Mathematical and Information Technologies
MIT-2016

Conference Proceedings

28 August – 5 September, 2016
Vrnjacka Banja, Serbia
Budva, Montenegro
Mathematical and Information Technologies

MIT-2016

Conference Proceedings

28 August — 5 September, 2016
Vrnjacka Banja, Serbia
Budva, Montenegro
Preface

This volume is published as the proceedings of the International Conference "Mathematical and Information Technologies, MIT-2016" held in Vrnjacka Banja, Serbia — Budva, Montenegro, from 28 August to 5 September 2016.

The conference was organized by Mathematical Institute SANU (Belgrade, Serbia), Institute of Computational Technologies of SB RAS (Novosibirsk, Russia), University of Pristina (Kosovska Mitrovica, Serbia), State University of Novi Pazar (Novi Pazar, Serbia), Al-Farabi Kazakh National University (Almaty, Kazakhstan), Mathematical Society of Kosovo and Metohija (Kosovska Mitrovica, Serbia), and Matrosov Institute for System Dynamics and Control Theory of SB RAS (Irkutsk, Russia).

The conference continued the good scientific traditions established in 2009 by cooperation mainly between Serbian and Russian scientists. The conference held fourth time and have been broaden its influence for the last years. The purpose of the conference is to bring together researchers and practitioners around the topics of mathematical modeling and combining information technologies. The areas of interested topics vary from the mathematical, computational and information methods to the modeling and simulation of challenging applications.

We are glad to see that the conference attracted more than 100 participants and that the talks led to many intensive discussions. The call for papers attracted 80 submissions. After a rigorous review process where each paper received at least two reviews, 46 papers were accepted for publications in the volume. The contributions of the proceedings were provided and edited by the authors. We hope the volume will be interesting for scientists in the field.

The conference has required a lot of effort by many people. The editors would like to thank to all participants of the conference and all people who helped us to make this event.

March 2017

Yurii Shokin
Hranislav Milosevic
Denis Esipov
Program Committee

Program Committee Chairs

Yurii Shokin Institute of Computational Technologies SB RAS, Novosibirsk, Russia
Srecko Milacic University of Pristina temporarily settled in Kosovska Mitrovica, Kosovska Mitrovica, Serbia
Cemal Dolichanin State University of Novi Pazar, Novi Pazar, Serbia

Program Committee

Ivan Arandelovic University of Belgrade, Belgrade, Serbia
Nebojsa Arsic University of Pristina temporarily settled in Kosovska Mitrovica, Kosovska Mitrovica, Serbia
Ranko Babic University of Pristina temporarily settled in Kosovska Mitrovica, Kosovska Mitrovica, Serbia
Milorad Banjanin University of East Sarajevo, East Sarajevo, Bosnia and Herzegovina
Natan Blaunshtein Ben-Gurion University of the Negev, Beersheba, Israel
Thomas Bonisch High Performance Computing Center in Stuttgart, Stuttgart, Germany
Igor Bychkov Institute for System Dynamics and Control Theory of SB RAS, Irkutsk, Russia
Leonid Chubarov Institute of Computational Technologies SB RAS, Novosibirsk, Russia
Donco Dimovski Saints Cyril and Methodius University of Skopje, Skopje, Macedonia
Rade Doroslovacki University of Novi Sad, Novi Sad, Serbia
Denis Esipov Institute of Computational Technologies SB RAS, Novosibirsk, Russia
Huse Fatkic University of Sarajevo, Sarajevo, Bosnia and Herzegovina
Mikhail Fedoruk Novosibirsk State University, Novosibirsk, Russia
Anatoly Fedotov Institute of Computational Technologies SB RAS, Novosibirsk, Russia
Ljiljana Gajic University of Novi Sad, Novi Sad, Serbia
Vidan Govedarica University of East Sarajevo, East Sarajevo, Bosnia and Herzegovina
Amanbek Jainakov Kyrgyz State Technical University named after I. Razzakov, Bishkek, Kyrgyzstan
Ditrich Kroner University of Freiburg, Freiburg, Germany
Almatbek Kydyrbekuly Al-Farabi Kazakh National University, Almaty, Kazakhstan
Zarko Mijailovic State University of Novi Pazar, Novi Pazar, Serbia
Hranislav Milosevic  
University of Pristina temporarily settled in Kosovska Mitrovica, Kosovska Mitrovica, Serbia

Gradimir Milovanovic  
Mathematical Institute of SASA, Belgrade, Serbia

Vladimir Moskvichev  
Institute of Computational Technologies SB RAS, Krasnoyarsk, Russia

Zoran Ognjanovic  
Mathematical Institute of SASA, Belgrade, Serbia

Zarko Pavicevic  
University of Montenegro, Podgorica, Montenegro

Branislava Perunicic  
University of Sarajevo, Sarajevo, Bosnia and Herzegovina

Dojcin Petkovic  
University of Pristina temporarily settled in Kosovska Mitrovica, Kosovska Mitrovica, Serbia

Vera Petrovic  
High school of Electro technical and Computers, Belgrade, Serbia

Milenko Pikula  
University of Republic Serbska, Sarajevo, Bosnia and Herzegovina

Stevan Pilipovic  
University of Novi Sad, Novi Sad, Serbia

Vadim Potapov  
Institute of Computational Technologies SB RAS, Kemerovo, Russia

Oleg Potaturkin  
Institute of Automation and Electrometry SB RAS, Novosibirsk, Russia

Jelena Protic  
University of Belgrade, Belgrade, Serbia

Bozidar Radenkovic  
University of Belgrade, Belgrade, Serbia

Miloje Rajovic  
University of Kragujevac, Kraljevo, Serbia

Michael Resch  
High Performance Computing Center in Stuttgart, Stuttgart, Germany

Karl Roesner  
Institute for Fluid Mechanics and Aerodynamics, Technical University of Darmstadt, Darmstadt, Germany

Sergei Rusakov  
Institute for Design Problems in Microelectronics RAS, Moscow, Russia

Nikita Sekutkovski  
Saints Cyril and Methodius University of Skopje, Skopje, Macedonia

Vladimir Shaidurov  
Institute of Computational Modelling, Krasnoyarsk, Russia

Nina Shokina  
Computing Center, University of Freiburg, Freiburg, Germany

Sergei Smagin  
Computer Center of FEB RAS, Khabarovsk, Russia

Viktor Soifer  
Samara National Research University, Samara, Russia

Aleksandr Stempkovskii  
Institute for Design Problems in Microelectronics RAS, Moscow, Russia

Svetlana Stevovic  
University Union, Belgrade, Serbia

Igor Turchanovskiy  
Institute of Computational Technologies SB RAS, Tomsk, Russia

Bakhytzhan Zhumagulov  
National Engineering Academy of RK, Almaty, Kazakhstan
<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zdravko Vitosevic</td>
<td>University Union, Belgrade, Serbia</td>
</tr>
<tr>
<td>Mirjana Vukovic</td>
<td>Academy of Sciences and Arts of Bosnia and Herzegovina, Sarajevo, Bosnia and Herzegovina</td>
</tr>
<tr>
<td>Duska Yovovich</td>
<td>University of Pristina temporarily settled in Kosovska Mitrovica, Kosovska Mitrovica, Serbia</td>
</tr>
<tr>
<td>Ziyavidin Yuldashev</td>
<td>National University of Uzbekistan named by after Mirza Ulugbek, Tashkent, Uzbekistan</td>
</tr>
<tr>
<td>Andrei Yurchenko</td>
<td>Institute of Computational Technologies SB RAS, Novosibirsk, Russia</td>
</tr>
<tr>
<td>Aleksandar Zoric</td>
<td>University of Pristina temporarily settled in Kosovska Mitrovica, Kosovska Mitrovica, Serbia</td>
</tr>
</tbody>
</table>
# Table of Contents

## Information technologies

GIS-technologies and mathematical simulation as tools for lightning-caused forest fire danger prediction  
*Nikolay Baranovskiy, Vladimir Barakhnin, Elena Yankovich*  
2

Evaluation of seismic hazard using seismic microzonation techniques  
*Evgeniy Bodyakin, Sergey Peretokin, Konstantin Simonov*  
16

3-D modeling of Angara river bed  
*Igor Bychkov, Andrey Gachenko, Gennady Rugnikov, Alezei Hmelnov*  
26

Methods for high performance graph processing  
*Mikhail Chernoskutov*  
33

Improving the efficiency of the human spine diagnostics systems  
*Nikolay Dorofeev, Konstantin Podmasteriev, Oleg Kuzichkin, Anastasia Grecheneva*  
41

Application of e-learning tools in different ways of implementation  
*Snezana Gavrilovic, Nebojsa Denic, Vukman Korac, Vera Petrovic*  
52

High technology software web-tools to solve environmental problems of coal region  
*Alexander Goudov, Valery Perminov, Yuri Filatov, Lee Hee Un, Sergey Zavozkin, Irina Grigorieva, Igor Sotnikov*  
61

The method of selection of the key geodynamic objects  
*Anastasia Grecheneva, Vladimir Eremenko, Oleg Kuzichkin, Nikolay Dorofeev*  
74

Decision support system for real-time diagnosis of musculoskeletal system  
*Anastasia Grecheneva, Igor Konstantinov, Oleg Kuzichkin, Nikolay Dorofeev*  
84

The study of inversion problems of cryptographic hash functions from MD family using algorithms for solving boolean satisfiability problem  
*Irina Gribanova, Oleg Zaikin, Stepan Kochemazov, Ilya Otpuschennikov, Alexander Semenov*  
98

Construction features and data analysis by BP-SOM modular neural network  
*Vladimir Gridin, Vladimir Solodovnikov*  
114

Synthetic data for AUV technical vision systems testing  
*Aleksandr Kamaev, Sergey Smagin, Viacheslav Sukhenko, Dmitry Karmanov*  
126

Methods and tools of parallel programming  
*Victor Kasyanov, Elena Kasyanova*  
141
The comparison of different SAT encodings for the problem of search for systems of orthogonal latin squares
*Stepan Kochemazov, Oleg Zaikin, Alexander Semenov*

Development of extended path-based role access control model for web applications
*Dmitry Kononov, Sergey Isaev*

Model of neuro-fuzzy prediction of confirmation timeout in a mobile ad hoc network
*Igor Konstantinov, Kostiantyn Polshchykov, Sergej Lazarev, Olha Polshchykova*

Implementation of Weibull’s model for determination of aircraft’s parts reliability and spare parts forecast
*Nataša Kontrec, Milena Petrović, Jelena Vujaković, Hranislav Milošević*

Webometric analysis of russian scientific and education web
*Denis Kosyakov, Andrey Guskov, Egor Bykhovtsev*

LiFlow: A workflow automation system for reproducible simulation studies
*Evgheniy Kuklin, Andrey Sozykin, Konstantin Ushenin, Dmitriy Byordov*

Parallel text document clustering based on genetic algorithm
*Madina Mansurova, Vladimir Barakhnin, Sanzhar Aubakirov, Yerzhan Khibatkhanuly, Aigerim Mussina*

Using non-negative matrix factorization for text segmentation
*Aliya Nugumanova, Madina Mansurova, Yerzhan Baiburin, Yermek Alimzhanov*

**Mathematical modeling**

Analysis and design of hybrid pressure vessels
*Evgeniya Amelina, Sergey Golushko, Andrey Yurchenko*

On the inverse problem of a creeping motion in thin layers
*Victor Andreev*

Tensor smooth length for SPH modelling of high speed impact
*Roman Cherepanov, Alexander Gerasimov*

Multiscale modeling of strength properties of dispersion-reinforced ceramic composite materials
*Yuriy Dimitrienko, Yulia Zakharova, Sergey Sborschikov*

Single and group impacts of high-speed elements on spacecraft
*Alexander Gerasimov, Yuriy Khristenko, Sergey Pashkov*

Hp-version of vollocation and least residuals method in mechanics of laminated composite plates
*Sergey Golushko, Semyon Idimeshev, Vasiliy Shapeev*
Computational modelling of metal vapor influence on the electric arc welding
Amanbek Jainakov, Rena Sultangazieva, Bubusara Medralieva

Supercomputer modeling of stochastic dynamics of the mercury ion array in an optical lattice
Leonid Kamenshchikov, Igor Krasnov

Congruent circles packing and covering problems for multi-connected domains with non-Euclidean metric, and their applications to logistics
Alexander Kazakov, Anna Lempert, Pavel Lebedev

Some exact solutions of a heat wave type of a nonlinear heat equation
Alexander Kazakov, Anna Lempert, Sergey Orlov, Svyatoslav Orlov

About verification of calculation methods of the shock waves
Valentin Kuropatenko, Elena Shestakovskaya

Applications of regularly varying functions in study of cosmological parameters
Žarko Mijajlović

Mathematical modeling of artificial mitral heart valve
Hranislav Miloshevic, Yury Zakharov, Yurii Shokin, Dmitry Dolgov, Irene Grigorieva

Forecast of thermokarst lakes dynamics in permafrost based on geo-simulation modeling and remote sensing data
Yury Polishchuk, Vladimir Polishchuk

Modification of Fourier approximation for solving boundary value problems having singularities of boundary layer type
Boris Semisalov, Georgy Kuzmin

Methods for optimal control of hydraulic fracturing process
Yuriii Shokin, Sergey Cherny, Vasily Lapin, Denis Esipov, Dmitriy Kuranakov, Anna Astrakova

Specification and analysis of transients in electrical power systems using the methodology of hybrid systems
Yury Shornikov, Dmitriy Dostovalov, Maria Nasyrova

Using of the mosaic-skeleton method for numerical solution of three-dimensional scalar diffraction problems
Sergei Smagin, Aleksei Kashirin, Mariia Taltykina

Numerical analysis of acoustic waves in a liquid crystal taking into account couple-stress interaction
Irina Smolekho, Oxana Sadowskaya, Vladimir Sadowskii

Biochemical processes of self-purification model in small rivers
Olga Taseiko, Tatiana Spitsina, Hranislav Milosevic, Dragan Radovanovic, Aleksandar Valjarevic
Modeling an initial tsunami waveform by inverting remote sea level records through the r-solution method
*Tatyana Voronina* 496

Effect of dielectric permittivity distribution in invaded zone on induction log data
*Igor Yeltsov, Timofey Eltsov, Aleksandr Makarov, Marina Nikitenko* 508

Computational investigation of turbulent flow impact on non-cohesive soil erosion near foundations of gravity type oil platforms
*Alexander Yushkov, Igor Nudner, Konstantin Semenov, Konstantin Ivanov, Nazim Geidarov, Sergey Stukolov, Yury Zakharov* 524

Numerical simulation of surface waves arising from underwater landslide movement
*Yury Zakharov, Anton Zimin* 535

Modeling of beryllia ceramics formation process
*Uzak Zhapbasbayev, Gaukhar Ramazanova, Bakhytzhan Assilbekov, Zamira Sattinova* 547

**List of Authors** 559
INFORMATION TECHNOLOGIES
GIS-Technologies and Mathematical Simulation as Tools for Lightning-Caused Forest Fire Danger Prediction

Nikolay Baranovskiy¹, Vladimir Barakhnin²,³, and Elena Yankovich¹

¹ National Research Tomsk Polytechnic University, Lenina ave. 30, 634050 Tomsk, Russia
² Novosibirsk State University, Pirogov st. 2, 630090 Novosibirsk, Russia
³ Institute of Computational Technologies SB RAS, Lavrent’ev ave. 6, 630090 Novosibirsk, Russia
firedanger@narod.ru

Abstract. New approach to forecasting of forest fire danger caused by storm activity is presented in the article. This approach is based on using the criteria of forest fire danger and physically proved mathematical models of forest fuel ignition. The formula of criterion is based on a probabilistic assessment of forest fire danger and uses the main theorems of probability theory. Data of a forest fire retrospective on the controlled territory are used to assess the members in probabilistic criterion. Timiryazevskiy local forestry of the Timiryazevskiy timber enterprise of the Tomsk region is considered as a typical territory. It is shown that it is not enough to use only statistical information on forest fires for an adequate assessment of the forest fire danger caused by action of storm activity. Visualization of data is carried out with the use of geoinformation technologies.

Keywords: GIS, mathematical simulation, forest fire danger, prediction, lightning.

1 Introduction

The remote areas of forested territories are characterized by a big share of the forest fires caused by action of storm activity [1]. The great value of the area passed by fire is noted for such fires [2]. Such fires in the forests are detected with delay when the ignition center already reached the big sizes. It is either impossible or ineffective to suppress ignitions in taiga zone. Fire fades in case of the beginning of long rains, or at burning out of all forest area before fire came across the natural barrier (for example, river). In such a situation, the most perspective approach is to forecast the forest fire danger and to carry out preventive measures in controlled forested territories [3]. There are various forest fire danger forecast systems taking into account storm activity developed in the different countries of the world [4–6]. However, all these systems have no physical basis and are based mainly on the analysis of statistical information on forest fires and characteristics of the forested territory [7].
2 Background

Storm discharges are one of the reasons for forest fires. Lightning is an electric discharge conditioned by the division into positive and negative discharges in the clouds that leads to a difference in potentials of the range 10-100 mV [8]. In order for the division into discharges to happen, it is necessary that water be present in all three phases: solid, liquid and gas [9].

According to the development conditions, storms are divided into the air-mass and frontal ones. Air-mass storms over a continent occur as the result of the local air heating from the ground surface that leads to a development of rising flows of the local convection and to a formation of heavy cumulonimbus clouds in it. The frontal storms occur on the borders of warm and cold air masses [10]. There may be the cloud-to-cloud and cloud-to-ground discharges. Around 90% of cloud-to-ground discharges are negative, and the nature of the remaining 10% of positive discharges is not fully clear [11]. The cloud-to-ground discharges, i.e. ground storm discharges, can cause forest fires [12]. The energy characteristics for positive and negative ground storm discharges are different, and these differences are substantial in terms of igniting the forest fuels. Due to the vast majority of positive discharges, all the energy reaches the surface in one stroke, and a multi-stroke is typical for the negative discharges [13].

Wide statistics on ground storm discharges has been collected within the functioning of the US National Lightning Detection Network [14]. This system may identify most ground storm discharges on USA and Canada territories with the spatial resolution of several kilometers and determination accuracy in time of 1 msec. Due to the system operation, the data on the stroke polarity, stroke peak current and stroke complexity are archived (if it is a single or multi-stroke) [13].

In Russia, between 1992 and 2000, storm-induced forest fires equaled 37 to 53 \% of the area where fire had spread, with a relative number of 8.817.5 \% [15]. Dry storms, producing mass ignitions on large spaces, often create a very intense situation [16].

Canadian Forest Fire Danger Rating System (CFFDRS) has two main sub-systems (modules) Canadian Forest Fire Weather Index (FWI) System and Canadian Forest Fire Behavior Prediction (FBP) System. Two other elements (Fuel Moisture System and Canadian Forest Fire Occurrence Prediction (FOP) System) are not developed for the whole country, but there are regional versions of these systems [17].

The Canadian method of forest fire danger prediction [14] is formed relying on analysis of a large number of statistical data according to which they formed the tables of fire danger dependence on different factors. Within the FWI sub-system, the moisture content of forest fuels is predicted depending on weather conditions; whereas within FBP, forest fire spots behavior is forecasted for different forest plant communities.

A logical structure of the system [15, 16] represents an abstract model of the impact of different factors and conditions on the process of how fire occurs and spreads.
The Canadian and American methods are similar to each other in their structure, approaches and fire danger index formation principles. Therefore, they have both similar advantages and disadvantages. European Forest Fire Information System EFFIS (Europe) [6]. The most progressive component of system repeats the subsystem of the Canadian Forest Fire Danger Rating System. This system has the same characteristics and uses Earth remote sensing data.

The work purpose is to create a new method for geospatial data analysis in order to monitor, assess and forecast the forest fire danger caused by storm activity.

3 Mathematical Methods

Using the basic principles of probability theory, we obtained a formula to assess the probability for the forest fire to occur for the j-th time interval of the forest fire season [18]:

\[
P_j = [P(A)P(A_j/A)P(FF/A, A_j) + P(L)P(L_j/L)P(FF/L, L_j)]P_j(D), \quad (1)
\]

where \(P_j\) is the probability of a forest fire to occur for the j-th interval in the controlled forest area; \(P(A)\) is the probability of anthropogenic impact; \(P(A_j/A)\) is the probability of a fire source presence on j-th day; \(P_j(FF/A, A_j)\) is the probability of a forest fire to occur from anthropogenic impact in the forest area; \(P(L)\) is the probability of dry thunderstorms to occur in the forest area; \(P(L_j/L)\) is the probability of ground lightning discharge; \(P_j(FF/L, L_j)\) is the probability of a forest fire to occur from lightning in case, if dry thunderstorms can happen in the forest area; \(P_j(D)\) is the probability of fire to occur due to weather conditions of forest fire maturation (the probability of the fact that the forest fuel layer will be dry); index \(j\) corresponds to a day of the fire danger season. To determine all multipliers in the formula (1), the author offers to use a definition of probability through frequency of events and to use statistical data for a concrete forestry. The formula (1) contains the following members [18]:

\[
P(A_j/A) \approx \frac{N_{FD}}{N_{FW}}, P(A) \approx \frac{N_A}{N_{FS}}, \quad (2)
\]

\[
P_j(FF/A, A_j) \approx \frac{N_{FA}}{N_{FT}}, \quad (3)
\]

\[
P(L_j/L) \approx \frac{N_{LN}}{N_{LD}}, P(L) \approx \frac{N_L}{N_{FS}}, \quad (4)
\]

\[
P_j(FF/L, L_j) \approx \frac{N_{FL}}{N_{FT}}, \quad (5)
\]

where \(N_A\) is the number of days during the fire danger season when the anthropogenic impact is enough for forest fuel ignition; \(N_{FA}\) is the number of
fires from anthropogenic impact; $N_{FT}$ is the total number of fires; $N_L$ is the number of days when lightning occurred (during dry thunderstorms); $N_{FS}$ is the total number of days in the fire danger season; $N_{FL}$ is the number of fires from lightning (during dry thunderstorms); $N_{FD}$ is the number of fires on a specific day of the week; $N_{FW}$ is the total number of fires for a week; $N_{LH}$ is a number of ground lightning discharges passed on the concrete hour, starting from 00:00 oclock; $N_{LD}$ is the total number of ground lightning discharges per day. Obviously, the more cases will be considered for this forestry, the bigger accuracy to determine the probability by formulas (2)-(5) will be. Therefore, in forestries, it is necessary to register all fire danger season parameters ($N_A$, $N_{FA}$, $N_{FT}$, $N_L$, $N_{FS}$, $N_{FL}$, $N_{FD}$, $N_{FW}$, $N_{LH}$, $N_{LD}$) every year.

Formula (1) contains the multiplier $P_J(D)$. This is the probability of fire danger from meteorological conditions. In early work, this probability was calculated through the time for the forest fuel layer to dry [19]. However, at present, it is hard to implement the method like this on the whole territory of Russian Federation, because in order to model the process of drying the forest fuel layer, it is necessary to have information about the initial moisture content of forest fuel. The present paper offers to use the compromise variant. We suggest calculating the probability by meteorological conditions using the Complex Meteorological Index, which was approved in the state standard. The range of this index starts from zero and has no upper border. However, it is possible to set its upper border as a maximum possible value during the fire danger season. To estimate the probability of forest fire danger, we normalize the complex meteorological index on figure of one [18]:

$$P_J(D) = \frac{NI_D}{NI_{max}},$$

where $NI_D$ is a value from the complex meteorological index for the day for which the forecast is realised; $NI_{max}$ is the maximum value of the complex meteorological index. Then, the range of variation of forest fire danger probability by meteorological conditions will be within 0 to 1.

The complex meteorological index is calculated by the formula [7]:

$$NI = \sum_n t(t - r),$$

where $t$ is the air temperature; $r$ is the dew point temperature; $n$ is the number of days after the last rain.

The dew point characterises the amount of moisture in the air. The higher the dew point, the greater the humidity is at a given temperature. The dew point temperature is determined as the temperature to which air must be cooled (at the constant pressure and constant water vapour content), in order to reach saturation and in order for its condensation process to start, that is, the dew to appear. The saturation state can exist only as long as the air contains the maximum possible amount of water vapour at the given temperature and pressure.

The work [20] suggests a simple mathematical model of tree ignition by the cloud-to-ground lightning discharge.
Electric current flow is various in the trunk of deciduous and coniferous trees [21]. This is due to the fact that in broad-leaved trees, moisture is transported in a massive central part. More damp central part is an electric current conductor. The analysis of the known information on wood properties of broad-leaved species shows, that it is necessary to consider moisture presence in the trunk wood structure. Even under the conditions of high-speed processes, moisture presence can essentially change the wood ignition conditions. Therefore, when setting the task for broad-leaved trees, it is expedient to consider the influence of moisture content on thermophysical characteristics of wood.

We consider the following physical model. A cloud-to-ground lightning discharge strikes in a tree trunk at the fixed moment of time. The electric current of the cloud-to-ground lightning discharge flows along the trunk. It is supposed, that the heat emits in the core according to Joule-Lenz law. It is supposed that in various trunk sections, the electric current has the same parameters. It is considered, that one can describe the moisture evaporation by Knudsen-Lengmuir equation [22]. As a result of electric current flow, the wood is warmed up due to the Joule heat emission and the wood ignites when achieving the critical thermal fluxes to ignition surface and critical temperature. It is supposed, that the formed vapor space is filled with water vapor. Changes of volume fractions of phases are reflected on thermophysical properties of internal wood part of the broad-leaved tree. The tree trunk is modeled by the cylinder. We consider the representative section of a trunk. Fig. 1 shows the decision area scheme.

![Decision Area Scheme](image)

**Fig. 1.** The decision area scheme: 1 core, 2 - bark.

The system of non-stationary differential equations mathematically describes the process how a cloud-to-ground lightning discharge warms up a tree trunk before ignition [20]:

\[
\rho_{ef} c_{ef} \frac{\partial T_1}{\partial t} = \frac{\lambda_{ef}}{r} \frac{\partial}{\partial r} \left( r \frac{\partial T_1}{\partial r} \right) + JU - QW \varphi_2, \tag{8}
\]
\[
\rho_2 c_2 \frac{\partial T_2}{\partial t} = \frac{\lambda_2}{r} \frac{\partial}{\partial r} \left( r \frac{\partial T_2}{\partial r} \right), \quad (9)
\]

\[
\rho_3 \frac{\partial \varphi_1}{\partial t} = 0, \quad (10)
\]

\[
\rho_4 \frac{\partial \varphi_2}{\partial t} = -W, \quad (11)
\]

\[
\sum_{i=3}^{5} \varphi_i = 1, \quad (12)
\]

\[
W = \frac{A(P^5 - P)}{\sqrt{2\pi R T M}}, \quad (13)
\]

\[
\rho_{ef} = \rho_3 \varphi_3 + \rho_4 \varphi_4 + \rho_5 \varphi_5, \quad \lambda_{ef} = \lambda_3 \varphi_3 + \lambda_4 \varphi_4 + \lambda_5 \varphi_5. \quad (14)
\]

Boundary conditions for the equations (1) - (2):

\[
r = 0, \lambda_{ef} \frac{\partial T_1}{\partial r} = 0; \quad (15)
\]

\[
r = R_1, \lambda_{ef} \frac{\partial T_1}{\partial r} = \lambda_2 \frac{\partial T_2}{\partial r}, T_1 = T_2; \quad (16)
\]

\[
r = R, \lambda_2 \frac{\partial T_2}{\partial r} = \alpha(T_e - T_{Rs}). \quad (17)
\]

Initial conditions for the equations (1) – (5):

\[
t = 0, T_i(r) = T_{i_0}, \varphi_1(0) = \varphi_{i_0}. \quad (18)
\]

Where \( T_i \) is temperature of internal part of tree trunk (\( i = 1 \)) and bark (\( i = 2 \)); \( \varphi_i \) - volume fraction: organic substance (\( i = 3 \)), water (\( i = 4 \)) and water vapor (\( i = 5 \)); \( \rho_i, c_i, \lambda_i \) is density, thermal capacity and heat conductivity of bark (\( i = 2 \)), organic substance (\( i = 3 \)), water (\( i = 4 \)) and water vapor (\( i = 5 \)); \( \rho_{ef}, c_{ef}, \lambda_{ef} \) - effective density, thermal capacity and heat conductivity of wood for internal part of trunk; \( \lambda \) - heat transfer factor; \( J \) - current strength; \( U \) - voltage; \( Q \) - thermal effect of moisture evaporation; \( r \) - coordinate; \( t \) - time. \( W \) - mass speed of water evaporation, \( A \) - accommodation coefficient, \( P^5 \) - pressure of saturated water vapor, \( P \) - partial pressure of water vapor in air, \( R \) - universal gas constant, \( M \) - molecular weight of water. Indexes \( Rs, e \) and 0 correspond to the parameters on the external border of tree trunk, the environment and to the parameters at the initial moment of time.

Formulated system of equations (8) – (14) with boundary and initial conditions (15) – (18) is solved by the finite difference method [23]. The double-sweep method in combination with the fixed point iteration method [23] was used to decide the difference analogues of one-dimensional equations.
The following ignition scenario was considered. The negative cloud-to-ground lightning discharge, with duration of $500\, ms$, with peak current of stroke in $23.5\, kA$ and voltage $100\, kV$, influences on a wide-leaved tree, for instance, birch. Fig. 2 shows the temperature distribution on the tree trunk radius in various moments of time before and at the moment of igniting by electric current (initial temperature $300\, K$).

![Temperature distribution on the tree trunk radius](image)

**Fig. 2.** Temperature distribution on the tree trunk radius at the various moments of time (discharge action duration is $500\, ms$): a - $t = 0.01\, s$; b - $0.1\, s$; c - $0.3\, s$; d - $0.5\, s$ [20].

Table 1 represents the lightning discharge parameters and ignition conditions depending on voltage of ground-to-cloud lightning discharge obtained by solving the problems (8) – (14). The analysis of dependences presented on Fig. 2 shows, that tree trunk is warmed up to ignition temperature (more than $1000\, K$) by the action of the considered cloud-to-ground lightning discharge. The analysis of results shows that for a typical cloud-to-ground lightning discharge, ignition conditions of wide-leaved tree are reached on critical temperature ($801\, K$) and value of thermal flux ($268\, kW/m^2$).

