. Связь безгистерезисной намагниченности $M_{\text{безгист}}$ с действующим полем описывается некоторой эмпирической формулой $M_{\text{безгист}} = M_{\text{безгист}}(H_e)$, вид которой в значительной степени зависит от свойств материала. Так, например, для изотропных ферромагнетиков Джайлсом предложена модифицированная функция Ланжевена $M_{\text{безгист}} = M_s \left[\coth\left(\frac{H_e}{a}\right) - \frac{a}{H_e} \right]$, где M_s – намагниченность насыщения, a – коэффициент формы (константа для материала, имеет ту же размерность что и действующее поле).

Для более точного описания гистерезиса необходимо учитывать влияние трения при смещении доменной стенки. Фрикционные силы возникают из-за закрепления доменных стенок на дефектах (дислокациях, примесях, границах зерен). Для их учета вводится дополнительный энергетический член E_p – энергия, затрачиваемая на отделение (или закрепление) доменных стенок. $dE_p = k dM$, где k – микроструктурный параметр, пропорциональный плотности закрепления стенок и энергии закрепления. Для одного материала его можно считать константой.

В процессе намагничивания или перемагничивания полная энергия материала равна энергии, как если бы материал был безгистерезисным, за вычетом энергии, потерянной на преодоление дефектов доменными границами. Общая намагниченность выражается суммой обратимой и необратимой намагниченности. Обратимая намагниченность возникает за счет изгибов доменных стенок без отрыва от дефектов. В итоге модель гистерезиса описывается следующей системой из трех дифференциальных уравнений:

$$(2) \frac{dM_{\text{необр}}}{dH} = \frac{M_{\text{безгист}}(H) - M_{\text{необр}}(H)}{\delta - \alpha(M_{\text{безгист}}(H) - M_{\text{необр}}(H))}$$

$$(3) \frac{dM_{\text{обр}}}{dH} = c \left(\frac{dM_{\text{безгист}}}{dH} - \frac{dM_{\text{необр}}}{dH} \right)$$

$$(4) \frac{dM}{dH} = \frac{dM_{\text{обр}}}{dH} + \frac{dM_{\text{необр}}}{dH}$$

где δ – переменная, обозначающая направление изменения поля (+1 во время увеличения H , и -1 в процессе уменьшения H), c – константа.

Для учета механических напряжений модель гистерезиса Джайлса-Атертона получила развитие в ряде работ Саблика,

Квуна и Буркхардта. Основная идея заключается во введении в формулу действующего поля дополнительного слагаемого $H_\sigma(\sigma, M)$, объясняющего изменение энергии элементарного объема под действием механических напряжений. В [7] показано, что для одноосного растяжения/сжатия слагаемое H_σ пропорционально напряжениям, намагнитченности и магнитострикции λ . В то же время магнитострикция может сложным образом зависеть от напряжений, а так же внешнего поля или намагнитченности. Таким образом, выбор функции, описывающей магнитострикцию, является одной из важнейших задач для построения адекватной модели.

Еще одно существенное дополнение модели представлено работе [5]. В ней рассмотрен случай приложения нагрузки несоосно внешнему магнитному полю. Формула действующего поля принимает вид:

$$(5) H_e = H \cos \beta + \alpha M + H_\sigma,$$

Здесь β – угол между направлением внешнего поля и суммарной намагнитченности, $H_\sigma = \frac{3}{2} \frac{\sigma}{\mu_0} \frac{\partial \lambda}{\partial M} (\cos^2 \varphi - \nu \sin^2 \varphi)$, где ν – коэффициент Пуассона.

Исследования влияния двухосных нагрузок на намагничивание, в приложении к модели Джайлса-Атертона, проведены в работах Саблика [3, 4, 6]. Несмотря на ограниченность экспериментальной базы (установка для двухосного нагружения позволяла проводить либо растяжение по обеим осям, либо сжатие по обеим осям, а комбинирование растяжения и сжатия моделировалось через отдельный эксперимент на кручение) в работах приводятся общие идеи, которые можно использовать для вывода формул при других видах деформирования.

3. Механическая формулировка

Идеальным (но неосуществимым) является эксперимент по независимому трехосному растяжению куба с одновременным измерением изменений магнитных характеристик. В данной работе объемное напряженное состояние достигается комбинированием таких видов нагружения, как одноосное растяжение/сжатие, кручение и внутреннее давление. Перечисленные виды нагружений по отдельности могут быть не характерны для конструкций. Их необходимо рассматривать именно с точки зрения возможности создания объемного (трехосного) напряженного состояния с взаимно-независимым изменением всех трех главных напряжений.

Расчеты напряжений вели в предположении об изотропности исследуемого материала. Нормальные напряжения σ_z , вызванные растяжением/сжатием вдоль оси образца, вычисляли по формуле:

$$(6) \sigma_z = \frac{F}{\pi(R_{out}^2 - R_{in}^2)},$$

где F – нагрузка, приложенная к образцу, R_{in} – внутренний радиус образца, R_{out} – внешний радиус образца. Изменением сечения образца при упругом деформировании пренебрегали. Усредненные по объему значения касательных напряжений:

$$(7) \tau = \frac{(R_{out} + R_{in})}{\pi(R_{out}^4 - R_{in}^4)} T.$$

Определение механических напряжений, возникающих под действием гидростатического давления, является решением классической задачи Ламе о толстостенной трубе. Вследствие симметрии приложенных нагрузок напряжения и деформации будут также симметричны относительно продольной оси симметрии цилиндра. Под действием внутреннего давления возникают растягивающие окружные напряжения σ_θ и

сжимающие радиальные напряжения σ_r . Их величины, как функции от радиуса r , определяются формулами:

$$(8) \begin{cases} \sigma_r = \frac{R_{in}^2}{R_{out}^2 - R_{in}^2} \left(1 - \frac{R_{out}^2}{r^2} \right) p, \\ \sigma_\theta = \frac{R_{in}^2}{R_{out}^2 - R_{in}^2} \left(1 + \frac{R_{out}^2}{r^2} \right) p \end{cases}$$

где p – величина гидростатического давления. Так как концы образца жестко зашпелены, то под действием кручения и/или гидростатического давления будут возникать продольные усилия. Однако устройство испытательной установки позволяет определять и компенсировать эти усилия контролируемой осевой нагрузкой F , которая входит в формулу (6). При комбинировании растяжения (сжатия), кручения и внутреннего давления тензор напряжений в цилиндрической системе координат имеет вид:

$$(9) A = \begin{pmatrix} \sigma_r & 0 & 0 \\ 0 & \sigma_\theta & \tau \\ 0 & \tau & \sigma_z \end{pmatrix}.$$

Для нахождения главных напряжений σ_k ($k = 1, 2, 3$) решаем характеристическое уравнение:

$$(10) |A - E\sigma| = 0,$$

где E – единичная матрица. Так как магнитные характеристики измеряли во всем объеме рабочей части образца, то нас будут интересовать усредненные по объему значения главных напряжений, которые примут вид:

$$(11) \begin{cases} \sigma_{j_1} = \frac{1}{R_{out} - R_{in}} \int_{R_{in}}^{R_{out}} \sigma_r dr \\ \sigma_{j_2} = \frac{1}{R_{out} - R_{in}} \int_{R_{in}}^{R_{out}} \frac{\sigma_\theta + \sigma_z + \sqrt{(\sigma_\theta + \sigma_z)^2 + 4(\tau^2 - \sigma_\theta \sigma_z)}}{2} dr \\ \sigma_{j_3} = \frac{1}{R_{out} - R_{in}} \int_{R_{in}}^{R_{out}} \frac{\sigma_\theta + \sigma_z - \sqrt{(\sigma_\theta + \sigma_z)^2 + 4(\tau^2 - \sigma_\theta \sigma_z)}}{2} dr \end{cases},$$

где индексы j_1, j_2 и j_3 принимают значения 1, 2, 3 из условия $\sigma_1 \geq \sigma_2 \geq \sigma_3$.

Для определения системы координат, в которой тензор напряжений имеет диагональный вид, необходимо определить направления главных напряжений. Поочередно подставим найденные главные напряжения в систему уравнений

$$(12) \begin{cases} (A - E\sigma_j) \vec{n}_j = 0, \\ n_{rj}^2 + n_{\theta j}^2 + n_{zj}^2 = 1 \end{cases},$$

где \vec{n}_j – вектор, с компонентами $n_{rj}, n_{\theta j}$ и n_{zj} , ($k = 1..3$). В этой системе три из четырех уравнений являются линейно независимыми. Таким образом, решив три системы, получим направляющие косинусы главных напряжений.

4. Модификация модели

Пусть в цилиндрической системе координат в каждом элементарном объеме действуют три взаимно-перпендикулярных главных напряжения $\sigma_1, \sigma_2, \sigma_3$. Каждое из них может быть направлено под произвольным углом ψ_j по отношению к направлению намагничивающего поля.

Для практики нас не интересуют обратимая и необратимая составляющие намагнитченности по отдельности. Подставив (2) и (3) в (4), и с учетом $M_{необр} = \frac{M - cM_{беззуст}}{1 + c}$, получим единое уравнение модели, описывающее изменение суммарной намагнитченности:

$$(13) \frac{dM}{dH} = \frac{c}{1+c} \frac{dM_{\text{безгист}}}{dH} + \frac{1}{1+c} \frac{M_{\text{безгист}}(H) - M(H)}{\delta k - \alpha(M_{\text{безгист}}(H) - M(H))}.$$

Для описания связи безгистерезисной намагниченности $M_{\text{безгист}}$ с действующим полем будем использовать модифицированную функцию Ланжевена, так как она достаточно хорошо подходит для конструктивных сталей:

$$(14) M_{\text{безгист}} = M_s \left[\coth\left(\frac{H_e}{a}\right) - \frac{a}{H_e} \right],$$

Магнитоупругая энергия под действием трех произвольных механических напряжений складывается из трех независимых частей, поэтому изменение энергии элементарного объема примет следующий вид:

$$(15) H_\sigma = \frac{3}{2\mu_0} \sum_{j=1}^3 \left[\sigma_j \frac{\partial \lambda_j}{\partial M} (\cos^2 \varphi_j - \nu \sin^2 \varphi_j) \right]$$

Здесь λ_j определяется по аналогии с [5], но вместо параметра b_φ используем $b_{\varphi_j} = \frac{2}{3} b(1+\nu) \left(1 - \frac{3}{2} \sin^2 \varphi_j \right)$:

$$(16) \lambda_j = \frac{2}{3} \frac{b}{|b|} \left(\sqrt{\left(\frac{b_{\varphi_j}}{Y} \right)^2 + \frac{2(\Phi_{\text{mag}}(M_s) - \Phi_{\text{mag}}(M))}{Y}} - \sqrt{\left(\frac{b_{\varphi_j}}{Y} \right)^2 + \frac{2\Phi_{\text{mag}}(M_s)}{Y}} \right),$$

где Y – модуль Юнга, $\Phi_{\text{mag}}(M) = \frac{1}{2} \alpha \mu_0 M^2$. Таким образом, действующее поле H_e (5) будет определено следующим уравнением:

$$(17) H_e = H \cos \beta + \alpha M + \frac{3}{2\mu_0} \sum_{j=1}^3 \left[\sigma_j \frac{\partial \lambda_j}{\partial M} (\cos^2 \varphi_j - \nu \sin^2 \varphi_j) \right].$$

Для нахождения углов φ_k необходимо решить три уравнения:

$$(18) \frac{d}{d\varphi_j} \left[\frac{1}{2} \alpha \mu_0 M^2 - \mu_0 H M \cos(\arccos n_j - \varphi_j) - \frac{3}{2} \lambda_j \sigma_j (\cos^2 \varphi_j - \nu \sin^2 \varphi_j) \right] = 0$$

где $j=1..3$.

Угол β между направлением внешнего поля и вектором намагниченности, необходимый для расчета действующего поля (17), можно выразить через найденные углы с учетом одинакового соотношения между углами от осей до вектора намагничивания и углами от осей до линии действия механических напряжений. Обозначим n_{rj}^* , $n_{\theta j}^*$, n_{zj}^* направляющие косинусы вектора намагниченности, если бы на него действовало только одно напряжение σ_j ($j=1..3$). То есть $\cos(\psi_j - \varphi_j) = n_{zj}^*$. Тогда

$$(19) n_{rj}^* = \cos\left(\frac{\varphi_j}{\psi_j} \arccos(n_{rj})\right)$$

$$(20) n_{\theta j}^* = \cos\left(\frac{\varphi_j}{\psi_j} \arccos(n_{\theta j})\right)$$

Для нахождения направления намагниченности под действием трех напряжений по координатно сложим три вектора. Без нормирования они примут вид:

$$(21) n_r^* = \sum_{j=1}^3 \cos\left(\frac{\varphi_j}{\psi_j} \arccos(n_{rj})\right)$$

$$(22) n_\theta^* = \sum_{j=1}^3 \cos\left(\frac{\varphi_j}{\psi_j} \arccos(n_{\theta j})\right)$$

$$(23) n_z^* = \sum_{j=1}^3 \cos(\psi_j - \varphi_j)$$

Искомый угол β определяется выражением

$$(24) \beta = \frac{1}{|n^*|} \arccos n_z^*,$$

где норма $|n^*| = \sqrt{(n_r^*)^2 + (n_\theta^*)^2 + (n_z^*)^2}$. А так же учитываем, что $\cos \psi_j = n_{zj}^*$ – известные величины, рассчитанные по (12).

Таким образом, моделирование гистерезиса при объемном напряженном состоянии необходимо выполнять в следующей последовательности:

1. Рассчитать главные напряжения (11) и их направления (12).

2. Выбрать диапазон изменений поля H и шаг dH .

3. На каждом шаге решать дифференциальное уравнение (13). В нем безгистерезисная намагниченность определена уравнением (14), где действующее поле задано выражением (17), магнитострикция определена равенством (16), углы между вектором намагниченности и направлениями действия главных напряжений φ_1 , φ_2 и φ_3 являются решениями системы трех уравнений (18), а угол β между векторами намагниченности и внешнего магнитного поля задан формулой (24).

5. Результаты и их обсуждение

Экспериментальные исследования проводили на образцах, вырезанных вдоль направления прокатки из сварных прямошовных труб стали контролируемой прокатки класса прочности Х80. Для испытаний на комбинированное нагружение по схеме, описанной в разделе 3 использовали полые цилиндрические образцы, внешний диаметр которых был равен 12 мм, а внутренний – 9 мм. Испытания проводили при комнатной температуре на универсальной испытательной машине с максимальным усилием растяжения 50 кН и максимальным крутящим моментом 200 Н*м. По достижению определенной степени деформации процесс нагружения приостанавливали без разгрузки образца, и при помощи магнитно-измерительного комплекса Remagraph C-500 в замкнутой магнитной цепи регистрировали петли магнитного гистерезиса. Магнитное поле напряженностью до 60 кА/м прикладывали вдоль оси образца.

Несмотря на то, что все параметры модели Джайлса-Атертона (и её модификаций) имеют физический смысл, их экспериментальное определение является не тривиальной задачей. Более практичными являются методы подбора значений параметров, по которым может быть построена петля гистерезиса, с достаточной точностью воспроизводящая петлю, полученную в эксперименте. Параметр M_s напрямую получали из определенной экспериментально предельной петли магнитного гистерезиса. Несмотря на то, что в эксперименте невозможно достижение абсолютного насыщения материала, отличие от него состояния технического насыщения (в поле напряженностью 60 кА/м) можно считать несущественным. Для нахождения оптимальных параметров минимизируется невязка одновременно по всем неизвестным:

$$(25) R = \sum_{i=1}^n (M_i^{\text{exp}} - M(H_i))^2$$

где (H_i, M_i^{exp}) – экспериментально полученные точки поле-намагниченность, n – количество точек в измеренной петле гистерезиса (~2000). Для каждого набора параметров и для

каждой точки значения $M(H_i)$ рассчитываются по математической модели, описанной в предыдущем разделе.

В качестве примера на рисунке 1 приведена серия петель гистерезиса исследованной стали при различных величинах максимального действующего поля.

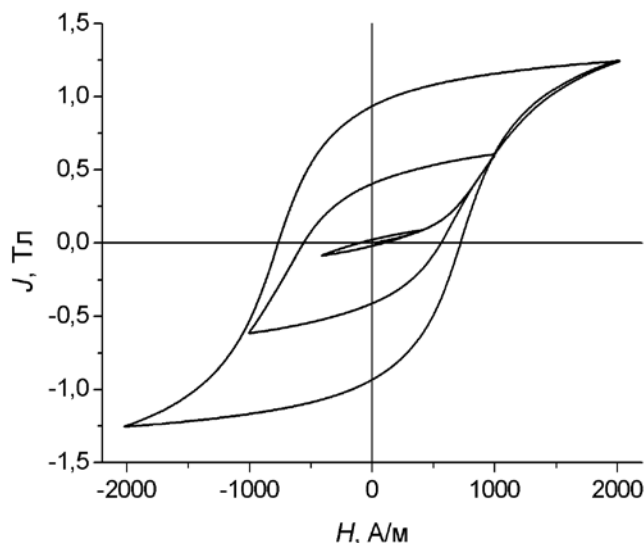


Рис. 1 Частные петли гистерезиса стали класса прочности X80.

Этим петлям соответствуют параметры математической модели $a = 385.577976$ (А/м), $k = 813.94$ (А/м), $c = 0.2$, $M_s = 1.663 \cdot 10^6$ (А/м), $\alpha = 1.968 \cdot 10^{-6}$. Относительное максимальное отклонение модели от экспериментальных значений не превышает 15%.

6. Заключение

Предложена модификация математической модели гистерезиса Джайлса-Атертона для случая трехосного напряженного состояния. Сравнение расчетных результатов с экспериментальными данными на стали класса прочности X80 показало адекватность модели.

Работа выполнена при поддержке гранта РФФИ №16-38-00598 мол_а.

7. Литература

1. Jiles D. C., Atherton D. L. Theory of ferromagnetic hysteresis // Journal of Applied Physics, 1984, Vol. 55, p. 2115-2120.
2. Liorzou F., Phelps B., Atherton D. L. Macroscopic Models of Magnetization // IEEE TRANSACTIONS ON MAGNETICS, 2000, Vol. 36, № 2, p. 418-428.
3. Sablik M.J., Burkhardt G.L., Kwun H. Application of Hysteresis Modeling to Magnetic Techniques for Monitoring Biaxial Stress. – Proc. Eleventh Symposium on Energy Engineering Sciences, Argonne, IL (CONF-9305134, DOE, 1993).
4. Sablik M.J., Riley L.A., Burkhardt G.L., Kwun H., Cannell P.Y., Watts K.T., Langman R.A. Micromagnetic model for biaxial stress effects on magnetic properties. - Journal of Magnetism and Magnetic Materials, 1994, v. 132, №1-3, p. 131-148.
5. Sablik M.J., Rubin S.W., Riley L.A., Jiles D.C., Kaminski D.A., Biner S.B. A model for hysteretic magnetic properties under the application of noncoaxial stress and field. - Journal of Applied Physics, 1993, v. 74, №1, p. 480-488.
6. Sablik M.J., Jiles D.C. Modeling the Effects of Torsional Stress on Hysteretic Magnetization. – IEEE Transactions on Magnetics, 1999, v. 35, №1, p. 498-504.
7. M. J. Sablik, H. Kwun, G. L. Burkhardt, D. C. Jiles. Model for the effect of tensile and compressive stress on ferromagnetic hysteresis // Journal of Applied Physics, 1987, Vol. 61, p. 3799-3801.

MATHEMATICAL MODEL OF SUSTAINABLE INTEGRATED BIOETHANOL SUPPLY CHAINS

Eng. Dzhelil Y., Prof. DSc Ivanov B., Eng. Ganey E., Assoc. Prof. PhD Dobrudzhaliyev D.
Bulgarian Academy of Sciences, Institute of Chemical Engineering, 1113 Sofia, BULGARIA,
E-mail: unzile_20@abv.bg, bivanov1946@gmail.com, evgeniy_ganayev@abv.bg, dragodob@yahoo.com

Abstract: This paper focuses on designing mathematical model a integrated bioethanol supply chain (IBSC) that will account for economic and environmental aspects of sustainability. A mixed integer linear programming model is proposed to design an optimal IBSC. Bioethanol production from renewable biomass has experienced increased interest in order to reduce Bulgarian dependence on imported oil and reduce carbon emissions. Concerns regarding cost efficiency and environmental problems result in significant challenges that hinder the increased bioethanol production from renewable biomass. The model considers key supply chain activities including biomass harvesting/processing and transportation. The model uses the delivered feedstock cost, energy consumption, and GHG emissions as system performance criteria. The utility of the supply chain simulation model is demonstrated by considering a biomass supply chain for a biofuel facility in Bulgarian scale. The results show that the model is a useful tool for supply chain management, including selection of the optimal bioethanol facility location, logistics design, inventory management, and information exchange.

KEYWORDS: BIOETHANOL SUPPLY CHAIN, MATHEMATICAL MODEL, ECONOMIC AND ENVIRONMENTAL ASPECTS

1. Introduction

Production and use of biofuels are promoted worldwide. Their use could potentially reduce emissions of greenhouse gases and the need for fossil fuels [1]. Accordingly, the European Union imposes a mandatory target of 10% biofuels by 2020 [2]. These fuels are produced from biomass. Their use for energy purposes has the potential to provide important benefits. Burning them releases such amount of CO₂ as was absorbed by the biomass in its formation [3]. Another advantage of biomass is its availability in the world due to its variety of sources. Despite the advantages of biomass with increasing quantities of biofuels to achieve the objectives of the European Union, this is accompanied by growing quantities of waste products. These wastes are related to the lifecycle of biofuels from crop cultivation, transportation, production to distribution and use. The main liquid biofuels are bioethanol and biodiesel. Depending on the raw material used, production is considered in three generations.

The first generation used as feedstock crops containing sugar and starch to produce bioethanol, and oilseed crops to produce biodiesel [4]. In the production of biodiesel, the advantage of these materials is that they can be grown on contaminated and saline soils, as the process does not affect the fuel production. The drawback is that they raise issues related to their competitiveness in the food sector. These materials also have a negative impact in terms of the quantity of water consumed. This is related to their cultivation that requires significant amounts of water resources. Excessive use of fertilizers, pesticides and chemicals to grow them also leads to accumulation of pollutants in groundwater that can penetrate into water courses and thus degrade water quality.

According to the second generation, bioethanol is produced by using as raw material waste biomass (agricultural and forest waste) [5], i.e. lignocellulose which is transformed into a valuable resource as bioethanol. Biofuel production second generation is an excellent way to deal with increasingly restrictive national and European regulations in this area and the use of organic waste for energy production and fertilizer as a byproduct. Logistics and use of these materials can be challenging due to the fact that they are usually dispersed. Another disadvantage from an environmental perspective is the need for further purification and processing.

The third generation comprises production from microalgae which occur as promising feedstock for biofuel production. The advantage of this biomass is that it is a year-round production and does not compete with the food industry.

The main technologies for production of bioethanol are fermentation, distillation and dehydration [6]. The wastes of biofuels are divided into production and performance. The technological waste is produced mainly in the creation of products that occur as waste production. The management of these wastes is related to their reduction, recovery and disposal. These guidelines are united in the idea of acquiring more sophisticated production processes. Efforts are focused on the use of new sources of raw materials, new processes, and new ways of realization of the side

products. The use of by-products as raw materials for other production closes cycle in the supply chain, reducing the price of obtaining fuel. Operational waste associated with gases and emissions released during operation and the burning of biofuels.

2. Aim

The present study deals with the issue of designing optimal Integrated Bioethanol Supply Chains (IBSC) for waste management in the process of biofuel production and usage. Tools have been developed for formulation of a mathematical model for description of the parameter, the restrictions and the goal function.

3. Problem statement

The problem addressed in this work can be formally stated as follows. Given are a set of biofuel crops that can be converted to bioethanol. These include agricultural feedstock's e.g. wheat, corn, etc. A planning horizon of one year for government regulations including manufacturing, construction and carbon tax is considered. A IBSC network superstructure including a set of harvesting sites and a set of demand zones, as well as the potential locations of a number of collection facilities and bio refineries are set. Data for biofuel crops production and harvesting are also given. For each demand zone, the biofuel demand is given, and the environmental burden associated with bioethanol distribution in the local region is known. For each transportation link, the transportation capacity, available transportation modes, distance, and emissions of each transportation type are known.

3.1. General Formulation of the Problem

The overall problem can be summarized, as follows:

- Optimal locations of biofuel production centers,
- Demand for petroleum fuel for each of the demand centers,
- The minimum required ratio between petroleum fuel and biofuel for blending,
- Biomass feedstock types and their geographical availability,
- Specific Green House Gas (GHG) emission factors of the biofuel life cycle stages,
- Potential areas where systems for utilization of solid waste from production can be installed.

The objectives are to minimize total cost of a IBSC by optimizing the following decision variables:

- Supply chain network structure,
- Locations and scales of bioethanol production facilities and biomass cultivation sites,
- Flows of each biomass type and bioethanol between regions,
- Modes of transport for delivery for biomass and bioethanol,
- The GHG emissions for each stage in the life cycle,
- Supply strategy for biomass to be delivered to facilities,
- Distribution processes for biofuel to be sent to demand zones.

4. Model formulation

The role of the optimization model is to identify what combination of options is the most efficient approach to supply the facility. The problem for the optimal location of bioethanol production plants

and the efficient use of the available land is formulated as a MILP model with the following notation:

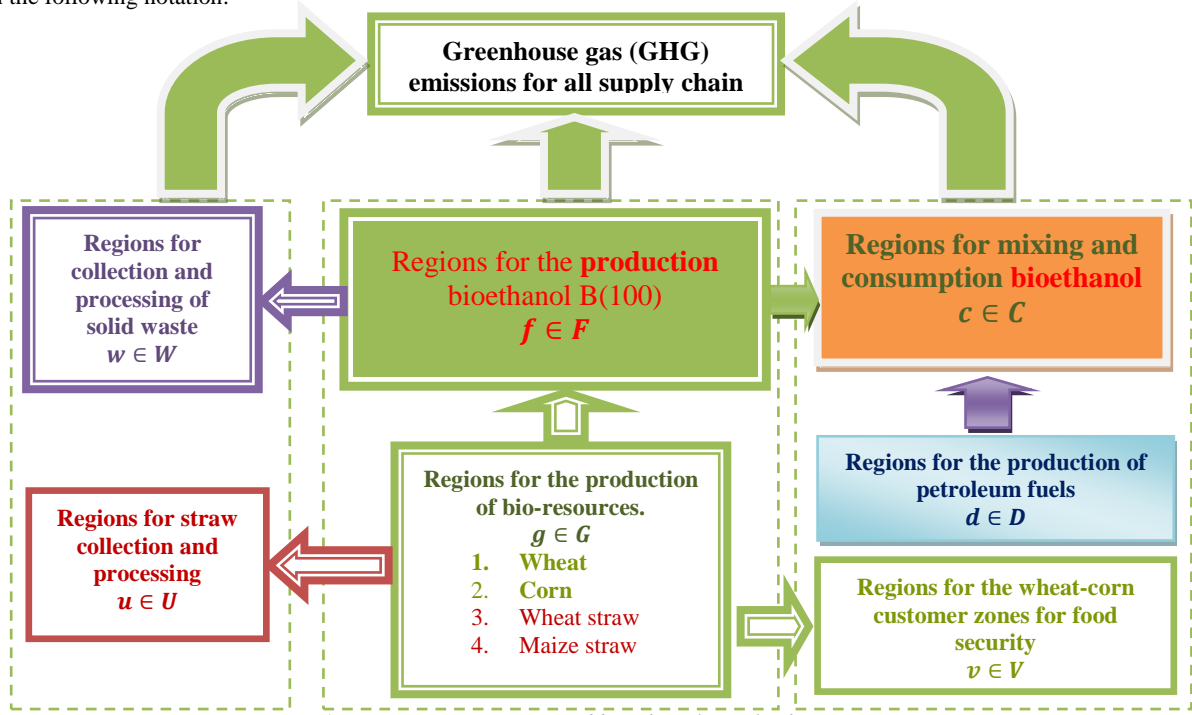


Figure 1. Superstructure integrated bioethanol supply chain (IBSC)

4.1. Mathematical model description

To start with the description of the MILP model, we first introduce the parameters, that are constant and known a priori, and the variables that are subject to optimization. Then we describe step by step the mathematical model by presenting the objective function and all the constraints. First of all, we introduce the set of time intervals of the horizon of planning $t = \{1, 2, \dots, T\}$.

In this article the mathematical model that is used in the network design is described. Before describing the mathematical model, the input parameters, the decision variables, and the sets, subsets and indices are listed below.

4.1.1. Sets, subsets and indices

The following sets and subsets are introduced:

Sets/indices

- I Set of biomass types indexed by i ;
- LF Set of transport modes indexed by lf ;
- P Set of plant size intervals indexed by $p = \overline{1, N_p}$;
- S Set of utilization plant size intervals indexed by $s = \overline{1, N_s}$;
- GF Set of regions of the territorial division indexed by gf ;
- K Set of proportion of bioethanol and gasoline indexed by k ;
- T Set of time intervals, indexed by t .

Subsets/indices

- B Set of transport modes for bioethanol and gasoline is a subset of LF ($B \subset LF$) indexed by b ;
- L Set of transport modes for biomass is a subset of LF ($L \subset LF$) indexed by l ;
- M Set of transport modes for solid wastes is a subset of LF ($M \subset LF$) indexed by m ;
- E Set of transport modes for straw is a subset of LF ($E \subset LF$) indexed by e ;
- Z Set of transport modes for wheat-corn for food security is a subset of LF ($Z \subset LF$) indexed by z ;
- F Set of candidate regions for bioethanol plants established, which is a subset of GF ($F \subset GF$) indexed by f ;
- C Set of bioethanol mixing and customer zones, which is a subset of GF ($C \subset GF$) indexed by c ;

- D Set for delivery and production gasoline, which is a subset of GF ($D \subset GF$) indexed by d ;
- W Set for regions for collection and processing of solid waste, which is a subset of GF ($W \subset GF$) indexed by w ;
- U Set for regions for straw collection and processing, which is a subset of GF ($U \subset GF$) indexed by u ;
- V Set for regions for the wheat-corn customer zones, which is a subset of GF ($V \subset GF$) indexed by v ;

4.1.2. Input parameters for the problem

Environmental parameters:

- $EFBP_{ip}$ Emission factor for bioethanol production from biomass type $i \in I$ using technology $p \in P$, [$kg CO_2 - eq / ton biofuel$];
- ESW Emission factor of pollution caused by one tone of solid waste if not used, [$\frac{kg CO_2 - eq}{ton solid waste}$];
- $EFDp_d$ Emission factor for gasoline production in region $d \in D$, [$kg CO_2 - eq / ton gasoline$];
- $EFTRA_{il}$ Emission factor for biomass $i \in I$ supply via mode $l \in L$, [$kg CO_2 - eq / ton km$];
- $EFTRB_b$ Emission factor for bioethanol supply via mode $b \in B$, [$kg CO_2 - eq / ton km$];
- $EFTM_{il}$ Emission factor of transportation of biomass $i \in I$ for mode $l \in L$, [$kg CO_2 - eq / ton km$];
- $EFTB_b$ Emission factor of transportation of bioethanol and gasoline for mode $b \in B$, [$kg CO_2 - eq / ton km$];
- $EFTRW_m$ Emission factor for transport of solid waste with transport $m \in M$, [$kg CO_2 - eq / ton km$];
- $EFTRU_e$ Emission factor for transport of straw with transport $e \in E$, [$kg CO_2 - eq / ton km$];
- $EFTRV_z$ Emission factor for transport of wheat-corn for food security with transport $z \in Z$, [$kg CO_2 - eq / ton km$];

ECB, ECG Emissions emitted during the combustion of CO_2 unit bioethanol and, gasoline, $[kg\ CO_2 - eq/ton\ bioethanol]$ or; $[kg\ CO_2 - eq/ton\ gasoline]$.

Monetary parameter:

$CosB_p, CosW_s$ Capital investment of bioethanol plant size $p \in P$ and capital investment of solid waste plant size $s \in S$, $[\$]$;

C_{CO_2} Carbon tax per unit of carbon emitted from the operation of the IBSC, $[\$/kg\ CO_2 - eq]$;

PG Price of gasoline, $[\$/ton]$;

$UTI_{il}, UTB_b, UTG_b, UTS_m, UTU_e, UTV_z$, Unit transport cost for biomass $i \in I$, via mode $l \in L$, bioethanol via mode $b \in B$, gasoline via mode $b \in B$, solid wastes via mode $m \in M$, straw via mode $e \in E$, wheat-corn for food security via mode z , $[\$/ton\ km]$;

Technical parameters:

PB_p^{MAX} / PB_p^{MIN} Maximum/Minimum annual plant capacity of size $p \in P$ for bioethanol production, $[ton/year]$;

ENO, ENB Energy equivalent unit of gasoline&bioethanol, $[GJ/ton]$;

$ADD_{dcb}, ADG_{gfl}, ADF_{fcb}, ADU_{gue}, ADW_{fwm}, ADV_{gvz}$ Actual delivery distance between grids via model of transport ($b \in B, l \in L, e \in E, m \in M, z \in Z$), $[km]$;

SW_{ip} The total amount of solid waste generated for production of bioethanol using biomass i for technology p , $[\frac{ton\ solid\ waste}{ton\ biofuel}]$;

$JobB_p, JobO_p$ The number of jobs needed to build and operation a bio-refinery with size $p \in P$ for year;

$JobG_{ig}$ The number of jobs required to grow a unit of $i \in I$ biosource in the region $g \in G$ per year.

Environmental parameters depending on the time interval:

$EFBC_{igt}$ Emission factor for cultivation of biomass type $i \in I$ in region $g \in G$ for each time interval t , $[kg\ CO_2 - eq/ton\ biomass]$;

TEI_t^{MAX} Maximum total environmental impact, $[kg\ CO_2 - eq\ d^{-1}]$.

Monetary parameters depending on the time interval

ζ_t Interest rate, %;

ε_t Discount factor;

M_{ft}^{const} Factor to the change of the base price, depending on the region $f \in F$ where the plant is installed, $[Dimensionless]$;

$Cost_{pft}^F$ Capital investment of plant size $p \in P$ for bioethanol production in each zones $f \in F$, $[\$]$;

INS_{ft} The government incentive includes construction incentive and volumetric from region $f \in F$, $[\$/ton]$;

UPC_{igt} Unit production costs for biomass type $i \in I$ in region $g \in G$ for each time interval $t \in T$, $[\$/ton]$;

UPB_{ipft} Unit bioethanol production cost from biomass type $i \in I$ at a biorafinery of scale $p \in P$ installed in region $f \in F$, $[\$/ton]$;

UPD_{dt} Unit gasoline production cost at a rafinery d , $[\$/ton]$.

Technical parameters depending on the time interval

K_{ct}^{mix} Proportion of bioethanol and gasoline subject of mixing for each of the customer zones, $[Dimensionless]$;

A_{gt}^S Set-aside area available in region $g \in G$ for biomass production for each time interval $t \in T$, $[ha]$;

A_{gt}^{Food} Set-aside area available in region $g \in G$ for food, $[ha]$;

β_{igt} Production rate of biomass i in region $g \in G$, $[ton/ha]$;

LT_t Duration of time intervals $t \in T$, $[year]$;

α_t Operating period for IBSC in a year, $[d/year]$;

γ_{ipt} Biomass to bioethanol conversion factor specific for biomass i using technology p , $[ton_bioethanol/ton_biomass]$;

YO_{ct} Gasoline demand in customer zones $c \in C$, $[ton/year]$;

$PBI_{igt}^{MIN} / PBI_{igt}^{MAX}$ Minimum/ Maximum biomass of type $i \in I$ which can be produced in the region, $g \in G$ per year, $[ton/year]$;

QT_{igt}^{MAX} Maximum flow rate of biomass i from region g , $[ton/d]$;

QB_{ft}^{MAX} Maximum flow rate of bioethanol from region f , $[ton/d]$;

QD_{dt}^{MAX} Maximum flow rate of gasoline from region d , $[ton/d]$;

QW_{ft}^{MAX} Maximum flow rate of solid wastes from f , $[ton/d]$;

QU_{gt}^{MAX} Maximum flow rate of straw from region $g \in G$, $[ton/d]$;

QV_{gt}^{MAX} Maximum flow rate of wheat-corn from $g \in G$, $[ton/d]$;

4.1.3. Decision variables for the problem X_t

To find the optimal configuration of the IBSC, the following decision variables are required:

A/ Positive continuous variables

PBB_{igt} Biomass i demand in region $g \in G$ at time interval $t \in T$;

QI_{igft} Flow rate of biomass $i \in I$ via mode $l \in L$ from region $g \in G$ to $f \in F$, for each $t \in T$, $[ton/d]$;

QB_{fcbt} Flow rate of bioethanol produced from all biomass $i \in I$ via mode $b \in B$ from region $f \in F$ to $c \in C$ for each $t \in T$, $[ton/d]$;

QBP_{ifcbpt} Flow rate of bioethanol produced from biomass i via mode b from f to c using technology p for each $t \in T$, $[ton/d]$;

QD_{dcbt} Flow rate of gasoline via mode $b \in B$ from region $d \in D$ to $c \in C$, for each time interval $t \in T$, $[ton/d]$;

QW_{fwmt} Flow rate of solid waste via mode $m \in M$ from region $f \in F$ to $w \in W$, for each $t \in T$, $[ton/d]$;

QU_{guet} Flow rate of straw collection and processing via mode e from region g to u , for each $t \in T$, $[ton/d]$;

QV_{gvzt} Flow rate of wheat-corn for food security via mode $z \in Z$ from region $g \in G$ to $v \in V$, for each $t \in T$, $[ton/d]$;

QED_{ct} Quantity of gasoline to be supplied to meet the energy needs of the region $c \in C$, for each $t \in T$, $[ton/year]$;

QEB_{ct} Quantity of bioethanol produced from biomass to be supplied to meet the energy needs of the region $c \in C$, $[ton/year]$;

A_{igt} Land occupied by crop $i \in I$ in region $g \in G$, $[ha]$;

A_{igt}^F Land by crops $i \in I$ needed for food security of the population in the region $g \in G$, for each $t \in T$, $[ha]$;

B/ Binary variables

X_{igft} 0-1 variable, equal to 1 if a biomass type i is transported from region g to f using transport l , and 0 otherwise at $t \in T$;

Y_{fcbt} 0-1 variable, equal to 1 if a bioethanol is transported from region f to c using transport b, l , and 0 otherwise at $t \in T$;

WS_{fwmt} 0-1 variable, equal to 1 if a solid waste is transported from region f to w using transport m and 0 otherwise for each $t \in T$;

WU_{guet} 0-1 variable, equal to 1 if a straw is transported from region g to $u \in U$ using transport $e \in E$ and 0 otherwise for each $t \in T$;

WV_{gvzt} 0-1 variable, equal to 1 if a wheat-corn is transported from region g to v using transport z and 0 otherwise for each $t \in T$;

ZW_{swt} 0-1 variable, equal to 1 if a solid waste utilization plant size s is installed in region w and 0 otherwise at time interval $t \in T$;

ZWF_{swt} 0-1 variable, equal to 1 if a solid waste utilization plant size s is to be working in region w and 0 otherwise at $t \in T$, which includes the plants installed in the previous time and the new ones built during this time which is calculate with equation $ZWF_{swt} = ZWF_{sw(t-1)} + ZW_{swt}$ for first year ($t=1$) configuration is set by initializing $ZWF_{sw'1} = ZW_{sw'1}$;

Z_{pft} 0-1 variable, equal to 1 if a bioethanol production plant size p is to be established in region f and 0 otherwise for each $t \in T$;

ZF_{pft} 0-1 variable, equal to 1 if a bioethanol production plant size $p \in P$ is to be working in region $f \in F$ and 0 otherwise at time interval $t \in T$, which includes the plants installed in the previous time interval and the new ones built during this time interval which is calculate with equation $ZF_{pft} = ZF_{pft(t-1)} + Z_{pft}$ for first year ($t=1$) configuration is set by initializing $ZF_{sw'1} = Z_{sw'1}$;

PD_{dt} 0-1 variable, equal to 1 if a gasoline is manufactured by the region $d \in D$ and 0 otherwise at time interval $t \in T$;

DT_{dcbt} 0-1 variable, equal to 1 if a gasoline is transported from region d to c using transport b and 0 otherwise for each $t \in T$.

4.1. Basic Relationships

As noted above, the assessment of IBSC production and distribution of bioethanol will be made by environmental and economic criteria.

4.1.1. Model of total environmental impact of IBSC

The environmental impact of the IBSC is measured in terms of total GHG emissions ($kg CO_2 - eq$) stemming from supply chain activities and the total emissions are converted to carbon credits by multiplying them with the carbon price in the market. The environmental objective is to minimize the total annual GHG emission resulting from the operations of the IBSC. The formulation of this objective is based on the field-to-wheel life cycle analysis, which takes into account the following life cycle stages of biomass-based liquid transportation fuels:

- biomass cultivation, growth and acquisition,
- biomass transportation from source locations to facilities,
- transportation of bioethanol facilities to the demand zones,
- local distribution of liquid transportation fuels in demand zones,
- emissions from bioethanol and gasoline usage.

Ecological assessment criteria will represent the total environmental impact at work on IBSC through the resulting GHG emissions for each time interval t . These emissions are equal to the sum of the impact that each of the stages of life cycle has on the environment. The GHG emission rate is defined as follows for each $t \in T$:

$$TEI_t = ELS_t + ELB_t + ELD_t + ETT_t + ESW_t + ECAR_t, \forall t \quad (1)$$

where

TEI_t Total GHG impact at work on IBSC [$kg CO_2 - eq d^{-1}$];

$\{ELS_t, ELB_t, ELD_t, ETT_t\}$ GHG impact of life cycle stages;

$ECAR_t$ Emissions from bioethanol and gasoline usage in vehicle operations [$kg CO_2 - eq d^{-1}$];

ESW_t Emissions from utilization solid waste for each $t \in T$.

Evaluation of environmental impact at every stage of life cycle is:

- Growing biomass ELS_t ;
- Production of bioethanol ELB_t ;
- Production of petroleum gasoline ELD_t ;
- Utilization of solid wastes ESW_t ;
- Transportation biomass ETA_t ;
- Transportation bioethanol ETE_t ;
- Transportation gasoline ETD_t ;
- Transportation of solid waste ETW_t ;
- Transportation of straw ETU_t ;

J. Transportation of wheat-corn for food security ETV_t ;

K. Usage bioethanol and gasoline $ECAR_t$.

1/ Greenhouse gases to grow biomass ELS_t , [$kg CO_2 - eq d^{-1}$]

GHG emissions resulting from the production of biomass depend on the cultivation practice adopted as well as on the geographical region in which the biomass crop has been established [7]. In particular, the actual environmental performance is affected by fertilisers and pesticides usage, irrigation techniques and soil characteristics. The factor may differ strongly from one production region to another. Accordingly, the biomass production stage is defined as follows:

$$ELS_t = \sum_{i \in I} \sum_{g \in G} \left(EFBC_{igt} \frac{\beta_{igt} A_{igt}}{\alpha_t} \right), \forall t, \quad (2)$$

2/ Total GHG emissions from bioethanol production ELB_t

The environmental impact of the bioethanol production stage is related to raw materials and the technology employed for the production of bioethanol.

$$ELB_t = \sum_{i \in I} \sum_{f \in F} \sum_{c \in C} \sum_{b \in B} \sum_{p \in P} (EFBP_{ip} QBP_{ifcbpt}), \forall t \quad (3)$$

Since only one of the technologies $p \in P$ can be selected for a region $f \in F$ (which is guaranteed by the condition $\sum_{p \in P} ZF_{pft} \leq 1.0 \forall t, f$), it QBP_{ifcbpt} is equal to "0" for all except

$p \in P$ for the selected technology. This is ensured by implementing the inequality $G^{MAX} ZF_{pft} \geq QBP_{ifcbpt}$, $\forall i, f, c, b, p, t$ where G^{MAX} there is a large enough number.

3/ Total GHG emissions from gasoline production ELD_t

$$ELD_t = \sum_{d \in D} \sum_{c \in C} \sum_{b \in B} EDP_{dt} QD_{dcbt}, \forall t \quad (4)$$

4/ The environmental impact of transportation ETT_t

The global warming impact related to both biomass supply and fuel distribution depends on the use of different transport means fuelled with fossil energy, typically either conventional oil-based fuels. The resulting GHG emissions of each transport option depend on both the distance run by the specific means and the freight load delivered. As a consequence, the emission factor represents the total carbon dioxide emissions equivalent accordingly:

$$ETT_t = ETA_t + ETB_t + ETD_t + ETW_t + ETU_t + ETV_t, \quad (5)$$

where,

$$ETA_t = \sum_{i \in I} \sum_{g \in G} \sum_{f \in F} \sum_{l \in L} (EFTM_{il} ADG_{gfl} QI_{igfl}), \forall t \text{ is environmental}$$

impact of transportation biomass [$kg CO_2 - eq d^{-1}$];

$$ETE_t = \sum_{f \in F} \sum_{c \in C} \sum_{b \in B} (EFTB_{fb} ADF_{fcb} QB_{fcbt}), \forall t \text{ is environmental impact}$$

of transportation bioethanol from zones $f \in F$ to $c \in C$ where

$$QB_{fcbt} = \sum_{i \in I} \sum_{p \in P} QBP_{ifcbpt} \text{ [kg CO}_2 - eq d^{-1} \text{];}$$

$$ETD_t = \sum_{d \in D} \sum_{c \in C} \sum_{b \in B} (EFTB_{db} ADD_{dcb} QD_{dcbt}), \forall t \text{ is environmental}$$

impact of transportation gasoline from zones $d \in D$ to $c \in C$;

$$ETW_t = \sum_{f \in F} \sum_{w \in W} \sum_{m \in M} (EFTRW_{fm} ADW_{fwm} QW_{fwm}), \forall t \text{ is environmental}$$

impact of transportation solid wastes from zones $f \in F$ to $w \in W$;

$$ETU_t = \sum_{g \in G} \sum_{u \in U} \sum_{e \in E} (EFTRU_{ge} ADU_{gue} QU_{guet}), \forall t \text{ is environmental}$$

impact of transportation straw from zones $g \in G$ to $u \in U$;

$$ETV_t = \sum_{g \in G} \sum_{v \in V} \sum_{z \in Z} (EFTRV_{gz} ADV_{gvz} QV_{gvzt}), \forall t \text{ is environmental}$$

impact of transportation wheat-corn from zones $g \in G$ to $v \in V$;

5/ Total GHG emissions from utilization solid wastes ESW_t

$$ESW_t = \left(\sum_{i \in I} \sum_{f \in F} \sum_{c \in C} \sum_{b \in B} \sum_{p \in P} SW_{ip} QBP_{ifcbpt} - \sum_{f \in F} \sum_{w \in W} \sum_{m \in M} QW_{fwm} \right) ESW, \forall t, \quad (6)$$

6/ GHG emissions from bioethanol and gasoline usage in vehicle operations $ECAR_i$

$$ECAR_i = ECB \sum_{f \in F} \sum_{c \in C} \sum_{b \in B} QB_{fcbt} + ECG \sum_{d \in D} \sum_{c \in C} \sum_{b \in B} QD_{dcbt}, \quad \forall t, \quad (7)$$

4.1.2. Model of total cost of a IBSC

The annual operational cost includes the biomass feedstock acquisition cost, the local distribution cost of final fuel product, the production costs of final products, and the transportation costs of biomass, and final products. In the production cost, we consider both the fixed annual operating cost, which is given as a percentage of the corresponding total capital investment, and the net variable cost, which is proportional to the processing amount. In the transportation cost, both distance-fixed cost and distance-variable cost are considered. The economic criterion will be the cost of living expenses to include total investment cost of bioethanol production facilities and operation of the IBDS. This price is expressed through the dependence [8] for each time interval $t \in T$:

$$TDC_t = TIC_t + TPC_t + TTC_t + TTAXB_t - TL_t, \quad \forall t \quad (8)$$

where

TDC_t Total cost of a IBSC for year [\$ year⁻¹];

TIC_t Total investment costs of production capacity of IBSC relative to the operational period per year [\$ year⁻¹];

TPC_t Production cost for biorefineries [\$ year⁻¹];

TTC_t Total transportation cost of a IBSC [\$ year⁻¹];

$TTAXB_t$ A carbon tax levied according to the total amount of CO_2 generated in the work of IBSC [\$ year⁻¹];

TL_t Government incentives for bioethanol production and use;

1/ Model investment costs for biorefineries by year TIC_t

A rational IBSC planning over the time is based upon the assumption that once a production facility has been built, it will be operating for the remaining time frame.

$$TIC_t = \varepsilon_t \sum_{f \in F} \sum_{p \in P} (Cost_{pf}^F Z_{pft}), \quad \forall t \quad (9)$$

where ε_t is calculate by equation (10):

$$\varepsilon_t = \frac{1}{(1 + \zeta_t)} \quad (10)$$

Capital cost of biorefinery for each region is determined by the equation:

$$Cost_{pf}^F = M_f^{cost} Cost_p, \quad \forall p \in P, \forall f \in F, \quad (11)$$

2/ Total production cost model of IBSC TPC_t [\$ year⁻¹]

Total production cost term, TPC_t consists of biomass cultivation TPA_t , bioethanol production costs TPB_t and production cost for gasoline TPD_t as follows for each time interval t :

$$TPC_t = TPA_t + TPB_t + TPD_t, \quad \forall t, \quad (12)$$

where the components of (12) are defined according to the relations:

$$\left. \begin{aligned} TPA_t &= \sum_{i \in I} \sum_{g \in G} (UPC_{igt} A_{igt} \beta_{igt}) \\ TPB_t &= \sum_{i \in I} \sum_{f \in F} \sum_{c \in C} \sum_{b \in B} \sum_{p \in P} (\alpha_i UPB_{ipft} QBP_{ifcbpt}) \\ TPD_t &= \sum_{c \in C} \sum_{b \in B} \sum_{d \in D} (\alpha_i UPD_{dct} QD_{dcbt}) \end{aligned} \right\}, \quad \forall t$$

3/ Total transportation cost model TTC_t [\$ year⁻¹]

With regard to transports, both the biomass delivery to conversion plants and the fuel distribution and transport of diesel to blending terminals are treated as an additional service provided by existing actors already operating within the industrial/transport infrastructure. As a consequence, TTC_t is evaluated as follows:

$$TTC_t = TTCA_t + TTCB_t + TTCD_t, \quad \forall t \quad (13)$$

where, $TTCA_t = \sum_{i \in I} \sum_{l \in L} \sum_{f \in F} \sum_{g \in G} (\alpha_i UTC_{igfl} QI_{igflt})$, $\forall t$ is transportation

cost for energy crops, $TTCB_t = \sum_{b \in B} \sum_{c \in C} \sum_{f \in F} (\alpha_i UTB_{fcb} QB_{fcbt})$, $\forall t$, for

bioethanol, $TTCD_t = \sum_{b \in B} \sum_{c \in C} \sum_{d \in D} (\alpha_i UTD_{dcb} QD_{dcbt})$, $\forall t$ and for

gasoline, where,

$$\left. \begin{aligned} UTC_{igfl} &= IA_{il} + (IB_{il} ADG_{gfl}) \\ UTB_{fcb} &= OA_b + (OB_b ADF_{fcb}) \\ UTD_{dcb} &= OAD_b + (OBD_b ADD_{dcb}) \end{aligned} \right\},$$

IA_{il} and IB_{il} is fixed and variable cost for transportation biomass type $i \in I$ and (OA_b, OB_b) is fixed and variable cost for transportation bioethanol.

The biomass transportation cost UTC_{igfl} is described by Börjesson and Gustavsson [9]. for transportation by tractor, truck and train UTB_{fcb} . They are composed of a fixed cost (IA_{il}, OA_b) and a variable cost (IB_{il}, OB_b) . Fixed costs include loading and unloading costs. They do not depend on the distance of transport. Variable costs include fuel cost, driver cost, maintenance cost etc.

4/ Government incentives for bioethanol production cost model TL_t , [\$ year⁻¹]

$$TL_t = \sum_{f \in F} \sum_{c \in C} \sum_{b \in B} (INS_{ft} \alpha_i QB_{fcbt}), \quad \forall t \quad (14)$$

5/ A carbon tax levied cost model $TTAXB_t$, [\$ year⁻¹]

Many countries are implementing various mechanisms to reduce GHG emissions including incentives or mandatory targets to reduce carbon footprint. Carbon taxes and carbon markets (emissions trading) are recognized as the most cost-effective mechanisms. The basic idea is to put a price tag on carbon emissions and create new investment opportunities to generate a fund for green technology development. There are already a number of active carbon markets for GHG emissions [10].

$$TTAXB_t = (\alpha_i TEI_t) C_{CO_2}, \quad \forall t \quad (15)$$

4.2. Restrictions

Plants capacity limited by upper and lower constrains

Plants capacity is limited by upper and lower bounds, where the minimal production level in each region is obtained by:

$$\sum_{p \in P} (PB_p^{MIN} ZF_{pft}) \leq \alpha_i \sum_{c \in C} \sum_{b \in B} QB_{fcbt} \leq \sum_{p \in P} (PB_p^{MAX} ZF_{pft}), \quad \forall f, t \quad (16)$$

$$\left. \begin{aligned} \sum_{m \in M} \sum_{w \in W} QW_{fwmt} &\leq QW_{ft}^{MAX}, \quad \forall f \\ \sum_{e \in E} \sum_{u \in U} QU_{guet} &\leq QU_{gt}^{MAX}, \quad \forall g \\ \sum_{z \in Z} \sum_{v \in V} QV_{gvzt} &\leq QV_{gt}^{MAX}, \quad \forall g \end{aligned} \right\}, \quad \forall t \quad (17)$$

Constraints balance of bioethanol to be produced from biomass available in the regions

$$\alpha_i \sum_{i \in I} \sum_{f \in F} \sum_{c \in C} \sum_{b \in B} \sum_{p \in P} \left(\frac{QBP_{ifcbpt}}{\beta_{igt} \gamma_{ipt}} \right) = \sum_{i \in I} \sum_{g \in G} (A_{igt}), \quad \forall t \quad (18)$$

A condition that ensures that the total amount of solid waste generated by all bio-refineries can be processed in the plants built for this purpose

$$\sum_{w \in W} \sum_{m \in M} QW_{fwmt} \leq \sum_{p \in P} \sum_{i \in I} \sum_{c \in C} \sum_{b \in B} (SW_{ip} QBP_{ifcbpt}), \quad \forall f, t \quad (19)$$

A restriction that ensures that the amount of solid waste processed at the plant is within its production capacity

$$\left. \begin{aligned} \sum_{s \in S} P_s^{MIN} ZWF_{swt} &\leq \alpha_i \sum_{f \in F} \sum_{m \in M} QW_{fwmt} \\ \alpha_i \sum_{f \in F} \sum_{m \in M} QW_{fwmt} &\leq \sum_{s \in S} P_s^{MAX} ZWF_{swt} \end{aligned} \right\}, \quad \forall t, w \quad (20)$$

Logical Constrains

- Restriction guarantees that a given region $f \in F$ installed power plant with $p \in P$ for bioethanol production.

$$\left. \begin{aligned} \sum_{p \in P} Z_{pft} &\leq 1 \\ \sum_{p \in P} ZF_{pft} &\leq 1 \end{aligned} \right\}, \quad \forall f, t \quad (21)$$

and for a utilization systems of solid wastes (21):

$$\left. \begin{aligned} \sum_{s \in S} ZW_{swt} &\leq 1 \\ \sum_{s \in S} ZWF_{swt} &\leq 1 \end{aligned} \right\}, \quad \forall w, t \quad (22)$$

- Limitation ensure the availability of at least one connection to a region of bioresources and region for biofuel

$$\sum_{g \in G} \sum_{l \in L} X_{igflt} \geq \sum_{c \in C} \sum_{b \in B} Y_{fcbt} \geq \sum_{p \in P} ZF_{pft}, \quad \forall i, f, t \quad (23)$$

- Limit which guarantees that each region will provide only one plant with a biomass type $i \in I$

$$\sum_{f \in F} \sum_{l \in L} X_{igflt} \leq 1, \quad \forall i, g, t \quad (24)$$

- Limitation of assurance that at least one region $f \in F$ producing bioethanol is connected to a costumer zones $c \in C$

$$\sum_{b \in B} \sum_{f \in F} Y_{fcbt} \leq 1, \quad \forall c, t \quad (25)$$

- Limitation of assurance that at least one region f is connected to a solid waste utilization plant located in region $w \in W$

$$\sum_{w \in W} \sum_{m \in M} WS_{fwmt} \leq 1, \quad \forall f, t \quad (26)$$

- Condition ensuring that the solid waste produced from a given bio-refinery will be processed in only one of the plants for use

$$\sum_{m \in M} \sum_{w \in W} WS_{fwmt} = \sum_{p \in P} ZF_{pft}, \quad \forall f, t \quad (27)$$

- Condition ensuring that a plant used in a given region will be connected to at least one plant in which solid waste is generated

$$\sum_{m \in M} \sum_{f \in F} WS_{fwmt} \geq \sum_{s \in S} ZWF_{swt}, \quad \forall w, t \quad (28)$$

Transport Links

Restrictions on transportation of biomass are

$$PBI_{ig}^{MIN} \sum_{l \in L} X_{igflt} \leq \alpha_i \sum_{l \in L} QI_{igflt} \leq PBI_{ig}^{MAX} \sum_{l \in L} X_{igflt}, \quad \forall i, g, f, t \quad (29)$$

Mass balances between bioethanol plants and biomass regions

The connections between bioethanol plants and biomass regions:

$$\sum_{l \in L} \sum_{g \in G} \sum_{i \in I} (QI_{igflt}) \leq \sum_{p \in P} \left(\frac{PB_p^{MAX} ZF_{pft}}{\gamma_{ipt}} \right), \quad \forall f, t \quad (30)$$

Mass balances between bioethanol plants and customer zones

$$\sum_{b \in B} \sum_{f \in F} (\alpha_i QB_{fcbt}) = QEB_{ct}, \quad \forall c, t \quad (31)$$

Energy Restriction

- limitation ensuring that the overall energy balance in the region is provided

$$ENO \sum_{c \in C} QEO_{ct} + ENB \sum_{c \in C} QEB_{ct} = ENO \sum_{c \in C} YO_{ct}, \quad \forall t \quad (32)$$

- limitation ensuring that each region will be provided in the desired proportions with fuels

$$ENB QEB_{ct} = K_{ct}^{mix} ENO YO_{ct}, \quad \forall c, t \quad (33)$$

4.3. Economic objective function

Objective function associated with the minimization of the economic costs includes all the operating costs of the supply chain, from the purchase of biomass feedstock to transportation of the final product, as well as the investment cost of biorefineries [11]. The costs of the supply chain are: the cost of raw material, the transport of raw material to the facilities, the cost of transport to the biorefineries, the cost of transformation into bioethanol and the cost of final transport to the blending facilities. The economic objective is to minimize the total annual costs. The terms of the cost objective

corresponding to the annual operation costs of the IBSC are described in the following equation:

$$COST = \sum_{i \in I} (LT_i TDC_i) \quad (34)$$

5. Optimization problem formulation

The problem for the optimal design of a IBSC is formulated as a MILP model for the objective function of Minimizing cost.

The task of determining the optimal location of facilities in the regions and their parameters is formulated as follows:

$$\left\{ \begin{aligned} &Find : X_i [Decision\ variables]^T \\ &MINIMIZE \{COST\} \rightarrow (Eq.34) \\ &s.t. : \{Eq.16 - Eq.33\} \end{aligned} \right\} \quad (35)$$

The problem is an ordinary MILP and can thus be solved using MILP techniques. The present model was developed in the commercial software GAMS [12]. The model chooses the less costly pathways from one set of biomass supply points to a specific plant and further to a set of biofuel demand points. The final result of the optimisation problem would then be a set of plants together with their corresponding biomass and biofuel demand points.

6. Discussion and conclusion

This paper studies the interactions among biofuel supply chain design, agricultural land use and local food market equilibrium. The study has been focused on the eco comparable behavior of the stakeholders in the biofuel supply chain incorporating them into the supply chain design model. The model includes the problem of crop rotation and solid waste utilization. The model is believed to be important for practical application and can be used for design and management of similar supply chains.

Acknowledgements

The authors would like to thank Bulgarian National Science Fund for the financial support obtained under contract DN 07-14/15.12.2016.

7. References

- [1]. IEA, *World energy outlook 2007*, Paris, France: International Energy Agency, 2007.
- [2]. European Commission, Directive 2009/28/EC of the European Parliament and of the Council of 23 April 2009 on the Promotion of the use of energy from renewable sources and Amending and Subsequently Repealing Directives 2001/77/EC and 2003/30/EC. Official J Eur Parliam, Brussels.
- [3]. O. Kitani, Carl HW. The state of food and agriculture. New York: Food and Agriculture Organization, Rome, Italy: FAO, 2008, vol. 5.
- [4]. Carlos Mireta, Philippe Chazara, Ludovic Montastruc, Stéphane Negny, Serge Domenech. Design of bioethanol green supply chain: Comparison between first and second generation biomass concerning economic, environmental and social criteria. *Computers and Chemical Engineering* 85 (2016) 16–35.
- [5]. S. Banerjee, S. Mudliar, R. Sen, L. Giri, D. Satpute, T. Chakrabarti, R.A. Pandey. Commercializing lignocellulosic bioethanol: technology bottlenecks and possible remedies, *Biofuels, Bioproducts and Biorefining* 4, pp 77–93, 2010.
- [6]. O. Akgul, A. Zamboni, F. Bezzo, N. Shah, L. Papageorgiou. Optimization-Based Approaches for Bioethanol Supply Chains, *Ind. Eng. Chem. Res.*, 50, 4927–4938, 2011.
- [7]. Zamboni A., Bezzo F., Shah N. Spatially explicit static model for the strategic design of future bioethanol production systems, 2. Multi-objective environmental optimization. *Energy and Fuels*, 2009, 23, 5134–5143.
- [8]. Ozlem A., Shah N., Papageorgiou L., Economic optimisation of a UK advanced biofuel supply chain, *Biomass and Bioenergy* 2012, 41, 57–72.
- [9]. Börjesson P., Gustavsson L., (1996). Regional Production and Utilization of Biomass in Sweden. *Energy*, 21, 747–764.
- [10]. Johnson E., Heinen R. Carbon trading: time for industry involvement. *Environment International*, 2004, 30, 279–288.
- [11]. Atif Osmania, Jun Zhang. Multi-period stochastic optimization of a sustainable multi-feedstock second generation bioethanol supply chain – A logistic case study in Midwestern United States, *Land Use Policy* 61 (2017) 420–450
- [12]. B. McCarl, A. Meeraus, Pvd. Eijk, M. Bussieck, S. Dirkse, P. Steacy, McCarl Expanded GAMS user Guide, Version 22.9. GAMS Development Corporation, 2008.

STABILITY ANALYSIS AND SIMULATIONS OF BIOREACTOR MODEL WITH DELAYED FEEDBACK

Assist. Prof., Ph.D. Borisov M.¹, Prof., Ph.D. Dimitrova N.¹, Prof., D.Sc. Krastanov M.^{1,2}

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences¹

Faculty of Mathematics and Informatics, Sofia University, Bulgaria²

E-mail: milen_kb@math.bas.bg, nelid@math.bas.bg, krastanov@fmi.uni-sofia.bg

Abstract: We consider a well known mathematical model of continuous methane fermentation, consisting of two nonlinear ordinary differential equations and one algebraic equation for the gaseous output. The model involves one microbial population and one substrate. We propose an output feedback including a discrete delay and use it for asymptotic stabilization of the model. Feedback control of bioreactor models provides many advantages in operating a plant, mainly by increasing its efficiency. We also propose a numerical model-based extremum seeking algorithm for maximizing the biogas (methane) flow rate in real time. Numerical simulations using this algorithm are included. The simulations are implemented in the Python programming language which is recently recognized as a powerful modern general purpose object-oriented language.

Keywords: BIOREACTOR MODEL, DELAY FEEDBACK CONTROL, EXTREMUM SEEKING ALGORITHM

1. Introduction

We consider a well known mathematical model of the continuous methane fermentation, consisting of two nonlinear ordinary differential equations and one algebraic equation

$$\begin{aligned} \frac{ds}{dt} &= -k_1\mu(s)x + u(s_{in} - s) \\ \frac{dx}{dt} &= (\mu(s) - \alpha u)x \end{aligned} \quad (1)$$

with gaseous output

$$Q(s, x) = k_2 \mu(s)x, \quad (2)$$

where $x = x(t)$ and $s = s(t)$ are state variables, x is biomass concentration [g/dm³], s – substrate concentration [g/dm³], u – dilution rate [day⁻¹], s_{in} – influent substrate concentration [g/dm³], k_1 – yield coefficient [-], k_2 – coefficient [(dm³)²/g], Q – methane gas flow rate [dm³/day]. The parameter $\alpha \in (0, 1)$ represents the proportion of bacteria that are affected by the dilution; $\alpha = 0$ and $\alpha = 1$ correspond to an ideal fixed bed reactor and to an ideal continuous stirred tank reactor, respectively. The dilution rate u is considered as a control variable. The model function $\mu(s)$ represents the specific growth rate of the biomass.

Assumption A1. We assume that μ is defined for $s \in [0, +\infty)$, $\mu(0) = 0$ and $\mu(s) > 0$ whenever $s > 0$, $\mu(s)$ is continuously differentiable for all $s > 0$.

We propose here an output feedback including a discrete delay and use it for asymptotic stabilization of the dynamic model (1). Feedback control of bioreactor models provides many advantages in operating a plant and is used to increase its efficiency. On the other hand, there is always a time delay between input and output measurements in industrial and biological systems [4].

We make the following assumption.

Assumption A2. Assume that lower bounds s_{in}^- and k_2^- for the values of s_{in} and k_2 respectively, and an upper bound k_1^+ for the value of k_1 are known.

Denote for simplicity

$$\beta^- = \frac{k_1^+}{k_2^- s_{in}^-}.$$

Define the following feedback control law:

$$k(s, x) = \beta k_2 \mu(s)x \text{ with } \beta \in (\beta^-, +\infty).$$

Replacing in the model (1)–(2) u by the feedback

$$k(s(t - \tau), x(t - \tau)),$$

where $\tau > 0$ is a discrete delay, we obtain the following control system:

$$\frac{ds}{dt} = -k_1\mu(s(t))x(t) + \beta k_2\mu(s(t - \tau))x(t - \tau)(s_{in} - s(t)) \quad (3)$$

$$\frac{dx}{dt} = \mu(s(t))x(t) - \alpha\beta k_2\mu(s(t - \tau))x(t - \tau) \quad (4)$$

Choose some $\beta \in (\beta^-, +\infty)$ and let

$$\bar{s} := s_{in} - \frac{k_1}{k_2\beta} \quad (5)$$

Obviously, \bar{s} belongs to the interval $(0, s_{in})$. Define

$$\bar{x} := \frac{1}{\alpha\beta k_2} \quad (6)$$

It is straightforward to see that the point

$$\bar{p}_\beta := (\bar{s}, \bar{x})$$

is an equilibrium point for (3)–(4).

2. Stability analysis of the model

Assumption A3. We assume that the following inequalities hold true

$$\mu(s^-) < \mu(\bar{s}) < \mu(s^+)$$

for each $s^- \in (0, \bar{s})$ and $s^+ \in (\bar{s}, s_{in})$.

Assumption A3 is technical. It is always fulfilled when the function $\mu(s)$ is monotone increasing (like the Monod specific growth rate, see Section 4). If the function $\mu(s)$ is not monotone increasing (like e. g. the Haldane law) then the point \bar{s} (i.e. β) has to be chosen in a proper way in order to satisfy Assumption A3.

Below we present a result about the local asymptotic stability of the equilibrium point \bar{p}_β with respect to the parameters of the closed-loop system (3)–(4). Denote for simplicity:

$$\begin{aligned} b &= k_1 \bar{x} \mu(\bar{s}) \mu'(\bar{s}) \\ d &= \mu(\bar{s}) \left(\frac{1}{\alpha} \mu(\bar{s}) - k_1 \bar{x} \mu(\bar{s}) \right) \end{aligned}$$

where the coefficients b and d depend on the parameter β , since \bar{s} and \bar{x} are functions of β , see (5) and (6).

Theorem 1.

(i) If $b \geq d$ then the equilibrium point \bar{p}_β is locally asymptotically stable for any value of the delay $\tau > 0$.

(ii) If $b < d$ then there exists a delay $\tau_0 > 0$ such that the equilibrium point \bar{p}_β is locally asymptotically stable for all values τ of the delay such that $\tau < \tau_0$; the equilibrium is locally unstable if $\tau \geq \tau_0$, and a Hopf bifurcation occurs at $\tau = \tau_0$.

The proof of the theorem is rather long and will not be presented here due to the restricted paper length. The proof is based on some ideas from [3], [5] and [6].

We suppose that in the cases when the equilibrium point \bar{p}_β is locally asymptotically stable then it is also globally asymptotically stable. The computer simulations confirm that assumption (see Section 4). Proving global stability of \bar{p}_β will be a subject of further studies.

3. Maximizing the biogas production via extremum seeking

Consider the equation (2) describing the process output, i.e. the methane (biogas) production. In this section we shall shortly present a numerical extremum seeking algorithm (ESA), cf. [1]–[2], to steer and stabilize the dynamics (3)–(4) towards a steady state, where maximum methane flow rate Q_{\max} is achieved. For that purpose we compute Q on the set of all equilibrium points, parameterized with respect to β . Denote the so obtained function by $Q(\beta)$, where $\beta \in (\beta^-, +\infty)$. The function $Q(\beta)$ is called input-output static characteristic of the model. Assume that the function $\beta \rightarrow Q(\beta)$, $\beta \in (\beta^-, +\infty)$, is strongly unimodal, i.e. there exists a unique point $\beta_{\max} \in (\beta^-, +\infty)$ where $Q(\beta)$ takes a maximum, $Q_{\max} = Q(\beta_{\max})$, the function strongly increases in the interval (β^-, β_{\max}) and strongly decreases in $(\beta_{\max}, +\infty)$.

Let

$$p_{\beta_{\max}} = (\bar{s}_{\max}, \bar{x}_{\max})$$

be the steady state where Q_{\max} is achieved. Our goal is to stabilize in real time the system (3)–(4) towards this (unknown) equilibrium point $p_{\beta_{\max}}$ and therefore to the maximum methane flow rate Q_{\max} .

The main idea of ESA is the following: we construct a sequence of points $\beta^1, \beta^2, \dots, \beta^n, \dots$ from the interval $(\beta^-, +\infty)$, each β^j being in the form $\beta^j = \beta^{j-1} \pm h$, $h > 0$, and such that the sequence $\{\beta^j\}$ tends to β_{\max} . Then by computing and comparing the values $Q(\beta^1), Q(\beta^2), \dots, Q(\beta^n), \dots$, we achieve the desired equilibrium point $p_{\beta_{\max}}$ and thus Q_{\max} .

In the computer implementation ESA is carried out in two stages. In the first stage, “rough” intervals $[\beta]$ and $[Q]$ are found which enclose β_{\max} and Q_{\max} respectively; in the second stage, the interval $[\beta]$ is refined using an elimination procedure based on the golden mean value (or Fibonacci search) strategy. The second stage produces the final intervals $[\beta_{\max}] = [\beta_{\max}^-, \beta_{\max}^+]$ and $[Q_{\max}]$ such that $\beta_{\max} \in [\beta_{\max}^-]$, $Q_{\max} \in [Q_{\max}^-]$ and $\beta_{\max}^+ - \beta_{\max}^- \leq \varepsilon$, where the tolerance $\varepsilon > 0$ is specified by the user.

The ESA was firstly developed for the model (3)–(4) with $\tau = 0$, cf. [1], [2]. Now the algorithm is modified for the delayed model and is implemented in the object-oriented language *Python*. Nowadays, this programming language has become an integral part of scientific and engineering computing due to the vast abundance of open source libraries and interfaces to major software tools.

4. Numerical Simulation

In the simulation process we consider the Monod specific growth rate

$$\mu(s) = \frac{m_1 s}{k_s + s},$$

where m_1 is the maximum specific growth rate of the microorganisms [1/day] and k_s is the saturation constant [g/dm³].

We use the following numerical values for the model parameters (cf. [1] and the references therein):

$$k_1 = 27.4, s_{in} = 3, m_1 = 0.4, k_s = 0.4, \alpha = 0.5, k_2 = 75.$$

With $k_1^+ = 27.6$, $s_{in}^- = 2.8$ and $k_2^- = 74.8$ we obtain that $\beta^- = \frac{k_1^+}{s_{in}^- k_2^-} \approx 0.1318$. Therefore the parameter β varies in the interval $(0.1318, +\infty)$.

For the above parameter values, the input-output static characteristic $Q(\beta)$ is strongly unimodal and takes its maximum for $\beta = \beta_{\max} = 0.16355$ (see Figure 1), thus $Q_{\max} = 3.21376$, $\bar{s}_{\max} = 0.76619$ and $\bar{x}_{\max} = 0.16305$.

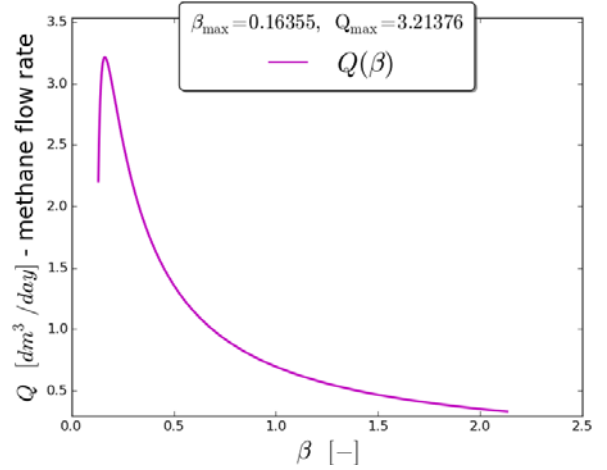


Fig. 1. The input-output static characteristic $Q(\beta)$

4.1. Simulation of the global dynamics behavior

First we shall demonstrate numerically the stability of the system (3)–(4) towards the equilibrium point \bar{p}_β for different values of the delay τ and the parameter β .

(i) $\beta = \beta_{\max} = 0.16355$

For this value of β , the inequality $b > d$ is fulfilled. According to Theorem 1, the equilibrium point $(\bar{s}_{\max}, \bar{x}_{\max}) = (0.76619, 0.16305)$ is stable for any $\tau \geq 0$ as it can be seen in the Figures 2 to 4 for the following values of the delay: $\tau = 1$, $\tau = 5$ and $\tau = 15$.

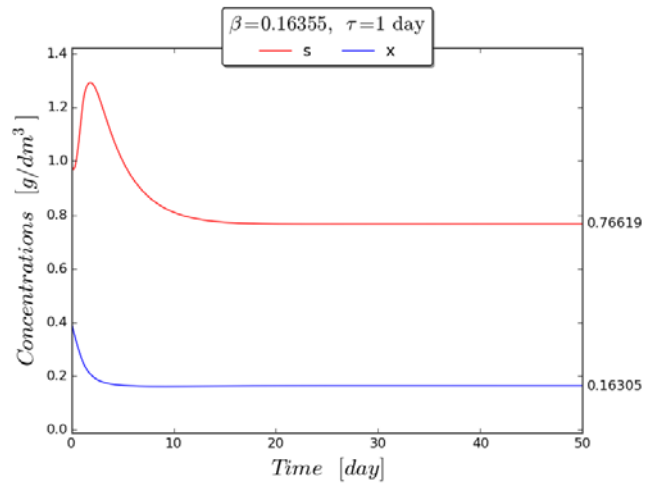


Fig. 2. $\beta = \beta_{\max} = 0.16355$, $\tau = 1$: Time evolution of s and x towards \bar{s}_{\max} and \bar{x}_{\max} respectively

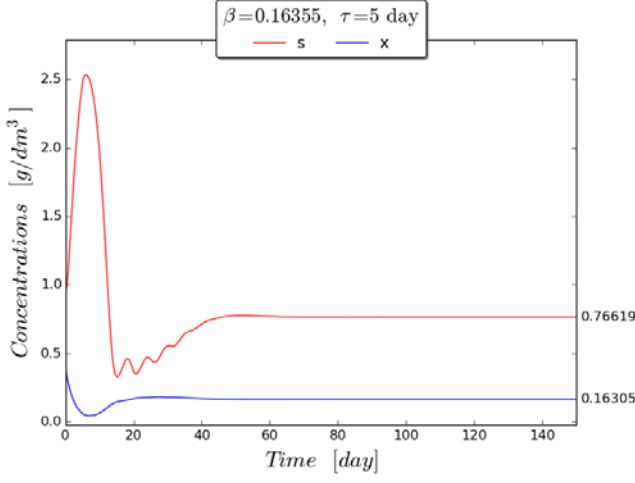


Fig. 3. $\beta = \beta_{\max} = 0.16355$, $\tau = 5$:
Time evolution of s and x towards \bar{s}_{\max} and \bar{x}_{\max} respectively

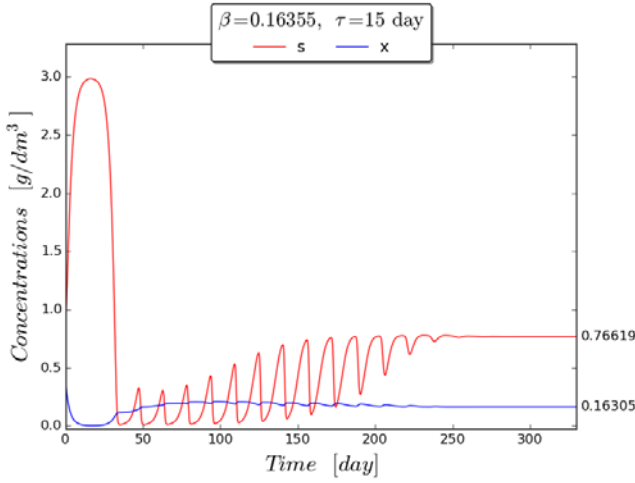


Fig. 4. $\beta = \beta_{\max} = 0.16355$, $\tau = 15$:
Time evolution of s and x towards \bar{s}_{\max} and \bar{x}_{\max} respectively

(ii) $\beta = 1.0$

Here $b < d$ is valid. In this case $\tau_0 = 4.70326$, $\bar{s} = 2.63467$ and $\bar{x} = 0.02667$. According to Theorem 1, the equilibrium point (\bar{s}, \bar{x}) is stable for any $\tau < \tau_0$ and unstable for any $\tau \geq \tau_0$; this is demonstrated in Figures 5 to 8 when the delay takes the values $\tau = 1$, $\tau = 4.5$, $\tau = 5$ respectively.