We established the ignition limits for tree trunk during the action of the electric discharge at various voltages (Table 1) and current. When the current is less than $15\, kA$ and voltage is $1 - 50\, kV$, ignitions fail to occur during the action of cloud-to-ground lightning discharge.
Table 1. Ignition condition of tree depending on voltage of the discharge at current \( J = 23.5kA \) [20]

<table>
<thead>
<tr>
<th>Voltage, ( U, kV )</th>
<th>Ignition conditions</th>
<th>Surface temperature, ( ^{\circ}C )</th>
<th>Heat flux from core to surface, ( kW/m^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–45</td>
<td>No</td>
<td>&lt; 801</td>
<td>&lt; 201</td>
</tr>
<tr>
<td>50</td>
<td>No</td>
<td>&lt; 801</td>
<td>252</td>
</tr>
<tr>
<td>55</td>
<td>Yes</td>
<td>801</td>
<td>268</td>
</tr>
<tr>
<td>60</td>
<td>Yes</td>
<td>801</td>
<td>268</td>
</tr>
<tr>
<td>80</td>
<td>Yes</td>
<td>801</td>
<td>268</td>
</tr>
<tr>
<td>100</td>
<td>Yes</td>
<td>801</td>
<td>268</td>
</tr>
<tr>
<td>110</td>
<td>Yes</td>
<td>801</td>
<td>268</td>
</tr>
</tbody>
</table>

3.1 GIS System

Program realization of mathematical model for quantitative assessment of probability of forest fire danger caused by storm activity is enabled in GIS.

Algorithms of geographical information system for quantitative assessment of forest fire danger are implemented in the Python language embedded into ArcGIS [24]. The quantitative assessment is carried out relying on the remote sensing data, land mensuration of forests and statistical information. The criteria to assess the forest fire danger are defined relying on the probability theory, and its values are within the range from 0 to 1. Calculations are made with accuracy up to 0.0001.

Below are the tables in the MS Excel format with forest mensuration descriptions on stratums (Table 2). Russian database on stratums description is used.

The program tool “FFstormactivity.tbx” solves the problem to forecast the fire danger of forest quadrant relying on the information about stratum composition and statistical information on fires caused by storm activity and the display of the obtained information on the electronic map. Python is the source language of the “FFstormactivity” program [24].

The program tool “FFstormactivity” contains 7 forms. It provides two variants to solve the task: complex and stage-by-stage with the control of result. Main stages:

1. Data import from the table Excel to the autonomous geodata base table.
2. Determination of fire danger for forest stratum.
3. Assessment of fire danger probability for forest quadrant according to forest mensuration descriptions.
4. Import of statistical data to geodata base.
5. Assessment of fire danger probability caused by storm activity.
6. Connection of attributive and autonomous tables.
7. Formation of the map according to a legend.
Table 2. Tables of forest mensuration data in the MS Excel format.

<table>
<thead>
<tr>
<th>forestry</th>
<th>quarter</th>
<th>site</th>
<th>area</th>
<th>composition</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaltayskiy</td>
<td>1</td>
<td>1</td>
<td>43.7</td>
<td>Grass</td>
<td></td>
</tr>
<tr>
<td>Kaltayskiy</td>
<td>1</td>
<td>2</td>
<td>7.6</td>
<td>7B2L1P</td>
<td></td>
</tr>
<tr>
<td>Kaltayskiy</td>
<td>1</td>
<td>3</td>
<td>14.8</td>
<td>7W3B</td>
<td></td>
</tr>
<tr>
<td>Kaltayskiy</td>
<td>1</td>
<td>4</td>
<td>19</td>
<td>7W3B</td>
<td></td>
</tr>
<tr>
<td>Kaltayskiy</td>
<td>1</td>
<td>6</td>
<td>18.3</td>
<td>8B2L</td>
<td></td>
</tr>
<tr>
<td>Kaltayskiy</td>
<td>1</td>
<td>7</td>
<td>5.6</td>
<td>5P21L2B</td>
<td></td>
</tr>
<tr>
<td>Kaltayskiy</td>
<td>1</td>
<td>8</td>
<td>5.3</td>
<td>5C2F1P2B</td>
<td></td>
</tr>
<tr>
<td>Kaltayskiy</td>
<td>1</td>
<td>10</td>
<td>1.2</td>
<td>Lake</td>
<td></td>
</tr>
</tbody>
</table>

First step is data import from the table Excel to the autonomous geodata base table.

Next step is determination of fire danger for stratum according to forest mensuration descriptions and assessment of fire danger probability of forest quadrant according to forest mensuration descriptions.

Third step is import of statistical data on storm activity to geodata base.

Algorithm of assessment of probability of the fire danger caused by storm activity presented on Fig. 3.

Last stages are connection of attributive and autonomous tables and formation of the map according to the legend.

The program tool FFstormactivity uses the following methods:

1. **AddField** is to add a field. The program adds a field.
2. **CalculateField** is to calculate the field value. To determine the fire danger of the stratum, to assess the probability of forest fire danger on the quadrant, of the level.
3. **Statistic_analysis** is the total statistics. To calculate the total quantity of stratums in each quadrant and quantity of the fire-dangerous ones.
4. **JoinField** is to connect the fields. A connection of two tables takes place on the basis of a key field
5. **ApplySymbologyFromLayer_management** is to add the layer symbols. To form the layer of quadrants according to the fire danger level.

The start-up of the program tool comes from ArcToolbox. It is necessary to specify the initial data in the dialog window that appears after start-up. Russian interface is used in current version of GIS-system.

Structures of tables with initial data on forest fires and forest mensuration descriptions of the territory are given below (Table 3 and Table 4).
The tool implementation results in creating the table with an assessment of probability of fire danger of forest quadrant caused by storm activity with regard to the forest vegetation conditions and the thematic map displaying the fire danger levels of forest quadrants (Fig. 4).

The forest fire danger levels in Fig. 4 correspond to the following gradation: 1 - 0,001852 - 0,030000 2 - 0,030001 - 0,060000 3 - 0,060001 - 0,090000 Minimum - 0,001852; maximum - 0,08333.

4 Discussion

The analysis of foreign forest fire danger forecast systems shows that in the territory of their states they show high operational qualities. However, it is difficult to apply them in the territory of other states, as it is necessary to carry out all range of works on the analysis and adjustment of empirical formulas for new forested territories.

All foreign systems finally offer an abstract index of forest fire danger. The present paper offers the new probabilistic approach to assess the most probable
Table 3. Structure of the table with statistical data on the fires in the MS Excel format.

<table>
<thead>
<tr>
<th>Name of forest</th>
<th>Name of forest district</th>
<th>N_F</th>
<th>A total number of the fires during a fire-dangerous season</th>
</tr>
</thead>
<tbody>
<tr>
<td>N_FA</td>
<td></td>
<td>N</td>
<td>Number of the fires from storm activity during a fire-dangerous season</td>
</tr>
<tr>
<td>N_D</td>
<td></td>
<td>D</td>
<td>Total number of days of a fire-dangerous season</td>
</tr>
<tr>
<td>N_A</td>
<td></td>
<td>A</td>
<td>Number of days during a fire-dangerous season when there is a storm activity sufficient for ignition of forest fuel</td>
</tr>
<tr>
<td>N_1</td>
<td>Monday</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_2</td>
<td>Tuesday</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_3</td>
<td>Wednesday</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_4</td>
<td>Thursday</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_5</td>
<td>Friday</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_6</td>
<td>Saturday</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_7</td>
<td>Sunday</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N_W</td>
<td>Total number of fires per a week</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Structure of the table with initial data on strata in the MS Excel format.

<table>
<thead>
<tr>
<th>forestry</th>
<th>quarter</th>
<th>site</th>
<th>area</th>
<th>composition</th>
<th>age</th>
</tr>
</thead>
</table>

scenarios of forest fire danger. The definite scenario can be calculated using a deterministic mathematical model of how the cloud-to-ground lightning discharge ignites a tree.

We developed GIS-system for forecasting the forest fire danger caused by storm activity. The system reserves the layers for the subsequent implementation of a deterministic component based on the mathematical model of tree ignition by cloud-to-ground lightning discharge.

Conclusion. As a result of research, we offered the new physically proved approach to forecast the forest fire danger caused by storm activity. The analysis of the methods based on statistical data shows that it is impossible to adequately assess the probability of forest fires caused by storm activity. We offered to use the deterministic models of tree ignition by a cloud-to-ground lightning
discharge and probabilistic criterion of forest fire danger assessment. The analysis of statistical approach is carried out in the territory of the Timiryazevskiy local forestry of the Timiryazevskiy forestry of the Tomsk region. Technologies of geographic information systems are used to visualize the spatial data. The program implementation of algorithms is enabled in the ArcGIS software.

**Acknowledgements.** This research implemented with financial support by Russian Foundation for Basic Research. Grant 16-41-700831.
References


Evaluation of Seismic Hazard Using Seismic Microzonation Techniques

Evgeniy Bodyakin¹, Sergey Peretokin², and Konstantin Simonov¹

¹ Institute of Computational Modeling of Siberian Branch of the Russian Academy of Sciences,
   Akademgorodok, ICM SB RAS, 660036, Krasnoyarsk, Russia
   bodyakinevgeniy@gmail.com, simonovkv@icm.krasn.ru

² Krasnoyarsk Branch Office of the Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences – Special Designing and Technological Bureau Nauka,
   Mira avenue 53, 660049, Krasnoyarsk, Russia
   ec_ropr@mail.ru

Abstract. It was shown the technique of seismic hazard assessment based on comprehensive use of methods of seismic microzonation. This technique consists of four steps. The first step is to collect geological, seismological, geophysical and topographic information. Each layer according to geological engineering survey and geophysical work are assigned physical and mechanical properties (density, limit shear stress) and the P- and S-wave velocity. Next (step 2) after visualization and examination input data using GIS technologies 3D modelling of the geological environment is performed (it is created a grid each point of which is referred to coordinates of the site). The number and depth of soil are set in each point based on geological drilling data. Then (step 3) at each point seismic intensity are calculated using instrumental methods including the method of acoustic impedance and computer simulation (GRUNT program). At the last stage according to the analysis of the results of theoretical and instrumental methods seismic microzonation map are created using GIS technologies. The procedure of constructing maps uses different methods of selection areas with the same seismic hazard (kriging, spline interpolation).

Keywords: earthquake, seismic microzoning, seismic intensity, GIS.

1 Introduction

Seismic microzonation is performed to quantify the influence of soil properties on seismic vibration within the area of specific facilities and within the residential areas. Selection of areas with different seismicity is carried out using comprehensive study of seismic properties of soils, geotechnical, hydro-geological and seismotectonic features of the territory. As a result, after performing seismic zoning the map of microzonation is created.
The technique of carrying out seismic microzonation can be divided into four steps. The first step (preparing) includes acquisition of geological, seismic, geophysical and topographic data. According this data 3D site model is created (the second step).

The third step is calculation of seismic intensity in each model point of site using instrumental and calculation methods of soil seismic properties assessment \[1, 2\]. The following methods are used when seismic hazard is estimated:

- Seismic microzonation using acoustic impedance method
- Seismic microzonation using earthquakes and explosions records
- Seismic microzonation using microtremors

These methods are normative methods and are subjected in \[2\].

The basis of numerical simulation of the geological environment reaction to strong earthquakes is the concept of the heed to take into account more than one possible earthquake effect, so ensemble of earthquakes effect with different spectral characteristics is estimated. Therefore for each model point of site numerical simulation is implemented on the full range of accelerograms.

Numerical simulation lets us to carry out calculation of peak ground acceleration, time stress and strain changes, Fourier spectra and response spectra on the site surface for a given input motion.

The final stage of seismic microzonation is creation of seismic microzonation maps that displays the regions with varying seismic intensity. Seismic microzonation maps provide the basis for seismic resistant construction, public safety, environmental protection and other actions aimed to reducing the damage caused by strong earthquakes.

2 Data acquisition

Reliability of hazard estimates directly depends on the quality and completeness of the initial engineering-geological and engineering-geophysical information in study area. In this regard, in the first step of seismic microzonation are used a complex engineering geophysical method to determine the characteristics of the ground layer, necessary to the implementation of instrumental and calculation methods of seismic microzonation.

Geological information consists of drilling data: coordinates, absolute elevation, thickness and a number of layers, the groundwater level. This information is taken from the geological engineering survey reports.

Each layer is assigned a number of engineering-geological elements (EGE). The EGE is some amount of ground with the same name-bearing type and uniform in properties and state. According to the EGE number each layer is assigned physical and mechanical properties (density, critical shear stress, etc.) and velocity of P- and S- waves. The velocities are obtained by processing and interpretation of seismic exploration materials (complex method of refracted waves, vertical seismic profiling). The abovementioned data should be presented as database or spreadsheet for further work in GIS. The example of input data is shown in Table 1.
Table 1: Example of one borehole input data

<table>
<thead>
<tr>
<th>Borehole</th>
<th>layer</th>
<th>EGE</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>depth, m</th>
<th>density</th>
<th>Vp</th>
<th>Vs</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>1</td>
<td>1</td>
<td>26436.6</td>
<td>21227.09</td>
<td>37.95</td>
<td>1.8</td>
<td>1.92</td>
<td>232</td>
<td>158</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>2</td>
<td>26436.6</td>
<td>21227.09</td>
<td>37.95</td>
<td>4.7</td>
<td>2.15</td>
<td>352</td>
<td>159</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>4</td>
<td>26436.6</td>
<td>21227.09</td>
<td>37.95</td>
<td>8</td>
<td>2.54</td>
<td>1025</td>
<td>488</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>5</td>
<td>26436.6</td>
<td>21227.09</td>
<td>37.95</td>
<td>10</td>
<td>2.72</td>
<td>2960</td>
<td>1968</td>
</tr>
</tbody>
</table>

At the same stage import and visualization of the boreholes data in the GIS environment are implemented. In addition to this information topographic data is displayed: lines of surface height, site boundaries, position of existing and planned buildings on site (Fig. 1).

![Primary visualization of the input boreholes data](image)

Fig. 1: Primary visualization of the input boreholes data

In most cases the input boreholes data may include data related to larger area than the area on which seismic microzonation is performed. Selection of the necessary data is carried out using GIS software tools which provide opportunities for the use of graphical tools, in particular, to highlight the geographical boundaries of site.

One more necessary data array for numerical simulation of the geological environment reaction to strong earthquakes is set of synthesized accelerograms or analogue accelerograms.

Calculated seismic impact can be modeled using accelerograms scaled to strength-level event (SLE) or ductility-level event (DLE). The accelerograms can be obtained from the following:
from time-history of earthquakes obtained on building site or development area
from time-history of earthquakes obtained on regions that have seismotectonic, geologic and other seismological conditions similar with area of site
synthesized time-history of earthquakes based on parameters of seismic impact for SLE and DLE

Time-history of earthquakes obtained on building site or development area scaled to SLE and DLE.

During whole period of field geophysical works in area of building site is carried out registration of seismic events from nearby seismogenic zone. The time-history of earthquakes obtained on site are scaled to SLE and DLE.

Time-history of earthquakes obtained on regions that have seismotectonic, geologic and other seismological conditions similar with area of site.

Ensembles of digital accelerograms of real seismic events obtained on regions that have seismotectonic, geologic and other seismological conditions similar with area of site are selected on the basis of SLE and DLE parameters (hypocentral distance and magnitude) obtained in detail seismic zoning. For this purpose are used databases contained time-history of real seismic events, for example the database of accelerograms of Japan seismic station (http://www.kik.bosai.go.jp/kik/search/index_en.html). When the accelerograms are selected from the database engineering-geological conditions, hypocentral distance and magnitude are taken into account.

Synthesized time-history of earthquakes based on parameters of seismic impact for SLE and DLE.

Synthesized accelerograms are calculated using programs SMSIM [4] for modeling synthesized time-history of earthquakes. The programs SMSIM are based on simulation of ground motion using the stochastic method. It is very convenient for modeling ground motion on studied site in engineering frequency band from earthquake with given parameters hypocentral distance and magnitude of earthquake. This method is taken into account regional features of seismic source as well as influence of the features of the geological environment on the propagation of seismic waves from the source to the site.

3 Building of geological 3D model of site

For the further computation on the basis of aforementioned date building of 3D model points of geological environment is carried out with certain step.

Using GIS tools a grid is created and each point of which is geographically linked height and coordinates of the site. Then each point is defined by the thickness and a number of layers based on drilling data and engineering stratigraphic columns (Fig. 2). These data are input for the subsequent calculation of seismic intensity.
4 Seismic intensity calculation

Seismic intensity calculation is carried out using acoustic impedance method and programs calculating the oscillation of surface. There are various methods for calculating the vibrations of layered soils based on linear equations [1] ; in this paper is used program GRUNT, which is based on the method of thin-layered media.

4.1 Seismic microzonation using acoustic impedance method

Instrumental evaluation of the velocity properties of site is considered as informational base for calculation of seismic intensity increment. Estimation of seismic intensity increments based on acoustic impedance method is carried out based on measuring the velocities of seismic waves and the density values in 10 meter soil thickness of studied aria and reference soil.

The calculations are carried out according to the equation:

\[ I = I_0 + \Delta I_S + \Delta I_B, \]  

(1)

where \( I \) is seismic intensity based on local conditions, \( I_0 \) is initial seismic in relation to the average ground conditions (II-category in seismic properties) according to the detail seismic zoning; \( \Delta I_S \) is seismic intensity increment because of acoustic impedance differences in the target and the reference soil:

\[ \Delta I_S = 1.67 \log(V_{(p,s)}\rho_R/V_{(p,s)i}\rho_i), \]  

(2)

where \( V_{(p,s)} \) and \( V_{(p,s)i} \) are weighted mean value of the propagation velocities of P- and S waves on the target and the reference soil, \( \rho_R \) and \( \rho_i \) are weighted mean value of density in the target and the reference soil; \( \Delta I_B \) is seismic intensity increment because of deterioration of seismic properties caused by water saturation.
4.2 Seismic microzonation using earthquakes and explosions records

Synchronous registration of earthquakes on the different parts of site provides a simple way to compare some properties of ground motion. In this approach direct problem of seismic microzonation is solved, i.e. seismic intensity increment in some areas of site selected according to engineering-geological study is estimated.

After getting synchronous time-history of earthquakes frequency-domain characteristics and seismic intensity increment with respect to the recording station, located on the reference ground are estimated.

Seismic intensity increment assessment is carried out according to the equation [2]:

$$\Delta I = 3.33 \log\left(\frac{A_i}{A_0}\right),$$

where $A_i$ and $A_0$ are the average amplitude of ground motions (from one earthquake) on the target and the reference soil respectively.

In many cases such assessment is not enough spectral analysis of the events is carried out. The seismic intensity increment is defined for the entire frequency range from 0.1 to 10 Hz, and separately for the low, mid and high frequencies.

Figure 3a shows the three-component record of ground motion of the regional earthquake registered on the base station situated on I-category soil and station 4 situated on II-category soil. Figure 3b shows the Fourier spectra of these events (root mean square was taken from three component records).

![Fig. 3: Three-component record of ground motion on the two stations (a) and Fourier spectra of records (b)](image-url)
4.3 Seismic microzonation using microtremors

This method is used as supplementary method in seismic microzonation.

To assess changes of strong earthquake intensity using peak of microtremors at a given period the following equation is used [2]:

\[ \Delta I = 2 \log\left( \frac{A_{max_i}}{A_{max_0}} \right) \]  

(4)

where \( A_{max_i} \) and \( A_{max_0} \) peak amplitude of microtremors on the target and the reference soil respectively.

All calculations are performed using the seismic processing program Geopsy ([http://www.geopsy.org](http://www.geopsy.org)). Average Fourier spectra with duration from 40 to 60 second on the interval of 10 minutes or more (depending on the recording quality noisy areas can be excluded) are calculated. All spectra are previously smoothed by the Horse and Omachi window [Konno, K. and Omachi, T., 1998, Bull. Seism. Soc. Am., 88, 228-241.] with a coefficient characterizing the bandwidth, \( b = 90 \). Each piece of the recording also is multiplied by 5% weight cosine function to reduce boundary effects.

Based on analysis of the fundamental frequency of the site maximum value of spectra is selected in a particular bandwidth. The last step is to calculate the seismic intensity increment using equation (4) relative to the base point. It should be noted that in microtremors formation along with natural sources involved numerous artificial sources, the impact of which cannot be controlled. Inability to comply with the necessary conditions of registration microtremors and strong variation of maximum amplitude values limit the use of microtremors to calculate seismic intensity increment. Thus, the use of method of microtremors registration is only suitable in combination with other instrumental methods.

4.4 Numerical simulation

One of the simplest one-dimensional soil models is KelvinVoigt viscoelastic model. Shear stress \( \tau \) in this case depends on the strain \( \gamma \) and its derivative \( \dot{\gamma} \) as shown below:

\[ \tau = G \gamma + \eta \dot{\gamma} \]  

(5)

where \( G \) is shear modulus, \( \eta \) is viscosity.

Schematic representation of KelvinVoigt model is shown in Figure 4.

The dynamics of the soil environment is described by the following equation:

\[ \frac{\partial \tau}{\partial z} = \rho \frac{\partial^2 u}{\partial t^2} \]  

(6)

Substitute the value \( \tau \) from (3) and taking into account that \( \gamma = \frac{\partial u}{\partial z} \),

\[ \rho \frac{\partial^2 u}{\partial t^2} = G \frac{\partial^2 u}{\partial z^2} + \eta \frac{\partial^3 u}{\partial z^2 \partial t} \]  

(7)
To calculate the spectral characteristics and accelerograms on the surface or in the interior of the multilayer inelastic (with absorption) media with plane boundaries is used thin-layered media (The Schmidt Institute of Physics of the Earth of the Russian Academy of Sciences). It is solved the two-dimensional problem of propagation of plane bodily waves in nonelastic layer pack with free upper boundary and underlying elastic half-space. From half-space to the lower boundary layer thickness at an arbitrary angle it is fallen P- or S- wave with unit amplitude at a given frequency, or wave with a given arbitrary shape. The model parameters are thickness and density of layers, P- and S- wave velocity, damping ratio. It can be used different models of the absorption mechanism in the medium (linear dependence of the absorption coefficient on the frequency or dependence described a linearly inelastic Gurevich model).

Output data depending on the conditions of the problem are obtained as follows:

- The amplitude-frequency characteristics of a subsurface formation according to a predetermined model
- Accelerograms on the free surface or inside the medium
- Seismic intensity increments calculated using relationship of the amplitude-frequency characteristics of the studied and reference models
- Response spectrum corresponding to the calculated accelerograms

This method is implemented in Grunt software. The algorithms have been adapted for problems of seismic microzonation allowing counting large amounts of information in an automatic mode.

Geological section is described as a set of enumerated and numbered from top to bottom layers (including the half-space) each of which has its mechanical parameters. Each layer can be divided into sub-layers of equal power with its mechanical layer parameters to determine the properties of motion in certain depth (all calculations are executed in the program for the top of layer).

Input motion is read from formatted files. The nonlinear and inelastic soil behavior under loads caused by strong movements is described by change in the modulus of elasticity and damping, which are due to deformation. Their values
are determined iteratively by leading maximum deformation to a uniform one to the entire layer.

For each model, automatic calculations are carried out using the synthesized accelerograms [3, 4], analogue accelerograms and regional records of earthquakes. The results of each calculation are numerical characteristics.

For each calculated pair (seismic event - model) are calculated set of characteristics including:

- Peak ground acceleration (PGA) expressed in g defined for a given return period (the maximum value of the modulus of the acceleration during earthquake)
- Response spectra calculated to the surface of site
- The duration of ground motion for a given return period;
- Period of oscillation with PGA;

To convert the peak acceleration in the seismic intensity the following equation is used[5]:

\[ I = 2.5 \log(PGA) + 1.25 \log(d) + 1.05, \quad (8) \]

where \( PGA \) is peak ground acceleration expressed in g, \( d \) is duration of ground motion expressed in second.

5 Creation of seismic microzonation maps

The final stage of seismic microzonation is creation of seismic microzonation maps based on analysis of the results of numerical and instrumental methods.

Creation contour seismic microzonation maps is performed using the grid with certain step (usually 25x25 meters). Nodes of the grid are assigned design parameters of seismic effects corresponding to zoning areas.

Based on the grid interpolation is performed and then surface in geotif format is received. The user may make use of various options for surface contouring (e.g. kriging, spline) offered by the GIS software. Because of the uniform grid REGULARIZED type of spline is used. This type creates a smooth, gradually changing surface.

The next step is to reclassify a raster values. The interval of raster values is assigned the mean value. The step of interval is taken equal to integer value of seismic intensity, and 0.1 of one.

After receiving the parameterized raster conversion to GIS polygons of this raster is performed. Groups of raster pixels with the same values of seismic intensity are combined into a single polygon with assigning this value of seismic intensity (Fig. 5)
6 Conclusion

The developed technique facilitates and accelerates research associated with seismic microzonation.

Creation of the 3D geological environment model lets us to increase the accuracy of defining seismic intensity of site.

Seismic intensity calculation is perfumed using instrumental methods and numerical simulation. GIS is the primary tool for creating seismic microzonation maps which is the final step of developed computational technique.

References

2. RSN 60-86 Engineering survey for building. Seismic microzonation. Standards of production work.
Abstract. This work is devoted to development of hybrid geoinformation system for making forecasts of possible flooded areas in the downstream water of the Irkutsk hydroelectric station with damages evaluation in conditions of extreme water content for lake Baikal and for effluents of the Angara river. For 3D model construction of the Angara river bed they used Atlas Map of the Angara River in Irkutsk hydroelectric power station up to 142 km on scale of 1: 10000 General plan of Waterways of the East Siberian basin ”, Irkutsk, 1994. Due to lack of a digital model (map), a paper-based map was digitized using vectorizer program Easy Trace 7.99. Depth contours, marks and bank lines were also digitized. To combine the data about land and underwater terrain a specific software was developed using Delaunay triangulation method. The role of data layers in the construction of triangulation is specified. The layers contain information about: land relief, standing water levels, shore lines, underwater topography. Auxiliary triangulation is made according to the data whereof further information is revealed regarding heights levels to which shore contours are attached and in reference to which depths are measured. Hybrid geographic information system allows to simulate different scenarios of flooding and to determine the flood zones with account for underwater and land relief.

Keywords: GIS, flooding, lake Baikal, river Angara, surface.

In the course of the project execution on calculation of flooding areas in the downstream of the Irkutsk hydraulic power station it was necessary to get a terrain model that takes into account the shape of the river bottom. The land relief with different mipmaping levels is shown on topographical vector electronic maps of different scales but the information on underwater terrain of the Angara River in soft copies wasn’t found. Existing data of some navigation systems on the depths are fragmented and experts comments are not precise. The only reliable source is an atlas that is distributed in paper form. In this regard the need of presented maps digitizing and its combining with the existing topographic base has arisen. Digitizing of paper Atlas of the Angara river maps was performed with the help of freeware Easy Trace 7.99 Vectorizer of cartographic representations. The following data were digitized: depths curves, marks of depth points, the coastline.
The Atlas sheets are randomly oriented aiming to the most close-together arrangements of the river sections on the pages. To use it further, we made rotation and shift to the maps reference system containing a terrain model. To select the rotation parameters we compared characteristic points of coastlines contours (capes, bays, islands tips, etc.) of two maps. Utility program was implemented, that calculates the transformation parameters for the given list of coordinates pairs of points and performs maps transformation. To find the transformation parameters they use the least square method. To use the method it is sufficient to specify two pairs of corresponding points, but the use of a larger number of them allows you to get a more exact result, as well as to evaluate the accuracy under the mean square deviation. The mean square deviation of the given examples in this case was 25-70 m. The deviation is due to the presence of significant divergence of contours, because these differences may affect also on characteristic points used for setting of maps conformance.

Attempts to combine contours of vectorized coastline to coastline of topographic base showed that the used atlas was prepared schematically without reference to any map (Figure 2). The atlas reflects the characteristic curves of the shoreline, but it is impossible to combine this line with a more precise contour by turning and shifting. User experience of the calculated under matched characteristics points of rotation and shift showed that the obtained result can not be significantly improved by clarifying the transformation settings. In order to make possible use of not enough precise data for terrain modeling software for electronic cards morphing have been made. To perform morphing it is necessary to find a continuous transformation of the plate area that can combine inexact contours of coastlines with more accurate ones. Layers of coastlines are used to align the maps because they exist on both cards. After that, the same transformation is applied to the other layers of the underwater contours maps (isobathic lines and depth marks). The resulting data are better consistent with the information about land relief, at least, transformed isobathic lines and depth marks are not beyond the shoreline. The transformed coastlines contours are not used as there are more exact coastlines available, so their deviation from the target lines is admissible, for example, by adding alignment interdiction. Principal is that the zones of contours deviation do not contain data of used layers of an inaccurate map. To perform a map morphing two separate software modules are implemented: for semi-automatic alignment parameter setting and for transformation performing.

The contours combination is implemented in the form of a special task in the program of electronic maps viewing IrkGV (Figure 3). While the task is performing the operator sets the directions, combining the maps of starting (inaccurate) and target (exact) maps. Then the contours combining starting and ending points of drawn by the user arrows are imaged on the closest to them contours points of relevant maps. Further they make a reflection of every contour line of the object of original map to the specified contour area of target map for each pair of arrows, that are adjacent to both contours. Thus, if, for example, to add an arrow combining this point to the third contour between the points of one
Fig. 1. Digitization result.

Fig. 2. Maps comparison.
Fig. 3. Contours alignment parameter setting. Accurate map is displayed with filling, movable as boundary representation.

Fig. 4. Maps overlap procedure.

contour the comparison of contours areas stops. Comparison of two closed loops is performed separately. Without regard to this case it would be necessary to set at least three pairs of corresponding points to connect contours lines totally. While selecting lines for comparison on less number of corresponding points it is needed to be taken into account direction of tracing; they compare contours areas that have the same direction of tracing. This allows to set correlation between contours using only one pair of points. Displaying of compared contour plots points is performed using a linear transformation of parametric coordinates of curves. In case if the obtained result is not satisfied, in order to ensure combination of some characteristic contour points it is necessary to match explicitly these points together. After entering the information on the maps alignment they form a displacement file that in addition to the explicit shifts contains displacements calculated for the intermediate contour points. Blockage of overlapping circuit plots is required when they compare plots represented with different refining degree. For example, there is a specified precise contour of bay, that is absent in
the less precise. In this case it is better to avoid comparison of corresponding bay plot with a straight line of a transformable contour. The corresponding operation is implemented in the program of triangulations construction to transform morphing. At that triangulation with constraints is formed according to the original points from the file of displacements. Hard edges are added for the nearly located

![Segment of formed triangulation, containing information on land and underwater terrain.](image1)

**Fig. 5.** Segment of formed triangulation, containing information on land and underwater terrain.

![Combined terrain (altitude is increased for clarity).](image2)

**Fig. 6.** Combined terrain (altitude is increased for clarity).
points of the same contour. Shift vector to the target point is stored additionally at every point of triangulation. After that, layers processing is performed in Shape format: each point is moved to the vector obtained by linear interpolation from the points of triangulation. For matching data on above and underwater terrains they modify form of a triangulation construction by maps layers. Now the role at construction of the triangulation is indicated. A layer can contain the following data: land relief, water edge mark, contours of coastlines, underwater terrain. To work with underwater terrain it is required to provide information about water edge marks. According to the data we construct auxiliary triangulation from which further information is extracted regarding altitude marks, to which coastline contours are attached and with regard to which depths are measured. To calculate flooded areas they use contour lines construction according to triangulation, that points altitude are determined by the difference in altitudes of points on the original terrain and are calculated with level of water in case of flooding. This approach allows to get more accurate results and to cover larger area than construction of contour lines according to terrain model for the altitude equal to the level of water at any point increased for the height of water rise at this point that is often used in such cases.

Formed terrain model is stored in a triangulation file (the files extension is .trg) and can be used to perform further analysis using dynamic library TrigLib.dll, designed to work with this data. This library allows to obtain the height of the terrain from triangulation at any point and to perform line tracing, i.e. to receive terrain section along any segment or a broken line. In addition to the basic triangulation, containing information about the terrain, auxiliary triangulation built according to water edge marks is also formed. It contains information about the water level surface used for the construction of a generalized terrain model. The program of triangulation construction includes morphopro file generation algorithms under the specified in the file coordinates, terrain contouring under the given triangulation, and subtraction of altitudes given by one triangulation from the heights of the other points of triangulation. Thus, for the construction of flood zones they make triangulation that takes into account the calculated water level. On the basis of dynamic library TrigLib.dll use we developed the program for altitude marks forming according to a randomly given sort of the Angara section.

**Conclusion.** As a result of the performed works on evaluation of extreme floods under different scenarios of extreme water availability in the basin of the Angara river and lake Baikal, they created hybrid geoinformation system, which allows to simulate different scenarios of floods and to identify flooded areas considering underwater and land relief.

**Acknowledgments.** The study was partially supported as part of integration project 4.10: "Development of unified methodological approaches in determining of anthropogenic impact on the surface water in the area of subaqueous discharge of contaminated sewage of soil water (using as example the area of Listvyanka village)" and RFBR grants 16-07-00411-a, 16-57-44034-mong-a, 14-07-00166-a.
References


Methods for High Performance Graph Processing

Mikhail Chernoskutov
Krasovskii Institute of Mathematics and Mechanics,
Ural Federal University,
Yekaterinburg, Russia
mach@imm.uran.ru

Abstract. This paper describes two methods for accelerating the processing of a large graph distributed in the memory of multiple nodes. The first allows to substantially reduce overheads connected with data transfer between different nodes. The second is designed to reduce workload imbalance amongst computational threads. Both methods are integrated into a breadth-first search algorithm and more than triple its performance.

Keywords: data intensive applications, graph processing, parallel algorithms.

1 Introduction

Graph algorithms are used in various fields of science and engineering applications. In many cases, big graphs can be processed in parallel by computational systems with multiple nodes [1]. However, the parallelization efficiency is impaired by intensive memory access and unknown (in general) exact location of data on systems with distributed memory. These obstacles turn graph algorithms into typical data-intensive applications [2].

Intensive memory access pattern is a crucial bottleneck in implementations of parallel graph algorithms because (in many cases) there is only a small number of computational operations in such classes of algorithms. On the other hand, graph algorithms tend to have a lot of operations related to access to small pieces of data in different parts of memory. Thus, they are less demanding as far as CPU capabilities are concerned but require much in the way of efficiency and bandwidth of the data transfer bus.

The fact that the distribution of data is not known in advance dramatically complicates efficient implementation of multi-node parallel implementations of graph algorithms. In the general case, the program might have to search for the data on some particular vertex across all the computational nodes.

The object of this research is a parallel breadth-first search algorithm on graphs with small diameter (generally, no more than 10) and skewed degree distribution. Such graphs arise in, for instance, social networks analysis, various mathematical and physical simulation applications [3], etc. The main feature of
these graphs is a relatively small number of vertices with highest degrees and large number of vertices with small degrees (with only few incident vertices).

2 Parallel Breadth-First Search

Breadth-first search may be parallelized with the aid of level-synchronous algorithms. These algorithms process each level (or iteration) separately and independently from each other. It means that (in case of breadth-first search) processing of the level $N + 1$ begins after the processing of the level $N$ has been finished, while each of these levels (i.e., all vertices and edges of the same level) may be processed in parallel.

Presently, there are two most common types of level-synchronous breadth-first search algorithms:

- direct traversal (top-down approach);
- inverse traversal (bottom-up approach [4]).

The direct graph traversal is the standard version of breadth-first search algorithm of this kind. It assumes that the vertices that are active on the current iteration would mark all their neighbors. The pseudocode of a parallel version of breadth-first search with the top-down traversal direction is presented on fig. 1.

In lines 1–4, the array of distances to source vertices is initialized. The initialization is followed by the main loop, which repeats while there exist unmarked vertices in the graph. First, in lines 6–14, the vertices stored at the current node are marked. Then, in lines 16 and 17, messages to other nodes are sent and received (point-to-point); these messages contain the data on vertices that must be marked on those nodes. Finally, in lines 19–22, the vertices the data on which was received from other nodes are marked. At the end of each iteration, the level counter in line 23 is incremented by one and then, at line 24, the algorithm checks if there are still unmarked vertices.

The bottom-up implementation of breadth-first search algorithm assumes that inactive vertices will be looking for active vertices amongst their neighbors. In case of presence of such type of vertices on the current iteration, the vertex is to mark itself. The pseudocode of parallel breadth-first search with inverse traversal is presented on fig. 2.

Like in the previous pseudocode, in lines 1–4, the breadth-first search is initialized. In lines 18 and 19, the level number is updated and it is tested if the algorithm is to terminate. However, there is a big difference in the vertex marking procedure—in case of bottom-up traversal, it is crucial to know the current state of all active vertices in the graph. It is most convenient to do this by means of a bitmap the length of which equals the number of vertices in the graph. Therefore, at each iteration, the data is synchronized by means of updating the bitmap through collective MPI communications (line 15 and 16).
for each u in dist
dist[u] := -1
dist[s] := 0
level := 0
do
    parallel for each vert in V.this_node
        if dist[vert] = level
            for each neighb in vert.neighbors
                if neighb in V.this_node
                    if dist[neighb] = -1
                        dist[neighb] := level + 1
                        pred[neighb] := vert
                    else
                        vert.batch_to_send.push(neighb)
                end
            end
        end
    send(vert.batch_to_send)
    receive(vert.batch_to_receive)
end
parallel for each vert in vert.batch_to_receive
    if dist[vert] = -1
        dist[vert] := level + 1
        pred[vert] := vert.pred
        level++
    end
while(!check_end())

Fig. 1. Parallel top-down breadth-first search algorithm pseudocode

for each u in dist
dist[u] := -1
dist[s] := 0
level := 0
do
    parallel for each vert in V.this_node
        if dist[vert] = -1
            for each neighb in vert.neighbors
                if bitmap_current.neighb = 1
                    dist[vert] := level + 1
                    pred[vert] := neighb
                    bitmap_next.vert := 1
                    break
            end
        end
    all_gather(bitmap_next)
    swap(bitmap_current, bitmap_next)
    level++
while(!check_end())

Fig. 2. Parallel bottom-up breadth-first search algorithm pseudocode
3 Performance Engineering of Parallel Breadth-First Search

To increase the performance of parallel breadth-first search algorithms, one could suggest the following two methods:

– hybrid graph traversal;
– workload distribution across computational threads.

3.1 Hybrid Traversal

This method is a combination of different types of parallel breadth-first search algorithm for executing different iterations. In particular, the top-down approach characterized by large computational workload and data transfer overheads on iterations in the middle. At the same time, the first and last iterations in the top-down approach execute faster and have almost no data transfers. The situation is different with the bottom-up approach due to the use of collective MPI data transfer operations. In its case, data transfer overheads are almost the same on each iteration. However, marking the vertices on the first iterations takes much more time than on the iterations in the middle or later.

Taking into account the fact that the data produced on each iteration is the same (regardless of traversal direction), we suggest to combine different types of graph traversal (with smallest possible data transfer overheads on each iteration) to achieve maximal performance. In this study, we propose the following scheme:

– top-down on first two iterations;
– bottom-up on next three iterations;
– top-down on all other iterations.

3.2 Workload Distribution

In processing of graphs with skewed degree distribution, it is not known in advance (in the general case) which vertices will be processed by a particular computational thread, the only thing that may be known in advance is, for instance, the total number of vertices that have to be processed by the current thread. However, the total workload is determined not by the number of active vertices but by the number of edges incident to them. This leads to workload imbalance amongst threads and big overheads during the runtime of parallel level-synchronous algorithm.

To avoid workload imbalance, we suggest a transition from “looking” through a vertex array to “looking“ through an array of edges. For this purpose, we logically divide the edges array into equal pieces holding max_edges elements. Each thread determines the corresponding vertex for all edges in every part of the edges’ array by using the part_column array, which contains the numbers of vertices incident to the first edge in the corresponding part of the edges’ array. The pseudocode for parallel filling of the part_column array is presented on fig. 3.
parallel for each i in V. this_node
first := V. this_node[i]
last := V. this_node[i + 1]
index := round_up(first / max_edges)
current := index * max_edges
while (current < last)
    part_column[index] := i
    current := current + max_edges
    index++

Fig. 3. Parallel filling of part_column array pseudocode

Pseudocode of a new version of breadth-first search algorithm that uses the part_column array is presented on fig. 4 (the top-down approach) and fig. 5 (the bottom-up approach).

4 Benchmarking

Both of the aforementioned methods were incorporated into a custom implementation of the Graph500 benchmark [5]. The kernel of this benchmark represents parallel breadth-first search on a big graph distributed in the memory of multiple nodes. The size of the graph is determined by the scale parameter, which is the logarithm base 2 of the number of vertices in the graph. The average degree of all vertices is 16. The main performance metric of this benchmark is the number of edges traversed per second (TEPS).

We developed a custom implementation that uses MPI (one process per computational node) for sending and receiving messages across all nodes and OpenMP (eight threads per node) to deal with shared memory within each node.

Benchmarking was carried out for graphs with different sizes on 1-, 2-, 4-, and 8-node configurations of the Uran supercomputer (located at the Krasovskii Institute of Mathematics and Mechanics). Each node had Intel Xeon X5675 CPU and 46 GB DRAM. Performance of the custom implementation was compared with the following reference implementations provided by Graph500:

- simple (represents top-down breadth-first search);
- replicated (represents bottom-up breadth-first search).

Benchmarking results are presented on fig. 6. As seen on the figure, our custom implementation substantially outperforms the simple and replicated implementations. In addition, in case of 8 nodes, it is clearly seen that the custom implementation scales much better than its counterparts (scalability of all implementations presented on fig 7).
1 // preparation...
2 parallel for each i in part_column
3     first_edge := i*max_edges
4     last_edge := (i+1)*max_edges
5     curr_vert := part_column[i]
6     for each edge in [first_edge; last_edge)
7         if neighbors of curr_vert in [first_edge; last_edge)
8             if dist[curr_vert] = level
9                 for each k in neighbors of curr_vert
10                    if dist[k] = -1
11                       dist[k] := level + 1
12                       pred[k] := curr_vert
13                       curr_vert++
14     // data synchronization...

Fig. 4. Parallel top-down breadth-first search algorithm pseudocode (with workload balancing)

1 // preparation...
2 parallel for i in part_column
3     first_edge := i*max_edges
4     last_edge := (i+1)*max_edges
5     curr_vert := part_column[i]
6     for each edge in [first_edge; last_edge)
7         if neighbors of curr_vert in [first_edge; last_edge)
8             if dist[curr_vert] = -1
9                 for each k in neighbors of curr_vert
10                    if bitmap_current.k = 1
11                       dist[curr_vert] := level + 1
12                       pred[curr_vert] := k
13                       bitmap_next.vert := 1
14                       break
15     curr_vert++
16 // data synchronization...

Fig. 5. Parallel bottom-up breadth-first search algorithm pseudocode (with workload balancing)
Fig. 6. Performance comparison

Fig. 7. Scalability comparison
5 Conclusion

Attempts at efficient parallelization of the breadth-first search algorithm with skewed degree distribution are hampered by the workload imbalance amongst computational threads and large amounts of data transfer only on few select iterations of the algorithm. This forms a bottleneck that makes it much more hard to make a high-performance implementation for this algorithm.

In this paper, we suggest a couple of methods for workload balancing and traversal hybridization, which allow to increase the performance (it is more than three times higher) of the parallel level-synchronous breadth-first search in comparison with the reference top-down and bottom-up procedures.

In our future work, we intend to focus on the research in scalability the suggested algorithm and testing it on graphs obtained from real-world applications. Another important task is to modify our custom implementation to use computational accelerators to improve its performance.

Acknowledgments. The reported study was partially supported by RFBR, research project No. 14-07-00435. The experiment was performed on the Uran supercomputer.

References

Improving the Efficiency of the Human Spine Diagnostics Systems

Nikolay Dorofeev, Konstantin Podmasteriev, Oleg Kuzichkin, and Anastasia Grecheneva

Murom Institute (branch) Vladimir State University, Murom, 602200, 23 Orlovskaya st., Russia
DorofeevNV@yandex.ru
http://www.mivlgu.ru

Abstract. The paper proposes a method of recording and measuring the spatial mutual movement and positioning of the spine segments. As a basic method, an application of the accelerometer measuring method of the rotation angles and acceleration is proposed. The research results show that the use of accelerometer sensor significantly increases the flexibility of the system (scalability and ease of interfacing the accelerometric diagnostic and rehabilitation system of the spine with other diagnostic systems) and the accuracy of the results (thanks to additional digital processing of recorded signals) due to the small size and high precision of sensors. The proposed method is passive (does not use additional exposure, such as in an ultrasound, X-ray and tomography) and has no negative impact on patient.

Keywords: biomechanics, human spine, diagnostic system, goniometric control, accelerometry.

1 Introduction

One important area of medical technology is diagnostic and rehabilitation equipment, especially spinal diagnosis and rehabilitation system [1]. Various deviations in the relative position of the spine segments affect the functionality of the musculoskeletal system, internal organs, eventually triggering a “chain reaction” of disorders of organs and systems of the body as a whole [2, 3].

To figure out timely corrections of the deviations in the spine, and preventing the formation of irreversible pathological processes, a high accuracy primary diagnosis is required. It should be noted, that during the period of spinal rehabilitation to assess the treatment and recovery results, an investigation of patient’s motion is required.

2 The current state of medical diagnostics of the spine

In the existing medical diagnosis methods, the presence of pain syndrome is a key reason for the prescription of rehabilitation procedures [4]. Pain is the basis
for hospitalization, medical appointments actions aimed at rehabilitation, and often for the adoption of expert solutions (diagnosis). Therefore, the prescriptions for rehabilitation are assigned to the patients that have a partial violation or substantial loss of functionality, i.e., spine pathology with pain syndrome and neurological manifestations. In practice, this means one must carry out rehabilitation effects only on the patients with acute syndromes and during recovery periods.

Diseases of the spine are accompanied by impaired mobility of vertebral joints. For example, studies were carried out on soldiers of the Air Force of the Ministry of Defense of Russia. The soldiers’ physical state were characterized by a high degree of physical fitness and health requirements. The results of the studies reveal a relatively high level of disorders of the spine functions in 56% of the cases. The same violations of the biomechanics of the spine were revealed in 61.3 ± 3.5% of the patients, with 19% of the violations were of moderate to severe degree [5].

For this reason, an important part of solving the problem of diagnosis of spinal diseases and their treatment is the development of an automated method of goniometric spine control. The method are to provide a high accuracy and sensitivity, and must be easy applied in practice for estimation of the functionality of the spinal column.

3 The accelerometer method of geometric spine control

We propose to use accelerometer method of bending angle measurement for assessing spinal configuration and range of its motion in the sagittal, frontal and horizontal planes as a method of goniometry in the automated control system [6]. This instrumental method of the investigation of the curvature and mobility of the spine is built on the principle of measuring the full acceleration vector of two accelerometers attached to the vertebrae controlled spinal segments (Fig. 1).

The output of the four signals proportional to the acceleration of the general point of biokinematic pairs $\ddot{a}$ are obtained as a result of measurement of the acceleration values of each accelerometer in two coordinate system.

\[
\begin{align*}
 a_x &= K_{ax}a \cos(\phi_A); \\
 a_y &= K_{ay}a \sin(\phi_A); \\
 b_x &= K_{bx}b \cos(\phi_B); \\
 b_y &= K_{by}b \sin(\phi_B).
\end{align*}
\]

(1)

where $a_x$, $a_y$, $b_x$, $b_y$ are the acceleration values of the first and the second accelerometers in the plane 2D coordinate system; and $\phi_A$, $\phi_B$ are the angle between the direction of the acceleration vector of a general point $O$ about a pair of adjacent vertebrae $\ddot{a}$ and measuring systems $(\ddot{x}_A, \ddot{y}_A)$ and $(\ddot{x}_B, \ddot{y}_B)$ the accelerometer, respectively; $K_{ax}$, $K_{ay}$, $K_{bx}$, $K_{by}$ are transform coefficients of the corresponding accelerometers.

The goniometric angle $\varphi = \phi_A - \phi_B$ is determined on the basis of relations between the components of the linear acceleration vector in the spine movement and displacement of the accelerometers. The angle is described by the formulas:
Fig. 1. The accelerometer method of geometric spine control

\[ \varphi_A = \cos \left( \frac{(b_x a_x)/(K_{bx} K_{ax}) - (b_y a_y)/(K_{by} K_{ay})}{(a_x/K_{ax})^2 + (b_x/K_{bx})^2} \right) \] (2)

\[ \varphi_B = \sin \left( \frac{(b_x a_y)/(K_{bx} K_{ay}) - (b_y a_x)/(K_{by} K_{ax})}{(a_x/K_{ax})^2 + (b_x/K_{bx})^2} \right) \] (3)

As a consequence, the desired value of the angle of rotation of the kinematic pair is determined from the expression

\[ \varphi = \arctg \left( \frac{a_y b_x - b_x a_x}{a_x b_x - a_y b_y} \right) \] (4)

However, in the process of calculation the resulting error might increase due to a possible division by zero in the trigonometric arc tangent function argument. Such an error occurs due to the indeterminacy of the patient motion, and, consequently, it is a result of receiving different acceleration values, including zero.

The solution of the problem is to use a phase-measuring method [7], whose implementation is to compensate the multiplicative instability of the accelerometers branches. Signals from the two-component accelerometers is converted into the phase sine wave by multiplying the signals \( a_x, a_y, b_x, b_y \) on phase quadrature signals \( \sin(\omega t) \) and \( \cos(\omega t) \), whose frequency is a multiple of the frequency of the reference generator (RG). The obtained signals will be as follows:

\[ a_x = U \sin(\omega t) K_{ax} a \cos(\varphi_A); \quad a_y = U \cos(\omega t) K_{ay} a \sin(\varphi_A); \]
\[ b_x = U \sin(\omega t) K_{bx} a \cos(\varphi_B); \quad b_y = U \cos(\omega t) K_{by} a \sin(\varphi_B), \] (5)

where \( U \) and \( \omega \) are the amplitude and the frequency of the quadrature phase signals.
Summing the signals in the adders for the object $A$ and object $B$, respectively, we obtain:

$$a_A = U K_{ax} a \cos(\omega t + \varphi_A + \varphi_{Kax}); \quad a_B = U K_{bx} a \cos(\omega t + \varphi_B + \varphi_{Kbx}).$$

(6)

where $\varphi_{Kax}$ and $\varphi_{Kbx}$ are the phases of misalignment of the measuring branches.

From the relations (5) and (6), the signal value at the input of the main phase detector will be as follows:

$$U_A = U K_{ax}(1 + \Delta K_1) a \cos(\omega t + \varphi_A + \varphi_{Kax});$$

$$U_B = U K_{bx}(1 + \Delta K_2) a \cos(\omega t + \varphi_B + \varphi_{Kbx}).$$

(7)

Since the serial connection circuit is selected, the transform coefficients of accelerometers are $K_{ax} = K_{ay}$ and $K_{bx} = K_{by}$, respectively, hence the phase error of the measuring branches $\varphi_{Kax} = \varphi_{Kbx} = 0$. As a result, phase detector generates a signal on the output. The signal is proportional to the angle of inclination of the pair of vertebrae, and it is finally independent of the instability influence of the coefficients of the measurement branches.

4 The algorithm of collecting and pre-processing the goniometric control data

Implementation the accelerometric method of rotation angle measuring in the goniometric control systems is based on an algorithm collecting dynamic data, which is based on the direct conversion of signals from two-component accelerometers in phase sine wave by multiplying the signal on the phase quadrature signals (PQS) with a frequency, which is a multiple of the reference oscillator frequency (RG) (Fig. 2) [8].

![Fig. 2. The algorithm of collecting and pre-processing goniometric control data](image)

According to the algorithm, the rotation angle of the biokinematic couples in the accelerometric goniometer is determined by the phase difference between
the measured and the reference signals. Therefore, a signal proportional to the angle of biokinematic pairs without affecting instability coefficients transmitter branches is generated at the output by summing the received harmonic signal.

The multiplicative error is eliminated by means of hardware implementation of the phase-measuring method, namely, by limiting the signal level by a limiter circuit and following detection of the phase signal [9].

5 The error compensation method of the digital accelerometers

The measurement error is determined at the error measurement phase, which is a part of the application of the phase method for measuring turning angle in the accelerometric goniometer. The measurement is based on the direct conversion of signals from two-component accelerometers in the phase sine wave. A sampling error occurs since the phase measurement is conducted at the discrete time moments [10]. The sampling error is depended on the frequency ratio of the reference sine wave produced by programmable generator and the sampling frequency of the signal obtained from the accelerometric goniometer, with an algorithm of approximation of the target signal being applied between adjacent sample time moments. Figure 3 shows the sampling error data of the accelerometric goniometer signal.

![Fig. 3. The sampling error of the accelerometric goniometer signal](image)

Traditionally, to calculate interval approximation between the samples $U_i$ (at the time moment $t_i$) and $U_{i+1}$ (at the next time moment $t_{i+1}$), algorithms of linear approximation are used:

$$U(t) = kt + b,$$  \hspace{1cm} (8)

but the real process is described by

$$U(t) = U_m \cdot \sin(\omega t + \varphi),$$  \hspace{1cm} (9)
where $\omega = 2\pi f$ is the frequency of reference oscillator, $\varphi$ is a variable phase. The approximation coefficients are determined by the following relations:

$$k = \frac{U_{i+1} - U_i}{t_{i+1} - t_i}, \quad b = \frac{U_i \cdot t_{i+1} - U_{i+1} \cdot t_i}{t_{i+1} - t_i},$$

(10)

The phase is determined on the base of the equation (10) by the time shift $\tau$, forming an error of determination of the angle of accelerometric goniometer through the phase error $\Delta \varphi$:

$$k(t_i + \tau(\Delta \varphi)) + b = 0.$$  

(11)

The time period expression that determines the measured phase on the base of $n$-periods of measurement is as follows:

$$T = \frac{(2\pi n + \varphi)}{\omega} = i \cdot \Delta t - \frac{U_i}{U_{i+1} - U_i} \cdot \Delta t,$$

(12)

where $f_d = 1/\Delta t$ is the sampling frequency.

Assuming that the frequencies are multiples of each other, the ratio of the oscillator and the sampling frequencies are defined by:

$$\frac{F}{f_d} = m.$$  

(13)

From (12) and (13), the expression for phase determination, which is based on the linear approximation, takes the form

$$\varphi = 2\pi \left( m \cdot i - \frac{U_i}{U_{i+1} - U_i} \cdot m - n \right)$$

(14)

The rotation angle of the biokinematic pair in a accelerometric goniometer is measured by the difference of $U_i$ and the reference phase signal $U^0$:

$$\alpha = \varphi - \varphi_0 = 2\pi \cdot m \left( \frac{U_i}{U_{i+1} - U_i} \cdot \frac{U^0_i}{U^0_{i+1} - U^0_i} \right)$$

(15)

For this case, the error in the determination of the angle for the real values of phases is determined on the base of direct calculation

$$\alpha = \arctan \left( \frac{U^0_i \sin(2\pi m)}{U^0_{i+1} - U^0_i \cos(2\pi m)} \right) - \arctan \left( \frac{U_i \sin(2\pi m)}{U_{i+1} - U_i \cos(2\pi m)} \right).$$

(16)

In this case, the analytical expression for estimating the angle measurement error in the accelerometric goniometer is as follows:

$$\max(\Delta \alpha/\alpha) = \frac{\sqrt{2\pi \cdot m - 1} - \arctg(\sqrt{2\pi \cdot m - 1})}{\pi \cdot m}.$$  

(17)

As seen from the expression (17), the angle error can be reduced by increasing the sampling rate with respect to the oscillator frequency.
6 The system of goniometric spine control

In the standard spine goniometry, there are so-called the support points, which correspond to the ends of the spinous processes of S4, L4, Th7 and C7 vertebrae [11].

The following algorithm provides a differentiated picture of movements in the different parts of the spine. At same time, due to the small size and high precision of accelerometer sensors and system flexibility, the reliability of the results are significantly increased especially thanks to an additional digital processing the recorded signals. In addition, our proposed approach is passive, i.e., it does not use additional exposure, such as in ultrasound or X-ray method, and has no negative effects on human body. Moreover, the accelerometer module interface can be coupled with other system modules because of compactness of the accelerometer module.

Existing systems of biomechanics spine control measure only kinematic parameters of the skeletal system without taking into account the patient’s neurophysiological parameters [12]. So, the process of diagnosis and rehabilitation is slower due to the absence of biofeedback data. Also, recording the parameters of evoked potentials makes it possible to determine the pain threshold more accurately, and to identify possible causes of motor neurophysiological abnormalities [13]. In general, the adaptive system of goniometric automated control can be represented as a block diagram shown in Figure 4.

![Fig. 4. The block diagram of hardware and software support of the automated goniometric control](image-url)
After the synchronous processing, the recorded parameters form time series, which are visualized with various degree of detail.

The time series are the basis of a model of patient. The model is processed by a neural network and is stored in the model data base. The model that most closely matches the time series is instantiated. Pain thresholds and threshold of sensitivity of the patient for generating control signals to the actuators are determined by neural network algorithms. This is possible via a feedback method (patient’s reactions to test stimuli). Based on the processed data, operation mode of the actuators is generated and selected from a database of test techniques.

It should be noted that the above adaptive system of goniometric control includes both stationary and mobile measuring systems. The number of monitored parameters is determined by the severity of the patient’s pathology. In case of injuries or low severity of the scheduled examination, the use of portable goniometers alone is sufficient, as it guarantees the freedom of patient’s movements. If the presence of more serious violations in the musculoskeletal system functioning is suspected, the use of the accelerometric goniometer coupled with X-ray and tomography is recommended.

7 The choice of informative variables for constructing diagnostic models

The main problem of a model determination for diagnosis of the patient’s health status is selection of informative variables.

In order to form adequate model instances, the sample data should be representative, the data should completely and correctly reflect the diagnostic object. The representativeness of the samples can only be achieved by selecting data objectively.

The sample data is represented as a matrix whose dimension is $N \times (n + 1)$:

$$W_N = [X_n|Y_m],$$

where $N$ is the number of cases (matrix rows); $n$ is the number of independent informative input variables; $m$ is the number of depended informative output variables; $X_n\{X_1, X_2, ..., X_i, ..., X_n\}$, $Y_m\{Y_1, Y_2, ..., Y_j, ..., Y_m\}$ are sets of vector values of inputs and outputs; $X_i\{x_{i,1}, x_{i,2}, ..., x_{i,N}\}$, $Y_j\{y_{j,1}, y_{j,2}, ..., y_{j,n}\}$ : $x_{i,N}$, $y_{j,N}$ are $i$-th value of the input and $j$-th value of the output variable of $N$-row sample data matrix.

Each data row contains sample values of the input (information on chronic diseases, data from patient history) and output (current symptoms, the results of goniometry (accelerometry)) informative variables describing the dynamic state of a particular patient. Sampling should include a training data sample, which is used at the stage of model construction of the real diagnosed object. An important condition for the use of samples is that the stored learning sample data set and testing data set must be different. This ensures reliability of the diagnostic decisions.
The quality of the learning data samples obtained in the spine diagnostic systems is evaluated according to the following criteria: representativity, informativity, and reliability.

### Table 1. Ranges of standards physiological fluctuations and degrees of spinal disorders in the goniometric studies

<table>
<thead>
<tr>
<th>Position of the spine</th>
<th>Angle</th>
<th>Norm</th>
<th>Degree of functional disorders</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>slight</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Less</td>
</tr>
<tr>
<td>Free vertical position, *</td>
<td>α</td>
<td>7-13</td>
<td>5-6</td>
</tr>
<tr>
<td></td>
<td>β</td>
<td>10-15</td>
<td>8-9</td>
</tr>
<tr>
<td></td>
<td>γ</td>
<td>9-14</td>
<td>7-8</td>
</tr>
</tbody>
</table>

| Maximum flexion, *     | α    | 60-80| 51-59 | 81-89   | 41-50 | ≤40     |
|                       | β    | 90-115| 81-89 | –       | 61-68 | –       |
|                       | γ    | 130-155| 121-129| 111-120| ≥110  | –       |

| Maximum extension, *    | α    | 0-3  | 4-6   | –       | 7-8   | ≥9      |
|                       | β    | 35-52| 25-34 | –       | 17-24 | –       |
|                       | γ    | 36-50| 26-35 | –       | 16-25 | –       |

| The slopes of the sides, * | bs   | 30-40| 20-29 | –       | 10-19 | –       |
|                           | bd   | 30-40| 20-29 | –       | 10-19 | –       |

To determine the degree of correlation between the informative variables, *i.e.*, to assess the informativity of sample data, rank correlations are used. The degree of the relationship is determined by the value of the correlation coefficient, which can range from -1 to +1 inclusive.

To calculate the minimal sample size, the following formula is used:

\[
N = n_{inf} \cdot (P \cdot \alpha / \eta^2),
\]

where \( n_{inf} \) is the number of informative variables, \( P \) is the representativeness of the sample data; \( \alpha \) is the required representativeness of the sample data; \( \eta \) is the resulting information content of the sample data.

The input sample data set consists of training and control subsets having the ratio of 2:1 and separated randomly to the subsets.

Usage of the accelerometer measurement method as a basis for the hardware and software system implementation of goniometric spine control requires continuous accounting assigned statistical range of physiological fluctuations of parameters with respect to the biokinematic norm and degrees of functional disorders of the spine (Table 1), as the system is intended to be used for optimization of medical biomechanical examination.