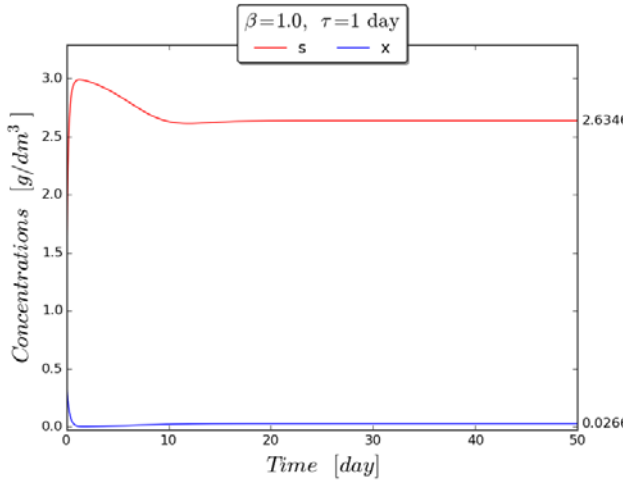


Fig. 5. $\beta = 1$, $\tau = 1$: Time evolution of s and x towards \bar{s} and \bar{x}

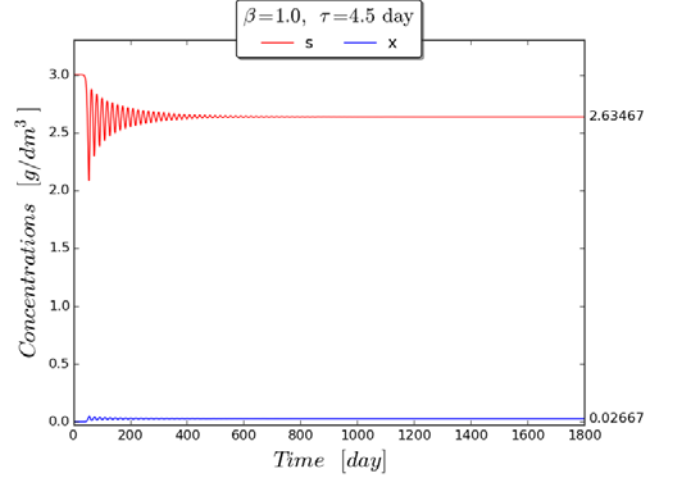


Fig. 6. $\beta = 1$, $\tau = 4.5$: Time evolution of s and x towards \bar{s} and \bar{x}

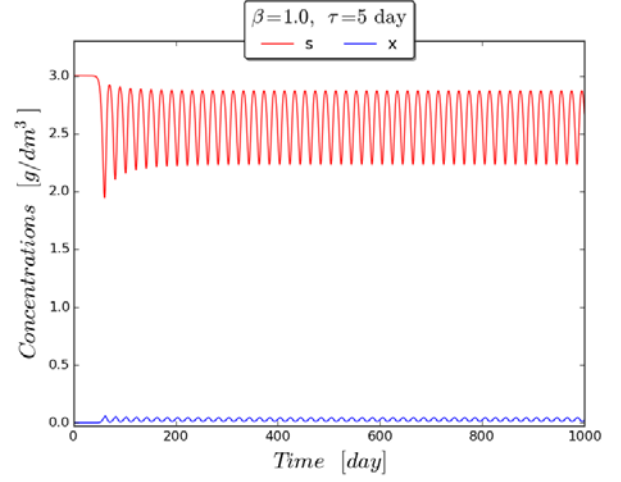


Fig. 7. $\beta = 1$, $\tau = 5$: Time evolution of s and x towards \bar{s} and \bar{x}

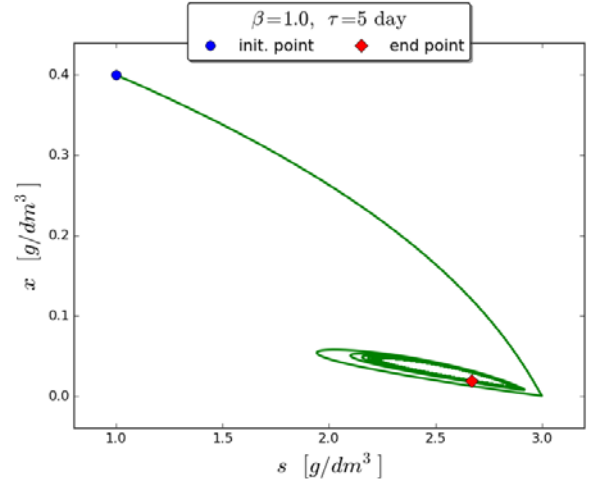


Fig. 8. $\beta = 1$, $\tau = 5$: A trajectory in the phase plane (s, x)
Existence of a periodic solution as a result of Hopf bifurcation

The graphic visualizations (see Figures 2 and 5), as well as other simulations, that are not presented in the paper, show that for a delay $\tau = 1$ [day] the dynamics is stabilized approximately within 10 days, which is practically reasonable for the modeled process. Increasing the delay leads to the appearance of damped oscillations and the time for the solution stabilization increases (see Figures 3 and 6 for $\tau = 5$ and $\tau = 4.5$ respectively). For extremely large (and practically unusual) value of the delay ($\tau = 15$ [days] in Figure 4), as well as for $\tau \geq \tau_0$ (e.g. $\tau = 5$ [days] in Figures 7 and 8) sustainable oscillations do occur.

4.2. Simulation results from the extremum seeking algorithm

The numerical results from the extremum seeking algorithm (ESA) are visualized in Figures 9 to 12 for small values of the delay $\tau = 1$ and $\tau = 3$.

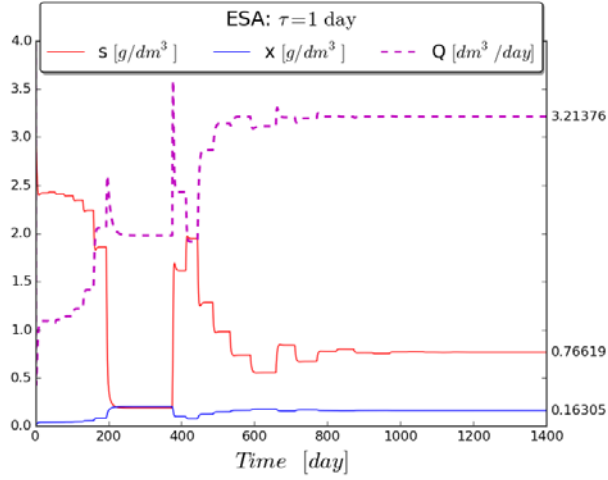


Fig. 9. $\tau = 1$. Visualization of the numerical results from the ESA. Time evolution of s, x and Q towards \bar{s}_{max} , \bar{x}_{max} and Q_{max}

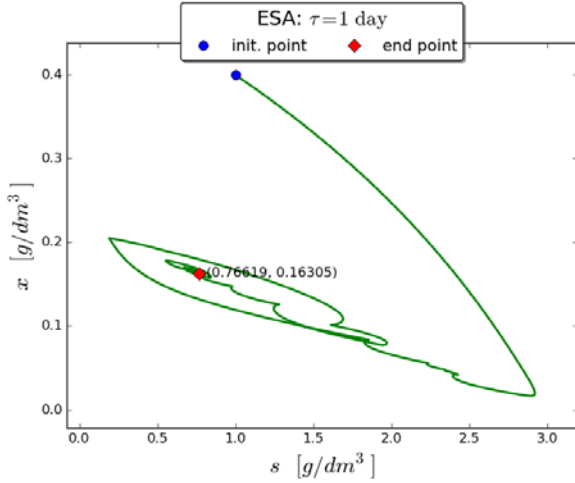


Fig. 10. $\tau = 1$. Visualization of the numerical results from the ESA. A trajectory in the phase plane (s, x) .

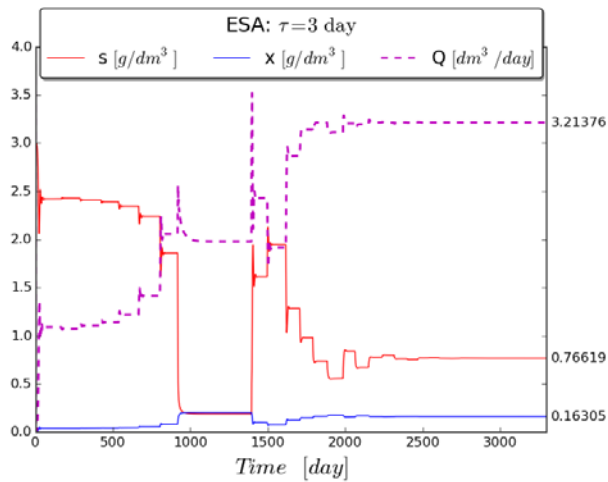


Fig. 11. $\tau = 3$. Visualization of the numerical results from the ESA. Time evolution of s, x and Q towards \bar{s}_{max} , \bar{x}_{max} and Q_{max}

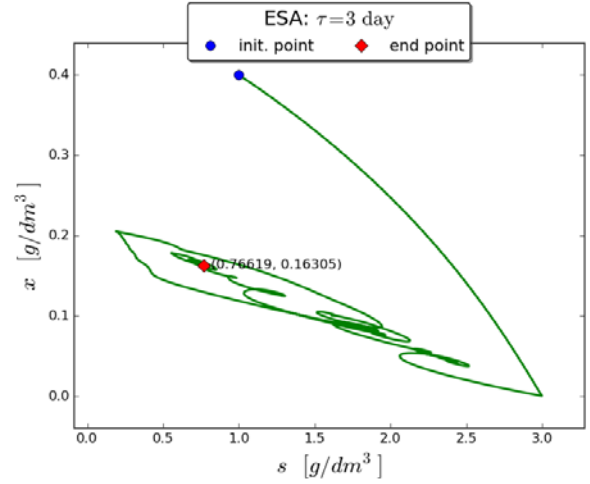


Fig. 12. $\tau = 3$. Visualization of the numerical results from the ESA. A trajectory in the phase plane (s, x) .

5. Conclusion

In this paper we investigate a nonlinear functional-differential model of an anaerobic bioreactor with methane (biogas) production, involving one microbial population and one substrate. The stabilization of the model solutions is carried out by a feedback control law involving one discrete time delay. We have established local asymptotic stability of the system dynamics towards a previously chosen equilibrium point as well as bifurcations of this equilibrium with respect to the time delay. The numerical simulations suggest that in the cases when the equilibrium point is locally asymptotically stable it is also globally stable. This allows us to apply a numerical model-based extremum seeking algorithm for stabilizing the dynamics towards the equilibrium point where maximum production of methane is achieved. The facilities of the algorithm are illustrated numerically as well.

Acknowledgements

The research of the first and the second author has been supported by the Bulgarian Academy of Sciences, Program for Support of Young Scientists and Scholars, grant № DFNP-17-25/25.07.2017. The work of the second and the third author has been partially supported by the Sofia University "St. K. Ohridski" under grant № 80-10-220/22.04.2017.

References

- [1] N. Dimitrova, M. Krastanov: Nonlinear Adaptive Control of a Model of an Uncertain Fermentation Process. *Int. Journ. Robust Nonlinear Control*, 20, 1001–1009, 2010.
- [2] N. S. Dimitrova, M. I. Krastanov: Adaptive Asymptotic Stabilization of a Bioprocess Model with Unknown Kinetics. *Int. J. Numerical Analysis and Modeling, Series B, Computing and Information*, vol. 2, No. 2–3, 200–214, 2011.
- [3] Y. Kuang: *Delay Differential Equations with Applications in Population Dynamics*. Academic press, 1993.
- [4] G. Robledo: Feedback Stabilization for a Chemostat with Delayed Output. Preprint INRIA Sophia Antipolis RR No 5844, 2004.
- [5] S. Ruan: On Nonlinear Dynamics of Predator-Prey Models with Discrete Delay. *Math. Model. Nat. Phenom.* 4 (2), 140–188, 2009.
- [6] H. Smith: *An Introduction to Delay Differential Equations with Applications to the Life Sciences*. Springer, 2011.

MODELING AND SIMULATION OF INDUSTRIAL PROCESSES

Christo Boyadjiev

Institute of Chemical Engineering, Bulgarian Academy of Sciences
chr.boyadjiev@gmail.com

Abstract: In the paper are presented the main theoretical techniques used for the modeling and simulation of industrial processes. The main focus is on the physical side of the theoretical techniques and their mathematical side is reduced to a reasonable minimum. Different theoretical approximations as thermodynamic and hydrodynamic levels are used.

KEYWORDS: MODELING, SIMULATION, THEORETICAL APPROXIMATION, THERMODYNAMIC LEVEL, HYDRODYNAMIC LEVEL, MECHANISM IDENTIFICATION, PARAMETERS IDENTIFICATION, STATISTICAL ANALYSIS.

1. Introduction

The modeling and simulation are a basic approach to quantifying processes and phenomena [1-3]. They have become realistic as a result of the development of computing and applied mathematics. In the industry, the modeling and simulation offer a quantitative description of the kinetics of processes and systems for the purposes of their optimal design or control. In the industrial systems, the process models of the individual devices are known in advance, and system models offer quantitative descriptions of the systems (process systems engineering).

In the industry, quantification of systems can also be used for various other tasks. For example, in periodically operating systems, the optimal schedules of the apparatuses (machines) for conducting different processes (operations) of different duration and different sequence to obtain different substances (machined details) can be determined. In these cases, the mass service theory offers models that allow for optimal solutions.

The fundamentals of the modeling and simulation, as a part of human knowledge and science, are related to the combination of intuition and logic. They are in different scales in the individual sciences [4, 5]. In the mathematics, logic dominates intuition, where intuition is the axiom (unconditional truth that is not proof able), and logic is the theorem (the logical consequences of the axiom). In the natural sciences (physics, chemistry, biology), the logic/intuition ratio is maintained, but axioms are usually conditional (principles, postulates, laws). This ratio goes back to the humanities and reaches the extreme in religion.

The modeling and simulation offer quantitative (mathematical) descriptions that have different degrees of detail. The lowest level is the thermodynamic (non-equilibrium thermodynamics) that examines the volume of the phase (gas, liquid, solid). The next level is the hydrodynamic, which examines the elementary phase volumes (mechanics of continua), which are much smaller than the phase volumes, but much larger than the intermolecular volumes, i.e. the molecules are indistinguishable. The highest level is the molecular (the kinetic theory of the ideal gas).

The modeling and simulation of industrial processes has a wide application, so the processes in the chemical industry and related biotechnologies and heating technologies, will be discussed. The major part of these processes is the transfer of mass and heat as a result of phase or phase boundary reactions. By reaction, will be understand the creation or disappearance of a particular substance (or amount of heat) as a result of a chemical reaction in the phase or on the phase boundary, interphase mass transfer, adsorption on the phase boundary or a liquid-vapor-liquid phase transition. These reactions result in varying concentrations and temperatures in the phases, i. e. to a deviation from the thermodynamic equilibrium and as a result of the mass transfer and heat transfer to restore the thermodynamic equilibrium. The models of the mass transfer and heat transfer are analogous and, therefore, the models of mass transfer in industrial processes will be presented.

2. Thermodynamic approximation

The reactions deviate the industrial systems from the thermodynamic equilibrium and the industrial processes for its recovery begin. The determining of the rate of these processes is a major problem in the industry, as it is the basis for their optimal design and control. This gives reason to use the thermodynamic laws of irreversible processes such as mathematical structures in the construction of the process models, described by extensive and

intense variables (in the case of merging of two identical systems, the extensive variables double their values, while the intensive variables retain their values).

The kinetics of the irreversible processes uses the mathematical structures, resulting from Onsanger's "linearity principle" [6]. According to him, the mean values of the derivatives at the time of the extensive variables depend linearly on the mean deviations of the conjugated intensive variables from their equilibrium states (values). This principle is valid in the vicinity of the equilibrium, and proportionality coefficients are kinetic (rate) constants.

According to the principle of linearity, the mass derivative at time

$J_0 = \frac{dm}{dt}$ [kg-mol.s⁻¹] depends linearly on the deviation from the

thermodynamic equilibrium Δc [kg-mol.m⁻³] of the concentration in two phase volumes or in one phase and the phase boundary, i.e.

$$J_0 = k_0 \Delta c, \quad (1)$$

where k_0 [m³.s⁻¹] is a proportionality coefficient.

Consider a system that contains two identical volumes in one phase $V_1 = V_2 = V$ [m³]. The system contains a substance whose masses

m_i [kg-mol] and concentrations $c_i = \frac{m_i}{V_i}$ [kg-mol.m⁻³] are

different in two volumes, $i = 1, 2, \dots$. The system is not in thermodynamic equilibrium. Let us assume for certainty $c_1 - c_2 > 0$, $i = 1, 2$. As a result, the mass of the substance starts to be transferred from volume V_1 to volume V_2 for to achieve the equilibrium. According to the principle of linearity, the mass transfer rate between the two volumes J_0 [kg-mol.s⁻¹] can be represented as:

$$J_0 = \frac{dm_1}{dt} = -\frac{dm_2}{dt} = k_0 (c_1 - c_2), \quad (2)$$

where k_0 [kg-mol⁻¹.m³.s⁻¹] is a proportionality coefficient. If we replace masses with concentrations $m_i = V_i c_i$, $i = 1, 2$, the mass transfer rate in one phase J [kg-mol.m⁻³.s⁻¹] between two points with different concentrations is:

$$J = \frac{dc_1}{dt} = -\frac{dc_2}{dt} = k (c_1 - c_2), \quad (3)$$

where k [s⁻¹] is a rate coefficient. This equation is capable of presenting the rate of interphase mass transfer in the case of adsorption or catalytic process, where C_1 is the concentration of the substance in the gas phase, while C_2 is the concentration of the substance in the gaseous portion of the solid (capillaries of the adsorbent or catalyst) phase.

In the cases, where the volumes $V_1 = V_2 = V$ are in different phases (for example, 1 is a gas phase and 2 is a liquid phase), the

thermodynamic equilibrium law has the form $c_1 - \chi c_2 = 0$, i.e. this is the Henry's law and χ is the Henry's number. If $c_1 - \chi c_2 > 0$ the mass transfer is from phase 1 to phase 2 and the mass transfer rate between phases is:

$$J = k(c_1 - \chi c_2), \quad (4)$$

where k [s^{-1}] is the rate coefficient of the interphase mass transfer. On the surface between two phases, the thermodynamic equilibrium is immediately established, practically, i.e. $c_1^* - \chi c_2^* = 0$, where $c_i^*, i=1,2$, are the equilibrium concentrations on the phase boundary. Thus, the mass transfer rate can be expressed by mass transfer rate in two phases:

$$J = k_1(c_1 - c_1^*) = k_2(c_2^* - c_2), \quad (5)$$

where $k_i, i=1,2$ [s^{-1}] are mass transfer rate coefficients.

The Onsanger principle of linearity represents the thermodynamic approximation of the mathematical description of the kinetics of irreversible processes, but it does not show the way to reach equilibrium, i.e. the mechanism of the process and as a result the rate coefficient is not known. Obviously, this "thermodynamic level" does not allow a real quantitative description of the kinetics of irreversible processes in industry and the next level of detail of the description, the so-called "hydrodynamic level", should be used.

3. Hydrodynamic approximation

The processes in the chemical industry and related biotechnologies and heating technologies are realized in one-, two- and three-phase systems (gas-liquid-solid). They are a result from the reactions, i.e. processes of disappearance or creation of any substance. The reactions are associated with a particular phase and can be homogeneous (occurring in volume of the phase) or heterogeneous (occurring at the interface with another phase). Homogeneous reactions are usually chemical, while heterogeneous reactions may be chemical, catalytic and adsorption. Heterogeneous reaction is the interphase mass transfer too, where on the interphase boundary the substance disappears (created) in one phase and creates (disappears) in the other phase.

The volume reactions lead to different concentrations of the reagents in the phase volumes and as a result two mass transfer processes are realized – convective transfer (caused by the movement of the phases) and diffusion transfer (caused by the concentration gradients in the phases). The mass transfer models are a mass balance in the phases, where components are convective transfer, diffusion transfer and volume reactions (volume mass sources or sinks). The surface reactions participate as mass sources or sinks in the boundary conditions of the model equations. The models of this complex process are possible to be created on the basis of the mass transfer theory, whose models are created by the models of the hydrodynamics, diffusion and reaction kinetics. The mass transfer theory combines the chemistry, physics and mathematics and builds its logical structures on three main "axioms":

1. The postulate of Stokes for the linear relationship between the stress and deformation rate, which is the basis of the Newtonian fluid dynamics models;
2. The first law of Fick for the linear relationship between the mass flow and the concentration gradient, which is the basis of the linear theory of the mass transfer;
3. The first law of Fourier for the linear relationship between the heat flux and the temperature gradient, which is the basis of the linear theories of the heat transfer.

These are the laws of the impulse, mass and energy transfer.

In Boltzmann's kinetic theory of the ideal gas, these axioms are replaced by the "elastic shock" axiom (in a shock between two molecules the direction and the velocity of the movement change, but the sum of their kinetic energies is retained, i.e. there is no loss of kinetic energy) and the rate coefficients are theoretically

determined by the average velocity and the average free run of the molecules.

The contemporary mass transfer theory is based of diffusion boundary layer theory (Landau, Levich [7]). This approach substitutes (physically justified) elliptic partial differential equations with parabolic partial differential equations, which facilitates their mathematical solution and offers a mathematical description of physical processes with free (not predetermined) ends.

The diffusion boundary layer theory is developed in the cases of drops and bubbles (Levich, Krylov [8]), film flows (Levich, Krylov, Boyadjiev, Beshkov[9, 10]), non-linear mass transfer and hydrodynamic stability (Krylov, Boyadjiev, Babak [11, 12]).

3.1. Mass transfer theory

The complex industrial processes are a collection of elementary physical and chemical processes. For example, the chemical absorption in a packed bed column represents a physical absorption of a gas phase component in the liquid phase and a subsequent chemical reaction with a component of the liquid phase. The gas moves in the column like jets and bubbles, while the liquid moves in the form of drops, jets, and flowing films on the surface of the packed bed. As a result, the chemical absorption in a packed bed column is a combination of many elementary physical and chemical processes, as absorption in the systems gas-liquid drops, liquid-gas bubbles, gas-liquid film flow, etc. As an example will be considered the gas absorption in liquid film with free interface.

Let us consider absorption of a slightly soluble gas in a laminar liquid film [9, 10] in a coordinate system (x, y) , flowing over a

flat vertical interface $(y=0)$. The hydrodynamic model has the form:

$$\begin{aligned} \nu \frac{\partial^2 u_x}{\partial y^2} + g &= 0, \quad \frac{\partial u_x}{\partial x} + \frac{\partial u_y}{\partial y} = 0; \\ y=0, \quad u_x &= 0, \quad u_y = 0; \quad y=h_0, \quad \frac{\partial u_x}{\partial y} = 0 \end{aligned} \quad (6)$$

and the velocity distribution is:

$$u_x = \frac{g}{2\nu}(2h_0y - y^2), \quad u_y \equiv 0. \quad (7)$$

I these conditions the convection-diffusion model has the form:

$$\begin{aligned} \frac{g}{2\nu}(2h_0y - y^2) \frac{\partial c}{\partial x} &= D \left(\frac{\partial^2 c}{\partial x^2} + \frac{\partial^2 c}{\partial y^2} \right), \\ x=0, \quad c &= c_0; \quad x \rightarrow \infty, \quad c = c^*; \\ y=0, \quad \frac{\partial c}{\partial y} &= 0; \quad y=h_0, \quad c = c^*, \end{aligned} \quad (8)$$

where the thermodynamically equilibrium exists at the film interface $(y=h_0)$ and c^* denotes the equilibrium concentration.

The solid interface $(y=0)$ is impenetrable for the diffusing substance with inlet concentration $c_0 < c^*$ (absorption). A film of length l will be considered.

The diffusion boundary layer thickness δ [2] is:

$$\delta = \sqrt{\frac{Dl}{u^*}} = \frac{l}{\sqrt{Pe}}, \quad Pe = \frac{u^*l}{D}, \quad \frac{\delta}{l} = \frac{1}{\sqrt{Pe}}, \quad u^* = \frac{gh_0^2}{2\nu}, \quad \frac{\delta^2}{h_0^2} = Fo = \frac{Dl}{u^*h_0^2}, \quad (9)$$

where u^* , Pe , Fo , are the film interface velocity, the Peclet and Fourier numbers.

The diffusion boundary layer thickness δ is less than liquid film thickness h_0 that permits the diffusion boundary layer

approximation to be applied. As a result, the next generalized variables can be introduced:

$$x = lX, \quad y = h_0 - \delta Y, \quad c = c_0 + (c^* - c_0)C, \quad (10)$$

where $0 = \frac{h_0}{l} \leq 10^{-2}$.

The introduction of (10) into (8) yields:

$$\begin{aligned} (1 + FoY^2) \frac{\partial C}{\partial X} &= \frac{Dl}{u^* \delta^2} \left(Pe^{-1} \frac{\partial^2 C}{\partial X^2} + \frac{\partial^2 C}{\partial Y^2} \right); \\ X = 0, \quad C = 0; \quad X \rightarrow \infty, \quad C = 1; \\ Y \rightarrow \infty, \quad C = 0; \quad Y = 0, \quad C = 1, \end{aligned} \quad (11)$$

where Fo is a small parameter:

$$\frac{\delta^2}{h_0^2} = Fo < 10^{-1}, \quad \frac{\delta^2}{l^2} = Pe^{-1} \leq 10^{-2}. \quad (12)$$

The problem (11) in the diffusion boundary layer approximation ($10^{-2} \geq Pe^{-1} = 0$) has the form: namely:

$$\begin{aligned} (1 + FoY^2) \frac{\partial C}{\partial X} &= \frac{\partial^2 C}{\partial Y^2}; \\ X = 0, \quad C = 0; \quad Y = 0, \quad C = 1; \quad Y \rightarrow \infty, \quad C = 0. \end{aligned} \quad (13)$$

The mass transfer rate (J) in the liquid film flow with a length l is the average value of the local mass flux through the face interphase ($y = h_0$). On the other hand this rate can be presented using the mass transfer coefficient k . As a result

$$J = \frac{D}{l} \int_0^l \left(\frac{\partial c}{\partial y} \right)_{y=h_0} dx = k(c^* - c_0). \quad (14)$$

In the generalized variables, from (14) the Sherwood number (Sh) is possible to be obtained:

$$Sh = \frac{kl}{D} = -\sqrt{Pe} \int_0^1 \left(\frac{\partial C}{\partial Y} \right)_{Y=0} dX, \quad (15)$$

where $C(X, Y)$ is the solution of (13) [9, 10] developed by a perturbation method [2]:

$$Sh = \sqrt{\frac{6Pe}{\pi}} \left(1 - \frac{Fo}{6} - \frac{19Fo^2}{120} \right). \quad (16)$$

The examined example shows the theoretical techniques, used for the modeling and simulation of the elementary processes in a complex industrial process. The presence of the models of all elementary processes in a complex industrial process does not give a practical opportunity to synthesize its model. The role of modeling and simulation of elementary processes is the theoretical explanation of a number of physical and chemical effects observed experimentally in industrial processes. An illustrative example in this respect is the nonlinear theory of the mass transfer.

3.2. Non-linear mass transfer theory

The theory of the diffusion boundary layer [2] is the basis of modern linear mass transfer theory, where the convection-diffusion equation in (8) is linear, i. e. the velocity does not depend on the concentration. In a number of cases, the experimental results for the mass transfer rate are higher than the predictions of the linear theory [11, 12]. This is due to nonlinear effects, where the mass transfer influences the hydrodynamics and the velocity begin to depend on concentrations. These non-linear effects are related to the induction of secondary flows at the interphase boundaries as a result of interphase mass transfer. Such effects are the effect of large concentration gradients [11], the effect of Marangoni and the effect of Stephan flow [12].

The large concentration gradients create an intensive diffusion flux that have a hydrodynamic character, and a secondary flow is induces, directed at the normal of the interphase boundary and results in an additional convective mass transfer.

The effect of Marangoni is a result of the gradient of the surface tension on the interphase surface, as a result of the surface gradient of the temperature or surface active agents concentration on the liquid-gas (liquid) interphase, and induces a tangential flow. As a result of the continuity of the flow, there appears to be a much lower flow in the direction of the normal of the interphase boundary and consequently an additional convective flow. Because of this, this effect is relatively weak and occurs in motionless or slow moving fluids.

The Stephan's flow is a result of a phase transition liquid-steam at the interphase surface when the volume of the liquid (steam) increases (decreases) a thousand times. As a result, there is a secondary flow, directed to the normal of the interphase boundary, and an additional convective mass transfer.

In the above three cases, an additional hydrodynamic effect appears very often because the secondary currents disturb the hydrodynamic stability of the flows and self-organizing dissipative structures occur, which further accelerate the mass transfer [9]. To these effects can be added the Benar instability [12] in the case of a positive vertical gradient of the density of gases or liquids resulting from concentration or temperature gradients.

The theory of mass transfer allows the construction of the process model if its physical mechanism is known. The model thus obtained allows for the identification of this mechanism, i.e. the determining of the significant physical effects and rejection of the insignificant, using the generalized analysis method [2].

4. Physical mechanism identification

The qualitative analysis of the models permits to be made the physical mechanism identification, using generalized variables [2]. They are dimensionless variables, whereas as the characteristic (inherent) scales are used the maximal or average values of the variables. The introduction of the generalized variables leads to dimensionless model. As a result the unity is the order of magnitude of all functions and their derivatives in the model, i.e. the effects of the physical and chemical phenomena (the contribution of the terms in the model), are determined by the orders of magnitude of the dimensionless parameters in the model. If all model equations are divided by the dimensionless parameter, which has the maximal order of magnitude, all terms in the model equations will be classified in three parts:

1. The parameter is unity or its order of magnitude is unity, i.e. this mathematical operator represents a main physical effect;
2. The parameter's order of magnitude is 10^{-1} , i.e. this mathematical operator represents a small physical effect;
3. The parameter's order of magnitude is $\leq 10^{-2}$, i.e. this mathematical operator represents a very small (negligible) physical effect and has to be neglected, because it is not possible to be measured experimentally.

After the physical mechanism identification the model contains a minimum number of parameters which must be determined experimentally.

5. Parameters identification

In general case, the identification of the parameters in the model is made by the minimization of the least squares function by the inverse identification problem solution [1-3]. The least squares function represents the sum of the squares of the differences between the calculated and the experimental values of the functions in the model and its minimum must be obtained with respect to the parameters in the model. This inverse identification problem is very often incorrect and needs special methods for the solutions [2].

6. Statistical analysis of the model adequacy

The stochastic nature of the errors during the experimental data determination leads to subsequent errors of the model parameters. The model is adequate if the variance of the statistical error of the model does not exceed the variance of the statistical error of the experimental data, i. e. the accuracy of the functions calculation by the model is not less than the accuracy of the function experimental measurement [2, 3].

7. Processes simulation

The process simulation uses the most common numerical methods of applied mathematics to solve model equations. For this purpose, commercial software is usually used. In many cases, however, it is necessary to introduce this software into specialized algorithms [2, 13].

8. Modeling of industrial mass transfer processes in column apparatuses

The diffusion boundary theory is not applicable for the modeling of chemical, absorption, adsorption and catalytic processes in column apparatuses, where the velocity distributions and interphase boundaries are unknown.

The use of the physical approximations of the mechanics of continua for the interphase mass transfer process modeling in industrial column apparatuses is possible if the mass appearance (disappearance) of the reagents on the interphase surfaces of the elementary physical volumes (as a result of the heterogeneous reactions) are replaced by the mass appearance (disappearance) of the reagents in the same elementary physical volumes (as a result of the equivalent homogenous reactions), i.e. the surface mass sources (sinks), caused by absorption, adsorption or catalytic reactions must be replaced with equivalent volume mass sources (sinks). The solution of this problem is related with the creation of new type of convection-diffusion and average-concentration models [13].

The convection-diffusion models permit the qualitative analysis of the processes only, because the velocity distribution in the column is unknown. On this base is possible to be obtained the role of the different physical effect in the process and to reject those processes, whose relative influence is less than 1%, i.e. to be made process mechanism identification.

The average-concentration models are obtained from the convection-diffusion models, where average velocities and concentrations are introduced. The velocity distributions are introduced by the parameters in the model, which must to be determined experimentally.

Conclusions

The theoretical foundations of modeling and simulation of the industrial processes are presented. The first step is the formulation

of the physical mechanism of the industrial process and the construction of a mathematical structure, containing the mathematical operators that quantitatively describe the individual physical effects in this mechanism. The introduction of generalized (dimensionless) variables through characteristic scales permits to be made a quality analysis of the industrial processes. The obtaining of the experimental data and using it to calculate the model parameters by solving a inverse identification problem leads to the final form of the mathematical model. The statistical analysis of the model adequacy leads to the practical applicability of the mathematical model. The presented results are published in 8 monographies (www.iche.bas.bg/Books_BG.htm).

References

1. Хр. Бояджиев, Основи на моделирането и симулирането в инженерната химия и химичната технология, ИИХ-БАН, София, 1993. (Fundamentals of modeling and simulation in chemical engineering and chemical technology - Bulgarian).
2. Chr. Boyadjiev, Theoretical Chemical Engineering. Modeling and simulation, Springer-Verlag, Berlin Heidelberg, 2010.
3. Хр. Бояджиев, Основи на моделирането и симулирането в химичната промишленост, Изд. БАН „Проф. Марин Дринов“, София, 2017. (Fundamentals of modeling and simulation in chemical industry - Bulgarian).
4. Е. Л. Фейнберг, Две культуры. Интуиция и логика в искусстве и науке, БЕК 2, Фрязино, 2004, 286 сс. (Logic and intuition in art and science - Russian).
5. Chr. Boyadjiev, Some Thoughts on Logic and Intuition in Science and Chemical Engineering, Open Access Library Journal, 1, №6, 1-5, 2014.
6. Keizer J., Statistical Thermodynamics of Nonequilibrium Processes, Springer-Verlag, New York, 1987.
7. L. D. Landau, E. M. Lifshitz, Fluid Mechanics, Pergamon, Oxford, 1989.
8. V. G. Levich, Physicochemical Hydrodynamics, Prentice Hall, New York, 1962.
9. Chr. Boyadjiev, V. Beschkov, Mass Transfer in Liquid Film Flows, Publ. House Bulg. Acad. Sci., Sofia, 1984.
10. Хр. Бояджиев, В. Бешков, Массоперенос в движущихся пленках жидкости. Москва, Изд. „Мир“, 1988. (Mass Transfer in Liquid Film Flows - Russian).
11. В. С. Крылов, Хр. Бояджиев, Нелинейный массоперенос. Новосибирск, Изд. Институт теплофизики СО РАН, 1996. (Non-Linear Mass Transfer - Russian).
12. Chr. B. Boyadjiev, V. N. Babak, Non-Linear Mass Transfer and Hydrodynamic Stability, New York, ELSEVIER, 2000.
13. Chr. Boyadjiev, M. Doichinova, B. Boyadjiev, P. Popova-Krumova, Modeling of Column Apparatus Processes, Springer-Verlag, Berlin Heidelberg, 2016.

APPLICATION OF A HYBRID MODEL: FRACTIONAL EXPERIMENTAL DESIGN + ARTIFICIAL NEURAL NETWORKS+ GREY RELATIONAL ANALYSIS METHOD ON OPTIMIZATION OF PROCESS PARAMETERS OF POWDER METALLURGY

S. Hartomacıoğlu. PhD.¹, H.O.Gülsoy. PhD.², B. Bakırcıoğlu Ms.C.³, S. Yuksel, Ms.C.⁴
Department of Mechanical Engineering – Marmara University, Turkey¹
Department of Metallurgy and Materials Engineering– Marmara University, Turkey²
Department of Mechatronics, Selcuk University, Turkey³
selimh@marmara.edu.tr

Abstract: In this study, a hybrid model was proposed to examine the expected properties of real working conditions of the part produced by powder metallurgy processes. The hybrid model consists three steps: 1) fractional experimental design, 2) artificial neural network, 3) grey relational analysis. In the proposed method, the effect of production parameters of synthetic diamond based cutting tool for concrete on wear and operation times were examined and optimized. As the production parameters, composition of binder, sintering temperature and holding time were defined. After the implementation of the proposed model steps, the percentage error of wear rate and operation time is 0.23 % and 5.98% in the test steps, respectively. Outside of training and test data group, 4 confirmation experiment predicted by using the model from new data group were conducted and the percentage error of wear rate and operation time were 2.69% and 7.35%, respectively. As a results of this study, this model can be used in powder metallurgy and other manufacturing methods in order to predict and optimized new results.

Keywords: FRACTIONAL EXPERIMENTAL DESIGN, ARTIFICIAL NEURAL NETWORKS, GREY RELATIONAL ANALYSIS

1. Introduction

More quality, lower cost ve shorter time of research and development of products have become more important due to fast developing technology and increasing competitive conditions. For this purpose, scientists have focused on methodologies that used in research and development processes and developed different methodologies and hybrid methods. In the literature, there are a lot of study about experimental design, optimization and multi criteria decision methods [1-3].

Powder metallurgy is one of production methods. There are a lot of parameters in production process. If someone want to develop new part that produced by using powder metallurgy, he must optimize the parameters, and examine the effect of product properties under real working conditions. In other words, the production parameters are modelled and optimized by using different methods.

In this study, a hybrid model that discipline the experimental studies of powder metallurgy processes is proposed and explained with case study. The hybrid method consists three methods; 1) fractional experimental design, 2) artificial neural network (ANN) method, and 3) grey relational analysis (GRA) method. In this method, fractional experimental design method was used to reduce the number of experiments, the artificial neural network was used to model the production parameters and the new results was predicted using this ANN structure, the grey relational analysis method that used to select optimal results in multi criteria decision making situation was used to sort / optimize the experimental results. The application of the hybrid model was conducted on optimization production parameters of synthetic diamond based cutting tool produced powder metallurgy method for concrete.

2. Proposed Hybrid Model

In the powder metallurgy processes, there are a lot of production parameters for example: compacting pressure, sintering temperature, holding time, additive powder ratio, heat rate, etc. In the experimental design, the input parameters correspond to production parameters. The output parameters correspond to the desired properties of products for example; density, hardness, wear, flexural strength, etc. For examination and optimization, the production parameters according to output parameters of powder metallurgy processes, if someone use the full-factorial experimental

design method, he must conduct a lot of number of experiments and spend a lot of time, then the situation lead to high cost. In this method, there are 5 steps; 1) design of experiments (DOE), 2) Developing the optimal ANN model, 3) Sorting results using GRA method, 4) Optimization experimental results, 5) Confirmation of the proposed model. In the first step, firstly, the input parameters of experiments and their levels were defined, and appropriate fractional experiment design table was selected using these parameters, results of step 1, the experimental table was prepared. In the next step, ANN step, the data was normalized between 0 and 1, and divided into training and testing data set, and the ANN topologies, training algorithms, and transfer functions were defined, and the optimal ANN structured was obtained. The new experimental conditions that untired were predicted by using the optimal ANN structure. In the GRA step, the grey relational degree of each experiments were calculated. In the optimization step, the experiments were sorted and optimized by using the grey relational degree of each experiments. In the end of step 4, the optimal results and worst results of experiments were defined. To prove the validity of the model, confirmation experiments were conducted. For this purpose, the 4 number of experiments from predicted results were selected, and the experiments were performed. Finally, for the confirmation of the model, the predicted results and confirmation experiments results were compared, and the percentage errors were calculated for each confirmation experiments results. Flow chart of proposed hybrid method was shown in Figure 1. I

2.1 Design of Experiments

Design of experiments (DOE) is a methodology that used in the processes of product developments. There are two approach in DOE; 1) full-factorial experimental design, 2) fractional experimental design. Especially, with the development of powerful prediction methods, the fractional experimental design method become widespread, so, the costs of research and development processes was reduced. In the fractional experimental design, the Taguchi method was generally used in many engineering fields.

In the experimental design, there are input and output parameters. The number of input parameters and their levels were selected by experienced engineers and results of preliminary experiment results, and the engineers prepared the experimental table using different fractional experiments methods. The output parameters were defined according to real working conditions of product. The experimental design and optimization results is a multi-criteria decision making process.

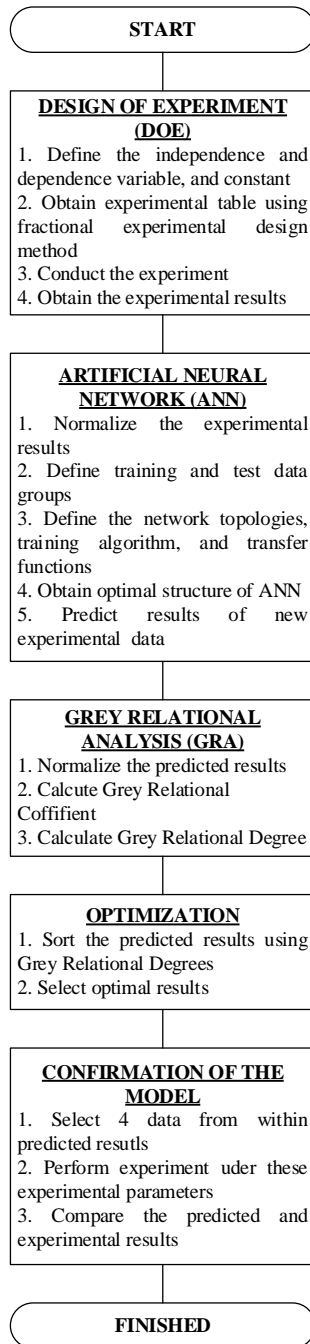


Figure 1 Flow chart of proposed model

In the experimental design, input parameters or in other words independent variables (X_{ij} (for $i=1, 2, \dots, m, j=1, 2, \dots, n$)), output parameters in other words dependent variables (Y_{kl} (for $k=1, 2, \dots, m, l=1, 2, \dots, n$)) were listed in Table 1.

Table 1 Experimental table

Input Variable				Output Variable			
X_1	X_2	X_j	X_n	Y_1	Y_2	Y_j	Y_k
X_{11}	X_{21}	X_{j1}	X_{n1}	Y_{11}	Y_{21}	Y_{j1}	Y_{k1}
X_{12}	X_{22}	X_{j2}	X_{n2}	Y_{12}	Y_{22}	Y_{j2}	Y_{k2}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
X_{1i}	X_{2j}	X_{ji}	X_{ni}	Y_{1i}	Y_{2j}	Y_{ji}	Y_{ki}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
X_{1m}	X_{2n}	X_{jm}	X_{mn}	Y_{1l}	Y_{2l}	Y_{jl}	Y_{kl}

2.2 Artificial Neural Networks

Artificial neural network (ANN) is a modeling and prediction tool, widely accepted as a technique offering an alternative way to tackle complex problems, and in addition, can learn exemplars and, one trained, can perform prediction and generalization at high speed [4]. Artificial neural network is a branch of artificial intelligent. The artificial neural network method was used in many engineering areas for purpose of prediction and classification. Many artificial neural networks structure have been developed so far, for example: single layer perceptron (SLP), multilayer perceptron (MLP), Hopfield network, and Radial Basis Network, etc. In the experimental design and optimization, multilayer perceptron structure is used more often to predict new results of experiments.

The ANN procedure has a series steps. First step, the data is normalized according to transfer function. In the proposed model, sigmoid is selected for transfer function and, the data is normalized ranging from 0 to 1. Next step, the data set was sorted randomly, and divided into training data set (75%) and testing data set (25%). After the preparation of data, training processes was carried out.

MLP generally has three layers: 1) input layer, 2) hidden layer, and 3) output layer. There is one or more artificial neuron in each layers. The artificial neuron develops from the biological neurons. The artificial neuron and biological neuron are shown in Figure 2. In the ANN structure, the input parameters of experiments correspond to number of neurons of input layers of MLP, the output parameters of experiments correspond to number of output layer of MLP. The number of hidden layer neurons is found by trial and error method. General MLP structure is shown in Figure 3.

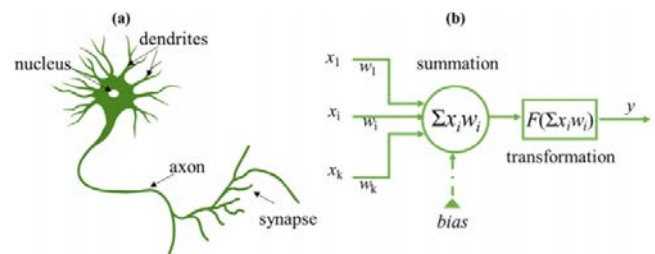


Figure 2 Biological and artificial neuron [5]

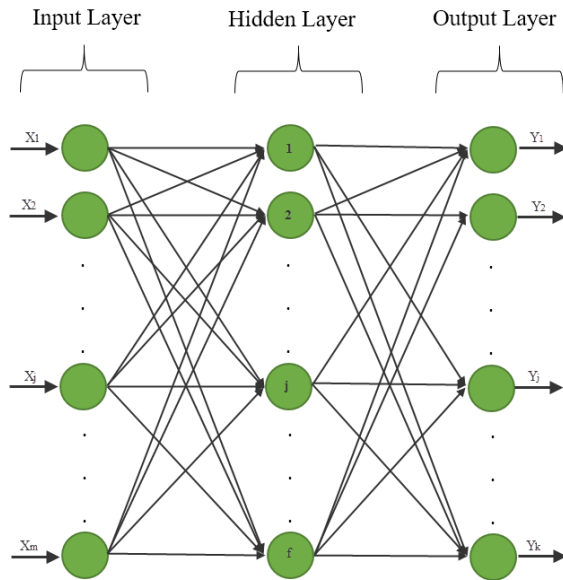


Figure 3 General MLP structure

Artificial neuron, in other words processes element, is basic element of ANN structure, and general procedure of artificial neuron and transfer function are shown in Figure 4.

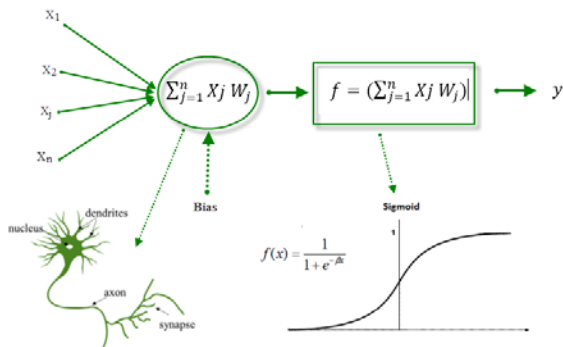


Figure 4 Principle of artificial neuron

In the training step, the performance evaluations are performed using mean square error (MSE) shown Equation 1. As the MSE value approaches zero, the error is minimized. In the training step, main goal is to obtain the minimal MSE error, and define the optimal ANN structure parameters. MLP is the feed-forward and back propagation network type. In this study, the momentum is used for back propagation algorithm. After the successful training step, the test step is carried out. In this step, firstly, the testing operations are performed using training data set, secondly the testing operations are performed using testing data set. The percentage error and correlation coefficient value of each testing operations is calculated.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (1)$$

Where y_i is output of network, \hat{y}_i is output of experimental results.

Final step of ANN procedure, new experimental table is prepared and the results of this table are predicted using obtained optimal ANN structure.

2.3 Grey Relational Analysis

The experimental results are multi-criteria and the multi-criteria decision method must be used for optimization of experimental results. The one of multi-criteria decision method is Grey Relational Analysis (GRA). GRA procedure consists 3 steps: 1) grey relational generating, 2) calculation of grey relational coefficient, 3) calculation of grey relational grade, 4) sorting experimental results.

First step of GRA is grey relational generating. In this step, the criteria's values are normalized by using 'the lower – the best', 'the highest – the best', 'the target – the best' approaches. For this, if there are m unit and n variables, i_{th} alternative is expressed as $Y_i = (y_{i1}, y_{i2}, y_{i3}, \dots, y_{ij}, \dots, y_{in})$. Where, y_{ij} is performance value of variables of j and alternative i . The normalization procedure of 'the larger-the better', 'the lower- the better', and 'the target – the better' approach is shown in Equation 2, 3, and 4, respectively [6].

$$x_{ij} = \frac{\max\{y_{ij}, i=1,2,\dots,m\} - y_{ij}}{\max\{y_{ij}, i=1,2,\dots,m\} - \min\{y_{ij}, i=1,2,\dots,m\}} \quad (2)$$

$$x_{ij} = \frac{y_{ij} - \min\{y_{ij}, i=1,2,\dots,m\}}{\max\{y_{ij}, i=1,2,\dots,m\} - \min\{y_{ij}, i=1,2,\dots,m\}} \quad (3)$$

$$x_{ij} = 1 - \frac{|y_{ij} - y_j^*|}{\max\{\max\{y_{ij}, i=1,2,\dots,m\} - y_j^*, y_j^* - \min\{y_{ij}, i=1,2,\dots,m\}\}} \quad (4)$$

Where, for $i = 1, 2, \dots, m, j = 1, 2, \dots, n$

The second step of GRA is calculation of grey relational coefficients. Firstly, the reference sequence is defined. The reference sequence is $X_0, (x_{01}, x_{02}, \dots, x_{0j}, \dots, x_{0n}) = (1, 1, \dots, 1, \dots, 1)$. The deviation sequence is calculated using reference sequence and normalized data. After this, the grey relational coefficient is calculated using Equation 5.

$$\gamma(x_{0j}, x_{ij}) = \frac{\Delta_{\min} + \zeta \Delta_{\max}}{\Delta_{ij} + \zeta \Delta_{\max}} \quad \text{for } i = 1, 2, \dots, m \quad j = 1, 2, \dots, n \quad (5)$$

Here, ζ is a distinguishing coefficient between 0 and 1. Studies demonstrate that the value of ζ does not affect the sorting that will occur after the calculation of the Grey Relational Degree. Δ_{ij} is the amount of deviation between the reference series and normalization values.

The third step of GRA procedure is calculation of grey relational grade. The grey relational grade is calculated using Equation 6.

$$r(X_0, X_i) = \sum_{j=1}^n w_j \gamma(x_{0j}, x_{ij}) \quad \text{for } i = 1, 2, \dots, m \quad (5)$$

Where $r(X_0, X_i)$ is expressed as grey relational grade. w_j is defined as weight of criteria.

Final step of GRA procedure, the experiments are sorted using the grey relational grade. The experiment with the highest grade is the best experimental conditions.

2.4 Confirmation of the model

To show the model's performance, the best 5 experiments and the worst 5 experiments are selected and the experiments are carried out. The results of experiments and model is compared and calculated the percentage error of the experiments. If the percentage error is below 15% error, the model is appropriate for modeling of process parameters of powder metallurgy method.

3. Case Study

The application of the hybrid model is performed to optimize production parameters of synthetic diamond based cutting tool for concrete. In the concrete drilling operations, HILTI DD350 machines was used and the machine has an auto-feed property. The machine working and cutting principle was shown in Figure 5.

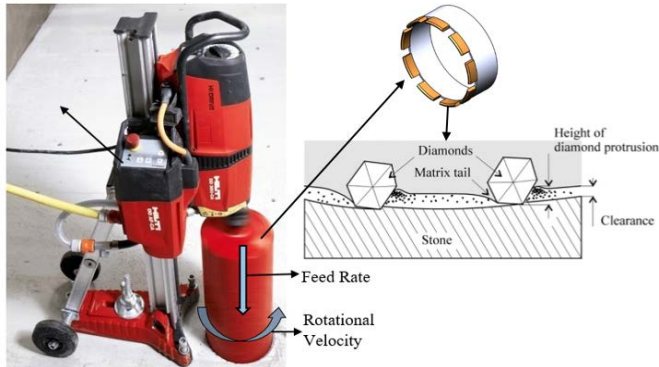


Figure 5 Principle of working and cutting of drilling machine

The segment that produced synthetic diamond based could be seen in Figure 5. The segments are produced by using powder metallurgy processes. The segments generally consists synthetic diamond (20 vol.%) and binder system (80vol%). The binder system components is bronze, cobber, cobalt, iron and lubricator. In the literature, different binding systems are used.

In this study, the bronze and cobalt composites were selected as binding systems. The zinc stearate was used as lubricator. Additional amonut of cobalt (wt.%), sintering temperature (°C), and holding time (second) were selected input parameters of experimental design. Wear rate and machining time were selected as output parameters of experimental design. The input parameters and their levels were listed in Table 2.

Table 2 The input factors and their levels

Factors	1	2	3	4	5	6
Add. Amount of Co(wt.%)	10	15	20	25	30	35
Sint. Temp., °C	800	850	900			
Holding Time, min	30	60	90			

The experimental table was prepared using Taguchi experimental design, and L18 orthogonal array was used according to number of input factors and their levels. In the experimental procedure, the powder and diamond mixture was obtained using 3D Turbula mixer. Then, the segments were pressed using the feed-stock using uniaxial hydraulic pressing machine under 10t load. The segments dimensions and mold were shown in Figure 6. The bit diameter of drilling machine is 100mm, and there were 10 segments of around of bit equidistant.

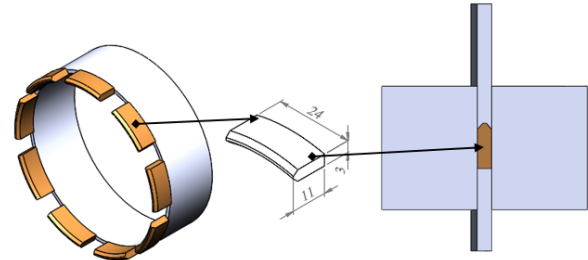


Figure 6 Dimension of segment and molding

After the cold pressing operations, the green segments were sintered using pressureless sintering method under hydrogen atmosphere at different sintering temperature according the experiment table. After the sintering operations, the segments were examined and some of the segments were selected as cutting performance experiments. For the drilling experiments, the segments were welded on bit. The drilling experiments were carried out using special test stands. The test stands consist C40 quality concrete and structural steel of 26mm. The drilling depth of concrete was 30cm. During the test operations, the operation time is determined. Before the experiments, the total height of bit was measured at top surface using CMM machine at three points, and after the machining operations, the total height of bit was also measured at same surface at same points. Total machining length is 100 cm.

L18 experiments table, results of experiments were listed in Table 3.

Table 3 Results of experiments

#	X1	X2	X3	Y1	Y2
1	10	800	30	5.08	105
2	10	850	60	3.22	138
3	10	900	90	2.55	143
4	15	800	30	4.82	115
5	15	850	60	3.11	148
6	15	900	90	1.72	155
7	20	800	60	4.35	128
8	20	850	90	2.95	161
9	20	900	30	0.95	169
10	25	800	90	4.18	132
11	25	850	30	2.88	172
12	25	900	60	0.77	175
13	30	800	60	3.95	145
14	30	850	90	2.22	171
15	30	900	30	0.44	180
16	35	800	90	3.88	168
17	35	850	30	1.88	188
18	35	900	60	0.38	213

After the normalization process, the data was divined into traning and testing data set. In the MLP structure, number of input layer neurons were 3, number of output layer neurons were 2, and hidden layer neurons number were defined trial error method ranging from 2 to 50 neurons. The maximum epoch and train times were 50 000 and 3, respectively. The ANN structure was developed using Neurosolutions software. In the training step, the MSE changing graphic according to hidden layer neurons was shown in Figure 7.

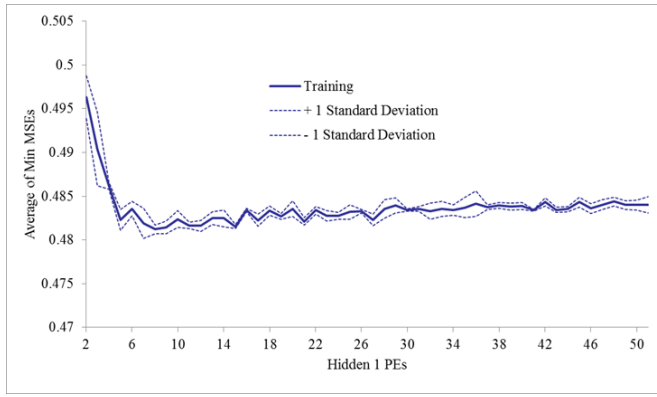


Figure 7 MSE value of training step

According to the Figure 7, the minimal MSE value was defined at 8 neurons number of hidden layer at 50 000 epochs. The MSE value of optimal ANN is 0.4806. In the optimal ANN structure, hidden layer neuron number, transfer function, and training algorithm is 8, sigmoid, and momentum, respectively.

In the testing step, firstly, the test operations were conducted using training data set. The results of this operations were shown in Figure 8, and 9 for wear rate (Y1) and operation time (Y2). The percentage error of experiments and output of network for wear rate and operation time is 0.042%, and 0.804%, respectively. This error values showed that the network was very powerful.

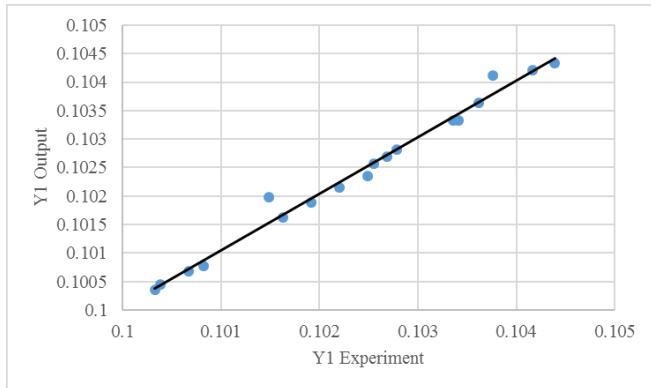


Figure 8 Testing results of training data set for wear rate

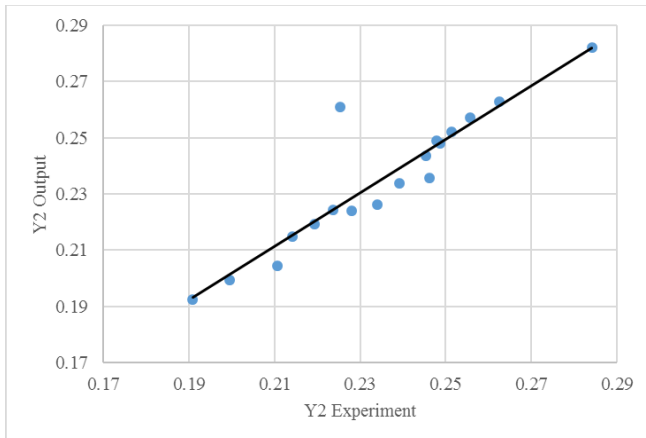


Figure 9 Testing results of training data set for operation time

For test operations, secondly, the test operations were carried out using testing data set. The percentage error of wear rate and operation time between results of experiments and output of networks were 0.23%, and 5.98%, respectively. The results show that the ANN structure could be acceptable because the error was below 10% error level. The results of test operations for testing data set was shown in Figure 10. The optimal ANN structure was shown in Figure 11.

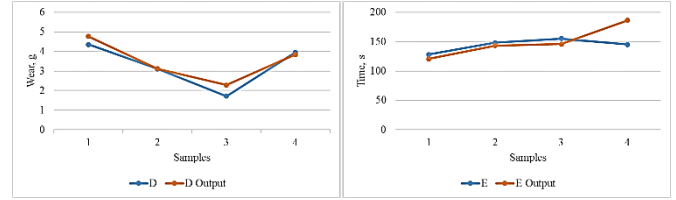


Figure 10 testing results of optimal ANN structure

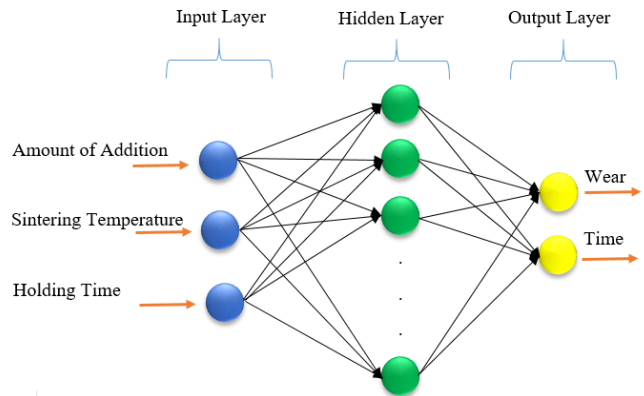


Figure 11 Optimal ANN structure

After the training and testing operations, the optimal ANN structure was defined, and new experiments table were prepared, and total experiments were 532. The results of new experiments conditions were predict using the optimal ANN structure. After the prediction steps, the results were optimized using GRA procedure. The best 5 experiments were listed in Table 4, and the worst experiments were listed in Table 5 according to grey relational degree value.

Table 4 The best 5 experiments

Predicted Results					G. R. Degree
X1	X2	X3	Y1	Y2	
15	925	40	1.6664	117.6528	0.7299
15	925	30	1.0991	130.4332	0.7234
20	925	50	1.0156	133.4620	0.7215
15	925	50	1.9599	115.8829	0.7170
20	925	40	0.7827	144.2946	0.7165

Table 5 The worst 5 experiments

Predicted Results					G. R. Degree
X1	X2	X3	Y1	Y2	
40	775	80	3.6786	192.3578	0.3974
30	800	50	3.6830	192.1116	0.3976
30	800	40	3.7369	190.0973	0.3984
40	775	70	3.5127	196.0550	0.3988
40	775	60	3.4658	196.4420	0.4002

The 4 experiments were selected during the best and worst experiments conditions to verification of the model. The selected experiments conditions, output and experimental results, and were listed in Table 6, 7, and 8.

Table 6 Selected experiments conditions for verification

#	X1	X2	X3
1	15	925	40
2	15	925	30
3	40	775	80
4	30	800	50

Table 7 Results of experiments and output of network

#	Y1 Out.	Y2 Out.	Y1 Exp.	Y2 Exp.
1	1.67	117.65	1.58	122.00
2	1.10	130.43	1.12	139.00
3	3.68	192.36	3.59	210.00
4	3.68	192.11	3.72	173.00

Table 8

#	Y1 Error, %	Y2 Error, %
1	5.19	3.69
2	2.18	6.57
3	2.41	9.17
4	1.00	9.95
Average, %	2.69	7.35

As a results of the confirmation experiments, the average percentage error of wear and operations time were 2.69% and 7.35%, respectively. This situation showed that the proposed model is very powerful and the model could be applied different engineering fields.

4. Conclusion

In this study, a hybrid method: fractional experiment design + artificial neural network and grey relational analysis method was developed and the method applied on optimization of production parameters of powder metallurgy processes. In the case study, the synthetic diamond based segments were produced by PM, and after the applying the model procedure, the prediction error of wear and operation time were 2.69% and 7.35%, respectively. This results showed that the model could be used for optimization and modeling of production parameters of powder metallurgy parts. The situation leads to reduction of research and development time and costs of product developments. In the future studies, the method will be applied in different engineering areas.

5. References

- [1] Reihanian, M., Asadollahpour, S.R., Hajarpour, S., Gheisari, Kh., 'Application of neural network and genetic algorithm to powder metallurgy of pure iron', *Materials and Design*, 32, 3183-3188, 2011
- [2] Evis, Z., Arcaklioglu, E., 'Artificial neural network investigation of hardness and fracture toughness of hydroxylapatite', *Ceramic International*, 1147-1152, 2011
- [3] Amirian, M., Khorsand, H., Siadati, M.H.S., Farsani, R.E., 'Artificial neural network prediction of Cu-Al₂O₃ composite

properties prepared by powder metallurgy method', *Journal of Materials Research and Technology*, 2 (4), 351-355, 2013

[4] Messikh, N., Bousba, S., Bougdah, N., 'The use of a multilayer perceptron (MLP) for modeling the phenol removal by emulsion liquid membrane', *Journal of Environmental Chemical Engineering*, 5, 3483-3489, 2017

[5] Yan, Z., Kim, Y., Hara, S., Shikazona, N., 'Prediction of Y_a0.6Sr0.4Co0.2Fe0.8O₃ cathode microstructures during sintering: Kinetic Monte Carlo (KMC) simulations calibrated by artificial neural networks', *Journal of Power Sources*, 346, (2017) 103-112

[6] Kuo, Y., Yang, T., Huang, G.-W., 'The use of grey relational analysis in solving multiple attribute decision-making problems', *Computers & Industrial Engineering*, 55, 80-93, 2008

SELECTION OF OPTIMAL EXPERIMENTAL CONDITIONS OF TURNING OPERATIONS By USING SATISFACTION FUNCTION AND DISTANCE BASED MULTI-CRITERIA DESISION MAKING METHOD

B. Bakircioğlu Ms.C.¹, S. Hartomacioğlu. PhD.², S. Yuksel, Ms.C.³, Ş. Yazman, Ms. C.⁴, S. Güvercin, Ms.C.⁵, A. Duran, Ph.D.⁶

Department of Mechatronics, Selcuk University, Turkey¹

Department of Mechanical Engineering – Marmara University, Turkey²

Department of Research & Development, Arkel Elektrik ve Elektronik San. ve Tic. A.Ş., Turkey³

Department of Mechanic, Selcuk University, Turkey⁴

Department of Mechanic, Amasya University, Turkey⁵

Department of Manufacturing Engineering, Gazi University, Turkey

bbakircioglu@selcuk.edu.tr

Abstract: In selection of experimental conditions, satisfaction function and distance measured based multi-criteria decision making method was used. The method was used to optimize the results of turning operations. In the experimental design, the cutting tool type, cutting speed and feed rate was selected as input parameters. Therefore, cutting force and surface roughness were defined as output parameters of experiments. After the implementation of the method procedure, the optimal input parameters were defined according to output parameters criteria's. The results of the method were compared with results of grey relational analysis method. As a results, the satisfaction function and distance measured based multi-criteria selection method could be successfully used in multi-criteria decision makeine for optimization of cutting parameters.

Keywords: TURNINIG OPERATION, SATISFACTION FUNCTION, DISTANCE BASED METHOD, OPTIMIZATION

1. Introduction

In the turning operations, experimental results must be sorted by using different multi-criteria selection method in case of multi criteria. Solving this problem, so far, a lot of method have been developed and were applied to optimize the results of experiments [1-3].

Satisfaction function and distance measure based multi-criteria decision making method was applied on selection of optimal cutting parameters of turning operations. Experiments were conducted at 2 axis CNC Turning machine, and 316L stainless steel material was machined with different of tool type, cutting speed and feed rate. In the experimental design, tool type, cutting speed and feed rate were selected as input parameters, cutting force and surface roughness were defined as output parameters of turning operations. The full factorial experimental design method was used, and results of experiments were sorted by the method.

2. Materials and Method

2.1 Experimental Studies

Full factorial experimental design method was used in experimental studies. The input parameters of experiments were cutting tool type, cutting speed, and feed rate. The levels of cutting tools types were SNMG 12 04 08-MM, SNMG 12 04 08-MR, SNMG 12 04 08-QM, SNMG 12 04 12 –MM, and SNMG 12 04 16-MM. The levels of cutting speed were 125, 150, 175, and 200 m/min. The levels of feed rate were 0.1, 0.2, and 0.3 mm/rev. The turning cutting tool form (SANDVIK) used in experiments is shown in Figure 1.

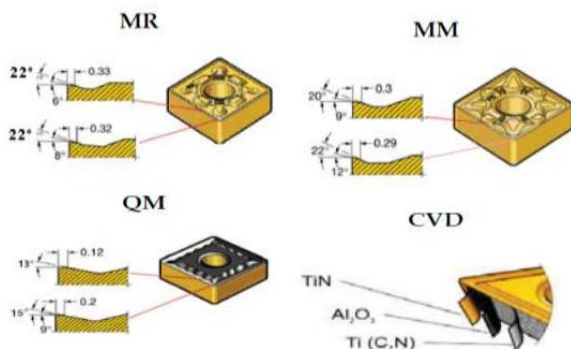


Figure 1 The turning cutting tool form used in experiments

The cutting speed and feed rate value used in experiments were listed in Table 1.

Table 1 cutting parameters of experiment conditions

Cutting Speed, (V, m/min)	Feed Rate, (f, mm/rev.)
125	0.1
	0.2
	0.3
150	0.1
	0.2
	0.3
175	0.1
	0.2
	0.3
200	0.1
	0.2
	0.3

316L stainless steel was used as workpiece materials. The material compositions were listed in Table 2.

Table 2 The chemical composition of 316L stainless steel

Quality	C	Mn	Si	Cr	Ni	P	S	Mo
316L	0.03	2.0	1.0	16-18	10-14	0.045	0.03	2-3

In the experiments, the 2 axis CNC turning machine was used. Total 60 experiments were carried out by using full factorial experimental design. The output parameters of experimental design were cutting force and surface roughness. The cutting forces of each experiments were measured using Kistler 9257B dynamometer. The cutting force measure system was shown in Figure 2.

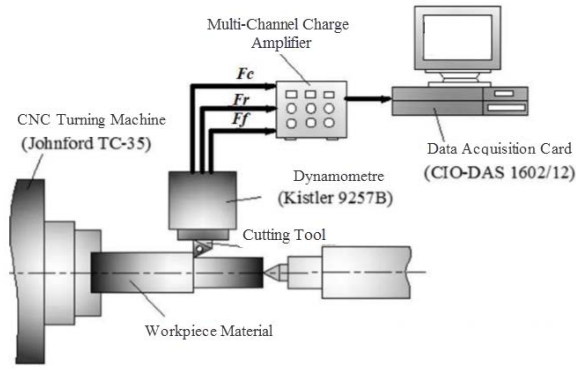


Figure 2 Cutting force measuring system

The surface roughness value of each machined parts were measured after machining by using Mahr Perthometer M1 at 5 points on each machined parts. The Ra values were calculated each parts.

2.2 Satisfaction Function and Distance Based Multi Criteria Decision Making Method

The satisfaction and distance based multi-criteria decision making method was explained in reference 1 [4]. The method was firstly applied on selection of robots by A. Kar and A. Kentli in 2011 [5]. The results of experiment results of turning operations were used, initially, the satisfaction value of each experiment number for every criterion were calculated. Then, distance measure values were calculated, and the experimental results were sorting by using this distance value. The minimum value of distance shows that the experiment is the best experimental condition.

The method, satisfaction function and distance based multi-criteria decision making method, consist two steps; 1) calculation of satisfaction values for using ‘the smaller- the better’ and ‘the larger- the better’ approaches, 2) calculation distance value, 3) Sorting experimental results using distance values.

In the calculation satisfaction value of experiments, the calculation procedure was shown in Figure 3.

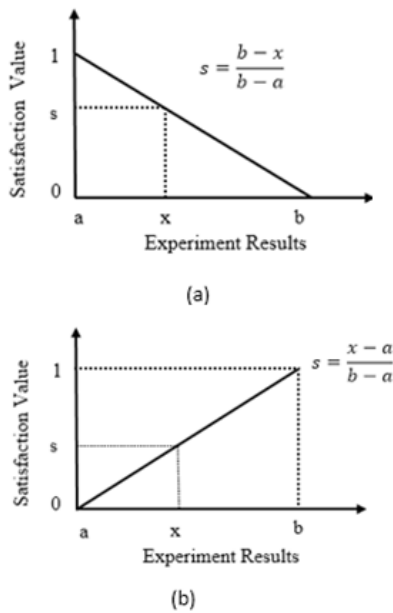


Figure 3 Calculation of satisfaction value a) ‘the smaller – the better’, b) ‘the larger – the better’ approaches,

In the experimental design, there are two criterion: 1) cutting force and surface roughness. The satisfaction value of cutting force and surface roughness were calculated by using ‘the smaller – the

better’ approach. After the calculation satisfaction value of criterion, the distance values were calculated by using Equation 1.

$$d = [\sum_{i=1}^n (1 - s_i)^2]^{\frac{1}{2}} \quad (1)$$

Where, d is the distance value, n is experiment number, s is satisfaction value.

3. Results and Discussion

During the experiments, cutting forces, F, of each experimental conditions were measured by Kistler 9257B Dynamometer, and after the experiments, the surface roughness value, Ra, were measured using Mahr Perthometer M1 devices. The experiments conditions, and the results of experiments were listed in Table 3.

Table 3 Results of experiments

#	Tool Type	V	f	F	Ra
1	SNMG 12 04 08-MM	125	0.1	771.77	1.58
2	SNMG 12 04 08-MM	125	0.2	1125.98	2.15
3	SNMG 12 04 08-MM	125	0.3	1496.98	4.34
4	SNMG 12 04 08-MM	150	0.1	813.38	2.09
5	SNMG 12 04 08-MM	150	0.2	1195.87	2.96
6	SNMG 12 04 08-MM	150	0.3	1560.75	5.15
7	SNMG 12 04 08-MM	175	0.1	777.96	1.43
8	SNMG 12 04 08-MM	175	0.2	1165.86	2.22
9	SNMG 12 04 08-MM	175	0.3	1518.31	4.45
10	SNMG 12 04 08-MM	200	0.1	727.76	1.74
11	SNMG 12 04 08-MM	200	0.2	1150.49	2.30
12	SNMG 12 04 08-MM	200	0.3	1506.08	3.91
13	SNMG 12 04 08-MR	125	0.1	1603.53	4.04
14	SNMG 12 04 08-MR	125	0.2	1250.35	2.85
15	SNMG 12 04 08-MR	125	0.3	1634.98	4.18
16	SNMG 12 04 08-MR	150	0.1	849.13	1.54
17	SNMG 12 04 08-MR	150	0.2	1295.25	2.51
18	SNMG 12 04 08-MR	150	0.3	1622.21	4.20
19	SNMG 12 04 08-MR	175	0.1	816.17	0.82
20	SNMG 12 04 08-MR	175	0.2	1251.24	1.79
21	SNMG 12 04 08-MR	175	0.3	1629.10	4.94
22	SNMG 12 04 08-MR	200	0.1	817.97	1.35
23	SNMG 12 04 08-MR	200	0.2	1218.58	1.89
24	SNMG 12 04 08-MR	200	0.3	1631.77	4.34
25	SNMG 12 04 08-QM	125	0.1	824.92	1.28
26	SNMG 12 04 08-QM	125	0.2	1192.42	2.46
27	SNMG 12 04 08-QM	125	0.3	1629.02	4.08
28	SNMG 12 04 08-QM	150	0.1	796.00	1.22
29	SNMG 12 04 08-QM	150	0.2	1165.78	2.26
30	SNMG 12 04 08-QM	150	0.3	1546.21	4.03
31	SNMG 12 04 08-QM	175	0.1	779.79	1.21
32	SNMG 12 04 08-QM	175	0.2	1215.84	2.16

33	SNMG 12 04 08-QM	175	0.3	1624.44	4.46
34	SNMG 12 04 08-QM	200	0.1	785.99	1.69
35	SNMG 12 04 08-QM	200	0.2	1203.99	2.65
36	SNMG 12 04 08-QM	200	0.3	1594.93	4.41
37	SNMG 12 04 12-MM	125	0.1	798.20	0.63
38	SNMG 12 04 12-MM	125	0.2	1158.55	1.30
39	SNMG 12 04 12-MM	125	0.3	1541.40	2.65
40	SNMG 12 04 12-MM	150	0.1	774.00	0.78
41	SNMG 12 04 12-MM	150	0.2	1140.12	1.30
42	SNMG 12 04 12-MM	150	0.3	1517.23	2.86
43	SNMG 12 04 12-MM	175	0.1	727.62	0.73
44	SNMG 12 04 12-MM	175	0.2	1116.02	1.41
45	SNMG 12 04 12-MM	175	0.3	1561.01	3.33
46	SNMG 12 04 12-MM	200	0.1	767.79	1.26
47	SNMG 12 04 12-MM	200	0.2	1114.85	1.25
48	SNMG 12 04 12-MM	200	0.3	1456.34	2.43
49	SNMG 12 04 16-MM	125	0.1	782.94	0.86
50	SNMG 12 04 16-MM	125	0.2	1188.20	1.60
51	SNMG 12 04 16-MM	125	0.3	1517.03	2.78
52	SNMG 12 04 16-MM	150	0.1	728.01	0.61
53	SNMG 12 04 16-MM	150	0.2	1176.22	0.80
54	SNMG 12 04 16-MM	150	0.3	1512.36	1.41
55	SNMG 12 04 16-MM	175	0.1	694.23	1.26
56	SNMG 12 04 16-MM	175	0.2	1072.19	2.14
57	SNMG 12 04 16-MM	175	0.3	1511.82	2.23
58	SNMG 12 04 16-MM	200	0.1	742.48	1.33
59	SNMG 12 04 16-MM	200	0.2	1127.84	1.37
60	SNMG 12 04 16-MM	200	0.3	1496.96	1.93

The next step of study, applying the method on the experiment results were performed. Firstly, the satisfaction values were calculated according to 'the smaller – the better' approach. The cutting force and surface roughness must be smaller. For this, the satisfaction values were calculated by using Figure 1. Secondly, the distance measures were carried out by using Equation 1, and finally, total distance of each criterion were calculated. The total distance values of each experiments were shown in Figure 4.

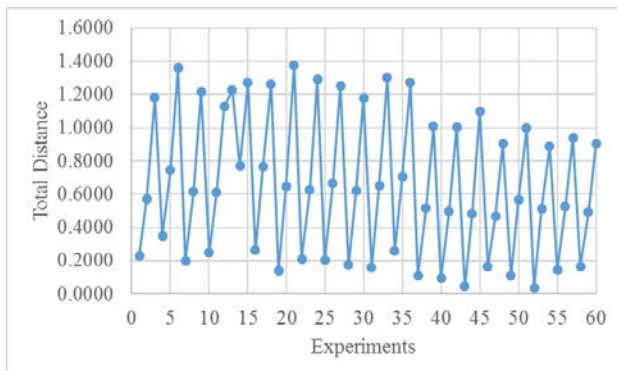


Figure 4 Total distance value of experiments

As seen the Figure 4, the minimum distance value was 0.0359 measured in experiment 52. The best and worst 4 experiments conditions were listed in Table 4 and Table 5, respectively.

Table 4 The best 5 experiments

#	Tool Type	V	f	d
52	SNMG 12 04 16-MM	150	0.1	0.0359
43	SNMG 12 04 12-MM	175	0.1	0.0450
40	SNMG 12 04 12-MM	150	0.1	0.0930
49	SNMG 12 04 16-MM	125	0.1	0.1093

Table 5 The worst 5 experiments

#	Tool Type	V	f	d
24	SNMG 12 04 08-MR	200	0.3	1.2911
33	SNMG 12 04 08-QM	175	0.3	1.3018
6	SNMG 12 04 08-MM	150	0.3	1.3596
21	SNMG 12 04 08-MR	175	0.3	1.3762

4. Conclusion

In this paper, application of satisfaction function and distance measured based multi-criteria selection method on selection optimal turning operation parameters was carried out. There are 3 input parameters and 2 output parameters of experiments. The input parameters were tool types, cutting speed, and feedrate. The tool types were SNMG 12 04 08-MM, SNMG 12 04 08-MR, SNMG 12 04 08-QM, SNMG 12 04 12 –MM, and SNMG 12 04 16-MM. The cutting speed levels were 125, 150, 175, 200 m/min. The feed rate levels were 0.1, 0.2, and 0.3 mm/rev. The workpiece materials was 316L quality of stainless steel. In the experimental design, full factorial experimental approach was used, and total 60 experiments were conducted. During the experiments, the cutting forces of each experiments were measured by using Kistler dynamometer. After the experiments, the surface roughness values of each experiments were measured by Mahr Perthometer M1. After the experimental procedure, firstly, the satisfaction value of each criterion were calculated. Then, using this satisfaction value, the distances measure was performed. Therefore, the experiments were sorted by using this distance value. Optimal experiments conditions, on other words, minimal distance value was 0.0359 measured in experiment 52. The worst experiments conditions, on other words, maximal distance value was 1.3762 measured in experiment 21. When the experimental results were examined in general and viewed in terms of the best and worst experimental conditions, it can be said that the application of the method is successful. As a results of this study, the satisfaction value and distance measured based multi-criteria decision making method can be used in turning operations to optimization of experimental results.

5. References

- [1] Jayaraman, P., Kumar, L.M., 'Multi-response optimization of machining parameters of turning AA6063 T6 aluminum alloy using grey relational analysis in Taguchi method', *Procedia Engineering*, 97, 197-204, 2014
- [2] Nayak, S.K., Patro, J.K., Dewangan, S., Gangopadhyay, S., 'Multi-objective optimization of machining parameters during dry turning of AISI 304 austenitic stainless steel using grey relational analysis', *Procedia Materials Science*, 6, 701-708, 2014
- [3] Camposeco-Negrete, C., 'Optimization of cutting parameters for minimizing energy consumption in turning of AISI 6061 T6 using Taguchi methodology', *Journal of Cleaner Production*, 53, 195-203, 2013
- [4] Hartomacioğlu, S., Kurt, M., Bakircioğlu, B., 'Application of Satisfaction Function and Distance Measure Based Multi-Criteria Decision Making Method on Selection Optimal Cutting Parameters', *Int. Journal of Latest Engineering Research and Applications*, 02, 01-05, 2017
- [5] A. Kentli, A.K. Kar, "A Satisfaction function and distance measure based multi-criteria robot selection procedure", *International Journal of Production Research*, Vol:49, No:19, 5821-5832, 2011

РАЗРАБОТКА ЭФФЕКТИВНОГО И УСТОЙЧИВОГО АЛГОРИТМА ДЕТЕКТИРОВАНИЯ ОБЪЕКТОВ В ПОСЛЕДОВАТЕЛЬНОСТИ КАДРОВ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ ВИДЕОНАБЛЮДЕНИЯ

Хлопин¹ С.В, Белов Н.А.²

¹ к.т.н, доцент Института Компьютерных наук и технологий СПбГПУ
e-mail: x@spbstu.ru

² аспирант Института Компьютерных наук и технологий СПбГПУ
e-mail: nik.belov@yahoo.com

Резюме: В данной работе рассматривается задача синтеза алгоритма детектирования пешеходов, устойчивого к частичному перекрытию объекта и плохому качеству изображения. Данный алгоритм разрабатывался для интеллектуальной системы видеонаблюдения, использующей роботизированную видеокамеру. Вычислительный эксперимент показал эффективность описанного метода, что позволяет предлагать использовать данный подход на практике.

Ключевые слова: ДЕТЕКТИРОВАНИЕ ОБЪЕКТА, МЕТОД ГИСТОГРАММ ОРИЕНТИРОВАННЫХ ГРАДИЕНТОВ, ДЕТЕКТОР ДВИЖЕНИЯ, АНАЛИЗ ИЗОБРАЖЕНИЯ, КОМПЬЮТЕРНОЕ ЗРЕНИЕ.

Введение

Одной из наиболее сложных задач в области автоматизированного слежения заключается в обнаружении объекта поиска в кадрах видеопоследовательности со сложной динамической сценой, а также нахождение параметров этого объекта, таких как размер, скорость, положение и ориентация в пространстве и так далее. Также среди прочих проблем нередко встречаются такие, как динамическое изменение освещенности сцены наблюдения, частичное или полное отсутствие информации об объекте, отсутствие информации о модели движения.

Для синтеза модели объекта необходимо составить комплексную систему признаков, обеспечивающую адекватную сложность при извлечении необходимой информации. Важной проблемой здесь является оценка ценности такой информации и оценка точности описания с помощью данной модели признаков. Из вышесказанного следует, что разработка новых методов для обнаружения объектов (в том числе подвижных) является важной и актуальной научно-исследовательской задачей.

Формулировка задачи

Целью исследования является синтез алгоритма детектирования пешеходов, устойчивого к частичному перекрытию объекта (не более 40%) и плохому качеству изображения. Важным требованием к алгоритму ставилось обеспечение высокой скорости обработки кадра для обеспечения работы в режиме реального времени (менее 100 мс. на один кадр).

Основная часть

Обычно на практике для решения задачи детектирования пешеходов используются методы категориального распознавания (классификация объектов) [1, 26]. Это обусловлено высокой точностью и скоростью алгоритмов при хороших условиях съемки (освещенность,

$$\|e\|_2^2 = \int_{-\infty}^{\infty} w(x-x', t-t') (f_1 g_x(x') + f_2 g_y(x') + g_t(x'))^2 dx' dt' \quad (2)$$

Задача решается с помощью метода наименьших квадратов. Затем используется процедура взвешенного усреднения, чтобы получить сокращение для уравнения свертки:

$$\|e\|_2^2 = (f_1 g_x + f_2 g_y + g_t)^2 \rightarrow \min \quad (3)$$

Решение для оптического потока может быть записано в следующем виде:

$$\begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = - \begin{bmatrix} g_x g_t \\ g_x g_x \\ g_y g_t \\ g_y g_y \end{bmatrix} \quad (4)$$

Можем выделить три случая:

отсутствие перекрытия объекта, статичность камеры). Однако, при несоблюдении таких условий методы категориального распознавания выдавали высокую погрешность. Для нивелирования этого недостатка необходимо внедрить дополнительный детектор для выявления косвенного признака. В качестве такого признака будем использовать движение объекта. Рассмотрим, каждый детектор более детально.

Для распознавания пешеходов (первый детектор) используется алгоритм, основанный на методе, использующий пирамидальные гистограммы ориентированных градиентов (PHOG) [2, 84]. Алгоритм построения дескрипторов представляет из себя совокупность различных техник:

1. Интегральное представление изображений;
2. Интегральные градиенты;
3. Прореживание;
4. Имитация алгоритма скольжения.

Для работы с PHOG дескрипторами использовался метод опорных векторов (SVM).

Для распознавания движения (второй детектор) использовался алгоритм, основанный на дифференциальных методах первого порядка, рассмотренный ниже.

Воспользуемся уравнением неразрывности оптического потока (1):

$$\frac{\partial g}{\partial t} + f^T \nabla g = 0 \quad (1)$$

где f – оптический поток, g – функция весовых коэффициентов пикселей.

Делается предположение, что оптический поток является постоянным в области [3, 236]. Имея ограничения неразрывности оптического потока во многих точках, такая система может быть решена с помощью минимизирования функционала ошибки.

1. $\overline{g_x g_x} > 0, \overline{g_y g_y} > 0$: пространственные изменения уровней яркости по всем направлениям. Обе компоненты оптического потока могут быть определены.

2. $\overline{g_x g_x} > 0, \overline{g_y g_y} = 0$: пространственные изменения уровней яркости только в направлении x (перпендикулярно контуру). Только компонента в направлении x может быть определена.

3. $\overline{g_x g_x} = 0, \overline{g_y g_y} = 0$: нет пространственных изменений уровней яркости в обоих направлениях. В случае постоянной области ни одна компонента оптического потока не может быть определена вообще.

Детектор движения является вспомогательным. Он предоставляет информацию о направлении движения в окрестности области предыдущего кадра, помеченной как область искомого объекта. Если детектор не смог распознать объект, второй детектор дает информацию о предполагаемом направлении движения объекта. Это снижает процент ошибок и увеличивает эффективность систем, использующих данную технологию.

Экспериментальное исследование

Для экспериментов использовалось изображение размером 1920×1080, шаг скольжения составлял 1×1, пропорции скользящего окна составляли 8×8. Алгоритмы реализовывались на языке C# с использованием библиотеки компьютерного зрения EmguCV.

Среднее время обработки одного кадра составило 68 мс. (59 мс. без вспомогательного детектора), средняя ошибка детектирования составила 0,12 (0,28 без вспомогательного детектора). Эксперимент проводился на вычислительной машине с процессором Intel Core i7-4790.

Заключение

Благодаря внедрению вспомогательного детектора движения, разработанный алгоритм детектирования имеет более высокую точность и лучше подходит для работы в плохих условиях съемки. Синтезированный алгоритм удовлетворяет всем предъявленным требованиям и может быть использован на практике при решении задач детектирования пешеходов.

Литература

1. P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. PAMI, 2012
2. Южаков Г.Б. Алгоритм быстрого построения дескрипторов изображения, основанных на технике гистограмм ориентированных градиентов // Труды МФТИ. М., 2013. Т.5. С.84-91
3. Яне. Б. Цифровая обработка изображений. М.: Техносфера, 2007. – 583 с.

THE EFFECT OF OPTICAL AND RECOMBINATION LOSSES IN $\text{Cu}_2\text{ZnSn}(\text{S,Se})_4$ -BASED THIN-FILM SOLAR CELLS WITH CDS, ZNSE, ZNS WINDOW AND ITO, ZNO CHARGE-COLLECTING LAYERS

M.Sc. Dobrozhan O.A.*, St. Danylchenko P.S., Ph.D. Grynenko, V.V., Prof. Dr. Opanasyuk A.S.**

Department of Electronics and Computer Technology

Sumy State University, 2, Rymyskogo-Korsakova st., 40007 Sumy, Ukraine

*dobrozhan.a@gmail.com, **opanasjuk_sumdu@ukr.net

Abstract: We reported the investigation the effect of the optical and recombination losses in solar cells (SCs) based on $n\text{-CdS}(\text{ZnSe}, \text{ZnS})/p\text{-Cu}_2\text{ZnSn}(\text{S,Se})_4$ heterojunctions with $n\text{-ITO}(\text{ZnO})$ frontal charge-collecting contacts on the internal (Q_{int}), external (Q_{ext}) quantum yields, short-circuit current density (J_{sc}) and maximum efficiency (η) of solar cells

KEYWORDS: $\text{Cu}_2\text{ZnSnS}_4$ FILMS; OPTICAL LOSSES; RECOMBINATION LOSSES; QUANTUM YIELD; SHORT-CIRCUIT CURRENT; EFFICIENCY.

1. Introduction

The future forecast for the renewable energy demonstrates that the solar power will become the dominant energy source from the middle of the 21st century. One of the most perspective routes to utilize the solar energy is its conversion into electricity by using the solar cells (SCs).

Nowadays, the silicon-based SCs are the most commercialized and widespread technology. But alternative are thin-film SCs based on $\text{CdS}/p\text{-CdTe}$ heterojunction (HJ) with $n\text{-ITO}$ frontal charge-collecting contact have also been widely spread on the photovoltaic market [1, 2]. However, the shortcomings, such as toxicity of Cd, high price and low abundance of In and Te, give rise to the search for the alternative materials to the functional layers in the photovoltaic devices. Nowadays, the compound $\text{Cu}_2\text{ZnSn}(\text{S,Se})_4$ (CZTSSe) is regarded as a promising substitute for the traditional Si, CdTe, $\text{Cu}(\text{In,Ga})(\text{S,Se})_2$ absorber layers. CZTSSe possesses the controlled band gap for the solar light absorption ($E_g^{\text{CZTSSe}} = 1.0\text{-}1.5$ eV), p -type conductivity and high absorption coefficients ($\alpha \sim 10^4\text{-}10^5 \text{ cm}^{-1}$) [3, 4]. The promising alternatives for the well-known SCs are considered the structure with ZnSe or ZnS window layers and ZnO charge collecting contact [5-7]. These structures contain only the abundant and non-toxic elements. ZnS, ZnO, ZnSe ($E_g^{\text{ZnS}} = 3.68$ eV, $E_g^{\text{ZnO}} = 3.37$ eV, $E_g^{\text{ZnSe}} = 2.67$ eV) are the wide-band gap semiconductors allowing to increase the number of the photons incoming to CZTSSe absorber layer.

According to Shockley-Queisser analysis, the maximum theoretical efficiency of the thin-film SCs with CZTSSe absorber layer is about (32-34) % [3, 8]. However, the experimental efficiency of CZTSSe SCs is only 12.6 % [9, 10]. The difference between the theoretical and experimental values of the efficiency can be explained by the optical, electrical and recombination losses which take place during the conversion of the solar energy into electricity.

The enhancement of the SC efficiency might be achieved by the minimization of the described losses by using the optimized structures based on the functional layers with the improved characteristics.

The foregoing discussions identified the main goal of this work – to determine and compare the optical and recombination losses in the SCs based on $n\text{-CdS}(\text{ZnSe}, \text{ZnS})/p\text{-CZTSSe}$ HJs with $n\text{-ITO}(\text{ZnO})$ frontal charge-collecting contacts.

2. Methodologies

The light flow, before reaching CZTSSe absorber layer in which the generation of the electron-hole pairs is occurred, passes through the $\text{ITO}(\text{ZnO})$ and $\text{CdS}(\text{ZnSe}, \text{ZnS})$ layers of the SCs. Herewith, the optical losses of energy as a consequence of the light

reflection from the air/ $\text{ITO}(\text{ZnO})$, $\text{ITO}(\text{ZnO})/\text{CdS}(\text{ZnSe}, \text{ZnS})$, $\text{CdS}(\text{ZnSe}, \text{ZnS})/\text{CZTSSe}$ interfaces and the light absorption of $\text{ITO}(\text{ZnO})$, $\text{CdS}(\text{ZnSe}, \text{ZnS})$ are observed.

The reflection coefficients at the interfaces of the contacted layers can be determined by using Fresnel equation [11]:

$$R = \left(\frac{n_i - n_j}{n_i + n_j} \right)^2 \quad (1)$$

where n_i, n_j – refractive indices of the first and second contacted materials, respectively.

In the case of the electrically conductive material, the reflection coefficients might be calculated by using the following expression [12]:

$$R_{ij} = \frac{|n_i^* - n_j^*|}{|n_i^* + n_j^*|} = \frac{(n_i - n_j)^2 + (k_i - k_j)^2}{(n_i + n_j)^2 + (k_i + k_j)^2} \quad (2)$$

where n_i^*, n_j^* – the complex refractive indices; k_i, k_j – the extinction coefficients.

The spectral dependencies of n and k were taken from the literature data on the refractive and extinction coefficients of ITO, ZnO, CdS, ZnSe, ZnS, CZTS [3, 13, 14]. It was assumed that the air has $n = 1$ and $k = 0$.

The transmission coefficients taking into account both the light reflection and absorption of the charge-collecting and window layers can be calculated using the expression [11, 15]:

$$T(\lambda) = (1 - R_{12})(1 - R_{23})(1 - R_{34})(1 - R_{45}) \exp(-\alpha_1 d_1) \exp(-\alpha_2 d_2) \quad (3)$$

where α_1, α_2 – the absorption coefficients of the charge-collecting and window layers; d_1, d_2 – the charge-collecting and window layer thicknesses.

The absorption coefficients of the materials $\alpha(\lambda)$, considering the extinction coefficient $k(\lambda)$, can be calculated by using the following equation [11]:

$$\alpha(\lambda) = \frac{4\pi}{\lambda} k \quad (4)$$

The modeling of the light reflection and absorption processes in the multilayer structures was carried out by using the different thicknesses of the window, $d_{\text{CdS}(\text{ZnSe}, \text{ZnS})} = (25\text{-}100)$ nm, and frontal charge-collecting, $d_{\text{ITO}(\text{ZnO})} = (100\text{-}200)$ nm, layers. These thickness values of the layers are typical for the practical SCs.

The important parameter for the analysis of the recombination losses in the SCs is the width of space charge region (w), in other words, the depletion region, occurring at the interface between the heteropairs, where the electrical field is acting as a separator for the photogenerated electron-hole pairs.

This width mainly depends on the concentration of uncompensated acceptors ($N_a - N_d$) (i.e., the difference between the acceptor and donor concentrations), locating in the semiconductor materials, and the contact barrier height. However, the latter value for the investigated junctions, unfortunately, was not known. This problem was solved by means of the construction of the energy band diagrams of the HJs.

It was considered that the small amount of the surface states exists at the interface between heteropairs. At the same time, the charge transport mechanism was described accordingly to Anderson model.

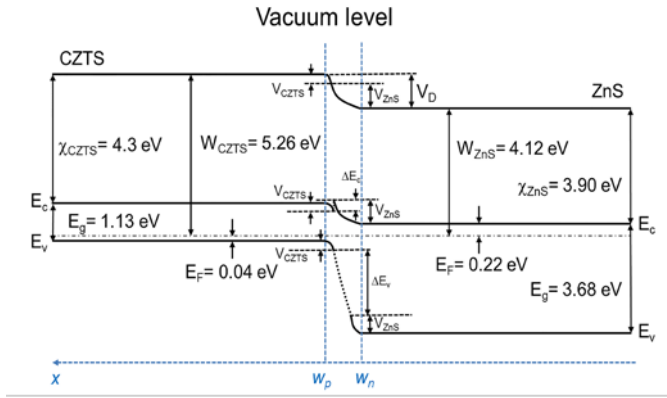


Fig. 1. Energy band diagrams of n-ZnS/p-CZTS HJ.

Unlike the fact that the charge transport processes at n-CdS/p-CdTe HJ are analogous to those occurring in the Schottky diodes [1], the same charge transport mechanisms for n-CdS(ZnSe, ZnS)/p-CZTS heterosystems are not acceptable. It is due to the fact that the doping levels in CZTS ($N_a = 10^{17}-10^{18} \text{ cm}^{-3}$ [16]) are higher than in CdTe material ($N_a = 10^{14}-10^{17} \text{ cm}^{-3}$ [17]) and even higher than in the window materials ($N_d = 10^{16}-10^{17} \text{ cm}^{-3}$ [14]). It means that SCR is located both in the window (w_n) and absorber (w_p) layers, and SCR width can be determined by the equations [18]:

$$w = \sqrt{\frac{2\varepsilon_n\varepsilon_p\varepsilon_0(V_D - qU)}{q^2} \left(\frac{1}{\varepsilon_n N_D} + \frac{1}{\varepsilon_p N_A} \right)} \quad (5)$$

where ε_n , ε_p – the relative permittivity of the window and absorber materials; ε_0 – the vacuum permittivity; $V_D = qV_{bi}$ – the contact barrier height (V_{bi} – the built-in potential); U – the applied external voltage; q – the elementary charge; N_A , N_D – the concentration of uncompensated acceptors and donors in the absorber and window layers.

Kosyachenko et al. showed that the solution of the continuity equation is effective for the determination of the drift component of the internal quantum yield (Q_{drift}) of the SC, while taking into account the recombination at the HJ interface and in SCR, by using the following equation [11, 17]:

$$Q_{drift\ p(n)} = \frac{1 + \frac{S}{D_{p\ p(n)}} \left(\alpha_{p(n)} + \frac{2 \cdot (V_D - qU)}{w_{p(n)} \cdot kT} \right)^{-1}}{1 + \frac{S}{D_{p\ p(n)}} \left(\frac{2 \cdot (V_D - qU)}{w_{p(n)} \cdot kT} \right)^{-1}} - \frac{\exp(-\alpha_{p(n)} w_{p(n)})}{1 + \alpha_{p(n)} \cdot L_{n\ p(p\ n)}} \quad (6)$$

where S – the recombination velocity of the charge carriers at the HJ interface and in SCR; $D_{p\ p(n)}$ – the diffusion coefficients of the holes (electrons) in the absorber (window) layers; $\alpha_{p(n)}$ – the light absorption coefficients of the absorber (window) layer; k – the Boltzmann constant; T – the temperature; $L_{n\ p(p\ n)}$ – the diffusion

length of the electrons (holes) in the absorber (window) layer ($L_{n(p)} = (\tau_{n(p)} D_{n(p)})^{1/2}$, where $\tau_{n(p)}$ – the lifetime of the electrons (holes), $D_{n(p)}$ – the diffusion coefficients of the electrons (holes) in the relevant layers).

It should be noted that the equation (6) does not consider the recombination in the quasineutral regions of the window and absorber materials and on the back surface of CZTS layer. To account these losses, the diffusion ($Q_{diff\ p(n)}$) component of the quantum yield can be evaluated by the following equation [17]:

$$Q_{diff\ p(n)} = \frac{(\alpha_{p(n)} L_{n\ p(p\ n)} / (D_{p\ p(n)} \cdot L_{n\ p(p\ n)} - 1)) \exp(-\alpha_{p(n)} w_{p(n)}) + \sinh((d_{p(n)} - w_{p(n)}) / L_{n\ p(p\ n)}) + \alpha_{p(n)} L_{n\ p(p\ n)} \exp(-\alpha_{p(n)} (d_{p(n)} - w_{p(n)}))}{(S_{n\ p(p\ n)} / D_{p\ p(n)}) \cosh((d_{p(n)} - w_{p(n)}) / L_{n\ p(p\ n)}) - \exp(-\alpha_{p(n)} (d_{p(n)} - w_{p(n)})) + \sinh((d_{p(n)} - w_{p(n)}) / L_{n\ p(p\ n)}) + \alpha_{p(n)} L_{n\ p(p\ n)} \exp(-\alpha_{p(n)} (d_{p(n)} - w_{p(n)}))} \quad (7)$$

where $d_{p(n)}$ – the thicknesses of the absorber and window layers; S_b – the recombination velocity in the quasineutral regions and on the back surface of the absorber layer.

The total internal quantum yield of the SCs is easy to determine as the sum of all quantum yields, considering the directions of the drift and diffusion currents in the space charge and quasineutral regions. The account of the optical losses owing to the reflection and absorption of the light by the auxiliary layers (ITO, CdS, ZnSe, ZnS) of the SCs gives the opportunity to determine the external quantum yield (Q_{ext}) of the device [11, 17]:

$$Q_{ext} = T(\lambda) Q_{int} \quad (8)$$

The optical losses described in the previous subsections are important for the analysis of the SC efficiency. As a consequence of its consideration, we built the spectral dependencies of the external quantum yield (Q_{ext}) for the investigated SCs. The calculations were carried out using the following physical values: $N_a = 10^{18} \text{ cm}^{-3}$, $N_d = 10^{17} \text{ cm}^{-3}$, $d_{ITO(ZnO)} = 100 \text{ nm}$, $d_{CdS(ZnSe, ZnS)} = 25 \text{ nm}$, $d_{CZTS} = 1 \mu\text{m}$. The concentrations of uncompensated acceptors and donors coincide with SCR widths which are close to the device thicknesses. At the same time, the thicknesses of all functional layers were taken close to those used in the practical SCs [2].

The short-circuit current density (J_{sc}) of the SCs was determined using the well-known formula:

$$J_{sc} = q \sum_i T(\lambda) \frac{\Phi_i(\lambda_i)}{h\nu_i} Q_{int}(\lambda_i) \Delta\lambda_i \quad (9)$$

where $\Phi_i(\lambda_i)$ – the spectral power density of the solar radiation; $\Delta\lambda_i$ – the interval between neighboring values of the wavelength; $h\nu_i$ – the photon energy.

The calculation of J_{sc} was carried out under AM 1.5G radiation conditions [19]. Herewith, the maximum short-circuit current density ($J_{max\ sc}$) can be obtained by neglecting the light losses owing to the absorption of the auxiliary layers, i.e. $T(\lambda) = 1$, and under the circumstance that every photon generates the electron-hole pair which reaches the charge-collecting contacts without recombination, i.e. $Q_{ext}(\lambda) = 1$. It was established, that the maximum value of the short-circuit current density of the investigated SCs is equal to $J_{max\ sc} = 34.82 \text{ mA/cm}^2$.

The solar cell efficiency (η) is determined by the well-known equation [10, 29, 38]:

$$\eta = \frac{U_{oc} \cdot J_{sc} \cdot FF}{P_{in}} \quad (10)$$

where U_{oc} – the open-circuit voltage; J_{sc} – the short-circuit current density; FF – the fill factor; P_{in} – the input power (100 mW/cm^2 , illumination AM 1.5G).

To determine the effect of the optical and recombination losses on the maximum efficiencies of the SCs with ITO(ZnO)/CdS(ZnSe, ZnS)/CZTS structures, the values of open-circuit voltage were taken as those that coinciding with the height of

the contact potential differences at the HJs: $U_{oc} = (0.72 \text{ V})_{\text{CdS}}, (1.07 \text{ V})_{\text{ZnSe}}, (1.14 \text{ V})_{\text{ZnS}}$, and the values of the fill factor that matching the maximum possible $FF = 89 \%$ [5]. Accordingly, it was found that the maximum efficiency of the single junction SC was 33.5% [5].

3. Results and discussions

The analysis of the optical losses owing the light reflection and absorption of the window and charge-collecting layers showed that, as it was expected, the replacement of the traditional window material (CdS) with wide band gap materials (ZnSe, ZnS) caused the increase of the transmission coefficients of the multilayer structures, primarily, in the short wave region with $d_{\text{CdS}}(\text{ZnSe}, \text{ZnS}) = (25-100) \text{ nm}$. This tendency was valid for applying ITO and ZnO layers with $d_{\text{ITO}(\text{ZnO})} = (100-200) \text{ nm}$ as the charge-collecting contacts.

ZnO layer is more attractive than ITO because it improves the light transmission coefficients toward CZTS absorber layer regardless of the considered window materials.

However, it should be noted that the values of T for the best and worst structures differed only in (5.2-13.5) %.

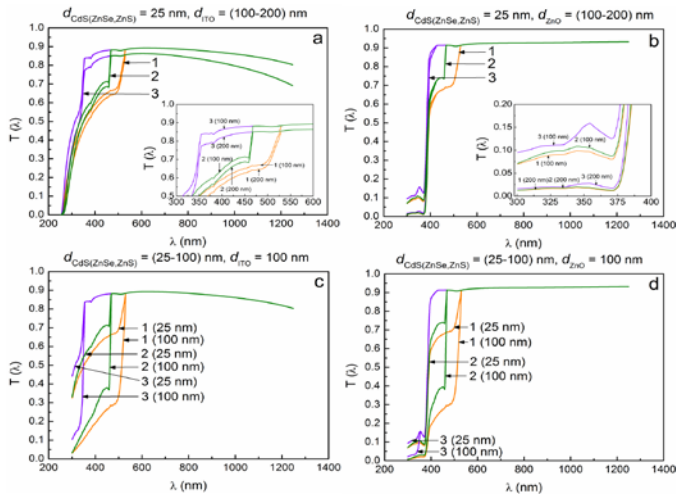


Fig.2. Spectral dependencies of the transmission coefficients of the SCs with the structures ITO/CdS/CZTS (1), ITO/ZnSe/CZTS (2), ITO/ZnS/CZTS (3) (a, c); ZnO/CdS/CZTS (1), ZnO/ZnSe/CZTS (2), ZnO/ZnS/CZTS (3) (b, d) with the different thicknesses of the window and charge-collecting layers. The light reflection from the interfaces and absorption of the auxiliary layers were taken into account.

It was established that the increase of the donor concentration in the window material at the constant values of N_a in the absorber layer resulted in the increase of the quantum efficiency of the SC based on n -CdS/ p -CZTS HJ in the photosensitive region for both CZTS and CdS materials. However, this increase had a weak influence on the internal quantum yield in the photosensitive region of the window materials in the SCs based on n -(ZnSe, ZnS)/ p -CZTS HJs. For the investigated HJs, the increase of the donor concentration caused the increase of the quantum yields in both middle and long wavelength regions, due to the extension of SCR in the absorber layer, and, as a consequence, reduced impact of the diffusion component on the total photocurrent (J_{ph}).

The analysis of the obtained dependencies showed that the values of Q_{ext} of the SC based on n -ZnS/ p -CZTS HJ were slightly higher than those of the structure with CdS and ZnSe window layers regardless of the material of the charge-collecting contacts. Thus, as it was expected, SCs with the window layers, which possess the higher values of the band gap, demonstrated the higher quantum yields.

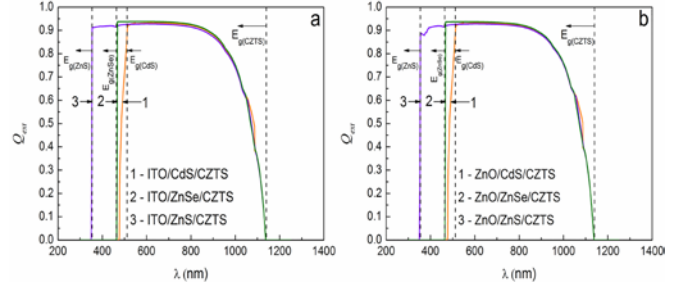


Fig. 3. Spectral dependencies of the external quantum yield (Q_{ext}) of the SCs based on n -CdS(ZnSe, ZnS)/ p -CZTS HJs with ITO (a) and ZnO (b) charge-collecting layers with: $N_a = 10^{18} \text{ cm}^{-3}$, $N_d = 10^{17} \text{ cm}^{-3}$, $d_{\text{ITO}}(\text{ZnO}) = 100 \text{ nm}$, $d_{\text{CdS}}(\text{ZnSe, ZnS}) = 25 \text{ nm}$, $d_{\text{CZTS}} = 1 \mu\text{m}$.

However, it should be noted that we neglected the inequality state at the interfaces of the different HJs. However, in reality, the mismatch density of the dislocations at the interfaces of the considered HJs is varied.

The dependencies of the SC efficiencies (η) on the thicknesses of the window (CdS, ZnSe, ZnS) and charge-collecting (ITO, ZnO) layers are presented in Fig. 3. As can be seen from Fig. 4, the best devices,

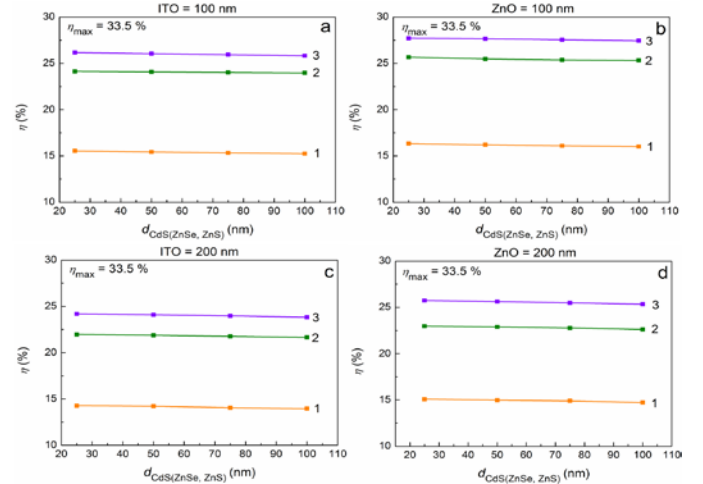


Fig.4. The effect of the optical and recombination losses on the efficiency of the SCs based on CdS/CZTS (1), ZnSe/CZTS (2), ZnS/CZTS (3) HJs with the variable thicknesses of the window layers and two constant thicknesses of the charge-collecting contacts: 100 nm (a, b) and 200 nm (c, d).

among the investigated SC structures, contain ZnS window layer ($\eta = 23.8-27.7 \%$), and the highest values of the efficiency were demonstrated by a device with ZnO/ZnS/CZTS structure ($\eta \sim 28 \%$ with $d_{\text{ZnO}} = 100 \text{ nm}$, $d_{\text{ZnS}} = 25 \text{ nm}$).

It should be mentioned, that the efficiency of the well-known SC with ITO/CdS/CZTS structure was about (13.9-15.5) %. These values are well correlated with the results obtained for the best SC with the analogous structure ($\eta = 12.6 \%$) [9]. The SCs with ZnSe window layer showed the quite high efficiencies as well, $\eta = (21.7-25.7) \%$.

4. Conclusions

It was found that, under the consideration of the losses owing to the reflection and absorption of the auxiliary layers of devices, the values of J_{sc} of the SCs with the ZnO/ZnS/CZTS ($d_{\text{ZnS}} = (25-100) \text{ nm}$, $d_{\text{ITO}(\text{ZnO})} = 100 \text{ nm}$) structure were higher in (3.06-3.27) mA/cm^2 than those obtained for devices with ITO/CdS/CZTS structure in the overall interval of the thickness alteration. The increase of the charge-collecting layer thickness up to 200 nm led to the decrease of J_{sc} and the difference between the best (ZnO/ZnS/CZTS) and worst (ITO/ZnSe/CZTS) structures of the SCs was found to be $\sim 3.15 \text{ mA/cm}^2$. It should be noted that the

optical and recombination losses caused the decrease of J_{sc} by (21.5-37.4) %.

The best devices, among the investigated SC structures, contain ZnS window layer ($\eta = 23.8\text{-}27.7\%$), and the highest values of the efficiency were demonstrated by the device with ZnO/ZnS/CZTS structure ($\eta \sim 28\%$ with $d_{ZnO} = 100\text{ nm}$, $d_{ZnS} = 25\text{ nm}$). The SCs with ZnSe window layer showed the quite high efficiency as well, $\eta = (21.7\text{-}25.7)\%$. It should be mentioned, that the efficiency of the well-known SC with ITO/CdS/CZTS structure was about (13.9-15.5) %. These values are well correlated with the results obtained for the best SC with the analogous structure ($\eta = 12.6\%$).

The presented results show the maximum values of the efficiencies of the SCs based on $n\text{-CdS}(\text{ZnSe}, \text{ZnSe})/p\text{-CZTS}$ HJs and open the way for the optimization of the practical thin film SCs.

5. References

- [1] S. M. Sze, and K. K. Ng, *Physics of Semiconductor Devices*, John Wiley & Sons, Hoboken (2006).
- [2] S. Girish Kumar, and K. S. R. Koteswara Rao. *Physics and Chemistry of CdTe/CdS Thin Film Heterojunction Photovoltaic Devices: Fundamental and Critical Aspects*, *Energy Environ. Sci.* vol. 7, 45 (2014).
- [3] K. Ito, *Copper Zinc Tin Sulfide-Based Thin Film Solar Cells*, John Wiley & Sons, Chichester (2015).
- [4] M. P. Suryawanshi, G. L. Agawane, S. M. Bhosale, S. W. Shin, P. S. Patil, J. H. Kim, and A. V. Moholkar. CZTS Based Thin Film Solar Cells: A Status Review. *Mater. Technol.* vol. 28, 98 (2013).
- [5] M. Nguyen, K. Ernits, K. F. Tai, C. F. Ng, S. S. Pramana, W. A. Sasangka, S. K. Batabyal, T. Holopainen, D. Meissner, and A. Neisser. ZnS Buffer Layer for $\text{Cu}_2\text{ZnSn}(\text{SSe})_4$ Monograin Layer Solar Cell. *Sol. Energy.* vol. 111, 344 (2015).
- [6] H. Katagiri, K. Saitoh, T. Washio, H. Shinohara, T. Kurumadani, and S. Miyajima. Development of Thin Film Solar Cell Based on $\text{Cu}_2\text{ZnSnS}_4$ Thin Films. *Sol. Energy Mater. Sol. Cells.* vol. 65, 141 (2001).
- [7] O. Dobrozhan, A. Opanasyuk, M. Kolesnyk, M. Demydenko, and H. Cheong. Substructural Investigations, Raman, and FTIR Spectroscopies of Nanocrystalline ZnO Films Deposited by Pulsed Spray Pyrolysis. *Phys. Status Solidi A.* vol. 212, 2915 (2015).
- [8] A. Polman, M. Knight, E. C. Garnett, B. Ehrler, and W. C. Sinke. Photovoltaic Materials: Present Efficiencies and Future Challenges. *Science.* vol. 352, aad4424 (2016).
- [9] W. Wang, M. T. Winkler, O. Gunawan, T. Gokmen, T. K. Todorov, Y. Zhu, and D. B. Mitzi. Device Characteristics of CZTSSe Thin-Film Solar Cells with 12.6% Efficiency. *Adv. Energy Mater.* vol. 4, 1301465 (2014).
- [10] M. A. Green, K. Emery, Y. Hishikawa, W. Warta, and E. D. Dunlop. Solar Cell Efficiency Tables (Version 48). *Prog. Photovoltaics.* vol. 24, 905 (2016).
- [11] L. A. Kosyachenko, E. V. Grushko, and X. Mathew. Quantitative Assessment of Optical Losses in Thin-Film CdS/CdTe Solar Cells. *Sol. Energy Mater. Sol. Cells.* vol. 96, 231 (2012).
- [12] L. A. Kosyachenko, X. Mathew, P. D. Paulson, V. Ya. Lytvynenko, and O. L. Maslyanchuk. Optical and Recombination Losses in Thin-Film $\text{Cu}(\text{In,Ga})\text{Se}_2$ Solar Cells. *Sol. Energy Mater. Sol. Cells.* vol. 130, 291 (2014).
- [13] S. O. Kasap, *Optoelectronics and Photonics: Principles and Practices*, Prentice Hall, Upper Saddle River (2001).
- [14] S. Adachi, Editor, *Handbook on Physical Properties of Semiconductors*, Kluwer Academic Publishers, Berlin (2004).
- [15] T. Mykytyuk, V. Y. Roshko, L. Kosyachenko, and E. Grushko. Limitations on Thickness of Absorber Layer in CdS/CdTe Solar Cells. *Acta Phys. Pol. A.* vol. 122, 1073 (2012).
- [16] J. P. Teixeira, R. A. Sousa, M. G. Sousa, A. F. da Cunha, P. A. Fernandes, P. M. P. Salomé, and J. P. Leitão. Radiative Transitions in Highly Doped and Compensated Chalcopyrites and Kesterites: The Case of $\text{Cu}_2\text{ZnSnS}_4$. *Phys. Rev. B.* vol. 90, 235202 (2014).
- [17] L. A. Kosyachenko. Problems of Efficiency of Photoelectric Conversion in Thin-Film CdS/CdTe Solar Cells. *Semiconductors.* vol. 40, 710 (2006).
- [18] M. Balkanski, and R. F. Wallis, *Semiconductor Physics and Applications*, Oxford University Press, Oxford (2000).
- [19] Reference Solar Spectral Irradiance at the Ground at Different Receiving Conditions, Part 1: Direct Normal and Hemispherical Solar Irradiance for Air Mass 1.5. Standard of International Organization for Standardization, ISO 9845-1:1992.

DECISION OF OPTIMIZATION PROBLEMS USING SYMMETRIC ALGORITHM OF HEAVY BALL METHOD

Kharlamova Y.N.
National Mining University, Dnipro, Ukraine
E-mail: harlamyshka@gmail.com

Abstract: The algorithm of the heavy ball method, based on the principle of symmetry, to find a global extremum is described. The computer simulation of the method for the three test functions (Ackley, Griewank and Schwefel) is carried out. The results of the study of the efficiency of this algorithm are given. The results of mathematical modeling in the graphs, describing the process of convergence of representative points to the global optimum point of test functions, are shown. Conclusions about the efficacy of the described algorithm applied to optimization problems are drawn.

KEYWORDS: HEAVY BALL METHOD, ENERGY INTERACTION BETWEEN THE TWO BALLS, EXTREMUM SEEKING PROCESS, CONCEPT OF SYMMETR

Y.

1. Introduction

One of the main factors that must be considered in solving the problems of synthesis of modern adaptive identification systems and information measuring systems is the inertia of the control object and measuring equipment. The creators of these systems use various optimization methods to ensure high efficiency of taken decisions.

In most cases, the seeking global extremum of multiextremal objective functions is carried out dynamically using the heavy ball method. But a moving heavy ball can have both a lack of kinetic energy and an excess of it. In the first case, it can stop at the one of the local extremes, before reaching the global extremum, and in the second case - to jump over it.

In the current situation, it is actual to organize the boosting of additional kinetic energy of the ball moving to the global extremum or to remove its excess.

2. Analysis of the literature sources

Today, both global optimization algorithms for solving a separate class of problems [1] and for more universal ones have been developed. When considering dynamic optimization problems, the relaxation method [2,3] is often applied, the realization of which is carried out by means of nonstationary processes that are described by vector differential equations of the form

$$(1) m \cdot x^{(2)}(t) + r \cdot x^{(1)}(t) + \text{grad} f(x) = 0, \quad m > 0, \quad r > 0;$$

$$(2) \frac{dx}{dt} + k \cdot \text{grad} f(x) = 0, \quad k > 0.$$

These processes to solve the task are eventually established.

The equation (1) describes the *heavy ball method*, which (with an appropriate choice of the parameters m and r) is referred to methods of seeking for a global extremum. The relaxation method, realized according to equation (2), is called the *steepest gradient descent method*, it is usually referred to methods of seeking for a local extremum of the function.

It is known that surface elongation of the objective function along one of the directions and its complex relief sharply worsens the effectiveness of these methods. The heavy ball method and the steepest gradient descent method in solving seeking global extremum problems, with an increase of the oscillations amplitude do not give a positive result, and the process of motion of the representing point stops at the first local extremum.

The purpose of the article is to construct algorithms for seeking for a global extremum of functions based on the principle of symmetry of the interaction of two heavy balls and to justify the advantage of this algorithm in seeking for a global extremum of the multiextremal function in comparison with other relaxation algorithms.

3. Results and discussion

The problem of seeking for the minimum of a multiextremal function can be solved by using the concept of symmetry, which has proven itself in such one-dimensional optimization methods such as methods of dichotomy, Fibonacci and golden section. In these methods, two *representative* points move symmetrically to the extremum of the function, significantly reducing the interval of uncertainty (localization).

Let us consider multidimensional methods improving process of function extremum seeking by applying the concept of symmetry [4,5].

Let us represent the expression of a downward-convex function $f(x)$ (x is a vector argument), which extremum is sought, in the form

$$(3) f(x) = 0.5 \left((x-x)^T Q(x-x) + f(x) + f(x) \right),$$

where Q is a positive definite symmetric matrix.

Then, replacing one of the vectors x with the vector y and the other with the vector z in the expression (3), we obtain an auxiliary function

$$(4) F(y, z) = 0.5 \left[(y-z)^T Q(y-z) + f(y) + f(z) \right],$$

the extremum of which will take place under the condition that $y=z=x^*$, where x^* is the value of the vector argument at which the function $f(x)$ takes an extreme value.

The motion to a minimum of the auxiliary function $F(y, z)$ is ensured as a result of a simultaneous coherent change of vector arguments y and z by any of the known extremum seeking algorithms.

The algorithm (1) of a heavy ball method [6,7] when working with an auxiliary function $F(y, z)$ will be such as

$$(5) \begin{aligned} m \frac{d^2 y}{dt^2} + r \frac{dy}{dt} + \frac{\partial F(y, z)}{\partial y} &= 0, \\ m \frac{d^2 z}{dt^2} + r \frac{dz}{dt} + \frac{\partial F(y, z)}{\partial z} &= 0. \end{aligned}$$

It should be noted that for the extremum seeking of the auxiliary function (5), both continuous and discrete algorithms of several converging points can be used. This allows them to overcome local extremes.

Let us consider the efficiency of the heavy ball method, based on the principle of symmetry, on the example of three standard test functions: Ackley, Griewank and Schwefel [8].

The Ackley function (Fig. 1, a) has many local extremums near the global optimum. On the interval $[-7; 3]$, the function takes a minimum value at the point 0, where $x=0$ and corresponds to the following description

$$(6) f(x) = 20 + e - 20 \cdot \exp(-0.2 \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}) - \exp(\frac{1}{n} \sum_{i=1}^n \cos(2\pi x_i)).$$

Then the auxiliary symmetric function $F(y,z)$ according to the equation (4) for the Ackley function (6) will have the following form:

$$(7) F(y,z) = 0.5(20 + e) - 20 \exp(-0.2 y^2) - \exp(\cos(2\pi y)) - 20 \exp(-0.2 z^2) - \exp(\cos(2\pi z)) + 0.5q(y-z)^2,$$

and the corresponding algorithm (5) for the function (7)

$$(8) \begin{cases} y_1'(t) = y_2, \\ y_2'(t) = -\frac{r}{m} y_2 - \frac{1}{m} \left[q(y_1 - z_1) + 2y_1 \sqrt{y_1^2} \exp(-0.2 \sqrt{y_1^2}) + \pi \sin(2\pi y_1) \exp(\cos(2\pi y_1)) \right], \\ z_1'(t) = z_2, \\ z_2'(t) = -\frac{r}{m} z_2 - \frac{1}{m} \left[q(z_1 - y_1) + 2z_1 \sqrt{z_1^2} \exp(-0.2 \sqrt{z_1^2}) + \pi \sin(2\pi z_1) \exp(\cos(2\pi z_1)) \right]. \end{cases}$$

where $y_1 = y, \quad z_1 = z$.

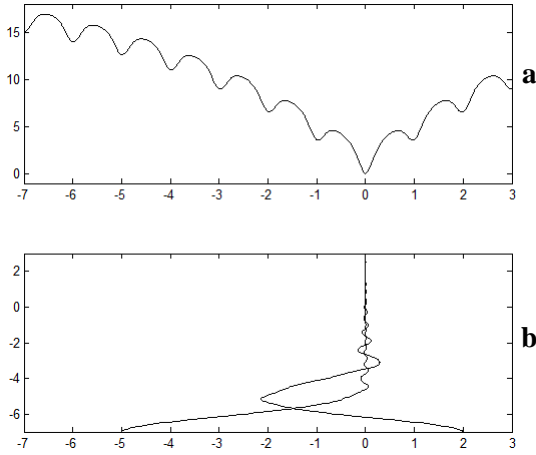


Fig.1. Graphs illustrating the process of the global extremum seeking of a function with use of the concept of symmetry: a - the type of the Ackley test function; b - solution of the system (8)

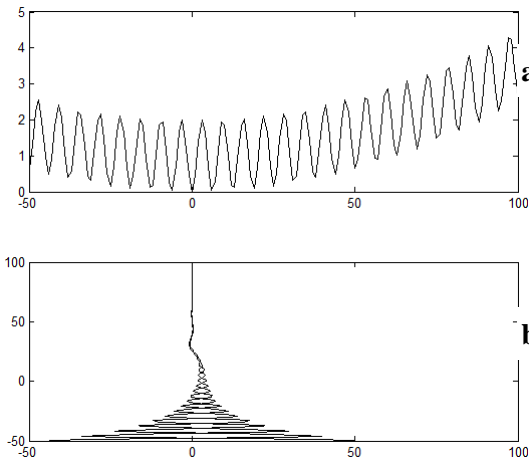


Fig.2. Graphs illustrating the process of the global extremum seeking of a function with the use of the concept of symmetry: a - the type of the Griewank test function; b - solution of the system (11)

For the solution of the system (8) initial conditions were set: $y_1(0) = -5, z_1(0) = 2$ and the following parameters were chosen: a mass of a heavy ball $m = 1.5$, damping coefficient $r = 2$ and $q = 2$. Fig. 1(b) shows the result of the solution of the system (8), which describes the process of convergence of the *representative* points to the point of the global optimum ($x^* = 0$). The following values are obtained: $y_1 = 0.000002$ and $z_1 = -0.001472$. The solution of the problem (6) with the usual heavy ball method according to the expression (1), with constant values of the parameters m and r , gave the following results: $x = -4.9862$ (from the initial point $x(0) = -5$) and $x = 1.9745$ (from the point $x(0) = 2$).

There are many local minimums in the *Griewank* function (Fig. 2, a). On the interval $[-50; 100]$, the function takes a minimum value at the point 0, where $x = 0$ and has such description

$$(9) f(x) = 1 + \sum_{i=1}^n \frac{x_i^2}{4000} - \prod_{i=1}^n \cos\left(\frac{x_i}{\sqrt{i}}\right).$$

Then the auxiliary symmetric function $F(y,z)$ for the *Griewank* function (9) is written so

$$(10) F(y,z) = 0.5(2 + \frac{y^2}{4000} - \cos(y) + \frac{z^2}{4000} - \cos(z)) + 0.5q(y-z)^2$$

and the corresponding algorithm (5) for the function (10)

$$(11) \begin{cases} y_1'(t) = y_2; \\ y_2'(t) = -\frac{r}{m} y_2 - \frac{1}{m} \left[q(y_1 - z_1) + \frac{y_1}{4000} + 0.5 \sin(y_1) \right], \\ z_1'(t) = z_2; \\ z_2'(t) = -\frac{r}{m} z_2 - \frac{1}{m} \left[q(z_1 - y_1) + \frac{z_1}{4000} + 0.5 \sin(z_1) \right]. \end{cases}$$

For the solution of the system (11) initial conditions were set: $y_1(0) = -47, z_1(0) = 53$ and the following parameters were chosen: a mass of a heavy ball $m = 7$, damping coefficient $r = 1$ and $q = 2$. Fig. 2(b) shows the result of the solution of the system (11), which describes the process of convergence of the *representative* points to the point of the global optimum ($x^* = 0$). The calculated values were: $y_1 = -0.000254$ and $z_1 = -0.000862$. The solution of the problem (9) with the usual heavy ball method according to the expression (1) gave the following results: $x = -43.9603$ (from the initial point $x(0) = -47$) and $x = 50.2403$ (from the point $x(0) = 53$).

The *Schwefel* function (Fig. 3(a)) as well as the two above-mentioned functions, has a local minimums near global optimum. On the interval $[-500; 500]$, the function takes a minimum value at the point 0, where $x = 420$ and corresponds to the following description

$$(12) f(x) = 418.9829 \cdot n + \sum_{i=1}^n (-x_i \cdot \sin(\sqrt{|x_i|})).$$

For the *Schwefel* function (12), the auxiliary symmetric function $F(y,z)$ will be

$$(13) F(y,z) = 0.5(2 \cdot 418.9829 - y \sin(\sqrt{|y|}) - z \sin(\sqrt{|z|})) + 0.5q(y-z)^2,$$

and the corresponding algorithm (5) for the function (13)

$$(14) \begin{cases} y_1'(t) = y_2; \\ y_2'(t) = -\frac{r}{m} y_2 - \frac{1}{m} \left[q(y_1 - z_1) - 0.5 \sin(\sqrt{|y_1|}) - \frac{y_1 \cos(\sqrt{|y_1|}) \cdot \text{sign}(y_1)}{4\sqrt{|y_1|}} \right]; \\ z_1'(t) = z_2; \\ z_2'(t) = -\frac{r}{m} z_2 - \frac{1}{m} \left[q(z_1 - y_1) - 0.5 \sin(\sqrt{|z_1|}) - \frac{z_1 \cos(\sqrt{|z_1|}) \cdot \text{sign}(z_1)}{4\sqrt{|z_1|}} \right]. \end{cases}$$

For the solution of the system (14) initial conditions were set: $y_1(0) = -200, z_1(0) = 480$ and the following parameters were chosen: a mass of a heavy ball $m = 57.8$, damping coefficient $r = 0.67$ and $q = 2$. Fig. 3(b) shows the result of the solution of the system (14), which describes the process of convergence of the *representative* points to the point of the global optimum ($x^* = 420$). The following values are obtained: $y_l = 420.0194$ and $z_l = 420.0862$. The solution of the given problem (12) by the heavy ball method according to the equation (1) from the initial point $x(0) = -4.5$ did not give a positive result, the representing point completed its motion at $x = -124.8345$, from the point $x(0) = 3.75$ the global optimum point is found with a low accuracy $x = 420.8741$.

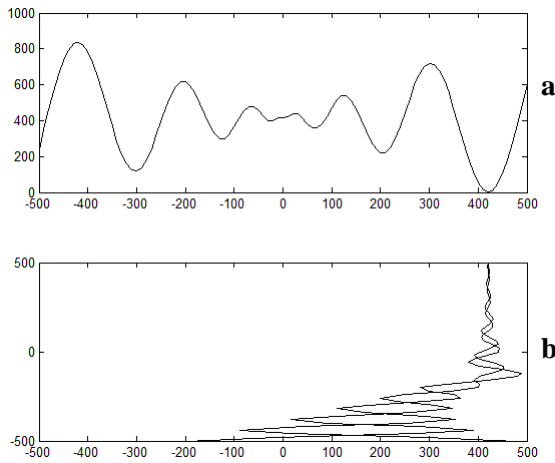


Fig.3. Graphs illustrating the process of the global extremum seeking of a function with the use of the concept of symmetry: a - the type of the Schwefel test function; b - solution of the system (14)

Let us analyze the work of the heavy ball method by comparing it with the method of coordinate descent and the gradient method with a constant step, applying the concept of symmetry to them. The algorithm of the steepest gradient descent method of when working with $F(y, z)$ is:

$$(15) \begin{cases} \frac{dy}{dt} = -kQ(y - z) - 0.5k \text{ grad } f(y), & y(0) = y_0 \\ \frac{dz}{dt} = -kQ(z - y) - 0.5k \text{ grad } f(z), & z(0) = z_0, \end{cases} \quad y_0 \neq z_0.$$

In contrast to the differential equation (1), which may be represented as a system of two first order differential equations and which describes the motion of one representative point, the algorithm (15) describes the energy interaction of two representative points. These points form a single system.

Let us study the work of the principle of the concept of symmetry by applying it to a function:

$$(16) f(x) = k \cdot (x - a)^2 - c \cdot \cos(2\pi x) + b,$$

the graph of which is shown in Fig.4(a). It can be seen from Fig. 4(a) that the function under consideration has local extremums, which are located near the global extremum, which is at the point with the coordinate $x = 4$.

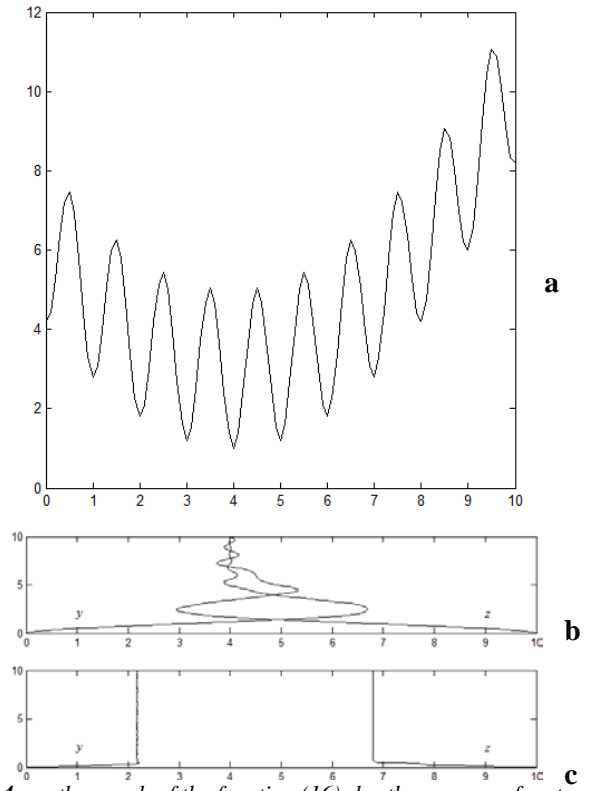


Fig. 4. a - the graph of the function (16); b - the process of motion of the representative points to the global extremum according to the algorithm of the heavy ball method, c - according to the gradient method.

Using equation (4), for the function (16), we obtain the auxiliary function $F(y, z)$:

$$(17) F(y, z) = 0.5 \left[k(y - a)^2 - c \cdot \cos(2\pi y) + k(z - a)^2 - c \cdot \cos(2\pi z) + 2b \right] + 0.5q(y - z)^2.$$

The motion to the minimum of the function $f(x)$ of two representative points with the initial values of the coordinates $y = 0$ and $z = 10$, belonging to the initial function (16), in accordance with the algorithm of the heavy ball method is shown in Fig. 4(b), of the gradient method - in Fig. 4(c).

To minimize the function (17) we'll apply the following methods: coordinate descent, gradient method with constant pitch and heavy ball method. Lines of the function level (17) with the motion path of representative points to its minimum, in accordance with the algorithms of the abovementioned methods (with $k = 0.2$, $a = 4$, $b = 3$, $c = 2$, and $q = 1$) are presented in Fig. 5-7.

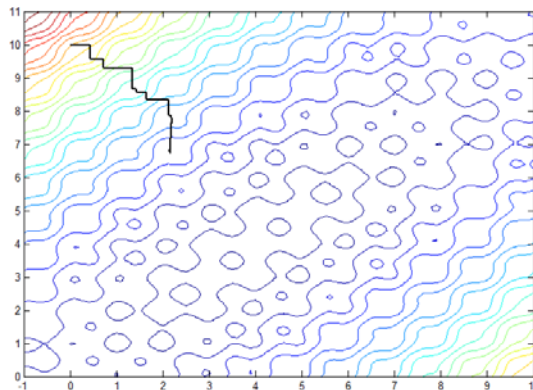


Fig. 5. Graphical illustration of the movement to a minimum by the coordinate descent method

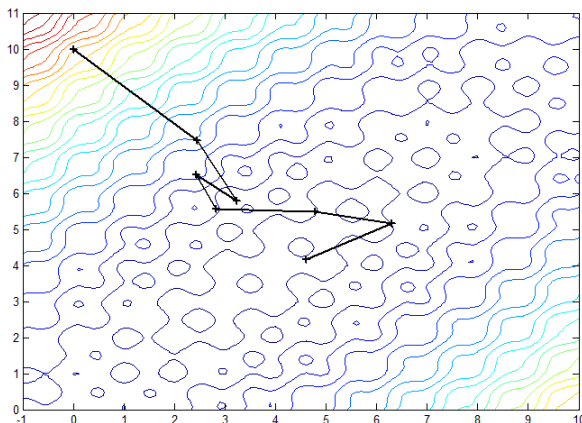


Fig. 6. Graphical illustration of the movement to a minimum by the gradient method with constant pitch

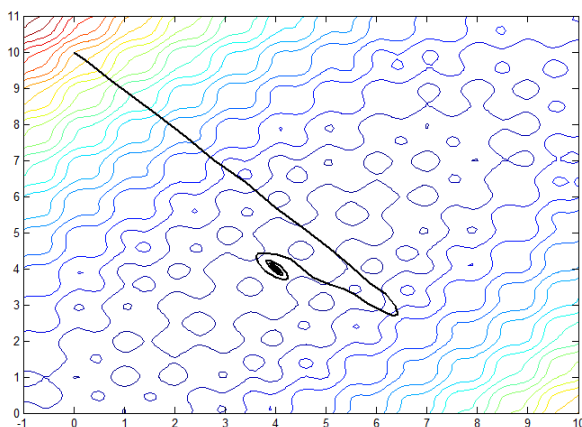


Fig. 7. Graphical illustration of the movement to a minimum by the heavy ball method

When the amplitude is increased in methods of coordinate descent and gradient descent with a constant pitch, the representative points end their motion in local extremums ($y = 6.8420$ and $z = 2.1491$ - coordinate descent method, $y = 4.1595$ and $z = 4.6124$ - gradient descent method with constant pitch). In the heavy ball method the process of motion of the representative points ends at the point of the global optimum $x^* = 4$. The calculated values were: $y = 4.0006$ and $z = 3.9973$, with a mass of heavy balls $m = 2.3$ and a damping factor $r = 2$.

4. Conclusion

The work of the algorithm of the modified heavy ball method, based on the principle of symmetry was investigated. The results of the research showed that for the test functions of Ackley, Griewank and Schwefel the seeking for a global extremum of a function using the usual heavy ball method does not give a positive result. However, the application of the concept of symmetry to the algorithm of the heavy ball method gave the desired result: the algorithms converge over time to the global minimum of multiextremal functions. When the amplitude of the oscillations of the multiextremal function (16) is increased, gradient methods of zero and first order are unsuitable for seeking for the global extremum of the function, since the representative points terminate at local minima. In addition, the considered algorithm of the heavy ball method with the application of the symmetry principle has good operability and high accuracy. Thus, the parallelization of the process of seeking for an extremum of a function based on the use of the concept of symmetry with regard to optimization problems will subsequently yield a number of positive results for estimating unknown parameters of objects in solving the problems of metrology of dynamic measurements and the synthesis of adaptive identification systems.

Bibliography

1. Стоян Б.Г. Решение некоторых многоэкстремальных задач методом сужающихся окрестностей / Б.Г. Стоян, В.З. Соколовский. – К.: Наук. думка, 1980. – 208 с.
2. Бахвалов Н.С. Численные методы (анализ, алгебра, обыкновенные дифференциальные уравнения) / Н.С. Бахвалов. – М.: Наука, 1973. – 632 с.
3. Васильев Ф.П. Численные методы решения экстремальных задач / Ф.П. Васильев. – М.: Наука, 1988. – 552 с.
4. Корсун В.И. Использование симметрии для распараллеливания процесса поиска экстремума целевой функции в задачах оптимального проектирования и адаптивной идентификации / В.И. Корсун // Математические модели и современные информационные технологии: сб. науч. тр. НАН Украины, Ин-т математики. – К.: 1998. – С.66-68.
5. Kharlamova Y.N. Study of the process of seeking global extremum of function by symmetric algorithms with parallel space / Y. N. Kharlamova, V.I. Korsun // Systems technologies — 2015. — № 5(100). — С. 151-161.
6. Korsun V. Paralleling of extremum seeking process in technical problems. / V. Korsun, Y. Harlamova // Science, Education and Culture in Eurasia and Africa: The 5th International Academic Congress (23-25 March 2015, Paris, France). – “Paris University Press”, 2015. – С. 112-116. – ISBN 978-2-547-75283-5.
7. Корсун В.И. Расширение возможностей методов установления при поиске глобального экстремума функции на основе концепции симметрии / В.И. Корсун, Ю.В. Жихарев, В.Л. Галюта // Матеріали міжнародної конференції «Математичні проблеми технічної механіки». – Д.: ДНВП «Системні технології», – 2005. – С.160.
8. Харламова Ю.Н. Оценка эффективности симметричного алгоритма метода тяжелого шарика при поиске глобальных экстремумов тестовых функций / Ю.Н. Харламова // Гірнича електромеханіка та автоматика – 2015. — №95 – С 31-35.

ENERGY AND EXERGY ANALYSIS OF SEA WATER PUMP FOR THE MAIN CONDENSER COOLING IN THE LNG CARRIER STEAM PROPULSION SYSTEM

PhD. Mrzljak Vedran¹, Student Žarković Božica¹, PhD Student Eng. Poljak Igor²
Faculty of Engineering, University of Rijeka, Vukovarska 58, 51000 Rijeka¹, Rožiči 4/3, 51221 Kostrena², Croatia
E-mail: vedran.mrzljak@riteh.hr, bozica.zarkovic@gmail.com, igor.poljak2@gmail.com

Abstract: Energy and exergy analysis of sea water pump which is used for the main condenser cooling at lower steam propulsion system loads on conventional LNG carrier is presented in this paper. By using the measured variables from the exploitation, it is presented different influences of pump used power on cumulative energy and exergy power inputs. Energy and exergy pump power losses are reverse proportional, while pump energy and exergy efficiencies are directly proportional to increase in pump load. The highest obtained pump energy efficiency amounts 59.61 %, while the highest obtained pump exergy efficiency amounts 60.25 %. From the viewpoint of efficiencies and power losses, it will be optimal that analyzed pump operates at the highest possible loads.

Keywords: ENERGY EFFICIENCY, ENERGY POWER LOSSES, EXERGY EFFICIENCY, EXERGY DESTRUCTION, SEA WATER PUMP

1. Introduction

Steam propulsion systems are dominant propulsion systems of LNG carriers [1]. Such steam propulsion systems have many components, not only for ship propulsion, but also for electricity and heat production [2]. One of the constituent components of such marine steam propulsion system is main condenser. Main condenser is cooled with sea water, while the condenser cooling system can be performed in two different ways. First one is cooling with the sea water pump (or more of them) only. Second is combination of cooling with sea water pump at the low steam system loads [3], while on the high steam system loads (at high ship speed) sea water pump is switched off and condenser cooling is performed with the scoop. The scoop is a tube mounted under the ship and it leads sea water directly to main condenser cooling tubes. Condenser cooling with a scoop is very dependable on ship current speed so it cannot be performed under optimal ship speed, defined with scoop producer specifications.

In this paper is analyzed sea water pump from a conventional LNG carrier, which uses main condenser cooling with sea water pump at low and with scoop at high steam system loads. Main characteristics of the LNG carrier in which steam propulsion system is mounted analyzed sea water pump are presented in Table 1.

Table 1. Main characteristics of the LNG carrier

Dead weight tonnage	84,812 DWT
Overall length	288 m
Max breadth	44 m
Design draft	9.3 m
Propulsion turbine	Mitsubishi MS40-2 (max. power 29.420 kW)
Turbo-generators	2 x Shinko RGA 92-2 (max. power 3.850 kW each)

The main goal of this paper is to perform analysis of sea water pump during its operation from the aspect of energy and exergy. It was investigated the influences of pump driving power on energy and exergy pump inputs and it was performed analysis of pump energy and exergy power losses, for the entire pump operation range. Based on a measured data from ship exploitation, it was observed the change in pump energy and exergy efficiencies from the lowest to the highest pump load.

2. Sea water pump energy and exergy analysis

2.1. General equations for energy and exergy analysis

The first law of thermodynamics defined energy analysis of any steam system component [4]. Mass and energy balance equations for a standard volume in steady state disregarding potential and kinetic energy can be expressed according to [5]:

$$\sum \dot{m}_{IN} = \sum \dot{m}_{OUT} \quad (1)$$

$$\dot{Q} - P = \sum \dot{m}_{OUT} \cdot h_{OUT} - \sum \dot{m}_{IN} \cdot h_{IN} \quad (2)$$

Energy power of a flow for any fluid stream can be calculated according to the [6] by using the equation:

$$\dot{E}_{en} = \dot{m} \cdot h \quad (3)$$

Energy efficiency may take different forms and types. Usually, energy efficiency can be written as [7]:

$$\eta_{en} = \frac{\text{Energy output}}{\text{Energy input}} \quad (4)$$

Second law of thermodynamics defines exergy analysis [8]. The main exergy balance equation for a standard volume in steady state is [9]:

$$\dot{X}_{heat} - P = \sum \dot{m}_{OUT} \cdot \varepsilon_{OUT} - \sum \dot{m}_{IN} \cdot \varepsilon_{IN} + \dot{E}_{ex,D} \quad (5)$$

where the net exergy transfer by heat (\dot{X}_{heat}) at the temperature T is equal to [10]:

$$\dot{X}_{heat} = \sum (1 - \frac{T_0}{T}) \cdot \dot{Q} \quad (6)$$

Specific exergy was defined according to [11] by an equation:

$$\varepsilon = (h - h_0) - T_0 \cdot (s - s_0) \quad (7)$$

The total exergy of a flow for every fluid stream can be calculated according to [6]:

$$\dot{E}_{ex} = \dot{m} \cdot \varepsilon = \dot{m} \cdot [(h - h_0) - T_0 \cdot (s - s_0)] \quad (8)$$

Exergy efficiency is also called second law efficiency or effectiveness [12]. It can be defined as:

$$\eta_{ex} = \frac{\text{Exergy output}}{\text{Exergy input}} \quad (9)$$

2.2. Sea water pump efficiencies and losses (energy and exergy)

Analyzed sea water pump is used in LNG carrier steam propulsion system for the main condenser cooling at low propulsion system loads. Main condenser cooling on this LNG carrier is performed in two ways: at low propulsion system loads (at low ship speed) main condenser is cooled with pump, while at high propulsion system loads (at high ship speed) main condenser is cooled with the scoop. According to producer specifications [13], main pump characteristics are:

- Pump maximum capacity: 6000 m³/h
- Pump maximum delivery height: 13 m
- Standard pump operation revolutions: 390 rpm

Power for sea water pump operation, in each operating point, is calculated according to producer specifications [13] by using sea water volume flow at the pump inlet. Sea water volume flow can be calculated by using measured sea parameters at the pump inlet: temperature, pressure and mass flow. Pump used power (P) is approximated by using sixth degree polynomial:

$$P = 1.3216361 \cdot 10^{-20} \cdot \dot{V}^6 - 2.4675409 \cdot 10^{-16} \cdot \dot{V}^5 + 1.7487692 \cdot 10^{-12} \cdot \dot{V}^4 - 6.0343228 \cdot 10^{-9} \cdot \dot{V}^3 + 1.0365022 \cdot 10^{-5} \cdot \dot{V}^2 - 4.2656487 \cdot 10^{-3} \cdot \dot{V} + 120.657 \quad (10)$$

where pump used power P is obtained in (kW) when in equation (10) is placed sea water volume flow at the pump inlet \dot{V} in (m³/h).

For the analyzed sea water pump, all necessary operating points were presented in Fig. 1. The required specific enthalpies and specific entropies as well as other thermodynamic properties were calculated from measured pressures and temperatures for each fluid stream by using NIST REFPROP software [14].

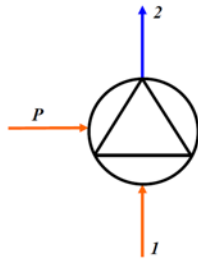


Fig. 1. Sea water pump symbol with marked inlets (inputs) and outlet (output)

Mass, energy and exergy balances for the analyzed pump, according to Fig. 1, are:

Mass balance:

$$\dot{m}_1 = \dot{m}_2 \quad (11)$$

Energy balance:

- Energy power input (only sea water flow):

$$\dot{E}_{en,IN,sw} = \dot{m}_1 \cdot h_1 \quad (12)$$

- Energy power input (cumulative):

$$\dot{E}_{en,IN,c} = \dot{m}_1 \cdot h_1 + P \quad (13)$$

- Energy power output:

$$\dot{E}_{en,OUT} = \dot{m}_2 \cdot h_2 \quad (14)$$

- Energy power loss:

$$\dot{E}_{en,PL} = \dot{E}_{en,IN,c} - \dot{E}_{en,OUT} = \dot{m}_1 \cdot h_1 + P - \dot{m}_2 \cdot h_2 \quad (15)$$

- Energy efficiency [15]:

$$\eta_{en} = \frac{\dot{E}_{en,OUT} - \dot{E}_{en,IN,sw}}{P} = \frac{\dot{m}_2 \cdot h_2 - \dot{m}_1 \cdot h_1}{P} \quad (16)$$

Exergy balance:

- Exergy power input (only sea water flow):

$$\dot{E}_{ex,IN,sw} = \dot{m}_1 \cdot \varepsilon_1 \quad (17)$$

- Exergy power input (cumulative):

$$\dot{E}_{ex,IN,c} = \dot{m}_1 \cdot \varepsilon_1 + P \quad (18)$$

- Exergy power output:

$$\dot{E}_{ex,OUT} = \dot{m}_2 \cdot \varepsilon_2 \quad (19)$$

- Exergy power loss (exergy destruction):

$$\dot{E}_{ex,D} = \dot{E}_{ex,IN,c} - \dot{E}_{ex,OUT} = \dot{m}_1 \cdot \varepsilon_1 + P - \dot{m}_2 \cdot \varepsilon_2 \quad (20)$$

- Exergy efficiency [16]:

$$\eta_{ex} = \frac{\dot{E}_{ex,OUT} - \dot{E}_{ex,IN,sw}}{P} = \frac{\dot{m}_2 \cdot \varepsilon_2 - \dot{m}_1 \cdot \varepsilon_1}{P} \quad (21)$$

Ambient conditions during measurements were:

- pressure: $p_0 = 0.1 \text{ MPa} = 1 \text{ bar}$,
- temperature: $T_0 = 25 \text{ °C} = 298.15 \text{ K}$.

3. Measurement results and measuring equipment of the analyzed sea water pump

Measurement results for sea water pump during pump operation are presented in Table 2. Pump operating parameters were presented in relation to propulsion propeller speed. Propulsion propeller speed is directly proportional to steam system load - higher propulsion propeller speed denotes a higher steam system load and vice versa. It also should be noted that the sea water pump load is directly proportional to steam system load, during the entire pump operating time.

Table 2. Measurement results for sea water pump during its operation

Propulsion propeller speed (rpm)	Sea water pump - inlet			Sea water pump - outlet	
	Temperature (°C)	Pressure (MPa)	Sea mass flow (kg/h)	Temperature (°C)	Pressure (MPa)
0.00	30	0.1	1120106	30.003	0.223
25.58	22	0.1	1346990	22.002	0.221
34.33	22	0.1	1795986	22.002	0.217
41.78	22	0.1	1795986	22.002	0.218
53.50	22	0.1	2244983	22.002	0.214
56.65	18	0.1	2471535	18.001	0.212

The measurement results were obtained from the existing measuring equipment mounted on the pump inlet and outlet. All measuring equipment is tested and calibrated by producers. List of all used measuring equipment is presented in Table 3.

Table 3. Used measuring equipment for the pump analysis

Sea temperature (pump inlet and outlet)	Greisinger GTF 401-Pt100 - Immersion probe [17]
Sea pressure (pump inlet and outlet)	Yamatake JTG940A - pressure transmitter [18]
Sea mass flow (pump inlet)	Promass 80F - Coriolis Mass Flow Measuring System [19]
Propulsion propeller speed	Kyma Shaft Power Meter (KPM-PFS) [20]

4. Results of pump energy and exergy analysis

At each operating point of the analyzed sea water pump is calculated sea volume flow, an essential element for pump used power calculation. Sea water volume flow is calculated for the pump inlet by using measured sea temperature and pressure along with measured sea mass flow, Table 2.

Fig. 2 presents the change in sea water volume flow at the pump inlet for each pump (steam system) load during pump operation. Increase in pump load causes a continuous increase in sea water volume flow from the lowest value (1125 m³/h at propulsion propeller speed of 0.00 rpm) to the highest value (2475 m³/h at propulsion propeller speed of 56.65 rpm).

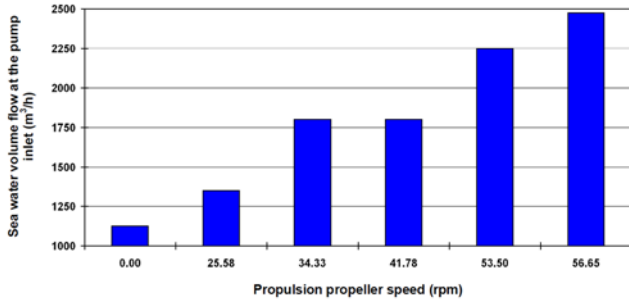


Fig. 2. Change in sea water volume flow at the pump inlet for various pump loads

Pump used power is calculated from the sea water volume flow at the pump inlet, according to equation (10). As presented in Fig. 3, pump used power change has the same trend as sea water volume flow. At the lowest load (0.00 rpm of the propulsion propeller) pump uses power of 122.77 kW, while at the highest load (56.65 rpm of the propulsion propeller) pump uses power of 127.85 kW. As expected, sea water pump used power increases during the increase in pump (steam system) load.

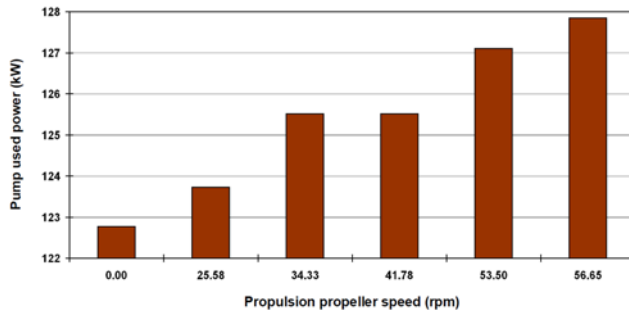


Fig. 3. Change in a pump used power for various pump loads

Energy power input and output of the analyzed pump in each observed operating point is presented in Fig. 4. According to equation (13), cumulative energy power input is a sum of only sea water flow power input and pump used power. At each pump operating point, pump used power influence on cumulative energy power input is very small, the dominant element of the cumulative energy power input is sea water flow energy power input. Only sea water flow energy power input amounts from 99.64 % to 99.78 % of cumulative pump energy power input.

In the entire pump operating range, cumulative energy power input amounts from 34686 kW up to 57732 kW, while energy power output amounts from 34608 kW up to 57676 kW. Energy power loss in each observed pump operating point is the difference between cumulative energy power input and energy power output. By taking into account the amount of cumulative energy power input and output, energy power loss for the observed pump will have small values in each operating point, which will not exceed 85 kW.

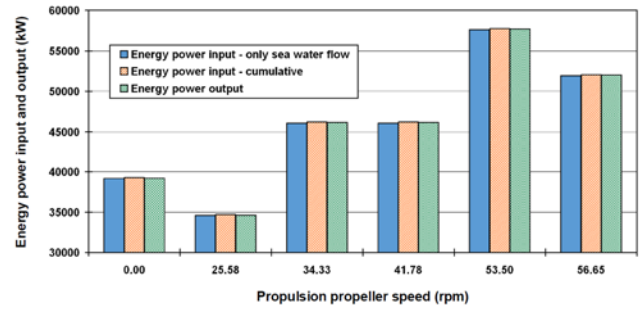


Fig. 4. Change in pump energy power input and output for various pump loads

Fig. 5 presents change in sea water pump energy power losses and energy efficiency. At the lowest pump load (0.00 rpm of the propulsion propeller) energy power loss is the highest and amounts 82.32 kW. Pump energy power losses continuously decrease during the increase in pump load and the lowest energy power loss was obtained at the highest pump load (56.65 rpm of the propulsion propeller) and amounts 51.64 kW.

The lowest pump energy efficiency was obtained at the lowest pump load and amounts 32.95 %, while the highest pump energy efficiency was obtained at the highest pump load and amounts 59.61 %. From Fig. 5 can be seen that pump energy efficiency continuously increases during the increase in pump load.

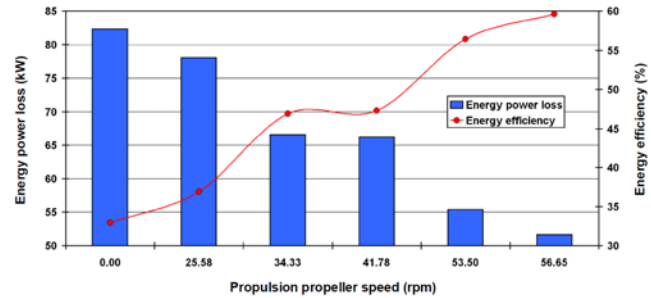


Fig. 5. Change in pump energy power loss and energy efficiency for various pump loads

Change in pump exergy power input and output is presented in Fig. 6. Pump used power has a strong influence on pump cumulative exergy power input. Only sea water flow exergy power input amounts from 16.12 % to 30.52 % of cumulative pump exergy power input for all pump operating points except for one at the highest pump load. It can be concluded that pump used power has a strong influence on pump cumulative exergy power input, while its influence on cumulative pump energy power input is low, Fig. 4.

For the whole sea water pump operating range, cumulative pump exergy power input amounts from 147.50 kW up to 367.64 kW, while pump exergy power output amounts from 69.15 kW up to 316.82 kW.

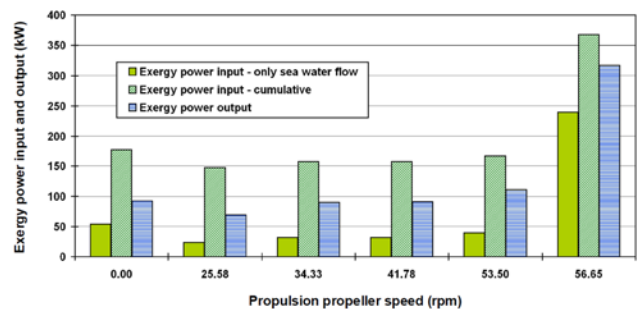


Fig. 6. Change in pump exergy power input and output for various pump loads

The change in sea water pump exergy power losses (exergy destruction) and exergy efficiency is presented in Fig. 7. The lowest pump exergy efficiency was obtained at the lowest pump load (0.00 rpm of the propulsion propeller) and amounts 31.31 %, while the highest pump exergy efficiency was obtained at the highest pump load (56.65 rpm of the propulsion propeller) and amounts 60.25 %.

At the lowest pump load exergy destruction has the highest value and amounts 84.33 kW. Pump exergy destruction continuously decreases during the increase in pump load and the lowest exergy destruction was obtained at the highest pump load and amounts 50.83 kW.

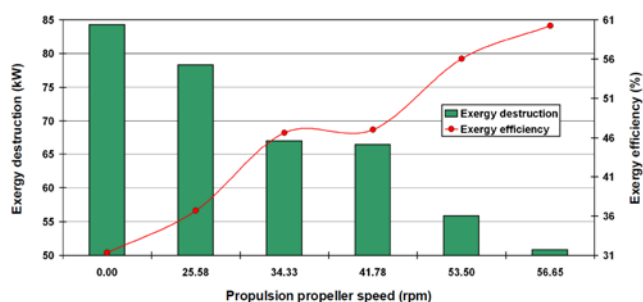


Fig. 7. Change in pump exergy destruction and exergy efficiency for various pump loads

5. Conclusions

The paper presents energy and exergy analysis of sea water pump which is used for the main condenser cooling at lower steam propulsion system loads on conventional LNG carrier.

By using the measured variables from the exploitation, it can be concluded that pump used power has a strong influence on pump cumulative exergy power input, while its influence on pump cumulative energy power input is minor, for the whole observed pump operation range.

Pump energy power losses and exergy destruction are reverse proportional to pump load. Increase in pump load also causes a continuous increase in both energy and exergy efficiencies. The highest pump energy efficiency amounts 59.61 %, while the highest pump exergy efficiency amounts 60.25 %. Increase in both efficiencies (and decrease in both power losses at the same time) will be possible if the pump operates at the highest possible loads, higher than presented ones from LNG carrier exploitation.

6. Acknowledgment

The authors would like to extend their appreciations to the main ship-owner office for conceding measuring equipment and for all help during the exploitation measurements. This work was supported by the University of Rijeka (contract no. 13.09.1.1.05) and Croatian Science Foundation-project 8722.

NOMENCLATURE		Greek symbols:	
Abbreviations:		ε	specific exergy, kJ/kg
LNG	Liquefied Natural Gas	η	efficiency, -
Latin Symbols:		Subscripts:	
\dot{E}	stream flow power, kJ/s	0	ambient state
h	specific enthalpy, kJ/kg	c	cumulative
\dot{m}	mass flow, kg/s or kg/h	D	destruction
p	pressure, MPa	en	energy
P	power, kJ/s	ex	exergy
\dot{Q}	heat transfer, kJ/s	IN	inlet
s	specific entropy, kJ/kg·K	OUT	outlet
\dot{X}_{heat}	heat exergy transfer, kJ/s	PL	power loss
T	temperature, °C or K	sw	sea water
\dot{V}	volume flow, m ³ /h		

7. References

- [1] Schinas, O., Butler, M.: *Feasibility and commercial considerations of LNG-fueled ships*, Ocean Engineering 122, p. 84–96, 2016. (doi:10.1016/j.oceaneng.2016.04.031)
- [2] Baldi, F., Ahlgren, F., Melino, F., Gabrielli, C., Andersson, K.: *Optimal load allocation of complex ship power plants*, Energy Conversion and Management 124, p. 344–356, 2016. (doi:10.1016/j.enconman.2016.07.009)
- [3] Fernández, I. A., Gómez, M. R., Gómez, J. R., Insua, A. A. B.: *Review of propulsion systems on LNG carriers*, Renewable and Sustainable Energy Reviews 67, p. 1395–1411, 2017. (doi:10.1016/j.rser.2016.09.095)
- [4] Kaushik, S. C., Siva Reddy, V., Tyagi, S. K.: *Energy and exergy analyses of thermal power plants: A review*, Renewable and Sustainable Energy Reviews 15, p. 1857–1872, 2011. (doi:10.1016/j.rser.2010.12.007)
- [5] Hafidhi, F., Khir, T., Ben Yahya, A., Ben Brahim, A.: *Energetic and exergetic analysis of a steam turbine power plant in an existing phosphoric acid factory*, Energy Conversion and Management 106, p. 1230–1241, 2015. (doi:10.1016/j.enconman.2015.10.044)
- [6] Mrzljak, V., Poljak, I., Mrakovčić, T.: *Energy and exergy analysis of the turbo-generators and steam turbine for the main feed water pump drive on LNG carrier*, Energy Conversion and Management 140, p. 307–323, 2017. (doi:10.1016/j.enconman.2017.03.007)
- [7] Mrzljak, V., Poljak, I., Medica-Viola, V.: *Dual fuel consumption and efficiency of marine steam generators for the propulsion of LNG carrier*, Applied Thermal Engineering 119, p. 331–346, 2017. (doi:10.1016/j.applthermaleng.2017.03.078)
- [8] Kanoğlu, M., Çengel, Y.A., Dincer, I.: *Efficiency Evaluation of Energy Systems*, Springer Briefs in Energy, Springer, 2012. (doi:10.1007/978-1-4614-2242-6)
- [9] Zisopoulos, F. K., Moejes, S. N., Rossier-Miranda, F. J., Van der Goot, A. J., Boom, R. M.: *Exergetic comparison of food waste valorization in industrial bread production*, Energy 82, p. 640–649, 2015. (doi:10.1016/j.energy.2015.01.073)
- [10] Ahmadi, G., Toghraie, D., Azimian, A., Ali Akbari, O.: *Evaluation of synchronous execution of full repowering and solar assisting in a 200 MW steam power plant, a case study*, Applied Thermal Engineering 112, p. 111–123, 2017. (doi:10.1016/j.applthermaleng.2016.10.083)
- [11] Ahmadi, G., Toghraie, D., Ali Akbari, O.: *Solar parallel feed water heating repowering of a steam power plant: A case study in Iran*, Renewable and Sustainable Energy Reviews 77, p. 474–485, 2017. (doi:10.1016/j.rser.2017.04.019)
- [12] Szargut, J.: *Exergy Method - Technical and Ecological Applications*, WIT Press, 2005.
- [13] *Final drawing for Main Sea Water Circulation Pump*, Shinko IND. LTD., Hiroshima, Japan, 2006., ship documentation
- [14] Lemmon, E. W., Huber, M. L., McLinden, M. O.: *NIST Reference Fluid Thermodynamic and Transport Properties-REFPROP*, Version 8.0, User's Guide, Colorado, 2007.
- [15] Çengel Y., Boles M.: *Thermodynamics an engineering approach*, Eighth edition, McGraw-Hill Education, 2015.
- [16] Moran M., Shapiro H., Boettner, D. D., Bailey, M. B.: *Fundamentals of engineering thermodynamics*, Seventh edition, John Wiley and Sons, Inc., 2011.
- [17] <https://www.greisinger.de> (accessed: 18.09.2017.)
- [18] <http://www.industriascontrolpro.com> (accessed: 18.09.2017.)
- [19] <https://portal.endress.com> (accessed: 22.09.2017.)
- [20] <https://www.kyma.no> (accessed: 02.10.2017.)