Selection of informative variables for the construction of multi-level diagnostic model is based on the medical reference book data (Table 2). These variables are to reflect the relations between the manifestations of a disease from the current symptoms and reported violations.
Table 2. Ranges of standards of physiological fluctuations and degrees of spinal disorders in the goniometric studies

<table>
<thead>
<tr>
<th>Vertebra</th>
<th>Parameters of vertebra</th>
<th>Parameters of segment</th>
<th>Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>b, mm</td>
<td>Fb, °</td>
<td>Gb, °</td>
</tr>
<tr>
<td>C2</td>
<td>66.3±1.2</td>
<td>21.4±10.9</td>
<td>19.4±5.2</td>
</tr>
<tr>
<td>C3</td>
<td>23.1±10.1</td>
<td>2.4±5.5</td>
<td>4.4 ±2.0</td>
</tr>
<tr>
<td>C4</td>
<td>21.4 ±0.1</td>
<td>2.0 ±3.5</td>
<td>3.2 ±3.6</td>
</tr>
<tr>
<td>C5</td>
<td>19.3 ±0.1</td>
<td>5.2±0.1</td>
<td>1.7±7.8</td>
</tr>
<tr>
<td>C6</td>
<td>20.7 ±0.2</td>
<td>7.0 ±6.2</td>
<td>5.3 ±1.2</td>
</tr>
<tr>
<td>C7</td>
<td>24.4 ±0.2</td>
<td>12.2 ±6.7</td>
<td>–</td>
</tr>
</tbody>
</table>

In Table 2, $Fb$ is the body inclination angle to the vertical; $Db$ is the body displacement of the overlying vertebra relative to the underlying one in the plane of the disc; $Hd$ is the disc height; $Gb$ is the angle between adjacent vertebrae.

8 Conclusion

The practical implementation of the proposed approach in the medical organization in goniometric spine control reveals new insights into the diagnosis and rehabilitation of the musculoskeletal system, particularly of the spine. The adaptability of the proposed approaches in early detection of motor function disorders, as well as its high accuracy, prevents the development of other diseases.

The proposed approach to automation of the spine diagnostics allows medical stuff to:
- continuously monitor the spine bending without attachment of patient to a stationary place;
- develop methods of evaluation of the degree of allowable deviation of the spine and vertebrae, accounting the age groups and possible bends of a healthy human spine, giving rise of the possibilities to automate the monitoring process that takes into account the patient’s anatomy;
- derive estimations of quantities characterizing the tolerances segments of the spine of healthy individuals for each type of bending of the back, increasing the efficiency of diagnosis and rehabilitation of the spine;
- assess the degree of friction during movement of the vertebrae under the influence of physical activity, and to identify the critical decrease (changes) of the intervertebral discs;
- derive estimates of the spectral components of an acoustic signal produced by the friction of the vertebrae, providing means of improvement of the health automated diagnostic systems based on acoustic methods of control;
- reduce significantly the risk of injury at the time of rehabilitation;
- optimize the rehabilitation process on the basis of the patient’s physical abilities and data control, allowing timely load adjustments, exclusion of the overloads and risks of injury.
Acknowledgments. This work was supported by RFBR grant 16-08-00992-a.

References

Application of E-Learning Tools in Different Ways of Implementation

Snezana Gavrilovic¹, Nebojsa Denic², Vukman Korac³, and Vera Petrovic³

¹Faculty of Information Technology, Alfa BK University, Palmira Toljatija 3, Belgrade, Serbia,
²Alfa BK University, Palmira Toljatija 3, Belgrade, Serbia
³School of Electrical and Computer Engineering of Applied Studies, Vojvode Stepe 283, Belgrade, Serbia
gavrilovicsnezana@yahoo.com

Abstract. In today’s stage of development information and communication technologies enable us to display a text, hear a sound material, see a static or dynamic visual material that can be pre-recorded or created on the computer. E-learning, in the form of independent use of the prepared material, often complements the teaching process being implemented in the classroom. The appropriate tools for e-learning are selected depending on the manner of implementation. The ways and possible tools for the implementation of e-learning are presented in the paper.

Keywords: E-learning, courseware, web, on-line, methods.

1 Introduction

In accordance with the development of information and communication technologies, education today is not limited only to that required, but what is more, the modern individual wants to develop and learn the possible and beyond. That is how the popular concept of lifelong learning was created. Learning theory is based on the notion that learning occurs when learners adopt new behaviours or demonstrate a change in behaviour as the result of an individuals response to stimuli.[1] In order to survive the abundance of information and knowledge, it is necessary to constantly improve and keep up with the world. One of the ways to satisfy the needs of the modern man is to develop e-learning which is growing and becoming increasingly popular in the world. E-learning certainly brings a great number of advantages in the educational process. This is not an alternative to the existing educational process, but it is also its integral part of the expansion and improvement. The combination of traditional paper-based learning material with digital one in a ubiquitous learning environment may offer great innovation in the delivery of education, to foster a student-centred approach, and to accommodate the needs of ubiquitous learners personal lifestyles. [2] The introduction of e-learning increases the role and importance of teachers, as mentors, coordinators and as well as participants in the educational process. E-learning enables that the student could be center of the educational process in which
he takes an active role and responsibility for its results. E-learning is a high education process in which all participants cooperate actively with the aim of achieving specific educational goals. During the process, modern information and communication technologies are used to create a flexible virtual environment.[3]

E-Learning falls into four categories, from the very basic to the very advanced.[4] The categories are:

- Knowledge databases: These databases are the most basic form of e-learning. These databases are seen on software sites offering indexed explanations and guidance for software questions, along with step-by-step instructions for performing specific tasks. These are usually moderately interactive.
- Online support: Online support comes in the form of forums, chat rooms, online bulletin boards, e-mail, or live instant-messaging support. This is slightly more interactive than knowledge databases.
- Asynchronous training: It includes access to instructors through online bulletin boards, online discussion groups and e-mail. Or, it may be totally self-contained with links to reference materials in place of a live instructor.
- Synchronous training: It is done in real-time with a live instructor facilitating the training. Everyone logs in at a set time and can communicate directly with the instructor and with each other.

2 Methods of implementation of e-learning

Web Based Training (WBT) is an application or set of applications contents of which are accessed using a Web Browser. Educational content of these applications is usually equipped and links to other educational resources. WBT systems offer much more than clear presentation of the learning content in a Web browser (eg, communication features, interactions, tests ...).[5]

In the, so called, teleconference classrooms equipped with cameras and equipment encoding and decoding video signals, lectures can be recorded and transmitted via the Internet to a remote location where there are participants.

Through the Internet we can publish and offer expertise in the educational market to a much wider circle of potential participants than those who geographically gravitate farther from the institution. It includes a publication of materials that are not available in print as well as other materials which are prepared for performance over the Internet.

Computer Based Training (CBT) is an application or set of applications that are educational facilities delivered via computer. This process includes lessons, exercises, simulation and testing. The CBT today means learning programs, which are not based on the Internet.[5]

The classrooms are equipped with one computer and projector, by which lectures can be run and which can visualize the process and facilitate the understanding of students.

In the computer classrooms, educational institutions can offer quality education to students requiring intensive computing (statistics, projecting, design). Testing can be performed using interactive tests with automatic evaluation.
Intelligent Tutorial System (ITS) like asynchronous e-learning systems represents a specialized software system intended for learning and teaching in the selected knowledge base. Unlike copyright or expert systems, thanks to elements of artificial intelligence, it has a possibility of detection and diagnosis of errors during the monitoring of students’ knowledge in the context of teaching strategies at their disposal.\[5\]

ITS has:

– regional knowledge (knowledge base created by the appropriate authority)
– the conclusion of the rules set out on the basis of domain knowledge
– assessment of students’ knowledge and
– teaching strategies

It is designed so that the teaching content in scope, structure and order can be adapted in accordance to the individual abilities of students. It is based on regional, methodical and pedagogical knowledge. It represents a kind of ”virtual teachers”. It analyzes the process of the current students’ knowledge and it has the ability to draw conclusions for instructing learning. Development of ITS began in the early seventies, it reached the popularity in the nineties, and recently it has appeared in the form of cognitive tutors.

The structure of the ITS consists of four related software modules:

– specialist module
– students module
– teachers module
– communication module (interface)

Specialized software environment enable professionals to build the base of domain knowledge, masters to construct the techniques for the course, and students to learn and assess. Video on Demand (VoD) and Video on Demand (AoD) are used for streaming multimedia technology and broadcasting audio and video data over the Internet in real time, for delivering educational content to desktop users upon request. There is no ”delayed broadcasts ”after a video is download from the server. It requires a quality link of the users on the network so that the required video footage could be aired at real time.

IP Television (IPTV) is an interactive television transmission using multicast technology for sending data from one user to a subset of users as a group concept of high quality transmitted video and audio on the user’s computer.

Virtual Classroom is a teaching environment that is not located in a brick building like traditional school, but in the computer-generated and communication supported systems. It represents the online seat of synchronous communication professors, instructors, mentors and students.\[6\]

Virtual Labs - free like a virtual classroom is a term that refers to the on-line position on which a project of software solution laboratories is realized with a view to students, instructors, professors, mentors in synchronously or asynchronously making preparations or realization of simulation of various laboratory tests to which the Virtual Lab is related.
Audio-video conferencing - as mentioned when we talked about WBT-in, in teleconference classrooms equipped with cameras and equipment for coding and decoding, lectures can be carried out and the AV signal is recorded and transmitted via the Internet to a remote location, thereby realizing a form of distance education. Interactive AV communication provides feedback to callers in real time, which both see and hear. Video technology increases productivity of the instructor and is particularly interesting to guest lectures and experts in certain areas, especially in the implementation of teamwork, which is dominant in the interactive distance learning.

Real-time on-line tutoring - we have already discussed this subject in the context of stories about AV conferencing, to which we can add what is a logical extension - computer online conference as a new form of educational interactions using modern, primarily telecommunications possibilities. To everything that has enabled AV conferencing infection, the ability to integrate with the content created via computer is added to the direct downloads. Lectures can be supplemented by different, for a specific topic prepared, software processed content (presentations, simulations, experiments and the like.). There are quantitative elements to consider, such as how many lessons, how much time per lesson, and how much material to cover. There are qualitative elements as well: What level is appropriate for the learner? What are the goals, objectives, and needs of the learner? [7] Tutoring fits in the topic of real time on line and what we have said for Virtual classroom (on-line seat synchronous communications) and chat.

Net-Learning, as well as learning through study of the content to be converted into digital formats to servers (or groups of servers) computer networks, allows the student, from school or any other location where there is a computer that is connected to one of the computer networks, to have access to these facilities at any time, as long as they want to meet their needs.

Searching knowledge base is a form of learning through the network. Searching information is carried out by entering keywords according to their interest entered by the user. ”Permanent assistance online” as opposed to the knowledge base, benefits through the possibility of obtaining replies on specific issues that are not provided by the standard knowledge bases. It is implemented through the use of the forum to network, ”interview room”, e-mailing a constantly active centers messages, blogs.

Learning on the Web is Net-Learning, or any Net-Learning does not have to be implemented Web technology. For example, learning by using a LAN of an educational institution that is organized so that the educational materials from the server and the network can be accessed to, for organized procedures by the aforementioned network administrator, may not be realized as WBT but it may be realized on the principle of file-transfer of materials mentioned on the local labor station (PC) and then it is still implemented. However, today, the Net-Learning increasingly includes Web Based Learning. The growing institutions, corporations, educational institutions, the banking group, state administration, etc., have organized Internet.
On-line Learning brings with it fundamental changes in educational psychology, methodology, didactics, communication ... whose achievements are necessary to implement a successful educational content. This is a much more complex task of digitization of standard textbooks and the use of computer communication with the mentor. On-line courses are educational technology of tomorrow.[8]

The contemporary university becomes increasingly permeated by digital mediation both on-line and face-to-face, relationships between digital media, time and socially-situated practices of meaning-making are foregrounded.[9]

One of the most accepted model for the development of on-line courses is a popular model ADDIE:

![The standard model ADDIE](image)

Fig. 1: The standard model ADDIE in the interpretation of the instructional designer Dhala Anglade

After much review, editing, proofreading and other control mechanisms, process making courses do not end with the beginning of classes. Most of them often provide the phase trial performance courses. In her project the team carefully supervises the course and impressions and reactions of the participants. After the last amendment and the full adaptation of the courses to real conditions, courses are fully completed.

Control of the online courses will not stop, and the connection of the topics with information and communication technologies directs the project team on a regular insight into compliance of courses produced with technological advances and changes according to it. It is a longer and more difficult mode compared to the traditional access and to the maintenance of educational materials. The difference in the quality of the courses in the obtained system training according to ADDIE model far exceeds the financial and time investment.

What we can point out as a problem when it comes to online courses is the percentage of the students who complete such courses. As this percentage varies, and in some cases falls to 25%, the question is why that happens with...
some on-line courses. Here are several factors which are important for how to make on-line course students be interested so they do not leave:

- design course
- informing the participants before the course
- skills on-line teaching
- technical support

Mobile Learning (M-Learning), as part of the accompanying e-learning, is one of the modern ways of training companies. The progress of information technology creates the conditions for the development of new forms of this type of education. Most young people know a lot about the use of mobile phones, which is the basis for a new line of e-learning. In addition to mobile phones, there is a growing number of people using laptops which represent an important factor of development of M-Learning.

Mobile devices such as tablets, mobile phone, Palm computing devices, portable computer (laptop, notebook) can be used to search for knowledge, regardless of where you are, even when you are in one of the vehicles, depending on the course of communication conditions available (usually wireless). If electronic learning can be defined as a method of education that performs the process of education beyond classroom, school or education camp, then M-Learning involves mobile way of education. M-Learning separates us from some fixed point. Where e-learning is involved, we believe that it is an alternative way of education, but on the other hand, M-Learning can be said to be a complementary activity to electronic education, as well as for traditional education. [10]

M-Learning is characterised by high mobility and low embeddedness: data storage and communication are easy, learning is enabled at anytime and anyplace through mobile phones, ipods, NintendoDS. Issues are that learning at anytime and anyplace is not enough and that personalisation according to the learners context is also important to provide the appropriate learning contents and to enhance learning in the real world. [2]

Today modern technology provides us with an opportunity to “pockets” bear many resources and have access to them whenever we need them. Let us mention a number of limitations that come with M-Learning: small screen size, limited processing power, reduced input capacity and the like. The foregoing means the need for adapting existing E-Learning services and content for M-Learning, which is not at all a trivial task.

3 E-learning tools

There are tools within which we can create, as an author or as a student, and use in the courses and as teaching materials, monitor the work, progress and interest of the students in topics that are object created courses. [11] Such tools are:

- courseware tools
- web discussion forums
Courseware-term obtained by merging concepts course (course-course) and software (program) is a computer software designed for education. Tools for electronic education can be offered on CD / DVD-ROM, website, as an instructional video or a program for learning. Tools for electronic education are often used for training people in the use of computer business applications, and often indicates the extra material in courses for computer use. Courseware tools are not a substitute for traditional books and textbooks, but should be seen as a new way of organizing teaching and learning, qualitative supplement (innovation) to existing (Traditional) ways, in order to successfully achieve the objectives such as: high quality education, lower cost of education, universal access to education.

Modern teaching only with a book or only with speakers is hard to imagine. It is hard to imagine distance learning courseware without the use of tools. They should not offer only a clearer presentation of the material, and the second mode of systematization of knowledge, but they should also serve as a place for communication. This does not mean that they will completely replace communication with “a human being”. These tools depend on the speakers. The lecturer who is not interesting in the classic form of teaching, most likely will not, without modification, be interesting even with the courseware tools. In addition to working on the materials, media, presentation mode, adapting to a new way of working, a lecturer using courseware can significantly improve the quality of teaching and the results of their work with students courses that he leads.

With the increasing popularity of Open Source and Freeware solutions and market courseware tools there are some free solutions. At this point the best representatives of the free courseware tools are Moodle, Claroline, A Tutor and Bazzar. These tools are not only free but they support the Open Source philosophy, so with them you get the source code. The advantage of such tools and Open Source philosophy is that that they will be noticed by people who want to adapt to their tools needs.

As a rule, large commercial tools justify their commerciality. Some of the most popular are: WebCT, Blackboard and Intralearn, a little less known FirstClass, e-Front, Lotus Learning Space, Lotus Sametime, and e-College e-Learner. With a good customer support, commercial tools often offer opportunities being free for now, but they are yet not able to offer: private space and settings, improved asynchronous and synchronous communication, internal e-mail, calendaring, selecting the interface layout, more features in checking the knowledge, support for audio and video, more information about the participation of students, support for content sharing.

Web discussion forums are sites or parts of a larger site where people interested in the subject of which the discussion is led may present their opinions and comments, ask questions or answer questions raised by other participants in the discussion. Forums are part of the Web and they are managed by a group that is responsible for the proper forum, so there is a reduced amount of advertising, enhanced security, programmable interface is the most common adjusted to pur-
poses (discussion among participants). Participant posted their assignments and provided peer feedback for each other regarding their design work in the discussion forums.[12] Forums and discussion groups and chat rooms can be classified as a "permanent online help" that gives ability to respond to specific questions and provides greater interactivity.

Chat is an electronic instrument which plays an important education role because it is one of the most common and most popular means of communication on the Web (especially among young people). In the chat room-in users "bring together" and discuss various topics, on-line, so that they are more engaged, which leads to a freer conversation, there are challenges, which are believed to be the best conditions of chat rooms with no more than five people. In such a situation, for example a very productive conversation can be led between the professor and two to three students.

Chat allows us: discussions about problems in real-time, game roles and simulations, exchanging opinions, debating and discussions in small groups, instructions and guidance in learning, group research, creation of an online community.

Blog, as a term, was created by shortening the two words Web Log (Web log). The phrase Web Log in Internet technology is used to define the statistical activities of one site (the number of visitors, the number of open pages site, etc.), but in the context of the phenomenon Blog, it means the guided log through the Web site.

Technically speaking, the blog is a type of automated Web site, where content (texts, images, multimedia) is entered and displayed chronologically, which is a significant difference compared to classic view news sites, where there is a particular emphasis on the most important news. This gives a chronological overview of all the articles that are published, whatever their importance.

Blogs can contain text, images, video, audio materials, musical and content, so-called micro-blogs.

4 Conclusion

The Internet has created great opportunities for informational and educational institutions to expand their database and the way of education, as well as the area in which they operate. Supporting the development and application of information and telecommunications technologies directly contributes to the development of economy and society and has the influence on the population.

The development of e-learning in the world today has reached a breakthrough. A great number of world-renowned institutions of higher education apply this category of learning in their curricula as a compulsory choice and modern way of education, as well as serious organized programs that are characterized by a large number of participants. The success of this type of education is guaranteed only when you align the way of implementation of e-learning with a choice of the proper tools. Combinations of different methods of implementation and tools for distance learning open a new chapter in the field of application.
of information technology in education. That is, as yet, under-explored and that will, in the near future, occupy an important place in the education system.

References

High Technology Software Web-Tools to Solve Environmental Problems of Coal Region

Alexander Goudov¹, Valery Perminov², Yuri Filatov³, Lee Hee Un³, Sergey Zavozkin¹, Irina Grigorieva¹, and Igor Sotnikov¹

¹ Kemerovo State University, 6, Krasnaya Street, Kemerovo, 650043, Russia
² National Research Tomsk Polytechnic University, 30, Lenin Avenue, Tomsk, 634050, Russia
³ NC VostNII JSC, 3, Institutskaya Street, Kemerovo, 650002, Russia

good@kemsu.ru, p_valer@mail.ru, main@nc-vostnii.ru, leeanatoly@mail.ru, shade@kemsu.ru, igriva@list.ru, mxtfonlife@mail.ru

http://www.kemsu.ru

Abstract. The paper is dedicated to the pilot system of the computer information portal, that is being designed in Kemerovo State University in order to enable engineers, students, postgraduate students and other users to get an expanded access to solutions of environmental applied problems in Kuzbass. The following elements are considered to be pilot system elements: solution of the grit motion in a flooded shaft problem, virtual laboratory of parallel programming, distributed computer resources access system. The present research is based on the task 2014/64 of state research Scientific research organization.

Keywords: computer information portal, mathematical modeling, high performance computing.

1 Introduction

Coal companies are known for their huge negative effect on Kuzbass environment and the adverse changes they cause. High cost of environmental facilities, lack of investment, unavailability of scientifically proven recommendations on how to reduce mine work adverse impact on environment and to eliminate that impact make the environmental situation even worse.

Coal companies development requires water consumption increase needed for coal production and further coal preparation. Coal companies are major environment contaminators with increasing waste water discharge. Various industrial methods of waste water treatment require great costs.

Waste water treatment by using flooded waste mining workings requires less costs compared to other methods. Kolchuginskaya coal mine was the first in the world to apply this method to purify Komsomolec (coal preparation plant) slurry water. Liquid wastes are pumped into worked-out area of a flooded mine. Natural purification of wastewater is supposed to take place in mine workings due to the water precipitation and mixture with influent underground water. Although the
method requires low costs it is essential to be researched to forecast possible effects of treatment processes in the flooded mine. Flooded mine working can be defined to be a black box with only input and output data possible to be estimated. Numerical simulation of the treatment process is almost the only possible way to estimate this method impact on the environment.

Taking into consideration current ecological situation the Russian government approved the state long term coal industry development program for the period until 2030, that includes measures aimed at coal consumption increasing by domestic power industry and scientific and technology capability development via the adoption of innovation technologies of coal processing and utilization such as underground coal gasification. This technology is considered to be the most ecologically friendly. Numerical simulation of underground coal gasification enables to identify qualitative composition of combustion gas in the coal mine under consideration.

Kemerovo State University develops the project targeted at high technology software complex development in order to solve environmental problems of the coal region that can be available for research community, managers, engineers, students and post graduates.

Software tools are often used by either developers or limited group of experts in the form of problem-oriented software. It happens because software is mostly aimed at single-discipline problem solving, sophisticated to use, and it requires upgrading in case the problem formulation changes. Such kind of software is considered to be unique and requires pricy operating license ($2500 for a processor or $900 - $1500 for a user). Modern information technologies enable to cut simulation experiment costs and make it available for more users by creating cloud computing and special purpose web-services.

The following tasks are solved:

- mathematical modeling to solve applied ecological problems,
- carrying out simulation experiments based on the multiple access computing center of high performance computing,
- development of applied software based on web-services (information-computer portal),
- creating virtual laboratory courses for educational process based on information-computer portal,
- leasing the developed software out.

Here are the main project streams:

- development of mathematical and software components to solve the problems of waste water treatment in flooded mines and underground coal gasification,
- development of software technology platform with set of services to perform functions of information-computer portal,
- KemSU computing resource (multiple access computing center of high performance computing) and other computing resources access arrangement, which are provided for users and based on external platforms or cloud.
2 The problem of waste water treatment in a flooded mine

Water body pollution by mining and quarry waters is a typical problem for Kuzbass and many other mining regions [3]. Mining waters usually contain particles of coal dust, clay, calcium compounds, magnesium, oil products, etc. Light substances (which density is less than water density) such as oil products accumulate on water surface while other particles remain suspended or sediment gradually. The problem of mining water treatment by pumping into abandoned mines and further use of it after precipitation of impurities (for heavy particles) or impurity floating up (for light particles) is of great interest.

The paper considers fluid flow containing impurity particles in a flooded mine. To analyze float impurities distribution a square form mine is under consideration. It has a ledge at the top (shown in Fig. 1).

Underground water inflows into the domain thorough the boundaries KD, CI and GN. Fluid leaves the domain though the boundary AB. Impurity layer stays inside the domain at the initial time. Influenced by a flow some impurities leave the domain while the remaining part stays in the domain. To describe this transfer process differential equation system is used. The equations express the laws of conservation of mass, momentum and elements concentration in the domain. Mathematically the following differential equation system for turbulent flow should be solved in the following way:

\[
\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x_j}(\rho u_j) = 0
\]  

(1)
\[
\frac{\partial}{\partial t}(\rho u_i) + \frac{\partial}{\partial x_j}(\rho u_i u_j) = -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j}(-\rho u'_i u'_j) - \rho Sc u_i |\overrightarrow{u}| - \rho g_i, \quad (2)
\]

\[
\rho \left( \frac{\partial Y_k}{\partial t} + u_1 \frac{\partial Y_k}{\partial x_1} + (u_3 - u_{3k}) \frac{\partial Y_k}{\partial x_3} \right) = \frac{\partial}{\partial x_j}(-\rho Y_k u'_j),
\]

\[
\frac{\partial}{\partial t}(\rho k) + \frac{\partial}{\partial x_i}(u_i \rho k) = \frac{\partial}{\partial x_i} \left[ \left( \frac{\mu_t}{\sigma_k} + \mu \right) \frac{\partial k}{\partial x_i} \right] - \mu_t \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \frac{\partial u_i}{\partial x_j}
- \beta g_i \frac{\mu_t}{Pr} \frac{\partial T}{\partial x_i} - \rho \varepsilon, \quad (4)
\]

\[
\frac{\partial}{\partial t}(\rho \varepsilon) + \frac{\partial}{\partial x_i}(u_i \rho \varepsilon) = \frac{\partial}{\partial x_i} \left[ \left( \frac{\mu_t}{\sigma_k} + \mu \right) \frac{\partial \varepsilon}{\partial x_i} \right] + C_1 \frac{\varepsilon}{k} (G_k + G_B) - C_2 \rho \frac{\varepsilon^2}{k}, \quad (5)
\]

\[
p = \rho R_0 T \sum_k \frac{Y_k}{M_k}, \quad \overrightarrow{g} = (0, g), \quad u_{3k} = \frac{gd^2}{18\nu} \left( \frac{\rho_k}{\rho} - 1 \right). \quad (6)
\]

Here, \(t, x_i\) - time and spatial coordinates \((i = 1, 3)\); \(u_i\) - velocity vector projection on the corresponding axis of cartesian reference system, \(p\) - pressure; \(g\) - gravitational acceleration, \(R_0\) - absolute gas constant, \(M_k\) - molecular weight of \(k\) - component, \(\rho\) - density of fluid and particles mixture, \(\nu\) - kinematic viscosity coefficient, \(D_t\) - diffusion coefficient, \(d_k, \rho_k, u_{3k}\) - diameter, density and velocity of particle settling, \(Y_k\) - mass concentrations of \(k\) - component \((k = 1\) - water, \(2\) - solid particles); \(\mu_t = \rho C_\mu k^2 / \varepsilon\) - coefficient of turbulent viscosity, \(k = \overline{u'_i u'_j}/2\) - turbulent kinetic energy; \(\varepsilon\) - its dissipation, \(C, \sigma_k, \sigma_\varepsilon, C_1, C_2\) - empirical constants, and \(G_k, G_B\) - turbulence caused by forced convection and natural convection.

Based on mathematical formulation of the problems (1)-(6) numerical calculations were made to determine the pattern of float impurity distribution process in a flooded mine [4].

Vector fields of velocity and impurity distribution at different time moments were obtained as the result of numerical integration of equation system (1)-(5). Side walls are considered not to influence the impurity distribution process and fluid flow. Thus the problem is solved in the two-dimensional domain \(X_1OX_3\). A mine (length \(-10\) meters horizontally, height \(-3\) meters) is under consideration (Fig. 1.). Underground water without any impurity enters the domain. Impurity concentration equals 1 inside the domain. Particles size is \(d_k = 5 \cdot 10^{-5}\) m. Impurity particles density is \(500 kgs/m^3\). The velocity of groundwater inflow from the upper layers is \(0.1m/s\). The distributions of impurity are numerically calculated at different time moments (Fig. 2-3). These figures show that the flow becomes stable and impurities accumulate in the upper part of the domain as time goes. It happens faster compared with the previous case because the particles density is two times less.
The mathematical model presented in this paper can be used to analyze mining water treatment process due to environment and evaluate its further possible improvements.

3 Problem of ignition and combustion of combustible gas and coal particles gas-dispersion mixture

Flame front distribution in gas-dispersion medium is under consideration, when exothermic chemical reactions take place in the gas phase and on the surface of disperse phase particles with one of gas phase components. Those processes accompany combustion process of methane-air mixture with coal particles in mixture of gases (oxidant, combustible gas and inert gas), where small coal particles that can be heterogeneously reactive with gas mixture oxygen are evenly distributed. The oxidant is supposed to be involved into the particle surface reaction. Gas-dispersion mixture has specified velocity. The particles have equal sizes and spherical shape. Heat exchange between particles and gas follows the Newtons law. The rate of gas chemical reactions and particle surface chemical reactions depends on the temperature according to the Arrhenius law. Gases are the resultants of heterogeneous reaction on particles. Thermal expansion of gas mixture can be ignored. The ignition source is situated at the boundary of $x = 0$ (combustion temperature is specified). Mathematical model of this mixture combustion takes into consideration the complexity of gas phase and
two-temperature medium [8, 9]. Taking into account the conditions above, the
equation system appears to be as follows:

\[ \frac{\partial}{\partial t} (\rho + mN) = 0 \]  

(7)

\[ \rho \left( \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} \right) = -\frac{\partial p}{\partial x} + \frac{\partial}{\partial x} \left( \mu \frac{\partial u}{\partial x} \right), \]  

(8)

\[ \rho c_p \left( \frac{\partial T}{\partial t} + u \frac{\partial T}{\partial x} \right) = \frac{\partial}{\partial x} \left( \lambda \frac{\partial T}{\partial x} \right) + q \rho^2 c_1 c_2 k_0 \exp(-E/RT) - 
-S \alpha N (T - T_S) - (c_p T - c_ST_S) N \frac{\partial m}{\partial t}, \]  

(9)

\[ m c_s \frac{\partial T_s}{\partial t} = S \alpha (T - T_S) - q_s \frac{\partial m}{\partial t}, \]  

(10)

\[ \rho \left( \frac{\partial c_1}{\partial t} + u \frac{\partial c_1}{\partial x} \right) = \frac{\partial}{\partial x} \left( \rho D \frac{\partial c_1}{\partial x} \right) - \rho^2 c_1 c_2 k_0 \exp(-E/RT), \]  

(11)

\[ \rho \left( \frac{\partial c_2}{\partial t} + u \frac{\partial c_2}{\partial x} \right) = \frac{\partial}{\partial x} \left( \rho D \frac{\partial c_2}{\partial x} \right) - \alpha_S \rho^2 c_1 c_2 k_0 \exp(-E/RT) + N \frac{\partial m}{\partial t}, \]  

(12)

\[ \frac{\partial m}{\partial t} = -\frac{S p c_2 R S \beta_m}{R_S + \beta_m}, \quad R_S = k_S \exp(-E_S/RT_S), \quad \beta_m = \frac{N u_D D}{d} \]  

(13)

where \( t \) - time, \( x \) - coordinate, \( u \) - velocity, \( T \) - temperature, \( \rho \) - density, 
\( p \) - pressure, \( q \) - heat of gas chemical reaction, \( q_S \) - heat of particle surface 
chemical reaction, \( E, E_S, k, k_S \) - activation energies and pre-exponential factors 
of gas and particle surface chemical reactions, \( d, S \) - diameter and particle surface 
area; \( N \) - number of particles per unit volume; \( \lambda, \alpha, \beta_m, D \) - coefficients of heat 
conduction, heat mass exchange and diffusion; \( R \) - absolute gas constant, \( c_1, c_2 \) 
concentrations of combustion gas and oxidant, \( Nu_D \) - Nusselt diffusion number, 
\( \alpha_S \) - stoichiometric coefficient, \( l \) - size of computational domain.

\[ t = 0 : \rho = \rho_0, T = T_0, c_1 = c_{10}, c_2 = c_{20}, m = m_0, \]  

(14)

\[ x = 0 : u = u_0, \quad \frac{\partial T}{\partial x} = 0, \quad \frac{\partial c_1}{\partial x} = 0, \quad \frac{\partial c_2}{\partial x} = 0, \]  

(15)

\[ x = l : \frac{\partial u}{\partial x} = 0, \quad \frac{\partial T}{\partial x} = 0, \quad \frac{\partial c_1}{\partial x} = 0, \quad \frac{\partial c_2}{\partial x} = 0. \]  

(16)

The indexes refer to: 1 - combustible gas, 2 - oxidizing agent (oxygen), s 
disperse phase, 0 - initial conditions. Basic data are found in [8–10].
Equation system (7)-(13) with initial and boundary conditions (14)-(16) was numerically solved. Control volume approach [4] was used to achieve discrete analog. Distributions of temperatures and components in the domain under consideration are identified by using numerical integration. The mixture of three gases is studied: overoxidized \( c_{10} = 0.0349, c_{20} = 0.15504 \), stoichiometric mixture \( c_{10} = 0.0402, c_{20} = 0.15504 \) and underoxidized \( c_{10} = 0.0405, c_{20} = 0.15504 \).

**Fig. 4.** Distribution of temperature (graph a), combustible gas concentration (graph b, curve 2), oxygen (graph b curve 1) and combustible gas (graph b line 1); \( t=15 \text{ s} \)

**Fig. 5.** Distribution of temperature (graph ), combustible gas concentration (graph b, curve 2), oxygen (graph b curve 1) and combustible gas (graph b line 1); \( t=18 \text{ s} \)
For example, on Fig. 4-5 show gas phase temperature distribution and distribution of combustibles and oxidizing agent due to different time moments. It is obvious that flame front develops because maximum temperature area moves. According to that, the concentration of the combustible $C_1$ is reduced and the concentration of the oxidant $C_2$ decreases to almost zero point.

4 Information-computer portal

Software (in the form of web-services integrated into engineer and computing portal) based on the problems mentioned in sections 2-3 was developed.

Fig. 6 shows the architecture of portal prototype as a deployment diagram.

![Diagram](image.jpg)

**Fig. 6.** The architecture of portal prototype as a deployment diagram

OpenLDAP Server is a server of LDAP-catalogue used to create single user base, web-service registry, business processes, high-end computing resources and some other data.
Client system is any external client system that interacts with web-services / business processes of the portal and/or with LDAP-catalogue.

User Client Workstation is a user personal computer, for example, a personal computer with web-browser.

Liferay Portal (https://www.liferay.com) is a configurable complex solution to develop web-portals. Portlet technology is used to create portal pages. Portlet is a web-application designed in accordance with JSR-168 or JSR-286 specification [5]. They generate portions of some content (usually fragments of HTML- or XML layout) embedded into a web-page. Such kind of a web page can include many portlets. Liferay enables each user to create personal pages. That opportunity is used to create personal user work space that aggregates services to meet his personal needs.

Apache Axis2 is an integration and web-service life-cycle management system (http://axis.apache.org/axis2/java/core/).

Apache ODE is an integration system of web-service orchestration (http://ode.apache.org) that is a mixture of web-service capabilities to compose new higher level web-service called business process. BPEL based on XML is a standard descriptive language for business processes.

Nginx is a high-end HTTP-server and reverse proxy, e-mail proxy server as well as general purpose TCP/UDP proxy server (http://nginx.org/ru/). In this context it is used as a proxy server to forward requests to other components (Apache Tomcat 7, Tornado PHP-FPM). Though, there are some exceptions (requests for static files): images, JavaScript-files, etc. the reason is that Nginx is optimized for such kind of requests.

Tornado is a web-server that hosts online development environment system Onlide [7].

Fig. 7. Web-service model providing access to the HPC resources (components and internal structure diagram)
PHP-FPM is a process manager FastCGI used to generate PHP dynamic content of the Virtual laboratory course system [6]. It is used due to the fact that Nginx has no native support for such content generation.

Apache Tomcat 7 is an application web-server with servlet specification support. It hosts Liferay, Axis2 and ODE.

Web-service HPCWebService is created to interact with high-end computing resource (HPC resources). Fig. 7 shows the service model.

Web service has the functions targeted at compute cluster-based sequential and parallel program compiling and running, program outputs, task monitoring, clearing up space for a user.

Depending on operating system and task management system (if there is any) the interaction with computing resources may vary. Therefore web-service interacts with HPC-resources via proxy agents.

Agents can be targeted at each individual resource in order to create multi-purpose agents interacting with many computing resources. For example, to interact with Linux-system and task management system Torque PBS clusters TorquePBSAgent is created. This agent is currently used to interact with main KemSU cluster (master.kemsu.ru).

Algorithms to solve applied problems, to access high performance resources and etc. are developed as web services. User interface used to interact with web services of the portal (solvers) is created either as portlets or with the help of purpose built format based on XML- SolverXML. The format is developed in the way that enables to use component kit to create solvers. This kind of approach enables to create reusable component-solvers which can be used to create fully-featured tools. It also makes it easy for a solver to meet users needs compared with its portlet-based implementation.

![Project Tree](image)

**Fig. 8.** An example of the solver interface to launch tasks on cluster, which is described with the help of SolverXML-format

SolverXML-description consists of not more than seven blocks: `<imports>`, `<js>`, `<gui>`, `<vars>`, `<externals>`, `<handlers>`, `<actions>`. 
The `<imports>` block lists imported JavaScript-files, user interface elements for data input/output (widgets) and SolverXML-descriptions of other solvers stored in the LDAP-catalogue.

The `<js>` block is used to add some JavaScript-based algorithms.

The `<gui>` block has user interface description.

The `<vars>` block has variable list which store or provide some value.

To make components interact with each other the `<externals>` block is used. It lists widgets and variables to address widgets and variables of other component.

The `<handlers>` block is used to specify handlers of widget and solver events.

The `<actions>` block describes the interaction with web services/business processes.

---

![Fig. 9. Architecture of the Onlide system as a deployment diagram](image-url)
Fig. 8 shows an example of solver interface based on the format under consideration.

The Onlide system is developed for remote development and launching of sequential and parallel programs at high-performance computing resources. It has the following functions: 1) development of software projects, consisting of many software texts based on different programming languages; 2) project editing; 3) project compilation and launching; 4) usage of extensions to increase capability options of development environment.

Fig. 9 shows the system architecture as a deployment diagram. LDAPMediator and HPCMediator (proxy agents) are developed in order to interact with LDAP-catalogue and HPCWebService. Onlide module is responsible for HTTP- and Ajax-requests processing and HTML-layout generation.

5 Conclusion

Multi-parameter model of incompressible fluid hydrodynamics is developed as a result of the project. Mathematical model of combustion of gas disperse phase with particles is presented. Numerical study of disperse phase and combustible gas and oxygen impact on the flame front rate in gas-disperse medium is carried out. According to calculation data the flame front rate depends on the gas and disperse phase parameters.

The algorithms are created and tested. Principles of information and computer portal based on service-oriented architecture are presented. The developed prototype of web-oriented technology software complex includes the following modules: software component of impurity motion in a flooded mine problem computation; virtual laboratory of parallel computing; component kit for interaction with distributed computing resources.

Practical relevance of the research enables to use prototype of technology software complex in order to conduct simulation experiments and teach high performance computing technologies to students and postgraduates.

The developed high technology product is supposed to attract additional investments in order to study new regional ecological problems.

The research is based on the state task № 2014/64, the state project Scientific researches organization. The results of the numerical calculations and problem formulation will be used by the educational resources information portal that offers students, postgraduate student and academic researches different educational services.

References


The Method of Selection of the Key Geodynamic Objects

Anastasia Grecheneva, Vladimir Eremenko, Oleg Kuzichkin, and Nicolay Dorofeev

Belgorod National Research University, Belgorod, 308015, 85 Pobedy st., Russia
1155464@bsu.edu.ru
http://www.bsu.edu.ru/bsu

Abstract. In this paper as an indicator manifestations of geodynamic processes in a large area are invited to select the most sensitive and informative geological structures to the appearance of endogenous and exogenous factors that contribute to the development of geodynamic processes and negative changes in the geological section. Such places are key geodynamic objects that can provide early warning of the beginning of the development of destructive geological processes that have no external signs of existence. Watching the local geodynamic key objects and with the involvement of the hydrology data, geology, meteorology and geo-information technologies, it is possible to form a forward-looking assessment of destructive geological processes over a large area. The paper proposes a method for detecting the key geodynamic objects, including the distributed processing algorithms informative sections of heterogeneous data, the temperature and the hydrological correction of the measurement results. The proposed approach is based not only on statistical methods and morphological analysis of the territory, but also on the use of mathematical models of the interaction of hydrological, geological and man-made environments.

Keywords: geoelectrical monitoring of geodynamic object, forecasting, localization of objects, key objects.

1 Introduction

It is known that the development of suffusion processes intensity of geodynamic changes of local sites of geological environment characterized by much greater performance than that of the total of its variations. Consequently, information about the occurrence of destructive processes through the use of selective geodynamic control can be obtained much earlier than in the monitoring geodynamic environment in general. Therefore, the practical use of geomonitoring systems built on the basis of geoelectric sounding methods is appropriate for monitoring the bearing capacity of overlying and underlying soil during the operation of industrial facilities, as well as to ensure the protection of natural and man-made objects from the possible consequences of accidents at suffusion danger [1,
2. Such systems through the application of information processing algorithms for heterogeneous monitoring data allow to register changes in the geodynamic control objects and obtain forecasts of the possibility of man-made disasters [3].

In this article, the example of suffusion processes the technique of constructing a regression geoelectric monitoring data processing algorithms with key geological objects in order to create predictive assessments of geodynamic.

2 The geological features of the site and the selection of the geodynamic control zones

Geodynamic monitoring carried out at the site of the alleged construction of Nizhny Novgorod NPP, which is located in the basin of the lower reaches of the river Oka (Figure 1a). The presence of low-mineralized water in the alluvial layer, lying in the valley r. Oka close to the surface, as well as the dominant stratum of carbonate and sulfate rocks, is the cause of the dynamics of the karst valley. Herewith, man-made increase in groundwater levels is the cause of the rapid process of karst formation and increase the risk of catastrophic situations at nuclear power plants.

The organization of the geodynamic control should take into account that there are two main types of geodynamic movements karst environment. This cyclic variation with varying intensity and duration of the period, characterized by cyclical changes in the structure of the medium, as well as the trend of variation which are of a pronounced character and having a constant direction for a long time, with the result that they are the main source of mechanisms of technological disasters [4]. Therefore, based on geological data it was determined optimum geoelectric zone control which will be geodynamic more pronounced than in other areas, for the same man-caused load. Monitoring of the local area will provide more accurate forecasts of geodynamic activity surrounding area (Figure 1b).

In addition geoelectric monitoring data by supplemented of stationary observations, including the monitoring of hydrogeological regime fracture-karst aquifer and overlying and geodetic monitoring of surface subsidence, changes in morphometric characteristics of the relief, the failures and deformations.

3 Key geological objects

For geodynamic control used multipolar equipotential electrical installation, developed together with IPE RAS. It is designed to monitor of the geodynamics of surface irregularities in the cases of the need provided increased sensitivity to the specific changes in the object of investigation. High efficiency is achieved by increasing the sensitivity of the measuring system, and the initial installation and operational positioning of the installation by controlling the sources of probing signals [5]. Operation is based on the fact that the source of the probing signals in the test environment is created in accordance with the principle of
superposition of a spatially-distributed signal forming a total zero signals in the measurement sensors of geoelectric field.

In this case the control signals of initial setting and positioning of geoelectric measuring systems, be formed in accordance:

\[ \bar{U}_i(t_0) = F_U(M_{Si}, \bar{U}^*(t_0)), \]  

where \( F_U \) the option forming of primary positioning on the control vector, by system \( \bar{U}^*(t) \) of space-time processing data control at start time \( t = t_0, M_{Si} \) a vector of model parameters.

Later the geoelectric measuring system is functions, directly, in the semi-automatic mode using the following algorithm:

\[ \bar{U}_i(t) = \bar{U}_i(t_0) + \Delta U(M_{Si}, \Delta \tilde{a}_i) + F_U(\Delta M_{Si}, \bar{U}^*(t)), \]  

where \( \Delta U(M_{Si}, \Delta \tilde{a}_i) \) the ongoing management of the positioning of the electrical installation of the vector of geodynamic variations \( \Delta \tilde{a}_i; \Delta M_{Si} \) the correction model.

Increase of sensitivity leads to an increase in noise level caused by thermal and tidal deformation effects. In addition, operational management of electrolocation signals is the presence of the trend component in the recorded signals, which is determined by the structural changes of the object [6].

Geoelectrical control method is based on the principle of linear and stationary of the geoelectric section, the transfer function \( \Delta H_{ij}(p, \alpha_1, ..., \alpha_l) \) is determined by a system of spatial functions of control object \( \psi_{ij}(p) \) with nominal geodynamic

**Fig. 1.** Geological features of the site and the selection of zones and geodynamic control
a) area of alleged placement of Nizhny Novgorod NPP; b) the geoelectric monitoring zone
parameters $\alpha_0$:

$$\Delta U_i = \sum_{j=1}^{n} \Delta H_{ij}(p) I_j(p),$$

(3)

$$\Delta H_{ij}(p, \alpha_1, ..., \alpha_l) = \frac{K(p)}{S_i(p)} \sum_{k=1}^{l} \left[ \frac{\partial \psi_{ij}(p, \alpha_1^0, ..., \alpha_l^0)}{\partial \alpha_k} \Delta \alpha_k \right],$$

(4)

where $I_i$ probe signal of $i$-th source; $\Delta U_i$ the response of $i$-th source; $K(p)$ Contrast Ratio of environs; $S_i(p)$ the dependence of the measurement channel gain.

These relations (1-4) makes it possible to solve the inverse problem - selection of properties of the local geodynamic object by adjusting the parameters of sensing sources, which is a key aspect of the organization of geodynamic control [7].

Monitoring of the key geological objects - places with an active geodynamics and the most sensitive to endogenous and exogenous factors, and further predicting of geodynamics on the entire territory requires a change in the structure of the geodynamic system of forecasting described in [7]. The main changes relate to the prediction block, its structure shown in Figure 2.

![Fig. 2. Structural blocks of the prediction unit](image-url)

Obviously, the key geodynamic objects for example, the suffusion processes, be chosen from the condition of the probability of the process itself: the presence of soluble species, and the solvent approach, removal of soluble species. Identify key geodynamic objects possible in rose histogram (Figure 2), which characterize the direction of the formation and propagation of failures, faults, the dominant
structure of the network of cracks, etc. The diagram shows not only the direction of education failures, but also their concentration on the area.

![Fig. 3. Rose histogram](image)

Expression (4) defines the principle of superposition of the probing signals by which to judge the possibility of providing separate characteristics of the environment (the object) by controlling the parameters of the source. This is one of the most important aspects of the organization of monitoring of geodynamic objects [10].

Based on the provisions described in this article, it is proposed to carry out the processing of heterogeneous data through specialized algorithms. Block diagram of the Information Technology Services of Geodynamic control system reflects the principle of joint processing of hydro-geological data (Figure 4).

The physical layer describes the physical methods of obtaining information that may be required to detect errors $\varepsilon_i$ and measurement errors in data analysis. On the same level a scheme of placing primary transducers (sensors, measuring tools and devices) are described. The main objective of this level is acquisition (measurement) of raw data $D_i$. This level is the hardware and hardware-software (in the case of digital sensors). Such devices as sensing devices, sensors, blocks of a positioning in space that define the coordinates $X_i, Y_i, Z_i$ measuring devices function in this level.

The link layer is represented by all kinds of measuring complexes, systems and instrumentation, and is a hardware-software. A modules and services related to prior and primary data processing, presentation and storage of primary $D_i$ and processed $D'_i$ data, supporting information: methods of measurement and processing, a model of locative level, required $X'_i, Y'_i, Z'_i$ and fixed $X_i, Y_i, Z_i$ positions of the primary converters in space are working in this level.

The link layer describes the working of the geographic information-analytical systems regulation and control (GIASC) of natural-technical systems (NTS) at the locative level, so there are also function modules and forecasting services, and the development of administrative decisions at the locative level. A control solutions are formed on the basis of received predictive estimates of $f$ and functioning models of natural, technical, natural-technical and social systems.
A errors forecasting and regulation at the NTS of the locative level are transferred network layer and serve as the basis for the correction components of operating at the link layer.

Fig. 4. Block diagram of information and technical support geodynamic monitoring system. \(\varepsilon_i\) the errors; \(\varepsilon'_i\) the compensation factor; \(D'_i\) the raw data; \(D_i\) the processed data; \(X_i, Y_i, Z_i\) the recorded position in space of the primary converters; \(X'_i, Y'_i, Z'_i\) the desired position in space of the primary converters; \(C'_i\) the synchronization signals and control.

One of the key factors determining the performance indicators of hemodynamic assessment at the geoelectric monitoring are used the earth models and models of geodynamic objects themselves. For a qualitative prediction of suffusion processes necessary to carry out an assessment of the expected location of displays and take into account their size, it is also necessary to take account of spatial-temporal geodynamic parameters. Therefore, the forecast is built on the basis of geomechanical models of different orders that can take into account the mechanism of interaction with the technosphere suffusion processes and geological conditions of its development.

For reasons of forecasting by the geoelectric monitoring necessary step is to establish the conformity of spatial functions in equation (4) for the transfer function of the geoelectric section geomechanical conditions of formation of local failures as described in [11,12].

This ratio can be set by considering the problem of the distribution of the geoelectric field of a point source field in the presence of a spherical in homogeneity, in which you can take as suffusion processes. The solution described in [13, 14, 15] to determine the characteristics of the occurrence of the ball on the observed distortions introduced them to the spatial distribution of potential geoelectric field.

The transfer function of the geoelectric section, which defines the displacement of equipotent lines i-source in space, taking into account the double anomalous component of the field is of the form:

\[
\Delta H_{ij}(p, a, h) = K(p)\psi_{ij}(a, h) = 2K(p)\frac{a^3 r_{ij}}{(r_{ij}^2 + h^2)^{3/2}}, \tag{5}
\]
where $h = z + a$ the depth of the sphere below the surface, $a$ the radius of the sphere, $r_{ij}$ the distance between the electrodes $i$ and $j$.

Depth assessment of the changes occurrence spherical near-surface heterogeneity and its size can be made on the basis of (5), using it to forecast future geodynamic suffusion processes as an assessment:

$$R_z = \sqrt[3]{\frac{3\sqrt{3}\Delta\hat{H}h^2}{2K(p)}}, \quad (6)$$

where $\Delta\hat{H}$ the maximum estimated offset value of the equipotent line. The results of the regression of processing time series are geodynamic background information for predictive modeling underlying the decision of geodynamic processes forecasting problem [16, 17].

Geoelectric model of geodynamics suffusion processes can be represented by a discrete linear system [11] defined by the difference equation:

$$Y_k[i] + \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij} Y_k[i - j] = S_k[i], \quad (7)$$

where $Y_k[i]$ the counts recorded geodynamic process on $k$-th registration point; $a_{ij}$ the model coefficients; $S_k[i]$ samples generated by a random process with geodynamic parameters $M\{S_k[i]\} = 0, M\{S_k[i]S_k[j]\} = \sigma_k^2 \delta_{ij}$ ($\delta_{ij}$ the weights of the model).

System regression of the original equations is formed on the basis of the expression (7):

$$Y^T[i] = F_a[i] + a^T[i] + s^T[i], \quad (8)$$

where $Y^T[i] = [Y[m+1],...,Y[l]]^T$;


$$a^T[i] = [a[1],...,a[m]]^T; \quad s^T[i] = [s[m+1],...,s[m+l]]^T; \quad l \text{ the depth of predictive estimates.}$$

Application in the analysis of suffusion processes predictive estimate of a regression model (8) allows you to make predictions that take into account not only the impact of cyclical planetary factors, but also man-made impacts. Figure 5 shows the preliminary interpretation of geological and geoelectric section in the area of geodynamic control.

On the basis of regime observations were interpreted registered signals geodynamic variations. Figure 6 shows the variations registered geodynamic gain bipolar equiopotential geoelectric installation during the annual observations from May 2013 to April 2014.
Fig. 5. The preliminary interpretation of geological and geoelectric section in the area of geodynamic control.
Conclusions

The data are in good agreement with the hydrological observations of the water level in the river Oka and calculated as the ratio of mineralized areas at the top and bottom of the river.

Based on these algorithms in this article was obtained prognosis estimation of dip by models karst suffusion processes. As a result, it was found that the use of these algorithms, the formation of forward-looking assessments in geoelectric monitoring promotes the release of a high degree of reliability and the conditions of dip karsting the development of suffusion processes. Increasing the depth of predictive assessments and improving the efficiency of the proposed method is achieved by increasing the number of sensing points of the geoelectric field and the number of sounding sources.

Acknowledgments. This work was supported by grants of the President of the Russian Federation № MK-7406.2015.8.

References


Decision Support System for Real-Time Diagnosis of Musculoskeletal System

Anastasia Grecheneva, Igor Konstantinov, Oleg Kuzichkin, and Nokolay Dorofeev

Belgorod National Research University, Belgorod, 308015, 85 Pobedy st., Russia
1155464@bsu.edu.ru
http://www.bsu.edu.ru/bsu

Abstract. A construction principle of a technical system for diagnosis and rehabilitation of the musculoskeletal system based on accelerometer method, together with synchronization algorithms measuring patient parameters, is considered. The optimal accuracy estimations of the technical parameters of the accelerometric goniometer system are determined; they are the sample rates of the accelerometer signal converters, the required sensitivity of the sensor, etc. The advantages of the proposed approaches to the construction of rehabilitation and diagnostic systems of the musculoskeletal system are adaptability and reliability of the diagnoses.

Keywords: biomechanics, information system, goniometric control, accelerometer, mathematical model.

1 Introduction

Accurate diagnosis and objective assessment of the treatment efficiency of motor function disorders to date remains one of the urgent problems of modern traumatology and orthopedics. The large number of evaluation approaches and techniques reveal a lack of reliability of the proposed criteria for diagnosis and assessing recovery efficiency. For example, the diversity of human movement is characterized by a number of parameters: torque, speed, complexity of trajectories, changes in the level neuromuscular and brain activity. In existing systems, goniometry and diagnosis of musculoskeletal system mainly take into account only the kinematic parameters of the skeletal system, regardless of the bone structure and neurophysiological parameters of state of the patient [1]. Therefore, in diagnosis, the study of the central control mechanisms of purposeful physical activity is of great importance, as well as the parameters of the skeletal system at the structural level [2,3].

2 Statistical basis of the goniometric measurements

Formation of the goniometric criteria and selection of the optimal working parameters of the system rehabilitation is carried out on the basis of statistical
clinical studies of patients under normal conditions and in the presence of deviations. In medical diagnostics, the assessment of the angular movement indicators and their permissible deviation from the normal ones is done according to the joints motion estimation table “Regulations on military-medical examinations” (approved by the Russian Federation Government Decree No 565 dated July 4, 2013) [4].

**Table 1.** Assessment of range of motion in the joints of the limbs

<table>
<thead>
<tr>
<th>Joint</th>
<th>Motion</th>
<th>Norm,°</th>
<th>Restriction of movement,°</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>slight</td>
<td>moderate</td>
</tr>
<tr>
<td>Shoulder to shoulder girdle</td>
<td>flexion</td>
<td>180</td>
<td>179-135</td>
</tr>
<tr>
<td></td>
<td>abduction</td>
<td>180</td>
<td>179-135</td>
</tr>
<tr>
<td>Shoulder (simple)</td>
<td>extension</td>
<td>60</td>
<td>59-40</td>
</tr>
<tr>
<td></td>
<td>internal rotation</td>
<td>90</td>
<td>89-45</td>
</tr>
<tr>
<td></td>
<td>external rotation</td>
<td>90</td>
<td>89-45</td>
</tr>
<tr>
<td>Elbow (simple)</td>
<td>flexion</td>
<td>30</td>
<td>31-70</td>
</tr>
<tr>
<td></td>
<td>abduction</td>
<td>180</td>
<td>179-150</td>
</tr>
<tr>
<td>Combined elbow</td>
<td>wrist pronation</td>
<td>90</td>
<td>89-45</td>
</tr>
<tr>
<td>radial shoulder</td>
<td>wrist supination</td>
<td>70</td>
<td>69-30</td>
</tr>
<tr>
<td>Carpal (combined)</td>
<td>flexion</td>
<td>105</td>
<td>106-145</td>
</tr>
<tr>
<td></td>
<td>extension</td>
<td>115</td>
<td>116-150</td>
</tr>
<tr>
<td></td>
<td>radial abduction</td>
<td>160</td>
<td>161-175</td>
</tr>
<tr>
<td></td>
<td>ulnar abduction</td>
<td>140</td>
<td>141-155</td>
</tr>
<tr>
<td>Hip (simple)</td>
<td>knee extension</td>
<td>90</td>
<td>91-120</td>
</tr>
<tr>
<td></td>
<td>knee flexion</td>
<td>60</td>
<td>61-90</td>
</tr>
<tr>
<td></td>
<td>extension</td>
<td>140</td>
<td>141-160</td>
</tr>
<tr>
<td></td>
<td>abduction</td>
<td>50</td>
<td>49-30</td>
</tr>
<tr>
<td></td>
<td>internal rotation</td>
<td>35</td>
<td>34-25</td>
</tr>
<tr>
<td></td>
<td>external rotation</td>
<td>45</td>
<td>44-25</td>
</tr>
<tr>
<td>Knee (complex)</td>
<td>flexion</td>
<td>135</td>
<td>134-90</td>
</tr>
<tr>
<td></td>
<td>abduction</td>
<td>180</td>
<td>179-170</td>
</tr>
<tr>
<td>Ankle (complex)</td>
<td>flexion</td>
<td>130</td>
<td>129-120</td>
</tr>
<tr>
<td></td>
<td>abduction</td>
<td>70</td>
<td>71-80</td>
</tr>
</tbody>
</table>

On the basis of the data presented in Table 1, it can be said that indicator deviations of joint angles by 1° are violations. Consequently, goniometric system must meet the requirements of measurement accuracy, with the threshold sensitivity of the measurement of mutual deviations and measuring range of motion being at least 1°. In this case, the measurement error must be smaller than this threshold. It should be noted that at present mechanical goniometers are widely used in medical diagnostics. Their accuracy does not meet the present requirements stated by the system of this class. Low accuracy of the measured parameters might be resulted from the design features of the device, and a high
degree of subjectivity of diagnosis due to professional experience and the influence of the human factor [5].

3 Information and technical support of the automated systems of diagnostics of the musculoskeletal system

The emergence of the movement results from neuro-cerebral human activity aimed at the implementation of any function. Therefore, the main objective of the design of goniometric systems is to assess the effectiveness of motor actions with respect to the application efforts of their execution [6].

For a comprehensive solution of this problem, we are to study design aspects of an automated diagnostic system of human musculoskeletal apparatus. The design is based on the synthesis of adequate informative physiological methods such as goniometry (accelerometer), computed tomography, electroencephalography (EEG) and electroneuromyography (ENMG). For this purpose, a medico-technical basis was formed. A block diagram of an automated system of complex real time diagnostics of the musculoskeletal system has been developed. The diagram consists of the circuit of accelerometric goniometer, an electroencephalograph and electromyography imaging is used as well (Fig. 1).

![Fig. 1. A block diagram of hardware and software of the automated system of goniometric control](image)

The above structure includes a measuring unit functioning in real time. It consists of a chain of inverters biokinematic driving parameters of the locomotor apparatus of man (accelerometric goniometer), recording units of psycho- and neurophysiological parameters (EEG and ENMG), and the registration equipment for bone and structural parameters (tomography).

The synchronous processing of the recorded parameters form time series, which are visualized with various degree of detail. The time series are the basis of a model of patient (Fig. 2). The model is processed by a neural network...
and is stored in the model database. The model that most closely matches the time series are instantiated. Pain thresholds and threshold of sensitivity of the patient for generating control signals to the actuators are determined by neural network algorithms. This is possible via a feedback method, whose implementation is carried out by the processing unit of evoked potentials of brain (patient’s reactions to test stimuli). Based on the processed data, operation mode of the actuators is generated and selected from a database of test techniques.

**Fig. 2.** The dynamic information model of patient

Thanks to neural network algorithms and decision support system (DSS) based on databases of the time series, diseases and evoked potentials, an approximate diagnosis of patient is determined.

It should be noted that the above adaptive goniometric control system includes both stationary and mobile measuring systems [6]. The number of monitored parameters is determined according to the severity of the patient’s pathology. In the case of low severity injuries, it is sufficient to use of a several portable accelerometric goniometers, guaranteeing the freedom of the patient’s movement. In the presence of more serious violations in the functioning of the musculoskeletal system is suspected, the accelerometric goniometers coupled with electroneuromyograph (Fig. 3a), EEG data and computing tomography (Fig. 3b) is recommended for use.

It is shown that the dynamic activity of brain neurons relating to the implementation of tool movements, typically starts 50-150 ms. prior to the occurrence of EMG activity and ended after a traffic stop. Thus, the joint reaches equilibrium during the dynamic development of the motor cortex neuronal activity phase long before the establishment of steady equilibrium level of neural activity. The maximum value of the mean frequency of neural activity in one bin of duration is 50 ms. In dynamic phase, reactions of neurons did not correlate with the magnitude of the equilibrium steady-articular angle (Fig. 5). At the same time, a positive correlation was revealed between the average frequency of neural
activity in the whole dynamic phase and magnitude response of articular angle [7, 8].

The presented results show that the maximum level of motor neuron activity depends primarily on the joint movement velocity and the duration of the movement. Thus, the obtained data contribute to definition of the criteria of permissible values, characterizing limits of the patient physiological parameters with respect to the normal ones; the limits are determined according to the degree of deviation and the conditions of pathology. Neurophysiological criteria are also formed based on statistical analysis of clinical studies of patients under normal conditions and in the presence of deviations.