THE PECULIARITIES OF THE METALLURGICAL DESIGN DEVELOPED THROUGH PROJECT MANAGEMENT PRINCIPLES

Prof. Tontchev N.¹, Assoc. Prof. Dimitrov D.¹ Prof. Hai Hao²
¹«Todor Kableshtov» Higher School of Transport, Sofia, Bulgaria
²Dalian University of Technology (DUT): 大连理工大学

Abstract. The tools for statistical expert evaluation of the influence of alloy composition elements on pre-selected quality indicators are described in order to improve the mechanical properties of the products through the prism of project management. Through the defined approach, it is possible to define a composition providing relatively best meanings of the values of the selected mechanical indices.

KEY WORDS. SIMULATION, ANN, MODELING, OPTIMIZATION, METALLURGICAL DESIGN.

1. Introduction

At the design stage, the material selection process reflects a whole strand of material science. The traditional alloy development strategy consists of producing multiple samples with varying composition and variations in chemical composition and processing mode to define an alloy with better properties [1]. This approach, known as “trial-error” leads to high experimental costs [2]. In most cases, experimental research may be difficult, too long and unacceptably expensive. An alternative effective approach is the use of previous experience data processed into a statistical model based on a large amount of data associated with composition, processing and properties. Compared to physical models, the advantage of statistical models lies in their ability to

obtain complex informative in a peculiar and effective way even when there are no well-established physical theories and models [3], [4].

We develop a system of methods for modeling the properties and optimization of the composition of different alloys. This paper attempts to point to a number of examples with the formulations of which, from various daughter projects, it is possible to develop further and improve the original idea developed in [7]. The formulated task stems directly from the subject of material science and it is already implemented for alloys of iron [7], titanium [9] and magnesium [9] base. In this regard, the graphical visualization of the basic problems of the material science – the basis of the projects for research of the relationship between properties and structure – is depicted in Fig. 1.

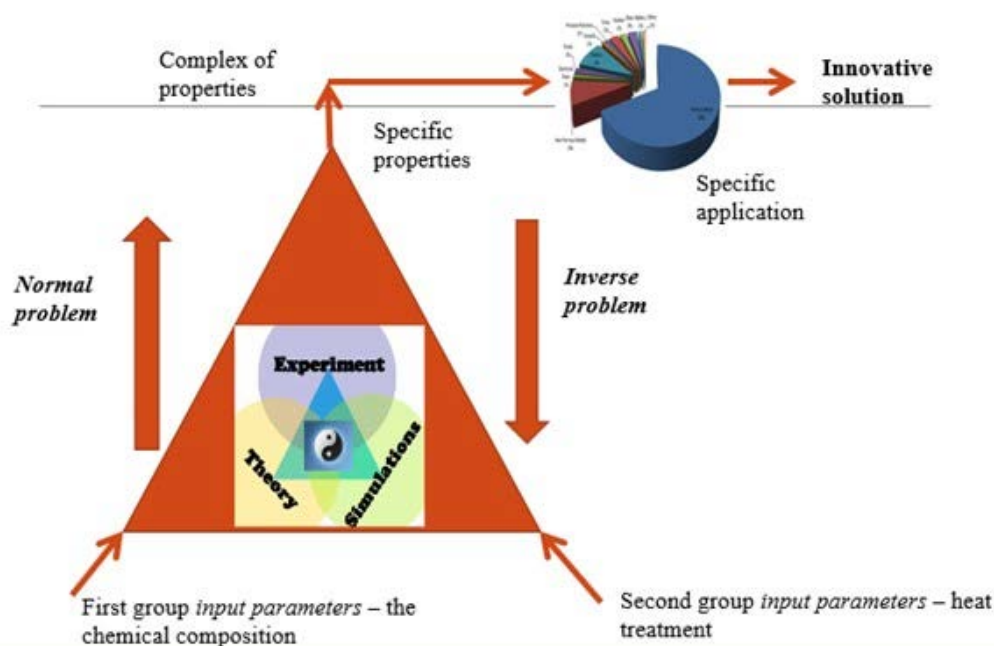


Fig.1. Relationship between the elements of the problems and the projects from material science.

The number of all parameters in the project is determined by the sum of the two sets of parameters (the parameters for the type, the number and the quantity of the alloying elements and the parameters of thermal, physical or mechanical impact).

The right task examines the impact of these parameters to synthesize new materials and in the inverse task and a defined and known application or characteristic values of the exploitation behavior, a rational type of material is determined with its composition and processing parameters.

The numeric metallurgical design is not only designing in the base of alloying elements by type, number and quantities but also taking into account their synergistic effect. It is necessary to select such a synergistic effect by a combination of elements in which the more expensive elements are in smaller quantities. Along with the accomplishment of this task, the metallurgical design has to fulfill the desired compromise between the properties of the product, depending on the processing parameters. These several groups of parameters should form the overall processing and composition parameters. The complexity of the so formulated problem lies in the large

number of input parameters and in the fact circumstance that the chemical composition participates implicitly via the synergistic effect in the system “composition – processing – structure – properties – price.

Alloy composition design and optimization of process parameters are directly related to resolving the trade-off between measured values relative to defined selected quality indicators for a set of materials in a surveyed group or class.

The process system also optimized the technological parameters of the processes, such as ionization, pouring, post-casting, as well as casting and thermal treatment of aluminum wheels.

With the quoted system, there have been optimized also the technological parameters of the processes such as ionization, plastering, post-casting, as well as casting and thermal treatment of aluminum wheels.

2. Classification of Material-Science Projects

The pursuit of strengthening in defining the specific purpose of a project in the field of material science is associated with its uniqueness. In this subject area, in the development of projects arising from the main task and also from “studying the relationship between the influence of chemical composition and heat treatment on mechanical properties in order to search for new applications”, the projects have a unique effect because – as a result of the optimization – a new innovative solution is usually obtained (Fig.1).

The specificity of these tasks is associated with a certain complex of properties, which analyzes several criteria (different qualitative indicators, which are most often contradictory in character), and the optimization of which determines solutions that are inherently effective. This goal is defined in the literature as the SMART goal, and the meaning of this abbreviation is described and explained in more detail in the table below.

Table 1. Defining the SMART-goal abbreviation

S	Specific
M	Measurable
A	Achievable
R	Realistic
T	Timed

The specificity of such projects is a consequence of the subject matter of the material science from which the basic relationship “structure-properties” derives. For the definition of qualitative indicators (most often mechanical characteristics) there are standardized methodologies that guarantee a measurable quality result. Many examples from our previous experience prove that the objectives of such a project are achievable. They depend on the database used, and the result of the project is to apply authoring methods that can determine chemical compositions outside the database to exceed the mechanical performance of the database. In this sense, the purpose of the project is realistic.

Determining the project over time depends on the volume of tasks set such as the size and the number of research quantities, and the organization of the implementation of the key activities. The main focus of this organization is at the stage of experimental testing of the prescribed compositions, which are determined by the application of the methodology as optimal.

Since the process of formulating the goal is not an easy process, there are projects that start with an undefined purpose. There is a certain categorization in this sense, which is presented in Fig. 2. It – besides an established or unsettled SMART goal – has also been implemented through known and unknown methods and technologies that will be implemented during the implementation of the project.

Unsettled technologies & unknown methods of work		“Open” projects
Established technologies & known methods of operation	“Closed” projects	
	Established SMART goal	Unsettled SMART goal

Fig.2. Classification of projects for clarified goals and used technologies

“Closed” projects are standard and routine projects. They are most likely to be successful and their performance is associated with the least risk. However, in any organization there is a “narrow” place, shaped as a project where the goal is difficult to be determined and the technology for its solution is unknown.

Unsettled technologies & unknown methods of work	A clear set of materials with a complex of properties with an unspecified test method	Synthesis of a new material with an undefined complex of properties
Established technologies & known methods of operation	Solving the classical task of material science by defining a qualitative indicator with an existing methodology	Use of material with a composition subject to investigation for which the test procedures are defined
	Established SMART goal	Unsettled SMART goal

Fig.3. Specifying the classification from Fig.1 for projects in the field of material science

In this sense, the project team has never solved such problems, and these projects are called “open” because the issue is open to both the goal and the technologies and methods to be used. The organization balances its activity by accepting to develop both “open” and “closed” projects. “Closed” projects are implemented in a short-term perspective, but similar projects are also being implemented by competitors so they are not very lucrative.

Fig.3 presents a specific application of the above Classification for Material Science projects to elucidate the structure-property relationship in order to provide a more rational application.

In the discovered project of the type described, there is a partial requirement for the material group and the conditions for the operational work in which the workpiece/product will work are known. If it is on a metal base, it is possible to apply a solution theoretically developed by the cited project, consisting of the following:

The existing system of numerical methods, by means of which it is possible to determine optimal thermal processing compositions and parameters, associated with a pre-search complex of properties. Various iterations, specifying shape and size, are performed with CAD / CAE system calculations. The search object in the project does not appear in the CAD system database. The calculations are performed with the material closest to the research class. On the basis of the strength calculations the complex of properties is determined. If there is no standardized test in the set of properties, a methodology is developed and experimental research is performed. Project calculations provide optimal formulations for the desired complex of properties. These compositions are validated as being appropriate after an experimental check. Following the verification, the database can be supplemented with the newly defined compositions.

The existing system of numerical methods, by means of which it is possible to determine optimal thermal processing compositions and parameters, associated with a pre-search complex of properties. In various iterations, specifying shape and dimensions, calculations for the strength are performed with the CAD/CAE system. The material being searched for in the project is not included in the CAD system database. The

calculations are performed with the material closest to the research class. On the basis of the strength calculations the complex of properties is determined. If there is no standardized test in the set of properties, a methodology is developed and experimental research is performed. Project calculations provide optimal formulations for the desired complex of properties. These compositions are validated as being appropriate after an experimental check. Following the check, the database with the newly defined compilations can be added.

The Project Life Cycle is related to the full set of stages (phases) of the project, the definition and number of which define the technology of production and the control needs of the organizations involved in the project.

“Open” and “closed” projects differ in their lifecycle approaches. Each project is fragmented in steps. Defining the lifecycle of the “closed project” bears the name “waterfall”, because the tasks are done in a top-down sequence and it resembles a cascade of water.

The example of the “closed” project authorizing the classical task “chemical composition – property” according to an approved methodology of the tested indicator is defined by the following phases of the “waterfall” life cycle:

1. Start the project.
2. Collecting existing experience, described by experimental data, expressing the relationship between the different chemical elements of the alloy and the
3. Data processing to obtain a regression [model] or neuronal model.
4. A single-criterion optimization of the derived model, resulting in a composition, guaranteeing an optimal value of the studied property.
5. Experimental examination of the theoretically determined composition.
6. Final of the project with a specified optimal chemical composition.

The definition of the waterfall life cycle model is related to the definition of the main project parameters (content, time, price cost and thus it determines the start of the project. Planning takes place for each project, and some of the phases can be run in parallel. The model is effective when it has a high degree of certainty.

However, the algorithm for “open” projects is [somewhat] different. The first necessary condition is to take steps to clarify the goals and technologies under which these projects will be implemented. However, this condition is not always possible to fulfill in conditions of great uncertainty and a constantly changing external environment. In this case, the “waterfall” life cycle ceases to work. If it is applied, time and resources may be lost and when the project finishes, it can be ascertained that the results obtained are not needed by our applicant.

For open projects where there is a large uncertainty, an iterative life cycle model is being developed. This model involves breaking the content of the project into small iterations. The end of each iteration is controlled by the applicant who has a feedback on the execution of the corresponding intermediate state. Then work continues and begins the next iteration, which in a certain sense is another mini project at the end of which a feedback from our applicant is expected again.

The result of the last stage may be unsatisfactory, which means it is possible to go back [with] a step back before proceeding with the new iteration. After a certain number of iterations, the applicant is fully satisfied with the outcome and this determines the end of the project. This result is ensured at all stages of the project. The end of the project was unknown and it was determined in the course of the work. The great

advantage of the iterative approach is to work with a large variability in the external environment.

The iterations ensure the improvement of the resulting product in the course of the project. Planning is performed only for the upcoming iteration. The disadvantage of the iterative approach is that re-expenditures are incurred for the individual iterations of the project itself. The team performs planned actions, some of which the applicant in the process of work may not accept, as a result, and then go another way.

The example of an iteration from the defined in Fig. 2 “open” project is limited to the following:

1. Sequential gradual clarification of the influence of individual elements of the chemical composition on the investigated properties. Determining the synergistic effect of the specified elements and selecting a combination that contains smaller quantities of costly elements. This synergistic effect should relate to the properties of the entire explored complex.

This example is typical for the application of the iterative life cycle, because the simulation uses the idea of a “trial-error” method to elucidate the effect of a combination of individual elements on a sought-after complex of properties.

3. Tools Used to Implement Material Science Projects

In the process of a numerical experiment, the solutions determining the predictive potential of the regression and neural models in the multi-criteria optimization of a number of properties are compared in varying the percentage of the alloying elements or the processing parameters.

The recommended tools are looking for and finding convergent and compromising solutions in the presence of diversity and heterogeneity of the output data. This is also done for the contradiction of the requirements to the designed objects. The influence and the interaction between the alloying elements and their synergistic effect related to the properties of the alloys and in particular their strength and performance properties are monitored.

The methodology of the research project consists of the following stages:

- Preliminary statistical analysis of the research data with visualization of the dependencies between observed quantities. This includes the determination of the basic (descriptive) statistical characteristics, the presence of a correlation of the parameters of the experimental studies.
- Simplification of the dependencies between the participating in the alloy chemical elements and its mechanical properties using neural models.

The most important task at this stage is to find an opportunity to find a possible relation between the independent parameters and the dependent characteristics in the experimental research. The statistical analysis allows determine the uncorrelated input parameters from the experiment that can be used to construct a regression model – the percentage of elements in the composition of the alloy under consideration. The limited number of observations sometimes does not make it possible to draw conclusions on the statistical distribution of the observed variables.

Then with these problems with the construction of the statistical models, approximating the influence of the predictors on the values of the result quantities, most often of R_m and A , there is a need to seek the modeling of the studied relationships with other means. To solve the specific problem, the most commonly used approach is the use of neuronal models of the “multilayer perceptron” (MMP) type. This approach is not a novelty, but it proves to be too effective in cases where classical statistical methods do not work. It allows for the compilation of approximation models in cases where the relations between the

observed values are considerably more complex and sometimes implicit.

In general, neural models are not related to statistical techniques, but they are regarded as self-contained outcomes of machine learning. However, many authors note the computational similarity of these models with the methods of statistical analysis. It is possible to use the Automatic Neural Network of the popular statistic package Statistica. It enables an experiment with more than 2000 models for each of the approximate searches, the selection being made according to the value of the correlation coefficients between the observed and the calculated result values (respectively for the learning, the test and the validation sets).

The analysis of the results of the approximation with the fixed neuronal models can be supplemented by a sensitivity analysis. The latter shows the relative importance of the predictors in the formation of the values of the resultant variables. There is a significant influence of the predictor if its value is greater than 1.

Following these statistical techniques, it is recommendable to use a Pareto front, which is an established traditional method in multi-criteria research. It makes it possible to assess the appropriateness and the expected effect of combining different preferences regarding the significance of the physico-mechanical parameters of the material, such as the tensile strength R_m and the relative elongation A . The classical formulation of the task of building a set of non-dominated solutions for the multi-criteria selection has the following appearance:

- Basic optimization task - presented in the following form:

$$\max(f_1(x), f_2(x), \dots, f_m(x)), \quad (1)$$

where:

$f_1(x), f_2(x), \dots, f_m(x)$ are the private optimization criteria (physico-mechanical characteristics of the material);

$$x \in D \subset R^n, m \leq 2, x = (x_1, x_2, \dots, x_n) \quad (2)$$

are the conditions defining the decision domain D of the solution vector $x = (x_1, x_2, \dots, x_n)$. In the specific problem the definition area is set in the form of interval limits:

$$D: x_k \in [l_k, h_k] \subset R, k = 1, \dots, n, \quad (3)$$

where l_k and h_k are respectively the lower and upper bounds of the change interval of each of the arguments (percentage of the chemical element in the material)

- The relation of dominance between a pair of solutions $x^{(1)}$ and $x^{(2)}$ ($x^{(1)} \prec x^{(2)}$) - is expressed as follows:

$$\begin{aligned} \forall i = 1, \dots, m \quad f_i(x^{(1)}) &\geq f_i(x^{(2)}), \\ \exists j = 1, \dots, m \quad f_j(x^{(1)}) &> f_j(x^{(2)}). \end{aligned} \quad (4)$$

The dominance of $x^{(1)}$ versus $x^{(2)}$ exists when at least one of the criteria has strict inequality (with the indicated orientation) for the two alternative solutions, and for the other criteria is in effect non-steady inequality (or at least equality).

The Pareto- optimal set is a set of solutions $P^* \subset D$ that are not dominated by any element of the D set. The Pareto-front can be graphically displayed in the space of the considered criteria $f_i(x)$.

The implementation of this experimental-computational approach requires the preliminary study research

of the experimental data in order to obtain approximating quantitative relationships between them and subsequently to include the established dependencies in a suitable optimization algorithm.

The steps involved in conducting the research are as follows:

- Creation of approximation models linking the parameters of the alloy composition with the resulting physico-mechanical parameters;
- Selection of a suitable algorithm for building a Pareto front;
- Incorporating the approximation models into the algorithm and realization as a programming tool.

All numerical approaches for the metallurgical design are set out in [7]

4. Conclusion

An approach has been described to offer a realistic opportunity to significantly reduce the cost and time to predict chemical concentrations for multiple properties of a class of alloys and technological processes so that under new conditions these materials have better properties. The obtained results from our previous research have shown that the proposed methodology can be fully used to determine the essential relations between the extreme values of the mechanical parameters of the investigated alloys.

Any addition to the experimental measurement database would allow the software model to be refined and further refine the results. The proposed approach and established dependencies can be used in engineering practice.

References:

1. Campbell C.E., G.B. Olson, "System design of high performance stainless steels I, Conceptual and computational design", Journal of Computer-Aided Materials Design 2001, 7:145-17
2. Vitos L., P.A. Korzhavnyi and B. Johansson, "Stainless steel optimization from quantum mechanical calculations", Nature Materials 2003, 2:25-28
3. Jones, J. , MacKay, D. J. C., "Neural Network Modeling of the Mechanical Properties of Nickel Base Superalloys", 8th Int. Symposium on Superalloys, Seven Springs, PA, eds. R. D. Kissinger et al., published by TMS, 1996, pp. 417-424.
4. Fujii, MacKay, D. J. C. and Bhadeshia, H. K. D. H., "Bayesian neural Network Analysis of Fatigue Crack Growth Rate in Nickel-Base Superalloys", ISIJ International, Vol. 36, 1996, pp. 1373-1382.
5. Borst R. de, "Challenges in computational materials science: Multiple scales, multi-physics and evolving discontinuities", Computational Materials Science 2008, 43:1-15
6. Fischer C.C, K.J. Tibbetts, D. Morgan and G. Cedar, "Predicting crystal structure by merging data mining with quantum mechanics", Nature Materials 2006, 5:641-646
7. Tontchev N. Materials science, Effective solution and technological variants. Lamber, 2014.
8. Razmov T., D. Dimitrov , Laboratory Exercise and Course Design on "Project Management" - Third Edition, Todor Kableskov University of Technology, 2012.
9. https://www.researchgate.net/profile/N_Tontchev/contributions

SOFTWARE ASSURANCE OF THE SYNTHESIS AND DESIGN OF HYPERBOLOID GEAR DRIVES

Assoc. Prof. Abadjieva E. PhD.^{1,2}, Prof. Sc. D. Abadjiev V. PhD.²
Graduate School of Engineering Science, Akita University, Japan¹
Institute of Mechanics, Bulgarian Academy of Sciences, Bulgaria²

abadjieva@gipc.akita-u.ac.jp

Abstract: *The study presents a brief description of the type's software, applicable to the synthesis and design of gear transmissions. The main accent is put on the approach to the computer synthesis, for which the optimization process is carried out by the method of directed search of the optimal variant of a synthesized mechanism. The specific features of the programs, applied by the authors, oriented to the synthesis of spatial gear mechanisms with linear contacting teeth with face meshing (Spiroid and Helicon) are studied in detail.*

Keywords: HYPERBOLOID GEAR DRIVES, MATHEMATICAL MODELLING, SYNTHESIS, SOFTWARE ASSURANCE

1. Introduction

The contemporary requirements to the accuracy and reliability characteristics of machine products for the techniques to a great extent dictate the applied scientific methods for the technological synthesis, design and manufacture of gear drives [1 - 3]. In the processes of synthesis and design of the different types of gears, it is necessary to be solved complex set of problems, which considered all together define the desired optimal construction. In this case, an optimal construction means a gear transmission, which is capable to ensure the preliminary given kinematic and strength characteristics at the minimum cost for realization and exploitation. In essence, this is a system of requirements, related to different quality characteristics of the gear, namely [1, 4]:

- **geometric ones**, which control the kinematic exactness, smoothness of the working process, the character of the contact (placement of the contact spot; the orientation of the contact lines and the radius of the curvatures at contact points), related to the loading capacity of the gear sets and etc.;
- **dynamic ones**, which have impact on the noise and vibrations of the gear drives, the conditions for appearance of resonant phenomena and etc.;
- **strength ones** determining the durability and reliability of gear sets, including the transfer of nominal power in the process of rotations transformation with avoidance of "scoring", "pitting" and etc. on the active tooth surfaces of the synthesized mechanism;
- **economic ones**, that define the production costs (e.g., per unit of power), energy loss for the motions transformation (coefficient of efficiency), etc.

The realization of an adequate approach for the creation of the real gear drives requires this approach to be a complex one. This is consisted in considering of the required quality characteristics of the created gear mechanisms with the existing specific technological and manufacturing capabilities.

The choice of the approach should be realized in the process of synthesis and design of gear transmissions.

2. Aspects of the Computer Design of Gear Drives

The wide variety of gear drives used in industry and transport as reduction drives and multipliers, as well as the continuous pursuit of researchers to create new and improved gear mechanisms on one hand and on the other - the different and rapidly vitiating approaches to the mathematical modeling, synthesis and design make it practically impossible to create universal CAD systems. In connection with the mentioned above, a special attention should be paid to the extremely dynamic development of the modern technical computational tools and software applications. This often requires a revaluation not only of the way in which computer programs are organized, but also leads to informal changes in applied mathematical models [5]. The computer design is evolved, forming three types of software [6], in order to realize scientific studies in

the field "Theory of gearing" and to provide an adequate scientific support for this type of manufacture.

First type. The programs, included here, are designed to study the influence of the different kinematic, constructive, technological and exploitation parameters on various quality characteristics of the studied gear drives. Essentially, this type of software is not subjected to a particular strategy, associated with the design of CAD systems. The elaborated mathematical models, algorithms and computer programs are designed to determine the influence of one or other real-life existing parameters on the qualitative characteristics of the concrete gearings. However, the programs created in this case can be used as software modules, which are elements of system of criteria for quality control of the synthesized gear mechanisms.

Second type. This group is consisted of computer programs organized on the basis of algorithms, which are contained in standardization documents [7], company methodologies [8] or handbooks [9, 10]. The program products, included here, are developed on the basis of algorithms for geometric and strength calculation of the traditional types of gear drives: cylindrical involute with external and internal mating gears, cylindrical worm gears, bevel gears with straight teeth and so on. It should be noted, that the algorithms used in these cases do not ensure the optimization in the synthesis and design of gear mechanisms. Secondly, this category of software can also include those products, through which the strength characteristics of the already geometrically and technologically synthesized gear drives [7] are examined. In that capacity, these computer programs can be treated as analysis instruments of the gear mechanisms.

Third type. The computer programs included in this category are those, which are based on the mathematical models, developed on the basis of the specially oriented scientific studies. For example, for Bulgaria, these are the computer programs that deal with the synthesis and design of Spiroid and Helicon gear sets [6, 11, 12], and with conical and hypoid gear mechanism – type Gleason [13-15]; and others. For the contemporary gear transmissions, including even classical gear mechanisms, which are treated in terms of current engineering requirements, the construction of new mathematical approaches to their geometrical, technological and strength synthesis is required. The optimization synthesis process in this case is realized by application of the **method of direct search**. This method gives opportunity to reduce the number of calculated gear pairs, which compose the synthesized gear mechanism. It will be reminded, that the essence of this method is as follows:

- input parameters are defined, as well as those that will not be changed throughout the whole synthesis process;
- the variables parameters are determined as well the way of their variation, respectively;
- the process of changing of the defined variable input parameters compared to their initially given value continues, until the introduced optimization criteria are fulfilled;

- from the calculated pairs of conjugated gear sets, a final variant is chosen for which, there is the best satisfaction of the additional conditions (restrictions) introduced in the mathematical model.

In other words, the process of optimization synthesis and design of the third type of software is based on adequate iterative procedures, by which the desired solution is found by changing certain parameters.

3. Constructing of Computer Programs for Calculation of Hyperboloid Gears with a Linear Contact

The computer programs designed for the synthesis and design of linear contacting hyperboloid gear mechanisms belong to the third type software. Taking into account, the known methodological limitations when constructing this type software, the following sequence is followed, for creation of the system for the computer design of hyperboloid gears with linear contacting tooth surfaces.

3.1. Mathematical Modeling for the Synthesis and Design

When profiling of the kinematically conjugated surfaces, upon which the rotations transformation between crossed axes is carried out, the basic observed principles are the principles of T. Olivier. Thoroughly discussed in [6], it will be summarized only that part of them, which is directly related to the construction of the concrete computer programs. Two applied approaches to the construction of mathematical models for synthesis are formulated here: mathematical modeling, upon which the geometric, technologic and exploitation characteristics of the designed gear sets in a small vicinity of the pitch contact point is defined and optimized and mathematical modeling related to the ensuring of the qualitative characteristics in the entire mesh region.

It is obvious that the methodological difference between the two approaches for the synthesis of spatial gear drives requires to define in advance the adaptability of the future designed hyperboloid gear sets to one of the two approaches. The determination of the adaptability of the planned procedure for building an adequate mechano-mathematical model is a complex creative process, requiring the knowledge of both the theoretical content of the approaches to the synthesis and the specific technological and exploitation requirements characterizing the created products.

3.2. Principles of Organization of the Design Process

Here, the focused will be paid only on those principles, which are determining for the construction of computer programs for the synthesis of Spiroid and Helicon gears.

Determination of the groups of independent and variable input parameters influencing the design conditions. To the group of independent input parameters should be included a set of standardization modules, that determines the technological capabilities of the hobbing machines; coefficients that define the tooth geometry as a function of the modules; coefficients of frictions between the different pairs of materials applicable for producing of the toothing of the conjugated gear pairs; coefficients, linear and angular values associated with the design of the instrumental equipment and etc.

To the input data parameters, among which the variable ones are chosen, as a rule are included those which define the overall geometry of the calculated gear system. Here belongs the parameters determining the dimensions of the gear structure: the offset, distances from the offset to the planes in which the pitch circles lie; the angles defining the orientation of the above said planes relative to the pitch normal and etc. The variable input data include also those, from which the geometry of the conjugated active tooth surfaces depends: the independent coordinates of the tooth surfaces; their helical parameters; parameters which determine face width of the teeth, etc.

Introducing basic analytical relations, which are based on the chosen approach to the mechano-mathematical modeling. Here are in included the solutions of the fundamental tasks of the synthesis upon a pitch contact point and upon a mesh region with the application of the adequate geometric interpretations of the basic equation of meshing, namely [6]: the task for the synthesis of pitch circles; the task for the definition of the geometry of the active tooth surfaces by their linear and angular characteristics in the pitch contact point; the task for defining the singularity in the pitch contact point (without describing the analytical type of the tooth surfaces); the task for analytically defining the entire mesh region; formulations of relations, which are used to determine the optimal dimensions and placement of the region of mesh on the active surface and etc. This principle of organizing the computer design includes also the introduction of geometric and kinematic relations, intended for the reduction of the input parameter sets.

Constructing the complex process for the synthesis and design of hyperboloid gear drives. This is accomplished by defining the separate stages of the synthesis and design in their sequence and interconnection. This principle, applied in the design of each computer program, is in direct dependency by the type of functioning set of criteria. Those criteria determine the defined characteristics of the quality of the gear mechanism in dependence of the accepted approach to the mathematical modelling. A distinctive characteristic of the accepted principle for construction of the complex process for synthesis and design is the chosen approach for the estimation of the calculated option of gear mechanism.

4. Software Programs for Geometric and Technological Synthesis of Spiroid Gears

The shown considerations for construction of software programs, applicable to the synthesis of spatial gear mechanisms [6, 16] are also realized when constructing three types' software products for the design of Spiroid gears, which functional relations are shown in Fig. 1. Each one of the three main directions, illustrated there, has its own importance. It means that the user can restrict himself to use the results of only one program; to analyze and interpret these results and then after an adequate assessment, to go through the entire process shown in the figure.

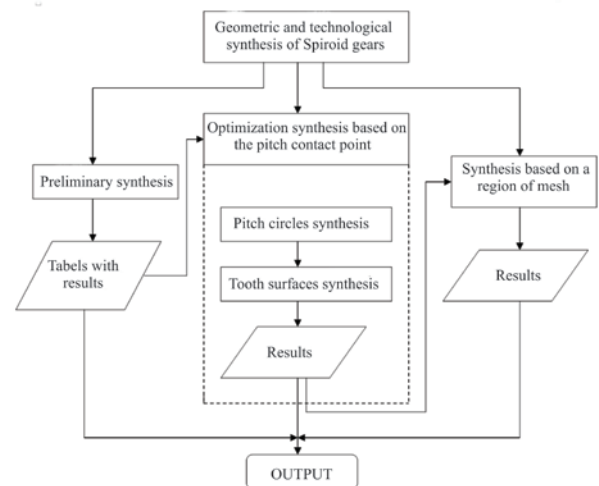


Fig. 1 General block-scheme of the approaches for the synthesis of Spiroid gears

Further bellow it will be treated only one of the directions - the central one, which includes the process of optimization geometric and technological synthesis of Spiroid gears.

4.1. Program for Optimization Geometric and Technological Synthesis of Spiroid Gears upon a Pitch Contact Point

This program consists of solving the following tasks:

- synthesis of the geometric pitch circles;

- synthesis of the active tooth surfaces of the Spiroid pinion and of the cutting tool (Spiroid hob);
- verifying the quality criteria to be fulfilled.

From the formulation of the defined tasks, it can be seen that the algorithm of this program corresponds to the approach to mathematical modeling for synthesis upon a pitch contact point. In this regard, when designing Spiroid gears, it is of particular importance to select the location of the pitch contact point in the fixed space. The placement of the pitch contact point (as a common point of the pitch circles and conjugated active tooth surfaces) affects on one hand on the common geometry of the designed gear system (overall dimensions of the gear pair) and on the other- on the geometry and proportions of the gear's teeth, as well as on the gears' quality (through the geometric, kinematic and strength characteristics of the conjugated gear pair).

For the purposes of this study, here will be briefly repeated some of the information contained in [6], which treats the geometric nature of the externally contacting geometric pitch circles. The pair of pitch circles ($H_1^c : H_2^c$) of the mentioned type are circles H_i^c ($i = 1, 2$), having only one contact point P . Their centers are places on the axes of rotations $i-i$ ($i = 1, 2$) of the movable links, and the corresponding circles are perpendicular to those axes. Mutual position of the crossed axes is uniquely determined by the angle $\delta = \text{constant}$ between the skewed axes and by the offset (center distance) $a_w = \text{constant}$. The position of the geometric circles H_i^c ($i = 1, 2$) in respect with the rotations axes $i-i$ ($i = 1, 2$) and with their offset line is defined by four parameters for each circle. Five independent scalar equations determine the condition - that two circles have one common point. Hence, the mutual position of two circles is not defined in a unique way. It is a function of 5 independent parameters. The synthesis of geometric pitch circles is preceded by the decision, which five of the eight parameter (for both circles) to be chosen for independent ones and how to choose the intervals for their variation.

These five independent parameters, for this case, are chosen as: an angle δ between the axes of rotations $i-i$ ($i = 1, 2$); the offset a_w ; the angle δ_1 , which is concluded between H_1^c and the pitch normal $m-m$ (half of the angle at the top of the pitch cone H_1^s of the Spiroid pinion); radius r_1 of H_1^c and the distance a_1 between the offset line of the gear and the plane in which H_1^c lies.

The ambiguity of the solution enables the possibility, that these free parameters to be changed discreetly within a certain limitations and among many pairs of geometric pitch circles to look for those ones, which parameters ensure that the preliminary defined requirements for the quality of the synthesized gear drives in the vicinity of the pitch contact point are satisfied. The criteria used in the program will be discussed below.

The program allows to choose the type of conic linear helicoid, applied as active tooth surfaces of the Spiroid pinion: convolute, Archimedean or involute ones. The calculation of the necessary and sufficient geometrical and technological parameters for the design of the Spiroid gear pair and cutting instruments is realized for the required type of conic helicoid.

The indicators that serve to control the quality of the gearing are significant for the design process. As it has already be mentioned, the dependence on the solution of the task for the synthesis of the pitch circles from concrete free parameters, should be searched among the optimal geometric, kinematic and technological quality characteristics in the vicinity of the pitch contact point. They will be briefly explained.

- **The basic technological criterion.** This is the main criterion related to the technology of elaboration of Spiroid gears. It is related to the decrease in cutting tools nomenclature by ensuring the conditions for the design of Spiroid hobs with standard modules (hob parameters are functions of this module). This causes the requirement that the calculated in the pitch contact point module to coincide (with a given exactness) with any of the modules contained in the input array of standard modules.

- **Criterion that controls the singularity in the pitch contact point.** The constructed criterion is analytically described in [6, 17]. The insurance of the performance of this criterion helps to reduce the ordinary nodes from the mesh region of the synthesized gear drive. Hence, it leads to the improvement of its loading capacity, of the efficiency and of the durability. It will be reminded, that the elimination of the singularity of first order, by this criterion, is guaranteed in vicinity of the pitch contact point. The optimization, when using this criterion, is realized by the verification of the analytical dependencies introduced for each of the selected combination of the five independent variables.

- **Criterion for controlling the transmission angle of the normal force (pressure angle).** This criterion provides optimization of gear sets in terms of the transmission of normal forces from the pinion to the big gear (crown), when the gear mechanism is operated under the conditions of the rotations transformation at low-side driving.

- **Criterion for controlling the value of the Spiroid pinion spiral angle.** This criterion controls the value of this angle in the pitch contact point. Its values have to belong to definite intervals in accordance to the purpose of the design of the gear mechanism.

Here, it should specially be noted, as it is shown in [6], that the choice of the appropriate values of the pressure angle and the Spiroid pinion spiral angle of the longitudinal line of the active tooth surface $\Sigma_1^{(l)}$ in the pitch contact point substantially affect the efficiency of the gear drives. Therefore, if these geometric characteristics of the tooth surfaces of the Spiroid pinion are appropriately chosen, then an indirect control of the gear mechanism' efficiency is achieved. It also should be mentioned, that from the calculated equivalent variants of the synthesized gear mechanisms, from a geometrical and technological view point, the program allows to select that one which has the highest efficiency value for the computational (pitch) contact point.

- **Criterion related to the durability of the gear drive.** It controls the magnitude of the sliding speed at the computational contact point, depending on the chosen material for the tooth of the Spiroid gear - different types of bronze.

- **Criterion controlling the hydrodynamic conditions of meshing.** This optimization aims that the synthesized gear set has to obtain a maximum as a value -summed circumferential velocity $|\bar{V}_\Sigma|$ in the pitch contact point and minimum value of the angle Ω , which \bar{V}_Σ concludes with the normal to the contact line in the pitch contact point.

- **Technological criteria for hobbing.** These criteria are related to the choice of the minimum value of the axial (normal) profile angle of the Spiroid hob, in order to provide optimal conditions for hobbing, both in terms of cutting the metal and in relation to the strength characteristics of the elements of the gear rack of the hob and others.

In number of cases of the design process, some of the initially independent parameters could be fixed due to the specific requirements (or example requirements for maximum sizes of the gear mechanism and the mutual position of the shafts of the gears), which results in reduction of the number of independent variables without limitation to search and find an optimal geometry of the tooth surfaces.

Input parameter of the programs are: number of Spiroid pinion threads; number of Spiroid gear teeth; offset; standard pressure angle; type of the Spiroid pinion (type of the tooth surfaces

of the Spiroid pinion); type of the bearing of the gear shafts (on two bearing supports or console); frequency of revolution and etc. Keys parameters will take values of 1 or 0 depending on whether a given criterion will be taken in consideration for the synthesis or not. For each of the free parameters should be chosen minimum and maximum values as well as the steps of variation. The independent cycles in the computer program are equal to the number of the free parameters.

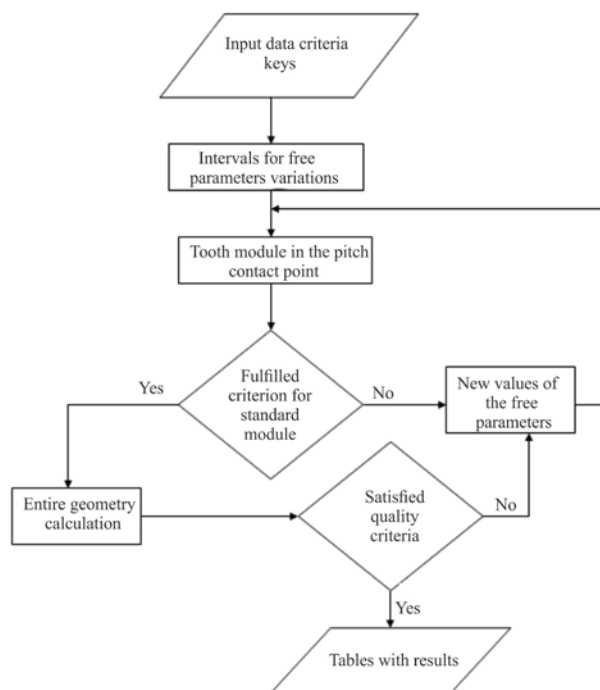


Fig. 2 Program scheme for an optimization synthesis of a Spiroid gear pair

In Fig. 2 the main block-scheme of the commented above program can be seen. The table of the results is consisted of: the basic geometric parameters of the Spiroid pinion and of the Spiroid gear, the constructive parameters of the Spiroid hob, geometric and constructive parameters of the gear pair, parameters related to the quality of meshing at the pitch contact point, such as forces in the pitch contact point, distances between bearings and pitch circles planes, efficiency, etc. It should be mentioned that, when the bearing supports are calculated, the distances between them and the pitch circles in real dimensions *mm* or dimensionless - as the ratio of the distances to the diameter of the pitch circle of the Spiroid gear should be given. In the program, all forces, that act in the pitch contact point and at the bearing supports are dimensionless in relation to the peripheral force, acting on the Spiroid gear. If in the beginning of the program, the torque on gear shaft or a torque on a pinion shaft is given, then this peripheral force has a concrete value and all forces and loads are calculated in [N]. Analogically, the sliding velocity is calculated referred to the Spiroid pinion angular velocity. If the Spiroid pinion numbers of driving motor revolutions per minute are given, then the sliding velocity is obtained in [*m/s*].

5. Conclusion

On the basis of the shown in the study basic principles of constructing computer programs for the gear drive synthesis, the accent is put on the description of the mathematical model and based on it software for an optimization synthesis upon a pitch contact point of gear pairs of type Spiroid and Helicon. The main tasks, which are solved upon this approach for synthesis and design, are shown briefly. The procedures described there are illustrated with an adequate block-scheme.

References

1. Abadjiev, V., D. Petrova. On the Synthesis and Computer Design of Spiroid Gears: A Review of the Bulgarian Approach. *Journal of Theoretical and Applied Mechanics*, Vol.33, No 3, Sofia, 2003, pp. 3-30.
2. Minkov –Petrof, K. Automated Technological Synthesis of Spatial Gearings. *Proceedings of the Sixth World Congress on "The Theory of Machines and Mechanisms"*, Volume II, December 15 – 20, 1983, Delhi, India, pp. 807-810.
3. Minkov, K., V. Abadjiev, D. Petrova. Some Aspects of the Geometric and Technological Synthesis of Hypoid and Spiroid Gears, *The Theory of Machines and Mechanisms, Proceedings of 7th World Congress of IFTOMM*, 17-22 September 1987, Seville, Spain, Pergamon Press, Volume 3, pp. 1275-1278.
4. Abadjiev, V. Bulgarian Approach to the Synthesis and, Design and Manufacture of Spiroid and Helicon Gear Drives, *Proc. from the First Conference with International Participation. Machine and Machine Elements*, 4-6 November 2004, Volume I, Sofia, 2004, pp. 70-78 (in Bulgarian).
5. Goldfarb, V. Aspects of Gears and Reduction Drives Automated Design. *Gearing and Transmissions*, Izhevsk, Assoc. Engineers, 1991, No 1, 20-24 (in Russian).
6. Abadjiev, V. Gearing Theory and Technical Applications of Hyperboloid Mechanisms, Sc. D. Thesis, Institute of Mechanics, Bulgarian Academy of Sciences, Sofia, 2007, 309p (In Bulgarian).
7. Bulgarian Standard 17108 (CT na CIV 5744 - 86). Cylindrical Involute Gears with Externally Meshing. *Standardization Strength Calculation of the Gear's Teeth*, G-15, 1999, Sofia, p. 124, (in Bulgarian).
8. Gear Engineering Standard. Bending, Stresses in Bevel Gear Teeth. Gleason Works, Rochester, N.Y., Copyright 1955, 1960, 1965, 28 p
9. Giznbrug, E., N. Golovanov, N. Firun, N. Hlebskiy. *Gears Handbook of Mechanical Engineering*. Leningrad, 1980, 416 p., (in Russian).
10. Dudley, D., J. Sprengrers, D. Schroder, H. Yamashina. *Gear Motor Handbook*, Springer – Verlag, Berlin Heidelberg, 1995, 607 p
11. Abadjiev, V. On the Synthesis and Analysis of Spiroid Gears. Ph. D. Thesis, 1984, Sofia, 158 p. (in Bulgarian).
12. Abadjiev, V., D. Petrova. Aided Design of Hyperboloid Gears with High Gear Ratios. *Theoretical and Applied Mechanics*, Vol. 4, Publishing House of the Bulgarian Academy of Sciences, Sofia, 1984, pp. 24-30.
13. Anchev, A., K. Minkov, D. Petrova, Og. Kumchev. *Synthesis and Analysis of Bevel Gears with Spiral Teeth*, Publishing House of the Bulgarian Academy of Sciences, Sofia, 1980, p. 160
14. Minkov, K. Synthesis and Analysis of Semi-Conjugated Conic and Hypoid Gears. Sc. D. Thesis, LITMO, Leningrad, 1971, p. 170
15. Minkov, K. Mechanical and Mathematical Modeling of Hyperbolic Gears. Sc. D. Thesis, Sofia, 1986, 330 p. (in Bulgarian).
16. Abadjieva, E., Mathematical Models of the Kinematic Processes in Spatial Rack Mechanisms and Their Application, Ph. D Thesis, Institute of Mechanics- BAS, Sofia, Bulgaria, 2010, 165 p.
17. Abadjieva E., V. Abadjiev. Synthesis of Hyperboloid Gear Drives: Controlling of the Singularity on the Active Tooth Surfaces. *Proc. of International Conference of Science, Engineering and Technology Innovations*. March 15-17, 2017 Taipei, Taiwan, (Published on CD)

INVESTIGATION OF SAMPLES ACCURACY TO MODEL THE PROCESSES IN 3D PRINTING

Assoc. Prof. MEng. Minev R. PhD., Assoc. Prof. MEng. Rusev R. PhD., MEng. Antonov S. PhD., MEng. Minev E. PhD.
"Angel Kanchev" University of Ruse, Bulgaria

eminev@uni-ruse.bg

Abstract: 3D printing also called Layer based technology, Freeform fabrication, Additive manufacturing or Rapid Prototyping technologies has undergone significant development over the last decades. The growth is related to the expansion of the range of materials used, application areas, and range of possible sizes from nanometer to tens of meters as well as increasing machine accessibility. There is a growing consensus that 3D printing technologies will be at the heart of the next major technological revolutions. At present there are some technological specifics and associated difficulties in 3D printing one of which is the accuracy of the manufactured product. Research in this area would allow modelling of 3D printing processes.

The article describes the possible types and sources of inaccuracies in 3D printing processes. The various types of test pieces used in practice are examined to quantify the errors in shape and sizes after building. Test pieces with predefined discrete points and methodology are provided to calculate inaccuracies. The results are presented in the terminology of "linear" and "shear" deformations. This gives opportunity to determine the variations in the shape and dimensions of the parts built by 3D printing. On the basis of the discrete results obtained, the possibility of 3D printing process modelling is discussed and presented.

Keywords: 3D PRINTING, ACCURACY, MODELLING, TEST PART, GRID METHOD

1. Introduction

Additive Manufacturing (AM or commonly known as 3D printing and before that rapid prototyping - RP) undergone significant development during the past decades in terms of production volume and technological achievements. However the AM industry only began to grow substantially after 2009 probably because the last major AM patent for Fused Deposition Modeling (FDM) expired in 2009. This way 3D printers could be produced without infringing on intellectual property, creating a newfound interest and investment in 3D printing [1]. This is particularly valid for the introduction of affordable consumer 3D printers into the general market on the basis of so-called RepRap (affordable open source replicating rapid prototyper). Hence the main drivers for the rapid growth are the reduction in cost to access the technology, an increase in applications [2] as well as expanding of the range of materials used and printed dimensions. Additive manufacturing has been called a disruptive technology [3] that will fundamentally influence many processes in production. Predicting of the future of AM on economic and social implications is constantly on the focus of the researchers [4, 5] and companies [6]. There is a growing consensus that 3D printing technologies will be one of the next major technological revolutions [7].

Although AM is a breakout technology, the implementation of it is still in infancy. There are numerous challenges in applying AM. One major obstacle is lower accuracy relative to other technologies [2]. The errors in sizes and shapes of produced component reflect various reasons during the build up stage. A possible general classification of error sources common for most popular AM processes comprises the following [8]: errors caused by laser scanning system; errors caused by material shrinkage; errors caused by spot size and heat effected zone; errors caused by Computer Aided Design (CAD), tessellation and slicing; random errors. The thermal nature of most of the process leads to shrinkage, distortion and warping of the built part. All of these factor can contribute to significant errors in the final component. Therefore it is especially important to perform an accuracy analysis that systematically reveals the type of errors, their sources and magnitudes, so that improvements in the production practice can be implemented. If enough data about accuracy of a given process is gathered, the process can be modeled in terms of its accuracy performance.

2. Prerequisites and means for solving the problem

The manufacturing parameters in AM have great significance on the part integrity, strength, density, surface quality and shrinkage for different processes and materials. The analysis of process parameters is therefore one of the main streams of RP research.

Once an optimum set of parameters are established they are kept as a material's "profile". In particular cases of unexpected changes to part quality, adjustments to the parameters can be made according to the experience of the machine operator or platform manufacturer recommendations. In most cases of AM it is practical to set the build parameters for optimum part strength, surface quality and build time. On this basis the accuracy in size and shape of the part can be targeted by applying scaling factors, beam offset compensation, part orientation and arrangement in the building chamber. As the manufacturing conditions at different areas of the build chamber vary, the overall accuracy of the process or platform is not uniform. The research in [9, 10] (Fig. 1) shows that the temperature variations of a SLS machine are considerable. It can be expected that the differences in shrinkage to room temperature and the ultimate accuracy of the build part will reflect these variations.

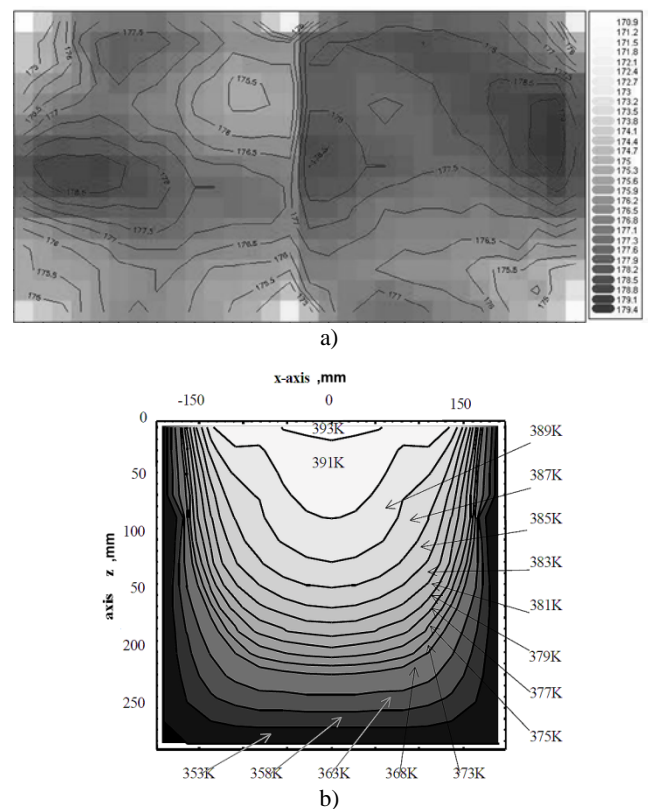


Fig. 1. Temperature distribution in the part bed - a) reported by [9] and b) reported by [10].

Essentially the practice of accuracy research in AM requires the building of various test parts or series of parts and evaluating the differences between their nominal and actual dimensions and shapes. Linear and angular dimensions, point's coordinates and surface roughness are the basic entities that are used to assess part geometry.

The elements that distinguish between different methodologies for accuracy studies of RP processes are test parts and the related measured geometrical entity. The most common types of test pieces can be classified as: pyramids (staircase); specifically designed parts; real parts. Typical examples of these are shown in fig. 2.

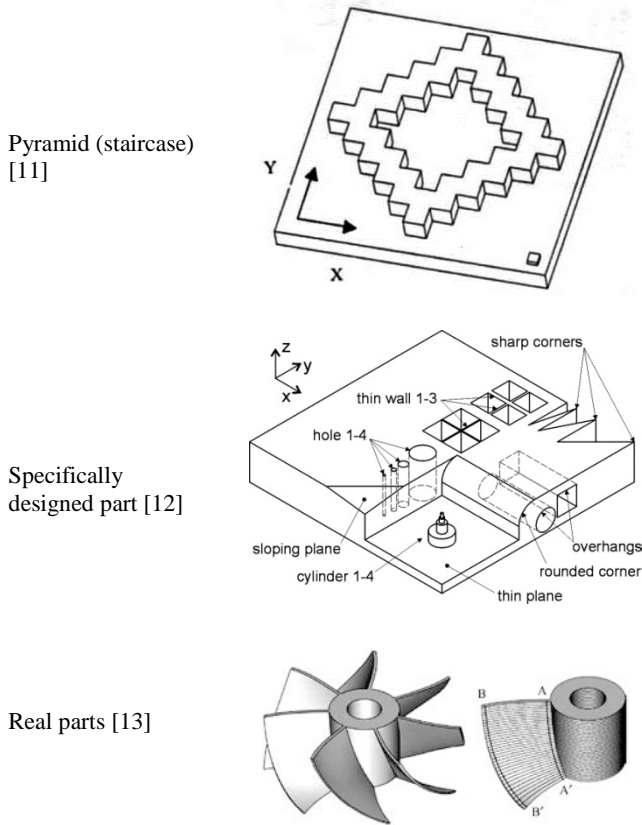


Fig. 2. Typical examples of test parts.

One of the drawback of all test pieces is that they don't cover entire surface of the part bed. Additionally they usually cross areas with different temperatures or with other manufacturing variations. As a result a continuous distribution of accuracies cannot be achieved. In case of the most popular type of test part - pyramid, the methodologies for accuracy investigation are based on measurements of the staircase dimensions and calculation of the material shrinkage in x, y or z direction. The measurements m_i are shown on fig. 3.

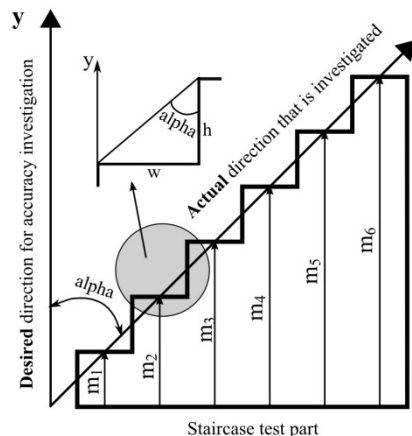


Fig. 3. Lengths and directions for measurements of pyramid shape test part.

The idea of such measurements is usually to determine the difference between the nominal dimension l_n and the final dimension l_f (produced by the machine) usually called the error (E):

$$E = L_f - L_n$$

The distribution of the errors E varies over the part bed defined as a plane (x, y). Geometrically the error is a function of two variables x and y :

$$E = f(x, y)$$

As illustrated in fig. 3, the direction in which the pyramid stairs advance is the "actual direction" defined by the angle α :

$$\alpha = \tan^{-1} \left(\frac{w}{h} \right)$$

This may not necessarily be the "desired direction" for investigation. Failure to observe this correctly can result in misleading directional results, particularly if z direction is investigated by vertical pyramid - fig. 4.

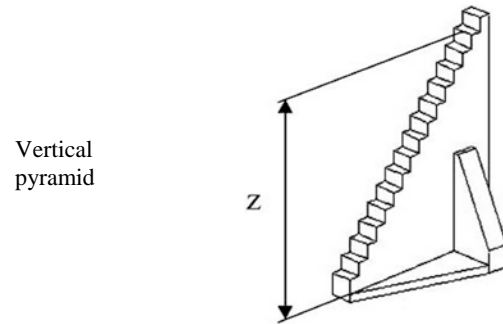


Fig. 4. Vertical pyramid for z accuracy investigation.

The most common approach of using the pyramid type of test piece is to compare the nominal (as in CAD model) and actual size of the steps as shown in fig. 5 [14]. The comparison of the results in coordinates "nominal dimensions - actual dimensions" is represented by an interpolation graph which slope is interpreted as the scaling factor. It has to be applied to the CAD model in order to achieve closer real part dimensions to the nominal. The intercept that the interpolation line cuts from the vertical coordinate axis of the graph gives the size of the systematic error due to the size of the extruded plastic jet or laser beam spot.

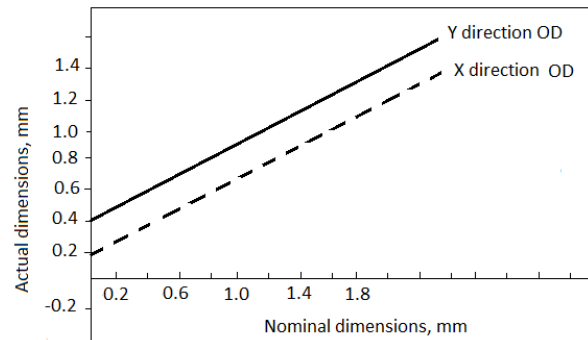


Fig. 5. Linear interpolation "nominal size - actual size" of the outside (OD) in X and Y direction of the pyramid.

In general practice it is important to know the length changes either in any directions or in some particular directions. In order to model the total accuracy of the AM process or specific platform or material behaviour in terms of shrinkage, constructing a map of entire field of distortions is essential. It is not achievable by utilising the test parts and measurement approach described above.

3. Solution of the examined problem

Regardless of the cause, it can be assumed that the change of dimensions or shape of a body manifests itself in the same geometrical appearance as strain. Strain is defined as being linear (ϵ) or angular (γ). The angular change in a right angle, known as

shear strain, is defined as the change in angle between two line segments which were previously mutually perpendicular. An advantage of describing the inaccuracy of parts manufactured by AM processes in terms of strains, is the possibility it allows to combine analysis of dimensional and shape deviation from nominal.

The analysis methodology that is proposed in this study for accuracy modeling of AM processes is based on measurement of coordinates x , y or z at many points throughout a test piece in the form of a flat plate. To expose the strains, a reference grid in some form is applied to the CAD model of the test piece beforehand. After the plate is manufactured the actual coordinates of grid points is measured and compared with the CAD model. As the produced grid differs from the nominal CAD grid it can be considered as deformed and therefore the strains calculated. Several ways to estimate the strains of a deformed object are known however the Coefficient or Square Grid Method is utilised here. This method was introduced by Bredendick in [15, 16].

The advantage of such a methodology is the simple geometry of the test part - a plate with reference points in form of hole, cross or other spot mark. The plate can be arbitrary in size, location and orientation within the build chamber (fig. 6). It is also suitable for handling and automatic measuring as well as being representative of various engineering parts.

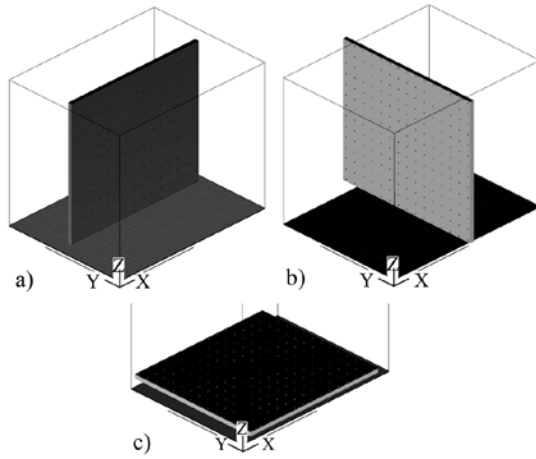


Fig. 6. Examples of test plate orientation.

If the plate is built horizontally the deviations ϵ_x and ϵ_y (fig. 6c) can be determined, if positioned vertically then either ϵ_z and ϵ_x (fig. 6a) or ϵ_z and ϵ_y (fig. 6b) can be calculated together with associated angular strains - γ_{xy} , γ_{zx} , γ_{zy} . By building a set of plates with a given orientation and in different positions in the build chamber, a three dimensional concept and model of the distribution of dimensional deviations over the entire build envelope can be created - fig. 7.

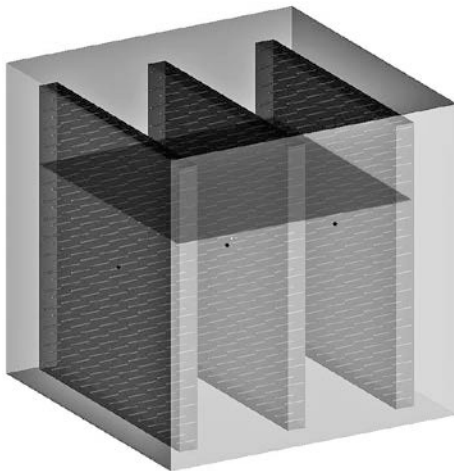


Fig. 7. Set of test parts for 3D concept and model of the accuracy distribution.

Accumulating a data from experimental set up similar to that shown on fig. 7 will allow interpolations of deformations from plates oriented in ($z - y$) to other directions by combining the points from different plates and forming virtual plates in perpendicular direction - the transverse surface on fig. 7.

4. Results and discussion

The experimental test plate was built by Selective Laser Sintering (SLS) process on DTM 2500 *Plus* machine according the settings on fig. 6a. The material used was polystyrene Castform. The hole centres was measured routinely by coordinate measuring machine (CMM) Mitutoyo QuickVision *Pro*. For that purpose a specialised software was developed that allows, together with Mitutoyo software, the total measurement to be done within minutes after the initial datum point and axis are set. The setting and the screenshot of the measuring software action are shown in fig. 8.

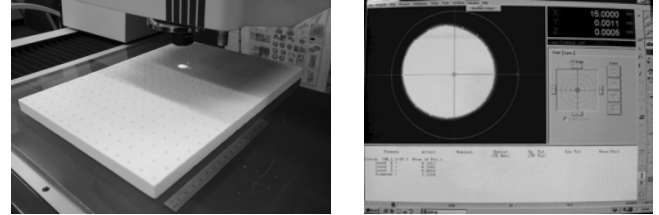


Fig. 8. Plate measuring process and screen shot of software window.

The accumulated data of x and y coordinates of grid points as holes is used as input data for linear and angular deviations from nominal calculations. For this purpose a VBA program for MS Excel spreadsheet was developed. The next step in data processing includes a MatLab program that was developed to analyse the results and to generate a 3D visualisation of the distributions of linear and shear deformations over the test part surface. The results are shown on fig. 9 and fig. 10.

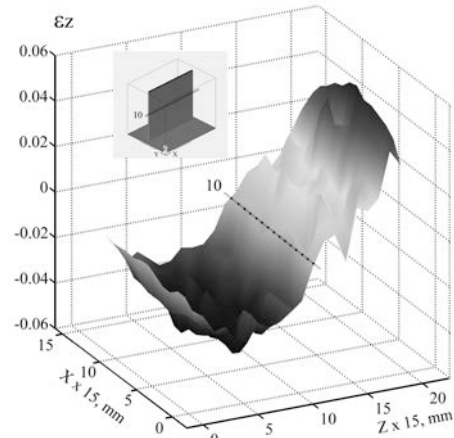


Fig. 9. 3D visualisation of ϵ_z (vertical size deviations from nominal) distribution in the (x , z) plane.

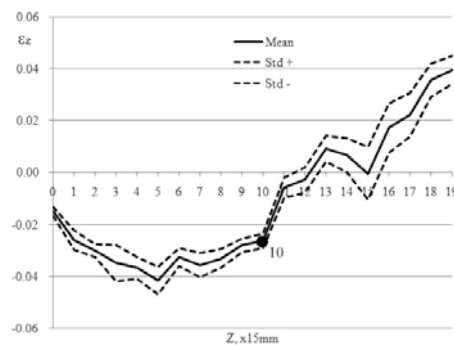


Fig. 10. Distribution of ϵ_z -vertical deviation of the sizes from nominal.

The mean value of all reference points along each row of the plate gives a point of the line shown on fig. 10. As an example the highlighted row "10" from fig. 9 gives the position of the point "10" on the fig. 10.

The results show that a single test piece and routine measurement procedure can be utilised to estimate the distribution of linear and angular deviations from the nominal sizes of AM parts. The geometry of the utilised test piece has advantages over the stair case (pyramid) type test pieces since it gives correct data about inaccuracy distribution along a particular direction or particular axes. The routine automatic measurements and data processing are possible due to the simple geometry of the samples and feasible software development.

5. Conclusion

Some estimates about accuracy of a process or platform can be done even with a single test piece. As it can be seen from fig. 9 and fig. 10 the described methodology can present the continuous distribution of distortions described in terms of linear and angular strains which can be interpreted as deviations from nominal in size and shape.

The data connection from several plates by interpolations is a step forward that can allow modelling of the entire build envelope, process or material accuracy. For that purpose a sufficient number of test plates over the entire build volume has to be produced and analysed.

The modelling of distribution of deviations from nominal, creates a possibility to reduce these deviations. One approach is a pre-deformation of the part data contrary to the expected defect. This can be implemented by using a free-form deformation (FFD) approach [17]. The FFD deforms the surrounding cuboid of an object in order to adapt the embedded object. Thus, an increased size and shape accuracy can be achieved by choosing suitable parameters for the FFD.

References

1. Van Lancker, P., 3D Printing: The Influence of Intellectual Property. CREAX, 12 August 2015. <https://www.creax.com/en/our-work/the-3d-printing-evolution-insights-on-the-influence-of-ip-on-technology-dev> (accessed 29/11/2017).
2. Attaran, M., Additive Manufacturing: The Most Promising Technology to Alter the Supply Chain and Logistics. *Journal of Service Science and Management*, 10, 2017, 189-205.
3. Petrick, I. J., T. W. Simpson, 3D printing disrupts manufacturing: how economies of one create new roles of competition. *Res. Technol. Manag.*, 56 (6), 2013, 12-16.
4. Ford, S., M. Despeisse, Additive manufacturing and sustainability: an exploratory study of the advantages and challenges. *Journal of Cleaner Production*, 137, 2016, 1573-1587.
5. Jiang, R., R. Kleer, F. T. Piller, Predicting the future of additive manufacturing: A Delphi study on economic and societal implications of 3D printing for 2030. *Technological Forecasting & Social Change* 117, 2017, 84-97.
6. Wohlers Report, 3D Printing and Additive Manufacturing State of the Industry Annual Worldwide Progress Report, <https:// Wohlersassociates.com/2013contents.htm> (accessed 26/11/2017).
7. Thierry, R., L. Striukova, From rapid prototyping to home fabrication: How 3D printing is changing business model innovation. *Technological Forecasting & Social Change*, 102, 2016, 214-224.
8. Tang, Y., H. T. Loh, J. Y. H. Fuh, Y. S. Wong, L. Lu, Y. Ning, X. Wang, 2004. Accuracy analysis and improvement for direct laser sintering, *Innovation in Manufacturing Systems and Technology (IMST) 01*, <http://hdl.handle.net/1721.1/3898> (accessed 26/11/2017).
9. Shwe, P., S., Quantitative analysis on SLS part curling using EOS P700 machine. *Journal of Materials Processing Technology*, Volume 212, Issue 11, November 2012, Pages 2433-2442.

10. Shen, J., J. Steinberger, J. Göpfert, R. Gerner, F. Daiber, K. Manetsberger, S. Ferstl, Inhomogeneous shrinkage of polymer materials in selective laser sintering. *Proceedings of Solid Freeform Fabrication Symposium*, 2000, Austin, Texas, pp. 298-305.
11. DTM Corporation, 1999. The Sinterstation System, Guide to Materials: CastForm PS. DCN: 8002-10006.
12. Kruth, J. P., B. Vandenbroucke, J. Van Vaerenbergh, P. Mercelis, Benchmarking of different SLS/SLM processes as rapid manufacturing techniques. *International Conference of Polymers & Moulds Innovations*, April 20-23, 2005, Gent, Belgium, pp.1-7.
13. Jiang Cho-Pei, Development of a novel two-laser beam stereolithography system. *Rapid Prototyping Journal*, 2011, 17 (2), pp. 148-155.
14. Minev E., E. Yankov, R. Minev, The RepRap Printer for Metal Casting Patternmaking - Capabilities and Application. *Труды VIII Международной научно-практической конференции „Прогрессивные литейные технологии“*, НИТУ МИСиС, 16-20.11.2015, Москва, стр.300-303, ISBN 978-5-9903239-3-3.
15. Bredendick, F., Zur ermittlung von deformationen an verzerrten gittern. *Wiss. Technical University Dresden*, 1967, v. 16, pp. 1473-1481.
16. Bredendick, F., Zur ermittlung von deformationen an verzerrten gittern. *Wiss. Technical University Dresden*, 1969, v. 16, pp. 1473-1481.
17. Schmutzlera, C., A. Zimmermannb, M. F. Zaeha, Compensating warpage of 3D printed parts using free-form deformation. *48th CIRP Conference on Manufacturing Systems - CIRP CMS 2015. Procedia CIRP* 41, 2016, 1017 – 1022.

ELECTRIC POWER SYSTEMS MODELING AND EDUCATION: SHORT TERM UNIT COMMITMENT AND ECONOMIC DISPATCH MODELING FRAMEWORK

M.Sc. Trashlieva V.¹, M.Sc. Radeva T. PhD.¹

Department of Electrical Power Engineering – Technical University of Sofia, Bulgaria
vesselina.trashlieva@gmail.com

Abstract: In this paper we try to build a framework in optimization model building for students without previous mathematical optimization training and experience. While modeling is a matter of art, approach and compromise between complexity and closeness to reality we need to spend some time on the basics of optimization modeling such as cost function, constraints, variables and coefficients. There is a huge variety of optimization problems solved in Electric Power Systems. The most complex of all is considered to be the Unit Commitment problem where at each iteration an Economic Dispatch is performed.

Keywords: OPTIMIZATION MODELING, ELECTRIC POWER SYSTEMS OPTIMIZATION

1. Introduction

Students mastering Electric Power Systems (EPS) often have no previous mathematical modeling and optimization experience but are required to solve certain optimizations problems for EPS operation. Such problems are the Unit commitment (UC) problem and the Economic dispatch (ED) problem. The UC is an optimization problem used to determine the operation schedule of the generating units at every hour interval with varying loads under different constraints and environments [1]. At each hour the output power of each generator is determined via a solution to an ED problem. An ED is the problem for determining the power output of each power plant, and power output of each generating unit within a power plant, which will minimize the overall cost of fuel needed to serve the system load [2]. In an optimization problem the goal to achieve might be a single or a multiple criteria formulation, or multiple criteria may be combined together in a single cost function. There are also two big groups of optimization problems concerning the implicit existence of constraints: constrained and unconstrained optimization problems. In EPS modeling constraints are often present describing the process dynamics and the physics of the systems' components. Using Lagrange multipliers a constrained problem may be transformed into an unconstrained one leading to a general nonlinear unconstrained optimization problem. The latter is very sensitive to initial conditions while nonlinear functions seem to have more than one local extreme values and most of the methods for solving nonlinear problems find a local not a global extreme. The inconvenience of providing initial values for the search method and the fact that the risk of finding a local extreme not the global one make the Linear Programming approach a leading one in EPS optimization modeling. Linear (affine) functions have only one minima and maxima that are global and provides for the omission of second order Karush-Kuhn-Tucker conditions acquisition. Certain binary modeling techniques provide for the modeling of a set of important nonlinearities that often appear in practice while the resulting problem is still linear. Cost function and constraints may be linear or nonlinear functions, variables under consideration may be continuous, integer or binary, bounded or unbounded.

So all these matters and topics presented in the last paragraph determine the opulence of approaches and methods for optimization problems solving. This same richness requires a systematic approach in the processes and elements presentation in the mathematical model [3-6].

2. EPS model elements

When optimizing the performance of the EPS the respective power levels of the different power plants and loads are considered. To present the amount of active power produced by a certain power plant a continuous variable is used. These are optimization variables with general lower and upper bounds resulting from the real plant's capacity. If different states in the power plant's operation are possible integer and binary variables come at hand to represent the different operation states. Such states are for example "on" and "off"

or a role in frequency control reserve: "primary", "secondary" and "tertiary". Generally the set of all possible states is a closed one, meaning it includes a finite number of elements. In practice, binary variables have proven themselves as the most convenient for power plants' states representation leaving the space for general integer variables for other purposes. When the set of all possible states includes more than two elements it can be divided in more than one binary sets. For example, a pump hydro-power station can be idle (non-operational because of a fault, a planned repair or simply because its power is not needed in the power balance), can work as a load or as a generator. These three states can be presented with three numbers: -1 for pump, 0 for idle, 1 for generator. Two of these states coded with non-zero numbers (-1 and 1) can be further represented with binary variables with logical ones for true and zeroes otherwise. It is obvious that with such approach of breaking the initial set into few binary sets there is an overlap with the two binary sets' false values and the zero state in the initial set. So such breaking technique must be used very cautiously in order not to achieve a redundancy of binary variables that make problem solving more difficult.

Branch and bound and branch and cut are the contemporary techniques for solving mixed-integer optimization problems while the relaxed solution might be achieved via different optimization methods. No matter of the algorithm used, branching is an exhaustive process so the number of the integer and binary variables make the problem solving more complex and time spending according to the number of possible integer states. This is the reason that a trend towards integer and binary variables number as well as the number of the possible integer values reduction exists in order to make the branching easier and faster.

General integer variables are mostly used in EPS modeling when there are more than one system elements with similar features. In certain cases different power plants might be grouped according to common or similar characteristics: type or costs for fuel, similar working ranges, etc. This is another manifest of the latter approach for an overall variables reduction that provides for the usage of a single variable for the whole group of plants while the participation of each element of a group is a matter of a different optimization problem (ED). This ED might be nonlinear and will not include binary or integer variables, so its solving is generally more easy. Practitioners in EPS optimization modeling often use such approach in order to reduce the number of variables.

In EPS optimization modeling the variables are:

- integer and continuous according to their feasible sets;
- real and artificial according to the nature of physical processes they represent: real variables model real powers, loads and volumes, while artificial variables model states and alternatives;
- optimization and dual according to the stage of analysis;

3. Loads modelling

Some loads are controllable and others are not. From the systems' point of view controllable loads *help* for the power balance while the fixed loads *determine* a part of power balance. The work of the uncontrollable (fixed) loads might be forecasted and they participate in power balance constraints as forecasted values on the right-hand sides.

The controllable loads are as important as the generators are especially in the current renewable power generators penetration growth. Base power plants are not capable of flexible load following in both up and down directions. Wind and solar power stations are considered as stochastic uncontrollable base power plants so the power they might inject into the EPS is also considered as a forecasted value and unfortunately inject most of their power in low load hours. In high load hours peak power plants as hydro-power stations help for preserving the power balance. In other low load hours, the power of the base loads such as nuclear power stations need to be decreased and sometimes for wind and solar power stations even to be rejected if there is not enough load.

In detailed modeling terms controllable loads are considered in the following groups:

- controllable (dispatch) over time
- controllable over power level
- controllable over time and power level
- loads with interruptible or non-interruptible work cycle

For example a pump hydro-power station is a load that is dispatchable over time because pumps can be turned on when needed. A smart household appliance might be considered as a dispatchable over time load with a non-interruptible power cycle if it can not be stopped once started until the cycle is completed. In general, controllable loads modeling require the introduction of additional binary and integer variables.

4. Power plants operation modelling

According to the flexibility and the load following abilities generators are divided into two big groups: those of base and peak power plants. As mentioned above, there is one more aspect of power plants work namely if the output power level is controllable or stochastic. Wind and solar power stations are considered uncontrollable therefore their power generation is forecasted before the building and solving of the UC or ED problem [2]. These forecasted values participate as right hand sides along with the forecasts of the uncontrollable loads. All thermal, nuclear and hydro-power plants are considered controllable. The function representing the relation between spent amounts of fuels and produced power is generally non-linear because the electric power production process is nonlinear. These functions are often captured in practice with direct observations and in order to achieve a functional expression the appropriate function approximation techniques must be applied. Such experimental data is approximated with polynomials of order $N=2$ or 3:

$$R(P) = \sum_{n=0}^N a_n P^n \quad (1)$$

Higher powers are suitable when higher precision is required or when inflexion points are more than 3. If the model has to be linear, a piece-wise approximation might also be used:

$$R(P) = \sum_{l=1}^L b_l P_l \quad (2)$$

In the latter expression $P = \sum_{l=1}^L P_l$ is the produced power in all

linear intervals $l = 1:L$, b_l is the fuel consumption growth in each consecutive interval l , and P_l is the power in the interval. So in a ED or UC optimization problem the number of continuous variables for each power plant that is modeled depends on the number of observation intervals, i.e. the value of L .

Table 1: Observed fuel consumption of a thermal power plant within its working range P_{min} P_{max}

Interval l	$P_{l,min}$	$P_{l,max}$	b_l
	MW	MW	tons
	P_{min}	-	23,88
1	60	70	26,68
2	70	80	29,54
3	80	90	32,44
4	90	95	34,03
5	95	100	35,63
7	100	105	37,85
8	105	110	40,31
9	110	P_{max}	45,87

Given the data that is experimentally carried out for the fuel consumption of a power plant (Table 1, Figure 1), a piece-wise fuel consumption function may be used. In this case the number of variables increases with the increase of the number of observations.

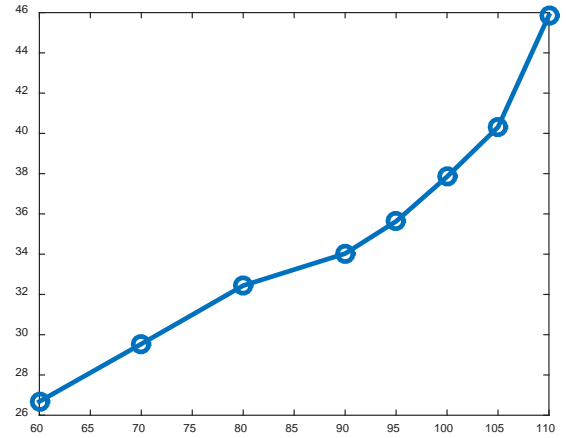


Fig. 1 Observed fuel consumption of a thermal power plant within its working range P_{min} P_{max} graphical representation

A single continuous optimization variable may be used when the fuel consumption curve is nonlinear (1). Each point of this polynomial representation of a nonlinear function has its coordinates (P_n, b_n) . In this case an appropriate curve fitting procedure has to be used that minimizes the square of the difference between the real observation (P_b, b_l) and the evaluated one (P_n, b_n) in order to evaluate the values of the coefficients a_n bringing the polynomial approximation (1) closer to the observed values:

$$\min \sum [F(P_i) - R(P_n)]^2 \quad (3)$$

Because the cost function of the latter unconstrained optimization problem is nonlinear the appropriate initial conditions a_0 have to be provided for the problem optimization (3). As mentioned above the number n (i.e. the polynomial order N) of the coefficients whose value has to be optimized depends on the number of the inflexions of the initial data shown on Fig. 1. The procedure of curve fitting is performed only once whereas the observation data is stationary meaning that the values in the observations onto the fuel consumption do not change over time.

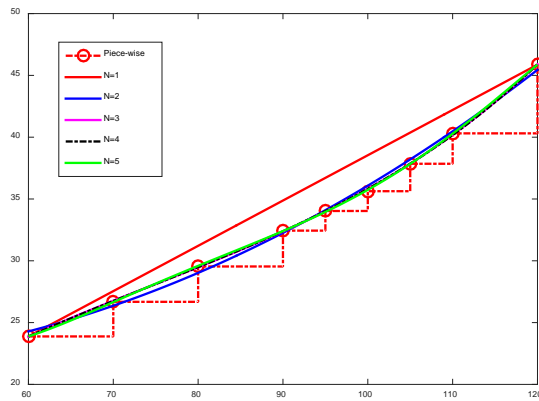


Fig. 2 Polynomial approximations of different orders

5. Costs structure and criteria

In optimization problems an optimal solution is found that fits best to certain criteria. In optimal power planning the most commonly used criteria is the minimal productions costs one. This is a variation of the minimal fuel costs (minimal raw materials consumption) criteria, that are determined by the consumption curve of each power plant. The fuel consumption curve multiplied by the price for the raw material (fuel) gives the fuel costs. These functions are generally nonlinear. Usually quadratic or polynomial approximations are used for the nonlinear case. In US problem on/off costs must be included. So two types of costs might be included in a cost function: costs that depend on the power plant's power level and costs that do not. This general idea for costs segregation introduces additional binary variables in the optimization models. More terms concerning different economic and financial relations may be also modeled as well as ecological criteria. All this makes the variety of cost functions wide enough to build different combined (multi criteria) goals.

6. Constraints

For the purpose of building a general idea and a modelling framework two large groups of constraints formulated in the optimization tasks in the EPS can be formulated. These are the system's conditions stating the balance requirement for each interval in the optimization horizon and the requirements for securing the reserves and the constraints arising from the production capacities. The second set of constraints may be further divided into subgroups according to the type of power plants (thermal, hydro), and a separate group of equality constraints responsible for water balances. Often in the optimization problems the working ranges of the plants are modelled by the simple bounds. Constraints that include efficiency coefficient η of the different cycles or mutually exclusive alternatives in the operation of units capable to work in reverse mode (pumps and turbines) form a separate subgroup:

minimize *Total Costs* subject to:

Balance constraints: $Production = Load + Losses$

Efficiency coefficient: $Accumulation. \eta = Generation$

Alternatives: $While\ true \leq Max\ variable\ value$

If false: Variable = 0

Water balance: *Waters used by pumps and turbines in different reservoirs and time intervals should equal or less than given constraints*

Simple bounds: $Min\ value \leq Variable \leq Max\ value$

Integer constraints: *Some variables are integer and some variables are binary {0,1}*

On the other part constraints generally are represented mathematically with equations and inequalities. In optimization equality constraints are called *tight* or *binding* because every change in the right-hand side of an equality constraints leads to inevitable changes in the total costs. When there is an inequality constraints that is satisfied in the optimal solution as an equality this constraint is also *binding* because there is no *reserve* for changes in its right-hand side. Constraints also may be explicit or implicit in the formulation with implicit constraints being generally the possible sets of the values of all integer variables and the simple bounds arising from the working ranges while the latter may be easily included in the models constraints while the integer ones require further mathematical knowledge and work.

7. Conclusion

The presented topics above give general idea on EPS optimization modeling basics and possible points of view. A certain classification of power plants, loads and costs criteria is given. Some optimization modeling milestones are mentioned as well as few techniques to model size reduction are given.

8. Bibliography

- [1] Saravanan, B., Das, S., Sikri, S. et al., A solution to the unit commitment problem - a review, *Frontiers in Energy* (2013) Vol 7: pp. 223-226, <https://doi.org/10.1007/s11708-013-0240-3>
- [2] Wang J., Botterud A., Miranda V., Monteiro C., Sheble G., Impact of Wind Power Forecasting on Unit Commitment and Dispatch
- [3] Wang Y., Niu D., Ji L., Short-term power load forecasting based on IVL-BP neural network technology, *Systems Engineering Procedia* 4 (2012), p. 168 – 174, The 2nd International Conference on Complexity Science & Information Engineering
- [4] Usaola J., Castronuovo E. D., *Wind Energy in Electricity Markets with High Penetration*, Nova Science Publishers Inc, ISBN: 978-1-60741-153-6, 2009
- [5] Viana A., Pedroso J. P., A new MILP-based approach for Unit Commitment in power production planning, *International Journal of Electrical Power & Energy Systems*, Volume 44, Issue 1, January 2013, p. 997-1005
- [6] Vieira F., Ramos H. M., Hybrid solution and pump-storage optimization in water supply system efficiency: A case study, *Elsevier Energy Policy*, ISSN: 0301-4215, Volume 36, Issue 11, November 2008, Pages 4142-4148

ELECTRIC POWER SYSTEMS MODELING AND EDUCATION: HYDRO POWER STATIONS MODELING FRAMEWORK

M.Sc. Trashlieva V.¹, M.Sc. Radeva T. PhD.¹

Department of Electrical Power Engineering – Technical University of Sofia, Bulgaria
vesselina.trashlieva@gmail.com

Abstract: *In this paper we try to bring a modeling framework in optimization problems for hydro-thermal coordination problems solved in electric power systems. There are certain specifics when dealing with hydro power plants and the reservoirs they are fed from. In many cases these water reservoirs are connected and this involves additional complications in optimization models building. We give guidelines for model size reduction and awareness. This framework is suitable for practitioners willing to be aware of hydro-power modeling and optimization but having no previous mathematical optimization training and experience.*

Keywords: HYDRO-THERMAL COORDINATION, WATER RESERVOIRS MODELING TECHNIQUES

1. Introduction

The unit commitment problem is the procedure for defining the optimal engagement of the units and a plan for optimal operation of the plants in an Electric Power System (EPS). The solution of this problem involves two things: the determination of the operating and reserved units and the output power of each operating unit (an economic dispatch)[1]. The cost function represents in most of the cases a striving to minimize the total operating and production costs while strictly respecting the balance between consumption and production, the reliability and security requirements, according to the constraints arising from the generating units themselves. Due to the nature of the different power stations, often the optimization problems involving the operation of both hydro-power and thermal power plants are subjected to a separate class of problems solved in EPS, called *optimal hydro-thermal coordination*.

2. The HPS and PHPS

Hydro power stations (HPS) use water as a primal source to produce electric power. The cost of water is negligible smaller related to the price for fuels used by the conventional thermal power plants. Water, however, is also a limited resource meaning that the available quantities for the individual planning periods must be used wisely. HPS are the most widely used plants for electricity generation from renewable energy sources, accounting for 16% of global electricity production by 2010, and is projected to grow by an average of 3.2% annually over the next 25 years [2,3]. HPS are used in over 150 countries where China leads with 721 TWh in 2010. There are four HPS with a capacity over 10 GW worldwide.

Since the output level of a hydropower plant depends on the water supply and head, the costs curves are nonlinear as well as in the conventional thermal power stations. The advantages that the HPP provide for the balance of the EPS are not limited only to the possibility of rapid change of their working outputs and fast automated start and stop. Although this flexibility is the main factor determining HPS as peak power plants. HPS also have very low operational (production-dependent) costs, as there are virtually no costs of purchasing and supplying fuel, and the aforementioned operating costs are related to the cost of providing ancillary services, operating and maintenance. Fixed costs, however, include the return of investments for construction and modernization, depreciation of facilities, insurance, etc. Sometimes such fixed costs might be significant, especially for brand new installations. HPS have a very long life cycle and do not require large operating labor costs as HPS are operated automatically. Water dams on the other hand are multifunctional facilities, which are used for the drinking water and domestic water supply purposes, irrigation, industrial water supply, fish-breeding, aquaculture, etc. besides for electricity production purposes (according to most legislations electricity production is of lowest priority among other water purposes). Regarding construction costs, for a hydropower plant the building of the water collection system and its derivative is more expensive than the actual construction and installation of the water turbines. In

fact, all deficiencies of HPS are mainly related to water collection and accumulation installments, not to the installed electricity generation apparel. Among the disadvantages the following are often stated: ecosystems damages, habitat fragmentation, land losses in the construction process of large reservoirs, congestion of large areas of land, forests and territories with rich soils, need for relocation of people, reorganization of road infrastructure, flows reduction, methane emissions of rotting organisms and flows risks. Since there is no fossil fuel combustion HPS are considered with no direct carbon dioxide production, except the relatively small quantities produced the construction phase. However, these quantities remain negligible compared to thermal plants.

The electricity production of HPS is determined by the available inlet water quantities. If there is a possibility of accumulation (sufficient useful volume of the water tank), the HPS can produce enough power when needed (under circumstances of electric power shortage). In Bulgaria HPS are mainly used as peak, balance and also accumulation capacities. Pump-storage hydropower plants (PHPS) are capable to accumulate the processed by a water turbine quantities (in the presence of a lower reservoir) and are used for balancing loads, being practically the most flexible and multifunctional elements of the EPS. During high load periods, PHPS operate as generators producing the famine power for the system's balance. During low load periods when the reduction of the power outputs of other generating units is practically impossible, PHPS operate in pumping mode like controllable loads by feeding the excess power to the pumps draining water from the lower into the upper reservoir. In addition to importing the required extra load into the system, they accumulate part of the surplus electricity, which can be later injected during the peak periods at a higher price. This reversible cycle (turbine-pump, load-generator) helps preventing frequent start-stop and the needs for frequent changes in the working levels of nuclear and thermal units.

The operation of PHPS is effective if the income of the operation of the plant is positive. This income is mainly determined by the difference between the sold electricity that is produced and the purchased electricity that is consumed by the pumps [4]. From a financial point of view, the accumulation process is an expense with a future return. If the accumulated energy is not realized in the future at better prices (this condition guarantees a positive revenue), these costs will also include the costs of lost financial benefits as the unit was idle and waited for the respective better financial period. Thus, in the end, the electric power accumulation may prove technically necessary and even crucial from the point of view of the EPS balance but economically unprofitable for the PHPS owner. Sometimes some water wastes from the upper reservoir might be required in order to provide for further load leveling. Energy storage is economically justified when the cost of the electricity used by the pumps is lower than the electricity generated and sold, taking into account the accumulation-generation cycle losses. System load demonstrates daily, weekly and seasonal changes that have to be followed by the power plants. A system operator having a PHPS accumulate the cheaper energy produced during low load

periods and deliver it to the grid during the peak periods when the process is cost-effective. If there is an opportunity to accumulate energy (large dams are built), generation may be postponed. This is not a "surplus" energy storage. This power can be used in the balancing market later. Power plants using renewables (except HPS of stored water) can not generally accumulate their primary source and by using their current availability they usually increase the EPS balancing problems [5-7].

With the increase in penetration of power stations using renewable power sources as wind and sun (also considered as stochastic power plants), PHPS become an indispensable element in the balance management of EPS, that present a buffer to reconcile the inconsistency of the renewable generation and load curves. In this paper, some attention is paid to the operation optimization of HPP and PHPS as balancing capacities and a means of minimizing total costs in the EPS[8-14].

3. Modelling interconnected HPS and PHPS

The modeling of the operation of the HPS and PHPS requires modeling of the plant's curve and associated water volumes. Some HPS operate in a cascade mode, i.e. their work is interconnected (consecutive and dependent) within a given terrain. So despite that the cost of the primary energy carrier for HPS is negligible, it is also necessary to model the "stock" of water availability. Water reservoirs may be of a complex purpose or electricity production only. The water reservoirs also imply a certain cycles for the water levels management (daily, seasonal, yearly) associated with the so-called water allowances. The inclusion of HPS and PHPS in the EPS optimization leads to two events: HPS allow for *peak shaving* to be adjusted, but also lead to the addition of additional constraints and variables for modeling the work of water power plants. The curve of water waste for a HPS is a function of the water quantity Q fed to the turbine: $P_{k,j} = f(Q)$.

Nomenclature for HPS and PHPS:

r - reservoirs

k - power stations (HPS and PHPS)

$V_{r,Usable}^{\min}, V_{r,Usable}^{\max}$ - minimal and maximal usable water reservoir r volume

$V_{r,j}$ - water reservoir r volume at the end of the time interval j

$F_{r,j}$ - water flow in r during a unit interval j

$R_{r,j}$ - unprocessed water from r in a unit interval j . Unprocessed water quantity includes controllable water release as well as uncontrollable losses such as evaporation.

$P_{pk,j}$ - power used by the pumps of k in a unit interval j

$P_{Hk,j}$ - power produced by the turbines of k in j

$P_{P,k}^{\min}, P_{P,k}^{\max}$ - minimal and maximal pump capacity of k

$P_{H,k}^{\min}, P_{H,k}^{\max}$ - minimal and maximal generating capacity of plant k

φ_{Hk} и φ_{Pk} - water consumption (m^3/MWh) for plant k in both generation and accumulation mode

$w_{k,j}$ - artificial binary variable for the operation mode of a PHPS k . $w_{k,j} = 1$ if the mode is pumping in j

The operational curve of a HPS with neglecting the water head may be expressed via the following linear functions:

$Q_{Hk,j} = \varphi_{Hk} P_{Hk,j}$ (1) - processed water quantity by the turbines of plant k in a unit interval j

$Q_{Pk,j} = \varphi_{Pk} P_{Pk,j}$ (2) - processed water quantity by the pumps of plant k in a unit interval j

The output level of all units / plants must be within the technological limits:

$$P_{H,k}^{\min} \leq P_{Hk,j} \leq P_{H,k}^{\max} \quad \text{and} \quad P_{P,k}^{\min} \leq P_{Pk,j} \leq P_{P,k}^{\max} \quad (3)$$

When modeling the PHPS operation, it is necessary to introduce restrictions for non-simultaneous operation of pumping and production capacities for each PSHP in a single interval:

$$P_{Pk,j} - w_{k,j} P_{Pk}^{\max} \leq 0 \quad \text{and} \quad P_{Hk,j} - (1 - w_{k,j}) P_{Hk}^{\max} \leq 0 \quad (4)$$

The volume of each reservoir in every unit interval must be within the actual water level limits:

$$V_{r,Usable}^{\min} \leq V_{r,j} \leq V_{r,Usable}^{\max} \quad (5)$$

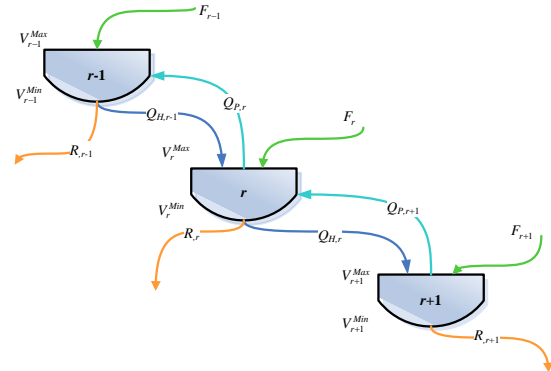
Values for the first ($j=1$) and last interval ($j=j_{\max}$) might be provided for the water level maintenance cycle:

$$V_{r,j=1} = V_{r,1} \quad \text{and} \quad V_{r,j=j_{\max}} = V_{r,N} \quad (6)$$

General form of the water balance constraints:

$$V_{r,j} = V_{r,j-1} + \left(- \sum_{m \in \Gamma_{from}} \varphi_{Hm} P_{Hm,j} + \sum_{n \in \Gamma_{in}} \varphi_{Hn} P_{Hn,j} - \sum_{q \in \Pi_{from}} \varphi_{Pq} P_{Pq,j} + \sum_{s \in \Pi_{in}} \varphi_{Ps} P_{Ps,j} \right) + F_{r,j} - R_{r,j} + \sum_{\rho \in in} R_{\rho,j} \quad (7)$$

, where Γ_{from} and Π_{from} are the sets of plants (with their pumps and turbines) that drain water from r during j . Γ_{in} and Π_{in} are the sets of plants that feed in water in r during j . So $\sum_{\rho \in in} R_{\rho,j}$ is the sum of controlled water wastes associated with the current reservoir r (both in and out) (see figure below).



In a short-term planning (day, week) of HPS and PSHP operation, water balance equations bring more security than complexity in the model, whereas in a medium-term planning (e.g. a year) the modeling of cascade-connected water reservoirs with complex purposes will increase complexity of the model and its size. Different purposes and usage of water volumes may be modeled with so-called quotas. When a large dams with yearly or seasonal water level management is available, a long-term strategy for the possible use of water needs to be applied.

Water balance constraints are build for each interval and each reservoir, i.e. their number depends on the number of unit intervals in the planning horizon. The cycle mater management implies the existence of a certain relationship between the functional dependencies at the beginning and end of the planning horizon. In other words, if modeling of water volumes is performed for a period of one year and the single interval's duration is 1 hour, for each reservoir, 365x24 or a total of 8 760 variables have to be introduced as well as 8 760 of water balance equations, and the constraints (6) must be respected for each equalization cycle.

For large reservoirs with annual management cycle this complication can be avoided by using long-term quotas i.e. by adding availability of water quantities by months, seasons, and even annually. This will result in the introduction of significantly fewer constraints. In the case of cascade-coupled different water volumes, the constraints of the upper and lower reservoirs (8) will keep the levels of the large tanks within their respective limits.

$$Q_r^z \leq Q_r^{z-1} + F_r^z + (1/z)V_r^{\max} - \sum_{k,j} Q_{Hk,j}^r + \sum_{k,j} Q_{Hk,j}^{r-1} + \sum_{k,j} Q_{Pk,j}^{r+1} - \sum_{k,j} Q_{Pk,j}^{r-1} \quad (8)$$

, where Q_r^z is the water volume in m^3 in reservoir r at the end of sub-period z , F_r^z is the flow in the reservoir during the sub-period and $F_r^z + (1/z)V_r^{\max}$ models the available water quantities for the whole sub-period.

The expression $-\sum_{k,j} Q_{Hk,j}^r + \sum_{k,j} Q_{Hk,j}^{r-1} + \sum_{k,j} Q_{Pk,j}^{r+1} - \sum_{k,j} Q_{Pk,j}^{r-1}$ models the operation of the pumps and turbines fed by and feeding in the reservoir r using the respective processed water quantities for the whole sub-period z . Thus, if a one-year horizon is considered for 1 hour single interval duration and seasonal plant operation is modeled, the water balance constraints for a reservoir r are reduced to four from the value 8 760. If an annual quota is used and seasonal interconnectivity is not required (or they are reported with similar indicators in terms of inflows, outflows and maximum volume), the constraint (8) will be only one. Thus, in an optimization problem for an optimal hydro-thermal coordination in which several cascades are modeled and the included water reservoirs are of a different purpose, volume, and management cycle it is possible to model the work of all hydro-power plants in the medium-term planning horizon and this is done without the addition of redundant complexity in the model only by using the appropriate generalization technique. Small reservoirs in the cascades are modeled by water balance equations of type (6) and (7), and the larger ones - with inequalities of the type (8).

4. Conclusion

The general idea of linear programming modelling of hydro-power plants is given including both turbines and pumps (if available), the importance of pump- hydro stations in power balance as well as the specific constraints deriving from the nature of hydro-power generation. Interconnected water reservoirs bring additional complexity and specifics in the optimization modelling. This modelling framework may be used for HPS forecasting, ED or in combination with the thermal framework for optimal coordinated work in an EPS with HPS integration. When the hydro-thermal coordination involves the optimization of PHPS a mixed-integer programming model is required. Some generalisation techniques are proposed for model size reduction that may be handy in many cases of optimization modelling.

5. Bibliography

- [1]. Wood A., Wollenberg B., Power Generation Operation and Control, 3rd edition, John Wiley & Sons, New York, 2013, ISBN: 978-0-471-79055-6, 656 pages
- [2]. International Hydropower Association, <https://www.hydropower.org/>
- [3]. World Energy Council, www.worldenergy.org/data
- [4]. Stoilov D., Yanev K., Electric Power Systems Regimes, Technical University of Sofia, 2011, 315 pages, ISBN 978-954-438-941-3 (in Bulgarian)
- [5]. Kaneva M., Popov Z., Stoilov D., Active Power Balance in Electric Power Systems with Significant Use of Wind Power Plants, Ecological Engineering and Environment Protection, 2013r., Issue 1, pp. 60-66, ISSN 1311-8668 (in Bulgarian)
- [6]. Stoilov D., Electric Power Systems Balancing and Reservation, Technical University of Sofia, 2013, pages 115, ISBN 978-619-167-084-0 (in Bulgarian)
- [7]. Delarue E., Bekaert D., Belmans R., D'haeseleer W., Development of a Comprehensive Electricity Generation Simulation Model Using a Mixed Integer Programming Approach, World Academy of Science, Engineering and Technology, International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering Vol:1, No:4, 2007, pp. 615 - 620
- [8]. Gjelsvik A., Mo B., Haugstad A., Long- and Medium-term Operations Planning and Stochastic Modelling in Hydro-dominated Power Systems Based on Stochastic Dual Dynamic Programming, Handbook of Power Systems I, p. 33-55
- [9]. Eid C., Koliou E., Valles M., Reneses J., Hakvoort R., Time-based pricing and electricity demand response: Existing barriers and next steps, Elsevier, Utilities Policy 40 (2016) pp. 15 - 25
- [10]. Anagnostopoulos J.S., Papantonis D.E., Pumping station design for a pumped-storage wind-hydro power plant, Energy Conversion and Management, Vol. 48, Issue 11, November 2007, p. 3009-3017
- [11]. Chen H., Cong T.N., Yang W., Tan C., Li Y., Ding Y., Progress in electrical energy storage system: A critical review, Progress in Natural Science, Vol. 19, Issue 3, 10 March 2009, p. 291-312
- [12]. Dinglin L., Yingjie C., Kun Z., Ming Z., Economic evaluation of wind-powered pumped storage system, Systems Engineering Procedia, Vol. 4, 2012, Pages 107-115
- [13]. Kapsali M., K. Kaldellis J. K., Combining hydro and variable wind power generation by means of pumped-storage under economically viable terms, Applied Energy, Vol. 87, Issue 11, ISSN: 0306-2619, November 2010, Pages 3475-3485
- [14]. Siegel N.P., Thermal energy storage for solar power production, Wiley Interdisciplinary Reviews: Energy and Environment, Vol. 1, Issue 2, September/October 2012, p. 119-131

ELECTRIC POWER SYSTEMS MODELING AND EDUCATION: THE CONTROLLABLE LOADS IN A SHORT TERM SYSTEM BALANCE

M.Sc. Trashlieva V. ¹, M.Sc. Radeva T. PhD.¹

Department of Electrical Power Engineering – Technical University of Sofia, Bulgaria
vesselina.trashlieva@gmail.com

Abstract: In this paper we try to build a framework in active power optimization model building when controllable loads are available for balancing purposes. A certain classification of non-fixed loads is given as well as the respective variables and constraints that have to be introduced in the mixed-integer linear programming model. A numerical example is also given illustrating the modeling approach. Some analysis on the presented numerical data is done in order to show sensitivity to certain environmental issues.

Keywords: OPTIMIZATION MODELING, ELECTRIC POWER SYSTEMS OPTIMIZATION, CONTROLLABLE LOADS

1. Introduction

When a load can be switched on or off when needed for system balance purposes, it is considered as a controllable load. An example of a controllable load in an EPS from the System's Operator point of view is the pumping capacities of an PHPS or a group of consumers whose power supply can be interrupted (and / or restricted) during peak load periods. For a micro-grid, a controllable load can be any consumer whose power can be managed and / or limited or its performance may be postponed (dispatched) over time. There are many loads whose operation extends over several time intervals. In this case for those loads is said that they have a working cycle. Since there are many possibilities in the handling of the controllable loads with a work cycle, it is necessary to introduce a certain classification for modeling purposes of such loads.

2. Loads classification and modeling framework

Type I. Controllable loads over time with an interruptible work cycle: These loads are assumed to have a fixed work cycle duration d_{l1*} and a fixed power level P_{l1*} in a single interval. The operation modeling of these loads requires the introduction of binary variables $v_{l1,j}$ having *true* value ($v_{l1,j} = 1$) when the load is switched to power level P_{l1*} in the interval j . Additions of the form $v_{l1,j}P_{l1*}$ must be introduced in the balance constraints. A constraint (1) handles the duration of the work cycle:

$$\sum_j v_{l1,j} = d_{l1*} \quad (1)$$

Type II. Controllable loads over time and consumption with an interruptible work cycle: The power consumed by this load $P_{l2,j}$ in the single interval j depends on the duration of the operating cycle $d_{l2}^{\min} \leq d_{l2} \leq d_{l2}^{\max}$. The total load consumption P_{l2}^{Σ} is known for a full work cycle. So the hour consumption $P_{l2,j}$ and the duration of the duty cycle d_{l2} are optimized introducing binary variables $v_{l2,j} = 1$ when the load is switched to power $P_{l2}^{\min} \leq P_{l2,j} \leq P_{l2}^{\max}$. In the latter the interval's limits are derived via: $P_{l2}^{\min} = P_{l2}^{\Sigma} : d_{l2}^{\max}$ and $P_{l2}^{\max} = P_{l2}^{\Sigma} : d_{l2}^{\min}$. Only the load consumption $P_{l2,j}$ is included in the balance constraint. For modeling the rest of the requirements, including the dependence between the cycle duration and the power level in the unit interval the following constraints are added:

$$P_{l2,j} - v_{l2,j}P_{l2}^{\max} \leq 0 \text{ and } P_{l2,j} - v_{l2,j}P_{l2}^{\min} \geq 0 \quad (2)$$

$$\sum_j v_{l2,j} = d_{l2} \quad (3)$$

$$\sum_j P_{l2,j} = P_{l2}^{\Sigma} \quad (4)$$

$$d_{l2}^{\min} \leq d_{l2} \leq d_{l2}^{\max} \text{ are integer and } v_{l2,j} \text{ are binary} \quad (5)$$

Type III. Controllable loads over time with a non-interruptible work cycle: These are loads that that once started can not be switched off until the whole fixed-duration work cycle is completed. It is assumed that the load consumption in the unit interval of the work cycle with a fixed duration d_{l3*} is P_{l3*} . Once the work cycle has begun, it can not be interrupted. To ensure this requirement, 3 sets of binary variables are introduced:

$s_{l3,j} = 1$ if the load starts a cycle at the beginning of j

$v_{l3,j} = 1$ if the load is running at power P_{l3*} in the interval j

$f_{l3,j} = 1$ if the load finishes its work cycle at the beginning of the interval j

Because the hour consumption of the load P_{l3*} is a constant additions of the form $v_{l3,j}P_{l3*}$ must be introduced in the balance constraints.

The constraint handling the duration of the work cycle remains unchanged:

$$\sum_j v_{l3,j} = d_{l3*} \quad (6)$$

Constraints ensuring the un-interruptible requirement for the working cycle are introduced:

$$s_{l3,j} - v_{l3,j} \leq 0 \quad (7)$$

$$s_{l3,j} - f_{l3,j} = v_{l3,j} - v_{l3,j-1} \quad (8)$$

$$f_{l3,j} + v_{l3,j} \leq 1 \quad (9)$$

$$s_{l3,j} + \sum_{k=j+1}^{j+d_{l3*}-1} f_{l3,k} \leq 1 \quad (10)$$

Type IV. Controllable loads over time and power level with a non-interruptible work cycle: In this case both the duration $d_{l4}^{\min} \leq d_{l4} \leq d_{l4}^{\max}$ of the work cycle is optimized and the power level at a unit interval $P_{l4}^{\Sigma} / d_{l4}^{\max} \leq P_{l4,j} \leq P_{l4}^{\Sigma} / d_{l4}^{\min}$. Constraints ensuring the un-interruptible requirement for the working cycle must be added using the three sets of binary variables:

$s_{l4,j} = 1$ if the load starts a cycle at the beginning of j

$v_{l4,j} = 1$ if the load is running at power $P_{l4,j}$ in the interval j

$f_{l4,j} = 1$ if the load finishes its work cycle at the beginning of the interval j

In the balance constraints only the power of the load $P_{l4,j}$ is added being and optimization variable. A constraint for the work

cycle duration is present with respect to the simple bounds $d_{l4}^{\min} \leq d_{l4} \leq d_{l4}^{\max}$:

$$\sum_j v_{l4,j} = d_{l4} \quad (11)$$

The following constraints ensure the un-interruptible work cycle and the power levels:

$$s_{l4,j} - v_{l4,j} \leq 0 \quad (12)$$

$$s_{l4,j} - f_{l4,j} = v_{l4,j} - v_{l4,j-1} \quad (13)$$

$$f_{l4,j} + v_{l4,j} \leq 1 \quad (14)$$

$$\sum_j P_{l4,j} = P_{l4}^{\Sigma}, P_{l4,j} - v_{l4,j} P_{l4}^{\max} \leq 0 \text{ and}$$

$$P_{l4,j} - v_{l4,j} P_{l4}^{\min} \geq 0 \quad (15)$$

3. A numerical example

To illustrate the controllable loads with a work cycle modeling techniques the following nomenclature and presumptions might be used. A semi-autonomous building has available controllable and uncontrollable loads and generators and can also purchase or sell power to the external grid. Forecasts for its own uncontrollable load and generation are available as well as prices for buying and selling power.

In the model formulation 'P' stands for 'Power', indices 'n' represent 'non', 'f' stands for 'fixed', 'l' stands for 'load', 'G' stands for 'Generation' and 'a' stands for 'accumulation'. The scenario under consideration includes equal hourly pricing for selling and purchasing power from the network ($c_{buy,t} = c_{sell,t}$) and relatively high price for the own generation units ($c_{nfG} > \max\{c_{buy,t}\}$). The numerical example uses a dataset (given $P_{fG,j}$ and $P_{fl,j}$) aiming at maximization of total profit:

$$\max J = \sum c_{sell,j} P_{sell,j} - \sum c_{buy,j} P_{buy,j} - \sum c_{nfG,j} P_{nfG,j} \quad (16)$$

The power of the controllable loads consists of total of seven members divided in four groups of loads according to the classification given in the beginning.

Two loads are from Type I with fixed work cycle duration that can be interrupted and power:

$$d_{l1,1*} = 3 \text{ hours}, P_{l1,1*} = 2,2 \text{ kWh}$$

$$d_{l1,2*} = 4 \text{ hours}, P_{l1,2*} = 2,5 \text{ kWh}$$

Two loads are from Type II with non-fixed power and duration of the work cycle and it can also be interrupted::

$$d_{l2,1} \text{ might be between 2 up to 4 hours, } P_{l2,1}^{\Sigma} = 8.2 \text{ kWh}$$

$$d_{l2,2} \text{ might be between 3 up to 7 hours, } P_{l2,2}^{\Sigma} = 14 \text{ kWh}$$

Two loads from Type III with fixed work cycle duration and power in a unit interval but the continuity of the cycle can not be interrupted:

$$d_{l3,1*} = 2 \text{ hours and } P_{l3,1*} = 1,8 \text{ kWh}$$

$$d_{l3,2*} = 4 \text{ hours and } P_{l3,2*} = 1,4 \text{ kWh}$$

A single load from the last Type IV is considered and its work cycle is un-interruptible. Its work cycle duration is optimized as well as its power in the unit time interval:

$$d_{l4} \text{ may be 3 up to 6 hours with } P_{l4}^{\Sigma} = 12 \text{ kWh}$$

The balance constraint includes all controllable loads additional variables:

$$P_{sell,j} + P_{a,j} + \sum_{l_1=1}^{L_1} v_{l1,j} P_{l1*} + \sum_{l_2=1}^{L_2} P_{l2,j} + \sum_{l_3=1}^{L_3} v_{l3,j} P_{l3*} + \sum_{l_4=1}^{L_4} P_{l4,j} + L_j = P_{buy,j} + P_{R,j} + P_{Ga,j} + P_{G,j} \quad (17)$$

The total power consumed by the controllable loads for the whole optimization period must be not less than the technological minima:

$$\sum_{l_1=1}^{L_1} v_{l1,j} P_{l1*} + \sum_{l_2=1}^{L_2} P_{l2,j} + \sum_{l_3=1}^{L_3} v_{l3,j} P_{l3*} + \sum_{l_4=1}^{L_4} P_{l4,j} \geq E_T \quad (17)$$

For each load the respective additional constraints (1-15) must be added: two constraints of type (2) for each controllable load $l_1 \in L_1$ (Type I), for each controllable load $l_2 \in L_2$ (Type II) a set of constraints (2) to (5), for each controllable load $l_3 \in L_3$ (Type III) a set of constraints (6) to (10) and for the single controllable load $l_4 \in L_4$ (Type IV) the constraints (11) to (15).

Table 1: Optimal values of the variables for accumulation, generation, buying and selling and own controllable generator

Hour j	$P_{a,j}$	$P_{Ga,j}$	$P_{buy,j}$	$P_{sell,j}$	$P_{G,j}$
Col. №	1	2	3	4	5
1	0	0	0	3,3	50
2	0	33,2	33,2	0	0
3	0	31,4	31,4	0	0
4	0	33,9	33,9	0	0
5	0	43,5	43,5	0	0
6	0	39,4	39,4	0	0
7	0	0	0	33,8	0
8	0	0	0	41,2	0
9	10	0	0	51,1	0
10	16	0	0	71	0
11	16	0	0	60	0
12	16	0	0	54,7	0
13	16	0	0	66	0
14	0	0	0	58	0
15	0	0	0	39,1	0
16	0	0	0	76,6	0
17	0	0	0	59,2	0
18	16	0	0	53,5	0
19	16	0	0	37,2	0
20	16	0	0	6,3	0
21	16	0	0	8	0
22	16	0	0	8	0
23	0	28	28	0	0
24	0	28	28	0	0

The constraint (19) stands for the accumulation-generation efficiency coefficient η . The fact that no purchasing and selling in a time interval is allowed, as well as accumulation and generation from accumulating units is not possible two sets of binary variables are introduced in the model ($v_j = 1$ when buying power from the external network and $u_j = 1$ when generating power from the accumulation units) to construct the mutually exclusive alternatives. Constraints (20) assure that no purchase and selling power onto the external grid will occur in a same time interval. Constraints (21) assure that no accumulation and generation will occur in a same time interval. The power of the controllable generator is fixed above (22).

$$\eta \sum P_{a,j} = \sum P_{Ga,j} \quad (19)$$

$$P_{sell,j} \leq (1 - v_j) P_{sell}^{\max} \quad \text{and} \quad P_{buy,j} \leq v_j P_{buy}^{\max} \quad (20)$$

$$P_{a,j} \leq (1 - u_j) P_a^{\max} \quad \text{and} \quad P_{Ga,j} \leq u_j P_{Ga}^{\max} \quad (21)$$

$$P_{nfG,j} \leq P_{nfG}^{\max} \quad (22)$$

The optimal values of the variables are given in Tables 1 and 2. The optimal working cycle durations for the loads of Groups II and

IV loads are 2 hours, 3 hours and 6 hours respectively. Since best prices are night, all controllable loads under consideration work at night (Table 2). With such levels of the fixed generation the building can not function autonomously, i.e. without buying and selling from the external grid, because the sum load from the controllables (18.7 kWh) and the maximum possible P_a^{Max} value can not equal the estimated difference between fixed generation and load (76.6 kWh at 16h).

Table 2: Optimal values for the controllable loads consumption over the optimization horizon

j №	$P_{1,1,j}$ 1	$P_{1,2,j}$ 2	$P_{2,1,j}$ 3	$P_{2,2,j}$ 4	$P_{3,1,j}$ 5	$P_{3,2,j}$ 6	$P_{4,1,j}$ 7
1	2,2	2,5	4,1	4,7	1,8	1,4	2
2	0	0	0	0	1,8	1,4	2
3	0	0	0	0	0	1,4	2
4	0	2,5	0	0	0	1,4	2
5	2,2	2,5	4,1	4,7	0	0	2
6	2,2	2,5	0	4,7	0	0	2
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0

This building can operate in a mode without buying from the external network ($P_{\text{buy},j} = 0$) but just selling ($0 \leq P_{\text{sell},j}$). With such a realization of the building's operation, the optimal timetable for the operation of the controllable loads shifts to the afternoon hours (Tables 3 and 4).

Table 3: Optimal values of the variables for accumulation, generation, buying and selling and own controllable generator when only selling is allowed

Hour j Col. №	$P_{a,j}$ 1	$P_{\text{Ga},j}$ 2	$P_{\text{buy},j}$ 3	$P_{\text{sell},j}$ 4	$P_{\text{G},j}$ 5
1	20	0	0	8,5	50
2	0	8	0	0	0
3	0	8	0	0	0
4	0	8	0	0	0
5	0	8	0	0	0
6	0	8	0	0	0
7	20	0	0	14,5	0
8	20	0	0	26,0	0
9	20	0	0	17,1	0
10	0	0	0	55	0
11	0	0	0	44	0
12	0	0	0	38,7	0
13	0	0	0	50	0
14	20	0	0	36,6	0
15	20	0	0	35,2	0
16	20	0	0	55,2	0
17	20	0	0	37,8	0
18	0	0	0	37,5	0
19	0	16	0	37,2	0
20	0	16	0	6,3	0
21	0	16	0	8	0
22	0	8	0	0	0
23	0	8	0	0	0
24	0	8	0	0	0

The accumulation and generations schedule changes as well as the optimal duration of the single load from the fourth group (L4) as it becomes 3 hours.

Table 4: Optimal values for the controllable loads consumption over the optimization horizon when only selling is allowed

j №	$P_{1,1,j}$ 1	$P_{1,2,j}$ 2	$P_{2,1,j}$ 3	$P_{2,2,j}$ 4	$P_{3,1,j}$ 5	$P_{3,2,j}$ 6	$P_{4,1,j}$ 7
1	2,2	2,5	4,1	4,7	0	0	0
2	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	2,2	2,5	4,1	4,7	1,8	0	4
8	2,2	2,5	0	4,7	1,8	0	4
9	0	0	0	0	0	0	4
10	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0
14	0	0	0	0	0	1,4	0
15	0	2,5	0	0	0	1,4	0
16	0	0	0	0	0	1,4	0
17	0	0	0	0	0	1,4	0
18	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0

4. Conclusion

An approach to controllable loads with a working cycle consisting of more than one unit interval is proposed. The same modeling technique is used for power generating units start-up and stopping cycles in problems for optimal thermal units maintenance medium-term planning. The latter is a modification of the unit commitment problem but it's result influence the UC solution as availability. The numerical example shows how binary and integer values handle the physical requirements of the controllable loads. It also shows the sensitivity towards simple bounds over one set of optimization variables and its interpretation.

5. Bibliography

- [1] Saravanan, B., Das, S., Sikri, S. et al., A solution to the unit commitment problem - a review, *Frontiers in Energy* (2013) Vol 7: pp. 223-226, <https://doi.org/10.1007/s11708-013-0240-3>
- [2] Стоилов Д. Г., Янев К. И., Избор на състава на работещите агрегати в електроенергийните системи чрез смесено-целочислено линейно програмиране, списание *Енергетика*, брой 5, 2000 г., стр 21-23
- [3] Стоилов Д., Ваковски Д., Алгоритми за оптимално разместване на планираните престои на агрегатите от КЕЦ в условия на дефицит, *Енергиен форум'2007*, Варна, юни 2007г., Сборник доклади том II, стр. 221-226.
- [4] Стоилов Д., Кънева М., Ваковски Д., Математически модел за оптимално разпределяне на планираните престои на кондензационните агрегати в условия на дефицит, *Списание Енергетика*, брой 6, 2007г.
- [5] 122 Viana A., Pedroso J. P., A new MILP-based approach for Unit Commitment in power production planning, *International Journal of Electrical Power & Energy Systems*, Volume 44, Issue 1, January 2013, p. 997-1005

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ОПТИМАЛЬНОГО ИСПОЛЬЗОВАНИЯ РИСОУБОРОЧНЫХ КОМБАЙНОВ В КЫЗЫЛОРДИНСКОЙ ОБЛАСТИ

MATHEMATICAL MODELING OF OPTIMAL USE OF RICE-CONTAINING COMBINES IN THE KYZYLORDA REGION

профессор Ж.Садыков, e-mail: sapa_kaz@mail.ru,
докторант Г.Д.Турымбетова
Казахский национальный аграрный университет, Алматы, Казахстан,
e-mail: gulzuhra62@mail.ru

доктор PhD, доцент Караиванов Димитър
Химико-технологического и металлургического университета г. София Болгария

Аннотация. В статье предлагается математический метод оптимизации распределения рисоуборочных комбайнов при выполнении уборочных работ по эксплуатационным показателям: расходу топлива, объему выполненной работы (размеру убранной площади), производительности комбайнов за 1 час сменного времени.

Ключевые слова: оптимизация, рисоуборочный комбайн, уборка, энергозатраты, производительность.

Abstract. The article proposes a mathematical method for optimizing the distribution of rice harvesters in the performance of harvesting operations: the fuel consumption, the amount of work done (the size of the harvested area), the productivity of combines in 1 hour of shifting time.

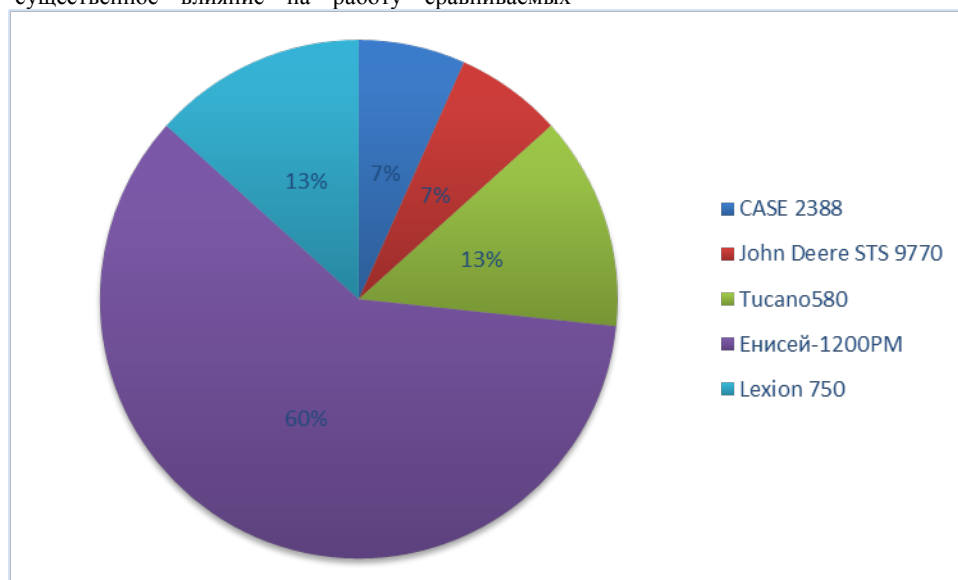
Keywords: optimization, rice harvesting combine, cleaning, energy consumption, productivity.

Введение. Основным способом уборки риса в Кызылординской области является раздельный. Этим способом убирается практически вся посевная площадь. Проведенные испытания показали, что до 90% и более посевов риса к уборке полегают, степень полеглости очень высокая. Жатки комбайнов не обеспечивают скашивания полеглого риса, поэтому полевые испытания на прямом комбайнировании не проводились. Условия испытаний, сложившиеся в этом регионе оказали существенное влияние на работу сравниваемых

комбайнов. На качество зерна также оказывают влияние резкие отличия между отдельными участками и чеками по урожайности, влажности зерна и соломы, что затрудняет регулировки комбайнов.

В настоящее время каждое из рисоводческих хозяйств Кызылординской области стремится функционировать в режиме минимализации затрат с целью получения наиболее высоких доходов. Для достижения этого необходимо учесть такое важное обстоятельство, как наличие той или иной уборочной техники.

В рисоводческом хозяйстве «Магжан и К» общая численность рисоуборочных комбайнов на 1.01.2017 года составляла 30 единиц различных марок и моделей. Парк рисоуборочных комбайнов в хозяйстве представлен 5 марками комбайнов как **российского, так и дальнего** зарубежного производства с различными уборочными параметрами (рис. 1).



Фиг. 1. Структурный состав парка рисоуборочных комбайнов в рисоводческом хозяйстве

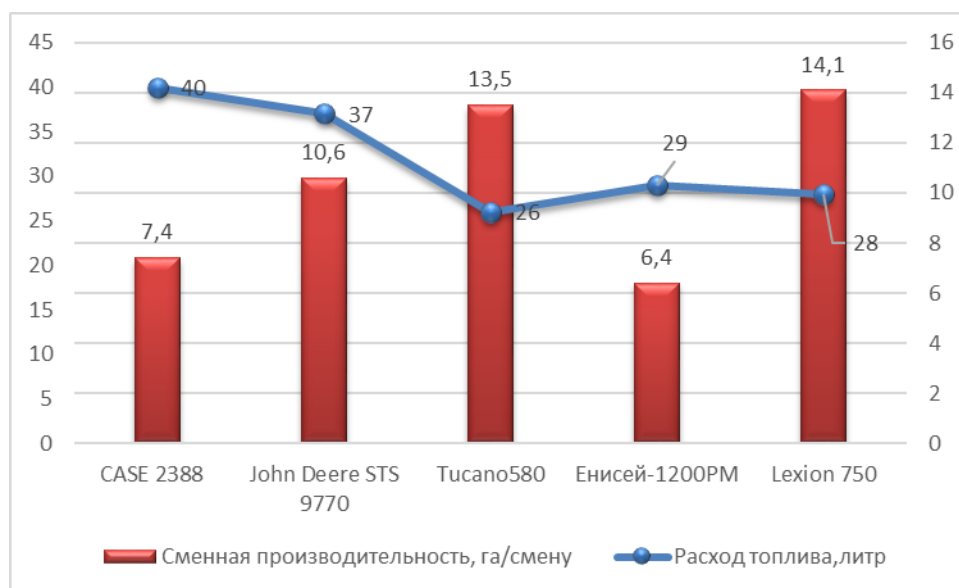
Результаты и дискуссия. Эффективность использования парка рисоуборочных комбайнов определяется комплексом показателей, полученных в результате проведения многократных сравнительных хозяйственных испытаний в реальных условиях эксплуатации. Согласно методике

исследований определяли: убранную каждым комбайном площадь за сезон, количество намолоченного зерна, количество отработанных дней, расход топлива на центнер убранного зерна и т.д. (табл.1).

Таблица 1. Результаты испытаний рисоуборочных комбайнов на уборке риса отдельным комбайнированием

Тип комбайна	Годы эксплуатации в уборке риса, лет	Намолот за время испытания, цн	Наработка за время испытания, га	Расходы топлива, л	
				за время испытания	на 1 ц, всего
CASE 2388	7	7326,7	126	5070	0,7
John Deere STS 9770	4	15399	265	9840	0,6
Tukano 580	2	19712,5	338	8780	0,4
Енисей-1200PM	6	9297,4	160	4650	0,5
Lexion 750	1	21394,7	367	10250	0,5

Одним из показателей уборочного процесса, влияющего на получение прибыли, является расход топлива.



Фиг. 2. Распределение рисоуборочных комбайнов на уборке риса по производительности и удельному расходу топлива

Анализируя данные (рис.2) необходимо отметить, что, на уборке риса рисоуборочный комбайн «Tucano 580» с производительностью по сменному времени – 13,5 га/ч имеет наименьший удельный расход топлива 26 л/га, тогда как удельный расход топлива «Lexion 750», с максимальной высокой производительностью по сменному времени – 14,1 га/ч, составляет 28 л/га. Рисоуборочный комбайн «Енисей-1200PM» с наименьшей производительностью по сменному времени – 6,4 га/ч имеет затраты топлива – 29 л/га, а комбайн

«CASE 2388» с производительностью по сменному времени – 7,4 га/ч имеет наиболее высокие затраты топлива – 40 л/га.

Снижения затрат топлива можно добиться путем оптимального распределения рисоуборочных комбайнов[1]. Для этого используем транспортную задачу по критерию минимального расхода топлива, при выполнении всего объема работ. Результаты хронометражных наблюдений за работой рисоуборочных комбайнов показаны в таблицах 2 и 3.

Таблица 2. Производительность рисоуборочных комбайнов

Тип комбайна	Площади полей,га					Имеется наличии комбайнов
	22	47	77	100	124	
	Производительность комбайнов за 1 час сменного времени, га					
Case 2388	1,137	0,783	0,77	0,592	0,775	2
John Deere STS 9770	0,597	0,752	0,858	0,919	0,933	2
Tucano580	1,11	1,175	1,283	1,1	1,24	4
Енисей-1200PM	0,504	0,36	0,393	0,424	0,526	18
Lexion 750	1,017	1,167	1,22	1,283	1,21	4
Требуемое количество гомбайнов для уборки Заданного объёма	2	5	6	9	8	30

Таблица 3. Удельный расход топлива рисоуборочными комбайнами

Тип комбайна	Площади полей, га				
	22	47	77	100	124
	Удельный расход топлива, л/га				
Case 2388	32,15	38,6	35,2	36,1	35,5
John Deere STS 9770	34,75	36	33,6	33,7	33,5
Tucano580	23,6	26,4	25,5	25,6	26,41
Енисей-1200PM	31,3	31,6	32,9	32,5	28,95
Lexion 750	20,2	29,5	33	32,16	31,7

В математической форме задача записывается следующим образом:

$$\text{Целевая функция: } Z = \sum_{j=1}^n g_j \cdot x_j \rightarrow \min$$

где: j - номер марки рисоуборочного комбайна

g_j - расходы топлива j -го комбайна ($j = 1; n$)

$$32,15 \cdot x_{11} + 38,6 \cdot x_{21} + 35,2 \cdot x_{31} + 36,1 \cdot x_{41} + 35,5 \cdot x_{51} + \dots + 31,7 \cdot x_{55} \rightarrow \min$$

Система переменных включает в себя группу переменных (x_{11} - x_{55}) обозначающих искомое число комбайнов соответствующего типа, используемых на уборке риса:

x_1 - Case 2388

x_2 - John Deere STS 9770

x_3 - Tucano580

x_4 - Енисей-1200PM

x_5 - Lexion 750

Сравнивая имеющиеся в наличии комбайнов с числом комбайнов на уборке заданного объема зерна, видим, что эти суммы совпадают. Следовательно, данная транспортная задача обладает закрытой моделью (табл.4).

Таблица 4. Результаты расчетов

Тип комбайна	Площади полей,га					Количество имеющихся в наличии комбайнов
	22	47	77	100	124	
	Удельный расход топлива, л/га					
Case 2388	32,15 X_{11}	38,6 X_{12}	35,2 X_{13}	36,1 X_{14}	35,5 X_{15}	2
John Deere STS 9770	34,75 X_{21}	36 X_{22}	33,6 X_{23}	33,7 X_{24}	33,5 X_{25}	2
Tucano580	23,6 X_{31}	26,4 X_{32}	25,5 X_{33}	25,6 X_{34}	26,41 X_{35}	4
Енисей-1200PM	31,3 X_{41}	31,6 X_{42}	32,9 X_{43}	32,5 X_{44}	28,95 X_{45}	18
Lexion 750	20,2 X_{51}	29,5 X_{52}	33 X_{53}	32,16 X_{54}	31,7 X_{55}	4
Требуемое количество комбайнов для уборки Заданного объёма	2	5	6	9	8	30

Первая группа ограничений обеспечивает выполнение заданных объемов работ имеющимися в наличии комбайнами.

$$\sum_{j=1}^n x_{ij} = b_j$$

$$\begin{aligned}
x_{11} + x_{21} + x_{31} + x_{41} + x_{51} &= 2 \\
x_{12} + x_{22} + x_{32} + x_{42} + x_{52} &= 2 \\
x_{13} + x_{23} + x_{33} + x_{43} + x_{53} &= 4 \\
x_{14} + x_{24} + x_{34} + x_{44} + x_{54} &= 18 \\
x_{15} + x_{25} + x_{35} + x_{45} + x_{55} &= 4
\end{aligned}$$

где b_j - количество имеющихся рисоуборочных комбайнов, шт;

Вторая группа ограничений показывает необходимость обязательного выполнения заданных объемов работ полностью.

$$\sum_{j=1}^n x_{ij} = a_i$$

$$\begin{aligned}
x_{11} + x_{21} + x_{31} + x_{41} + x_{51} &= 2 \\
x_{12} + x_{22} + x_{32} + x_{42} + x_{52} &= 5 \\
x_{13} + x_{23} + x_{33} + x_{43} + x_{53} &= 6 \\
x_{14} + x_{24} + x_{34} + x_{44} + x_{54} &= 9 \\
x_{15} + x_{25} + x_{35} + x_{45} + x_{55} &= 8
\end{aligned}$$

Таблица 5. Результаты расчетов

Операция		Площади полей,га					Количество имеющихся в наличии комбайнов
		22	47	77	100	124	
		Удельный расход топлива, л/га					
Марка комбайна	<div><div>V_j</div><div>U_i</div></div>	10	11,6	12,9	12,5	8,95	
Case 2388	23,6	32,15	38,6	35,2	36,1	35,5	2
John Deere STS 9770	21,2	34,75	36	33,6	33,7	33,5	2
Tucano580	12,6	23,6	26,4	25,5	25,6	26,41	4
Енисей-1200PM	20	31,3	31,6	32,9	32,5	28,95	18
Lexion 750	10,2	20,2	29,5	33	32,16	31,7	4
Требуемое количество гомбаинов для уборки Заданного объёма		2	5	6	9	8	30

Построим исходный опорный план методом минимального элемента[2].

Для исследования плана на оптимальность необходимо найти оценки свободных клеток. Для этого надо знать потенциалы

U_i и V_j , которые определяются в результате решения системы уравнений

$$\left\{ \begin{aligned}
U_1 + V_3 &= 35,2 \\
U_2 + V_3 &= 33,6 \\
U_3 + V_3 &= 25,5 \\
U_3 + V_4 &= 25,6 \\
U_4 + V_2 &= 31,6 \\
U_4 + V_4 &= 32,5 \\
U_4 + V_5 &= 28,95 \\
U_5 + V_1 &= 20,2 \\
U_5 + V_2 &= 29,5
\end{aligned} \right. \quad (1)$$

составленных по заполненным клеткам. Придадим одному из неизвестных определенное числовое значение, $U_1 = 0$. Тогда

остальные неизвестные находятся из системы (1): Получаем:

$$\begin{aligned} U_1 &= 0, & V_1 &= 20,2 + 4,9 = 25,1 \\ U_2 &= 33,6 - 35,2 = -1,6 & V_2 &= 31,6 + 2,8 = 34,4 \\ U_3 &= 25,5 - 35,2 = -9,7 & V_3 &= 35,2 \\ U_4 &= 32,5 - 35,3 = -2,8 & V_4 &= 25,6 + 9,7 = 35,3 \\ U_5 &= 29,5 - 34,4 = -4,9 & V_5 &= 28,95 + 2,8 = 31,75 \end{aligned}$$

Теперь можно найти оценки свободных клеток:

$$\begin{aligned} S_{11} &= C_{11} - (U_1 + V_1) = 32,15 - 25,1 = 7,05 \\ S_{12} &= C_{12} - (U_1 + V_2) = 38,6 - 34,4 = 4,2 \\ S_{14} &= C_{14} - (U_1 + V_4) = 36,1 - 35,3 = 0,8 \\ S_{15} &= C_{15} - (U_1 + V_5) = 35,5 - 31,75 = 3,75 \\ S_{21} &= C_{21} - (U_2 + V_1) = 34,75 - (-1,6 + 25,1) = 11,25 \\ S_{22} &= C_{22} - (U_2 + V_2) = 36 - (-1,6 + 34,4) = 3,2 \\ S_{24} &= C_{24} - (U_2 + V_4) = 33,7 - (-1,6 + 35,3) = 0 \\ S_{25} &= C_{25} - (U_2 + V_5) = 33,5 - (-1,6 + 31,75) = 3,35 \\ S_{31} &= C_{31} - (U_3 + V_1) = 23,6 - (-9,7 + 25,1) = 8,2 \\ S_{32} &= C_{32} - (U_3 + V_2) = 26,4 - (-9,7 + 34,4) = 1,7 \\ S_{35} &= C_{35} - (U_3 + V_5) = 26,41 - (-9,7 + 31,75) = 4,36 \\ S_{41} &= C_{41} - (U_4 + V_1) = 31,3 - (-2,8 + 25,1) = 9 \\ S_{43} &= C_{43} - (U_4 + V_3) = 32,9 - (-2,8 + 35,2) = 0,5 \\ S_{53} &= C_{53} - (U_5 + V_3) = 33 - (-4,9 + 35,2) = 2,7 \\ S_{54} &= C_{54} - (U_5 + V_4) = 32,16 - (-4,9 + 35,3) = 1,76 \\ S_{55} &= C_{55} - (U_5 + V_5) = 31,7 - (-4,9 + 31,75) = 4,85 \end{aligned}$$

Поскольку в табл. 5 свободных клеток с отрицательными оценками нет, то опорный план является оптимальным.

Заключение

Для уборки урожая на полях в 22 га потребуется два зерноуборочных комбайна Lexion 750; 47 га – два зерноуборочных комбайна Lexion 750 и три Енисей-1200РМ; 77 га – два Case 2388, два John Deere STS 9770 и два Tucano580; 100 га – два Tucano 580 и семь Енисей-1200РМ, а для 124 га – восемь рисоуборочных комбайнов Енисей-1200РМ. При таком распределении рисоуборочных комбайнов по работам расход топлива будет минимальным. При другом распределении агрегатов по работам расход топлива будет больше по сравнению с расчетным. Следовательно, оно не оптимально.

Литература

1. Кидяева Н.П., Щитов С.В., Жирнов А.Б. Оптимизация выбора комбайна по расходу топлива при уборке сельскохозяйственных культур // Техника и оборудование для села. – 2013, №1. – С. 18-22.
2. Экономико-математические методы и прикладные модели / В.В. Федосеев [и др.]. М.: ЮНИТИ, 1999. 391 с.

КОЛЕБАНИЯ В САМОУСТАНОВЛИВАЮЩИХСЯ МЕХАНИЗМАХ КОНСТРУКЦИИ МАЯТНИК

OSCILLATIONS IN SELF-ADJUSTING MECHANISMS OF CONSTRUCTION PENDULUM.

Nauryzbaev R., Sadykov Z., Sansyzbayev K., Toilybaev M., Koshanova S.
Kazakh National Agrarian University – Almaty, Kazakhstan

Annotation: An exact analytical solution of the nonlinear differential equation is found. On the basis of the formula of the modern theory of mechanisms defined mathematical model constructs a self-aligning mechanism of a physical pendulum. The established parameters, basic modes of operation rational design and regular construction of mechanisms such as a pendulum.

Keywords: Self-stabilizing mechanisms of pendulum constructions, physical pendulum mechanisms, full elliptic integral of the first kind, swing, uniform rotation, non-uniform rotation, inertial perturbation.

Самоустанавливающиеся механизмы конструкции маятник – это статически определимые механизмы, т.е. механизмы без избыточных связей. В общем случае число

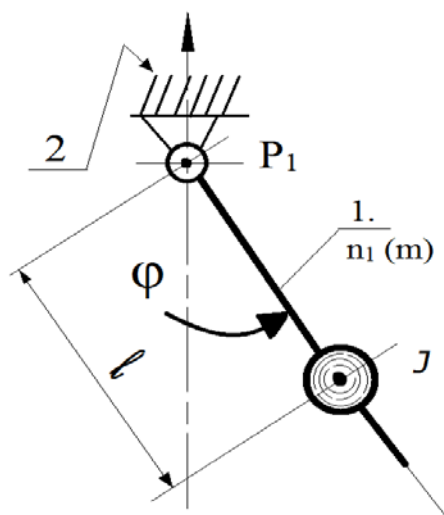
степеней свободы самоустанавливающейся кинематической цепи механизма определяется по формуле [1,2,3]:

$$W = m \cdot (n + n_1 + n_2 - 1) - \sum_{k=1}^{m-1} (m - k) \cdot P_k \quad (1)$$

Для конструкций цепей третьего семейства по систематизации академика И.И. Артоболевского в формуле (1) необходимо принять $m=3$.

Исследуемый механизм простейшей конструкции (рис.1), двухзвенный, самоустанавливающийся, лишен избыточных связей, не имеет лишних степеней свободы:

$$W = 3 \cdot n_1 - 2 \cdot P_1 = 3 \cdot 1 - 2 \cdot 1 = 1 \quad (2)$$



Основа структуры механизма – однозвенная незамкнутая, простая кинематическая цепь с $W=1$, она (n_1) присоединяется к стойке 2 не изменяя числа степеней свободы.

В научно – теоретических курсах теории механизмов и механики машин известен как механизм физического маятника, маятник Фроуда, механизм качели, механизм I^{го} класса и др.

Уравнение движения механизма физического маятника имеет вид записи:

$$J \cdot \ddot{\varphi} + m \cdot g \cdot l \cdot \sin \varphi = 0, \quad (3)$$

где I – момент инерции механизма физического маятника относительно оси вращения (шарнира - P_1), [1/сек].

Тогда момент, создаваемый силами инерции, равен:

$$M_H = J \cdot \ddot{\varphi}, \quad [H \cdot M] \quad (4)$$

m – масса вращающегося звена (n_1) механизма физического маятника, [кг];

g – ускорение свободного падения ($g=9,81 \text{ м/сек}^2$).

Тогда сила тяжести звена (n_1) механизма физического маятника равна формуле следующего вида:

$$P = m \cdot g, \quad [H] \quad (5)$$

l – расстояние от центра тяжести до оси вращения звена (n_1), [м];

φ – обобщенная координата ведущего звена механизма физического маятника. [рад.]

$$\text{Пологая } \omega^2 = \frac{m \cdot g \cdot l}{2a}, \quad (6)$$

Получим уравнение вида:

$$\ddot{\varphi} + \omega^2 \cdot \sin \varphi = 0 \quad (7)$$

Дифференциальное уравнение (7) имеет точное аналитическое решение [1,2,3]. Период колебания звена (n_1), соответствующий изменению угла φ на 2π , равен записи вида:

$$\dot{\varphi}_0 \cdot t = 4 \cdot \omega \cdot K(q) \quad (8)$$

где $K(q) = F[\frac{\pi}{2}, q]$ - называется полным эллиптическим интегралом первого рода, а величина q - его модулем ($0 < q < 1$). Численное значение полного эллиптического интеграла первого рода по заданному модулю определяется по таблицам эллиптических интегралов.

При значениях модуля полного эллиптического интеграла первого рода $q^2 = 1$ механизм физического маятника еще не переходит во вращательное движение.

При значениях модуля полного эллиптического интеграла первого рода $q^2 = 0$ звено (n_1) механизма физического маятника (рис.1) вращается равномерно.

Запишем (8) с учётом формулы (6) и тогда получим выражение следующего вида:

$$\dot{\varphi}_0 \cdot t = 4 \cdot \frac{\sqrt{m \cdot g \cdot l}}{J} \cdot [F\frac{\pi}{2}, q] \quad (9)$$

Третий режим работы механизма физического маятника показывает неравномерное вращение ведущего звена (n_1). Этот режим работы механизма физического маятника общим. В данной постановке задачи возмущение в системе чисто инерционное.

Значение критической угловой скорости механизма физического маятника определяется формулой вида

$$\omega_k = \frac{\sqrt{m \cdot g \cdot l}}{J}, [1/сек] \quad (10)$$

Если начальное значение угловой скорости (Y_0) звена (n_1) механизма физического маятника (Рис.1) меньше ω_k ,

$$\dot{\varphi}_0 < \omega_k \quad (11)$$

тогда звено (n_1) совершает периодические колебания. Механизм физического маятника работает в колебательном режиме - в режиме качания. Если же начальное значение угловой скорости (φ_0) звена (n_1) механизма физического маятника больше чем (ω_k), т.е.

$$\dot{\varphi}_0 > \omega_k \quad (12)$$

тогда свободное движение звена (n_1) механизма физического маятника носит ротационный характер, с периодом изменения угловой скорости по формуле следующего вида:

$$T = \frac{4 \cdot K(q)}{\dot{\varphi}_0}, [сек] \quad (13)$$

Степень равномерности вращения ведущего звена (n_1) механизма физического маятника зависит от модуля (q)

полного эллиптического интеграла первого рода $F[\frac{\pi}{2}, q]$, тем выше, чем больше значение модуля.

Если начальное значение угловой скорости (φ_0) звена (n_1)

механизма вида: $\dot{\varphi}_0 \gg 2\omega_k$,

то при малых значениях модуля (q) полного эллиптического интеграла первого рода $F[\frac{\pi}{2}, q]$ имеем:

$$K = \frac{\pi}{2} (1 + \frac{1}{4} q^2 + \dots) = \frac{\pi}{2} (1 + \frac{\omega^2}{\dot{\varphi}_0^2} + \dots), \quad (14)$$

а формула периода изменения угловой скорости ведущего звена (n_1) механизма физического маятника будет иметь следующий вид:

$$T_1 = \frac{2\pi}{\dot{\varphi}_0} (1 + \frac{\omega^2}{\dot{\varphi}_0^2}), [сек] \quad (15)$$

С учетом зависимости (6) имеем:

$$T_1 = \frac{2\pi}{\dot{\varphi}_0} (1 + \frac{m \cdot g \cdot l}{J \cdot \dot{\varphi}_0^2}), [сек] \quad (16)$$

Таким образом, положение или состояние колеблющегося звена (n_1) механизма физического маятника (Рис.1) определяется обобщенной координатой - (φ).

Величина, обратная периоду колебания (T), называется частотой колебания (f) и равняется числу колебаний в секунду:

$$f = \frac{1}{T}, [1/сек] = 1/\Gamma_{\varphi} \quad (17)$$

Под круговой частотой понимается число колебаний за 2π секунд:

$$\omega_c = 2\pi \cdot f = \frac{2\pi}{T}, [1/сек] = [рад/сек] \quad (18)$$

С учетом (18), формулы (16) запишем круговую частоту колебания ведущего звена (n_1) механизма физического маятника в форме записи:

$$\omega_c = \dot{\varphi}_0 \cdot \left(1 + \frac{\varphi_0^2 \cdot J}{m \cdot g \cdot l}\right), [рад/сек] \quad (19)$$

По формуле (19) рассчитывается круговая частота собственных (свободных) колебаний механизма физического маятника в режиме неравномерного вращения ведущего звена (n_1). Время, в течении которого совершается одно полное, называется периодом (T), параметры: ω_c и T - не зависят от начальных условий и являются неизменными характеристиками колеблющейся системы. Уравнение (7) нелинейное, если ограничится малыми углами отклонения ведущего звена (n_1) механизма физического маятника $\varphi \ll 1$, то можно её упростить. При этом уравнение (7) становится линейным:

$$\ddot{\varphi} + \omega^2 \cdot \varphi = 0 \quad (20)$$

Собственными колебаниями являются движения, совершаемые колебательной системой, которая после

кратковременного внешнего возмущения представлена самой себе. При этом происходят периодические переходы одного вида энергии в другой т.е. потенциальная энергия (определяемая положением системы) и наоборот.

Если сумма этих энергий в процессе колебаний сохраняется, то колебания будут недемпфированными (незатухающими) и система в этом случае называется *консервативной*.

Если энергия системы уменьшается (например, из-за наличия трения), то происходят демпфированные (затухающие) колебания и система называется *неконсервативной*.

Круговая частота собственных колебаний ведущего звена (n_1) механизма физического маятника в режиме качания равно :

$$\omega_c = \omega = \sqrt{\frac{m \cdot g \cdot l}{J}}, \text{ [рад/сек]} \quad (21)$$

Период собственных колебаний ведущего звена (n_1) механизма физического маятника в режиме качания:

$$T = \frac{2\pi}{\omega_c} = 2\pi \cdot \sqrt{\frac{J}{m \cdot g \cdot l}}, \text{ [сек]} \quad (22)$$

При значениях параметров: $\pi=3,14$; $g=9.81 \text{ м/сек}^2$ можно считать

$$J = \frac{m \cdot l^2}{3}, \text{ [кг} \cdot \text{м}^2 \text{]}$$

Умножим нелинейное дифференциальное уравнение (20) на (\dot{x}) и проинтегрируем:

$$\ddot{x} \cdot \dot{x} + \omega^2 x \cdot \dot{x} = \frac{d}{dt} \left(\frac{\dot{x}^2}{2} \right) + \omega^2 \frac{d}{dt} \left(\frac{x^2}{2} \right) = 0, \quad (23)$$

или

$$\dot{x}^2 + \omega^2 \cdot x^2 = \text{const.} \quad (24)$$

Полученное уравнение представляет собой уравнение фазовой траектории, т.е. устанавливает зависимость между x и \dot{x} .

Литература.

1. Наурызбаев Р.К. и др. Теория самоустанавливающихся кинематических цепей пространственных исполнительных механизмов: - Монография, Алматы.: «Тауар», 2000. - 494с.
2. Наурызбаев Р.К. и др. Современная прикладная механика. - Алматы: Серия: «Машиностроение», 2004. - 464с.
3. Наурызбаев Р.К. Анализ, синтез и разработка самоустанавливающихся шарнирно-стержневых механизмов с гибкими связями: Дисс..... докт. техн. наук., Алматы, 1993. - 484с.

NUMERICAL MODEL FOR SIMULATION OF THE VELOCITY FIELDS FOR THE EXPLOSIVELY FORMED PENETRATOR

ЧИСЛЕН МОДЕЛ ЗА СИМУЛАЦИЯ НА СКОРОСТНИТЕ ЗОНИ ПРИ ЗАРЯДИ ОТ УДАРНО ЯДРО

M.Sc. Hutov I. PhD.¹, Prof. M.Sc. Lilov I. PhD.²

Joint Forces Command – Sofia, Bulgaria¹

“Vasil Levski” National Military University, Veliko Turnovo, Bulgaria²

i.hutov@armf.bg

Abstract: The current paper presents numerical approach of velocity performances estimations for the EFP (Explosively Formed Projectiles). The proposed method mathematically develops velocities parameters of a particular segment for EFP liner propelled by explosive process. The numerical method is developed, to provide estimations about behavior of projectile vs. time in the EFP forming process powered by explosion. The model is valid for performances estimations of EFP warheads and design data for optimal EFP configuration. Simulations are supported by the software Autodyn for numerical modeling respectively. The obtained numerical results are compared with the available experimental data.

KEYWORDS: EXPLOSIVE FORMED PENETRATOR, NUMERICAL SIMULATION, VELOCITIES DISTRIBUTION, AUTODYN,

1. Introduction

Nowadays, the EFP warheads are present in many systems that expect appropriate modernization and/or optimization; as artillery sub-munitions, antitank missiles, mines etc. Approaches which define the processes of explosively formed projectiles [1-4] are one of the most sophisticated problems of rigid body mechanics based on the elastic to plastic theory. The distinguishing problem of the EFP projectiles is the velocity of the EFP liner. This velocity is generated in the explosively driven process and the dynamics of their evolution is the main topic of this paper. Recently, most papers are based on numerical methods [5-10] which determine the projectiles velocity performances based on detailed modeling of the loadings and deformation process during explosion. Numerical software, particularly Autodyn, which is often used for detailed analyses in numerical simulations, require comprehensive preparation of the expected initial data but some others methods as it analytical are less precise but enough reliable and provides much faster data obtaining for the applications of warheads performances estimations.

The current paper presents software based on the previously studied analytical method as a solution to provide the ability to preliminarily estimations as well as numerical solution of the same rooted liners velocities. This methodology provides ability to analyze the adopted design of warhead's performances by more precise numerical software like Autodyn.

The research based on the analytical models presented in papers [1-5], provides crucial information about the EFP performances in a short time without required comprehensive initial data preparation. The algorithm presented in further papers provides the possibility to directly export the adopted geometry of EFP liners integrated with warheads into Autodyn numerical software, from the software package Matlab, which considerably decreases preparation time.

The results of numerical method contribute in improving the accuracy of EFP velocity estimations. This is achieved by an appropriate augmentation in the number of the grid elements for the method used.

2. Numerical approach

Numerical approach based on the finite element method is used in this research in order to be compared with experimental data.

The properties of the adopted simulation model mesh [12-19] are given in Table 1. The mesh density is determined taking into account accuracy as well as reasonable simulation run time within available computer facilities.

Figure 1a and 1b shows configuration of EFP warhead as well as appearance of created mesh for each component separately.

The simulation sample volume in numerical approach is observed as the quarter shown on the figure 1a and 1b.

Presented analysis uses fully Lagrangian solver, where after 35 μ s, detonation products are not influenced into forming processes. But that average liners final velocity comparative with analytical modeling corresponds not to the 35 μ s instant of forming time then about 70-150 μ s where dynamical process is fully completed (figures 5 and 6).

A wide range of metal powders (from light alloys through steels to super-alloys and composites) is currently available for DMLS process and other new materials are under development. Table 1 lists mechanical properties of selected powder materials.

Table 1: Grid properties of the numerical approach [3]

Conditions	Type 1		Type 2*	
	1	2	1	
Liner	7776	6125	7776	6125
Explosive	10496	9000	10496	9000
Cover	15006	12000	-	-
Back plate	768	450	768	450
1 – nodes; 2 –elements;				

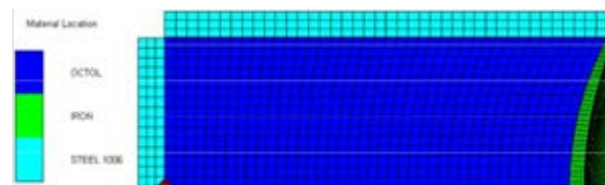


Figure 1a: Geometrical configuration of EFP sample type1 (with cover) and finite elements mesh



Figure 1b: Geometrical configuration of EFP sample type2 (without cover) and finite elements mesh

The loading forces distribution model is expressed by the detonation pressure products and is determined according to Jones-Wilkins-Lee [1] by the equation of state:

$$p = K \left(1 - \frac{\omega}{R_1 V} \right) e^{(-R_1 V)} + K_1 \left(1 - \frac{\omega}{R_2 V} \right) e^{(-R_2 V)} + \frac{\omega E}{V} \quad (1)$$

where V and E are represented as $V = \rho_0 / \rho$, $E = \rho_0 e$, ρ_0 is the current density, ρ is the reference density, e is the specific internal energy and K , K_1 , R_1 , R_2 and ω are constants for the given explosive material [1,2].

3. Simulation model

The comparison of these methods is performed on the sample design fig 2, with accepted, fixed EFP liner form and explosive charge, with and without metal cover. Adopted explosively driven projectile model and its elements of geometry, presented in the paper [2], and design characteristics of testing sample as in the [14] are shown in Fig. 2. The model does not include the fuze and wave shaper integrated in the warhead design and influenced on the real performances modeling.

The properties of explosive and other materials used in simulations are given in Table 1 [14]. In tested examples, the initiation point is located on the warhead bottom and lies on axis of symmetry [14] (Fig. 1).

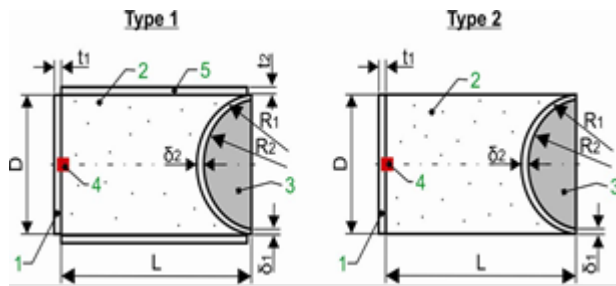


Figure 2: Types of testing sample and their basic dimensions: 1-back plate, 2-explosive charge, 3-liner, 4-initiation point, 5-cover.

Table 2: Geometrical parameters for EFP sample models [3]

Design parameter		Type 1	Type 2*
Length of charge	L [mm]	85	85
Caliber	D [mm]	57.2	57.2
Thickness of back plate	t1 [mm]	5	3
Cover thickness	t2 [mm]	5	-
Inner radius	R1 [mm]	60.4	71.3
Outer radius	R2 [mm]	60.4	71.3
Thickness of liner edge	delta1 [mm]	1.5	1.5
Thickness of liner center	delta2 [mm]	2.7	2.7
Type initiation		p.	p.
* -experiment; p. -point initiation			

Analytical and numerical approach used Octol as explosive material with density of 1.82 g/cm³ and detonation velocity 8480 m/s as well as steel as cover and iron as liner material. The experimental sample was tested on the proving ground as a type 2 [14] in Table 2.

4. Results and discussion

Two types of simulation samples of the liners and explosives

integrated have been considered through represented modeling in numerical approach.

Figures 3 and 4 show energy distribution vs. time during projectile forming. Kinetic energy represents penetration capability of formed projectiles.

The plastic works, is important for liners' design and for selection of appropriate material. Figure 3 and 4 represents nonlinear and uniform distribution of plastic energy. It means that liner during formation had proper deformation also influenced on the velocities distribution. If that curve in initial phase of formation has no permanent increase, this indicates the liner had the fracture.

Table 3 shows differences in the energy distribution obtained by the numerical and analytical approach. In table 3 are presented next values: absolute initial velocity V_0 [m/s], kinetic energy E_k [J], axial deformation energy ADE [J], radial deformation energy RDE [J] and plastic deformation energy/plastic work PW [J].

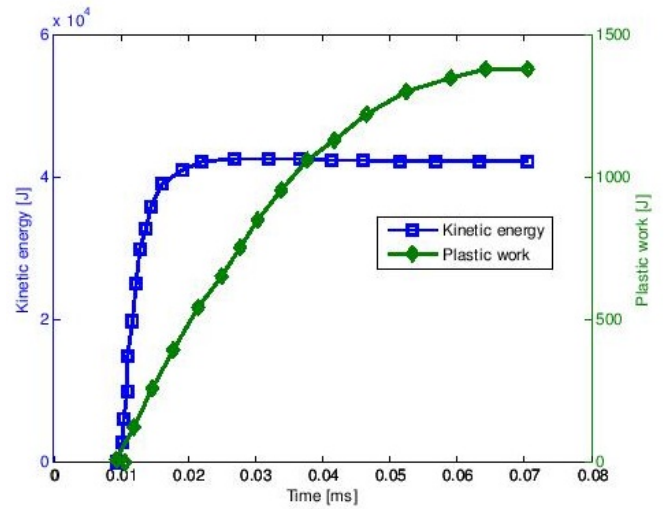


Figure 3: Energy distribution during time of the forming of explosively formed projectile, sample type 1, obtained by numerical method [3]

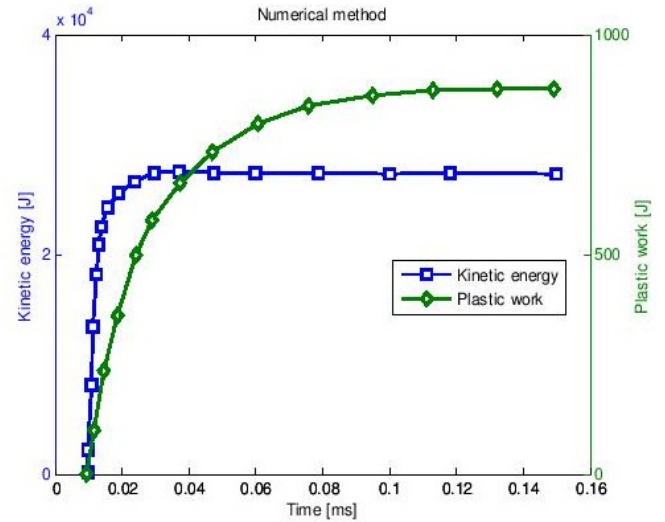


Figure 4: Energy distribution during time of the forming of explosively formed projectile, sample-type 2, obtained by numerical method [3]

These parameters are collected as the consequence of considering problems of deformation energy in the numerical and in the analytical models. Differences between two types of samples show that cover of the explosive sample influences as to increase of kinetic energy of projectile and also the increase of total plastic deformation work [1,2,10,13,20].

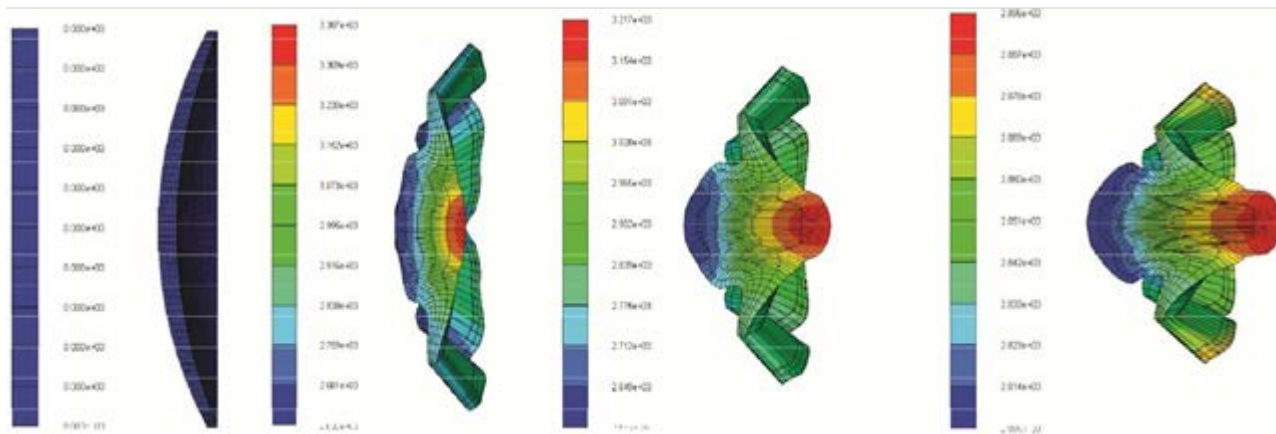


Figure 5: Shape of projectile configuration during forming to the final shape in 70 μ s of sample-type 1

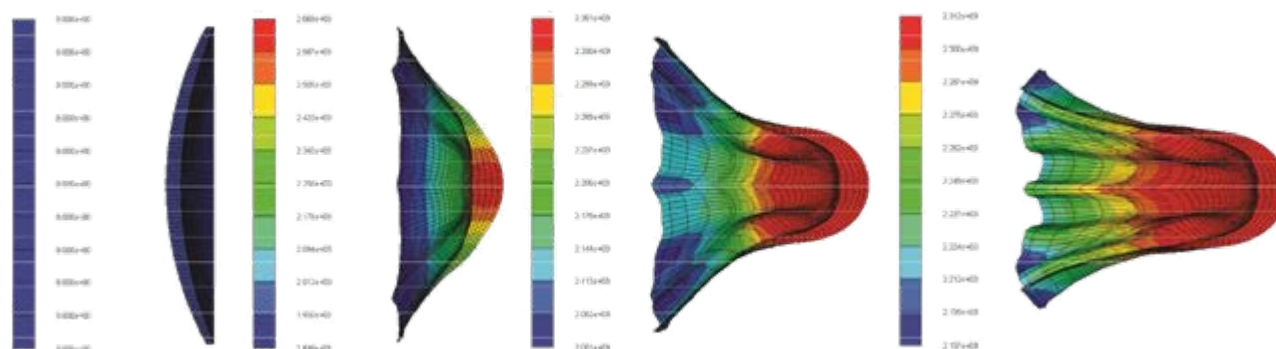


Figure 6: Shape of projectile configuration during forming to the final shape in 150 μ s of sample-type 2

Numerical simulation also reproduces expected shapes of projectiles at the end of forming process shown in (Figs. 5 and 6). For the sample type 1 (Fig. 5) projectile is formed with its final shape after $t=70.5 \mu$ s at the distance 265.31mm, realizing final velocity of about 2860 m/s. For the sample type 2 (Fig. 6) these values are corresponding to the instant $t=150 \mu$ s, at the distance 418.2 mm and velocity 2435 m/s. That means that sample type 2 has much less coefficient of energy efficiency than covered warhead charges [10]. The final projectile shape joint with considered velocity performances influences two basic performances important for EFP warhead design – penetrability and precision.