4 The DSS for setting of a diagnosis

Development of DSS based on the dynamics analysis of the recorded time series for goniometric, kinematic and neurophysiological parameters is a complex and multicriterial problem. The algorithms of the DSS are based on Bayes’ rule. The rule accounts various heterogeneous types of input data expressing many kinds of deceases of the musculoskeletal system and a large number of symptoms. Bayes’ rule in a generalized form is as follows [9]:

\[
P (d \mid S_k \cap \cdots \cap S_1) = \frac{P (S_1 \cap \cdots \cap S_k \mid d) \cdot P (d)}{P (S_1 \cap \cdots \cap S_k)},
\]

where \( P(d) \) is a priori probability of the diagnosis \( d \), and \( S_1 \ldots S_k \) are functional physiological parameters.
Fig. 4. Interpretation of results. Angle support limb (40 degrees) is not greater than (equal to) the angle of the working limb (40 degrees). The right lower limb: abduction angle in the hip joint is 40 degrees, the average EMG gluteus 1 is 563.6 mV. The left lower limb: abduction angle in the hip joint is 40 degrees, the average EMG gluteus is 893.5 mV.; muscle operation mode stabilizing. The ratio of the two coefficients of reciprocity for medium gluteus is 1.75.

Fig. 5. The activity of neurons of the motor cortex and variations of the articular angle measured during the flexion movements at different joint velocity. Legend: 40° is the variation of the articular angle. The line parallel to the y-axis is the average frequency of pulses in the bin. The line parallel to the x-axis indicates the average frequency of the background activity of the neuron. Vertical lines indicate the boundaries of dynamic and stationary phases of motion. 1,2,3 are joint flexion speeds, N is number of iterations.
This formula requires \((m \cdot n)^2 + m^2 + n^2\) calculations of probability estimates, where \(m\) is the number of possible diagnoses, and \(n\) is the number of different variations. In order to calculate the total probabilities \(P(S_1 \cap \cdots \cap S_k)\), we are to calculate \(P(S_1/S_2 \cap \cdots \cap S_k) \cdot P(S_2/S_3 \cap \cdots \cap S_k) \cdots \cdot P(S_k)\).

Therefore, the model \(P\) for automated diagnostic expert system will be based on

\[
P(d \mid S) = \frac{P(S \mid d) \cdot P(d)}{(P(S \mid d) \cdot P(d) + P(S \mid d) \cdot P(d))}.
\]

The probability of the hypothesis \(d\) in the presence of certain abnormalities in the recorded data \(S\) is calculated based on the prior probability of the hypothesis without confirming abnormalities and the likelihood of having abnormalities in conditions that hypothesis is correct (event \(d\)) or incorrect (event \(\bar{d}\)). Therefore, for the problem of diagnosis of diseases of the musculoskeletal system, it appears that

\[
P(d \mid S) = \frac{P_{yes} \cdot P(d)}{(P_{yes} \cdot P(d) + P_no \cdot P(d))}.
\]

Let the pathology probability \(P(d)\) be equal to \(P\). The program generates a condition (parameters in the presence of pathology) and calculates the probability \(P(d \mid S)\) depending on the it’s implementation. The answer “Yes” (\(P_{yes}\)) confirms the above calculations, the answer “No” (\(P_no\)) does it too but with probability \((1 - P_{yes})\), and \((1 - P_no)\) instead \(P_{yes}\) and \(P_no\). Thereafter, the \(a\ priori\) probability \(P(d)\) is replaced with \(P(d \mid S)\). The program execution is cyclic, with the constant value \(P(d)\) refining at each iteration. The general scheme of the diagnosis selection algorithm is shown in Fig. 6.

The diagnosis selection algorithm structure consists of several branches:

Step 1. Enter the input data – a set of biometric, goniometric, neurophysiological and structural parameters; then the program retrieves information on the number of the diseases recorded having the corresponding symptoms from in the database (\(N\) is the number of relevant deviations disease, \(n\) is the number of the disease in question: \(0 \leq n \leq N\)).

Step 2. Set counter of a disease incrementally from the initial state \(n = 0\) till \(n \leq N\).

Step 3. Traverse all the \(a\ priori\) probabilities \(P(d)\), relating to the input data set and to the selected disease, to prioritize detected deviations. Deviations with the minimal likelihood are excluded from the probability set (\(J\) is the selected number of deviations in the set, \(j\) the number of the current deviation \(0 \leq j \leq J\)).

Step 4. Set the deviation counter incrementally from the initial state \(j = 0\) to \(j \leq J\).

Step 5. Select the deviation with the greatest probability of presence.

Step 6. Evaluate the degree of reliability of the diagnosis according to the interval \([-5, +5]\) (a scale). If the value belongs to the interval, then the program calculates the proportion of the degree of affiliation to a particular diagnosis parameters, using the corresponding weighting coefficients.

Step 7. Poll of the counter of registered deviations. If there is no new events of a decease then process the next unprocessed selected departure, go to step 4.
Enter the initial information: physiological parameters, the number of diseases (N)

\[ n = n + 1, \quad 0 \leq n \leq N \]

Viewing probability \( P(d) \), with the exception of deviations minimum probability

\[ j = j + 1, \quad 0 \leq J \leq N \]

Selection of deviation \( j \) with max value of a priori probability of the formation of the corresponding issue

\[
P(d|\bar{S}) = \frac{(P(S|d)P(d))}{(P(S|d)P(d) + P(S|\bar{d})P(\bar{d}))}
\]

Selecting the max value of an array of prior probabilities of diagnoses and the formation of an advisory diagnosis

Selecting the most probable diagnosis

Values calculated probabilities based on the works of deviations probabilities

End

**Fig. 6.** The scheme of algorithm of diagnosis selection
Step 8. Figure out new probabilities in Bayes rules. Specify the minimal and maximal values of the probabilities for each disease based on the currently existing *a priori* probabilities and assumptions that the remaining evidence will speak in favor of the hypothesis or contradict it. This step calculates the total conditional probabilities for each deviations. The hypotheses whose the minimal values are above a certain thresholds are considered as possible outcomes (possible diseases) and are subject to further diagnosis.

Step 9. Check the counter of the registered diseases.

Step 10. Sort the outcome list according to the probabilities, display subset with the maximal probability values as a recommendation for the diagnosis mentioned symptoms to physician.

Step 11. Display the recommended diagnosis.

Medical information system, which implements this algorithm, produces a finite set of the recommendations for doctor, emphasizing the presence of deviations registered with the diagnostics system sensors. Limiting the selection decision by a finite set of possible diagnoses is to reduce the probability of setting wrong preliminary diagnosis, eliminate human factor and increase the objectivity of the diagnosis of diseases.

5 DSS based on fuzzy logic and artificial neural networks

In order to develop the control unit of diagnostics system of the musculoskeletal system, we propose a method of computer support of diagnosis setting based on fuzzy logic and artificial neural networks. The method is represented in the form of two structural units: decision-making unit and knowledge base.

The decision (a diagnosis) is produced in two stages [10]. At the first stage – the preliminary diagnosis–, the system determines in advance the possible pathology and generate diagnostic recommendations based on data from the medical records and X-ray images. Then, at the stage of the goniometric diagnosis, together with EMNG, the recommendations are confirmed or rejected on the basis of the analysis of the obtained information, with immediate neural network processing by the diagnosis system.

The advantage of fuzzy logic is the ability to describe the operation of the system by means of fuzzy production rules (FPR) [11, 12]. The initial values of the parameters (used in FPR) for the normal cases or the pathological ones are determined at the beginning on the base of experts’ opinions. The values are adjusted with neural network engines.

A distinctive feature of this set of rules is the allocation of a separate group of so-called factors of pain diagnosed by EEG. Pain is one of the most important factors in the diagnosis of diseases of the musculoskeletal system. On the basis of these features we justify the choice of the rules of fuzzy productions of the form:

\[
\text{IF } \text{<Condition1} = \text{true} > \text{AND } ... \text{AND } \text{<ConditionN} = \text{true} > \\
\text{THEN } \text{<Consequent1} = \text{true} > \text{AND } ... \text{AND } \text{<ConsequentN} = \text{true} >
\]
Necessary linguistic input variables \((IN_LV)\) are selected according to a statistical base of the analysis of medical records and printed materials [13].

Membership functions (MF) \(\mu\) of a strict magnitude to a fuzzy term set (corresponding to one of the values of the input linguistic variables) were determined by means of the following expert evaluation techniques. Let expert \(E_1\) believe that the specific value of \(x^*\) belongs to the fuzzy term set at \(a_1 \leq x^* \leq b_1\); expert \(E_2\) at \(a_2 \leq x^* \leq b_2\); \ldots; \(E_g\) expert at \(a_g \leq x^* \leq b_g\). Then the term of MF \(\mu\) is obtained in the form shown in Figure 7.

\[\text{Fig. 7. The membership function of a term. } IN_LV \text{ is input linguistic variable, } LOut \text{ is a linguistic output}\]

In the figure, the horizontal axis shows the value of strict variable under fuzzification, where \(i = 1, g\) is an expert number. The vertical axis displays \(\mu\) fraction of all the experts who believe that its value of a \(x\) belongs to this linguistic value of the linguistic variable. This initially plots the MF’s, which is obtained at \(x \in (\min(a_i), \max(a_i)) \cup (\min(b_i), \max(b_i))\); the plot is curvilinear and will get a linear form if the least squares method is used.

The next stage is the aggregation construction applied to the RFP’s whose terms contain more than one sub-conditions. The conditional part of the rules is as follows:

\[
\text{IF } \langle IN_{LV1}=VALUE_{11} \rangle \text{ OR } \ldots \text{ OR } \langle IN_{LV1}=VALUE_{1m} \rangle \text{ AND } \langle IN_{LVn}=VALUE_{n1} \rangle \text{ OR } \ldots \text{ OR } \langle IN_{LVn}=VALUE_{nm} \rangle
\]

Each of the \(n\) terms \(\langle IN_{LVi}=VALUE_{i1} \rangle \text{ OR } \ldots \text{ OR } \langle IN_{LVi}=VALUE_{ij} \rangle\) consists of \(m\) subconditions \(\langle IN_{LVi}=VALUE_{ij} \rangle\), where \(VALUE_{ij}\) is the \(j\)-th value of the \(i\)-th \(LV\) in subconditions. Its number is determined by the number of input values \(LV : \ ij\). Let the truth degree of subconditions with the number \(ij\) be, respectively, \(\mu_{ij}\). The following RFP matrix \(M\) is formed for all the subconditions:

\[
M = \begin{pmatrix}
\mu_{11} & \mu_{12} & \cdots & \mu_{1m} \\
\mu_{21} & \mu_{22} & \cdots & \mu_{2m} \\
\vdots & \vdots & \ddots & \vdots \\
\mu_{n1} & \mu_{n2} & \cdots & \mu_{nm}
\end{pmatrix}.
\]
Using this matrix at the 3-6-th stages, we get the formula for calculating the confidence coefficient $\chi$ – the correctness precision of the system solution, – which is calculated for each of the possible diseases identified by DSS.

$$\chi_u = \frac{\min_{y_u} \left( \mu_{uv} \cdot \frac{\sum_{k=1}^{q_{uv}} F_k \cdot \min_i (\max_j (\mu_{ij}(x_i)))}{\sum_{k=1}^{q_{uv}} F_k} \right) dy_u}{\max_{y_u}} / \frac{\min_{y_u} \left( \mu_{uv} \cdot \frac{\sum_{k=1}^{q_{uv}} F_k \cdot \min_i (\max_j (\mu_{ij}(x_i)))}{\sum_{k=1}^{q_{uv}} F_k} \right) dy_u}{\max_{y_u}},$$  \hspace{1cm} (5)

where $\min$ and $\max$ are the left and right limits of the carrier interval of fuzzy set $LOu_{\omega_u}$ under consideration; $F_k$ are the weighting coefficients of the rules, $k = 1, q_{uv}$; $q_{uv}$ is the number of RFP, which is determined in the consequent of the $u$-th term of MF $LOv$; $\mu_{uv}(y_u)$ is the antecedent MF $v$-th term of $u$-th MF $LOu$.

It should be noted that the weights of the rules vary depending on the occurrence of new facts and results of fuzzy inferences at the previous stages. To resolve this uncertainty in form of an adjustment of RFP weights, an inference system is represented as a hybrid, i.e., fuzzy neural network (Fig. 8). Its structure is identical to the multilayer network, but its layers correspond to the stages of fuzzy inference, which has continuously carry out the following procedures:

- Input layer performs fuzzification function based on the specified input membership functions;
- Output layer implements the defuzzification function;
- Hidden layers reflect the totality of the RFP and the output stages: aggregation, activation and accumulation.

![Fig. 8. The structure of hybrid neural network](image-url)
Neurons \( \min, \max \) and \( \text{sum} \) indicated in Fig. 8 act as appropriate mathematical functions. Neurons, marked with “\( \times \)” is transmitted to the output product of the input signals. Symbol “\( \equiv \)” marks neurons establishing a correspondence between \( LV \) and an intermediate \( LOut \), and symbols “\( \wedge \)” are neurons realizing operation fuzzification for each term of \( \text{INLV} \). Node “\( a/b \)” denotes division of the input value by the sum of the weights of active rules. Neuron “\( \text{def} \)” implements the function of defuzzification, with applying the method of gravity center.

The fuzzy rule selection engine is represented as \( \text{INLV} \) inputs having “0” (a rule is selected) and “1” (a rule is not selected) logic levels weights multiplied by the corresponding membership functions \( \mu_{ij}(x_i) \). Here index \( i \in \{1, n\} \) is the number of \( I\!N_{LV} \) and index \( j \in \{1, m\} \) is the number of its term.

A neural network is trained by the algorithm of error back-propagation modified for use in the fuzzy neural networks. The layers of neurons with specified parameters are represented by one layer with a complex activation function, fuzzy artificial neural network (ANN) is identical a three-layer ANN with one hidden layer. Thus, network training is reduced to a three-layer perceptron learning. It is worth noting that the fuzzy neural network is used only in the case of changes of the DSS structure (change aggregate \( RFP \), input or output \( LV \)), and in the case of the appearance of new evidence proving or disproving the previously known data in the literature or medical practice.

6 The results of research

This section is devoted to the results of the benchmark tests of the calculated model; the results are obtained with an installed bodily machinery of the human skeleton. During the investigation, we used the method of mechanical goniometry in collaboration with an orthopedic doctor. The method of accelerometer goniometry usage is proposed in [5].

### Table 2. Average values of measurement errors for accelerometric and mechanical goniometers

<table>
<thead>
<tr>
<th>Motion pattern</th>
<th>Error of an accelerometric goniometer</th>
<th>Error of a mechanical goniometer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bending</td>
<td>( \pm 0.02 )</td>
<td>( \pm 1.00 )</td>
</tr>
<tr>
<td>Abduction</td>
<td>( \pm 0.02 )</td>
<td>( \pm 1.00 )</td>
</tr>
<tr>
<td>Extension</td>
<td>( \pm 0.02 )</td>
<td>( \pm 1.00 )</td>
</tr>
<tr>
<td>Internal rotation</td>
<td>( \pm 0.02 )</td>
<td>( \pm 1.00 )</td>
</tr>
<tr>
<td>External rotation</td>
<td>( \pm 0.02 )</td>
<td>( \pm 1.00 )</td>
</tr>
<tr>
<td>Tremor imitation</td>
<td>( \pm 0.02 )</td>
<td>( \pm 3.50 )</td>
</tr>
</tbody>
</table>

The results in Table 1 confirmed that the application of the accelerometric goniometer improves the diagnosis accuracy in average \( 1^\circ 44' \) compared to mechanical goniometer.
In addition, the diagnosis was carried on 20 patients who underwent rehabilitation after complicated shoulder injury and wrist. It should be noted that in all cases the diagnosis was a medical opinion on the normal rates. However, based on the neural network processing of the recorded data of the goniometric control, the electroneuromyographic and the tomographic control, a dial of a estimation of a motion in the joints from the Table 1, it was found that in 14 cases the medical diagnosis coincided with the diagnosis of DSS, in 4 cases were diagnosed of a functional deviation of the work wrist in a small extent, and in 2 cases were more prominent deviation of the combining elbow-shoulder joint in small extent.

7 Conclusion

Data processing algorithms and approaches to designing a system of diagnostics of the musculoskeletal system are presented in the article. The obtained results allow one to
- define the evaluation criteria of the “current state” of the musculoskeletal system;
- determine the severity of biomechanical disorders with a high degree of confidence;
- predict the biomechanical disorders of the musculoskeletal system;
- have the possibility of a science-based rehabilitation prognosis;
- create and optimize individual training programs that promote the advancement of technical training of athletes and prevent the diseases.

The results show that the diagnosis of the functional state of the musculoskeletal system based on the proposed system is an informative method of detecting violations. This technique is recommended to be used as a supplement to conventional methods of examination of the musculoskeletal system condition, as well as a stand alone technique.

Acknowledgments. This work was supported by RFBR grant 16-38-00704 mol_a.

References

The Study of Inversion Problems of Cryptographic Hash Functions From MD Family Using Algorithms for Solving Boolean Satisfiability Problem

Irina Gribanova, Oleg Zaikin, Stepan Kochemazov, Ilya Otpuschenikov, and Alexander Semenov
Matrosov Institute for System Dynamics and Control Theory SB RAS, Irkutsk, Russia
\{the42dimension, zaikin.icc, veinamond\}@gmail.com, \{otilya, biclop.rambler\}@yandex.ru

Abstract. In this paper we present the results of application of state-of-the-art SAT solvers to inversion of cryptographic hash functions from the MD family. In particular we consider the problems of finding preimages and collisions for MD4 and MD5. To solve them we use the approach based on reducing the original problems to Boolean satisfiability problem (SAT). The propositional encoding of the algorithms specifying the considered functions was performed using the Transalg software system. The features of this system make it possible to effectively augment the SAT encodings for MD4 and MD5 hash functions with various additional constraints that improve the effectiveness of SAT solvers on corresponding instances. The effectiveness of the proposed algorithms is better than that in a number of preceding papers. We used the developed algorithms to find new families of two-block collisions for MD5 and to construct new differential paths for finding single-block collisions for MD4.

Keywords: SAT, MD4, MD5, cryptanalysis, preimage attack, collision, inversion problem, hash functions, Transalg.

1 Introduction

Let us denote by \( \{0,1\}^k, k \in \mathbb{N} \) the set of all binary words of length \( k \). By \( \{0,1\}^* \) we denote the set of all binary words of an arbitrary finite length (i.e. \( \{0,1\}^* = \bigcup_{k=0}^{\infty} \{0,1\}^k \), where \( \{0,1\}^0 = \emptyset \)).

Let us remind [10] that a hash function is a total computable discrete function of the kind \( \chi : \{0,1\}^* \to \{0,1\}^C \), where \( C \) is some constant representing a length of hash value. Hash functions are used in various areas of computer science. In particular, they are applied to speed up access to large data sets. In cryptography, the scope of hash functions applications is also quite wide and has both theoretical (to construct proofs in random oracle model [4]), and practical
aspects. For example, in all state-of-the-art algorithms for constructing digital
signatures, the message to be signed is first hashed. Cryptographic hash func-
tions have to meet additional requirements, which consist in the fact that the
inversion problems of such functions should be computationally hard. In partic-
ular, the problem of finding a preimage for a given hash value and the problem
of finding collisions should both be hard. Let us remind that if there exists a
pair of messages $x_1, x_2 \in \{0, 1\}^*$, $x_1 \neq x_2$ such, that $\chi(x_1) = \chi(x_2)$, then these
messages form a collision for hash function $\chi$.

In the present paper we describe the results of applying algorithms for solving
Boolean satisfiability problem (SAT) to finding preimages and collisions of some
cryptographic hash functions. These results are a part of actively developing di-
rection of research known as SAT-based cryptanalysis. In particular, within the
context of mentioned problems we present the results of SAT-based cryptanal-
ysis of MD4 and MD5 cryptographic hash functions. Hereinafter by inversion
problems we mean both the problem of finding preimages and the problem of
finding collisions for considered hash functions.

Let us give a brief outline of the paper. In the next section we present the
theoretical foundations of SAT-based cryptanalysis. In the third section we con-
sider finding preimages and collisions for MD4 and MD5 as special cases of SAT,
paying special attention to characteristic features of obtained SAT instances. In
the fourth section we describe new methods and results obtained using them.
The fifth section is a short review of the related works preceding our research.

2 Theoretical Foundations of SAT-based Cryptanalysis

Boolean Satisfiability Problem (SAT) consists in the following: for an arbitrary
Boolean formula $F$ to decide whether it is satisfiable or not, i.e. if there exists
a set of values of Boolean variables from this formula that makes it TRUE. In
general case this problem can be effectively (in polynomial time on the size of
code of $F$) reduced to SAT for formula $C_F$ in the Conjunctive Normal Form
(CNF). That is why nowadays for convenience SAT is usually considered for
some CNF. Also SAT is often understood as a search variant of the problem:
given a CNF $C$ to decide if it is satisfiable and if the answer is ‘Yes’ to provide
a corresponding satisfying assignment.

In the last 15 years there was achieved a dramatic progress in the devel-
opment of SAT-solving algorithms and techniques [6]. Today these algorithms
are successfully applied in symbolic verification, computational combinatorics,
bioinformatics and many other areas. In the recent years there can be seen a
growing interest to applying SAT solvers to cryptanalysis problems. The corre-
sponding direction is now known as SAT-based cryptanalysis and the number of
works related to it is steadily growing.

SAT-based cryptanalysis relies on the ability to effectively reduce inversion
problems of discrete functions to SAT. The effectiveness of these algorithms,
called propositional encoding procedures, follows from the Cook theorem [8]. Let
us consider some discrete function $f : \{0, 1\}^* \rightarrow \{0, 1\}^*$ specified by a polynomial
algorithm $A$. For each $n \in N$ this algorithm specifies a function of the kind: $f_n : \{0,1\}^n \rightarrow \{0,1\}^*$. Hereinafter by inversion problem for discrete function $f_n$ we mean the following problem: given $y \in \text{Range} f_n$ and known algorithm $A$ to find such $x \in \{0,1\}^n$ that $f_n(x) = y$. In the context of this general problem it is possible to consider many cryptanalysis problems. For example, suppose that $A$ is an algorithm which specifies hash function $\chi : \{0,1\}^* \rightarrow \{0,1\}^C$. Then for a specific $n$ program $A$ specifies a function $\chi_n : \{0,1\}^n \rightarrow \{0,1\}^C$ and for a given hash image $y$ of some message $x$ we can consider the problem of finding a preimage $x'$, $\chi_n(x') = y$, where $x'$ does not have to be equal to $x$.

The advantage of the presented above style of formulating cryptanalysis problems lies in the fact that we can effectively reduce any such problem to SAT for some satisfiable CNF $C(f_n, y)$. In recent years there appeared several different automated software systems designed to perform the corresponding reductions. From our point of view the most interesting of them are URSA [21] and Cryptol [18, 19]. In our computational experiments we used the Transalg system [29, 30] developed by us. Let us briefly describe its functional capabilities.

The Transalg system was specifically developed to automate reducing the inversion problems of discrete functions, specified by algorithmic descriptions, to SAT. It is based on Cook’s ideas on propositional encoding of algorithms [8] and King’s ideas on symbolic execution [24]. The Transalg system uses the domain specific TA-language with C-like syntax [23] to specify considered discrete functions. In the process of translating TA-programs to SAT it uses standard techniques from the compilation theory. The result of the translation is not an executable code but a set of Boolean formulas that naturally corresponds to a system of Boolean equations. Using Tseitin transformations [37] we can construct a SAT encoding from this set.

Thus, the initial stage of constructing a SAT instance encoding an inversion problem of some discrete function consists in describing this function on a specialized TA-language. Since TA-language is a procedural C-like language, it is usually enough to make small changes to the existing implementation of an algorithm written in the C language to obtain the corresponding TA-program. The TA-language supports basic constructions typical for procedural languages (variable, array and function declarations, assignment operators, conditional operators, loops, function calls), various bit and integer operations including bit shifting and numerical comparison.

Hereinafter, by translation we mean the process of constructing the propositional encoding of a discrete function for a given TA-program. The translation of any TA-program consists of two main stages. At the first stage Transalg parses the source code of a TA-program and constructs a syntax tree using standard techniques of the compilation theory [1]. At the second stage the system employs the concept of symbolic execution [24] to construct a propositional encoding of a considered TA-program. The propositional encoding can be provided in several standard formats: CNF, DNF (Disjunctive Normal Form), ANF (Algebraic Normal Form). Transalg has an option of applying minimization to
constructed Boolean formulas. For this purpose it uses the ESPRESSO Boolean minimization library which was embedded into system as one of its modules.

Let us briefly discuss the SAT solving algorithms used in SAT-based cryptanalysis. According to the Cook theorem, SAT is NP-hard problem. That is why it is unlikely that there are polynomial deterministic algorithms for its solving. Nevertheless, ‘effective’ (that make it possible to solve practical instances in reasonable time) algorithms for solving SAT are in high demand in a number of important areas of science, first of all in formal verification. In recent years this demand led to a ‘boom’ in the development of strategies and heuristics for solving SAT.

There are a number of general concepts underlying state-of-the-art SAT solvers. The CDCL (Conflict Driven Clause Learning) concept first described in [25] became one of the most fruitful. Since CDCL-solvers show the highest effectiveness in application to inversion of cryptographic functions, let us briefly describe their design features. The CDCL-solver is based on DPLL algorithm (Davis-Putnam-Logemann-Loveland) [13], complemented by Clause Learning technique, which allows to store the information about traversed fragments of the search space in the form of conflict clauses. A conflict clause is a logical implication of an original CNF $C$, so the conjunction of this clause with $C$ gives CNF $C'$, that is satisfiable exactly on the same assignments as $C$ (or both are unsatisfiable). Conflict clauses are used to derive new information and therefore direct the search to a new path. In the worst case scenario both DPLL algorithm and DPLL + CDCL algorithms are exponential. It follows from the fact that the resolution method is exponential [20] and from the results of [3]. Nevertheless, state-of-the-art CDCL-solvers are surprisingly successful when dealing with ‘industrial’ SAT problems of large dimensions (tens of thousands of variables, hundreds of thousands of clauses). As it was already mentioned, it is the CDCL-solvers that perform best on cryptographic tests, including the inversion problems for hash functions.

3 Inversion Problems of MD4 and MD5 as SAT

In the present paper we apply SAT-based cryptanalysis to the problems of finding collisions and to inversion (i.e. finding preimages) of cryptographic hash functions MD4 [32] and MD5 [31]. Let us briefly discuss design features of these functions.

There is a number of algorithmic constructions used for building cryptographic hash functions. One of the most commonly used is the Merkle-Damgard construction [27, 12], underlying such well-known families of cryptographic functions as MD and SHA. This construction implies that an original message is split into blocks of fixed size (if necessary the length of a message can be extended using the so-called ‘padding’). During the processing of the initial message each block is treated one at a time using special function called compression function. It converts each block into a short message, stored in specially allocated memory registers. The values of the registers are considered as the values of the so-called ‘chaining variables’. After processing all message blocks the concate-
nation of current values of chaining variables gives us the hash value. At the initial moment the chaining variables are assigned with initial values, which are commonly known and usually are fixed in the algorithm specification. The general scheme of Merkle-Damgard construction is shown in Figure 1 (the message blocks that form the original message are denoted by $M_1, \ldots, M_N$, the initial values of chaining variables are denoted by $IV$, and $H_N$ is a final hash value).

Let us illustrate the work of the Merkle-Damgard construction using the MD5 algorithm as an example. The MD5 algorithm begins with padding an original message. In this process we add to it a special bit sequence $p_1, \ldots, p_q$, $q \geq 65$. The bit $p_1$ is always equal to 1, while bits $p_{q-63}, \ldots, p_q$ are the binary representation of 64-bit number equal to the number of bits in the original message. Bits $(p_2, \ldots, p_{q-64})$ if they are present, are equal to 0. The parameter $q \geq 65$ is chosen in such a way that the length of the padded message is a multiple of 512 bits. After padding the message is divided into 512-bit blocks $M = (M_1, \ldots, M_N)$. According to the Merkle-Damgard construction an iterative process of calculating the hash value can be described by the equation:

$$H_i = f(H_{i-1}, M_i), 1 \leq i \leq N,$$

where $H_0 = IV$ is the initial value and $H_N$ is the value of hash function. For each $i \in \{1, \ldots, N\}$ the value $H_i$ is the concatenation of 32-bit values of chaining variables. At the initial moment these variables have the following values (according to the specification of the MD5 algorithm): $a=0x67452301$, $b=0xefcdab89$, $c=0x98badcfe$, $d=0x10325476$. Concatenation of these values, i.e. the word $a\|b\|c\|d$, forms IV.

Schematically the compression function $f_{MD5}$ used in the MD5 algorithm is presented in Figure 2. Let us comment the Figure 2. The function $f_{MD5}$ receives as input an array with the current values of chaining variables – $a_0, b_0, c_0, d_0$. The corresponding values are given in form of 32-bit words. Also compression function $f_{MD5}$ operates with 512-bit input message block $M_i$, divided into 16 32-bit words: $M_i = (m_1, \ldots, m_{16})$. The process of computing $f_{MD5}$ can be divided into 4 stages, called rounds. In each round an iterative recalculation of chaining variables is performed, the value of each variable is updated 4 times (in Figure 2, each round is represented by transformation $\Phi^j, j \in \{1,2,3,4\}$). The transformations corresponding to the first update of chaining variables values in
the first round are specified by the following equations:

\[
\begin{align*}
    a_1 &= b_0 + ((a_0 + \phi^1(b_0, c_0, d_0) + m_1 + t_1) \ll s^1_a), \\
    d_1 &= a_0 + ((d_0 + \phi^1(a_0, b_0, c_0) + m_2 + t_2) \ll s^1_d), \\
    c_1 &= d_0 + ((c_0 + \phi^1(d_0, a_0, b_0) + m_3 + t_3) \ll s^1_c), \\
    b_1 &= c_0 + ((b_0 + \phi^1(c_0, d_0, a_0) + m_4 + t_4) \ll s^1_b).
\end{align*}
\] (1)

Here, ‘+’ means addition modulo $2^{32}$, ‘$\ll s$’ means $s$-fold circular left shift of a 32-bit word, $s^1_\zeta$, $\zeta \in \{a, b, c, d\}$ are known constants defined in the algorithm specification (for example, $s^1_a = 7$, $s^1_d = 12$, $s^1_c = 17$, $s^1_b = 22$), constants $t_k$, $k \in \{1, 2, 3, 4\}$ are also known. The result of second recalculation of chaining variables values is the set of values $a_2, b_2, c_2, d_2$ computed using formulas that are different from (1) in that their right parts contain $a_1, b_1, c_1, d_1$ instead of $a_0, b_0, c_0, d_0$ and $m_5, m_6, m_7, m_8$ instead of $m_1, m_2, m_3, m_4$. Also the constants $t_k$, $k \in \{5, 6, 7, 8\}$ are used. The round ends after four recalculations, i.e. after all the words $m_1, \ldots, m_{16}$ have been used. The equations used to recalculate chaining variables values at further rounds are generally similar to (1) with the difference that they use other constants and functions $\phi^j$. The functions $\phi^j$, $j \in \{1, 2, 3, 4\}$ are functions of the kind:

\[
\phi^j : \{0, 1\}^{32} \times \{0, 1\}^{32} \times \{0, 1\}^{32} \rightarrow \{0, 1\}^{32}.
\]

They are called round functions and are specified as follows:

\[
\begin{align*}
    \phi^1(X, Y, Z) &= (X \land Y) \lor (\neg X \land Z) \\
    \phi^2(X, Y, Z) &= (X \land Z) \lor (Y \land \neg Z) \\
    \phi^3(X, Y, Z) &= X \oplus Y \oplus Z \\
    \phi^4(X, Y, Z) &= Y \oplus (X \land \neg Z)
\end{align*}
\]

In these formulas it is assumed that Boolean operations are performed componentwise over 32-bit words.