5. Conclusion

The next conclusions are presented as the result of this study:

The numerical approach is a well-designed tool for the EFP velocity and energy modeling and estimations.

Numerical method gives more accurate results regarding velocity in comparison with analytical methods and these results are very close to experimental data, with error of less than 1.5%. It should be noted that numerical method is useful for the shortening the development time of EFP warheads during design and reduces the cost of their experimental testing.

The same configuration of liners and explosive charges with and without metal covers produced different shapes of explosively formed projectiles. Sample type 1 produced EFP as the plastic solid shape less adoptable for distance flight, and sample type 2 produced EFP with more adoptable shape for distance flight regarding aerodynamical drag.

6. Literature

- [1] Орленко, Л.П. : *Физика Взрыва*, Главная редакция физико-математической литературы, Москва, 2004.
- [2] Sharma, VK, Kishore, P., Bhattacharyya, AR, Raychaudhuri, TK, Singh S.: *An Analytical Approach for Modeling EFP Formation and Estimation of Confinement on Velocity*, International Ballistics Society, pp.565-574.
- [3] Marković, M., Elek, P., Jaramaz, S. Milinović, M., Micković, D.. “Numerical and analytical approach to the modeling of explosively formed projectiles”, 6th International Scientific Conference OTEH 2014, October Belgrade, pp. 9-10, 2014.
- [4] Marković, M.: *Explosively Formed Projectiles*, MSc Thesis, University in Belgrade, Mechanical Engineering, Weapon Systems Department, 2011.
- [5] Markovic, M., Milinovic, M., Elek, P., Jaramaz, S., Mickovic, D. “Comparative approaches to the modeling of explosively formed projectiles”, Proceedings of Tomsk State University, Serie Physics and Mathematics, V.293, 2014.
- [6] Markovic M., Milinovic M., Jeremic O., Jaramaz S.. “Numerical modeling of temperature field on high velocity explosively formed projectile”, 17th Symposium On Thermal Science And Engineering Of Serbia, October 20–23, pp.175-181, 2015.
- [7] Pappu, S., Murr, L.E.: *Hydrocode and microstructural analysis of explosively formed penetrators*, Journal of Materials Science, pp. 233-248, 2002.
- [8] Hussain, G., et al.: *Gradient Valued Profiles and L/D Ratio of Al EFP With Modified Johnson Cook Model*, Journal of Materials Science and Engineering 5, pp. 599-604, 2011.
- [9] Fedorov, S.V., Bayanova Ya. M., Ladov, S.V. “Numerical analysis of the effect of the geometric parameters of a combined shaped-charge liner on the mass and velocity of explosively formed

compact elements”, Combustion, Explosion, and Shock Wave, Vol. 51, No. 1, pp. 130-142, 2015.

[10] Hussain, G., Hameed, A., Hetherington, J.G. “Analytical performance study of explosively formed projectile”, Journal of applied mechanics and technical physics, Vol.54, No.1, pp.10-20, 2013.

[11] Bender, D., Corleone, J. “*Tactical Missile Warheads - Explosively Formed Projectile*”, American Institute of Aeronautic and Astronautic, Washington, 1993.

[12] Teng, T.L., Chu, Y.A., Chang, f.a., Shen, B.C. “*Design and Implementation of a High Velocity Projectile Generator*”, Fizika Goreniya I Vzryva, Vol. 43, No. 2, pp. 233-240, 2007.

[13] Weimann, K. “Research and development in the area of explosively formed projectiles charge technology”, Properllants, Explosives, Pyro-technics, Vol. 18, pp. 294-298, 1993.

[14] Jones, D.A., Kemister, G., Borg, R.A.J.: *Numerical Simulation of Detonation in Condensed Phase Explosives*, Weapons Systems Division Aeronautical and Maritime Research Laboratory, August 1998.

[15] Lam, C., McQueen, D.: *Study of the Penetration of Water by an Explosively Formed Projectile*, Weapons Systems Division Aeronautical and Maritime Research Laboratory, June 1998.

[16] Fong, R., Kraft, J., Thompson, L.M., Ng, W.: *3D Hydrocode Analysis of Novel Asymmetrical Warhead Designs*, Proceedings for the Army Science Conference (24th), 29 November - 2 December, Florida, pp.1-3, 2005.

[17] Weibing, L., Xiaoming, W., Wenbin, L.: *The Effect of Annular multi-point Initiation on the Formation and Penetration of an Explosively Formed Penetrator*, International Journal of Impact Engineering 37, pp. 414-424, 2010.

[18] Church, P.D., Cullis, I.: *Development and Application of High Strain Rate Constitutive Models in Hydrocodes*, Journal de Physique IV, Vol. 1, pp. 917-922, October 1991.

[19] Luttwak, G., Cowler, M.S.: *Advanced Eulerian Techniques for the Numerical Simulation of Impact and Penetration using AUTODYN-3D*, International Symposium of Interaction of the Effects of Munitions with structures, Berlin, 3-7 May, 1999.

[20] Jianfeng, L., Tao, H., Longhe, L., Bing, H.: *Numerical Simulation of Formation of EFP With Charge of Aluminized High Explosive*, International Symposium on Ballistics, Tarragona, Spain 16-20 April, 2007.

MODELING OF PRODUCTION PARAMETERS OF B₄C + ZrO₂ COMPOSITES VIA ARTIFICIAL NEURAL NETWORKS METHOD

S. Hartomacıoğlu. PhD.¹, H.O.Gülsoy. PhD.², B. Bakırcıoğlu Ms.C.³
Department of Mechanical Engineering – Marmara University, Turkey¹
Department of Metallurgy and Materials Engineering– Marmara University, Turkey²
Department of Mechatronics, Selcuk University, Turkey³
selimh@marmara.edu.tr

Abstract: In this study, the effect of production parameters of B₄C + ZrO₂ composites on density was modelled by using Artificial Neural Network (ANN). The composites were produced by using powder injection molding method (PIM). In the sintering stage, pressureless sintering method under argon atmosphere was used. As the production parameters, amount of additional (A, wt.%) and sintering temperature (T, °C) were defined. The main aim of the study is to obtain the experimental conditions giving maximum density. As a results of this study, the production parameters of hard sintered materials like B₄C + ZrO₂ could be modelled by using ANN method to optimize and predict because the prediction error is blow percentage of 10%. Therefore, the research and development time and cost can be reduced by using this method.

Keywords: POWDER INJECTION MOLDING, ARTIFICIAL NEURAL NETWORK, MODELING

1. Introduction

In the literature, most of article about sintering of boron carbide is related to fabrication of boron carbide via hot isostatic pressing and hot pressing method. There are several studies in the literature about the pressureless sintering of boron carbide with additive or non-additive. T.K. Roy et al produced boron carbide ceramics with and without additives (C, TiB₂, and ZrO₂) by pressureless sintering method to obtain dense samples for use as neutron absorber in fast breeder reactors in 2006. They investigated the effect of particle size and sintering temperature on density and microstructure. The compacts were fired in the temperature range of 2225-2375 °C under vacuum [1]. The effect of ZrO₂-3 % Y₂O₃ addition on densification, sintering behavior and mechanical properties of B₄C was studied by H.R. Baharvandi et al in 2006. The adding amount of ZrO₂-3 % Y₂O₃ was 0-30 wt.% and sintering temperature was between 2050-2150 °C [2]. In the 2007, the B₄C-metal boride composites were derived from B₄C/metal oxide mixture by A. Goldstein et al. The green pellets were sintered from room temperature to 2180 °C under Ar atmosphere via pressureless sintering method. In this study, TiO₂, ZrO₂, V₂O₅, Zr₂O₃, Y₂O₃ and LaO₃ were used for addition [3].

In this study, B₄C + ZrO₂ composites based part produced by powder injection molding method were investigated Density properties of the sintered products were evaluated in sintered condition. The production parameters were modelled using artificial neural network method.

2. Materials and Method

Materials

Boron Carbide (B₄C) is an important non-metallic hard material with high melting point (2450 °C), high hardness (25 to 35 GPa- next only to diamond and cubic boron nitride), high elastic modulus (450 GPa), high flexural strength (350-500 MPa) and low density (2.52 g/cm³) [1, 4]. ZrO₂ is the ceramic material with adequate mechanical properties. In this thesis study, commercially available B₄C powder (ABSCO Co, UK) was used. Additive powder, ZrO₂ Stack, Germany) powders were used. For molding the mixing powders, binder materials and their rates must be defined. The primary required the binder is to allow flow of the particles into the cavity. Paraffin Wax (PW), Carnauba Wax (CW), Polypropylene (PP) and Stearic Acid (SA) were selected for binder system.

Some properties of binder system components are shown in Table 1.

Table 1: Some properties of the binder system components

Binder Type	Density (g/cm ³)	Melting Point °C
Paraffin Wax (MERC)	0.9	90
Carnauba Wax (MERC)	0.97	112
Polypropylene (MERC)	0.89	161
Stearic Acid (MERC)	0.85	73

Method

Powder Injection Molding (PIM) which enables to produce products with high dimensional accuracy in such a way to have excellent, fine grain structure and non-anisotropic mechanical properties [5] is an advanced manufacturing technology. The other words, PIM method is a new technology and uses the shaping advantage of injection molding but it is applicable to metals and ceramics. Complex, precision, and net shape components are produced by PIM method from metal or ceramic powder. In recent years, PIM has established it-self as a cost-effective production technique derived from plastic injection molding, allowing large scale production of complex part [6]. The product produced by PIM is expected to have more homogenous microstructure since hydraulic pressure is filled up uniformly. Also. Fabrication cost could be eliminated significantly by reducing machining and recycling use of feedstock [7].

PIM method has four steps: 1) Preparation of feedstock, 2) Injection of molding, 3) Solvent and thermal debinding, 4) Sintering [8, 9]. The flow chart of PIM method is shown in Figure 1.

The first mold was designed based on Metal Powder Industries Federation (MPIF) standard 50. The mold was shown in Figure 2.

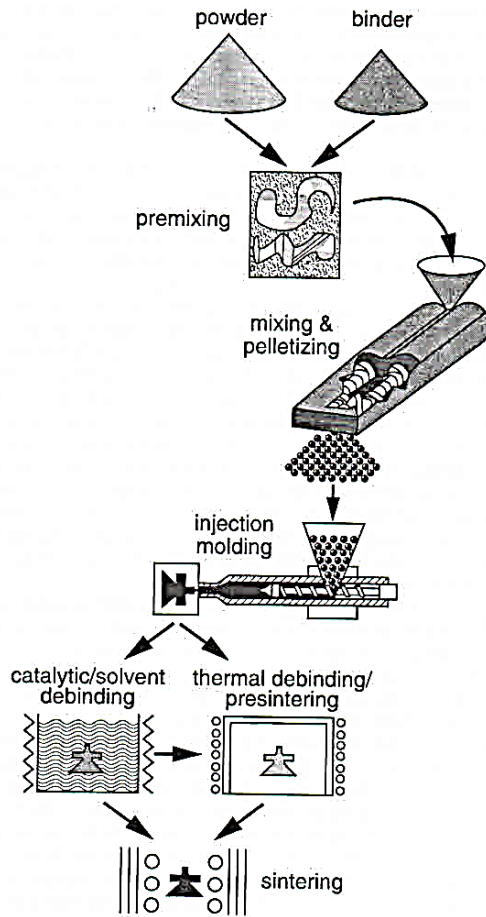


Figure 2 Flow chart of powder injection molding method [7]

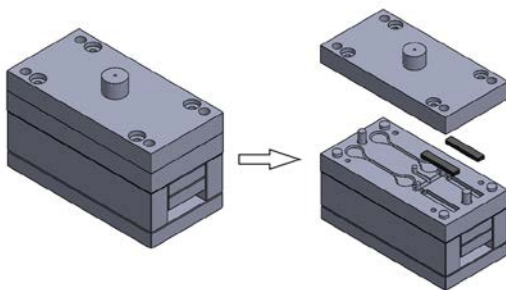


Figure 3 The mold based on MPIF 50 and 59 standard

As a results of pre-experimental studies of PIM methods, the optimal binder materials ratio was defined and listed in Table 2.

Table 2 Mixture ratio of binder system

Binder Types	PW (wt. %)	CW (wt. %)	PP (wt. %)	SA (wt. %)
Mixture Ratio	69	10	20	1

The full factorial experimental design was used as preparation experimental table. The Input parameters and their levels of experimental design was shown in Table 3.

Factors	1	2	3	4	5	6
A: Amount of Ad. (wt.%)	0	2	4	8	12	16
B: Sint. Temp.(°C)	2000	2100	2200			

The total experimental number according to full factorial experimental design is 18 experiments. The experiment conditions of each experiment number was prepared. The main powder and additive powder was mixed using Turbula Mixer with 3 dimensional motion to be obtained homogenous mixing. The next step, the feedstock was prepared for powder injection molding procedure. In this step, special custom made device was used. The mixed powder and binder system materials was mixed under 130 °C temperature and 200 rev./min mixing speed. The granules were obtained by hand and the standard samples were injected by custom made powder injection molding device. In the injection molding device, molding pressure was 15 bar, clapping pressure 10 bar, barrel temperature was 160 °C, and Molding speed was 15 sec. From the injected parts, the binder other than Polypropylene was achieved debinding in solvent. In this step, to provide uniform temperature, hot water was circulated continuously around the container of samples with heptane. Thermal debinding process was performed in furnace under Ar atmosphere. After the debinding processes, the samples were sintered using custom made sintering furnace with graphite resistance under Ar atmosphere.

After the sintering, the samples were cut by means of diamond cutting disc. The density of each part was measured by Archimedes Method.

Artificial Neural Network

After the experimental studies, the experimental numerical results were obtained. In the ANN processes, first step the data was normalized between 0 and 1. Then, the data was divided into training data set (75%) and testing data set (25%). After this step, the ANN topology was prepared. In this study, multilayer perceptron (MLP) of ANN structure was used. In this structure, there are three layers: input layer, hidden layer and output layer. In the input layer, the number of processes element (artificial neuron) corresponds to the input parameters of experiments. The output layer processes elements correspond to the output parameters of experiments. In this study, there are 3 artificial neurons in input layer, and there are 1 artificial neuron in output layer. The number of hidden layer neurons was found to be between 1 and 50 by trial and error method under 50 000 epochs. In this study, there are 3 step of ANN procedure, 1) training, 2) testing, and 3) production step. In the training step, optimal ANN structure was obtained using mean square error (MSE) performance criterion. The next step, testing step, the test performance of optimal ANN structure was tested and evaluated. Final step of ANN procedure, the new experimental results were predicted using this optimal ANN structure. The ANN structure of this study was shown in Figure 4.

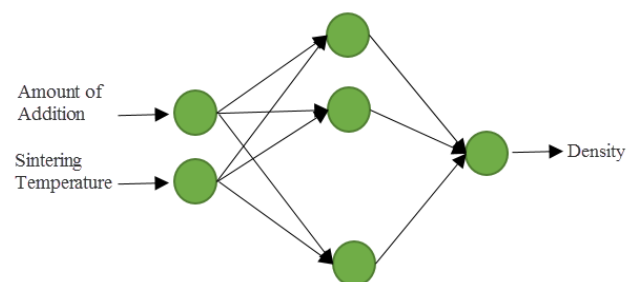


Figure 4 The ANN structure of this study

3. Results and Discussion

After the experimental studies, the experimental results were obtained and the relative density was calculated for each experimental conditions. The experimental results were listed in Table 3.

Table 3 The experimental results

#	A [wt.%]	B [°C]	Y1[%]
1	0	2000	58
2	0	2100	63
3	0	2200	68
4	2	2000	61
5	2	2100	95
6	2	2200	85
7	4	2000	57
8	4	2100	96
9	4	2200	73
10	8	2000	69
11	8	2100	72
12	8	2200	75
13	12	2000	66
14	12	2100	73
15	12	2200	72
16	16	2000	64
17	16	2100	67
18	16	2200	71

After the applying ANN procedure, the training results was shown in Figure 5. The optimal ANN structure was define using this results. In the optimal ANN structure, the number of hidden layer neurons is 20. This value was obtained at 50 000 iterations. Final MSE is 0.46.

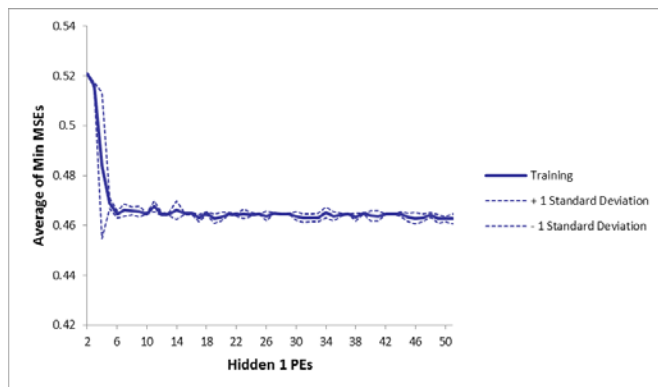


Figure 5: Training results of ANN

In the test steps, the optimal ANN was evaluated by two methods. In the first method, the testing operations was conducted by using training data set, the second method, the testing operations was performed by using test data set. The training data set testing results was shown in Figure 6. In this operations, the correlation coefficient is 0.99, and the percentage error is 0.028%.

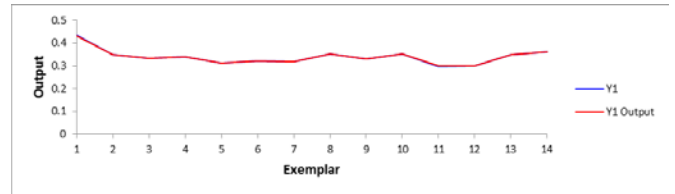


Figure 6 results of training data set testing

The second test operations results were shown in Figure. In this step, the percentage error is %9.36, and correlation coefficient is 0.95.

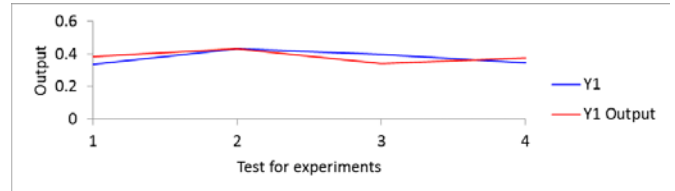


Figure 7 Results of testing data set testing

4. Conclusion

In this study, the production parameters of boron carbide composites were studied and evaluated. Firstly, the powder injection molding method was applied. In this step, the main powder and additive powder was mixed, and then the feedstock was prepared using binder systems. After the granulation, the standard samples were injected. The solvent debinding and thermal debinding processes was applied and then the debinded parts were sintered using different sintering temperature. The densities of sintered samples were measured by using Archimedes Methods, and the theoretical density and relative density values were calculated related formula. After the experimental studies, the optimal ANN structure was obtained using training, and testing operations. In the testing steps, the percentage error of training data set was 0.029%, and correlation coefficient 0.99, the percentage error of testing data set was 9.36% and correlation coefficient is 0.95. The optimal ANN stricter consist three layer; 1) input layer (3 artificial neurons), 2) hidden layer (20 artificial neurons), and 3) output layers (1 artificial neurons). As a results of this study, the artificial neural networks method was used to modeling the production parameters of hard sintered materials like boron carbide composites.

5. Acknowledgements

This work is support by Scientific Research Project Program of Marmara University (Grand: **FEN-K-110117-0019**).

6. References

- [1] Roy, T.K., Subramanian, C., Suri, A.K. (2006) Pressureless Sintering of Boron Carbide. *Ceramic International*, 32, 227-233
- [2] Baharvandi, H.R., Hadian, A.M., Abdizadeh, A. (2006) Investigation on Addition of ZrO₃-3 mol% Y₂O₃ Powder on Sintering Behaviour and Mechanical Properties of B₄C. *Journal of Materials Science*, 41(16), 5269-5272
- [3] Goldstain, A., Yeshurun, Y., Goldenberg, A. (2007) B₄C/Metal Boride Composites Derived from B₄C/Metal Oxide Mixtures. *Journal of the European Ceramic Society*, 27, 695-700
- [4] Beaudet, T.D., Smith, J.R., Adams, J.W. (2015) Surface energy and relaxation in boron carbide (1011) from first principles. *Solid State Communications*, 219, 42-47
- [5] Gülsoy, H.Ö., Özgün, Ö., Bilketaş, S. (2016) Powder injection molding of Stellite 6 powder: sintering, microstructural and mechanical properties. *Material Science & Engineering A*, 651, 914-924

- [6] Bleyan, D., Hausnerova, B., Svoboda, P. (2015) the development of powder injection molding binders: A quantification of individual interactions. *Powder Technology*, 286, 84-89
- [7] German, R.M. (1996) *Sintering Theory and Practice*, Wiley, San Diego
- [8] Baojun, Z., Xuanhui, Q., Ying, T. (2002) Powder injection molding of WC-8%Co tungsten cemented carbide. *International Journal of Refractory Metals & Hard Materials*, 20, 389-394
- [9] Nyberg, E., Miller, M., Simmons, K., Weil, K., S. (2005) Microstructure and mechanical properties of titanium components fabricated by a new powder injection molding technique. *Materials Science and Engineering C*, 25, 336-342

IN SILICO MODELING AND EVALUATION OF BASFIA SUCCINICIPRODUCENS FOR 1,4-BUTANEDIOL PRODUCTION FROM RENEWABLE RESOURCES

Dr. Zsolt, Bodor Assistant professor
Sapientia Hungarian University of Transylvania, Romania

Abstract: *Today's biology has become a data-rich field and the utilization of these data to the deep understanding of function and structure of biological systems is mandatory. Hence, genome-scale metabolic models (GEMs) are created in order to analyse complex biological systems and incorporate all the genomic, proteomic and metabolic data available into mathematical models. GEMs can be used for example to predict metabolic fluxes or redesign the metabolism of microorganisms to create industrially important strains, capable of producing various natural and non-natural metabolites. The extending concept of bio-based economy is based on the bioconversion of renewable feedstocks to value added chemicals such as succinic acid or 1,4-butanediol (BDO). Basfia succiniciproducens is a relatively unexplored succinic acid producing bacterium with advantageous features such as broad substrate utilization or facultative anaerobic metabolism, with application possibilities in today's metabolic engineering-based biotechnologies. The purpose of this study was to investigate in silico the BDO production potential of B. succiniciproducens using formerly designed biosynthetic pathways of BDO starting from different carbon sources. To our best knowledge, this is the first attempt to analyse the host strain metabolic potential for BDO production from glucose, glycerol and xylose in this host chassis using a systems biology approach.*

МОДЕЛИРОВАНИЕ ПОЛНОЙ МГД-ЗАДАЧИ В СПИРАЛЬНОМ ТОРОИДАЛЬНОМ ПОТОКЕ

SIMULATION OF A FULL MHD-PROBLEM IN A HELICAL TOROIDAL FLOW

Чупин Антон Викторович кандидат, физ. - мат. наук, Научный сотрудник

РЕЗЮМЕ: Магнитогидродинамическое динамо, заключающееся в самогенерации магнитного поля в определённых трёхмерных потоках проводящей среды, в настоящее время является главным претендентом на описание магнитных процессов, происходящих на Солнце и внутри Земли. Однако полностью этот эффект не исследован, в частности, чрезвычайно трудно его воспроизведение в лабораторных условиях. В лаборатории физической гидродинамики Института механики сплошных сред Пермского федерального исследовательского центра проводится динамо-эксперимент, основой которого является создание спирального течения жидкого металла в тороидальном канале с электропроводящими стенками. Имеющиеся численные эксперименты показывают теоретическую возможность генерации магнитного поля, однако они все предполагают кинематический подход (без обратного влияния магнитного поля на течение), т. е. описывают эффект только в начале зарождения поля. Дальнейшее описание совместной эволюции магнитного поля и поля скорости жидкости невозможно без решения связанной ("полной") магнитогидродинамической задачи.

В докладе будут описаны подходы к решению полной задачи о генерации магнитного поля и спирального потока несжимаемой электропроводящей жидкости в тороидальном канале и приведены результаты численного моделирования. Поскольку динамо — пороговый эффект, для каждого течения существует минимальная интенсивность (определяемая обычно магнитным числом Рейнольдса), при которой он возникает. В сверхкритических режимах амплитуда магнитного поля сначала возрастает экспоненциально (кинематическое динамо), а затем наступает т. н. насыщение — магнитное поле изменяет гидродинамическое так, что скорость генерации становится равной 0. В докладе будут представлены установившиеся конфигурации магнитного и гидродинамического полей для нескольких начальных параметров. Будет показано сравнение результатов, полученных с помощью прямого численного моделирования ламинарного течения и с помощью моделирования турбулентного течения в свободном гидродинамическом пакете OpenFoam.

KEYWORDS: MHD-DYNAMO, PERM DYNAMO, TOROIDAL FLOW, HELICAL FLOW, HELICITY, DIRECT NUMERICAL SIMULATION, DYNAMO SATURATION

MODELLING AND EDUCATION: THE ROLE OF MATHEMATICAL MODELLING IN THE REALIZATION OF CONTINUITY OF THE STOCHASTIC LINE IN THE SCHOOL COURSE OF MATHEMATICS

МОДЕЛИРОВАНИЕ И ОБРАЗОВАНИЕ: РОЛЬ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ В РЕАЛИЗАЦИИ НЕПРЕРЫВНОСТИ СТОХАСТИЧЕСКОЙ ЛИНИИ ШКОЛЬНОГО КУРСА МАТЕМАТИКИ¹

Щербатых Сергей Викторович

доктор педагогических наук, доцент, проректор по учебной работе Елецкого государственного университета им. И.А. Бунина
399770, Липецкая область, г. Елец, ул. Коммунаров, д.28
+74746720275; shcherserg@mail.ru

Резюме: Потребность в использовании практических материалов при обучении стохастике определяется тем, что возникновение, формирование и развитие основных стохастических понятий и идей имеют своим источником чисто человеческие ощущения и восприятия.

Понятия, созданные современной стохастикой, порой кажутся весьма далёкими от реального мира. Однако именно с их помощью людям удалось постичь тайны строения атомного ядра, познать ход химических реакций, рассчитать движение космических кораблей, создать весь тот мир техники, на котором основано современное производство. Одним из основных методов познания природы является опыт. С помощью опытов были установлены многие законы природы. Но не всегда целесообразно проводить опыт. Одним из наиболее плодотворных методов стохастического познания окружающей действительности является метод построения математических моделей изучаемых реальных объектов или объектов, уже описанных в других областях знаний, с целью их более глубокого изучения и решения всех возникающих в этих реальных ситуациях задач с помощью математического аппарата.

Включение математического моделирования в учебный процесс делает его более рациональным и одновременно активизирует познавательную деятельность учащихся. Следовательно, на уроке математики осуществляется развитие учащихся. Моделирование отражает теоретический стиль мышления, который содействует развитию учащихся и приобщает их к научному стилю мышления, поэтому понятие математической модели и некоторые общие положения, связанные с ним, должны в той или иной форме иллюстрироваться на протяжении всего периода обучения стохастике в школе.

Доклад будет посвящен проблеме реализации метода математического моделирования при непрерывном обучении новому компоненту школьного математического образования – стохастике.

КЛЮЧЕВЫЕ СЛОВА: МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ, ОБЩЕОБРАЗОВАТЕЛЬНАЯ ШКОЛА, НЕПРЕРЫВНОСТЬ СТОХАСТИЧЕСКОЙ ЛИНИИ.

Abstract: The need for the use of practical materials when teaching stochastics is defined by the fact that the emergence, formation and development of the basic stochastic concepts and ideas have the source purely human feelings and perceptions.

The concepts created by the modern stochastics sometimes seem very far from the real world. However with their help people succeeded to comprehend the mysteries of the structure of an atomic nucleus, to learn the course of the chemical reactions, to calculate the movement of spaceships, to create all that world of the equipment on which the modern production is founded. One of the main methods of the knowledge of nature is experience. By means of experiences many laws of nature have been established. But it isn't always expedient to make experiment. One of the most fruitful methods of stochastic knowledge of the surrounding reality is the method of the creation of the mathematical models of the studied real the objects or objects which are already described in other fields of knowledge, for the purpose of their deeper studying and the solution of all arising in these real situations of tasks by means of the mathematical apparatus.

The inclusion of mathematical modelling into the educational process makes it more rational and at the same time it stirs up the pupils' cognitive activity. Therefore, at a lesson of mathematics the development of pupils is carried out. Modelling reflects the theoretical style of thinking which promotes the pupils' development and acquaints them with the scientific style of thinking therefore the concept of mathematical model and some related general provisions have to be illustrated in this or that form throughout the entire period of teaching stochastics at school.

The report will be devoted to the problem of realization of the method of mathematical modelling while teaching continuously a new component of the school mathematical education – stochastics.

KEYWORDS: MATHEMATICAL MODELLING, COMPREHENSIVE SCHOOL, CONTINUITY OF THE STOCHASTIC LINE.

¹ Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (отделение гуманитарных и общественных наук). Проект 17-36-01004 «Теоретико-методические основы реализации непрерывности и преемственности в развитии стохастической линии школьного курса математики в русле идей системно-деятельностного подхода».

Через понятие математической модели раскрывается двойная связь математики с реальным миром. С одной стороны, математика служит практике по изучению и освоению объектов окружающего реального мира, с другой – сама жизнь, практика способствуют дальнейшему развитию математики.

Придерживаясь мнения авторитетных учёных [1-4], основу метода математического моделирования составляют следующие этапы процесса математизации:

- этап перехода от ситуации, которую необходимо разрешить, к формальной математической модели этой ситуации, к чётко поставленной математической задаче – этап формализации;

- решение поставленной математической задачи методами, развитыми в самой математике для задач данного типа, составляет содержание второго этапа – этапа решения задачи внутри построенной математической модели;

- интерпретация полученного решения математической задачи, применения этого решения к исходной ситуации и сопоставления его с нею.

Коротко эти три этапа можно назвать:

- 1) построение математической модели;
- 2) получение математических результатов;
- 3) принятие решения (выводы в реальном мире).

Следует отметить, что недооценка каждого из рассмотренных этапов приводит к существенным затруднениям в использовании метода математического моделирования.

Наиболее ответственным является первый этап – построение математической модели. Оно осуществляется логическим путём на основе глубокого анализа изучаемого явления и требует умения описать явление на языке математики (в частности, стохастики). В процессе построения модели можно выделить несколько шагов.

Первый шаг – индуктивный. Он заключается в отборе наблюдений, относящихся к тому процессу, который предстоит в последующем моделировать. На этом шаге формулируется проблема, то есть принимается решение относительно того, что следует принимать во внимание, а чем можно пренебречь.

Второй шаг заключается в переходе от определения проблемы к построению модели, пусть ещё неформальной. На данном шаге рассматривается ряд наборов неформальных допущений, способных объяснить одни и те же данные. Таким

образом, рассматриваются несколько различных моделей и решается, какая из них лучше всего отображает изучаемый процесс.

Третий шаг заключается в переводе полученной, ещё «неформальной» модели в математическую модель, и включает в себя рассмотрение словесного описания этой модели, а также поиск подходящего математического аппарата. Следует отметить, что это самый сложный этап во всём процессе моделирования. При этом стадия перевода таит в себе ряд опасностей. С одной стороны, сами по себе «неформальные» модели часто неоднозначны, и поэтому существует несколько способов перевода «неформальной» модели в математическую. С другой стороны, язык математики (в частности, стохастики) лишён двусмысленностей и более точен, чем житейский язык. Благодаря ему исследуется скрытый смысл различий в формулировках, который практически недоступен исследованию посредством житейского языка.

Следующий этап – получение математического результата, т.е. решение задачи в рамках математической теории. Он является решающим в математическом моделировании. На данном этапе применяются все известные ученику стохастические методы с целью формального вывода следствий из исходных допущений модели. На стадии получения математического результата имеют дело с чистыми математическими абстракциями и используют одинаковые математические средства. Данный этап представляет собой дедуктивное ядро процесса математического моделирования.

На последнем этапе полученные в ходе решения выводы проходят через ещё один процесс перевода – с языка чистой стохастики на житейский язык.

Примером, иллюстрирующим сказанное, является следующая задача.

Задача. При посеве и последующем самоопылении партии семян гороха, полученного от скрещивания растений с жёлтыми гладкими и зелёными морщинистыми семенами, было получено: 876 растений с жёлтыми гладкими, 321 – с жёлтыми морщинистыми, 298 – с зелёными гладкими и 115 – с зелёными морщинистыми семенами. Можно ли считать, что взятые для посева семена являлись носителями «чистых» линий и можно ли оставшуюся партию использовать в дальнейших генетических исследованиях?

Решение. Построение математической модели. Пусть $f_{эмп_i}$ – количество растений гороха, полученных в опыте и наделённых определённым признаком. Тогда:

$f_{эмп_1}$ – количество растений с жёлтыми гладкими семенами;

$f_{эмп_2}$ – количество растений с жёлтыми морщинистыми семенами;

$f_{эмп_3}$ – количество растений с зелёными гладкими семенами;

$f_{эмп_4}$ – количество растений с зелёными морщинистыми семенами, где объём выборки равен

$$n = 876 + 321 + 298 + 115 = 1610.$$

$$\text{Тогда } f_{эмп_1} = 876, f_{эмп_2} = 321, f_{эмп_3} = 298, f_{эмп_4} = 115.$$

О «чистоте» линий можно судить лишь только в том случае, если до скрещивания материал был гомозиготным по двум парам аллелей. Тогда во втором поколении следует ожидать расщепление в отношении 9:3:3:1, что согласуется с 3-им законом Г. Менделя. Таким образом, задача сводится к установлению того: подчиняется ли подобное расщепление 3-му закону Г. Менделя (дигибридное скрещивание) или нет?

Выдвигаем гипотезы:

$$H_0 = \{\text{полученное расщепление – дигибридное с расщеплением 9:3:3:1}\};$$

$$H_1 = \{\text{расщепление не является дигибридным}\}.$$

В качестве критерия значимости выбираем критерий «хи-квадрат» Пирсона: $\chi^2_{эмп} = \sum_{i=1}^n \frac{(f_{эмп_i} - f_{теор_i})^2}{f_{теор_i}}$ (объём вы-

борки $n = 1610 \geq 20$) – математическая модель рассматриваемой ситуации. Здесь $f_{теор_i}$ – предполагаемое количество растений гороха, наделённых определённым признаком и согласующихся с теоретическим распределением.

Получение математических результатов. Для применения данного критерия необходимо рассчитать теоретические частоты (частоты контрольной выборки), согласующиеся с дигибридным скрещиванием.

Так как $9 + 3 + 3 + 1 = 16$, то

$$f_{теор_1} = \frac{9}{16}n = \frac{9}{16} \cdot 1610 \approx 905,6;$$

$$f_{теор_2} = \frac{3}{16}n = \frac{3}{16} \cdot 1610 \approx 301,9;$$

$$f_{теор_3} = \frac{3}{16}n = \frac{3}{16} \cdot 1610 \approx 301,9;$$

$$f_{теор_4} = \frac{1}{16}n = \frac{1}{16} \cdot 1610 \approx 100,6.$$

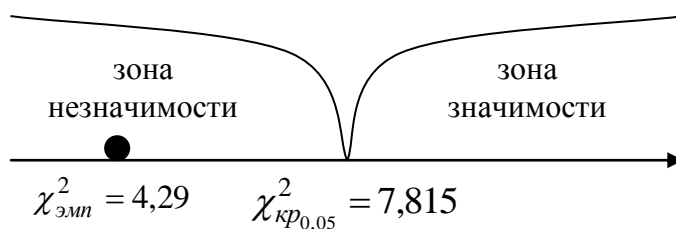
Рассчитаем эмпирическое значение критерия:

$$\chi^2_{эмп} = \frac{(876 - 905,6)^2}{905,6} + \frac{(321 - 301,9)^2}{301,9} + \frac{(298 - 301,9)^2}{301,9} + \frac{(115 - 100,6)^2}{100,6} \approx 4,29.$$

Чтобы найти критическое значение критерия, необходимо знать число «степеней свободы» и уровень значимости. В данной задаче $\nu = 4 - 1 = 3$, так как имеем 4 класса-интервала и одно ограничение (объём выборки должен быть равен n). В качестве уровня значимости выберем $\alpha = 0,05$.

По статистической таблице критерия находим критическое значение: $\chi^2_{кр0,05} = 7,815$.

Строим «ось значимости»:



В нашем случае $\chi^2_{эмп} = 4,29$ попало в зону незначимости. В соответствии с правилом принятия решения мы оставляем гипотезу H_0 и отвергаем H_1 .

Принятие решения (выводы в реальном мире). При переводе результата, полученного в ходе математических вычислений, заключаем, что данное расщепление в полной мере удовлетворяет 3-му закону Г. Менделя, и поэтому мы можем утверждать, что взятые для посева семена являлись носителями «чистых» линий. Таким образом, оставшуюся партию семян можно использовать в дальнейших генетических исследованиях.

В непосредственной практике традиционного обучения мир реальных объектов подменяется изучением соответствующих им понятий и других продуктов познания, полученных не учениками, а специалистами, учёными или авторами учебного материала.

Подобное положение усугубляется при изучении основ науки о случайном. К примеру, классическая вероятность, базируемая на гипотезе равновероятности исходов, является математической моделью, отсечённой от этапов формализации и интерпретации. Вычисление вероятностей с помощью комбинаторных правил, их косвенный подсчёт на основе заданных, неизвестно откуда взятых вероятностей – всё это представляет собой оперирование внутри математических моделей.

При таком обучении некоторые понятия (вероятность, математическое ожидание, дисперсия, мода, медиана, статистиче-

ский критерий и т.п.) воспринимаются как искусственные и чуждые по отношению как к самой математике, так и к жизни. Не случайно многие обучаемые испытывают внутреннее психологическое сопротивление этой науке. Поэтому стохастику необходимо изучать не как чисто математическую дисциплину, а как прикладную при явном вовлечении этапов формализации и интерпретации в процесс решения задач, причём процесс этот должен идти непрерывно – на всём этапе изучения стохастики. При таком изучении вырабатываются умения решать стохастические задачи, выдвигаемые практикой, что является критерием достижения поставленных целей.

Литература

1. Плоцки А. Вероятность события в стохастической линии школьного математического образования // Математика в школе. 1997. № 2. С. 24-28.
2. Плоцки А. Стохастика в математике «для всех». Краков, 1988. 244 с.
3. Терешин Н.А. Прикладная направленности школьного курса математики. М.: Просвещение, 1990. 96 с.
4. Фирсов В.В. О прикладной ориентации курса математики // Математика в школе. 2006. № 6. С. 2-9.

THE SUBJECTIVE MODEL OF RATIONAL CHOICE IN MULTI-AGENT SYSTEMS

Gennady Pavlovich Vinogradov
Тверской государственный технический университет
Tver
wgp272ng@mail.ru

Abstract: The paper considers the problem of modeling agent's choice, which allows explaining made decisions, as well as predicting possible options. The approach is based on the ideas of a subjectively rational choice. The subjectively rational choice supposes that a choice motivation is determined by external and internal factors. Internal factors represent the interests of a subject induced by his needs and ethical system he follows. External factors are induced by accepted obligations. The agent can estimate his satisfaction with the current goal-seeking state situation. The estimation might lead to changing a structure of interests, so the agent can choose it. The paper shows that when making decisions the agent uses three sets of alternatives as follows: controlling, structural and identification. This supposes the existence of three virtual sides, which choose relevant alternatives. The selection rules for such alternatives depending on subject's awareness of the situation and the structure of his interests are formed by finding a compromise.

KEYWORDS: REFLEXIVE CONTROL, DECISION MAKING, MODEL, COMPROMISE.

1 Introduction

Agent's behavior control, when the agent has mentality and will, has become possible after appearing of decision-making models, which take into account his subjective understanding of a choice situation. These models helped estimating control efficiency before controlling [3]. It should be noted, that a decision-making process was an uncontrolled factor in a normative decision-making theory.

The development of the idea of subjective rational choice [3] helped to:

- 1) explain decision-making by a subject in specific situations;
- 2) predict possible reactions of the other subject in various situations (for a decision-maker);
- 3) solve a problem of active prediction when a control side creates an appropriate image of the future for a controlee.

A subjective rational choice supposes that the choice motivation is determined by external and internal factors. Internal factors reflect subject's interests, which are induced by his needs and an ethical system. External factors are connected with obligations. Affective evaluation of the level of subject's satisfaction of current choice situation, as shown in [4], might lead to changing the structure of subject's interests, so he is able to choose it. Subject's preferences in a selection process reflect his interests, thus we can determine a number of G alternatives of a preference structure (structural alternatives according to [1]).

2 Background assumptions

1. Subject's choice is based on his view of a choice situation.
2. The components of his view reflect various aspects of understanding choice situation by the subject and create an information structure of representations. Many optional representations will be X .
3. For many surrounding conditions S a set of observed surrounding conditions meets the condition $S \cap X \neq \emptyset$, i.e. subject's view might include an objective part, as well as a phantom one.
4. The subject chooses structural alternatives depending one valuation of the level of satisfaction of choice situation property values.
5. Formation of the view is based on the procedures of perception, awareness and analysis according to subject's cognitive capabilities.

According to the abovementioned assumptions, the subject uses three sets of alternatives in decision-making: control C (modes of action), structural G and identification X . Therefore, it is possible to assume the existence of three virtual sides, which choose corresponding alternatives. A choice rule for such alternatives depending on understanding of the surrounding and the structure of interests by a subject will be called *strategies* here.

Let us assume that decision-making has several cyclical steps,

and modes of action are chosen at each step $n = 1, 2, \dots$ from the set C depending on the view of surrounding conditions $x \in X$. This is due to the fact that a joint over conscious (intuitional) and conscious (formal) analysis of surrounding conditions allow firstly accepting a vaguely realizable decision in multiple iterations, then more clear and grounded decision. There are also some restrictions $C_x \subseteq C$ to allow the choice of alternatives depending on the vision of surrounding conditions $x \in X$. Process dynamics in subject's surrounding is out of reach for direct perception, therefore its representations are formed using identification with the idea of choosing a vision alternative depending on the observed conditions. Here there are restrictions $X_s \subseteq X$ to allow representations as identification alternatives depending on the observed conditions $s \in S$.

Proceeding from these assumptions, according to [1], we introduce the strategy determinations.

A single-valued transformation $\lambda : X \rightarrow C$ so that $\lambda(x) \in C_x$, $x \in X$, is called a choice or control function; an ordered set $(\lambda_1, \dots, \lambda_n) \equiv \lambda_1^n$ is called a choice strategy on the length horizon $n < \infty$; $\lim \{\lambda_1^n\} = \lambda_1^\infty$ when $n \rightarrow \infty$ is a strategy directed to achieving a local ideal, which determines subject's reason for existence.

A single-valued monotone transformation $\xi : S \rightarrow X$ so that $\xi(s) \in X_s$, $s \in S$, is called an identification function; an ordered set $(\xi_1, \dots, \xi_n) \equiv \xi_1^n$ is called an identification strategy on the length horizon $n < \infty$; a consecutive order $\{\xi_1^n, n = 1, 2, \dots\}$ is called an identification strategy on the limited horizon. Due to the fact, that the subject tends to forming useful vision, so there is $\lim \{\xi_1^n\} = \xi_1^\infty$ при $n \rightarrow \infty$.

As the sets S and X meet the condition $|S| > |X|$, therefore a single-valued transformation $\xi : S \rightarrow X$ causes separation of the set S on subsets

$$\xi^{-1}(x) = \bigcup \{s \in S : \xi(s) = x\} \subset S, \quad x \in X.$$

Subsets $\xi^{-1}(x) \subset S$, $x \in X$ are associated sets, i.e. any element $s \in \xi^{-1}(x)$ uniquely determines appropriate representation $x \in X$. Therefore, it can be said that subsets $\xi^{-1}(x) \subset S$, $x \in X$ form classes of equivalent representations. It allows using the methods of the fuzzy sets theory to formalize subject's representations, e.g. as described in [4].

A structure alternative $\gamma_n \in G$ chosen at the moment n is a *structure choice* at n -th step of decision-making; an ordered set $(\gamma_n, \dots, \gamma_1) \equiv \gamma_1^n$ is a *structure choice* strategy on a decision-making horizon of the length $n < \infty$; a consecutive order $\{\gamma_1^n, n = 1, 2, \dots\}$ is a structure choice strategy on the limited horizon. As the

subject tends to the correspondence between its structure of interests and requirements of the accepted ethical system, so there is $\lim \{\gamma_1^n\} = \gamma^\infty$ when $n \rightarrow \infty$.

3 A decision-making model with changing preference structure

According to [4] a selection criteria for a control strategy have the meaning of a desired specific value of a purposeful condition based on the result with formalization, which has the formula for the utility function $E\varphi^g(C \times S \times X)$ that depends on the structure alternative $g \in G_s$ on a parameter. As the control process starts with a certain situation $x \in X$, then the criterion $E\varphi_n(\lambda_1^n | \gamma_1^n)$ will also depend on the situation $x \in X$ as from the initial condition. A number of situation X is finite, so the criterion $E\varphi_n(\lambda_1^n | \gamma_1^n)$ will be definitely represented by a vector in the space R^X of the dimension $|X|$. Its parts will be written as $E\varphi_n(\lambda_1^n | \gamma_1^n)(x)$, $x \in X$. According to the choice result, the subjects offers emotionally, so the quality of the structure choice strategy γ_1^n should be described as a criterion meaning “choice results satisfaction”. Therefore, the quality of the structure γ_1^n is natural to describe by a folding of an expected utility vector $E\varphi_n(\lambda_n | \gamma_1^n) \in R^X$ in a certain composite function $\mu: R^X \rightarrow R^1$. Then a strategy quality criterion γ_1^n might be written as

$$\mu_n(\lambda_1^n | \gamma_1^n) = \mu(E\varphi_n(\lambda_1^n | \gamma_1^n)) \in R^1.$$

The subject associates its representation quality with evaluation of possibilities of achieving desired conditions when controlling $c \in C$, as well as with the possibility of extending the number of $C \uparrow$ by including efficient alternatives. The paper [6] uses the terms of the linguistic variable “utility”, which are based on the values $E\varphi_n(\lambda_1^n | \gamma_1^n)$, as a representation estimation criterion. In these conditions utility estimates will depend on control strategies λ_1^n , a structure choice γ_1^n as on postulated conditions. The “utility” criterion will be labeled as follows $\psi_n(\xi_1^n | \lambda_1^n, \gamma_1^n)$. As identification starts with a certain state $s \in S$, this criterion will depend on the state $s \in S$, assigned as the initial condition. Here the set of states S is finite, so the identification criterion will be represented by a vector $\psi_n(\xi_1^n | \lambda_1^n, \gamma_1^n)$ in the space R^S of the dimension $|S|$.

In a goal-seeking state situation of the quality of control strategies and structure choice is described by criteria $E\varphi_n(\lambda_1^n | \gamma_1^n) \in R^X$ and $\mu_n(\gamma_1^n | \lambda_1^n) \in R^1$ respectively. They have a meaning of specific value by a result and satisfaction with choice results. The identification strategy quality is described by the criterion $\psi_n(\xi_1^n | \lambda_1^n, \gamma_1^n) \in R^S$, which has a meaning of representation utility to achieve desired states. The use of introduced criteria assumes determination of corresponding information structures or models that allow making an appropriate choice.

Let us assume the existence of an information structure of representations I , which reflects subject’s knowledge and experience on: modes of action (control), his own interests and preferences, dynamics of surrounding transition into different states. Therefore, it is likely that there is a structural transformation of this structure into an information structure, which enables creating a

specific value criterion $E\varphi_n(\lambda_1^n | \gamma_1^n)$ and a domain model. Let us call such transformation a “specific value transformation”, and the induced information structure will be called “information structure of a specific value of a goal-seeking state situation by a result” and designated as $U = U(I)$.

In a similar way, if there is a structural transformation of the structure I into an information structure, which enables creating an identification criterion $\psi_n(\xi_1^n | \lambda_1^n, \gamma_1^n)$ and identification procedure models, then we will call such transformation “identification transformation” and designate it as R , an induced information structure will be called “identification information structure” and designated as $R = R(I)$.

Subject’s representation about a goal-seeking state situation is subjective and qualitative, based on observations and analysis of the surrounding transition process affected by control $c \in C$ into various states $s \in S$. Let us indicate the rule of such transition using $q^g(S | S \times C)$ from $S \times C$ into S . Actually, the subject for estimation of possible result value uses the model $Q^g(X | X \times C)$ from $X \times C$ into X constructed by identification strategy results ξ_1^n . When constructing we take in to account control strategies λ_1^n , a structure choice γ_1^n , or it is defined by such strategies. It means that a transformation of the operational function $q^g(S | S \times C)$ into function of subject’s understanding of the surrounding processes $Q^g(X | X \times Y)$ is possible only in the aposterior mode depending on used strategies $(\lambda_1^n, \gamma_1^n, \xi_1^n)$. Such transformation and construction of the desired specific value criterion $E\varphi_n(\lambda_1^n | \gamma_1^n)$ is possible when “utility” information structures are formed successively depending on used strategies. This condition is written as $U_n = U(\lambda_1^n, \gamma_1^n, \xi_1^n)(I)$, $n = 1, 2, \dots$. As this condition is necessary for forming the desired utility criterion and a domain model, so it should be pointed out every time it is used. Note that the criterion $E\varphi_n(\lambda_1^n | \gamma_1^n)$ tacitly depends on the identification strategy ξ_1^n due to introduction an induced structure U_n into the choice model. As it was stated above, the criterion $\mu_n(\gamma_1^n | \lambda_1^n) \in R^1$ of the structure choice quality is determined by a criterion folding $E\varphi_n(\lambda_1^n | \gamma_1^n) \in R^X$. Generality of information structure of their formation allows writing

$$\begin{cases} E\varphi_n(\lambda_1^n | \xi_1^n) \\ \mu_n(E\varphi_n(\xi_1^n | \lambda_1^n)) \\ U_n = U(\lambda_1^n, \gamma_1^n, \xi_1^n)(I). \end{cases}$$

To construct identification criterion we need to use a specific function, which would have a meaning of “utility”. For this purpose it is necessary to construct verbal estimates on the values of a function $E\varphi^g(S \times X \times Y)$. A required transformation exists and might be performed in the aprior mode (i.e. before choosing decisions). Such transformation is determined by the subject as regard to a fuzzy measure, which might be constructed when the defined function is $q^g(S | S \times C)$ from $S \times C$ into S . As its analogue has a form of $Q^g(X | X \times C)$ in subject’s mind, and he can define it in a unique manner in the information structure I , therefore, there is no need in additional transformations. “Representation utility” function construction depletes a necessary structural transformation. We will call it identification structural

transformation and define as R . The induced information structure will be called “Representation utility” information structure and defined as $R = R(I)$.

Taking into account these reasons an identification criterion is written as follows:

$$\begin{cases} \psi_n(\xi_1^n | \lambda_1^n, \gamma_1^n) \\ R = R(I) \end{cases}$$

The induced definitions and constructions show that quality criteria for strategies are different and interdependent. Therefore the choice problem has game meaning and is reduced to searching as table compromise between aiming at maximizing desired specific value of a goal-seeking state by a result and minimizing possible loss due to wrong actions. Such compromise is called *balance*.

It should be noted that the information structure $U_n = U(\lambda_1^n, \gamma_1^n, \xi_1^n)(I)$, which is a base for the criterion

$\mu_n(E\varphi_n(\gamma_1^n | \lambda_1^n))$, must be formed consequently depending on used strategies. Thus, required balances will be interdependent not only at each step $n = 1, 2, \dots$ of forming decisions, but they will also depend on the decisions chosen at the previous steps. Considering this fact, it is natural to call balances *dynamic*.

The triple of strategies $\{\lambda_1^n, \gamma_1^n, \xi_1^n\}$, which meet the conditions

$$\begin{cases} E\varphi_n(\lambda_1^n | \gamma_1^n) \geq E\varphi_n(\lambda_1^n | \gamma_1^n) \quad \forall \lambda_1^n, \\ \mu_n(\gamma_1^n | \lambda_1^n) \geq \mu_n(\gamma_1^n | \lambda_1^n) \quad \forall \gamma_1^n, \\ U_n = U(\lambda_1^n, \gamma_1^n, \xi_1^n)(I) \\ \psi_n(\xi_1^n | \gamma_1^n, \lambda_1^n) \geq \mu_n(\xi_1^n | \gamma_1^n, \lambda_1^n) \quad \forall \xi_1^n, \\ R = R(I), n = 1, 2, \dots \end{cases}$$

are called *dynamic balances*.

According to the abovementioned assumptions, the number of cycles that form decisions is unlimited. Therefore, dynamic balances must be meaningful, including the situation when $n \rightarrow \infty$.

For this purpose, it is natural to require fulfillment of the following additional conditions:

- 1) when $n \rightarrow \infty$ strategy quality criteria must tend to specific limits;
- 2) such limits cannot depend on the initial conditions.

As criteria are not assigned in an explicit form, then realization of these properties is not explicit. It requires as signing necessary properties and then indicating criteria in the explicit form, which satisfies these properties.

According to the induced assumptions, quality criteria of stationary strategies $\lambda^n, \gamma^n, \xi^n$ when $n \rightarrow \infty$ have limits. Therefore

the triple of stationary strategies $(\lambda^\infty, \gamma^\infty, \xi^\infty)$ is called stationary balances if there are limits that meet the conditions:

$$\begin{cases} \varphi(\lambda^\infty | \gamma^\infty) \geq \varphi_n(\lambda^\infty | \gamma^\infty), \forall \lambda^\infty \\ \mu(\gamma^\infty | \lambda^\infty) \geq \mu_n(\gamma^\infty | \lambda^\infty), \forall \gamma^\infty \\ U = U(\lambda^\infty, \gamma^\infty, \xi^\infty)(I) \\ \psi(\xi^\infty | \lambda^\infty, \gamma^\infty) \geq \psi(\xi^\infty | \lambda^\infty, \gamma^\infty), \forall \xi^\infty; \\ R = R(I) \end{cases}$$

As a result, the content of the choice modeling problem consists in finding a compromise between aiming at achieving a maximal desired specific value by a result and minimal loss from wrong representations taking into account their mutual dependence. According to the equilibrium solution principle, compromise must be “unimprovable” equally by all parts of interests.

When achieving such compromise it is fair to say that subject’s interests are materialized with “the best result”. Provided that dynamic balances meet the requirements of asymptotic stationary, it is also fair to say that subject’s interests are materialized with “the best result” on the unlimited horizon, including $n \rightarrow \infty$. It follows that dynamic balances determine the meaning and the method of interests materializing with “the best result”. Thus, dynamic balances naturally determine *internal aim* when making decisions.

4 Information structures in decision-making

The formal descriptions introduced determine not only the conditions of decision-making, but also proper a priori information carriers. Together, they form a set of the following formal properties:

S is an environment state set; $\beta(S)$ is a priori possibility distribution for a state set; X is a situation set; $X_S \cap X \neq \emptyset$ stands for the limitations determining the presence of the “right” ideas as diagnostics alternatives depending on $s \in S$ states; C is a control alternative set; $C_x \subseteq C$ means control alternative feasibility limitation depending on $x \in X$ situations; G is a structural alternative set; $q^g(S | S \times C)$ is a transitional function of $S \times C$ to S ; $E\varphi^g(C \times (S \times X))$ stands for the utility function representing a priori preferences for $c \in C$ alternatives depending on $s \in S$ states, $x \in X$ situations and $g \in G$ structural alternatives.

This set defines an a priori information structure which is to be set according to decision-making rules.

The peculiarity of conditions of an information structure is that it is supposed to include a task of both states and situations, the choice of control actions depending on situations which, being qualitative characteristics, representing relationship to a state, are inaccessible for direct observation and need preventive maintenance.

Under these conditions, the regularity of situation dynamics cannot be set a priori. Therefore, decision-making rules suggest posing the laws of dynamics states only, defined by the

$q^g(S | S \times C)$ transition function from $S \times C$ to S . In this sense, information given a priori is minimal.

In the conditions of a priori information deficiency, the minimum structure can be incomplete. Then it is necessary to introduce plausible assumptions (in the form of hypothesis set I) which would allow formulating a problem definition as some approach of the initial task [2]. Let us assume, for example, that in

the basic information structure transfer function $q^g(S | S \times C)$ is not given, but there is set of hypotheses I of it. Then technically it can be assumed that transfer function $q^{(g, \gamma)}(S | S \times X)$ depends on some γ parameter getting values from given set I , but the true value of the parameter is unknown.

It is obvious that it will also demand the choice of the, in a sense, "best" hypothesis of a transitional function. At the same time expanded information structure completeness can be observed only according to final results of the problem research.

5 Game approach to formalization of the choice problem

Assumptions of the choice specify the existence of two aspects of an agent's interests, one of which is determined by the bias in the management of a desirable object evolution, and the other - by the choice of a preference structure. The purposeful control concept defines the third aspect of interests associated with the need for the diagnostics of the situation depending on the observed condition. In compliance with these three aspects three sets of alternatives are to be assigned: set C - action modes, set G - structural alternatives and set X - diagnostic alternatives. It is also assumed to assign the utility function $E\varphi^g(C \times S \times X)$ and the transient function $q^g(S | S \times C)$ of $S \times C$ to S . Assigning these objects suggests the possibility of forming the qualitative mode choice criterion as postulated by the situational control concept that makes sense of expected utility and the purposeful state situation model choice quality criterion that concerns the risk. These criteria are obviously different and in a way interdependent. The natural presumption is that in order to choose structural alternatives the corresponding quality criterion may be introduced. It differs from the rest of the criteria and in some way is dependent on the choice of other alternatives. It is commonly known that in similar conditions the problem of mode choice has a gaming intension [2]. Then each set of alternatives can be formally linked with a party concerned (a player), whose interests are related to the choice of alternatives from the corresponding set of alternatives according to their individual quality criterion. Within a set of alternatives every party has the freedom of choice. Since interests of each party represent a certain component of agent's interests, parties are to comply with the common to them agent's interests, when selecting alternatives. Therefore the problem of mode choice acquires a gaming content in relation to corporate interests [6], and the subject of interest plays the role of a center. He can accept the proposed trade-off alternative if it is hardly possible to improve it without infringing at least one component of interests. The compromise that meets this requirement will be called a "corporate stable equilibrium."

Conclusions

The paper considers a decision-making model for an agent, who can form internal aim and uses subjective representations on a choice situation.

It is shown that the aim of choice is to maximize a specific value of the choice situation by a result. The choice result is determined by agent's representations on the choice situation and on his own interests. When making a decision the agent uses three sets of alternatives: controlling C (mode so faction), structural G and identification X . Therefore, it is possible to assume the existence of three virtual sides, which make a choice of corresponding alternatives that are balanced strategies.

References

- [1]. Baranov V.V. Dynamic Equilibria in Problems of Stochastic Control and Decision Making under Uncertainty. *Journ. of Computer and Systems Sciences Int.* 2002, no. 3, pp. 409–425.
- [2]. V.V. Baranov, "Methods of static equilibriums in the problems of dynamic decision making under uncertainty about the state," / V.V. Baranov // *Bulletin of RAS. Theory and managerial systems*, vol. 5, pp. 45-59, 2001.
- [3]. Lefevr V.A. *Konfliktuyushchie struktury* [Conflicting Structures]. Moscow, Sovetskoe radio Publ., 1973, 158 p.
- [4]. Vinogradov G.P., Kuznetsov V.N. Modeling agent's behavior taking into account subjective representations about a choice situation. *Iskusstvenny intellekt i prinyatie resheny* [Artificial Intelligence and Decision-Making]. 2011, no. 3, pp. 58–72 (in Russ.).
- [5]. Vinogradov G.P., Borisov P.A., Semenov N.A. Integration of Neural Network Algorithms, Nonlinear Dynamics Models, and Fuzzy Logic Methods in Prediction Problems. *Journ. of Computer and Systems Sciences Int.* 2008, no. 1, pp. 72–77.
- [6]. Vinogradov G.P., Shmatov G.P., Borzov D.A. Formation of agent's representations of the domain in a situation of choice. *Programmnye produkty i sistemy* [Software & Systems]. 2015, no. 2 (110), pp. 83–94 (in Russ.).

TRENDS IN DATA ANALYSIS: STATE, DEVELOPMENT PROSPECTS

Doctor in Economic sciences, Prof., I. A. Katsko; PhD in Economic sciences, P. Yu. Velichko;
Student, M. Nikogda
Kuban State Agrarian University – Russia, e-mail: stat@kubsau.ru

On the banks of the Rhine for many centuries was towering beautiful castle. The spiders, which dwelt in the cellars of the castle, tightened all its aisles with cobwebs. Once a strong gust of wind destroyed the thinnest threads of the web, and the spiders began to recover the gaps: they believed that the lock was kept on their web!

M. Klain [10]

Abstract. The article suggests the consideration of data analysis ideology in the context of knowledge creation process and the technological patterns of social development. The problems of singularity (human misunderstanding of data processing results) associated with increase in data variety, volume and further intellectualization of the corresponding technologies for their processing are proposed to be solved by creating new formalization techniques that allow retransmission.

KEY WORDS: DATA ANALYSIS, TECHNOLOGICAL PARADIGM, ANALYTICS, KNOWLEDGE, MACHINE LEARNING, BIG DATA AND THE INTERNET OF THINGS, MEGATRENDS, AMICABLE INTELLIGENCE OF THE HUMAN LEVEL, FORMALIZATION, CONSTRUCT, SCIP

1. Introduction. Data analysis (applied statistics) as development of statistics ideology, probability theory and mathematical statistics intensively developed over the last two centuries is naturally considered in the context of the socio-economic development of society with regard to solution of management and decision-making problems [16-18]. In this case, it seems interesting as a context to lead the ideology of technological paradigm¹, suggested by D.S. Lvov, S.Yu. Glazyev, G.G. Fetisov and essentially relied on larger cycles by N.D. Kondratiev (the phase of new ideas emergence - lasts about 10 years, the phase of paradigm growth - does about 40 years, the maturity phase lasts about 10 years more) [4, 5, 12]. Using the data analysis as an example, it is easy enough to trace the tendencies of the decision-making support ideology that are based on empirical observations characteristic for one or another way (Table 1)

2. Data analysis – contextual approach.

In postindustrial society, the cognitive revolution, which began in the 1950s and 1960s, manifested itself. They can talk about two periods of its development. I cognitive revolution, where a person is a carrier and a generator of knowledge, and a computer, the Internet and software are tools based on the machine learning ideology (the artificial intelligence implementation in a weak version - machine intelligence). II cognitive revolution - new knowledge is generated by the computer (the artificial intelligence formation).

The first scientific paradigms had a material basis, which is very important for the social adaptation of man in the real world. Cognitive paradigm, based on machine intelligence (and in the long term amicable *artificial intelligence of the human level* - AIHL (DIYCH-in Russian), called the fourth industrial revolution, is focused on accelerating all processes by integration at the expense of information technology and the Internet of all things, the basic megatrends of modern society (physical, digital and biological) [1, 19, 22-24].

The purpose of data analysis is to obtain new knowledge about the studied system using observations or differently to convolve (compress) existing information for solving applied problems of

analysis and explaining the features of the studied system functioning, management, forecasting (predication) and decision-making.

The main difference between applied statistics (data analysis) from mathematical statistics is the consideration of not only probabilistic but also geometric and logical nature of data, as well as the obtaining of convolutions by both formal algorithmic methods (classical methods of *multidimensional statistical analysis* - MSA) and not formal ones (machine learning and adapted methods of MSA).

According to E. Toffler's studies, nowadays, power in society is based on three basic elements: strength, money and knowledge. [21] Moreover, knowledge becomes a universal tool that can replace all others. This is why Russell Ackoff's definition, which characterizes the process of the knowledge formation, acquires a special meaning, which in our formulation is expressed in the following way [2]

Facts – Information – Data – Knowledge – Understanding – Wisdom.

One of the forms of knowledge representation contributing to thinking formation and worldview of human has always been mathematics, which allowed to form a chain of thinking levels (*recognition - reproduction of model situations - atypical situations analysis - creativity*). The most important stage in the process of knowledge formation is understanding - a person easily perceives and uses in practice what is understandable. From the point of view of data analysis, the stages of the knowledge formation can be disclosed in the following way [2, 6]:

- *facts – events, that have already happened;*
- *information – facts characteristic;*
- *data – facts, described quantitatively or qualitatively, presented in the form of tables «object – property (feature)» or «question - answer»;*
- *knowledge – rules «If ..., so ...», which can be used in decision-making;*
- *understanding – presentation about functional features of studied object, managing possibilities, foresight (predication) and decision-making;*
- *wisdom – ability to use the reached understanding in future.*

Table 1 – Data analysis in context of socio-economical development

Technological way	Main formalization form (approach)	Data processing way
I mechanization 1770-1830	Mathematical analysis (data – realization results of mathematical laws of natural sciences)	Descriptive statistics, differential and integral calculus
II steam engines, railways 1830-1880		
III electricity, metallurgy 1880-1930		
IV oil, mass production, nuclear power 1930-1980	Probability theory and mathematical statistics (data – random processes realization, which submit to certain distribution laws – parametrical statistics, or nonparametrical statistics)	Selective method, Convolution of information. Formalistic algorithmic approaches, solving problems: Data description, visualisation, classifications and dimension decrease, search of dependences (multidimensional statistical analysis)
V informatization, telecommunication 1980-2020	Data analysis (applied statistics) (any data nature; probabilistic, geometric, - data form in multidimensional attribute space «compact» (clots), logical – this not only quantitative, but also non-numerical (qualitative) form patterns – interrelations not always explainable at the quantitative level)	Analytics 1.0 (descriptive analytics) OLAP cubes, convolution procedures that do not allow an algorithmic approach - Exploratory data analysis (EDA), also based on the computer training ideology (Data Mining) as an option for implementing EDA based on information technology, web-sites scraping
VI nano-, bio-, info-, cognitive-, socio-technology – NBICS 2020-2060	<i>Big Data, Internet of Things (IoT)</i> (data nature is any of the above, including visual, textual, sound, video-audiofiles and others)	Analytics 2.0 (predictive analytics) Analytics 3.0 (prescriptive analytics - the basis of CRM) Analytics N.0 (analitics, supporting typical solutions, having opportunities to search and process information on-line/interface (analogue of modern app Siri from Apple))
VII human – main technology subject 2060-2100	... Data nature is any of the above, including not representable in semiotics systems projection foresight	Analytics NBICS, based on technologies of amicable intelligence of human level

In order to use methods of applied statistics, the data must be measured with qualitative or quantitative scales. The measurement process is accompanied by problems: heterogeneity, quality,

limitations, subjectivity of perception and thinking. Moreover, person is limited in perception of the surrounding world. As it is known, it is characterized by the number of J. Miller (1956) (7 ± 2) - according to it, there is a need to compress large volumes of information and representation in the form of (preferably understandable) models. This goal is devoted to the work of decision support systems that allow solving tasks: descriptive statistics (OLAP cubes), classification and diminution of dimensions, search for dependencies, prediction, etc. implemented in KDD class systems and Data Mining.

The implementation of machine learning methods does not allow us to realize the "understanding" stage in R.Akoff's knowledge formation process and shows the practical application of Braimean's uncertainty principle, which is an analogue of Heisenberg's uncertainty principle in data analysis context:

$$\begin{aligned} &\llbracket \text{accuracy} \times \text{interpretability} = \\ &= \text{Braimean's constant} \rrbracket. \end{aligned}$$

The famous futurist E. Toffler talks about three waves in the development of society: agrarian, industrial, informational, which their traditional education systems conformed to [19]. For several decades, we have witnessed the transformation of education traditional system for an industrial society. Data analysis has always evolved in the direction of meeting the needs of society. If there was enough descriptive statistics in the agrarian society, in industrial - analytical statistics, the postindustrial society and the expected information extended the applied statistics with machine learning methods using both structured and unstructured data (Data Mining, Text Mining, Web Mining, Social Mining, Big Data and the Internet of things), thus, the demand for data analysis methods is determined by the level of development of the society, for example, business intelligence technologies that have long been used in Moscow and St. Petersburg, are developing in the regions only in recent years, becoming one of the costly business articles (Table 1). The development of the digital sector of the economy is being discussed in the world, therefore, it becomes necessary to systematically comprehend the possibilities, limitations and prospects for data analysis at the present stage of society development.

3. Megatrends. Today's data processing methods are based on the ideology of probability, statistics, data analysis and, among other tasks, allow solving the problem of finding "megatrends" (conditional coordinate system) at different levels of society at a new level that allows explaining many phenomena in the socio-economic space. The number of "megatrends" for a society is the same as for a person (7 ± 2) and corresponds to one of the theories claiming the theory of "great unification" in physics - superstring theory, which requires about 10 measurements. At present, one of the obvious "megatrends" explaining the transformation of the education system is that "the transition to an information society requires getting rid of the outdated educational system that trained cadres for the industrial society." Analysis of printed sources allows us today to talk about the following "megatrends" of our society, conditioned by the digital revolution and carrying a disruptive influence (a form of natural selection in biology), "tearing" the homogeneous aggregate (usually) into two extreme variants and not contributing to the average state, which explains the appearance in the socio-economic systems of power law distribution laws, such as the Pareto or Zipf law (for example, the result of such an impact can be considered the stratification of society: rich and poor, "golden" Billion and others, etc.) [1, 2, 8, 9, 19-24] :

- 1) Mankind today lives in a consumer society and gradually loses the distinction between the real and the virtual.
- 2) Priority in life is obtained by "physical" people, without moral and moral obligations.
- 3) The society increasingly depends on information technology (3D-printing, development of 4D-printing

technology, the production of any services and goods in online mode).

- 4) Information technology is becoming one of the subjects of everyday life of modern man (unmanned vehicles, robotics, new materials).
- 5) People pay and receive money for virtual actions that do not give a sense of physical incarnation, which negatively affects the human psyche.
- 6) New approaches to interaction and cooperation at all levels of the society are being developed. For example, "distributed databases" are *block chains* that represent a data store that is available for verification to anyone (for example, *Bitcoin*).
- 7) Particular attention is paid to *Big Data* concepts and the "*Internet of all things*" in the study of social networks, industry, business.
- 8) The growing opportunities for biological engineering require the development of a normative, ethical and legal framework.
- 9) The transition to an information society requires transforming an outdated education system that trained personnel for an industrial society by selecting a goal at the state level (for example, harmonious development of a person) instead of replenishing a certain labor market (or "human cloud").
- 10) The actual sector of the economy, based on the "human cloud", is becoming relevant.

Modern information technologies for data analysis (*web mining* and *text mining*, etc.) make it possible to find an alternative to classical content analysis when searching for "megatrends" including without human participation (scraping *web-sites*).

"*Megatrends*" change over time and the past ("megatrends") "twists" along the new ones, relying on the general direction (mainstream) of the 21st century - a digital revolution that, like the communist movement 100 years ago, will change the landscape of the planet. Perhaps right now there is a tectonic gap between the "golden billion" and the rest of the world's population, although the maturing changes are not universal even for developed countries. For example, the politics of the consumer society is unacceptable in the Arab world and can have a different form (as in India and China). In Russia there is a centro-peripheral model of socio-economic space (N. Zubarevich) [8]:

- post industrial Russia (federal cities with a million population with postindustrial economy),
- industrial Russia (industrial cities with a population of up to 250 thousand people),
- rural periphery
- agrarian Russia (the main part of the country and residents of settlements with a population of less than 20 thousand people) ,
- patriarchal republics, based on their own values (the North Caucasus, Southern Siberia).

And all four of Russia perceive different future changes and react differently.

At different levels of the hierarchy of society, there are their own tendencies to change the world, so at the state level (in most countries) first of all one can distinguish: bitcoin, crypto-currencies, cyberwar, fakes.

4. The problem of translation of knowledge. Current trends in the development of analytics are directly related to the achievements of human intellect, which cause a futurist (fear of the future). Mankind is preoccupied with its own ideas about the power of computers and the alleged consequences of the emergence of artificial intelligence (AI). Many scientists expect the emergence of the point of "singularity" - the moment when the possibilities and results of the activity of artificial intelligence systems in the narrow sense (understood as the realization of the ideology of machine learning) will surpass the possibilities of human understanding [1,

9]. In addition, artificial intelligence is expected to reach the human level. The only thing people hope for is that they will be a *friendly human intellect (AIHL)*.

Intelligence of information systems is achieved today through the use of the Internet (the Internet of things), the ideology of machine learning, and computational capabilities. Thus, we are not talking about the presence of consciousness, but it is possible that the opportunity will come of "creating the effect of consciousness" through the effect of replacing the computational abilities. Then only the initial education of unconditional friendliness to man will pass the point of singularity.

In fact, we are talking about machine intelligence, which, due to the processing power, can generate new knowledge in the form of patterns and (or) constructs that can not be explained by humans on the basis of available information (including using *Big Data* and the Internet of all things). The traditional sequence of the process of forming knowledge according to R. Akoff (Facts - Information - Data - Knowledge - Understanding - Wisdom) will be broken. For if traditionally for a person of "knowledge" are products (rules) "if ..., then ..." that allow to realize the stage of "understanding", then the question arises about the need for a new round of development of mathematics and information technologies oriented to "not ... exclusion" of man from the processes of management and decision-making in the socio-economic space, since the knowledge obtained by the *AIHL* can go beyond the boundaries of human understanding (the rules" if ..., then ... "). Data analysis today is a way of compressing large volumes of information to support decision-making processes, which allows to identify patterns in data and to present them in the form of: graphs, tables, formulas, various dependencies obtained using machine learning methods. It is assumed that at the point of singularity, knowledge will go beyond these limits.

The problem of understanding the intellectual systems of the future is similar to the problems of the middle of the last century, when the possibilities of interaction with computers were formed through programming languages, which today number more than 8000.

We believe that in order to solve the problem of "retransmission" of knowledge to a person, the potential capabilities of the *AIHL* must presuppose the possibility of synthesizing a number of subject areas (SA) into which new knowledge can be projected. The lexicographic ordering of the SA will allow to identify and rank the consequences of applying new knowledge in different areas.

Thus, it becomes urgent to develop an understandable ideology of the description of the SA. As a possible example, consider, following the work of L.S. Bolotova [3], the method of situational analysis and design of the design of the SA model, on the basis of the set-theoretical (relational) approach, in the form of a complex of invariant constructs as applied to the description of the SA for the new knowledge.

4.1. The domain model. The synthesized object (system) should be created under the condition of the existence of an external environment that is characterized by a certain subject area (SA) - a part of the real world within the given context (industrial, agricultural, financial, computer, etc. corresponding to the direction of knowledge). Each subject area has its own language, which can be formalized using binary relations [3]. Usually, a system is understood as the set of a related set of objects. Most often, two sides of connectivity are considered: as a fact of the existence of a relationship between individual elements of the system - realizing the cognitive conceptual aspect (cognitive maps); as a description of the process of the corresponding connectivity of elements - a functional, information or behavioural aspect (semantic networks, frames, products, methods of situational modelling). Both approaches are considered rather rarely.

Modelling SA of an arbitrary nature is connected, first of all, with the analysis of the categories describing it. A category is understood as a construct or otherwise some abstract container, with some objects entering into it, and others not (this is the postulate of

our thinking for more than two millennia). It is assumed that the categories that a person operates on can be arranged in the following hierarchy: the higher level - the base level - the lower level. The base level is the level at which most of our knowledge is structured. It gives an opportunity to perceive geometric visualization of the conceptual structure of an object.

4.2. Algebraic models of SA. At present, algebraic language and style of thinking are the standard approach to the representation of data and knowledge in information systems. The question of the possibility of a correct description of the model of the subject domain associated with the person (observer) in the form of a system of objects with certain relations can be investigated only by means external to this system (that is, in some other theory), as follows from K. Godel's theorem on incompleteness of formal arithmetic (to which almost all mathematical theories can be reduced).

D. Hilbert describes the interpretation of the formalization of mathematical theory, as well as the method that makes the formal system the subject of the study of mathematical discipline - metamathematics or the theory of evidence [11].

The introduction of a system of S objects according to the ideology of metamathematics, assuming the existence of a non-empty set of objects between which certain relations are established, can proceed from two methods characterizing two main trends in modern mathematics - constructivism (modern predecessor of which is A. Poincaré and which was basic in ancient science, for example, in Euclid) and formalism, which implies a complete abstraction from the meaning.

In an axiomatic method, the axioms underlying the formal approach are used as assumptions about the system of S objects. Then we examine the consequences of the axioms, which form a theory with respect to the system of S objects under consideration.

A constructive (genetic) method involves constructing objects in a certain order. S. Klini characterized this approach as a method of substantive or material axiomatics. To describe the system from the point of view of the observer observing the system from the outside, a "formal system object" is introduced - the meta-set of the research system S , which allows describing systems based on logical and mathematical methods [13].

The presence of experts with exact, technological, effective thinking, free "from traditions and cognitive prejudices", which will interact with *AIHL*, is postulated. To do this, special (psychological) work with experts and subjects of the problem of creating a new object is supposed, accustoming them to operate with their "constructs".

The basic representation of the construct is the meta-set

$$S < X_{as}, X_a, X_{ao}, \cup X_{ac_i} >,$$

where X_{as} - (action subject), X_a - (action), X_{ao} - (action object), $\cup X_{ac_i}$ - (action components). All elements of meta-set have (property):

$X_{as} = X_{as}(p_{s1}, \dots, p_{sl})$, $X_a = X_a(p_{a1}, \dots, p_{aq})$, $X_{ao} = X_{ao}(p_{o1}, \dots, p_{ot})$, $X_{ac_i} = X_{ac_i}(p_{ac_i1}, \dots, p_{ac_ih})$, the results of relations of which among themselves within the framework of our task realize TER (technological, technical, operational, economic, environmental requirements for the subject area, which are based on the normative provisions defined by the person).

A logical representation of a construct implies two components:

1) functional - Φ , regarded to the purposes of building a new object and described as a union of binary relations (R)

$$\Phi = R_{as}(X_{as}, X_a) \cup R_{ao}(X_a, X_{ao});$$

2) providing, (achievement of the goal) - Q ,

$$Q = R_{as}(X_{as}, X_a) \cup R_{ac_i}(X_a, X_{ac_i}).$$

$$K = \Phi \cup Q.$$

Each construct K is a kind of domain concept that is open to expansion and modification, which is intended for reusable use in designing, obtaining production rules, and so on.

Combining all constructs:

$$U = \cup K_j$$

gives us the universum U , a structure called a polyhedron in topology, describing an SA of arbitrary nature.

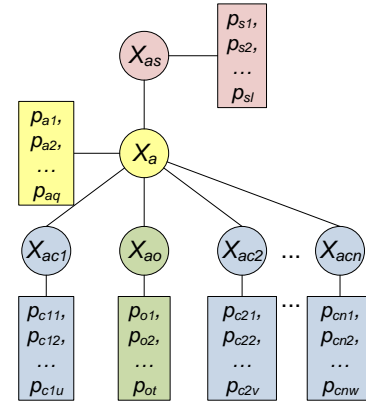


Figure 1 – Construct model – Subject area extract [3]

The universum is a generalized model of a specific subject area, which can be represented as a basis for concepts (*ontologies*) designed for reusable, multipurpose use in various applications and relationships between them that implement production rules [14]. Consideration of ontologies with selection functions and mechanisms for their implementation allows us to talk about a knowledge base that potentially allows the formation of products (rules) that are understandable to man [15].

5. Conclusions. The problems of human interaction and the human-level friendly intellect require the creation of means for "retransmitting" to a person new knowledge obtained by advanced methods of data analysis (structured, weakly structured, unstructured). At the author's level of vision, developments are required that allow translating, obtained knowledge (patterns, constructs, etc.) into an "understandable" kind of person, for example, projecting into several mutually complementary subject areas, which can be realized (*AIHL*) described by the method of situational analysis and design of the design of the SA model. Within the framework of the domain model - the creation of a new object (or description of the situation) is reduced to operations over the relations between constructs and their elements. Thus, one of the variants of "passing the point of singularity" is the formation of knowledge bases related to a certain subject area and an assessment of the consequences of the implementation of new knowledge through the use of scenarios based on the SA model.

The emergence of a new scientific paradigm in science and society forms a new space in which the previous includes (due to) a new generalization, the postulates (principles) change and shrink, with an increase in the coverage of the phenomena described (the scaling effect of socio-economic space) [7], which is realized in an explicit form using the example of data analysis. The further development of intelligent information systems leads to the automation of the work of analysts and other professionals, the emergence of new knowledge and the ability to present them for adequate human perception, and the need to create ethics councils, to address employment and social issues is already generally recognized. To preserve one's identity in the future information society, a person needs to solve many problems: preserve human culture, universal values, find for himself and implement new means of formalization (adapt or develop new mathematics for translation of knowledge obtained by the *AIHL*), etc. Everyone should understand that it is he who builds the future world and how he will solve it.