The sequence of data transformations listed above forms the algorithm that defines the hash function $\chi_{MD5} : \{0, 1\}^* \rightarrow \{0, 1\}^{128}$. When we fix the length of an original message we obtain functions, the corresponding inversion problems of which can be reduced to SAT using the techniques described above. In what follows we use the TRANSALG system to perform this reduction. Below we will
briefly consider the important stages of the process of constructing propositional encoding of the algorithm specifying $\chi_{MD5}$.

As it follows from the above, the value $\chi_{MD5}$ is formed as a result of an iterative recalculation of chaining variables values. The operations used in this process are integer addition modulo $2^{32}$, circular bit shift operation and bitwise conjunction, disjunction and addition modulo 2. Let $a = (a_n, \ldots, a_1)$, $b = (b_n, \ldots, b_1)$ be $n$-bit numbers (let us use the notation in which the most significant bit is the leftmost bit). Then the integer addition $a + b$ is encoded by the following set of Boolean formulas:

$$

c_1 \equiv a_1 \oplus b_1 \\
p_1 \equiv a_1 \land b_1 \\
c_i \equiv a_i \oplus b_i \oplus p_{i-1}, i = 2, \ldots, n \\
p_i \equiv a_i \land b_i \lor a_i \land p_{i-1} \lor b_i \land p_{i-1}, i = 2, \ldots, n \\
c_{n+1} \equiv p_n
$$

Here, $p_i, i = 1, \ldots, n$ are carry bits, $c = a + b = (c_{n+1}, c_n, \ldots, c_1)$ represents the result of addition. In case of integer addition modulo $2^{32}$ numbers $a, b$ are represented by 32-bit vectors ($n = 32$), and as a result of the operation we take 32 less significant bits of vector $c$.

As it was shown in [29], when encoding the circular shift operation it is not necessary to create new Boolean variables. This is a consequence of the fact that the circular shift can be considered as a process of redefinition of relations between variables, already used at the previous stages of construction of propositional encoding.

Operations that appear in round functions are bitwise. Thus, for example, propositional code for the function $\phi^1(X, Y, Z) = (X \land Y) \lor (\neg X \land Z)$ is based on the set of formulas of the kind:

$$
\upsilon_i \leftrightarrow (x_i \land y_i) \lor (\neg x_i \land z_i), i = 1, \ldots, 32,
$$

where $\upsilon_i$ is a Boolean variable encoding $i$-th component in the vector $\phi^1(X, Y, Z)$, $X = (x_1, \ldots, x_{32}), Y = (y_1, \ldots, y_{32}), Z = (z_1, \ldots, z_{32})$.

All of these procedures can be described using the TA-language. The result of the translation of the constructed TA-program is the CNF $C(\chi_{MD5}, m)$ that encodes the calculation of $\chi_{MD5}$ for an input of a fixed length $m$. To obtain the problem of inversion for this function in the form of SAT instance one has to add to this CNF the specific value of the hash, for which we want to find a preimage, in the form of unit clauses. To construct the SAT encoding for finding a collision (in general case, the $k$-block collision) two similar TA-programs are used, resulting in the construction of two CNFs $C_1(\chi_{MD5}, m)$ and $C_2(\chi_{MD5}, m)$ over disjoint sets of variables. Then we construct the CNF $C_1(\chi_{MD5}, m) \land C_2(\chi_{MD5}, m)$ to which we append the conditions specifying that functions outputs are equal, while functions inputs are not. From the properties of translation procedures used in Transalg, it follows that the resulting CNF, hereinafter denoted by $\tilde{C}$, has as many satisfying assignments as there are collisions of length $m$. We will say that CNF $\tilde{C}$ encodes the problem of finding collisions of the corresponding hash function (MD5 or MD4).
4 Solving Inversion Problems for Hash Functions from MD family via SAT

Apparently, the first example of applying SAT solvers to cryptanalysis of hash functions can be found in paper \[22\]. In this work, there were considered the problems of finding collisions of hash functions from MD family in form of SAT. However, in \[22\] the collisions for the unweakened algorithms have not been found. From this perspective the first successful experience was presented in \[28\]. In this paper there was implemented the SAT-version of the famous differential attack by X. Wang. Using this attack, first introduced in \[38, 39\], one can effectively construct single-block collisions for MD4 and two-block collisions for MD5. Below let us briefly consider the main stages of the Wang scheme.

Thus, consider two messages $M = (M_1, \ldots, M_N)$, and $M' = (M'_1, \ldots, M'_N)$, $M \neq M'$. Let

$$H_i = a | b | c | d, \quad H'_i = a' | b' | c' | d'$$

be the values of chaining variables after calculating hash values for message blocks number $i \in \{0, 1, \ldots, N\}$. Consider the following relation:

$$\Delta H_i = a - a' | b - b' | c - c' | d - d'$$

in which we use integer differences modulo $2^{32}$ between the corresponding values of chaining variables. Given this notation a differential path for a considered hash function can be defined as follows:

$$\Delta H_0 \xrightarrow{(M_1, M'_1)} \Delta H_1 \xrightarrow{(M_2, M'_2)} \cdots \xrightarrow{(M_{N-1}, M'_{N-1})} \Delta H_{N-1} \xrightarrow{(M_N, M'_N)} \Delta H_N = \Delta H$$

It’s obvious that $\Delta H_0 = 0$. If $\Delta H = 0$, then the corresponding messages $M$ and $M'$ form an $N$-block collision. If for some reasons there were selected a particular type of differences $\Delta H_i$, $1 \leq i \leq N$ and of differences between the intermediate values of chaining variables during the calculation of $H$ and $H'$, then the problem of finding collisions transforms into the problem of selecting a pair $(M, M')$ that satisfies the resulting differential path. This is the main idea of differential attacks on MD family presented in \[38, 39\]. More specifically, in \[38, 39\] for the MD family hash functions there have been proposed differential paths and described the collision finding method which consists of two simple steps: random message selection and deterministic modification of messages to fit a particular differential path. This method has allowed to construct single-block collisions for the MD4 algorithm and two-block collision for the MD5 algorithm in a reasonable time. The attacks of the described type were often applied in later papers. Among the latest achievements in this direction there should be noted the paper \[36\], in which single-block collisions for the MD5 hash function were built using differential attacks.

As we noted above, in \[28\] there was implemented a ‘SAT-version’ of Wang attack. More precisely, the Boolean constraints encoding the differential paths from \[38, 39\] were added to the propositional encoding for finding collisions for the MD4 and MD5 hash functions. To find one MD4 collision it took them...
about 10 minutes (500 seconds on average) using Minisat solver [16]. Finding two-block collisions for MD5 proved to be much more difficult.

In our opinion the main disadvantage of [28] is the use of highly specialized tools to construct propositional encodings, because that makes it impossible to reproduce their experiments. In our experiments we used the Transalg system. The encodings obtained were more compact compared to encodings presented in [28]. Using cryptominisat solver [35] one collision for MD4 was found in a fraction of a second and to find 10,000 collisions it took less than 2 hours. To find two-block MD5 collisions using the encodings constructed by Transalg we employed Plingeling and Treengeling solvers [5], the winners of the latest SAT competitions. Using Treengeling solver we managed to isolate one special class of two-block collisions for MD5 with the first 10 zero bytes. To find one collision on one node of ‘Academician Matrosov’ computing cluster of Irkutsk Supercomputer Center SB RAS (two 16-core Opteron 6276 processors + 64 Gb RAM)\(^1\) it took from a few hours to a couple of days. There were found dozens of collisions of this type. In Figure 3 we present one of them.

<table>
<thead>
<tr>
<th>M(_1)</th>
<th>00 00 00 00 00 00 00 00 20 74 67 a6 f5 48</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cb c1 6d a5 3e f7 b8 bc 67 8d 9a 3b 67 c5</td>
</tr>
<tr>
<td></td>
<td>ff d1 2e c9 a0 80 9a b3 7d e7 f5 bc 2a 4e</td>
</tr>
<tr>
<td></td>
<td>2d b8 d4 13 4c cc 7b 1b 00 29 8b f5 53 7a</td>
</tr>
<tr>
<td></td>
<td>5d f7 b7 7a af 36 ce 08 1e 44 a2 d0 84 8d</td>
</tr>
<tr>
<td></td>
<td>c5 4c 28 89 75 b3 84 ac 97 7f f2 7e 50 4d</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>M(_2)</th>
<th>00 00 00 00 00 00 00 00 20 74 67 a6 f5 48</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cb c1 6d 25 3e f7 b8 bc 67 8d 9a 3b 67 a6</td>
</tr>
<tr>
<td></td>
<td>db 4c 6f 4e 6d f0 8d b4 7d e7 f5 bc 2a 4e</td>
</tr>
<tr>
<td></td>
<td>ff d1 2e c9 a0 80 9a b3 7d e7 f5 bc 2a 4e</td>
</tr>
<tr>
<td></td>
<td>2d b8 d4 93 4c cc 7b 1b 00 29 8b f5 53 7a</td>
</tr>
<tr>
<td></td>
<td>5d f7 b7 7a af 36 ce 08 1e 44 a2 d0 84 8d</td>
</tr>
<tr>
<td></td>
<td>c5 4c 28 89 75 b3 84 ac 97 7f f2 7e 50 4d</td>
</tr>
</tbody>
</table>

\(\text{Hash}\) | c22664780a9766ce57065eba36af06b |

Fig. 3. An example of two-block MD5 collision with first 10 zero bytes.

Next we applied SAT approach to generate new differential paths (different from that proposed by X. Wang et. al.) for finding collisions for the MD4 hash function. Thus, consider the problem of computing MD4 hash values for two different messages \(M\) and \(M'\). As we mentioned above, by \(H\) and \(H'\) we denote hash registers, where \(H\) is a hash image of \(X\), and \(H'\) is a hash image of \(X'\). We assume that hash images for \(X\) and \(X'\) are calculated simultaneously. The data recorded in \(H\) and \(H'\) in a particular time moment can be considered as an

\(^1\) http://www.hpc.icc.ru/
instantaneous configuration of memory registers of the computing device that implements two independent copies of the MD4 algorithm. In total in the given algorithm there are 48 such instantaneous configurations corresponding to basic steps – each round consists of 16 steps (in the case of MD5 there are 4 rounds of 16 steps, making it a total of 64 steps). Let us denote these configurations by \( h_k \) and \( h'_k \) for registers \( H \) and \( H' \), respectively, \( k \in \{1, \ldots, 48\} \) \( (k = 0 \) corresponds to an initial configuration). Thus, on the \( k \)-th step the configurations \( h_k \) and \( h'_k \) contain four 32-bit numbers each, corresponding to the values of chaining variables from the considered step. During the transition from the configuration \( h_k \) to the configuration \( h_{k+1} \) only one of the four chaining variables values is recalculated (the same holds for the transition from \( h'_k \) to \( h'_{k+1} \)). Let us consider the difference between 32-bit integers modulo \( 2^{32} \) only for the values of chaining variables that were recalculated during the transition \( k \to k + 1 \) and denote it as \( \delta_{k+1} \).

Let \( \tilde{C} \) be a CNF encoding the problem of finding collisions for the MD4 hash function. By \( C_\Delta \) we denote a CNF, which takes the value 1 (true) if and only if hashed messages \( M \) and \( M' \) satisfy a fixed differential path (for example, the one from [38]). Let \( C_{\delta_k = c} \) be a CNF which takes the value 1 (true) if and only if the value of the difference \( \delta_k \) is equal to \( c \) (here \( c \) is some 32-bit number). Since in each particular case \( c \) is a 32-bit constant it is not difficult to go through all possible values of \( c \) and consider the SAT for corresponding CNFs of the kind:

\[
\tilde{C} \land C_\Delta \land C_{\delta_k = c}.
\]

In our experiments we found that for the vast majority of possible values state-of-the-art SAT solvers can prove unsatisfiability of CNF \( \tilde{C} \land C_\Delta \land C_{\delta_k = c} \) very fast (in a fraction of second). The solving of SAT for CNFs of the kind \( \tilde{C} \land C_\Delta \land C_{\delta_k = c} \) for which we could not obtain the answer in a relatively short period of time (a few seconds), were interrupted. Assume that \( c' \) is the value of \( \delta_k \) for which the solving of SAT instance \( \tilde{C} \land C_\Delta \land C_{\delta_k = c} \) was interrupted and that this \( c' \) is not equal to corresponding value of \( \delta_k \) in the path from [38]. Then this value can be considered as a candidate for the value of \( \delta_k \) in a new differential path. To solve the corresponding SAT instance it is possible to spend more effort using embarrassing parallelism if necessary. If \( \tilde{C} \land C_\Delta \land C_{\delta_k = c'} \) turns out to be satisfiable, then it means that we constructed a new differential path that is different from the Wang path in the value of \( \delta_k \).

This algorithm has been implemented in the form of parallel MPI-program, and launched on the computing cluster ‘Academician Matrosov’ of Irkutsk Supercomputer Center SB RAS. As a result, several differential paths for MD4, that are different from the Wang path, have been found. Below in Figure 4 we present one of the found paths, which is different from the Wang path in values \( \delta_k \) for \( k \in \{13, 17, 20, 21\} \). Using this differential path and message difference from [38] we constructed CNF \( \tilde{C}' \) and used CRYPTOMINISAT solver [35] to evaluate its effectiveness. It took the solver 416 seconds to find 1000 collisions. When applied to the CNF \( \tilde{C} \) based on the Wang differential path it took CRYPTOMINISAT 520 seconds to find 1000 collisions.
We also applied the SAT approach to the problem of finding preimages for the weakened version of MD4 algorithm. Despite the fact that the problem of finding collisions for MD4 can be effectively solved, the problem of finding preimages for the full version of this function remains computationally hard. We started our research in this direction from reviewing the results of [15], where there were showed that two-round version of MD4 can be easily inverted. In [14] there was used the SAT approach to find preimages of weakened versions of MD4. Again, the results of computational experiments from this work are hard to reproduce due to the use of specialized tools for constructing propositional encodings of the considered functions. In our experiments we used the propositional encoding constructed by means of the Transalg system.

The attack on MD4, described in [15] is an iterative procedure, in which each iteration includes two stages: random selection of some data and calculating hashed message blocks based on the selected data. Randomly generated data is actually supposed to fill the contents of some of the hash registers at certain steps. Thus, the values corresponding to the chaining variable $a$, assumed to be equal to a constant $K$ on steps 12, 16, 20, 24. This and similar constraints are assumed also regarding other chaining variables (in total 12 constraints). The constant $K$ is chosen randomly at each new iteration. Besides $K$ at each iteration the value of chaining variable $b$ previous to the calculation of a hash value is randomly selected (we remind that two-round version of MD4 is considered, i.e. 32 steps of algorithm). Then using the available data the blocks of the hashed message are calculated. For this purpose the equations for calculating chaining variables at corresponding steps are used. It is possible that as a result the value of $b$ will not coincide with its randomly selected value. In this case the iteration is considered unsuccessful. If the corresponding values coincide, then the preimage for the hash was found.

In [14] there was proposed a SAT-attack on MD4, based on the ideas of [15]. More precisely, in [14] the constant $K$ was fixed to be equal to 0. The values of the variable $b$ in steps preceding the calculation of a hash were not fixed. Also, the authors of [14] rejected one of the conditions from [15]. Thus, in [14] there were used 11 conditions. The conditions of ‘Dobbertin’ type were added to the propositional encoding of the MD4 algorithm in the form of unit clauses. The best result presented in [14] was successful solving of inversion of 39-step version of MD4, which took Minisat SAT solver [16] about 8 hours.

Fig. 4. Differential path for finding single-block collisions for MD4. Only values of differences for chaining variables that are different from that in the path from [38] are presented.

| Step 13 | 0x62c00000 |
| Step 17 | 0xb6000000 |
| Step 20 | 0xe0000000 |
| Step 21 | 0xf0000000 |
In order to develop the results of [15] and [14] we built a special technique that makes it possible to add to a propositional encoding various additional constraints (such as the conditions of ‘Dobbertin’ type), vary it and find the best ones in the sense of the effectiveness of inversion problem solving. The main idea of this approach consists in the following. Let $C$ be a CNF that encodes the inversion of some function and $X$ be a set of Boolean variables from $C$. Assume that we need to add to $C$ new constraints that specify some predicate over variables from a set $\tilde{X}$, $\tilde{X} \subseteq X$. Let $R(\tilde{X})$ be a formula specifying this predicate. Now let us introduce new Boolean variable $u$, $u \notin X$. Consider formula $C' = C \land (\neg u \lor R(\tilde{X}))$. It is clear that the constraint $R(\tilde{X})$ will be inactive when $u = 0$. Let us call the variables similar to $u$ the switching variables. Varying values of switching variables it is possible to find a set of additional constraints that makes it possible to significantly decrease the time required to find the preimage of the considered hash image. To work with switching variables one can use various ‘learning’ techniques embedded into state-of-the-art SAT solvers. In particular it is possible to consider the values of switching variables as the so-called assumptions [17].

Hereinafter we considered the inversion of MD4 hash image 

$$0xffffffff, 0xffffffff, 0xffffffff, 0xffffffff.$$ 

At the initial stage of our experiments we implemented the search over combinations of Dobbertin-like conditions with $K = 0$. Thus we tried all possible combinations of values of switching variables for 28 steps of the first two rounds (2$^{28}$ variants in total).

The exhaustive search was performed using ‘Academician Matrosov’ computing cluster of Irkutsk Supercomputer Center SB RAS (in the experiment we employed 160 cores of Opteron 6276 processors). Each SAT instance produced by fixing active switching variables had a time limit of 0.005 seconds. In total the experiment took about 10 minutes. As a result there were synthesized 5 sets of values of switching variables for which the SAT solver managed to find satisfying assignment within the time limit. Less than 5 % of all SAT instances were interrupted due to the time limit, and more than 94% turned out to be UNSATISFIABLE. Among the 5 found sets of values one of the sets exactly coincided with Dobbertin conditions with $K = 0$. Thus, it can be said that we managed to synthesize the conditions presented in [15] automatically using the SAT solver. After this we considered the problem of inversion for 39-step version of MD4 using Transalg encodings and state-of-the-art multithreaded SAT solvers. The best results were shown by the Treengeling solver [5]: on average to solve one inversion problem for 39-step version of MD4 it took less than 10 minutes of work using one node of ‘Academician Matrosov’ cluster (32 cores). It is quite surprising that we did not manage to solve inversion problem for 40-step version of MD4 in reasonable time. We believe that we will manage to do it in the nearest future thanks to generation of new ‘Dobbertin-like’ conditions. For this purpose we will employ the SAT-based cryptanalysis techniques presented in this paper.
5 Related Works

The idea of applying SAT solving algorithms to cryptanalysis problems was first suggested, apparently, in [9]. It is usually assumed that the first practical implementation of this idea was performed in [26]. From that moment on the flow of research in the direction of SAT-based cryptanalysis has been constantly increasing. The first example of successful application of SAT-based cryptanalysis to the cryptographic functions used in the real world can be found in [28]. In a number of later works SAT-based cryptanalysis has also been applied quite successfully. Thus, in [34] the problem of cryptanalysis of the unweakened A5/1 algorithm used to encrypt GSM-traffic was solved using the SAT solvers. In [35] there were given the complexity estimations of SAT-based cryptanalysis of the Bivium cipher. These estimations were better than similar ones for other known methods of cryptanalysis of this cipher. In a recent paper [33] estimations from [35] have been improved.

For the first time the method that used differential paths to construct collisions of cryptographic hash functions was presented in the papers by X. Wang et.al. [38, 39]. Later there appeared a lot of different works following [38, 39]. Probably one of the most interesting of them is the differential attack by M. Stevens [36] that makes it possible to construct single block collisions for the MD5 hash function.

As we already mentioned, for the first time the problem of finding collisions for cryptographic hash functions was considered as SAT in [22]. However, real progress in this direction has been achieved a year later in [28]. The decisive factor in this direction was the use of SAT encodings of differential paths given in [38, 39].

In the context of finding collisions of hash functions from the MD family the main result of the present paper consists in using SAT to find a previously unknown class of two-block collisions for MD5 where first 10 bytes are zeroes. Also we described the SAT-based method for generating new differential paths based on the known ones. Using one of the paths found by this method, we managed to enumerate MD4 collisions faster than we could do it using the Wang path.

The nontrivial attack on incomplete version of MD4 was first proposed by H. Dobbertin in [15], where it was shown that the first two rounds of MD4 are not one-way. In our paper we automatically synthesized SAT equivalents of Dobbertin constraints by showing that in some sense these constraints are the best possible ones. For this purpose we used a special parallel SAT solver.

6 Conclusions

In this paper we presented the results of application of SAT-based cryptanalysis to inversion problems of the cryptographic functions MD4 and MD5. With regards to finding collisions (single-block for MD4 and two-block for MD5) the effectiveness of proposed methods exceeds that of previous papers. We showed
that SAT approach is applicable to the construction of differential paths for finding collisions of cryptographic hash functions: we managed to construct a new differential path for finding MD4 collisions. Also we applied SAT approach to construct additional constraints that make it possible to improve the effectiveness of finding preimages of truncated variants of the MD4 hash function. Among the automatically synthesized constraints there were constructed the constraints used in [15, 14]. As a concluding remark we would like to note, that SAT-based cryptanalysis is a rapidly developing area. SAT solvers can be used not only for cryptanalysis, but also as oracles for solving various auxiliary problems, for example, to solve nonlinear equations over finite fields in algebraic cryptanalysis [2]. The evident progress of computing architectures, and especially intensive development of parallel computing technologies gives us hope that new results in SAT-based cryptanalysis, including the problems related to cryptographic hash functions, would not take long to appear.

Acknowledgments. The research was supported by Russian Science Foundation (grant №16-11-10046).

References

Abstract. Data are a valuable resource that keeps a great potential for recovery of the useful analytical information. One of the most promising toolkits to solve the problems of data mining could be the usage of neural network technology. The problem of initial parameters values and ways of neural network construction on example of a multilayer perceptron are considered. Also information about the task and available raw data are taken into account. The modular BP-SOM network, which combines the multi-layered feed-forward network with the Back-Propagation (BP) learning algorithm and Kohonens self-organising maps (SOM), is suggested for visualization of the internal information representation and the resulting architecture assessment. The features of BP-SOM functioning, methods of rule extraction from trained neural networks and the ways of the result interpretation are presented.

Keywords: neural network, multilayer feedforward network, Kohonen self-organizing maps, modular network BP-SOM, rules extraction.

1 Introduction

Data analysis processes are often related to the tasks which are characterized by the lack of information about the sample structure, dependencies and distributions of analyzed indicators. One of the closest correspondences to this condition could be the usage of an approach based on neural network technology. The ability of neural networks to train, simulate nonlinear processes, deal with noisy data, extract and generalize the essential features from the incoming information makes them one of the most promising toolkits in solving data mining problems. However, there are several difficulties in using of this approach. In particular there is the problem of choosing an optimal network topology, parameter values and structural features that would best meet the problem being solved on available raw data. This is due to the fact that different neural networks can show very similar results on the samples from training set and significantly different when working with new, never shown data. Designing the optimal topology of the neural network can be represented as a search of architecture that provides the best (relative to the chosen criterion) solution of a particular problem. Usually the particular architecture and structural features could be selected on the
results of the assessment based on knowledge of the problem and the available source data. After that, the training and testing processes are taking place. Their results are used in the decision-making process, that the network meets all the requirements.

Another complication in using neural network approach could be related with the results interpretation and its preconditions. Especially clearly, this problem appears for the multilayer perceptron (MLP) [1]. In fact, neural network acts as a "black box", where the source data are sent to the input neurons and the result is got from the output, but the explanation about the reasons of such a solution is not provided. The rules are contained in the weight coefficients, activation functions and connections between neurons, but usually their structure is too complex for understanding. Moreover, in the multilayer network, these parameters may represent non-linear, non-monotonic relationship between input and target values. So, generally, it is not possible to distinguish the influence of a certain characteristic to the target value, because the effect is mediated by the values of other parameters. Also, you may experience some difficulties in using the learning algorithm of back-propagation BP, both with local extremes of the error function, and with the solutions of a certain class of problems.

2 The architecture and initial values choice for the neural network

The choice in favor of neural network architecture can be based on knowledge of the problem being solved and the available source data, their dimension and the samples scope. There are different approaches for choosing the initial values of the neural network characteristics. For example, the "Network Advisor" of the ST Neural Networks package offers by default one intermediate layer with the number of elements equals to the half of the sum of the quantity of inputs and outputs for the multilayer perceptron. In general, the problem of choosing the number of hidden elements for the multilayer perceptron should account two opposite properties, on the one hand, the number of elements should be adequate for the task, and on the other, should not be too large to provide the necessary generalization capability and avoid overfitting. In addition, the number of hidden units is dependent on the complexity of the function, which should be reproduced by the neural network, but this function is not known in advance. It should be noted that while the number of elements increases, the required number of observations also increases. As an estimate, it is possible to use the principle of joint optimization of the empirical error and the complexity of the model, which takes the following form [1]:

\[
\min \{\text{description of the error} + \text{description of the model}\}. \tag{1}
\]

The first part of this expression is responsible for the accuracy of the approximation and the observed learning error. The less it is - the less bits it is needed to correct the model predictions. If the model predicts the data accurately, then the
length of the error description equals to zero. The second part makes sense to the amount of information needed to select a specific model from the set of all possible. Its accounting allows to apply the necessary constraints on the complexity of the model by suppressing an excessive amount of tuning parameters.

The accuracy of the neural network function approximation increases with the number of neurons in the hidden layer.

When there are \( h \) neurons the error could be estimated as \( O\left(\frac{1}{h}\right) \). Since the number of outputs in the network does not exceed, and typically much smaller than the number of inputs, so the main number of weights in the two-layer network would be concentrated in the first layer, i.e. \( w = i \cdot h \), where \( i \) - is the input dimension. In this case, the average approximation error is expressed by the total number of weights in the network as follows: \( O\left(\frac{1}{w}\right) \).

The network description is associated with the models complexity and is basically comes down to the consideration of the amount of information in the transmission of its weights values through some communication channel. If we accept the hypothesis \( \psi \) about the network settings, its weights and the number of neurons, the amount of information (in the absence of noise) while transferring the weights will be \( -\log (Proab) \), where \( Proab \) is the probability of this event before the message arrives at the receiver input. For a given accuracy this description requires about \( -\log P(\psi) \sim w \) bit. Therefore, a specific error for one pattern associated with the complexity of the model could be estimated as: \( \sim \frac{w}{p} \), where \( p \) is the number of patterns in the training set. The error decreases monotonically with increasing the number of patterns. So Haykin, using the results from the work of Baum and Hessler, gives recommendations about the volume of a training sample relative to the number of weighting coefficients and taking into account the proportion of errors allowed during the test, which can be expressed by the inequality: \( p \geq \frac{w}{\varepsilon} \), where \( \varepsilon \) is the proportion of errors which allowed during testing. Thus, when 10% of errors are acceptable then the number of training patterns must be 10 times greater than the number of available weighting coefficients in the network.

Thus, both components of the network generalization error from expression (1) were considered. It is important that these components are differently depend on the network size (number of weights), which implies the possibility of choosing the optimal size that minimizes the total error:

\[
Expression \ (1) \sim \frac{i}{w} + \frac{w}{p} \geq 2 \cdot \sqrt{\frac{i}{p}}, \tag{2}
\]

Minimum error (equal sign) is achieved with the optimal number of the weights in the network: \( w \sim \sqrt{p \cdot i} \) which corresponds to the number of neurons in the hidden layer:

\[
h = \frac{w}{i} \sim \sqrt{\frac{p}{i}}, \tag{3}
\]

where \( i, h \) are the quantity of neurons in input and hidden layers, \( w \) is the number of weights and \( p \) is the amount of patterns in the training sample.
3 Evaluation of the resulting architecture and parameters

After the network architecture was selected the learning process is carried out. In particular, for a multi-layer perceptron this may be the error back-propagation (BP) algorithm. One of the most serious problems is that the network is trained to minimize the error on the training set, rather than an error that can be expected from the network while it will process completely new patterns. Thus, the training error will differ from the generalization error for the previously unknown model of the phenomenon in the absence of the ideal and the infinitely large training sample [2].

Since the generalization error is defined for data, which are not included in the training set, the solution may consist of separation of all the available data into two sets: a training set, which is used to match the specific values to the weights, and validation set, which evaluates the predictive ability of the network and selects the models optimal complexity. The training process is commonly stops with consideration of ”learning curves” which track dependencies of learning and generalization errors according to the neural network size [3, 4]. The optimum matches to local minima and points, where the graphs meet asymptote. Figure 1 shows the stop point, which corresponds to the validation error minimum (dash-dotted line), while the training error (solid line) keeps going down.

![Fig. 1. The stop training criterion at time t*.](image)

Another class of learning curves uses the dependencies of the neural network internal properties to its size, and then mapped to an error of generalization. For example, in [3] the analysis of the internal representation of the problem being solved, relationship of the training error and the maximum sum of the synapse weights modules attributable to the neuron of network are carried out. Also there are variants of generalized curves, which are based on dependence of the wave criterion from the neural network size [5] or perform a comparison of the average module values of synapse weights [6].

In simplified form, the following criteria could be formulated to assess the already constructed neural network model:
If the training error is small, and testing error is large, it means that the network includes too much synapses; if the training error and testing error is large, then the number of synapses is too small; if all the synapse weights are too large, it means that there are too few synapses.