6. Literature

1. J. Barrat, The last invention of mankind: Artificial intellect and the end of the era of *Homo sapiens*. /from English. Natalia Lisova. - M.: Alpina non-fiction, 2015. - 304 p.
2. P.S. Bondarenko Theory of Probability and Mathematical Statistics: Textbook / P.S. Bondarenko, G.V. Gorelova, I.A. Katsko; Ed. by I.A. Katsko, A.I. Trubilina. Moscow: KNORUS, 2017. - 390 p.

3. L.S. Bolotova. Conceptual design of the domain model with the help of software systems for the development of knowledge bases for intelligent decision support systems / L.S. Bolotova, V.A. Smolyaninova, S.S. Smirnov // High technology: scientific. -technical. - 2009. - T. 10 No. 8. - P. 28-36.
4. S. Yu. Glazyev, D.S. Lvov, G.G. Fetisov. Evolution of technical and economic systems: the possibilities and boundaries of centralized regulation. - M.: Science. - 1992. - 208 p.
5. S.Yu. Glazyev. Strategy of advanced development of Russia in the conditions of global crisis. - M.: Economics, 2010. - 255 p.
6. N.G. Zagoruyko. Applied methods of data analysis and knowledge. - Novosibirsk, Uni of maths, 1999. - 270 p.
7. V. A. Zorich Mathematical analysis of natural science tasks - M.: MCNMO, 2017. - 160 p.
8. N. Zubarevich. «Four Russia-s» and new political reality. Web: http://polit.ru/article/2016/01/17/four_russians/
9. M. Kaku, Mind future / from eng. by N. Lisova, - M.: Alpina non-fiction, 2015. - 502 p.
10. M. Kline. Mathematics. Loss of certainty. - The World, 1984. - 134 p.
11. S. Klini. Introduction to metamathematics: - M.: Librocom, 2009. - 528 p.
12. N.D. Kondratiev, Large cycles of conjuncture and theory of foresight / edited by AL Albakina. - M.: "Publishing house" Economics ", 2002. - 767 p.
13. V.V. Kulba etc. Information processes and information management / Human factor in management c6. Articles of the ISP RAS ed. O.N. Abramova, K.S. Ginsberg, D.A. Novikov and others - M.: KomKniga, 2006. - 496 p.
14. Y. Lyapar. Theory of System-Structural Design - the Basis of Intellectualization of Modelling and Decision Support Systems / Yu.I. Lyapar, I.A. Katsko, G.F. Bershitskaya - Krasnodar.: KSAU, 2010. - 49 p.
15. Yu. I. Lyapar. Synthesis of knowledge bases of analog electronic devices. // Ulyanovsk: Works of the Intern. Conf. "Continual and Algebraic Logic, Calculus and Neuroinformatics in Science and Technology", vol. 3, 2006, p.120-128.
16. A.V.Maltseva, N.E. Shilkina, O.V. Mahnitkina Data mining sociology: Experience and outlook for research // Social surveys 2016-January(3), p. 35-44.
17. D.A. Novikov. Cybernetics: Navigator. History of cybernetics, current state, development prospects. - M.: LENAND, 2016. - 160 p. (Smart Management Series)
18. A.I. Orlov. The main features of the new paradigm of mathematical statistics // Polytematic network electronic scientific journal of the Kuban State Agrarian University (Kuban State University, Krasnodar, KubSAU, 2013. № 06 (090)., P. 188-214 <http://ej.kubagro.ru/>
19. D. Rose, The Future of Things: How a Fairy Tale and Fantasy Become a Reality / 3rd edition / translation from the English Seeds of Sheshenin. - M.: Alpina non-fiction, 2017. - 344 pp.
20. E. Toffler. The Third Wave - M.: AST, 2010. - 784 pp.
21. E. Toffler. Metamorphoses of Power - M.: AST, 2004. - 672 pp.
22. M. Ford. Technologies, who will change the world. Alexandra Kardash, M.: Mann, Ivanov and Ferber, 2014. - 268 p.
23. M. Ford. Robots come: Technology development and the future without work / Translated from English by Sergey Chernin. - M.: Alpina non-fiction, 2016. - 430 p.
24. K. Schwab. The Fourth Industrial Revolution / K. Schwab - M.: "E", 2017. - 208 p.

INTEGRAL ASSESSMENT OF ENVIRONMENTAL QUALITY AND THE QUALITY OF LIFE OF THE POPULATION OF THE ARCTIC REGIONS OF RUSSIA IN THE PERIOD FROM 2003 TO 2015

Prof. Dr. V.V. Dmitriev
Institute of Earth Sciences – Saint Petersburg State University, Saint Petersburg, Russia

E-mail: vasily-dmitriev@rambler.ru

Abstract: *The integral assessment of environmental quality and quality of life of the population of 9 regions of the Arctic zone of the Russian Federation in the period from 2003 to 2015 is considered. To build integrated indicators, we used: summary indicators, randomized summary indicators, and "ASPID methodology" (analysis and synthesis of indicators in the information deficit). When calculating weights, incomplete, inaccurate, non-numeric information was taken into account) about the criteria and priorities of the evaluation. To assess the quality of the environment, 8 parameters were used. To assess the quality of life of the population, the state of three subsystems was taken into account: ecological (8 parameters), economic (5 parameters), social (5 parameters). The choice of criteria was made taking into account the information available on the website of the Federal Service of State Statistics of the Russian Federation, in the collections "Regions of Russia" and in state reports "On the state of the environment ...". To assess the quality of the environment in the regions and the quality of life of the population of the regions, five quality classes were introduced (I - high, II - above average, III - average, IV - below average, V - low). In constructing integral indicators, the sum of the normalized values of the indicators within subsystems (blocks) and between them was used as a synthesizing function, taking into account the equilibrium or nonequilibrium setting of priorities. When assessing the quality of the environment, all regions fall into the third class (middle - the right border of the class) with a slight temporal change. In evaluating the quality of life, three groups of regions were identified. In the first group, the quality of life for the period under review improved by 10 percent or more. The second group includes regions with an improvement in the quality of life by 5-10%, the third group includes regions with an improvement in the quality of life up to 5%. For the same time interval, the quality of life of the APR regions was compared with the regions of Central Russia (Tver Region). The forecast scenarios of a possible change in the quality of the environment and the quality of life of the population in the regions are considered. The studies were carried out with the support of the RFBR grant No. 16-05-00715-a.*

Keywords: INTEGRATED ASSESSMENT, QUALITY OF ENVIRONMENT, QUALITY OF LIFE, ARCTIC REGIONS

1. Introduction

The relevancy of the work is accounted for by the necessity to develop the theory and practice of the evaluation of the state of complex systems in nature and society, their non-additive (emergent) properties, and the system simulation of natural and social transformation of eco-, geo-, and socio-systems. The recent recommendations in the sphere of global, regional, economic, and social development are specified in the works of the commission of Stiglitz – Sen – Fitoussi (see www.stiglitz-sen-fitoussi.fr) [1]. The Commission is more known as the Stiglitz Commission, was set up in February 2008 at the initiative of the President of the French Republic under the guidance of the Nobel prize winner in economics Joseph Stiglitz, with participation of the Nobel prize winner Amartya Sen. The commission was authorized to identify the suitability of using the existing national indicators of development and progress, including such indicator as gross domestic product (GDP). It was necessary to validate the economic development and social progress parameters, to study what additional information could be required to form a more adequate picture, to discuss how to present this information correctly, and to check the feasibility of the suggested tools of measurement. The Commission submitted its first report on 14 September 2009 and called upon national statistics authorities under the aegis of international ones to focus their efforts on the development of new indicators of social progress for more adequate assessment of quality of people's life in countries and regions. To arrange its activities, the Commission was broken into three working groups which studied respectively the traditional GDP evaluation issues, life quality and sustainability issues. The working groups submitted recommendations for each of these spheres [1], which have come to be known as 12 recommendations for the fundamental amendment of the state statistics basics in France and the entire world.

The main conclusions of the report E/CN.3/2011/1 of the UN National Institute of Statistics and Economic Studies in 2011 (section B. Basic conclusions of the report) state that "well-being includes both economic resources such as income, and non-economic aspects of people's life (what they do and what they can do, how they feel, in what natural environment they live)". The

sustainability of these levels of well-being depends on our ability to pass on to future generations the accumulated assets which are significant for our life (natural, physical, human, social). Therefore it is important to discriminate between evaluation of the current well-being and evaluation of its sustainability in time" [1, p.2].

It is in this connection that the resolutions can be mentioned which were passed by the UN General Assembly on 25 September 2015 "Transforming our world: Agenda of sustainable development for the period until 2030" and "Report of the interdepartmental group of experts for indicators of objective fulfillment in the sphere of sustainable development" at the UN forty-seventh session 8-11 March 2016 [2,3].

The specific feature of the modern stage is not only the validation of the representative criteria or groups of criteria for the evaluation of the state of natural and socio-ecological-economic systems, but also the development of models of analysis and synthesis of indicators taking into account the use of incomplete, inaccurate, non-numerical information on the evaluation criteria and priorities [4].

The article discusses evaluation of the state of socio-ecological-economic systems (SEES) and quality of life of people of the RF regions. The state of SEES is believed to be the characteristic of the system at a certain moment of time. The focus is on the comparative evaluation of the quality of life in the regions of the Arctic zone of the RF (AZR) from 2003 to 2015.

The key point of our publications on the integral evaluation of the state of eco-, geo-, socio-systems and their emergent properties [6-10] is the following conclusion: in the multi-criteria evaluation of the state of the systems with an indicator approach the incomparability of the obtained assessments is revealed when according to one criterion (indicator) or a group of criteria the system is referred to one class, and according to another (others) it is referred to another (other) class (-es). Thus, with the indicator evaluations of SEES states uncertainties arise in the treatment of the obtained results. The authors have to write on what number of criteria a system can be referred to each of the classes, and more frequently, without introducing classes, the results are just ranked for each of the indicators, determining the place of SEES in question in the list of similar national or regional systems. In the

same way the SEES state is evaluated by each of the indicators recommended by them or national statistics authorities. The indicators are not generalized, and if they are, it is on the additive (grade) basis, without taking into account the priority of indicators or their trustworthiness. Then, it is often noted that the objective information of national statistics authorities is unsuitable for evaluation, because it does not contain a number of indicators already tested abroad or recommended recently. At that, the indicators taking into account the perception by people of their position in society, i.e. subjective indicators, are often used as the basic ones. We noted that in order to take such indicators into account, it is logical to use non-numerical (ordinal), inaccurate (interval), incomplete information (so-called *nnn-information*) which is necessary to take into account both for specifying the indicators and for determining the evaluation priorities. It was recommended to use multi-criteria and multi-level evaluation accounting for simulation of evaluation priorities inside levels (subsystems) and between them on the basis of “nnn-information”. The levels can include groups of criteria based on the data of national statistics authorities, as well as subjective data obtained in the statistical polls of the people. The following is recommended as the methods: a method of consolidated indicators (MCI), a method of randomized consolidated indicators (MRCI) and its modern version named by the author “ASPID methodology”, the methodology of analysis and synthesis of indicators in the information deficit [4]. In all cases a possible change in the priorities for evaluation inside the groups and between them is taken into account.

2. Experimental procedure

The modern foreign level of research is characterized by the developed methods of analysis of target indicators used to characterize the state of complex systems (mostly economic or socio-systems and their subsystems) and, to a less degree, their emergent properties. The present time is characterized by accumulation of methodological and practical experience in the research of the state of complex systems in nature and society and their separate subsystems. The method of making up a “web diagram” (“rose diagram” in Russian publications) is often used which units in a single picture the information on a great number of indicators [5]. As a result, there is an analog of the natural Hutchinson niche or socio-ecological-economic niche visually characterizing the aggregate of the conditions of existence of the system. In case of transformation system its visual image, a niche, is also transformed and its GIS-image reflects the result of impact on the system (its area (volume) is changed, the system acquires a new, predominant vector of development which is revealed and visualized graphically).

The other approach [5] uses the method of building a composite indicator that is a union of the aggregate of the used parameters into a single composite indicator (composite sustainability indicator - CSI) in which the parameters are taken into account with their weights reflecting the priority of each of them (Fig.1).

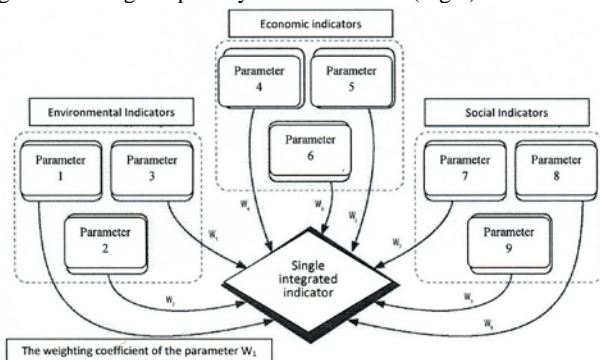


Fig.1 The structure of the integrated stability indicator (by Noam Lior, 2015)

Thus, the building of an integral indicator is implemented at one level of convolution of indicators as a sum or product of the characteristics taken with their weights. At that, the author does not consider peripheral issues of the rating of indicators taking into account the type of connection and its non-linearity; simulation of priorities (weighting factors) of the characteristics; creation of the effect of hierarchy on account of introduction of multi-level convolutions; investigation with the use of integral indicators of emergent properties of the systems (stability, well-being etc.).

In our case the building of classification models containing several levels of investigation and convolutions of indicators (Fig.2) is implemented. This figure represents one of the models that we used to evaluate the quality of life in the RF regions [6,7]. These works and Fig.2 give the units, the parameters of each unit and the results of evaluation of the quality of the environment and the quality of life of people of the arctic regions of the Russian Federation. All indicators were sampled from the data of the Rosstat website (“Regions of Russia” collections) for the period from 2003 to 2015. The basic objective of the investigations was to perform a convolution of indicators at the first and second levels and the identification of the situations in which SEES cannot retain its properties and mode parameters with a certain hypothetic influence on it in separate subsystems and the system as a whole [6].

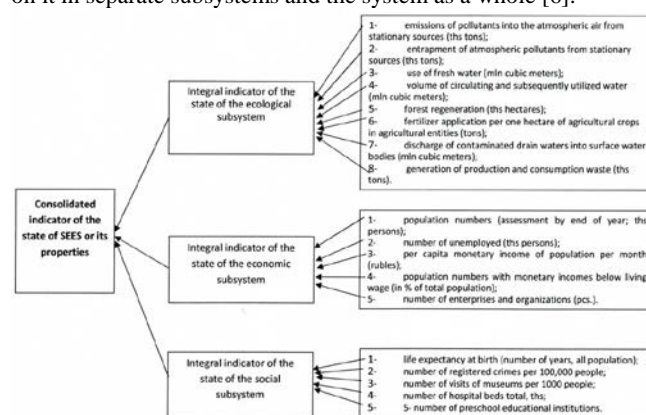


Fig. 2 Building a model for integral assessment of the SEES state, quality of the environment, and quality of life of the region's population[6].

The integral indicator Q_i was constructed so that it depended not only on the rated values of the initial characteristics q_i , but also on their priorities determined by the weights p_i , the sum of which should equal 1.0 ($0 \leq p_i \leq 1$). As an expression for the integral indicator, the linear convolution was used in the form of: Q_i

$$= \sum_{i=1}^n q_i p_i, n \text{ is the number of the evaluation criteria. The state of}$$

the system and the quality of life of people of the region were evaluated for 5 classes (I – high; II – above average; III – average; IV – below average; V – low). The proximity of the integral indicator to 0.0 evidenced high quality of life of people, the proximity to 1.0 evidenced low quality.

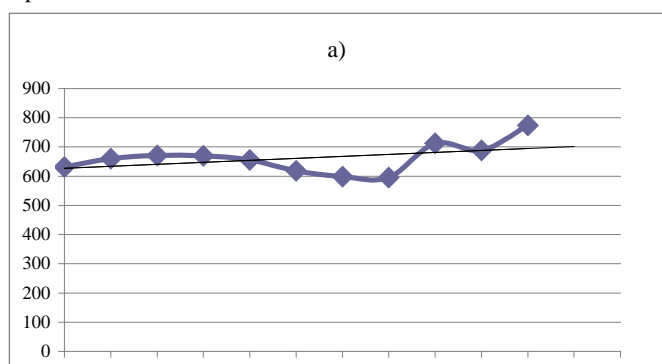
In [6,7] we investigated a change in the quality of life in 9 regions of the arctic zone of the Russian Federation (AZR) for 10-12 years, and the results of the experiments in a hypothetic change of the situations in each of the units and in all units simultaneously were described. As a result, the integral indicators of the quality of life of people for 8 scenarios for the first (inside the subsystems) and the second (between the subsystems) levels of the convolution of the indicators was calculated. The results of the evaluation of the quality of the environment and the quality of life were compared for 2003 and 2013. In these models the liner rating functions were used in the rating of the indicators with the equal weight of the evaluation parameters inside the three subsystems (ecological, economic, social) and between them. In the experiments with loads the results of the options with 30% and two-time deterioration of

the situation inside and between the units against the background of 2013 were described.

In general, the subsystem of social conditions was found to be the most sensitive subsystem. The maximum increase in the effect of impact of both on separate subsystems and on the socio-ecological-economic system as a whole (summary evaluation). With

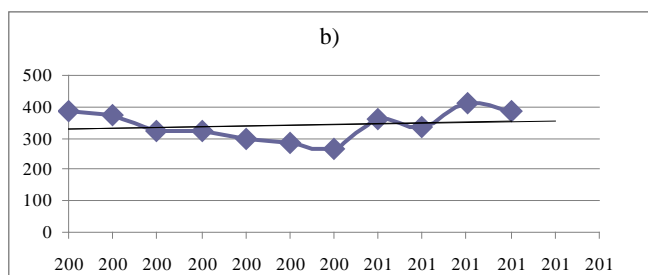
3. Results and discussion

In [7] the main drawback of the experiments with a hypothetical change in the situations in the regions was noted. It consisted in the fact that under real conditions it is not logical to expect a simultaneous change in the load by 30%, 50%, two times, etc. inside one of the units or in all the subsystems simultaneously. Each parameter chosen as a representative criterion will have its rates and direction of the changes. The situation is complicated by the different rates and direction of such changes that may be noted in different regions. Therefore, in the next stage of the investigations it was necessary to study the temporal change of each of 18 criteria and obtain the trends of these changes for the regions. Fig.3 (a,b,c,d) shows the examples of such changes for four parameters of the environmental quality subsystem (ecological unit) for the Republic of Komi for 2003-2013.

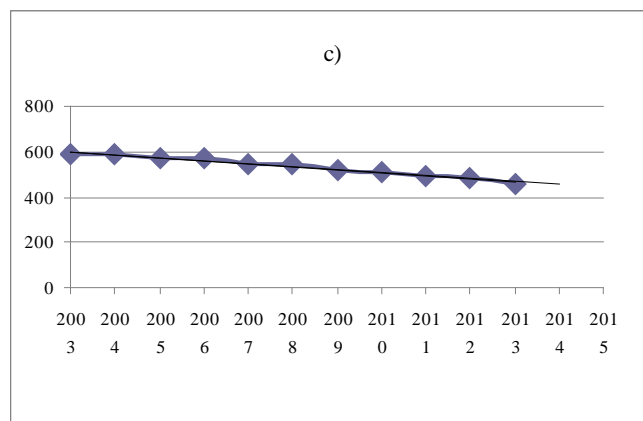


a) Contaminants emitted into atmosphere from stationary sources (ths tons/year)

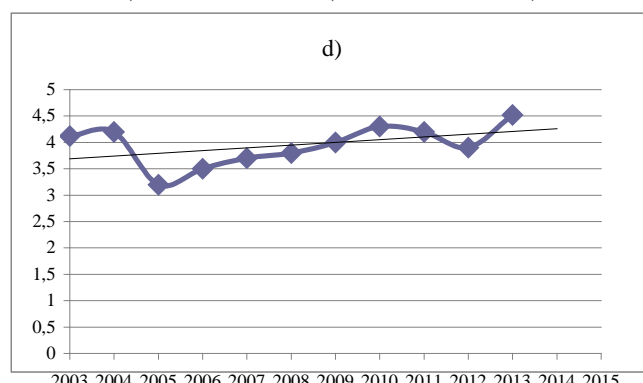
Fig.3 Tendency of changes in certain parameters of the evaluation of the quality of the environment for the Republic of Komi for 2003-2013.



b) Entrapment of atmospheric contaminants emitted by stationary sources (ths tons/year)



c) Use of fresh water (million cubic meters)



d) Fertilizer application per one hectare of sowed crops in agricultural crops in agricultural entities (tons)

Analysis of the trends of change of separate characteristics, as was expected, showed that the rates and the direction of their change are different. Thus, for the ecological unit for the region of the Republic of Komi for 5 parameters out of 8 in 2014 a decrease in the indicators in percent compared to 2013 (indicator No. is given in Fig.2) was expected by: 1-9,5; 2-8,3; 4-0,5; 5-5,5; 6-6,0. For the rest of the parameters an increase in the indicators compared to 2013 was obtained by: 3-3,1; 7-1,9; 8-54,9. Thus, the maximum increase in respect of 2013 is noted for parameter No.8, generation of production and consumption waste, by 54.9%, the minimum decrease is noted for parameter No.4, the volume of the recycled and successively utilized water, by 0.5%.

The forecast change of the integral indicator of the ecological unit for 2014 with the quality of the weights yielded the value of the integral indicator $Q_I=0,49$, which allowed the quality of the environment to be referred to quality class III with the width of the class interval 0.37-0.56. The statistical reporting data (Rosstat website, "Regions of Russia" collections) for 2014 and 2015 confirmed the forecast calculations of the integral indicators for these years. The integral indicators of the quality of life of people in the AZR regions for the period from 2003 to 2015 (the second level of the convolution of the indicators) are given in Table 1. The scale of an integral indicator of the second level of convolution for equal priorities at the first and second levels: I – high (0-0.16); II – above average (0.16-0.36); III – average (0.36-0.56); IV – below average (0.56-0.79); V – low (0.79-1).

As a result, we note that in 8 regions there is a tendency for improvement of the quality of life of people. In Murmask region, the Republic of Komi, Khanty-Mansiysk Autonomous Okrug Yugra, Republic of Sakha (Yakutia) there is an improvement of the

quality of life of people by 7-10%. In Arkhangelsk region, Nenets Autonomous Okrug, Chukotka Autonomous Okrug, Yamalo-Nenets Autonomous Okrug the improvement of the quality of life of people is by 10-12%. The unchanged is the quality of life in Taymyr Dolgano Nenets Autonomous Okrug in the period from 2003 to 2005 (Table 1).

To compare the quality of life of the Arctic regions of the RF with the regions of the central part of the Russian Federation, the

quality of life of people in Tver region for 2003 and 2013 was examined. By the value of the integral indicator the quality of life of people in Tver region from 2003 (0.64) to 2013 (0.57) was improved by 11%. This indicates the close rates of changes in the quality of life in the regions being compared.

Table 1: Integral indicators of quality of life of people in the regions of the arctic zone of the RF for the period from 2003 to 2015 (the second level of the convolution of indicators).

Region / Year	2003	2005	2010	2013	2015
Arkhangelsk region	0,65 (IV)	0,64 (IV)	0,61 (IV)	0,58 (III - IV)	0,57 (III - IV)
Murmansk region	0,65 (IV)	0,65 (IV)	0,61 (IV)	0,58 (III - IV)	0,60 (IV)
Nenets Autonomous Okrug	0,66 (IV)	0,62 (IV)	0,59 (IV)	0,63 (IV)	0,55 (III)
Taymyr Dolgano Nenets Autonomous Okrug	0,66 (IV)	0,66 (IV)	-	-	-
Chukotka Autonomous Okrug	0,63 (IV)	0,61 (IV)	0,60 (IV)	0,57 (III - IV)	0,55 (III)
Republic of Sakha (Yakutia)	0,62 (IV)	0,61 (IV)	0,59 (IV)	0,55 (III)	0,56 (III - IV)
Yamalo-Nenets Autonomous Okrug	0,63 (IV)	0,65 (IV)	0,60 (IV)	0,57 (III - IV)	0,55 (III)
Republic of Komi	0,67 (IV)	0,67 (IV)	0,62 (IV)	0,60 (IV)	0,61 (IV)
Khanty-Mansiysk Autonomous Okrug Yugra	0,60 (IV)	0,64 (IV)	0,60 (IV)	0,54 (III)	0,55 (III)

Note. 1. The table gives the value of integral indicator; in brackets the class of quality of life; 2 – for Taymyr (Dolgano-Nenets Autonomous Okrug) the data prior to 2005 are given, for by the results of the referendum held on 17 April 2005, from 1 January 2007 Taymyr (Dolgano-Nenets) Autonomous Okrug was abolished, and the municipal Taymyr Dolgano-Nenets Autonomous region was made part of Krasnoyarsk Territory as an administrative-territorial unit with a special status.

4. Conclusions

To conclude, we will note the advantages of using the examined approach for evaluation of integrative properties of complex natural and social systems and the quality of life of people. When building classification models, the investigator introduces the classes of states of the systems and the quality of life; uses the axiological approach and axiometry (ecological qualimetry), validates the type of an integral indicator, solves the problem of rating initial data taking into account the type of communication (direct, reverse) and its non-linearity, takes into account *nnn*-information on evaluation priorities; works with evaluation scales of the necessary and sufficient evaluation criteria, though may also use qualitative scales for evaluation; can introduce several levels of convolution of indicators, specifies or simulates weights (priorities) of evaluation inside groups, subsystems (levels) between them and can change them when necessary.

The use of models in GIS environment allows territories to be divided or zoned by values of integral indicators; the temporal dynamics and spatial differences of integral indicators to be traced, a conclusion on the ability of the systems to safeguard or change the class of state (quality) in time to be drawn. The flexibility of the model algorithms allows a hierarchical system of multi-level evaluation of the state of systems when there is an uncertainty at each level of hierarchy to be organized, and the mapping of integral indicators, division and zoning of territories on this base determines new capacities of the evaluation of states of regions and quality of life of their population.

5. Acknowledgements

This research was financially supported by the Russian Foundation for Basic Research by the following grant projects: 16-

05-00715-a“Development and testing of models of integrated assessment of the sustainability of land and aquatic landscapes and socio-ecological-economic systems”.

References

- [1] United Nations Organization. Economic and Social Council. Statistical Commission Forty-second session 22–25 February 2011 Clause 4(o) of the preliminary agenda E/CN.3/2011/1. Questions for information: measuring efficiency of economics and social progress. Report of National Institute of Statistics and Economic Studies. I. Recommendations of Stiglitz Commission. P.2.
- [2] Transforming our world: Agenda of sustainable development for the period until 2030. The resolution taken by the UN General Assembly 25 September 2015, A/RES/70/, 44 p.
- [3] Report of the interdepartmental group of experts for indicators of objective fulfillment in the sphere of sustainable development. UNO. Economic and Social Council. Forty-Seventh session. 8-11 March 2016. E/CN.3/2016/2, 46 p.
- [4] Khovanov N.V. Analysis and synthesis of indicators in the information deficit. SPb.: SPbSU Publishers, 1996. 196 p.
- [5] Noam Lior. Quantitative indicators of sustainable energy development. Energy Bulletin, 2015, No.19, p.8-29.
- [6] Dmitriev V.V., Osipov G.K. INTEGRAL ASSESSMENT OF STABILITY OF SOCIAL-ECOLOGICAL-ECONOMIC SYSTEM AGAINST CHANGES IN ITS FUNCTIONING CONDITIONS. 17th International multidisciplinary Scientific Geoconference SGEM 2017. Conference Proceedings. Volume 17. Ecology, Economics, Education and Legislation. ISSUE 52. Ecology and Environmental protection. 29 June – 5 July, 2017.Albena, Bulgaria, 565-572 pp. DOI: 10.5593/sgem2017/52.

[7] Dmitriev V.V., Lobacheva S.V., Chistilina V.S. Assessment of the state of socio-ecological-economic systems of the regions of Russia: AZR regions. Materials of the International scientific and practical conference "Modern Ecology: Education, Science, Practice", 2017, Voronezh, p.297-301.

[8] Dmitriev V.V. & Kaledin N.V. Integrated assessment of regional socio-ecological-economic systems and the quality of life (case study the constituent entities of the North-Western Federal District of Russia), Vol. 8, No.2, The Baltic region, Russia, 2016, pp.125-140.

[9] Vasiliy Dmitriev, Vladimir Kaledin, Nikolai Kaledin
THEORY AND PRACTICE OF INTEGRATED ASSESSMENT

OF THE STATE OF COMPLEX SYSTEMS IN NATURE AND SOCIETY 16th International multidisciplinary Scientific Geocoference SGEM 2016, Book 5, Ecology, Economics, Education and Legislation, www.sgem.org, SGEM2016 Conference Proceedings, ISBN 978-619-7105-66-7; ISSN 1314-2704, 30 June - 06 July, 2016, Albena, Bulgaria, Vol. II, 871-877 pp. DOI: 10.5593/SGEM2016/B52/S20.113.

[10] Dmitriev V.V. & Kaledin N.V. Integrated assessment of regional socio-ecological-economic systems and the quality of life (case study the constituent entities of the North-Western Federal District of Russia), Vol. 8, No.2, The Baltic region, Russia, 2016, pp.125-140

CONCEPTUAL CYBERNETIC MODEL OF TEACHING AND LEARNING

Prof. Ing. Veronika Stoffová, CSc.
Faculty of Education – Trnava University in Trnava, Slovakia
NikaStoffova@seznam.cz

Abstract: The article describes conceptual models of learning and teaching as a managed cybernetic system to achieve the knowledge and skills required by the standards. In the article, the learning process is examined as a system for building a knowledge system of a learner actively which is managed and directed by the tutor or teacher to achieve specified learning goal. The information received by the learner has a multimedia nature. The recipient receives the information through various channels (sensors), sight, hearing, smell, taste and touch. The received information arrives in the short-term memory where it is processed - confronted with the previous knowledge in long-term memory of the individual. New information and knowledge can reinforce and extend the recipient's long-term knowledge system if they are consistent logically. If there is a conflict between new information and old knowledge and information - the conflict needs to be solved. The solution can be the clarification of the knowledge system, correction of misconceptions so they are in line with objective reality and relevant knowledge about the subject of learning. It means, learning is not a constant storage of isolated information and information units in memory, but their transformation into active knowledge which is connected to logical structures and forms the knowledge system of the learner. The actual knowledge system enables an individual to solve non-standard problems not only in everyday life but also in science and technology and in various research areas from which the necessary, suitable and useable knowledge (expert) system is built.

Keywords: MODEL OF LEARNING, MODEL OF MULTIMEDIA LEARNING, CYBERNETIC MODEL OF LEARNING, COGNITIVE PEDAGOGICAL AND PSYCHOLOGICAL MODEL OF LEARNING

1. Introduction

Based on the theory of cognitive learning, learning is a complex process of acquiring knowledge and improving the cognitive abilities of a person. The aim of general education is the wisdom – so the learner becomes a wise man to be able to successfully integrate into other people's society, to share the achievements, material and spiritual values of human society, and to be able to contribute to their development, preserve and transfer them.

The success of human society consists of the fact that it can effectively transfer its knowledge to other generations which are able to further develop it. The continuity in the learning of humanity is ensured by education. A wise society creates a stimulating environment for individuals to learn and to ensure their future development. It creates an access to information sources, provides help, motivation and positive patterns. Education is a lifetime process and it can take various forms. It can be organized and institutionalized which means to take place in specially designed institution (e.g. school) or to take place outside the educational institution during everyday life consciously or unconsciously. The institution confirms formal education obtained at a school with a certificate of education. Knowledge, skills and other personal qualities, which society considers to be significant, are informal evidence of the education of an individual. A person, who wants to use the achievements of human society, must be aware of them. Therefore, the acquisition of knowledge is one of the main aims of education. It is obvious that the knowledge of the individual cannot contain everything. A person chooses what he/she wants and needs to learn. His/her choices are influenced by affects, surroundings and life situations in which he/she is involved.

It is necessary to realize that the knowledge is not transferred to a student, but it is created by thinking. Each information and knowledge acquired by a learner must be transformed into a knowledge by the learner – incorporating them into the learner's own knowledge system. In order for education to be successful, the cognitive abilities and thinking of the learners must be developed by the teachers. It depends on the learners to what depth and quality the knowledge will be achieved, in what way an individual will be able to use it in his/her life, and whether he/she will be able to contribute to its creative development.

Teaching in the school environment will be considered as a controlled learning that aims to achieve the prescribed standard of knowledge. Each individual builds his/her knowledge system for his/her whole life. The knowledge system of an individual is not an encyclopaedia of isolated information units, but a system of knowledge linked to relations expressing the context into one

organic whole. In school education, there is a selection of knowledge that a student should master, contained in a curriculum which is created by experts in education in each country. The role of a school is to create a stimulating environment in which the student is motivated to accept this selection of knowledge [14]. School education is the most effective form of education since it focuses on fulfilling the prescribed standards.

2. The Cybernetic Model of Education

It is possible to present the model of the educational system using a control system where the real system is the learner himself, in fact his knowledge system. The process which we manage is student's learning, that means building of the knowledge system of the learner who is supposed to achieve standards in each of the parameters or the prescribed level of the profile of the graduate.

The motivation, the work of the teacher, the acquisition and processing of information (a process of learning) changes the real object. The ideal system which is compared to the real system is expressed by the standards. In case that the learner does not reach the prescribed level of knowledge, based on the difference in knowledge, it is necessary to apply the feedback regulator represented by the teacher, by the influence of the environment and the surroundings, by acquiring new knowledge from different sources, that means learner's process of learning. The feedback regulator does not respond to the positive difference, the situation when the learner knows more than it is prescribed by the standards. However, this may have a positive effect on the activities of the learner and may increase the motivation to self-study [13], [14].

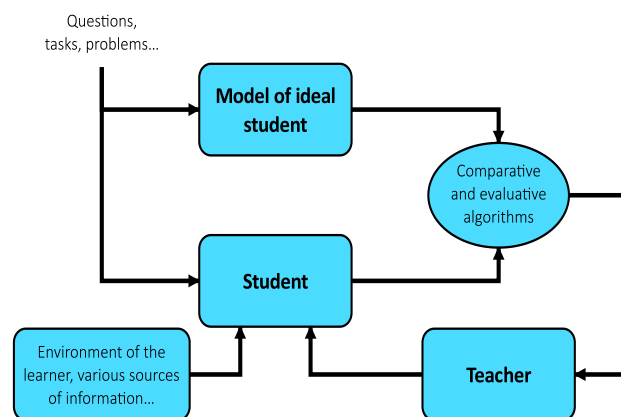


Fig. 1 Cybernetic model of school education

3. The model of multimedia teaching and learning

Learner perceives information and knowledge by several senses. Information is mostly gathered by *sight - eyesight*. According to Fredmann, almost 83% of information we sense with eyesight. Even our research, as well as other researches approved this hypothesis. Graphical information in form of a picture, an animation, a graph etc. can be considered as concentrated form of presenting particular knowledge. Hearing is second mostly used sense for gathering information with 11%. Other senses, such as touch or taste are not as important as sight and hearing. Researches show that we are able to remember only 20% of what we have heard or read, 30% of what we have seen. We are able to remember almost 70% of all information by integration of previous activities with feedback. To be able to remember 80% and more information it is necessary to effectively use gathered information and transfer it into personal system of knowledge. This will help with deduction of new knowledge [03], [04], [16].

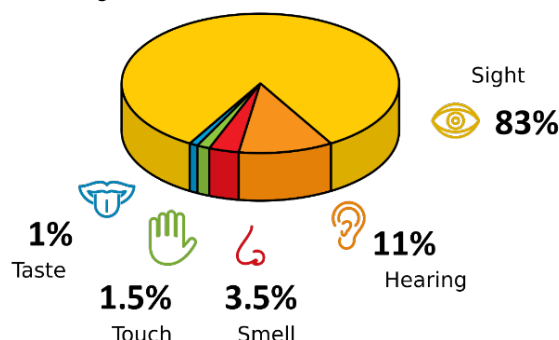


Fig. 2 How do we get information and knowledge

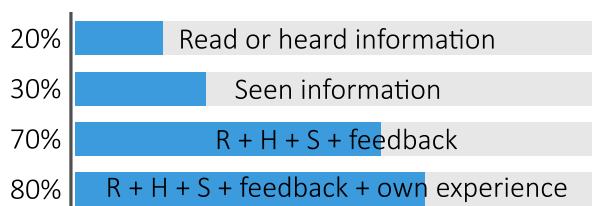


Fig. 3 How much we remember from the acquired knowledge

This theory and experimental results are in accordance with Niemierko's and Bloom's taxonomies of educational goals [09], [13].

General taxonomy of cognitive objectives describes 7 levels of how to handle a topic - gathering knowledge. 1. Knowledge, 2. Comprehension, 3. Application, 4. Analysis, 5. Synthesis, 6. Evaluation, 7. Imagination.

Niemierko's taxonomy of cognitive educational goals consists of four levels and is used especially for technical and exact sciences. **Level of knowledge:** 1. remembering; 2. understanding. **Level of competences:** 3. Specific transfer - application of acquired knowledge according presented tasks, in standard situations; 4. Non-specific transfer - an application in new and non-specific - real problem situations.

Niemierko's taxonomy of cognitive aims can be transformed into the knowledge requirements of the learner. The learner knows:

- 1) To list, to reproduce, to repeat, to name;
- 2) To explain, to describe, to give examples, to say/express with his/her own words;
- 3) To apply, to calculate, to demonstrate, to quantify;
- 4) To compare, to judge, to defend, to justify, to draw conclusions, to generalize, etc.

Bloom's Taxonomy (created in 1956) in order to promote higher forms of thinking in education, such as analysing and evaluating concept, process, procedures and principles, rather than just remembering facts (rote learning) [02]. It contains 6 linear consecutive cognitive levels: Evaluation, Synthesis, Analysis, Application, Comprehension and Knowledge. The cognitive domain involves knowledge and the development of intellectual skills. This

includes the recall or recognition of specific facts, procedural patterns, and concepts that serve in the development of intellectual abilities and skills. This six major categories of cognitive processes, starting from the simplest evaluation to the most complex knowledge (see the Fig. 4) for an in-depth coverage of each category. The categories can be thought of as degrees of difficulties. That is, the first ones must normally be mastered before the next one can take place.

The original Bloom's taxonomy was revised after 45 years [01]. The order of synthesis and evaluation categories was exchanged. The synthesis has been replaced by the "create" category, which is not understood as a rebuilding of individual elements, but includes a creative element along with the evaluation (the usage of critical thinking). The category of comprehension was renamed to understanding. The cognitive dimensions are expressed in the verb form, while the knowledge dimensions are in the form of a noun (substantive). Bloom's Taxonomy is mostly used when designing or modelling educational, training, and learning processes

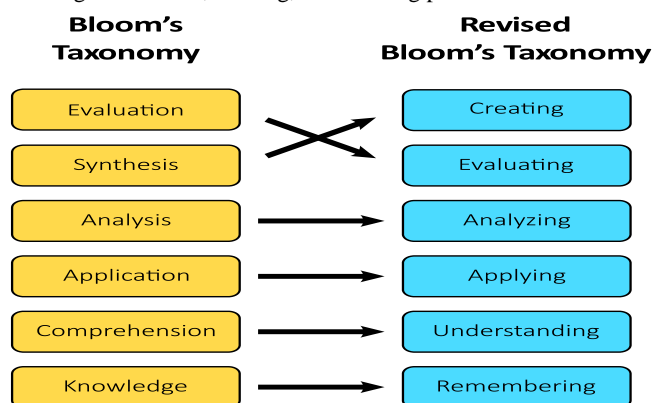


Fig. 4 Bloom's original and revised taxonomy

The chart (Fig. 4) compares the original taxonomy with the revised one changing the names in the six categories from noun to verb forms.

Revised Bloom's taxonomy recedes from the hierarchical order of the categories from the lowest to the highest. Overlapping may occur mainly during the educational activities, e.g. the learner sometimes evaluates even though he/she had not analysed the subject matter, or he/she creates new knowledge on the basis of partial knowledge. However, all dimensions should be shown. Table 1 provides examples of possible student activities for each of the component of the knowledge and cognitive dimension. Insertion the objective into the table is not always so easy and unequivocal. The formulation of the educational objectives and their taxonomic classification helps the teacher to choose the teaching methods (methods and forms of teaching) as well as it helps him/her in assessment, which is appropriate to the nature of the objectives [09]. The activity of the teacher and the learner should correspond to the chosen educational objective. Today, Bloom's taxonomy is easily understood and probably the most often applied one.

Table 1: examples of possible student activities for each of the component of the knowledge and cognitive dimension

The Knowledge Dimension	Remembering	Understanding	Applying	Analyzing	Evaluating	Creating
Factual Knowledge	List	Summarize	Sort, classify	Order	Choose	Combine
Conceptual Knowledge	Describe	Interpret, recognize	Experiment	Explain, compare	Estimate, determine	Plan, outline
Procedural Knowledge	Organize	Predict	Calculate, solve	Differentiate, depict	Conclude	Compose, design
Meta-Cognitive Knowledge	Appropriate Use	Process	Construct	Create	Perform, express	Actualize, improve

The effectiveness of learning is closely related to the participants' activity. Already, Edgar Dale (1900-1985), an American educational researcher, has found this in conjunction with developing the "Cone of Experiences" [05], which shows the effectiveness of the different teaching methods. From this we can see that the best methods are based on being more reliant on learner activity (Fig. 5). For this reason, today's most popular educational methods are all focused on activities of learners such as Inquiry-Based Learning, Problem-Based learning or Project-Based Learning. At the top of the pyramid are the less-active methods, where we listen and read the learning materials. In order for the educational environment to be effective, action, interaction and creativity should be introduced into education. This methods are located at the bottom of the cone [03], [04], [05], [06], [15].

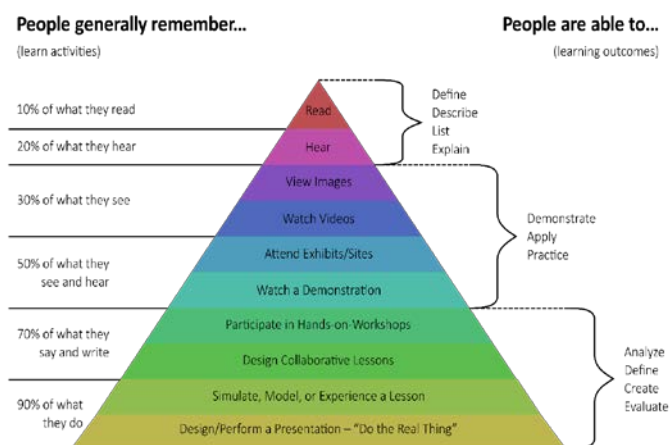


Fig. 5 Cone of Experiences

4. The Model of Knowledge System Creation

The teacher in the educational process affects the learner with the intention of achieving educational goals. Teaching is considered to be an optimal way of getting necessary knowledge. In implementing the learning model, it is important to apply both the taxonomy of objectives as well as the effectiveness of the various sources of knowledge and information that are acquired through different channels and the existence of different learning methods and styles. An important part of the learning model is the **presentation of knowledge** and information which are the subject of learning. This subsystem is part of the **model of the (ideal) teacher**. Learning material should be presented with the appropriate pedagogical transformation, in a multimedia structured form, with the respect to the student's mental level and his/her active involvement in the learning process. Another model of the subsystem of the teacher is **the management of the learning process**. It is necessary to identify learners, their learning style, their current level of knowledge, etc. to optimally manage the process of learning in order to individualize the learning process and shorten the time that is necessary to process gained information and knowledge. Incorporating new knowledge into the learning system of the learner means confronting new knowledge with knowledge in one's own knowledge system. In the case that the new knowledge is in accordance with the existing ones, the knowledge is deepened, and its durability is increased [16], [19]. If a conflict arises in assessing and processing new information and knowledge with the old ones, it needs to be solved. Thus, the individual constantly builds and improves their own knowledge system not only by consciously searching for new information and by checking and using it in practice, but also by the influence of the environment where the learner lives and in contact and communication with others. The objective of building the knowledge system is to make the individual wise, to change the quantity into the quality, so that understood (passive) information and knowledge become active, and therefore the individual is convinced of their relevance.

The knowledge system of an individual is not an isolated information and knowledge, but a system of knowledge – knowledge which is interconnected by relations that express the connection between knowledge. The knowledge system contains knowledge that is useful for solving everyday life problems, problems in professional life, employment problems and the problems of society we live in.

The learner analyses the acquired information, confronts them with knowledge of his/her own knowledge system, tests them by active usage, and incorporates them into his/her own knowledge system by synthesis, in other words the new information become knowledge. (See the Fig. 5.) It is according the rule of learning where the ultimate goal is a wise learner. „From information and knowledge to wisdom“.

In the Fig. 5 can be seen how the learner processes information and knowledge that is presented in different forms and how he/she creates his/her own knowledge system.

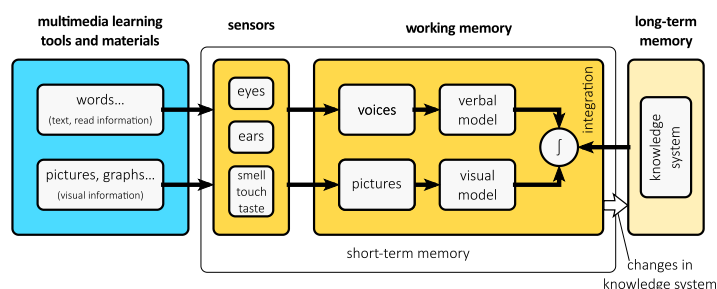


Fig. 6 Model of knowledge system creation

5. The Knowledge System Building

In constructing the graphical model of building the knowledge system, we proceeded from the cognitive theory of multimedia learning of Mayer. Mayer, based on his research, developed the Cognitive Theory of Multimedia Learning, based on 12 principles, which should be taken into account when developing multimedia teaching materials. These principles were grouped into three main categories [08].

- I. Reducing processing of information not covered (Coherence Principle; Signalling Principle; Redundancy Principle; Spatial Contiguity Principle; Temporal Contiguity Principle).
- II. Control the processing of relevant information (Segmenting Principle; Pre-training Principle; Modality Principle).
- III. To promote constructive processing of information (Multimedia Principle; Personalization Principle; Voice Principle; Image Principle).

In the Fig. 5 is illustrated how a knowledge system is built and it is also easy to see how the human brain processes several types of information.

On the basis of the results of Mayer's research, the basic principles of multimedia learning have been developed. Those results are confirmed by our research results in the area of the usage of the multimedia learning aids for programming in forms of animation-simulation models, which we have accomplished during the past years at J. Selye University in Komárno as well as at Trnava University in Trnava. The animations were developed in HTML5 using JavaScript technologies. We also used the CreateJS libraries (www.createjs.com) for animating the objects.

Our research results have shown that when the explanation of teaching material is in textual form, it is important to give students enough time to read and comprehend it. According to the principles of Mayer's multimedia learning, it is not recommended to show the text and animate the objects on the screen at the same time. It is better to let students start the animation themselves after reading and understanding the explanation. Another possible solution could be to use narration during the animation instead of textual explanation [16], [17], [18], [19].

The results of several experiments emphasized that the animations may be used more efficiently when they are not displayed alone, but when they are part of a learning environment. This environment could be an electronic textbook with hypertext structure, enriched with embedded animations, diagrams, and examples [19].

The limited scope of the article does not allow detailed analysis of the individual subsystems of the model. We plan to publish this analysis in the next article about the modelling of the learning process.

6. Conclusion

The model of learning – building of the knowledge system of the learner can be implemented in various ways. It might be built as a learning expert system [12], a virtual “second life” university or as a learning environment created with the support of LMS and CMS, where the presentation of knowledge can have a different form, e.g. multimedia interactive animation-simulation model of the studied phenomenon, or dynamic process [10], [11]. The management and optimisation of the learning process should be realised on the basis of the logical hyper-structure of the knowledge of the studied object, which is the subject of the education. In that case the logical hyper-structure of the knowledge contributes to the systematisation of the knowledge and to the correct improvement and building of the knowledge system of the learner. These kinds of education systems must be implemented in a such way that their utilization reduces the time needed to understand new knowledge, to process them and to convert them into active knowledge, that are useful and easily accessible in the long-term memory of the learner. They help him/her solve various problems, not only in the professional field, but also in everyday life. If we want educational software systems, which support learning, to deliver the expected effect, we must respect the rules of creating and building a knowledge system and the pedagogical, psychological and didactic rules of learning.

References

- [01] Anderson, L. et al. (2001) A Taxonomy for Learning, Teaching a Assessing of Educational Objectives. New York: Longman, 2001. 352 s. ISBN 0-321-08405-5.
- [02] Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *The Taxonomy of Educational Objectives, The Classification of Educational Goals, Handbook I: Cognitive Domain* (B. S. Bloom Ed.). New York: David McKay Company, Inc.
- [03] Czakóová, K. (2016) Creation small educational software in the micro-world of small languages. In: *Teaching Mathematics and Computer Science*. 14th volume, issue one, 2016/1, p. 117. Debrecen : University of Debrecen, 2016. ISSN 1589-7389
- [04] Czakóová K. (2016) Felfedezésen alapuló aktív tanulás mikrovilág környezetben. (Discovery-based active learning in microworld environment). In: *New methods and technologies in education and practice : 19. DIDMATTECH 2016*. Budapest : Eötvös Loránd University, 2016. 168-173. s. ISBN 978-963-284-799-3
- [05] Cone of learning: (On-line <http://bit.ly/1TNR2o6>, last access: 2017. 11.10)
- [06] Illés, Z., H. Bakonyi V. (2016) Experiment for increasing equal opportunity in university with the support of a BYOD system, SzámOkt 2016, Kolozsvár október 8-9. ISSN 1842-4546
- [07] Illés, Z. (2017) Valós idejű rendszerek és megjelenésük az oktatásban - Real-time systems and their appearance in education. (Habilitation theses). ELTE- FI, Budapest 45. p. 2017
- [08] Mayer, R. E. (2009). *Multimedia Learning* (second ed.). New York, USA: Cambridge University Press.
- [09] Majherová, J.: Revidovaná Bloomová taxonómia a kompetencie pre používanie IKT - Revised Bloom's Taxonomy and Competencies for Use of ICT, on-line access - https://www.pdf.umb.sk/~lrovnanova/taxonomia_ciele_Anderson.pdf
- [10] Stoffa. V. (2004) Modelling and simulation as a recognising method in the education. *Educational Media International*, 41(1):51-58, 2004.
- [11] Stoffa V. (2008) Az animáció szerepe az elektronikus tankönyvekben. *Információs társadalom*, VIII(3):113-125, 2008.
- [12] Stoffová, V., Gergelová, Z. (2001) Expert systems in the systematisation of the knowledge at primary schools (Expertné systémy v systematizácii poznatkov na ZŠ. *Technológia vzdelávania : Zväzok 3 : Educational Technology : Volume 3*. 1. vyd. Nitra, Pedagogická fakulta UKF v Nitre, 2001, s. 160-165.
- [13] Stoffová, V. (2004) *Počítač – univerzálny didaktický prostriedok* 1. vyd. Nitra : Fakulta prírodných vied UKF v Nitre, 2004. 172 s. ISBN 80-8050-450-4
- [14] Stoffová, V. (2006) The Importance of Didactic Computer Games in the Acquisition of New Knowledge. In: *The European Proceedings of Social & Behavioural Sciences EpSBS*. pp. 676-688. eISSN: 2357-1330. (on-line access: <http://dx.doi.org/10.15405/epsbs.2016.11.70>)
- [15] Stoffová, V. - Czakóová, K. (2016) Prostredie na učenie sa bádaním. In: *Úvod do programovania v prostredí mikrosvetov : vysokoškolská učebnica*. Komárno : Univerzita J. Selyeho, 2016. 8-33. s. ISBN 978-80-8122-170-5
- [16] Végh, L. – Stoffová, V. (2016) An interactive animation for learning sorting algorithms: How students reduced the number of comparisons in a sorting algorithm by playing a didactic game. In: *Teaching Mathematics and Computer Science*. Debrecen : Institute of Mathematics – University of Debrecen, 14th volume, issue one, 2016/1, s. 45–62. ISSN 1589-7389.
- [17] Végh, L. (2011) Animations in teaching algorithms and programming (Animácie vo vyučovaní algoritmov a programovania). In Jiří Dostál, editor, *Nové technologie ve vzdelávání*, pages 47-51, Olomouc, CZ, 2011. Palacký University, Olomouc.
- [18] Végh, L. (2011) From bubblesort to quicksort with playing a game (Hravou formou od bublinkového triedenia po rýchle triedenie). In Jiří Neubauer and Eva Hájková, editors, *XXIX. International Colloquium on the Management of Educational Process*, pages 539-549, Brno, CZ, 2011. University of Defence.
- [19] Végh, L. (2017) A programozás tanulásának és tanításának támogatása elektronikus tananyagba beépíthető interaktív animációs modellekkel. (PhD theses) ELTE - Faculty of Informatics, Budapest 202. p. 2017

SIMULATION MODELING OF AUDITORY FUNCTION

Ass. Prof., Dr. Eng. Donii O.¹, Dr. Med. Pisanko V.², Ass. Prof., Dr. Eng. Kulinich A.¹, Ass. Kotliar S.¹
National Technical University of Ukraine "Kyiv Polytechnic Institute named after Igor Sikorsky" - Kyiv, Ukraine¹
SI"O.S.Kolomiychenko Institute of Otolaryngology Of National Academy of Medical Science of Ukraine - Kyiv, Ukraine²
Email: dosha@iff.kpi.ua

Abstract: The hypotheses concerning encoding of information in the peripheral part of the human's auricular analyzer are presented. A brief critical analysis of contemporary trends in theoretical concepts concerning principles of work of the cochlea of the inner ear are conducted. Prerequisites for the construction of an alternative theory of coding of information in it are formulated in order to optimize the design and software of cochlear implants. The principle of constructing an imitation model of the generation of electric signals formed in the cochlea of the inner ear are proposed.

KEYWORDS: COCHLEAR, INNER EAR, BASILAR MEMBRANE, HEARING, SIMULATION MODELING

1. Introduction

Hearing plays an important role in the development of speech, intelligence and formation of the human's psyche. The information which is received through hearing, is not less important than the information which is perceived by sight. Loss of functions of hearing by a human largely limits his/her communication and leads to serious psychological and moral problems. Therefore, the task of full or partial, but socially adequate, recovery of hearing is an urgent one. To solve this problem hearing aids, as well as cochlear implants, are used in modern medicine. The task of hearing aids is simple and consists in shaping the frequency response of the amplifier to compensate for a decrease in the sensitivity of the patient's auditory system in a certain range of frequency. Implants of cochlea are constructed on the principle of separating the audio signal into several signals in separate frequency bands, followed by further direct electrical stimulation of the auditory nerve by each of them. They are used in case of violation of the peripheral part of the auditory analyzer, more specifically - of the patient's inner ear. Nowadays medicine has made significant progress in the use of cochlear implantation. Surgical techniques of their implantation are well established. However, there are certain difficulties related, mainly, with the perception of signals from the implant. These difficulties are caused by principles of implants' work, which are based on a fairly rough analysis of the audio signal's spectrum and selection of several frequency channels, each of which excites a certain portion of the auditory nerve. This ideology of the cochlear implantation is based on the current understanding of work of the peripheral part of the auditory analyzer, which is based on theory of Békésy [1], concerning the running wave on the basilar membrane (BM), and on "frequency-place" principle. There is enough experimental data that can be interpreted as a confirmation of this theory. However, there are also experimental data which contradict it [2]. One can say that today there is no consistent view on the functioning of the cochlea of the inner ear and representation of the audio signal in the structures, which stimulate the auditory nerve. Therefore, in this paper, a brief critical analysis of contemporary trends in the theoretical concepts of the peripheral part of the auditory analyzer is given and prerequisites are formulated for the construction of an alternative hypothesis on the coding of information in auditory analyzer. Creation of the experimentally proved theory on the basis of this hypothesis will provide an opportunity for further critical rethinking of principles of the cochlear implants' design in order to improve the adaptation of patients with deafness caused by damage of the inner ear's cochlea.

2. Preconditions for resolving the problem

One of the known effects, inexplicable from the standpoint of the theory of "frequency-place" is the binaural effect. Evaluation of the time difference of arrival of the same wave's phases to both ears may occur, evidently, only in the brain centers, that means that the periodic nature of the sound process should somehow appear in the neural processes of the cortex. Meanwhile, the theory of "frequency-place", as the theory of "peripheral analyzer" refers assessment of sound solely to the excitation of nerves in the given area of the cochlea. This leads to the emergence of new theories of hearing. One of such theories is the theory of G. Fletcher [3].

According to this theory, it is not individual strings of basic membrane respond to the audio waves, but peri- and endolymph of the cochlea do this. Plate of the stapes transmits sound vibrations of the cochlea's fluid to the BM, at that the maximum of amplitude of these oscillations at higher tones lies closer to the base of the cochlea, at lower ones - closer to its top. Nerve fibers which end in the main membrane, including the organ of Corti (according to some authors) resonate only at frequencies above 60 - 80 Hz. There are no fibers, which receive more lower frequency on the main membrane. Nevertheless, feeling of heights up to 20 Hz is formed in the brain. It appears like the combination of high tone harmonies. Thus, from the Fletcher hypothesis's viewpoint, perception of the low tones' pitch is explained by perception of the whole complex of harmonic overtones, and not only by perception of the frequency of the main tone, as it was usually taken so far. And as content of overtones to a large extent is dependent on the intensity of the sound, then it becomes clear a close relationship between the three subjective qualities of the sound - its height, volume and timbre. All these elements, each of them individually, are dependent on the frequency, strength and composition of the sound's overtones. According to the Fletcher's hypothesis, resonant properties are inherent to the mechanical system of the cochlea as a whole, not just to the main membrane's fibers. Under the influence of a certain pitch, not only fibers, which resonate with this frequency, oscillate, but the entire membrane as well, and also this or that amount of fluids in cochlea. High tones force to drive only a small mass of liquid near the base of the cochlea, low ones - are fixed closer to helicotrema. Fletcher also overcomes the main difficulty of the resonance theory associated with explanation of a large range of volume. He believes that the volume is determined by the total number of nerve impulses, coming to the brain from all the excited nerve fibers of the basal membrane. Fletcher's theory, in general, does not deny existence of "frequency-place" theory and it can be attributed to the theories of "peripheral analyzer".

Theories of "central analyzer" or so-called "telephone theory", form another group of theories [4]. According to these theories, audio vibrations are converted by cochlea into synchronous waves in the nerve and transmitted to the brain, where their analysis and perception of level of tone takes place. J. Ewald's theory, which was proposed in the late of 19th century, also belongs to this group of theories. According to this theory, under effect of the sound, standing waves are formed in the cochlea with a length which is determined by the frequency of the sound. The level of the tone is determined by the perception of the shape of the pattern of the standing waves. The feeling of a certain tone corresponds to the excitation of one part of the nerve fibers, and the feeling of a different tone corresponds to the excitation of another part. Analysis of sound is performed not in the cochlea but in the central areas of the cortex. Ewald succeeded to build a model of BM, with the size which approximately corresponded to the real ones. In his experiments the entire membrane began to vibrate when it was excited by the sound. The "audio picture" appears in the form of standing waves with the length, which is as smaller, as the sound is higher. Despite the successful explanation of some embarrassing particulars, Ewald's theory (as well as other theories of "the central analyzer") hardly corresponds to the latest physiological researches

of the nerve impulses' nature. In the [3] the dual point of view is expressed, namely, the explanation of the perception of high tones is given in the sense of "peripheral analyzer" and of low ones - from the perspective of "the central analyzer".

American researchers, who first-ever implanted microelectrodes in cat's cochlea, registered electrical potentials which arised in the cochlea [5]. Based on their observations, they created an electrophysiological theory of hearing. According to this theory, every hair of hair cells of organ of Corti is similar to the piezoelectric crystal. As it is known, these crystals have an interesting property - upright they are neutral, but when they are bent even a little bit then electric charge appears immediately. In case of fluctuations of BM, hair cells, naturally, begin to oscillate also. But the tectorial membrane pushes on top of the hair, they bend, causing an electrical charge. Thus, under the influence of the deformation of receptor cells' hairs in sync with the sound's vibrations, the electrical energy is released, and biological currents appear. These biological currents stimulate the thinnest endings of branches of auditory nerve, which criss-crosses the hair cells. Through this nerve and conductive pathways of medulla oblongata excitation is transferred to the cortex of the temporal lobes of the brain, where the analysis and synthesis of audio stimuli takes place. Thus, at the moment it is common to speak about the duality of the mechanisms of perception of the pitch: in the high-frequency range the most acceptable is the principle of "place", in the area of lower frequencies - a modified principle of "bursts". Despite the long history of the discussions on the principles of the functioning of the auditory system as a whole and its peripheral parts in particular, and the availability of a huge number of researches related to the study of the perception of pitch, it is obvious that the mechanism of coding the information in the peripheral part of the auditory analyzer is not completely elucidated and requires further intensive research.

3. Formulation of a hypothesis for model development

From the perspective of physics, dynamic range of 125 dB is a unique parameter of the human auditory system. Thus, the maximum amplitude of the audio signal at the system's input (eardrum) differs from the minimal amplitude in trillion times. Such value of the dynamic's range leads to suggestion that there is not only one, but several mechanisms of perception of the sound in the peripheral part of the auditory analyzer, which is consistent with the views of a number of scientists. For example, it is known [6], that the subcutaneous plate of stirrup at high intensity of the input's signal, close to the maximum intensity, moves from the translational vibrations to the vibrational-rotational ones, thus preventing the entire system of middle and inner ear from the mechanical damage. Then this is logical to assume the existence of several reactions to sounds, which are different not only by frequency, but also by intensity. Thus, it is interesting to compare processes of formation of auditory images in the cochlea of the inner ear at maximum and minimum levels of input signals.

Taking into account the assumption that there are several mechanisms of the inner ear's functioning while converting sounds of different intensity into the electrical activity of the auditory nerve, let us consider the mechanical processes in the cochlea when it is exposed to weak (low intensity) beeps.

As it was mentioned above, Békésy G. substantiated hypothesis, called the "theory of running waves", which states that the vibrations are distributed through BM in the form of running waves having a maximum of the envelope at a certain point of membrane, whose place varies depending on frequency of the acting signal. A number of authors who used different methods for registration of membrane's oscillations, experimentally confirmed the basic provisions of G. Békésy's hypothesis [7 - 10]. However, direct observations of BM's vibrations, executed by G. Békésy, were carried out at a power of the sound of 90 - 120 dB. Using an extrapolation of the existing experimental data it is possible to calculate that the amplitude of BM's movement in the place of the running wave's envelope's maximum is about 10^{-12} sm [1, 7]. This value is much smaller than the amplitude of the thermal motion of

the molecules of cochlea fluids. Furthermore, in order to have any fluctuations of BM, the pressure of sound's signal must overcome cochlea's elasticity and its inertia of rest. And, obviously, because of its stiffness, it will not respond to such a small amount of energy. So, it can be postulated that BM is stationary at low intensities of the sound's signal. And this gives a basis to suppose the absence of influence of BM's mechanics on the analysis of the sound's information at small acoustic signals. At the same time, the presence of a running wave can say about ability of the cochlea's structures to absorb the excess of energy at high levels of the sound. This assumption, as well as high sensitivity of acoustic analyzer, give reasons to look for other mechanisms of perception of the sound in the cochlea, especially at low levels of intensity of acoustic signals. In a number of works [8, 9], where the amplitude-frequency characteristics of the BM's vibrations were investigated, it turned out that they are less selective than it was expected based on the values of differential threshold of frequency. This allowed to suppose the presence of so-called "the second filter", which is located, according to the researchers' viewpoint, at the junction of the tectorial membrane - wax cells of the organ of Corti or hair cells - the fibers of the auditory nerve Ψ [11 - 13], although such a filter was not detected experimentally [11]. Also in the work [14] the hypothesis is suggested concerning the existence of molecular resonance mechanism, which is localized in the tectorial membrane. Due to different speeds of sound's movement in the perilymph and in the tectorial membrane, concentration of the acoustic energy takes place in the latter, resulting in the conversion of mechanical energy into biochemical one through resonance and movement of ions, which cover complex of the tectorial membrane - the hair cells of organ of Corti. There is a hypothesis, according to which the flanking sprouts of Deiters cells and cuticular region of the hair cell form a pair, called "Auron" by the authors, that makes primary frequency analysis of sounds using resonant oscillations (hypothesis of "auron resonance"). A number of authors [4, 15] express the opinion that the frequency analysis in the cochlea is carried out by means of resonant vibrations of the hair cells and the tectorial membrane. These assumptions are confirmed by correlation between the lengths of hair cells' stereocilia and characteristic frequencies [4, 15]. Thus, modern researches have led to the need for a thorough study of the "thin" structures of the organ of Corti.

Imagine the possible mechanism of the cochlea's mechanical part under the influence of weak acoustic signals (0-20 dB above threshold). In this case, taking into account the above mentioned extrapolation, let us assume that BM is immovable. If we are adhere to the concept of a mechanical nature of transformation of sound's vibrations into receptors' potentials of the hair cells, it is necessary to find in the structure of the organ of Corti those oscillating elements, that have to "feel" the impact of such a small incentive. To do this, we consider a schematic cross-sectional view of the organ of Corti in a single turn of the cochlea (Figure 1). The figure shows that the hair cells are connected with the tectorial membrane (this is confirmed experimentally by [1]), and the latter, in turn, is fixed in such a way that it can form a very sensitive lever system in vertical direction regarding to the BM's plane. The schematic connection of the covering membrane with hair cells is shown in Figure 2, at that it is assumed that the hair cells have their own elasticity (coefficients K_j). When the mechanical structure of the cochlea's membrane was studied, it was found that the tectorial membrane really moved easily in the direction which is perpendicular regarding to BM [1].

In the experiments described in the works of [16], it is shown that the stereocilia of hair cells are rather hard. It can be assumed that they alone or in a bundle together with the mass of the coating membrane form sensitive, and perhaps the resonance system. In the [4, 15], for example, the possibility of resonance in such a system is shown. This is confirmed also by the structure of the coating membrane: it consists of thin transverse fibers, which generally have a radial direction along the axis of the cochlea, and there is a transparent gluing substance between the the fibers [17]. Reissner's membrane is a very thin film of the same elasticity over the entire length. At that, its elasticity is small in comparison with the

elasticity of BM, through which transmission of oscillations passes freely. Its role is to separate endo- and perilymph and it seems, that it does not participate in the analysis of vibrations.

Based on the above mentioned facts, one can imagine a process of the membranes' vibrations in the cochlea in terms of action of acoustic signals with low levels of sound's pressure as follows: sound's pressure is transferred to a tectorial membrane through the perilymph, endolymph and Reissner membrane. Because of its structure, tectorial membrane responds to minimum pressure by substantial displacement in a plane, which is perpendicular to the BM, as compared with the displacement of MB (or its absence) at the same levels of input signals. The mass of the tectorial membrane and elasticity of the hair cells can form the sensitive system, which reacts to this movement. Moreover, taking into account the structure of the tectorial membrane, it can be assumed, that together with the hair cells, it forms the whole system of sensitive elements, which, in principle, can work as unbound tuned resonators. Taking into account calculations contained in [4, 15], their amplitude and frequency characteristics will be sharper than the same specifications for MB. BM in this case remains stationary and does not participate in the analysis of signals with such levels. It is obvious, that while increasing the sound's level, starting from some particular magnitude of the acoustic signal, vibrations of the BM become relevant and can not be ignored anymore. In this case the nature of excitation of hair cells, which determines the information transmitted by the auditory nerve, is to be altered.

This situation can be considered in more details. Under the influence of a weak signal BM is stationary. Sound's pressure, according to the laws of physics, is transmitted through the perilymph in all directions. At that, if the forward movement of the stapes is very slow (very low frequency), a column of liquid should overflow from channel to channel through helicotrema without creating significant pressure on the side walls. In this case, hair cells do not respond to the signal. By increasing the speed of movement of the stapes (increasing the frequency of the signal) perilymph does not have time to go through an opening with a limited area (helicotrema) and the radial component of pressure arises, which affects the complex "tectorial membrane - hair cells", generating an electric signal (it is possible that the area of helicotrema defines the lower boundary of perception of the signal). This pressure should be distributed over the entire length of the channel, that corresponds to the assumptions of the "telephone theory"

Thus, when the input signal has low intensity, the dotted excitation response of hair cells is possible. However, considering the fact, that pressure in liquid circulates in all directions simultaneously, and that the helicotrema restricts the flow of the perilymph from channel to channel, one may assume the simultaneous stimulation of hair cells in the area of a certain length.

The increase in signal intensity leads to the excitation of BM's vibrations and the appearance of a running wave on it. A running wave should lead to the displacement of structures of organ of Corti and irritation of hair cells throughout its whole length, not just at the point of its maximum amplitude. So, it is possible to propose a hypothesis about the formation of a some spatio-temporal pattern, which represents the excitation of the auditory nerve. At that, taking into account the form of envelope of a running wave, the excitation should reflect the reaction of hair cells in three spatial coordinates, as well as changes in the spatial "images" in time. Thus the spatial-temporal signal, having four coordinates: length, width, depth and time, is formed. Obviously, in such way principle of "frequency-place" and elements of the telephone theory are combined.

Considering the mechanisms of forming the excitation in the cochlea of the auditory analyzer, the question must arise about the nature of the signal applied to the input of the system in experimental studies. Usually it is a pure tone, which represents a sine wave with a certain amplitude and frequency. It is believed that the "frequency-place" principle works in this case. When the composite signal is supplied, then since the Helmholtz times in most studies it is suggested that the ear analyzes in this or that way its spectrum, represented as a set of sinusoids. However, representation of the composite signal as a set of sine and cosine,

which is widely used in electronics and acoustics, is a comfortable way, but not the only one. For example, according to the approximation theorem of Weierstrass any complex function can be approximately described by a polynomial of degree n , i.e. as a sum of exponential functions with different coefficients of [18]. And taking into account the fact that in the natural environment you can hardly meet pure tones, the natural response of the auditory system is the analysis of complex sounds. It is doubtful the possibility of the membrane of the cochlea of the inner ear to decompose non-periodic audio signal in Fourier series. Thus, it is not necessarily that the inner ear works as a spectrum analyzer, allocating and fixing the harmonic components of the audio signal. Hence the assumption arises (which supports the expressed above hypothesis) that the audio signal is not decomposed in the inner ear into components, but is perceived as a whole. Of course, the principle of "frequency-place" can not be denied, as it was confirmed experimentally. However, its confirmation was received when exposed to high intensity signals.

4. Computer model of signal formation in the cochlea

These assumptions should be confirmed by experimental verification. At the same time, it should be noted that the experimental study of structures (and particularly of "thin" structures) of the inner ear's cochlea in real objects are very labour-consuming and not very informative, even when using modern technology and equipment. Therefore simulation is one of the main methods of investigation of the inner ear. Unfortunately, among a large number and variety of existing models of cochlea, there are just a few of them which take into account not only the behavior of the BM, but also of other membranes, but even at that case they are used only with the aim to clarify its vibrational characteristics.

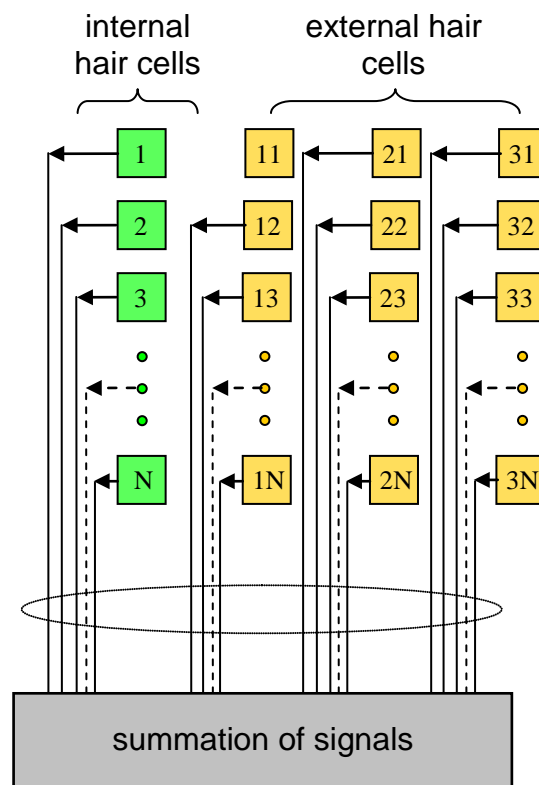


Fig. 1. Structure of the imitation model of the inner ear function

Taking into account complexity of the structure of the inner ear's cochlea, simulation modeling is conveniently chosen as a method of constructing of the model, since this technique was developed for studying of complex systems. Organization of hydromechanical part of the cochlea is complex, not linear and poorly researched. However, its response to the sound signal causes the excitation of hair cells, both internal and external. This excitation forms a common electrical signal, which is transmitted to the upper parts of

the brain along the auditory nerve. Based on the proposed hypothesis of the formation of some spatio-temporal pattern of oscillations in the cochlea, it is possible to model different variants of hair cells' excitation, and, accordingly, of their reaction as an aggregated electrical signal. This electrical signal can be registered in a real experiment. Comparing signals which were simulated by the model with the experimentally recorded ones, one can select the most similar ones and thereby make a conclusion about the real mechanisms of the inner ear's functioning. A simplified diagram of the model of the formation of an electrical signal of the hair cells' reaction to an external signal is shown in Fig. 1. Inner and outer hair cells are denoted by multicolored squares. In this variant, each cell (which can be regarded as a sensor) generates a single pulse regardless of the parameters of the input signal. All pulses which were generated at a given moment of time are summed in the summation block. Thus, a signal is generated, which is similar to that one, which can be registered in a real experiment. The fragment of output results of modeling of hair cells' excitation under influence of the input signal which was generated by means

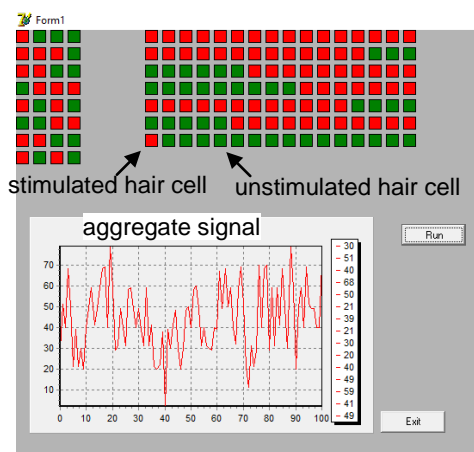


Fig. 2. The computer simulation result of the formation of an electrical signal at the "exit" of the cochlea of the inner ear

of the generator of random numbers is presented in Fig. 2. As it can be seen from Fig. 2, the resulting signal is similar to the microphone potential which was measured in a real experiment. Thus, one can see the prospects of developing of this approach for modeling of the auditory function. One can draw a conclusion about the principles of coding of information in the peripheral part of the human auditory system by analyzing the mechanisms of functioning of the inner ear's cochlea from the point of view of existing theories and developing models of signals that stimulate hair cells and comparing the results of modeling with the results of the experiment. Moreover, it seems expedient to select tests of subjective and objective researches of hearing, on the basis of which it is possible to create an adequate model of inner ear's functioning both in conditions of norm and pathology. This will deepen our knowledge of coding of auditory information and optimize design and software of cochlear implants on their basis. In case of confirmation of mechanisms of the inner ear's functioning in framework of the above mentioned hypothesis, the design of cochlear implants must be based on other ideas about coding of signals in the peripheral part of the auditory analyzer other than decomposition of the input signal into several frequency ranges.

5. Conclusion

1. The hypothesis of stimulation of hair cells of the inner ear's cochlea is proposed, in which the formation of a spatio-temporal picture of excitation as a reaction to an external sound signal is suggested. This hypothesis makes it possible to explain the sensitivity of the auditory analyzer at low intensities of input signal and, on the whole, does not contradict to the existing ideas about mechanisms of hearing's perception.
2. An imitation model of formation of an electrical signal of the hair cells' reaction to an external signal is proposed, in which each

cell generates a pulse and all pulses are summed in the summation block. In this way the electric reaction of a cochlea is simulated, which is similar to that one, which can be recorded in a real experiment.

3. It seems appropriate to conduct tests of subjective and objective researches of hearing for the analysis of modeling results on the basis of which the adequacy of the proposed hypothesis of functioning of the inner ear can be confirmed in conditions of norm and pathology. This will allow to deep knowledge about coding of auditory information and optimize on this basis design and software of cochlear implants.

6. Literature

1. Bekesy G. Experiments in Hearing. N.Y., McGraw-Hill, 1960.
2. Rubinshtein S.L. Foundations of General Psychology, St. Petersburg, Peter, 2002. (Russian)
3. Fletcher H. On the dynamics of the cochlea. - J. Acoust. Soc. Amer., 1951, v. 23., p. 637 - 645.
4. Zwislocki, J.J. Theory of cochlear mechanics. Hear. Res., 2, 1980, P. 171 - 182.
5. Kiang N.Y.-S., Watanabe T., Thomas E.C., Clark L.F. Discharge patterns of the single fibers in the cat's auditory nerve. Cambridge, Massachusetts, 1965.
6. Gelfand S.A. Hearing: An Introduction to Psychological and Physiological Acoustics, Marsel Dekker, INC, New York and Basel, 1981.
7. Johnstone B.M., Taylor K.J., Boyle A.J. Mechanics of the guinea-pig cochlea. J. Acoust. Soc. Amer., 47, 1970, p. 504 - 509.
8. Johnstone B.M., Yates G.K. Basilar membrane tuning curves in the guinea pig. J. Acoust. Soc. Amer., 55, 1974, p. 584 - 587.
9. Kohlöffel L.U.E. Observations of the mechanical disturbances along the basilar membrane with laser illumination. In: Basic mechanisms in hearing. New York; London, 1973, p. 95 - 117.
10. Rhode W.S. Some observations on cochlear mechanics. J. Acoust. Soc. Amer., 64, 1978, p. 158 - 176.
11. Labutin V.K., Molchanov A.P. Models of the mechanisms of hearing. M., 1973. (Russian)
12. Duifhuis D. Cochlear linearity and second filter: Possible mechanisms and applications. J. Acoust. Soc. Amer., 59, 1976, p. 356 - 358.
13. Hall J.L. Spatial differentiation as an auditory "second filter": assesment on a nonlinear model of the basilar membrane. J. Acoust. Soc. Amer., 61, 1977, p. 520 - 524.
14. Naftalin L., Path F.R.C. The Peripheral Hearing Mechanism: A Biochemical and Biological Approach. Annals of Otolaryngology, Rhinology & Laryngology, v. 85, №1, 1976, p. 38 - 42.
15. Zwislocki, J.J. Five decades of research on cochlear mechanics. J. Acoust. Soc. Amer., 67, 1980, p. 1679 - 1685.
16. Flock A. Physiological properties of sensory hairs in the ear/ In: Psychophysics and physiology of hearing. London, 1977.
17. Vinnikov Y.A., Titova L.K. Corti's organ. Histophysiology and histochemistry. M.-L., 1961. (Russian)
18. Fikhtengolts G.M. Course of differential and integral calculus. FIZMATLIT, v. 3, 2001. (Russian)

GENERATION OF AN ATLAS-BASED FINITE ELEMENT MODEL OF THE HEART FOR CARDIAC SIMULATION

M.Sc. Vasiliev Evgeny
Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia
eugene.unn@gmail.com

Abstract: In this paper an algorithm for creating an atlas-based finite element heart model for cardiac simulation is described. This model is used to simulate the propagation of electrical impulses of the heart. An important feature of this model is that it contains conductive paths and fibrous tissue, which makes it possible to make more realistic calculations of the propagation of electrical signals. The model created from anatomical segments of the heart surface is defined by a polygonal mesh. The algorithm presented in the article offers a means to create models of various accuracy.

Keywords: HEART MODEL, CARDIAC SYSTEM, RECONSTRUCTION, RAY CASTING, FE-MODEL, TETRAHEDRAL MESH.

1. Introduction

In order to simulate an accurate electrical system of the human heart, it is necessary to have a detailed heart model containing different types of tissues: the direction of the muscle tissue, conducting fibers, fibrous rings, etc.

Nowadays there is a number of finite element models (FE models) of heart for modeling electricity signals. For example, in some systems of heart electrical modeling, the rabbit heart model is used [R. Bordasa, 2011] [Arevalo HJ, 2016]. The rabbit heart model was constructed from a high resolution MRI dataset, with the use of an intensity based level-set filter.

The first developed 3D models of cardiac anatomy were simplistic models based on geometric shapes, this approach is still in use for applications where the anatomical validity is not important for the purpose of the model [Colli Franzone P, 1998] [Sermesant M, 2006]. Currently, researches use simple geometric shapes and make parameterized models based on segmented heart images [P. Lamata et al, 2014].

In many articles, studies are based on models built on the segmentation of CT [Deng D, 2012] [Aslanidi OV, 2013] or MRI [Plotkowiak M, 2008] [Lopez-Perez & Sebastian, 2015] scans. Models of this type very anatomically accurate, but usually they describe only one heartbeat phase, rather than the whole cycle.

Our algorithm enables automatic marking of vertex the type of tissues after the grid generation, because manual marking of hundreds of thousands and more tetrahedrons makes no sense.

2. Source data

As the initial data, anatomical segments of heart surface from Plastic boy anatomy (Plastic boy store) were used. We have several types of meshes (Figure 1):

- the outer surface of the heart;
- left atrium;
- left ventricle;
- right atrium;
- right ventricle;
- mitral valve;
- tricuspid valve;
- atrioventricular node;
- sinoatrial node;
- His bundle;
- Bachmann bundle;
- Signal ways.

We use the heart surface polygonal model to generate a tetrahedral FE model. We can vary the total number of vertices and tetrahedrons in the FE model.

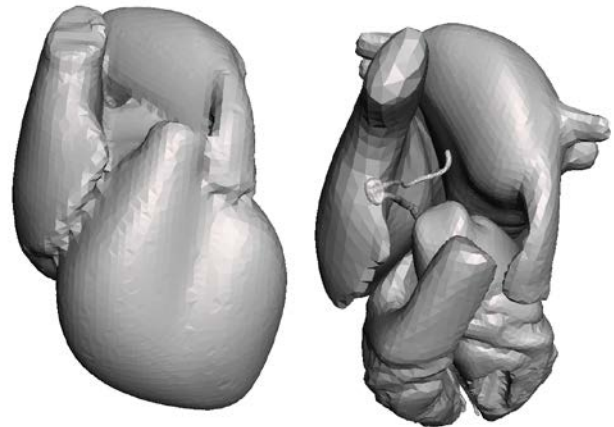


Fig. 1 The outer surface of the heart and inner tissues.

3. Algorithm of model generation

The algorithm consists of several parts: FE model generation; classification of atriums and ventricles; removal of cavities; cardiovascular system classification.

Step 1. Generation of the FE model.

Generation of a three-dimensional finite element mesh is labour-intensive process, that is why there are few high-quality open-source 3D mesh generators. We used Netgen for mesh generation, because Netgen is an open-source framework with LGPL v2 license, it is well-tested, and has its own user community. Netgen uses the outer surface of the heart in stl format, and yields a FE model consisting of tetrahedrons (Figure 2).

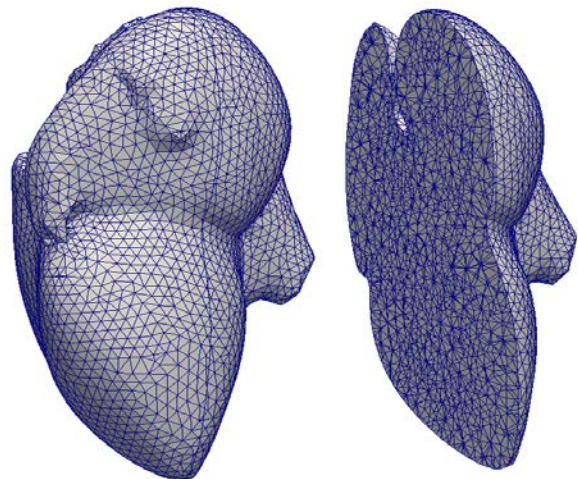


Fig. 2 Generated FE-model

Step 2. Removing tetrahedrons of atria and ventricles cavities.

After the first step we have the FE model, which consists of tetrahedrons inside atria and ventricles cavities and we delete them. To remove them we need to mark vertices for deleting. We mark vertices located inside tissue meshes (Figure 1) via the ray casting algorithm (J. D. Foley, 1990). A two-dimensional version of the algorithm is shown in Figure 3.

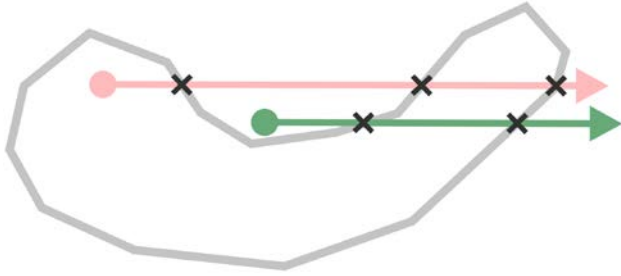


Fig 3 Even-rule algorithm (2D schema). If the point is on the inside of the polygon then it will intersect the edge an odd number of times, and outside otherwise.

We start the ray from a tested vertex in any fixed direction and calculate the number of intersections with heart tissue shells. If the point is outside of the shell the ray will intersect it an even number of times. If the point is inside the shape then the ray will intersect the edge an odd number of times.

Having tested vertices with all shapes, we can delete tetrahedrons consisting of marked vertices. After deleting we obtain the model shown in Figure 4.

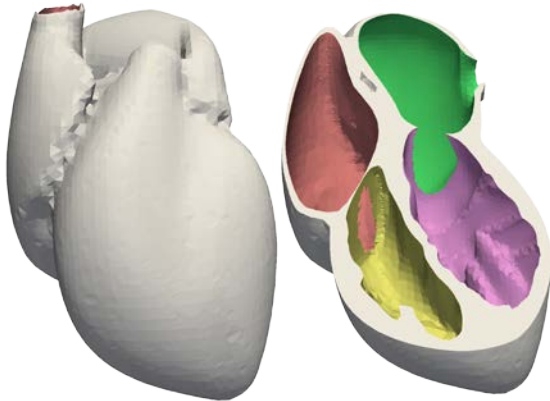


Fig. 4 FE-model with removed cavities

Step 3. Marking up the electrical conduction system of the heart.

To mark up the cardiac conduction system we use the ray casting algorithm for checking whether the point is inside or outside the shell again. As a result we have a model with marked atriums, ventricles and cardiac conduction system.

4. Parallel realization

The brunt of the calculation is marking points according to tissue type. We calculate it with the ray casting algorithm. This algorithm is good for parallel calculations, because we can test a lot of points with the same shell together. Parallel realization of the ray casting algorithm: we divide all vertices into blocks whose size equals size of the GPU thread block and perform calculations for blocks independently; inside the block, the ray casting algorithm is performed for all vertices simultaneously.

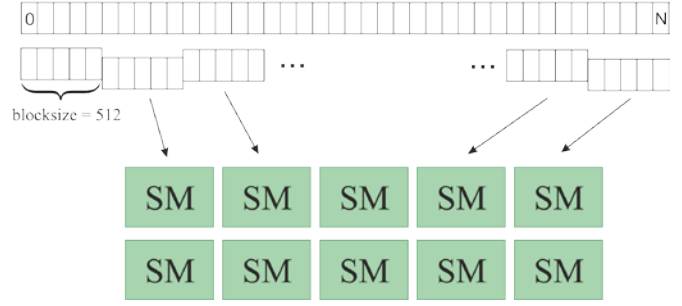


Fig. 5 Data parallelization scheme for the ray casting algorithm. SM - Streaming Microprocessor

Parallelization scheme is shown in Figure 5. All vertices are divided into blocks of the same size. On a GeForce GTX 680 (GK104), the optimal block size was found to be 512 threads. Scheduler allocates vertex blocks for processing on Streaming Multiprocessors automatically.

As the maximum length of the tetrahedron edge decreases, the number of vertices and tetrahedrons and the time for their processing increase exponentially (Figure 6), and it takes more than 6 hours for sequential algorithm to process 2 million vertices.

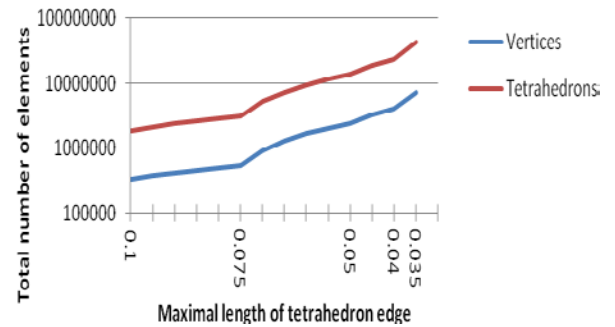


Fig. 6 Ratio between the length of tetrahedron edge and the number of elements. Y axis has logarithmic scale

We used CUDA for the parallel version. To measure the performance, a computer with the following characteristics was used: CPU Intel Core-i7 3820 3.6 GHz and GPU Nvidia GTX 680 4Gb 1006 MHz 1536 cores. The results are shown in the table below. Algorithms were tested on a mesh consisted of 2389529 vertices and 13690149 tetrahedrons.

Table 1: Speed-up of the ray casting algorithm with CUDA

Type of shape	Count of triangles	CPU msec	CUDA msec	Speed-up
AV	328	30995	100	310
SAV	590	54228	165	328
SA	792	82990	223	372
B	830	74979	221	339
F	2454	222461	606	367
RA	2646	247647	645	383
RV	3026	329501	771	427
His	4216	449535	1044	430
LA	5686	473853	1318	359
LV	7180	714429	1735	411
Total	27748	26800618	6828	392

We can see from the table that the speed-up when using the CUDA realization of the algorithm compared to the serial version is about 400x. Accelerating structures will increase the speed of the algorithm even more.

5. Conclusions

Having completed all the steps of the algorithm, we obtain a model of the heart, consisting of elements that belong to one of 10 types of tissue, to simulate the propagation of electrical heart signals along conductive paths of the cardiac system. With an optimized algorithm we can easily update our mesh after each tissue model correction, because the updating process takes some minutes rather than hours as is the case with sequential algorithms.

6. Acknowledgements

The Authors acknowledge Ministry of Education and Science of Russia (Contract № 02.G25.31.0157).

References

- Arevalo HJ, B. P. (2016). Computational rabbit models to investigate the initiation, perpetuation, and termination of ventricular arrhythmia. *Progress in biophysics and molecular biology* , 185–194.
- Aslanidi OV, N. T. (2013). Application of micro-computed tomography with iodine staining to cardiac imaging, segmentation, and computational model development. *. IEEE Trans Med Imaging.* , 32-8.
- Colli Franzone P, G. L. (1998). Spread of excitation in 3-D models of the anisotropic cardiac. *Effects of fiber architecture and ventricular geometry* , 131-71.
- Deng D, J. P. (2012). An image-based model of the whole human heart with detailed anatomical structure and fiber orientation. *. Comput Math Methods Med.* , 2012-16.
- J. D. Foley, A. v. (1990). *Computer Graphics: Principles and Practice. The Systems Programming Series.* Addison-Wesley, Reading.
- Lopez-Perez, A., & Sebastian, R. (2015). Three-dimensional cardiac computational modelling: methods, features and application. *BioMedical Engineering OnLine* .
- P. Lamata et al. (2014). An automatic service for the personalization of ventricular cardiac meshes. *J. R. Soc. Interface* .
- Plastic boy store.* (n.d.). Retrieved august 2017, from Plastic boy anatomy models store: <http://www.plasticboy.co.uk/store/>
- Plotkowiak M, R. B. (2008). High performance computer simulations of cardiac electrical function based on high resolution MRI datasets. *Int Conf Comput Sci 2008* , 571-80.
- R. Bordasa, K. G. (2011). Rabbit-Specific Ventricular Model of Cardiac Electrophysiological Function including Specialized Conduction System. *Prog Biophys Mol Biol* .
- Sermesant M, M. P.-M. (2006). Cardiac function estimation from MRI using a heart model and data assimilation: advances and difficulties. *Med Image Anal* , 642–56.

ILLUSTRATION OF MODEL CREATION ON EXAMPLE OF APPROXIMATIONS TO THE STEADY STATE CURRENT OF CHEMICAL CYCLIC PROCESSES

Assoc. Prof. Dimitrov A.G. PhD

Institute of Biophysics and Biomedical Engineering – Bulgarian Academy of Sciences, Sofia, Bulgaria

e-mail: agd@biomed.bas.bg

Abstract: An approach for creation of biophysically based models for the steady state current of cyclic processes is investigated. When the process (like chemical reactions) can be described by a system of linear ordinary differential equations, an analytic expression for its steady state exists. The analytic expression is especially simple for the current of single cycle processes. In biologic context, concentrations of many substances change in a very restricted (patho)physiological range. This allows neglecting some terms of the analytic expression and thus obtaining biophysically based models that are both simple and adequate for description of currents produced by enzymes, pumps or transporters. The approximations obtained could be reduced to the existing empirical models. A clear way of expanding a specific empirical model for obtaining the desired quality and range of validity is also represented. The described approach is general and can be useful for creating biophysically based models of other types of processes.

KEYWORDS: MATHEMATICAL MODEL, CYCLIC PROCESSES, STEADY STATE CURRENT, ENZYME, TRANSPORTER, PUMP

1. Introduction

To construct an empirical model of some process one has to fit the parameters of a predefined function to some available experimental data [1]. One could construct many functions that are close to each other for some range of data. The choice of a suitable function is complex and requires taking into account specific information. The range of possibilities spans from a lucky guess to a detailed study of the underlying processes. A comparison of various alternative models may illustrate the advantages and the drawbacks of different approaches. The purpose of this paper is to illustrate important aspects of model creation on the example of the most studied chemical cyclic process – the Na⁺/K⁺ ATPase (NKA).

2. Analysis of existing NKA models

2.1. Empirical models

Cyclic processes are abundant in biologic environment. Actions of enzymes, transporters, pumps are all cyclic process. All of them collect some resources, perform the appropriate actions, release the products and then they are ready to start again thus forming a cycle. Typically, one is interested in the cumulative effect of many cycles. As a result, the steady state current or at least the average turning rate is used to characterize the cycle. The first great success was description of the steady state enzyme reaction rate – v , as a function of substrate concentration – $[S]$, performed by Michaelis and Menten (1913) [2] and represented by eq. (1).

$$v = \frac{d[P]}{dt} = \frac{[S] \cdot V_{\max}}{[S] + K_m} \quad (1)$$

where V_{\max} is the maximal reaction rate, $[P]$ is the product concentration, $1/K_m$ is the substrate's affinity.

The Michaelis – Menten relation becomes a basis for many empirical models of protein kinetics. The models of DiFrancesco and Noble [3] eq. (2) and of Luo and Rudy [4] eq. (3) are those for NKA. The Luo and Rudy model has become the standard model of choice for NKA [5-12]. Empirical equations like eq. (2) and (3) are simple, easy to use and to calculate. Unfortunately it is not clear when the approximation is valid, and how to modify the equation to expand the range of validity.

$$I_{\text{pmp}} = P_0 \frac{[K_o]}{[K_o] + p_{K_o}} \frac{[Na_i]}{[Na_i] + p_{Na_i}} \quad (2)$$

$$I_{\text{pmp}} = P_0 \frac{[K_o]}{[K_o] + p_{K_o}} \frac{1}{1 + \left(\frac{p_{Na_i}}{[Na_i]}\right)^{1.5}} * \frac{1}{1 + 0.1245 e^{-0.1 \frac{FV}{RT}} + 0.0365 \sigma e^{-\frac{FV}{RT}}} \quad (3)$$

where P_0 is the maximal current, $\sigma = \frac{1}{7} (e^{\frac{[Na_o]}{67.3}} - 1)$, $[K_o]$, $[Na_i]$ and $[Na_o]$ are external potassium, internal and external sodium concentrations, respectively; p_{K_o} and p_{Na_i} are corresponding affinities; V is membrane potential; F and R are the Faraday and universal gas constants, T is the absolute temperature.

2.2. ODE models

An approach to modeling the processes that is alternative to creation of empirical models is based on studying and description of the underlying chemical reactions. The chemical reactions are usually described as series of transitions between (often unknown) states. Let us denote the probability of the i -th state by $[T_i]$. Each transition between the states i and j , is described by a rate constant – $\alpha_{i,j}$. This results in a system of linear ordinary differential equations (ODE) that potentially fully describes the reaction – eq. (4).

$$\frac{d[T_i]}{dt} = \sum_j \alpha_{i,j} [T_j] \quad (4)$$

To perform direct numerical integration one must provide all the rate constants that describe transitions between the states.

For cyclic processes it is beneficial to distinguish forward rate constants ($\alpha_{i,j}$) and backward rate constants ($\beta_{i,j}$). For NKA, there are at least 30 rate constants (15 forward and 15 backward ones) [13]. However, only some of them have been estimated [14-18]. The most common approach used in this case is the reduction of the number of states in the cycle [13, 19, 20]. Initially, Chapman et al. [21] constructed a 6-state cycle; however, even in this case, some of the 12 rate constants had to be guessed [21]. Later, a complete set of rate constants was obtained for a 4-state NKA cycle only [13, 19, 22].

The rate constants have a close relation with the thermodynamic force that drives the reaction [23, 24]. For the rate constants that form a complete cycle one could write:

$$\frac{\alpha_{1,2}\alpha_{2,3}\dots\alpha_{n-1,n}\alpha_{n,1}}{\beta_{1,2}\beta_{2,3}\dots\beta_{n-1,n}\beta_{n,1}} = \exp(Y/RT) = e^X \quad (5)$$

where Y is the total, associated with the transported substances free energy. Further for clarity we would redefine $X=Y/RT$ and would name X as total (driving) force.

To compensate simplifications, the reduction of the number of states in the cycle should be accompanied by more complex expressions used to define individual rate constants. Attempts were made to increase complexity of the 4-state NKA model as the

simplified models faced problems to reflect inter-tissue and inter-species differences [19, 20].

2.3. Analytic models

In the steady state conditions a system of differential equations is transformed into a system of algebraic ones. This allows an exact analytic description of the steady state [23-25]. Some models are based on this observation [26-29].

The steady state current for a cyclic process looks like rational function with two terms above and n^2 terms below the line, where n is the number of the cycle states. For a 3 – state cycle system the solution is like eq. (6):

$$I_{st} = \frac{\alpha_{1,2}\alpha_{2,3}\alpha_{3,1} - \beta_{1,2}\beta_{2,3}\beta_{3,1}}{(\alpha_{2,3}\alpha_{3,1} + \alpha_{3,1}\beta_{1,2} + \beta_{1,2}\beta_{2,3}) + (\alpha_{3,1}\alpha_{1,2} + \alpha_{1,2}\beta_{2,3} + \beta_{2,3}\beta_{3,1}) + (\alpha_{1,2}\alpha_{2,3} + \alpha_{2,3}\beta_{3,1} + \beta_{3,1}\beta_{1,2})} \quad (6)$$

The general solution for an n-state cycle [23, 24] is a direct expansion of eq. (6):

$$I_{st} = \frac{\alpha_{1,2}\alpha_{2,3}\dots\alpha_{n-1,n}\alpha_{n,1} - \beta_{1,2}\beta_{2,3}\dots\beta_{n-1,n}\beta_{n,1}}{(\alpha_{2,3}\dots\alpha_{n-1,n}\alpha_{n,1} + \alpha_{3,4}\dots\alpha_{n,1}\beta_{1,2} + (n-2) \text{ other terms in a group}) + (n-1) \text{ other groups}} \quad (7)$$

Where, I_{st} as well as the rate constants have dimension of s^{-1} .

Unfortunately the rate constants in eq. (7) are still mainly unknown. The first approach to this problem is to reduce the number of states in the cycle to 2-6 and then to solve the system of ODE analytically. So obtained approximation is further transferred into an analytical function of concentrations and potential [26-28].

Another approach to the problem is also possible [29]. Instead of decreasing the number of states and complicating the individual rate constants, one could do just the opposite. That is to simplify individual rate constants by using the maximal number of states the transporter can occupy. Then, according to the mass action law, those rate constants would have a linear dependence on concentrations:

$$\alpha_k = \alpha'_k [Sb_k] \quad \beta_l = \beta'_l [Sr_l] \quad (8)$$

where α'_k and β'_l do not depend on concentrations, $[Sb_k]$ is concentration of the k-th binding substance, $[Sr_l]$ is concentration of the l-th released substance.

Eq. (8) could transform the terms under the line in eq. (7) into a multidimensional polynomial. Then, combining equations (5, 8 and 7) one could obtain for the current:

$$I_{st} = \frac{e^X - 1}{Ae^X + B} \quad (9)$$

$$A = a_0 \left(1 + \sum \frac{a_{sbk}([S])}{[Sb_k]} + \sum \frac{[Sr_l]}{a_{srl}([S])}\right)$$

$$B = b_0 \left(1 + \sum \frac{b_{srl}([S])}{[Sr_l]} + \sum \frac{[Sb_k]}{b_{sbk}([S])}\right)$$

where a_0 , b_0 have dimension of seconds, $[S]$ represents concentrations of all the substances, $a_{sbk}([S])$, $a_{srl}([S])$, $b_{sbk}([S])$ and $b_{srl}([S])$ have dimension of concentration, X is the driving force [29]. For NKA the driving force would be:

$$X = \frac{dG_{atp}}{RT} + \ln \left(\frac{[ATP]}{[ADP][P][H]} \right) - 3 \ln \left(\frac{[Na_o]}{[Na_i]} \right) + 2 \ln \left(\frac{[K_o]}{[K_i]} \right) + (3z_{na} - 2z_k) \frac{FV}{RT} \quad (10)$$

We see that many different models for a single underlying process could be created.

3. Discussion and recommendations

The steady state current depends on the driving force X, and concentrations of the related substances (eq. 9). Specific approximations would be appropriate in various conditions.

The processes where the driving force is close to zero would be reversible processes, where current could change direction. For them small changes in concentrations could cause large relative changes in the driving force X (eq. 10) and thus in the current. As a result, the effect of the driving force on the current could be significant for reversible processes and should be explicitly present in the model.

For irreversible processes the size of the driving force is significant ($|X| \gg 0$). Then the effect of the driving force on the current would be small (eq. 9). As a result it is often neglected like in the cases of empirical models described above (eq. 2, 3).

The driving force calculation should not be a problem. An explicit presence of the driving force in eq. (9) guarantees that the direction of the current would always be correct. So models that explicitly represent the effect of the driving force would have potentially larger range of validity.

In biologic environment, the concentrations of many substances change in very restricted ranges. This could be combined with the observation that the reaction rate in the Michaelis – Menten relation (eq. 1) has a weak sensitivity to the substrate concentration. The rate is 0.1Vmax when $[S] = K_m/9$, and it is 0.9Vmax when $[S] =$

9Km. In other words an 81-fold increase in substrate concentration is required for increasing the rate from 10% to 90% of the limit [30]. So by neglecting the affinities to substances whose concentrations vary significantly less than 81 times, only a small error would be introduced in a model for the steady state current (eq. 9). If only substances, whose concentrations vary within narrow ranges, are used in the cycle, the current could often be described by a single parameter – its maximal turning rate. This is usually treated as an oversimplification. Therefore, in such cases, the models like those described by eq. (2) and eq. (3) are used, where contributions of the most significant terms are present and contribution of all other terms is neglected.

Concentration of some other substances could vary much more significantly (like 81-fold or similar). That could be intracellular Ca^{++} , pH or other signaling molecules. If that substance is part of the cycle, good reasons should be present to neglect its affinity. That affinity may have complex dependence on concentrations [29].

The processes in the cycle that we have studied so far may have nothing common with the regulation of the cycle. For example, in skeletal muscles, changes in concentration of cyclic adenosine mono phosphate (cAMP) activates the related kinase that in turn, modifies NKA to augment NKA current [31, 32]. Thus to obtain a realistic model, one may need to consider several cycles with possible ligand or voltage gated transitions between them.

A proper system of differential equations could potentially fully describe the chemical processes under study when the necessary states and rate constants are available. However obtaining the states and rate constants could be problematic. When all the substances have rather stable concentrations, the steady state current would be almost equal to its maximal turning rate. Finding many rate constants from what is essentially a single current value is an impossible task. To make situation even more complex external regulation by some signaling molecules or by membrane voltage is likely to be present in that case. To take information on the rate constants, one must perform measurements of the steady state current and of the concentrations of all of the involved substances with great precision. So obtaining rate constants to construct quantitative ODE model may be a very tricky task that is not always possible.

In such complicated cases, when creating quantitative ODE model is not possible, analytic approach that lead to eq. (9) provides a valuable alternative. Thus equation (9), turns out to be a more universal candidate for a quantitative model than a system of ODE. Nevertheless a system of ODE is a great starting point for an analysis of the underlying process [29].

4. Conclusion

Life is more complex than the steady state current of cyclic processes. Models for many processes have to be created. However creation of a useful quantitative model is not an easy task. Different types of models reveal different aspects of the underlying processes. Cyclic chemical processes are sufficiently well studied to illustrate many important aspects of model creation. I hope that the analysis represented here would be helpful in creation of new models.

5. References

1. Press, W.H., et al., *Numerical recipes in C: the art of scientific computing* 1992: Cambridge University Press.
2. Michaelis, L. and M.M.L. Menten, *The kinetics of invertin action*. FEBS letters, 2013. **587**(17): p. 2712-2720.
3. DiFrancesco, D. and D. Noble, *A model of cardiac electrical activity incorporating ionic pumps and concentration changes*. Philos Trans R Soc Lond B Biol Sci, 1985. **307**(1133): p. 353-98.
4. Luo, C.H. and Y. Rudy, *A dynamic model of the cardiac ventricular action potential. I. Simulations of ionic currents and concentration changes*. Circ Res, 1994. **74**(6): p. 1071-96.
5. Wallinga, W., et al., *Modelling action potentials and membrane currents of mammalian skeletal muscle fibres in coherence with potassium concentration changes in the T-tubular system*. Eur Biophys J, 1999. **28**(4): p. 317-29.
6. Pandit, S.V., et al., *A mathematical model of action potential heterogeneity in adult rat left ventricular myocytes*. Biophys J, 2001. **81**(6): p. 3029-51.
7. Bondarenko, V.E., et al., *Computer model of action potential of mouse ventricular myocytes*. Am J Physiol Heart Circ Physiol, 2004. **287**(3): p. H1378-403.
8. Hund, T.J. and Y. Rudy, *Rate dependence and regulation of action potential and calcium transient in a canine cardiac ventricular cell model*. Circulation, 2004. **110**(20): p. 3168-74.
9. Shannon, T.R., et al., *A mathematical treatment of integrated Ca dynamics within the ventricular myocyte*. Biophys J, 2004. **87**(5): p. 3351-71.
10. ten Tusscher, K.H., et al., *A model for human ventricular tissue*. Am J Physiol Heart Circ Physiol, 2004. **286**(4): p. H1573-89.
11. Fortune, E. and M.M. Lowery, *Effect of extracellular potassium accumulation on muscle fiber conduction velocity: a simulation study*. Ann Biomed Eng, 2009. **37**(10): p. 2105-17.
12. Grandi, E., et al., *Interplay of voltage and Ca-dependent inactivation of L-type Ca current*. Prog Biophys Mol Biol, 2010. **103**(1): p. 44-50.
13. Smith, N.P. and E.J. Crampin, *Development of models of active ion transport for whole-cell modelling: cardiac sodium-potassium pump as a case study*. Prog Biophys Mol Biol, 2004. **85**(2-3): p. 387-405.
14. Schulz, S. and H.J. Apell, *Investigation of ion binding to the cytoplasmic binding sites of the Na,K-pump*. Eur Biophys J, 1995. **23**(6): p. 413-21.
15. Schneeberger, A. and H.J. Apell, *Ion selectivity of the cytoplasmic binding sites of the Na,K-ATPase: I. Sodium binding is associated with a conformational rearrangement*. J Membr Biol, 1999. **168**(3): p. 221-8.
16. Holmgren, M., et al., *Three distinct and sequential steps in the release of sodium ions by the Na⁺/K⁺-ATPase*. Nature, 2000. **403**(6772): p. 898-901.
17. De Weer, P., D.C. Gadsby, and R.F. Rakowski, *Voltage dependence of the apparent affinity for external Na⁺ of the backward-running sodium pump*. J Gen Physiol, 2001. **117**(4): p. 315-28.
18. Gadsby, D.C., et al., *The dynamic relationships between the three events that release individual Na⁺ ions from the Na⁺/K⁺-ATPase*. Nat Commun, 2012. **3**: p. 669.
19. Garcia, A., et al., *Kinetic comparisons of heart and kidney Na⁺/K⁺-ATPases*. Biophys J, 2012. **103**(4): p. 677-88.
20. Lewalle, A., S.A. Niederer, and N.P. Smith, *Species-dependent adaptation of the cardiac Na⁺/K⁺ pump kinetics to the intracellular Na⁺ concentration*. J Physiol, 2014. **592**(24): p. 5355-71.
21. Chapman, J.B., E.A. Johnson, and J.M. Kootsey, *Electrical and biochemical properties of an enzyme model*

- of the sodium pump. *J Membr Biol*, 1983. **74**(2): p. 139-53.
22. Oka, C., C.Y. Cha, and A. Noma, *Characterization of the cardiac Na⁺/K⁺ pump by development of a comprehensive and mechanistic model*. *J Theor Biol*, 2010. **265**(1): p. 68-77.
 23. Hill, T.L., *Free energy transduction in biology: the steady-state kinetic and thermodynamic formalism*. 1977, New York: Academic Press. xii, 229 p.
 24. Hill, T.L., *Free energy transduction and biochemical cycle kinetics*. 1989, New York: Springer-Verlag. 119 p.
 25. King, E.L. and C. Altman, *A Schematic Method of Deriving the Rate Laws for Enzyme-Catalyzed Reactions*. *The Journal of Physical Chemistry*, 1956. **60**(10): p. 1375-1378.
 26. Hernandez, J., J. Fischbarg, and L.S. Liebovitch, *Kinetic model of the effects of electrogenic enzymes on the membrane potential*. *J Theor Biol*, 1989. **137**(1): p. 113-25.
 27. Gadsby, D.C. and M. Nakao, *Steady-state current-voltage relationship of the Na/K pump in guinea pig ventricular myocytes*. *J Gen Physiol*, 1989. **94**(3): p. 511-37.
 28. Sagar, A. and R.F. Rakowski, *Access channel model for the voltage dependence of the forward-running Na⁺/K⁺ pump*. *J Gen Physiol*, 1994. **103**(5): p. 869-93.
 29. Dimitrov, A., *An approach to expand description of the pump and co-transporter steady-state current*. *Journal of theoretical biology*, 2017. **412**: p. 94-99.
 30. Cornish-Bowden, A., *Fundamentals of enzyme kinetics*. 2012, Weinheim, Germany: Wiley-Blackwell.
 31. Clausen, T., *The sodium pump keeps us going*. *Ann N Y Acad Sci*, 2003. **986**: p. 595-602.
 32. Clausen, T., *Quantification of Na⁺,K⁺ pumps and their transport rate in skeletal muscle: functional significance*. *J Gen Physiol*, 2013. **142**(4): p. 327-45.

APPLICATION OF FUZZY MODELING TO PREDICT THE DISEASE OF STAFF FROM EXPOSURE TO WORKING CONDITIONS

Klimova I. V., Smirnov Yu. G.

Department of Informatics, Computer Technologies and Engineering Graphics - Ukhta State Technical University, the Russian Federation

bgd4@mail.ru, ysmirnov@ugtu.net

Abstract: A fuzzy model for determining the morbidity rate of employees of a refinery with diseases of the respiratory organs is analyzed on the basis of an analysis of the concentrations of pollutants in all occupational environments using a mathematical apparatus of fuzzy sets. The results of visualization of the developed fuzzy model in the MATLAB Fuzzy Logic Toolbox medium are presented.

Keywords: FUZZY MODELING, MODEL, WORKING CONDITIONS, HARMFUL PRODUCTION FACTORS, MORBIDITY, STAFF

1. Introduction

Fuzzy modeling is the most promising direction for scientific research in the field of analysis, forecasting and modeling of various processes. This is especially important for assessing occupational risks for the health of personnel, where there is insufficient data on the connection of certain diseases with working conditions. The existing assessment of working conditions makes it possible to determine the "verbal" level of risk on the basis of the established class of working conditions. At the same time, the range of production factors under consideration is constantly narrowing, excluding even the receipt of additional payments for harmful working conditions. Thus, with the invariance of working conditions, the class of working conditions is reduced only by changing the methodology by which labor conditions are evaluated.

The purpose of this work is to assess the applicability of existing models and methods of fuzzy logic to modeling the impact of harmful substances on the health of personnel of "RN - Komsomolsk Oil Refinery".

The object of the study is the personnel of the plant exposed to harmful substances in three environments for the period from 2004 to 2010. The subject of the study is inhalation non-carcinogenic risks for the health of the personnel of the oil refinery.

2. Problem discussion

Consider the problem of constructing the dependence of the morbidity of diseases of the respiratory organs of the personnel of the oil refinery on the indices of the non-carcinogenic hazard of chemicals and suspended substances. The nosological group of "respiratory diseases" was chosen on the basis of an earlier analysis of the impact substances and the peculiarities of their effect in three environments: industrial-technological, industrial and urban [1].

The production technological environment is the most polluted - it is the territory of technological workshops and installations. The production environment is a less polluted environment within administrative buildings, as well as the rest of the plant's territory. The urban environment in turn is a combination of household and environmental environments in the city of Komsomolsk-on-Amur. Such a division into environments is made with the aim of clarifying the concentrations of harmful substances, and, accordingly, a more detailed assessment of the health risks of workers [1].

A total of 66 substances participated in the analysis. Figure 1 shows the quantitative ratio of substances affecting personnel. Of them, substances affecting the respiratory system - 30 items, 7 has a primary effect on the respiratory system (Figure 2) [2].

All harmful substances that affect the respiratory system can be divided into two groups: due to their specific effects: chemical substances and suspended substances. Therefore, the model of the dependence of the morbidity of personnel with respiratory diseases on two parameters will be constructed: the index of non-

carcinogenic hazard for chemical substances and the index of non-carcinogenic hazard for suspended substances.

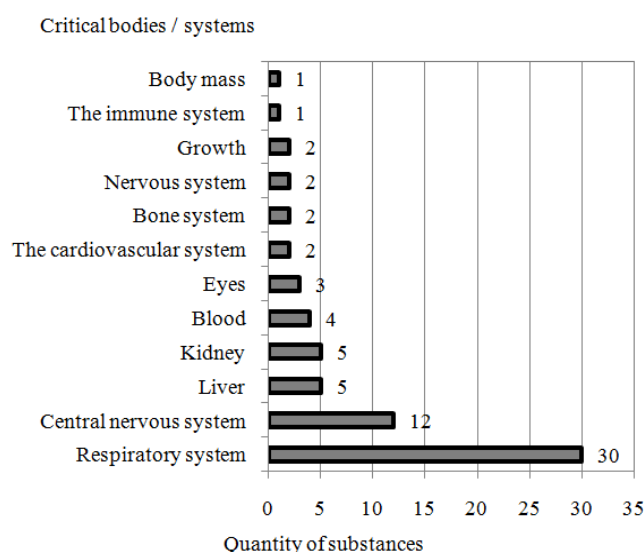


Fig. 1. Ratio of substances affecting critical bodies / systems

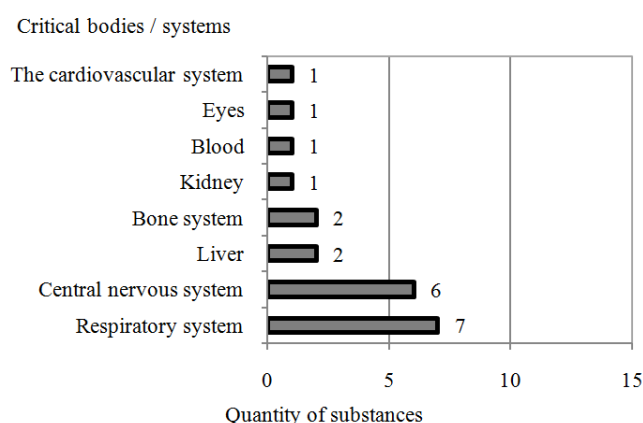


Fig. 2. Ratio of substances according to their primary effect on critical organs / systems

Indices of non-carcinogenic hazards are quantitative estimates of the amount of harmful substances affecting the worker's body throughout the day in three environments and are calculated first for each substance (1) and then for the group of effects on the particular organ or system as a whole (2):

$$(1) HQ_i = AC / RfC,$$

$$(2) HI = \sum HQ_i,$$

where HQ_i are the hazard ratios for the individual components of the mixture of agents; HI - index of non-carcinogenic hazard for a critical organ or system; AC - average concentration of substance, mg/m^3 ; RfC - reference (safe) concentration, mg/m^3 [1].

It is worth noting that the average concentration of each substance involved in the calculation was averaged over 24 hours, depending on each type of medium, the concentration of the substance in it, and the time of exposure to the worker.

After the HI indices for respiratory organs were calculated at each workplace, the data were averaged over 29 occupational groups. The main criteria for the formation of groups of personnel: belonging to the facility or shop and the value of the received non-carcinogenic risk HI .

The minimum values of the indices were $HI_{min\ chem} = 23.87$ and $HI_{min\ dust} = 2.56$, which exceeds the permissible value $HI_{norm} = 1$. The maximum values of the indices were $HI_{max\ chem} = 1621.58$ and $HI_{max\ dust} = 5.01$.

Next, consider such an output parameter as the state of health of personnel, which is a reflection of a complex set of phenomena in the environment. The process of its formation is influenced by a number of industrial, socio-economic, as well as biological, anthropogenic, natural climatic and other factors that together determine the ecological environment in which the person is during the day, food and water. Most of the harmful substances a person receives with inhaled air, this is about 80% of all intake doses.

Anticipate the corresponding incidence rate of personnel with respiratory diseases is quite difficult, especially with the seasonality of this phenomenon. In this connection, it is expedient to use linguistic variables, that is, variables whose values are not numbers, but words in natural or formal language.

The data on the morbidity of the personnel of the plant on the nosological form of "respiratory disease" from 2004 to 2010 were broken down into the allocated 29 occupational groups, and then averaged over 7 years and included the number of cases per 1000 workers and the number of days of incapacity for work 1000 working. Further, data are used only for the number of cases of diseases: the minimum value is 125 diseases, the maximum is 495.

Comparison of calculated HI indices and morbidity leads to the receipt of 29 points, which are difficult to describe with the help of the equations of dependence, but nevertheless, a directly proportional relationship is clearly visible: the greater the hazard index, the greater the response.

For the transition to fuzzy logic, an additional conversion of the input parameters (HI indices) was performed. Since all indices exceeded the allowable value, an attempt was made to select a so-called "acceptable" value that would guarantee minimal deviations in the state of health.

For this purpose, an additional criterion "severity of disease progression" was introduced and a corresponding group was chosen whose indices are accepted for an "acceptable" new standard. This made it possible to obtain new ranges of values of $HI_{chem} \in [0.12, 8.05]$ and $HI_{dust} \in [1.0, 1.96]$.

3. Methodology

The modeling process can be presented in large form in the following sequence of actions:

- 1) awareness of the problem;
- 2) highlight the main factors that determine the problem, which should serve as output parameters of the model;
- 3) highlighting the defining input variables of the model;
- 4) the actual development of a mathematical model;
- 5) identification of the model (parametric or structural-parametric);
- 6) conducting numerical experiments with the model and, if necessary, statistical processing of the obtained data;

7) determining the composition of the quality parameters that characterize the problem being solved (using simulation data and a priori information);

8) formalization of particular quality criteria based on quality parameters;

9) determination of parameters characterizing the relative importance of particular criteria for solving a common problem;

10) formalization of the generalized quality criterion for solving a problem on the basis of aggregation of particular criteria, taking into account their relative importance;

11) solving the problem of choosing the best alternative or multi-criteria optimization, depending on the type of problem being solved.

In essence, it is some detail of the generally accepted scheme: the formulation of the problem \rightarrow model building and identification \rightarrow optimization.

The development of fuzzy models of sanitary and toxicological safety of the personnel of industrial enterprises makes it possible to obtain a numerical estimate of occupational risk. The mathematical apparatus of fuzzy logic is usually used in those cases when the available quantitative information is insufficient, or it is not complete enough to obtain reliable statistically significant conclusions [3-7].

Along with classical analytical methods, it is advisable to use the fuzzy set device implemented in particular in the MATLAB computer simulation system [8], which allows developing a fuzzy-multiple model for the estimation, analysis and visualization of professional risk indicators.

Since there is a need to take into account the multitude of indicators that are dissimilar in physical nature and dimensions, it is advisable to bring them to a dimensionless form by rationing, for example, as follows:

$$(3) s = \frac{s_i - s_{min}}{s_{max} - s_{min}},$$

where s_i is a normed index; s_{max} , s_{min} - the maximum and minimum value of the criterion in the sample according to the normed indicator.

The system of fuzzy inference in the general case includes the following stages:

1. Phasing (reduction to fuzziness). At this stage, the exact set of input data is converted into a fuzzy set, which is determined using membership functions.

2. Construction of the base of rules for fuzzy products.

3. Composition using aggregation methods.

4. Dephasing (reduction to clarity). At the stage of dephasing, the fuzzy system's executive module, on the basis of many fuzzy conclusions, forms an unambiguous decision with respect to the input variables.

Let us consider in more detail the initial stage on which phasing is carried out. Denote by d the input variable "suspended matter", which reflects the dustiness of the company's air environment. The corresponding term-set will be denoted by:

$$T1 = \{\text{low, medium, high}\} = \{D1, D2, D3\}.$$

The second input variable x - "chemical substances" - reflects the chemical contamination of the air environment. It corresponds to the analogous term set:

$$T2 = \{\text{low, medium, high}\} = \{X1, X2, X3\}.$$

The output variable y (the level of morbidity of workers) is also comparable to the analogous term set:

$$T3 = \{\text{low, medium, high}\} = \{Y1, Y2, Y3\}.$$

The next stage is the construction of a database of rules for fuzzy products. Most often, the Mamdani model is used as a model of fuzzy inference, the feature of which is the fact that its rules of inference contain fuzzy meanings in its consecutive clauses. In our case this is the membership function of the term-set $T3$.

In the chosen notation, we give for example some of these rules:

IF d IS $X1$ AND x IS $D1$ THEN y IS $Y1$

IF d IS $X2$ AND x IS $D1$ THEN y IS $Y2$

IF d IS $X3$ AND x IS $D1$ THEN y IS $Y3$

As a tool for implementing this approach, it is convenient to use the Fuzzy Logic Toolbox extension of the MATLAB computer mathematics environment, which allows creating fuzzy inference and fuzzy classification systems. The main interactive tool of the Fuzzy Logic Toolbox is the FIS inference editor, which contains tools for the functional mapping of input and output variables [8].

4. The discussion of the results

The stages of the simulation are illustrated in Figures 3-7 below.

Figure 3 shows a view of a software window with a schematic model.

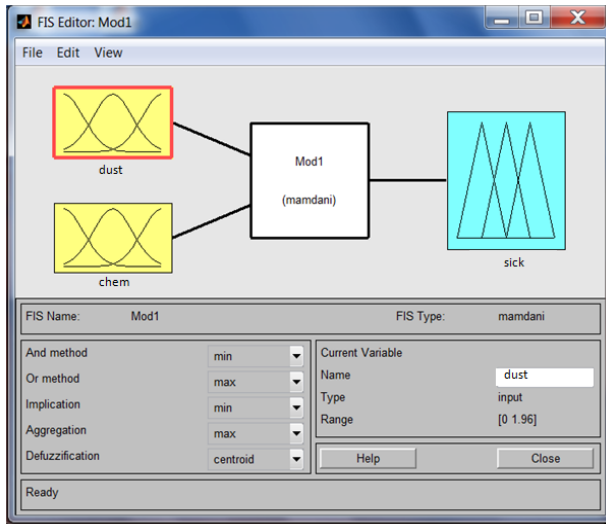


Fig. 3. Software window view

Figure 4 shows the graphs of the membership functions for the input terms "dust" (d), "chemical" (x) and output "sick" (y) of linguistic variables.

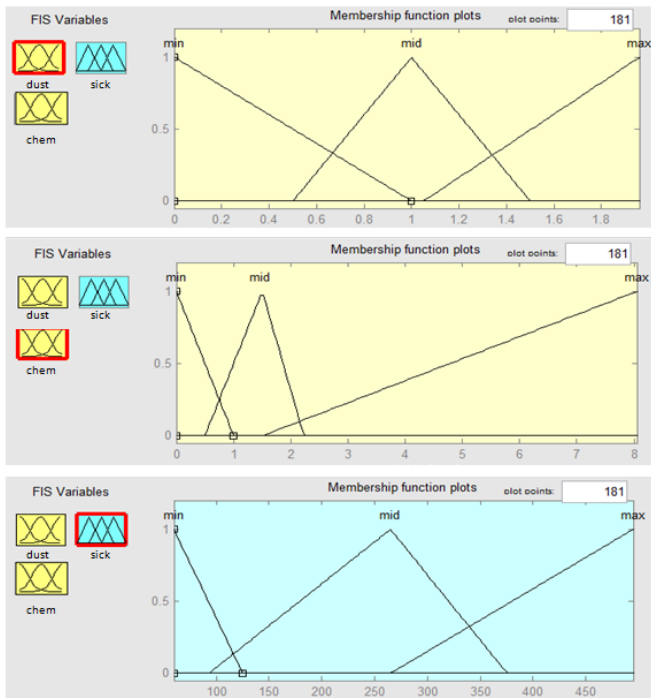


Fig. 4. Type of membership functions in the FIS editor of the Fuzzy Logic Toolbox extension package: "dust" and "chemical" - for indicators of non-carcinogenic risk for suspended and chemical substances, respectively; "sick" - for the number of cases of respiratory diseases per 1000 workers

Figure 5 shows a fragment of the window for determining the base of production rules for fuzzy inference.

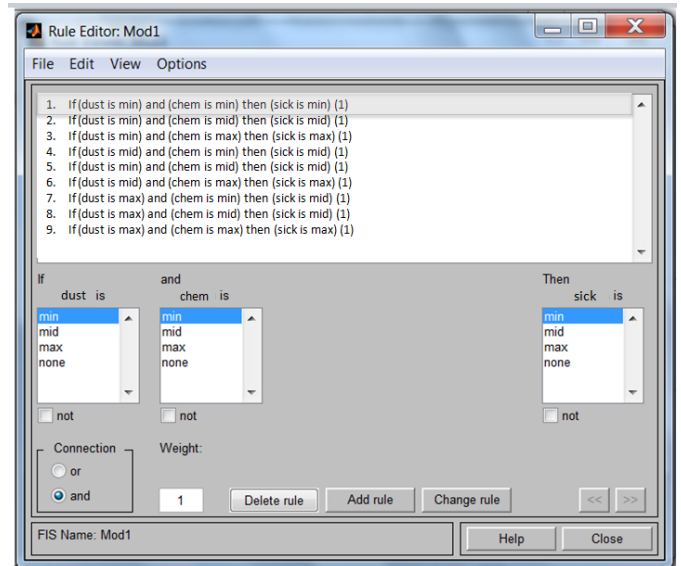


Fig. 5. The form of the product rules definition window in the FIS editor

Figure 6 shows the view of the fuzzy output viewer in the Mamdani model for the problem under consideration. Here, the aggregation of fuzzy rules is shown for two input variables, "sick" and "chemical". This uses a logical product, which corresponds to the operation min. Aggregation of implication concerning rules is carried out by logical summation, which corresponds to the operation max.

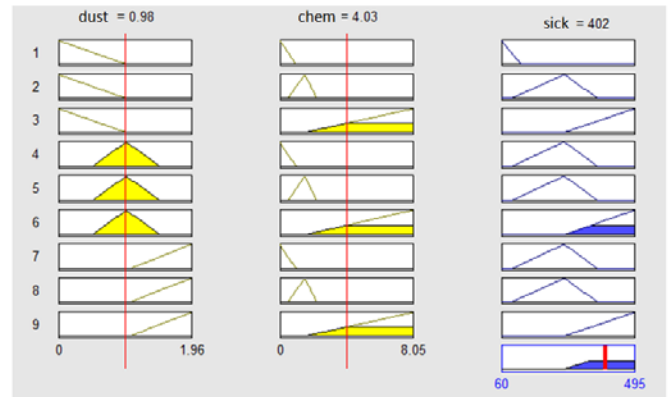


Fig. 6. View of Mamdani's fuzzy output viewer

Since the max operator is used as an aggregation operator, and the min operator is used as an implication operator, the procedure for obtaining a fuzzy output value is a composition of max-min.

After receiving a fuzzy output (y), it is necessary to go to the phase of dephasing, which has the corresponding clear value y_{out} (4). As a method of dephasing, we used the method of the center of gravity:

$$(4) \quad y_{out} = \frac{\sum_{i=1}^n y_i \mu(y_i)}{\sum_{i=1}^n \mu(y_i)},$$

where $\mu(y_i)$ is the membership function of the i -th rule, and n is the number of fuzzy products rules.

Finally, Figure 7 shows the surface of the fuzzy output for the developed fuzzy model. This type serves for a general assessment of the adequacy of the constructed fuzzy model, and also allows analyzing the influence of the values of input variables on the value of the output variable.

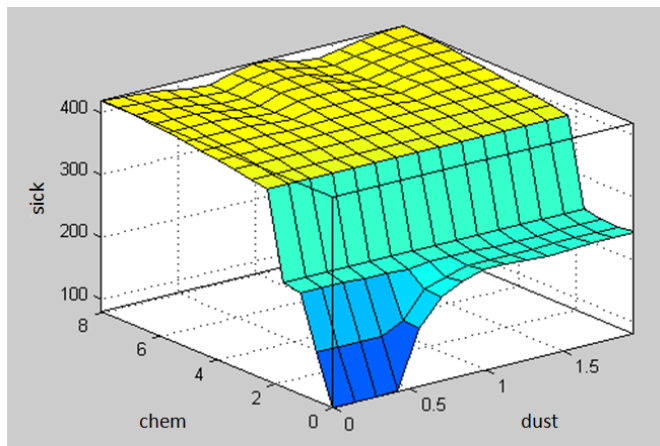


Fig. 7. The surface of fuzzy output for the developed fuzzy model

5. Conclusion

1. There are 29 professional groups of plant personnel, the data on which are converted into 2 input parameters d (suspended substances), x (chemical substances) and 1 output parameter y (the incidence of personnel with respiratory diseases).

2. A fuzzy model of the influence of pollutants of two groups on the health of personnel based on the Mamdani model, namely, on the respiratory system, has been developed and mathematically described.

3. A variant of phasing the three identified parameters based on an acceptable risk to staff health is suggested as "low", "medium", "high".

4. In the interactive mode, the fuzzy output system of the solved task is developed and visualized using the graphical tools of the Fuzzy Logic Toolbox extension package of the computer mathematics environment MATLAB.

5. It has been revealed that all personnel of the plant are exposed to a sufficiently large exposure to non-carcinogenic risk, the likelihood of harmful effects on the worker (respiratory diseases) increases in proportion to the increase in the coefficient of non-carcinogenic HI hazard.

6. The developed model can be easily supplemented by new indicators of air pollution or other production factors (linguistic variables) and new output parameters (fuzzy inference rules).

6. Literature

1. Assessment of non-carcinogenic risk for professional groups of "RN-Komsomolsk Oil Refinery" / I.V. Afanasyeva, V.V. Anisimov // Scientists notes Komsomolsk-on-Amur State Technical University. Sciences about nature and engineering, 2012. № II-1 (10). P. 89-96.

2. Identification of hazards in the assessment of health risks of personnel of RN-Komsomolsk Oil Refinery / I.V. Afanasyeva, V.V. Anisimov // Materials of the 11th International Scientific and Practical Conference in the Field of Ecology and Safety life activities "Far Eastern Spring - 2011". Komsomolsk-on-Amur : GOUVPO "KnAGTU", 2011. P. 141-149.

3. Diligenskiy N.V., Dymova L.G., Sevastyanov P.V. Fuzzy modeling and multicriteria optimization of production systems in conditions of uncertainty: technology, economics, ecology. M. : "Publishing Machine-Building-1", 2004. 398 p.

4. Mazorchuk M.S., Simonova K.A., Grekov L.D. Application of methods and models of fuzzy logic for modeling of economic processes // Systems of information retrieval, 2007. № 9 (67). P. 159-162.

5. Rogachev A.F., Melikhova E.V. Fuzzy-plural modeling and assessment of ecological safety of agricultural lands with radiation contamination // Global Nuclear Security, 2016. № 1 (18). P. 7-18.

6. Sereda S.N. Assessment of environmental risk through fuzzy models // Mechanical engineering and life safety, 2013. № 3. P. 15-20.

7. Tindova M.G. Fuzzy model of economic assessment of environmental damage // Economics: yesterday, today, tomorrow, 2012. № 3-4. P. 129-139.

8. Leonenkov A.V. Fuzzy modeling in MATLAB and fuzzyTECH. S-Pb. : BHV-Petersburg, 2005. 736 p.

EXACT RECONSTRUCTION VERSION OF RADON TRANSFORMATION IN TOMOSYNTHESIS

Morgun O., Nemchenko K., Vaisburd A. *, Viktynska T*.
 "Radioprom" LTD, Dostoevskogo Str. 1, Kharkov, Ukraine 61011
 *Karazin Kharkov National University, Svobody Sq.4, Ukraine 61022

Abstract: The purpose of this study is a method of exact reconstruction in the Radon problem, which consists in refusal of the approximate transformation kernel usage. A comparison of the methods that are currently used in tomosynthesis was conducted. Model experiments were performed; the results of the application of the proposed method in real tomography studies in tomography are given.

The traditional methods of Radon transformation [1] do not make it possible to accurately reconstruct the initial distribution function due to the divergence of the integral expression. In this

paper, we suggest alternative method aiming to get rid of this defect by changing the order of integration using re-grouping the integrand functions in the result of transformation:

$$f(x, y) = \int \frac{|Q|dQd\alpha dX}{(2\pi)} F(X, \alpha) \exp(-iQ(X - x \sin \alpha + y \cos \alpha)). \quad (1)$$

Here $f(x, y)$ is the desired density distribution in the layer, $F(X, \alpha)$ is the projections made under angle α , and X is the coordinate in the detector. Usual way to use (1) is introduce function (filter)

$$G(z) = \int \frac{|Q|dQ}{(2\pi)} \exp(-iQz) = \frac{1}{\pi} \int_0^\infty |Q| \cos(Qz) dz, \quad (2)$$

which describes the connection between $f(x, y)$ and $F(X, \alpha)$:

$$f(x, y) = \int d\alpha dX F(X, \alpha) G(X - x \sin \alpha + y \cos \alpha). \quad (3)$$

The main problem here is that the integral (2) contains a singularity and, therefore, approximate functions [2] are used instead of it, which reflect the main features of the function (2), but lead to an approximate recovery.

In this paper, we propose to change the order of integration to avoid this problem. In this case none of the integrals into this

expression (1) diverges within the limits. As a result of taking the integral in this way, we eventually have to obtain the distribution function equivalent to the original one.

Let's consider the case where the initial distribution function is the delta function, which is determined by the following expression

$$f(x, y) = \rho_0 a^2 \delta(x - x_0)(y - y_0) \quad (4)$$

which corresponds to the projection:

$$F(X, \alpha) = \rho_0 a^2 \delta(x_0 \sin(\alpha) - y_0 \cos(\alpha) - X) \quad (5)$$

Whatever divergences arise in the integral expression (1), we shall not get it to the form containing the kernel of the Radon transformations (2) in explicit form, but immediately substitute the expression (5) into it. Thus, the relation (1), taking into account the integration limits, can be written in the form:

$$f(x, y) = \frac{\rho_0 a^2}{(2\pi)^2} \int_{-\infty}^{+\infty} Q dQ \int_0^\pi d\alpha \int_{-\infty}^{+\infty} dX \delta(x_0 \sin(\alpha) - y_0 \cos(\alpha) - X) * e^{-iQX} e^{iQ(x \sin(\alpha) - y \cos(\alpha))} \quad (6)$$

Now integrating

this expression first with dX , getting rid of the delta functions. As a result, we get:

$$f(x, y) = \frac{\rho_0 a^2}{(2\pi)^2} \int_{-\infty}^{+\infty} Q dQ \int_0^\pi d\alpha e^{iQ((x-x_0) \sin(\alpha) - (y-y_0) \cos(\alpha))} \quad (7)$$

Now we return to other integration variables to avoid divergence and make integration possible in a different order. It is logical to go over to the variables (\vec{k}_x, \vec{k}_y) , for this we write the Jacobian of the transition:

$$dQd\alpha = \left| \begin{array}{c} \widetilde{k}_x = Q \sin(\alpha) \\ \widetilde{k}_y = Q \cos(\alpha) \end{array} \right| = d\tilde{k}_x d\tilde{k}_y \frac{\partial(Q, \alpha)}{\partial(k_x, k_y)} = \frac{d\tilde{k}_x d\tilde{k}_y}{Q}, \quad (8)$$

taking this into account, expression (7) can be written as follows:

$$f(x, y) = \frac{\rho_0 a^2}{(2\pi)^2} \int_{-\infty}^{+\infty} e^{i\tilde{k}_x(x-x_0)} d\tilde{k}_x \int_0^\infty e^{-i\tilde{k}_y(y-y_0)} d\tilde{k}_y. \quad (9)$$

It is easy to see that the two integrals in the expression are just the Fourier representation of the delta function, similar to the representation:

$$\delta(x - x_1) = \int \frac{dQ}{2\pi} e^{-iQ(x-x_1)} \quad (10)$$

Thus, we get:

$$f(x, y) = \frac{\rho_0 a^2}{(2\pi)^2} 2\pi \delta(x - x_0) 2\pi \delta(y - y_0) \quad (11)$$

Hence, we get the answer:

$$(x, y) = \rho_0 a^2 \delta(x - x_0) \delta(y - y_0) \quad (12)$$

We can see that this function completely identical to the function (4) introduced by us, moreover, we had no need to enter redefinitions anywhere, since all the transitions were initially equal. As a result, we showed that the delta distribution function can be accurately reconstructed by the proposed method.

After the prove above that the delta function can be accurately reconstructed, natural to assume that any other function can also be reconstructed without loss of precision. We can prove it

by taking into account that any distribution function can be represented as a continuous set of delta functions, and all the resulting interim expressions are additive quantities, which gives the right to put the sum or integral sign before the reconstructed distribution function.

This explanation can be considered as self-explanatory. Nevertheless, we strictly prove this statement without referring to the result already obtained. To do this, we return to the previously obtained formula (1) and change the order of integration:

$$f(x, y) = \frac{1}{(2\pi)^2} \int_0^\infty |Q| dQ \int_0^\pi d\alpha \int_{-\infty}^{+\infty} e^{-iQX} e^{iQx \sin(\alpha) - iQy \cos(\alpha)} F(X, \alpha) dX, \quad (13)$$

now the integration over dQ will be performed last.

To converge this integral, we recall the linear distribution

$$F(X, \alpha) = \int f(x, y) dl, \quad (14)$$

and represent

$f(x, y)$ through the delta function as follows:

$$f(x, y) = \int \int d\tilde{x} d\tilde{y} \delta(x - \tilde{x}) \delta(y - \tilde{y}) f(\tilde{x}, \tilde{y}), \quad (15)$$

Then we use the following presentations

$$F(X, \alpha) = \int dx dy \delta(x \sin(\alpha) - y \cos(\alpha) - X) f(x, y) \quad (16)$$

$$dl = dx dy \delta(x \sin(\alpha) - y \cos(\alpha) - X) \quad (17)$$

Putting this into (13) we get:

$$(x, y) = \frac{1}{(2\pi)^2} \int_0^\infty |Q| dQ \int_0^\pi d\alpha \int_{-\infty}^{+\infty} dX e^{-iQX} e^{iQx \sin(\alpha) - iQy \cos(\alpha)} \int \int \int \int d\tilde{x} d\tilde{y} \\ * \delta(x - \tilde{x}) \delta(y - \tilde{y}) f(\tilde{x}, \tilde{y}) dx dy \delta(x \sin(\alpha) - y \cos(\alpha) - X) \quad (18)$$

Integrating by $d\tilde{x} d\tilde{y}$, we get:

$$f(x, y) = \frac{1}{(2\pi)^2} \int_0^\infty |Q| dQ \int_0^\pi d\alpha \int_{-\infty}^{+\infty} dX e^{-iQX} e^{iQx \sin(\alpha) - iQy \cos(\alpha)} \int \int d\tilde{x} d\tilde{y} f(\tilde{x}, \tilde{y}) \\ * \delta(\tilde{x} \sin(\alpha) - \tilde{y} \cos(\alpha) - X), \quad (19)$$

then integrate over dX and get:

$$f(x, y) = \frac{1}{(2\pi)^2} \int_0^\infty |Q| dQ \int_0^\pi d\alpha e^{iQx \sin(\alpha) - iQy \cos(\alpha)} \iint e^{-iQ\tilde{x} \sin(\alpha) + iQ\tilde{y} \cos(\alpha)} f(\tilde{x}, \tilde{y}) d\tilde{x} d\tilde{y} \quad (20)$$

Now we make the transition to new variables completely analogous to the transition (8), so we easily receive from (7) the expression:

$$f(x, y) = \frac{1}{(2\pi)^2} \int dk_x dk_y e^{ik_x(x-\tilde{x})} e^{ik_y(y-\tilde{y})} \iint f(\tilde{x}, \tilde{y}) d\tilde{x} d\tilde{y} \quad (30)$$

Let's consider the representation of the delta function again:

$$\delta(x - x_1) = \int \frac{dQ}{2\pi} e^{-iQ(x-x_1)} \quad (31)$$

Applying it we get:

$$f(x, y) = \iint d\tilde{x} d\tilde{y} \delta(x - \tilde{x}) \delta(y - \tilde{y}) f(\tilde{x}, \tilde{y}) \quad (32)$$

Comparing this result, we can see that this expression is fully equivalent to the expression (15).

Finally, we conclude that the proposed method absolutely allows reconstructing the initial distribution, regardless of what function it describes. Let us highlight, that this proof doesn't contain any reference to the result of the reconstruction of the delta-like distribution function, so the proof done in the previous section can be considered as a particular case.

REFERENCES

1. Radon, Johann (1917), "Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser

Mannigfaltigkeiten", *Berichte über die Verhandlungen der Königlich-Sächsischen Akademie der Wissenschaften zu Leipzig, Mathematisch-Physische Klasse [Reports on the proceedings of the Royal Saxonian Academy of Sciences at Leipzig, mathematical and physical section]*, Leipzig: Teubner (69): 262–277; Translation: Radon, J.; Parks, P.C. (translator) (1986), "On the determination of functions from their integral values along certain manifolds", *IEEE Transactions on Medical Imaging*, 5 (4): 170–176, PMID 18244009, doi:10.1109/TMI.1986.4307775.

2. Herman, Gabor T. (2009), *Fundamentals of Computerized Tomography: Image Reconstruction from Projections* (2nd ed.), Springer, ISBN 978-1-85233-617-2

STOCHASTIC COMPUTER SIMULATION OF THE IONIC DIFFUSION THROUGH BIOLOGICAL TISSUES UNDER THE EFFECT OF DIRECT ELECTRIC FIELD

MSc. Fawzy HM¹, Associate Prof. Dr. Salem NM², Prof. Dr. El Sheikh SM¹ and Prof. Dr. El-Messierey MA²

Department of Physics, American University in Cairo, AUC Avenue, P.O. Box 74, New Cairo 11835, Egypt¹

Department of Engineering Mathematics and Physics, Faculty of Engineering, Cairo University, Giza 12613, Egypt²

nmsalem@aucegypt.edu

Abstract: The effect of the exposure of biological tissues to external electric fields is still a potent source of controversy. This work addresses the diffusion of ions under the effect of an external DC electric field. This was done by studying the diffusion coefficient D as an indicating parameter for such effects. The work was based on a stochastic computer simulation in which the tissue was considered as a matrix containing the elements under study. The size of the matrix was up to $30,000 \times 30,000$. A two-dimensional honey comb cellular pattern was simulated such that it allowed six maximum possible element-to-element communications. The effect of vacancy concentration and annealing time were tested firstly in the absence of electric field. Then different values of the electric field were applied. Moreover, different vacancy concentrations were studied under the effect of the electric field. The results showed that the ionic penetration increases proportionally with the strength of the electric field as well as the percentage of the available vacancies in the host medium.

Keywords: IONIC DIFFUSION, BIOLOGICAL TISSUES, STOCHASTIC SIMULATIONS

1. Introduction

Many theoretical efforts have been devoted to studying diffusion in different media. Electric fields enhance the diffusion of charge carriers in disordered materials. It was shown that the diffusion coefficient in one-dimensional hopping depends linearly on the electric field, while in three dimensions the dependence is quadratic [1].

Diffusion takes place in biological tissues as well. The effective diffusion coefficient, D_{eff} , of salt into a biological tissue due to the application of an electric field at different temperatures showed an increase of ion transport [2]. The effective diffusion coefficients of K^+ and Cl^- ions are appreciably reduced in narrowing the channel in the cell membrane when subjected to an external field. The extent of the reduction is similar for both the anionic and cationic species [3].

Diffusion in the extracellular space (ECS) is constrained by the volume fraction, hence a modified diffusion equation was proposed to govern the transport behavior of many molecules in the brain [4].

Liu and Shi [5] developed two-dimensional and three-dimensional FEM model to study the transport of ionic species in an externally applied electric field. They found that more chlorides are driven out of samples with increasing direct current density and treatment time.

In the present work we introduce a stochastic model to follow and determine the diffusion coefficient of an ionic tracer through a biological tissue. Hence, the effect of the application of an external DC field on the diffusion and the penetration depth of these ions in biological tissue is studied.

2. Computer Model

The present model simulates the ionic diffusion under the following assumptions: A part of the biological tissue is represented as a 2D matrix of sizes up to $30,000 \times 30,000$ elements. Each element represents either a host particle, a vacancy or a tracer ion. The diffusants (tracers), are considered as positive ions.

At zero time, the diffusants occupy the first row of the matrix and are in a continuous flow; each particle that leaves the surface

into the matrix is replaced by another one. The rest of the matrix is either occupied by the host elements (biological cells) or with vacancies that are randomly distributed throughout the matrix. The characteristics of each element are represented by one byte which contains information of the type (host, vacancy or tracer), spatial location and the time elapsed since diffusion starts. The biological tissue could be simply modeled as a close-packed spherical array of cells as shown in Fig. 1.

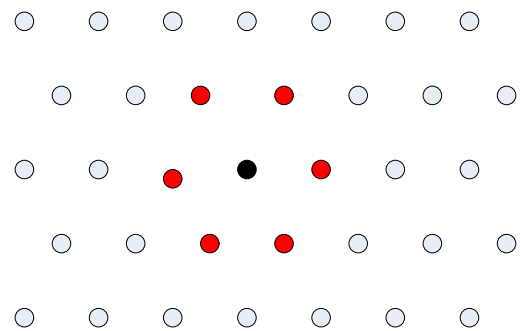


Fig. 1. The six neighbors' structure in a hexagonal matrix modelling a biological tissue. The middle black element is the tracer with six neighboring vacancies represented by red circles.

Since the diffusion process follows a random walk procedure, in the present model the tracer ion diffuses through the vacancies available in the nearest sites such that the jump follows a random choice of the accessible vacancies. The effect of the electric field is represented as a controlled bias in the jump direction. The field direction affects the randomization process of jumping so that along the direction of the field there is a higher probability than other directions. Fig. 2 shows a visualization of the system under consideration, for both cases of free diffusion and the diffusion under the effect of a DC electric field.

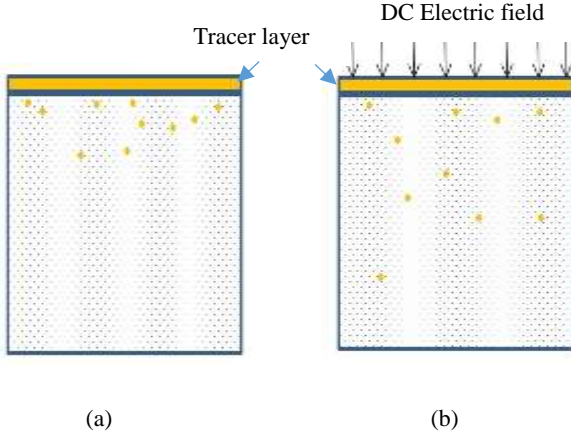


Fig. 2. The diffusion process (a) in the absence of external field and (b) under the effect of DC electric field. The traces are represented by the larger circles.

Hence the concentration $c(x,t)$ of the ions/tracer after a certain time t and a location x is obtained from the following

$$c(x,t) = \text{constant} \times e^{\frac{-x^2}{4Dt}}$$

The mean squared radius of the diffusion pattern is represented by:

$$\langle R^2 \rangle = \sum_{i=1}^N R_i^2 / N$$

where N is the number of jumps and R_i is the individual displacement.

The diffusion coefficient D can be calculated from the slope of $\langle R^2 \rangle$ versus the annealing time t :

$$D = \langle R^2 \rangle / 4t$$

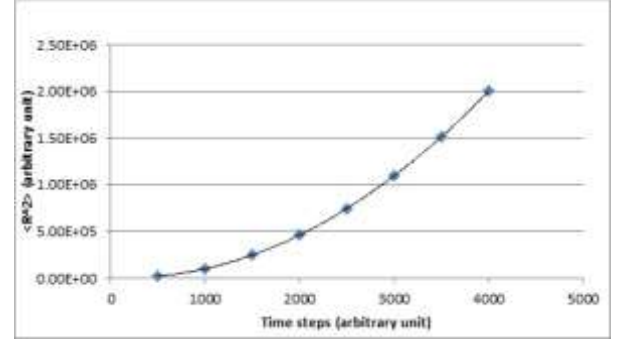
The concentration is calculated by sectioning the matrix to a certain number of rows and calculating the number of the tracer ions in each section. As the thickness of the layers becomes smaller, the accuracy of the penetration profile which describes the diffusion increases. Plotting the logarithm of the concentration $c(x,t)$ against x^2 for different time steps and different vacancy concentrations is used to obtain the diffusion coefficient D according the equation

$$\text{Slope} = -1/4Dt$$

The annealing time is taken as the total number of iterations which is varied up to 500,000 time steps. The vacancies are randomly distributed, and their concentration is varied from 5% to 80%.

3. Results

Firstly, the diffusion pattern is investigated in the absence of an external field and assuming that all the sites are vacant. Figs. 3 and



4 show the relation of the mean squared displacement, $\langle R^2 \rangle$, and the diffusion coefficient, D , as a function of time.

Fig. 3. Variation of the mean squared radius and time steps in the case of free diffusion.

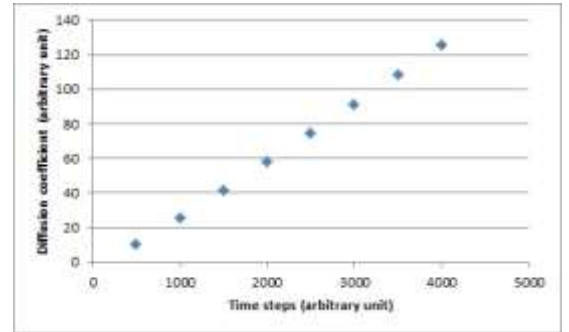


Fig. 4. The diffusion coefficient with time steps.

In both figures, the classical pattern of the random walk diffusion in the absence of external field is preserved. The diffusion coefficient increases linearly with annealing time. The penetration depth $\langle R^2 \rangle$ of the tracers is studied in the presence of randomly distributed vacancies of different ratios in a biological tissue. Fig. 5 shows the increase in the penetration distances of the tracers with annealing time at a vacancy concentration of 50%. In this figure the time steps are increased up to 1.2 millions so that the asymptotic level of the curve is obtained; at this level the penetration distance is about to reach the boundaries of the matrix.

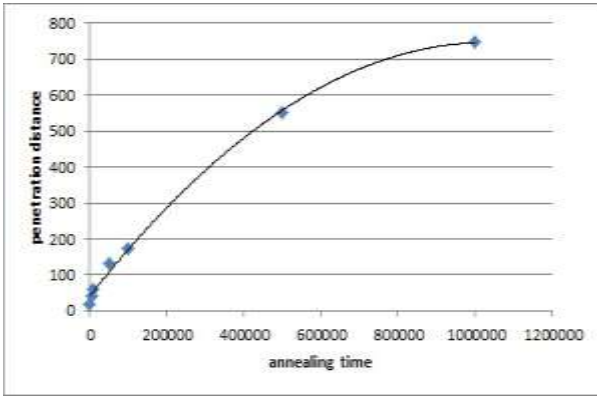
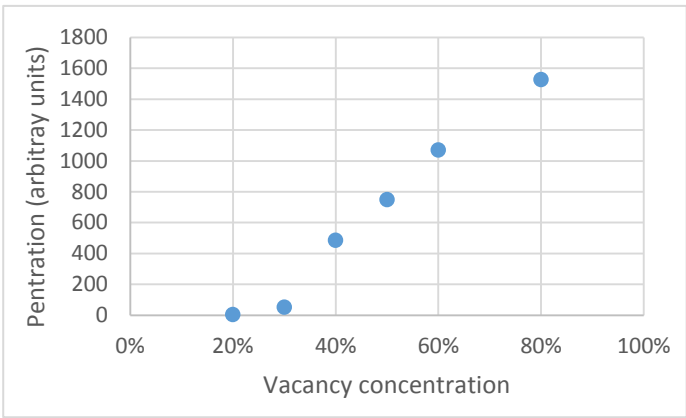


Fig. 5. The variation of penetration distance (arbitrary units) with the annealing time.

Fig. 6 show the mean penetration distance with vacancy concentration. It is obvious that at low vacancy percentage up to 20% , the tracers hardly invade the host matrix. Then, $\langle R^2 \rangle$ increases almost linearly. Similar behavior is observed when the diffusion coefficient is plotted versus the vacancy concentration. Fig. 7 illustrates this relation in absence of external field and for a



fixed annealing time.

Fig. 6. The variation of the penetration with vacancy concentration.

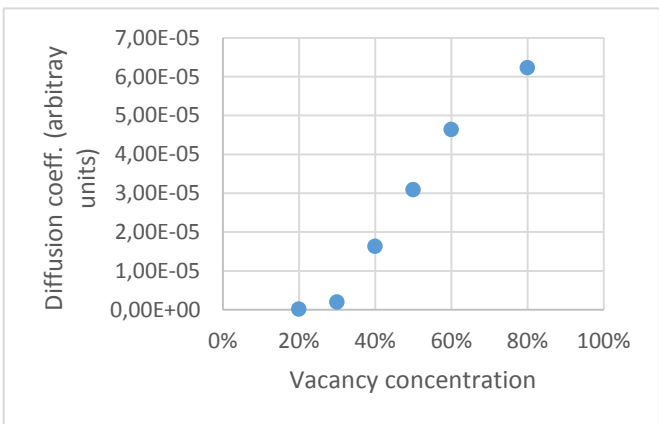
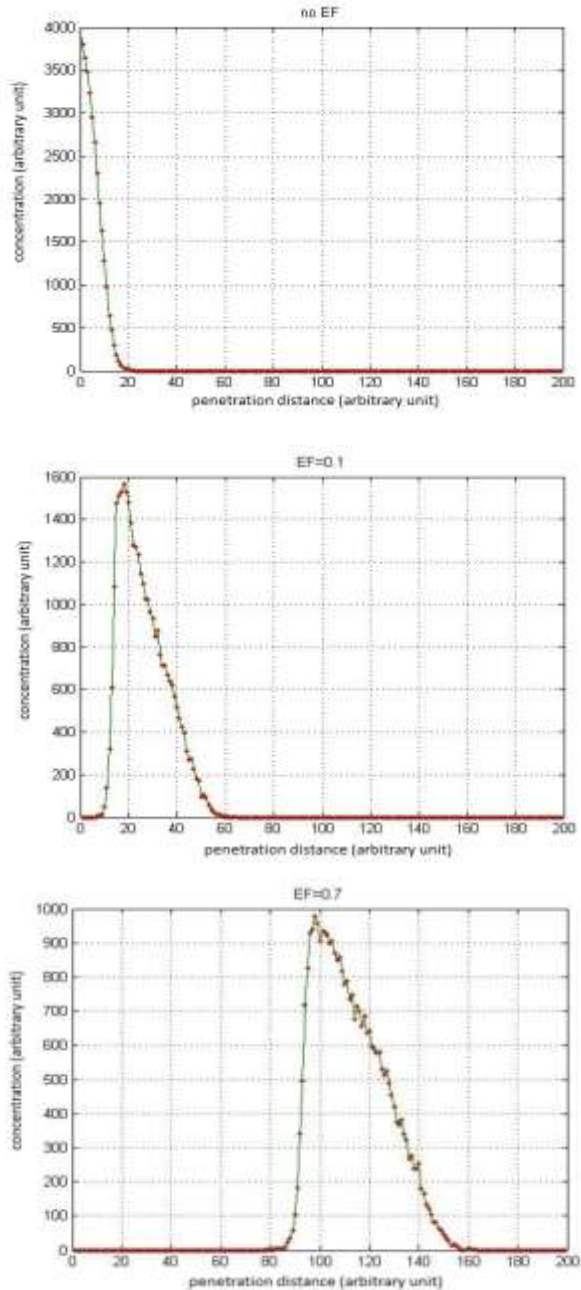


Fig. 7. The variation diffusion coefficient with vacancy concentration.

The diffusion of positive ions in a biological host under the effect of DC electric field is then examined for different field strengths, different annealing times and varying vacancy concentrations. Considering the matrix having a 90% vacancy and the DC electric field, EF, is increasing gradually (EF=0 % , 10%, and 70%), the diffusion pattern of the tracer in the host matrix is illustrated in Fig. 8. We infer that the concentration of the diffused



ions travels deeper as a result of increasing the electric field strength.

Fig. 8. Penetration profiles for ions in a 90% vacant matrix under the effect of direct electric field for different strengths, namely, EF=0 % , 10%, and 70% and for constant annealing time.

Fig. 9 illustrates the increase in the diffusion coefficient with the strength of the external field at a constant vacant percentage and for the annealing time.

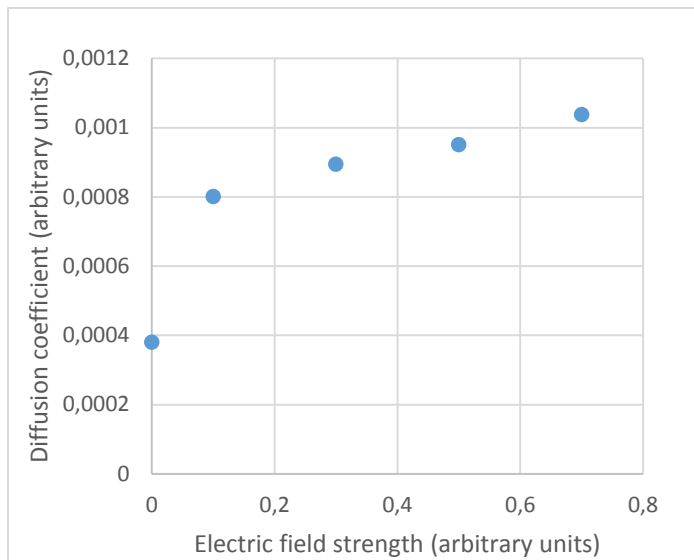


Fig. 9. Variation of diffusion coefficient with electric field strength.

4. Discussion and Conclusion

In the present work we have considered the profile of ionic diffusion in a biological tissue under the effect of DC electric fields. Initially, in the absence of an electric field and the vacancies occupy 0-30% of the host tissue, we found that the penetration of diffusing ions and the diffusion coefficient doesn't increase much. As the percentage of vacancies increases more than 30%, the penetration of the diffusing ions and the diffusion coefficient increase linearly with the vacancies concentration.

Preliminary results showed that when the matrix is 50% vacant, as the EF increases the penetration increases up to a point after which it decreases again. This happens because there is no much space for the movement of the ions, they are hindered by the matrix structure. As the percentage of vacancies increases the penetration of the diffusants ions increases. This increase is accelerated with the applications of an external field EF. The positive ions have more space and more probability to jump forwarded aided by the applied field.

In conclusion the present work introduces a stochastic model that tackles the problem of ionic diffusion in a biological tissue. Emphasis is given to the effect of both the existence of different percentage of vacancies available for random jumps and the effect of an external DC field of different strengths.

References

- [1] F. Jansson, A.V. Nenashev, S.D. Baranovskii, F. Gebhard, and R. Österbacka, "Effect of electric field on diffusion in disordered materials.," *Annalen der Physik*, vol. 18, pp. 856 – 862, 2010.
- [2] C. Kusnadi and S. K. Sastry, "Effect of moderate electric fields on salt diffusion into vegetable tissue.," *Journal of Food Engineering*, vol. 110, no. 3, pp. 329-336, 2012.
- [3] G. R. Smith and M. S. P. Sansom, "Effective diffusion coefficients of K⁺ and Cl⁻ ions in ion channel models. ," *Biophysical Chemistry*, vol. 79, no. 2, pp. 129-151, 1999.
- [4] E. Syková and C. Nicholson, "Diffusion in brain extracellular space.," *Physiological Reviews*, vol. 88, no. 4, pp. 1277-1340, 2008.
- [5] Y. Liu and X. Shi, "Ionic transport in cementitious materials under an externally applied electric field: Finite element modeling," *Construction and Building Materials*, vol. 27, no. 1, pp. 450-460, 2012.