After evaluation of the neural network model, the decision-making process is taking place about the necessity of changing the number of hidden elements in one or another direction, and the learning process should be repeated. It is worth to mention that modified decision trees, which are based on the first-order predicate logic, could be applied as a means of decision making support and neural network structure construction automation.

4 Rules extraction and results interpretation

Generally speaking, there are two approaches to extract rules from the multilayer neural networks. The first approach is based on extraction of global rules that characterize the output classes directly through the input parameter values. An alternative is in extraction of local rules, separating the multilayer net on a set of single-layer networks. Each extracted local rule characterizes a separate hidden or output neuron based on weighted connections with other elements. Then rules are combined into a set, which determines the behavior of the whole network.

The NeuroRule algorithm is applicable for rules extraction from the trained multilayer neural networks, such as perceptron. This algorithm performs the network pruning and identifies the most important features. However, it sets quite strict limitations on the architecture, the number of elements, connections and type of activation functions. As an alternative approach may be highlighted TREPAN type algorithms which extract structured knowledge not only of the extremely simplified neural networks, but also arbitrary classifiers in the form of a decision tree [7]. However, this approach does not take into account structural features that can introduce additional information.

The decision of such kind of problems could be based on the usage of the modular neural network BP-SOM[8].

5 Modular neural network BP-SOM

5.1 Network architecture

The main idea is to increase the reactions similarity of the hidden elements while processing the patterns from the sample, which belong to the same class. The traditional architecture of the direct distribution network [1], in particular the multi-layer perceptron with the back-propogation learning algorithm (BP), combines with the Kohonen self-organizing maps (SOM) [9], where each hidden layer of the perceptron network is associated with a certain self-organizing map. The structure of such a network is shown in Figure 2.
5.2 Learning algorithm

BP-SOM learning algorithm largely corresponds a combination of algorithms which are specific to the learning rules of its component parts [8]. First, the initial vector from the training sample is fed to the input of the network and its direct passage is carried out. At the same time, the result of neurons activation in each hidden layer is used as a vector of input values for the corresponding SOM network. Training of SOM components is carried out in the usual way and ensures their self-organization. In further, this self-organization is used to account classes, which tags are assigned to each element of the Kohonen maps. For this purpose, a counting is taking place, which purpose is to get the number of times the SOM-neuron became the winner and determine what class the initial vector of training sample belongs to. The winner is chosen from the SOM-neuron, whose weights vector is the closest in terms of the Euclidean distance measure to the output values vector of the hidden layer neurons. The most common class is taken as the mark. Reliability is calculated from the ratio of the number of class mark occurrences to the total number of victories of the neuron, i.e. for example, if the SOM-neuron became 4 times winner of class A and 2 times for class B, class A label with certainty 4/6 is selected. The total accuracy of the self-organizing map is equal to the average reliability of all elements of the card. Also, SOM allows data visualizing and displays areas for the various classes (Fig. 2). Learning rule of the multilayer perceptron component part is carried out by the similar to BackPropagation (BP) algorithm, minimizing aggregate square error:

\[ BP_{Error} = \sum_i (d_i - y_i)^2 , \]

where index \( i \) runs through all outputs of the multi-layer network, \( d_i \) is the desired output of the neuron \( i \), \( y_i \) is the current output of the neuron \( i \) from
the last layer. This error is transferred over the network in the opposite direction from the output to the hidden layers. Also, an additional error component $\text{SOMError}$ for neurons from hidden layers is introduced, which is based on information about the particular class of input vector and taking into account the self-organizing map data. Thus, in the SOM, which corresponds to the current hidden layer, the searches for an special element $V_{\text{SOM}}$ is taking place. This element should be closest, in terms of Euclidean distance, to the output vector of the hidden layer $V_{\text{hidden}}$, and be the same class label as the input vector. The distance between the detected vector $V_{\text{SOM}}$ and the vector $V_{\text{hidden}}$ is taken as the error value $\text{SOMError}$ and accounted for all of the hidden layer neurons. If the item $V_{\text{SOM}}$ is not found, then the error value $\text{SOMError}$ is assumed to be 0. Thus, the total error for the neurons of the hidden layer takes the form:

$$BP_{\text{SOM}}\text{Error} = (1 - \alpha) \cdot BP\text{Error} + r \cdot \alpha \cdot \text{SOMError},$$

(5)

where $BP\text{Error}$ is the error of perceptron (from BackPropagation algorithm); $\text{SOMError}$ is the error of the winner neuron from Kohonen network; $r$ is the reliability factor of the winner neuron from Kohonen network; $\alpha$ is the influence coefficient of the Kohonen network errors (if the value is equal to 0, then it will become original BP).

The results of the Kohonen maps self-organization are used for changing of the weight coefficients in the process of network training. This provides an effect, in which the activation of neurons in the hidden layer will become more similar to all other cases processing vectors of the same class [10]. The SOM with the dimension 7 per 7 is shown in Figure 3.

![Fig. 3. Coloring of the hidden layer for the Kohonen self-organizing maps, depending on the learning algorithm.](image)

It characterizes the reaction of the hidden layer neurons of the BP-SOM network, which was trained to solve the problem of classifying for two classes. Map on the left corresponds to the base algorithm of back-propagation (BP), the right - with the influence of the Kohonen map, i.e. BP-SOM training. Here white cells correspond to class A, and black cells to class B. In turn, the size of the shaded region of the cell determines the accuracy of the result. So, completely white or completely black cell is characterized by the accuracy of 100% .
This could ensure structuring and visualization of information extracted from data, improve the perception of the under study phenomenon and help in the network architecture selection process. For example, if it is impossible to isolate the areas at the SOM for the individual classes, then there are not enough neurons and their number should be increased. Moreover, this approach can simplify the rules extraction from already trained neural network and provide the result in a hierarchical structure of the consistent rules such as "if-then".

5.3 Rules extraction

As an example, let’s consider a small test BP-SOM network that will be trained to solve the classification problem which is defined by the next logical function [11]:

\[ F(x_0, x_1, x_2) = (x_0 \land \bar{x}_1 \land \bar{x}_2) \lor (\bar{x}_0 \land x_1 \land \bar{x}_2) \lor (\bar{x}_0 \land \bar{x}_1 \land x_2) \]

This function is set to True (Class 1) only in case where one of the arguments is True, otherwise the function value is False (Class 0). Two-layer neural network could be used for implementation, which consists of three input elements, three neurons in the hidden layer, and two neurons in the resulting output layer. The dimension for the Kohonen map for the intermediate layer is 3x3 (Figure 4).

![Kohonen map, class labels and reliability.](image)

Four elements of the Kohonen map acquired class labels with certainty 1 and 5 elements were left without a label, and their reliability is equal to 0 after training (Figure 4).

One of the methods for rules extraction from such a neural network could be the algorithm, which was designed for classification problems with digital inputs. It consists of the following two steps [10,11]:

1. Searching for such groups of patterns in the training set, which are potentially could be combined into individual subsets, each of which is connected to one element of the Kohonen map (Table 1).
2. Then, each of the subsets is examined to identify the values of the inputs that have constant value in the subset. For example, in the subgroup associated with the element \( k_1 \) all the attributes \( x_0, x_1 \) and \( x_2 \) have a constant value 0, and for the element \( k_3 \) value 1 respectively.
Table 1. Splitting the training set into groups with the reference to a specific element of the SOM.

<table>
<thead>
<tr>
<th>SOM element</th>
<th>Class</th>
<th>$x_0$</th>
<th>$x_1$</th>
<th>$x_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$k_3$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$k_7$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$k_9$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Thus, it is possible to get the following two rules from the Table 1:

- $IF(x_0 = 0 \land x_1 = 0 \land x_2 = 0) THEN (Class = 0);$  
- $IF(x_0 = 1 \land x_1 = 1 \land x_2 = 1) THEN (Class = 0);$  

However, it is rather problematic to use this way to distinguish rules for elements $k_7$ and $k_9$. Of course, it is possible to compose a disjunction of all the possible options for each SOM-element, but these rules would be complicated in perception.

Additional available information consists of the attributes sum for each patterns from the training sample. It is easy to notice that each SOM-element is responsible for a certain obtained sum value (Table 2).

Table 2. The sum of the attribute values for each patterns from the training sample concerning the SOM-elements.

<table>
<thead>
<tr>
<th></th>
<th>$k_1$</th>
<th>$k_3$</th>
<th>$k_7$</th>
<th>$k_9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Class</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

These considerations could be summarized. We need to find constraints on the attribute values of the input vectors and their weight coefficients for extraction rules, which corresponds to Kohonen map elements. This may be done by back-propagation of minimum and maximum values of the neuron activation back to the previous layer, i.e. we have to apply the function $f^{-1}(V^C_{i,\text{Cur}})$ (inverse to activation function) to the output neuron value \cite{[12]}.

$$f^{-1}(V^C_{i,\text{Cur}}) = f^{-1}(f(\sum_j w_{ji}V^P_{j,\text{Prev}} + \text{bias}_i)) = \sum_j w_{ji}V^P_{j,\text{Prev}} + \text{bias}_i,$$

where $V^C_{i,\text{Cur}}$ is the $i$-th neuron output from the current layer, $V^P_{j,\text{Prev}}$ is the output of the $j$-th neuron of the previous layer. Assuming that sigmoid was used...
as the neurons activation function, then in this case we will get:

\[ f^{-1}(V_{i\text{Cur}}) = -\ln\left(\frac{1}{V_{i\text{Cur}}} - 1\right). \]  

(7)

Additionally, it is known that self-organizing map elements, which are connected to the elements of the first hidden layer of the perceptron, will respond to proximity of the weight vectors and outputs of the hidden layer neurons. Therefore, it is proposed to replace the back-propagation neural activation of the hidden layer to the weight vector values of the self-organization map element during the rules construction for each SOM-element. For example, if the hidden layer contains neuron \( A \), and the weight between this neuron and SOM-element \( k \) was denoted by \( w^k_A \), then the restriction takes the form:

\[ f^{-1}(w^k_A) - \text{bias}_A \approx \sum_j w_{jA} x_j \text{ or } 1 \approx \left( \sum_j w_{jA} x_j \right) / (f^{-1}(w^k_A) - \text{bias}_A). \]  

(8)

Such restrictions, which were obtained for all the neurons of the hidden layer, could be used for construction of following rules:

\[ IF \left( \land_i(f^{-1}(w^k_A) - \text{bias}_A \approx \sum_j w_{jA} x_j) \right) THEN \ (Class = Class(k)). \]  

(9)

Similar rules are required for all SOM-components, the reliability of which exceeds a certain threshold.

If we apply the considered method for the initial example, then four sets of restrictions would be obtained. Each set includes three restrictions, which corresponds to the number of neurons in the hidden layer of the perceptron.

For SOM-element \( k_1 \):

- \( 1 \approx 687 \times x_0 + 687 \times x_1 + 687 \times x_2; \)
- \( 1 \approx 738 \times x_0 + 738 \times x_1 + 738 \times x_2; \)
- \( 1 \approx 1062 \times x_0 + 1062 \times x_1 + 1062 \times x_2. \)

Taking into account that \( x_0, x_1, x_2 \in \{0, 1\} \) then the best results may be achieved if \( x_0 = 0 \), \( x_1 = 0 \) and \( x_2 = 0 \).

For element \( k_3 \) all restrictions coincide and have the form:

- \( 1 \approx 0.33 \times x_0 + 0.33 \times x_1 + 0.33 \times x_2; \)

Thus, the values are as follows: \( x_0 = 1 \), \( x_1 = 1 \), \( x_2 = 1 \).

For element \( k_7 \) restrictions coincide and take the form:

- \( 1 \approx x_0 + x_1 + x_2; \)

This corresponds to the case when only one of the attributes is equal to 1.

Restrictions for element \( k_9 \):

- \( 1 \approx 0.5 \times x_0 + 0.5 \times x_1 + 0.5 \times x_2; \)
This condition characterizes the case when two of three attributes are 1.

Thus, it is obtained the following set of rules, if all restrictions would be generalized:

- \( IF \ (x_0 + x_1 + x_2 \approx 0) \ THEN \ (Class = 0) \)
- \( IF \ (x_0 + x_1 + x_2 \approx 3) \ THEN \ (Class = 0) \)
- \( IF \ (x_0 + x_1 + x_2 \approx 1) \ THEN \ (Class = 1) \)
- \( IF \ (x_0 + x_1 + x_2 \approx 2) \ THEN \ (Class = 0) \)

It is easy to notice that this set correctly describes all the elements of a self-organizing network.

6 Conclusion

Approaches for the selection of the initial values of the neural network parameters were considered. The analysis of the training process, its stopping criteria and evaluation of the received architecture were carried out. Combining different neural network architectures, such as multi-layer perceptron with the back-propagation learning algorithm, and Kohonen’s self-organizing maps, could bring additional possibilities in the learning process and in the rules extraction from trained neural network. Self-organizing maps are used both for information visualization, and for influencing the weights changes during the network training process. That provides an effect, in which the neurons activation in the hidden layer will become increasingly similar to all the cases processing vectors of the same class. This ensures the extracted information structuring and has the main purpose to improve the perception of the studied phenomena, assist in the process of selecting the network architecture and simplify the extraction rules. The results could be used for data processing and hidden patterns identification in the information storage, which could become the basis for prognostic and design solutions.

Acknowledgements. Work is carried out with the financial support of the RFBR (grant 15-07-01117-a).

References

Synthetic Data for AUV Technical Vision Systems Testing

Aleksandr Kamaev¹, Sergey Smagin¹, Viacheslav Sukhenko², and Dmitry Karmanov¹

¹Computing Center of FEB RAS, Khabarovsk, Russia
²Pacific National University, Khabarovsk, Russia
kamaev_an@mail.ru

Abstract. Development of the autonomous underwater vehicle technical vision systems is impossible without precise debugging and testing. Due to the factors described in the paper, such testing in most cases cannot be implemented with the help of real AUV. Therefore, the usage of the procedurally generated virtual testing areas is suggested. The algorithm for generation and visualization of the seabed 3D model that is suitable for AUV technical vision systems testing and debugging is described. This algorithm allows building of high detailed surface of the seabed, where each part is absolutely unique and does not contain repeating texture patterns. Also, software system “AUV Vision Debugger” that consists of seabed generator and AUV simulator is considered. The simulator provides interaction between generated seabed, AUV model, and technical vision system that is currently being tested.

Keywords: AUV, computer vision, procedural, texturing, height map, tessellation, fractal noise, seabed, simulator.

Introduction

Complexity of the underwater navigation and inability to organize communication with high capacity data exchange between autonomous underwater vehicle (AUV) and operator leads to necessity of onboard technical vision systems development to improve underwater vehicle autonomy. High reliability and quality operation in various conditions is required from such systems. Their development and further usage is impossible without precise debugging and testing. Currently testing process organization and testing data obtaining becomes a serious problem.

At present time, testing areas with markers and targets located on seabed and specifically equipped pools are used for tasks of testing and debugging. Such methods are well suited for testing AUV devices and equipment, but they are not applicable for AUV technical vision systems testing and debugging because of following reasons.

1. Obtaining data from AUV is a time consuming and expensive process. It includes not only mission accomplishment itself, but also transportation to
testing area and device launching that is unacceptable especially at the early stages of development.

2. It is not always possible to obtain necessary data for testing from areas with different types of relief due to geographic reasons.

3. It is impossible to evaluate vision system accuracy while using data from real testing areas, because there are no other ways to measure seabed with the required accuracy level (up to cm) on large areas.

4. It is impossible to interfere in the vision system work at any required time.

5. There is a risk to lose the AUV due to technical vision system errors.

To avoid all mentioned problems of real data usage in AUV technical vision system testing and debugging processes, it is suggested to replace real experiments with tests in virtual environment. The software system “AUV Vision Debugger” that allows holding such tests is considered in the paper.

1 Related Work

Usage of the virtual simulations for tasks related to AUV is an actively developing research area. Currently a lot of modeling software for different purposes is developed. The workbench [1] supports 3D visualization and can be applied for missions planning. The simulator [2] is intended for AUV control systems developing. The modeling software system [3] is designed to be used for AUV operators learning. The system [4, 5] allows debugging of different systems and devices of AUV in virtual environment. Since AUV technical vision systems are using feature points and their descriptors [6–8], such requirements as high levels of detail (up to one pixel), absolute uniqueness of all relief parts and absence of repeated texture patterns are imposed to seabed model. These requirements make it impossible to use existing modeling software for AUV vision systems testing and debugging.

Generation and visualization of the seabed 3D model, that has good enough quality for AUV vision system testing and debugging, is the most complex task. Despite large number of existing methods for landscape generation [9], they cannot be directly applied for procedural seabed model synthesis. Methods based on fractal brushes [10] allow obtaining a landscape of needed form, but they require long time of artist work. Approaches based on multifractals [11, 12] do not provide enough control over the result that is necessary for generating the seabed of required form. Methods based on modeling of the physical process of landscapes formation [13, 14], do not take in account the specifics of the underwater relief formation processes. The methods of fractal interpolation [15] require basic constraints, but they do not solve problem of obtaining such constraints. Creation of the seabed 3D model that will be suitable for AUV technical vision system testing and debugging requires complex approaches and combinations of different generation methods.
2 Program System “AUV Vision Debugger”

Structural scheme of the developed program system for AUV technical vision algorithms testing and debugging – “AUV Vision Debugger” is shown in Fig. 1. This system consists of two programs: seabed 3D model generator and AUV simulator. The generator is a GUI program that allows building a seabed 3D model from the set of user defined parameters using approaches described in section 3. The simulator is a program that simulates AUV dynamic in virtual environment and allows user to observe all AUV movements by controlling the view camera. The physical model used by simulator is described in section 4.

![Fig. 1. Structural scheme of the program system “AUV Vision Debugger”](image)

One of the simulator’s capabilities is a selection of the controlling program – technical vision system combined with control system. The procedure of vision system testing and debugging may be different for different development stages and tasks that system must solve. Therefore, specific control system is required to implement testing of concrete vision system task. “AUV Vision debugger” does not impose any limitations on organization of technical vision and control systems and their interaction, and provides universal interface with simulator.

The simulator runs the controlling program and scans its standard output stream with approximately 50 Hz frequency to detect passed commands. As an answer to the commands the simulator writes messages to the standard input stream of the vision system and saves graphical information in the file system if required.

Commands supported by simulator can be divided into three groups: control commands, state requests and image requests. Control commands serve to
change such states of AUV parts (see section 4.1 for parts description) as position, orientation, power and etc. State requests can be used to know information about AUV position, orientation and velocity, distances to objects (if AUV equipped with sonar) and etc. As an answer to state requests, the simulator passes state information to the standard input stream of AUV technical vision system. Image requests are necessary for obtaining images with debug information from AUV onboard cameras. Images and debug information are not passed directly through standard input stream, but stored in file system. Instead, a notification that images are ready to use is passed to the input stream. The debug information contains accurate camera external and internal parameters and distances from camera to image pixels. Using this information, the controlling program can evaluate results produced by the technical vision system and make conclusions about the system’s accuracy.

Thus, “AUV Vision Debugger” provides just a physical and visual model of environment and AUV. Development of the control system for testing AUV technical vision algorithms is the user’s responsibility.

3 Generation and Visualization of the Seabed 3D Model

The algorithm that was developed for “AUV Vision Debugger” to generate and visualize the seabed 3D model has five stages: generation of low frequency relief map (section 3.1), fractal noise computation to increase level of details (section 3.2), 3D model mesh building (section 3.3), model refinement during visualization (section 3.4) and texturing (section 3.5).

3.1 Relief Map

Relief map $H = (H_{ij}), i = 1, 2, ..., h, j = 1, 2, ..., w$, where $w$ and $h$ - map’s size, defines relief height $H_{ij}$ in point $(j \cdot res, i \cdot res)$, where $res$ is map resolution specified in meters per point (m/p). The maps with $res = 10$ m/p and $w, h \leq 1000$ were used in the “AUV Vision Debugger”. Such maps can describe the seabed with length up to 10 km that is more than enough for AUV technical vision algorithms testing and debugging.

The relief map contains such basic seabed elements as coast, shelf, continental slope, ocean bed, submarine canyons and mountains (Fig. 2a). These elements are defined by user through setting a few parameters’ values, and they do not require specific artist knowledge.

Contours of the basic elements are specified with fractal lines and heights are interpolated, using different interpolation functions, based on distances to defined contour lines (Fig. 2b).

3.2 Fractal Noise

The relief details that have frequency more than $1/(2res)$ m$^{-1}$ cannot be represented by $H$ map. Meanwhile, to provide correct work of technical vision algorithms on distances from seabed that could be reached with onboard light
equipment, accuracy up to millimeters is required. To improve relief level of details we will use fractal noise that consists of summed Perlin noise functions, taken at following frequencies:

\[ f_l = (f_{li}) = \frac{2^{i-1}}{res}, \quad f_h = (f_{hj}) = \frac{2^{j+n_l}}{res} \]

where \( i = 0, 1, \ldots, n_l, j = 0, 1, \ldots, n_h, n_l = \lfloor \log_2(2 \cdot res) \rfloor, \]
\[ n_h = \lfloor \log_2(p_w^{-1}2^{-n_l res}) \rfloor, \]
\( p_w \) — drawing fragment size (pixel of the screen or AUV camera) in meters. If fragment size could not be computed, the desired accuracy should be directly assigned to \( p_w \). Low frequencies \( f_l \) will be added to seabed 3D model and high frequencies \( f_h \) will be utilized during visualization process.

Amplitudes on the same frequencies may be different for different seabed parts, depending on their slope roughness and some random factor. Let \( S = (S_{ij}) \) be the slope map (Fig. 3a) and \( R = R_{ij} \) be the roughness map (Fig. 3b), where \( i = 1, 2, \ldots, h, j = 1, 2, \ldots, w, \)

\[ R_{ij} = \min\left(1, \frac{\sigma(H_{ij})}{\sigma_{max}}\right), \]
\[ S_{ij} = \frac{2\arccos(n_{ij,z})}{\pi}, \]

where \( \sigma(H_{ij}) \) is standard deviation of high frequency part of height map in 9 × 9 neighborhood (weighted with Gauss function with \( \sigma = 2 \)) of the point \( H_{ij} \), \( \sigma_{max} \) is a normalization factor and \( n_{ij} = (n_{ij,x}, n_{ij,y}, n_{ij,z}) \) is a normal to low frequency part of height map in point \( H_{ij} \).

We divide the frequencies range \( (f_l, f_h) \) into three groups each of which is controlled by own basis function \( \phi_i(f), i = 1, 2, 3, \) — Fig. 4a. The logarithmic scale is used on the charts; at this scale functions \( \phi_i(f) \) are piecewise-linear. As well we introduce functions of roughness and slope influences on the 1st and 2nd frequency groups:

\[ I_{R1}(r) = 0.45r^2 + 0.05, \quad I_{S1}(s) = \begin{cases} 
6s^2 - 8s^3, & s \leq 0.5 \\
(8s - 2)(s - 1)^2, & s > 0.5
\end{cases} \]
Fig. 3. a – slope map $S$, b – roughness map $R$.

$$I_{R2}(r) = \frac{r^2}{2}, \quad I_{S2}(s) = \frac{8}{9} \max(0, s - 0.25)^2.$$ 

Plots of these functions are presented at Fig. 4b, c. We define a united influence function for $i$-th frequency group that unites roughness, slope and random factors:

$$\psi_i(x) = \max(0, \min(1, I_{Ri}(R(x_1, x_2)) + I_{Si}(S(x_1, x_2)) + \lambda_i P_i(x))), \quad (1)$$

where $x = (x_1, x_2, x_3)$ is a space point, $i = 1, 2$, $R(x_1, x_2)$ and $S(x_1, x_2)$ are the values obtained from $R$ and $S$ by means of bilinear interpolation, $\lambda \in [0..1]$ is user defined coefficient of the random influence on $i$-th group,

$$P_i(x) = \sum_{j=0}^{4} \frac{1}{2^j} P\left(\frac{2^j}{\text{res}} (M_i x + T_i)\right),$$

where $i = 1, 2, 3$, $P(x)$ is a Perlin noise function [16], $M_i$ is a random rotation matrix, and $T_i$ is a random translation vector. The noise amplitude in $x$ point corresponding to frequency $f$

$$A(x, f) = \frac{\gamma}{f} \left(\sum_{i=1}^{2} \phi_i(f) \psi_i(x) + \phi_3(f) \left(\sin\left(\frac{2\pi P_3(x)}{2}\right) + 1\right)\right), \quad (2)$$

where $\gamma$ – coefficient defining general roughness of height differences on all frequencies. Recommended value lies in range $\gamma \in [0.5..1]$. Using amplitude (2) we define noise in $x$

$$N(x, t) = \sum_{i=0}^{n_t} A(x, f_{iti})P(f_{iti}x). \quad (3)$$

### 3.3 Building of 3D Model

For mesh construction we will use approach [17], but unlike [17], we will build mesh during generation step on CPU and not at rendering time on GPU. We
Fig. 4. a – functions $\phi_i(f)$, b – functions $I_{R1}(r), I_{R2}(r)$, c – functions $I_{S1}(s), I_{S2}(s)$.

define relief density in point $x = (x_1, x_2, x_3)$

$$\rho(x) = x_3 - H(x_1, x_2) + N(x, l), \quad (4)$$

where value $H(x_1, x_2)$ is obtained from $H$ by means of bilinear interpolation. Surface $\rho(x) = 0$ defines seabed. To obtain 3D model of seabed surface $\rho(x) = 0$ is approximated using marching cubes algorithm [18]. Figure 5 depicts result of such approximation.

Fig. 5. Seabed 3D model.
3.4 Visualization

It is necessary to tessellate mesh during visualization to provide size close to one pixel for all visible triangles. The most of modern GPUs supported hardware accelerated tessellation with subdivision by up to 64 parts for each triangle side. Such subdivision will not be enough for mesh triangles located near the camera. Therefore, tessellation based on precalculated triangles sets (PTS) is suggested for these triangles.

Each triangle side could be divided into $2^{n_i}$ parts, where $i = 1, 2, 3$ is an index number of triangle side (counterclockwise numeration), and $n_i = 0, 1, ..., n_{\text{max}}$. AUV operation distance to seabeed and onboard cameras resolution allows to choose $n_{\text{max}} = 9$ that leads to $N = (n_{\text{max}} + 1)^3 = 1000$ of different variants of tessellation. Let us consider $j$-th tessellation set, $j = 1, 2, ..., n$, that consists of $N_j$ triangles. Each triangle of $j$-th set is described by three vertices $v_{jkt} = (v_{jkt1}, v_{jkt2})$, $k = 1, 2, ..., N_j$, $t = 1, 2, 3$ (counterclockwise numeration). Values $v_{jkt1}$ and $v_{jkt2}$ are coordinates of $t$-th vertex of $k$-th triangle of $j$-th set, they defined in the coordinate system with basis vectors represented by 1st and 3rd sides of $k$-th triangle and origin in the intersection point of these sides – Fig. 6a.

![Fig. 6. a – PTS-based tessellation, b – invisible blocks clipping.](image)

Using such coordinate system we can draw $j$-th tessellation set instead of some mesh triangle that have coordinates $p_1, p_2, p_3$, by projecting local $j$-th set coordinates to the global space:

$$x_{jkt} = v_{jkt1}(p_2 - p_1) + v_{jkt2}(p_3 - p_1) + p_1. \quad (5)$$

The fact that tessellation sets store coordinates independent from mesh coordinates allows us to put these sets into video memory once, and then just use through one function call. The attempt to draw a triangle is replaced by suitable
tessellation set drawing under control of vertex shader that computes global vertices coordinates using (5). Coordinates $p_1$, $p_2$, $p_3$ are passed to vertex shader as uniform parameters. The number of tessellation set $j$ is defined based on triangle size and its remoteness from the camera to provide pixel accuracy in the screen space.

To speed up visualization process the seabed model is divided into rectangular blocks, so that blocks number and triangles number per block would allow iterating through them before each frame rendering. Blocks located behind camera’s clipping planes are discarded. For rendering of the triangles located in blocks that are close to the camera, the PTS-based tessellation is used. If the block is far enough for using hardware accelerated tessellation only, it renders by single drawing function call. The process of seabed subdivision is shown in Fig. 6b.

After tessellation the fractal noise (3) is added to each vertex in the direction of the interpolated normal $n$, computed in this vertex:

$$\tilde{x} = x + n (N(x, h) + \rho(x)),$$

where $\rho(x)$ is a density function (4). The example of tessellated and amplified with high frequency noise seabed model is presented in Fig. 7.

![Tessellated seabed 3D model.](image)

**Fig. 7.** Tessellated seabed 3D model.

### 3.5 Texturing

Texturing of landscapes, developing for testing AUV technical vision system has its own specific:
1. Only procedural textures could be used, because using of the bitmap images leads to the appearing of repeated texture patterns that makes impossible the testing of the algorithms based on feature points [6–8].

2. Complex relief makes impossible calculation of correct 2D texture coordinates, so just 3D textures could be used.

3. Textures should be correctly represented on all scale levels that are used during testing.

Surface type and, respectively, texture type is defined based on functions $\psi_1(x)$ and $\psi_2(x)$ (1). The sand procedural texture with weight $\alpha$ is mixed with the stone procedural texture with weight $1 - \alpha$, where

$$\alpha(x) = \prod_{i=1,2} (1 - \text{smoothstep}(\psi_{ic} - \delta, \psi_{ic} + \delta, \psi_1(x))).$$

Smoothstep is a standard GLSL function. Constants $\psi_{ic}$ define boundaries between relief types and $\delta$ is a transition width. These constants are set during seabed generation. The Fig. 9 is obtained with $\psi_{1c} = 0.2$, $\psi_{2c} = 0.12$ and $\delta = \max(0.0025, 0.05p_w)$, where $p_w$ – is a drawing fragment size in meters.

Creation of procedural texture for each type of surface requires individual approach. Sum of Perlin noise functions or cellular basis [11] is used to create different unique texture patterns at different scales. When the screen’s pixel size becomes too large to depict a pattern, this pattern is replaced by its average color and intensity. A single texture could have up to three different patterns at different scales. An example of sand and stone textures used in AUV Vision debugger is presented in Fig. 8.

![Fig. 8. Textures at different scale: sand(above) and stone covered with moss (below).](image-url)
4 AUV Simulator

The main tasks of the simulator are visualization of seabed and AUV models, simulation of AUV dynamics, and computation of interactions with the environment. Simulator also provides the possibility to observe the mission process. “AUV Vision Debugger” was created mainly for testing and debugging of technical vision system and not for testing AUV control systems and devices, therefore simplified AUV model (section 4.1), its dynamics (section 4.2) and interactions with seabed (section 4.3) are used.

4.1 AUV Model

AUV model consists of parts, described in text file of the model. Each part has its own type: shell, engine, control surface, sonar, floodlight and camera. In the model description there could be only one shell and arbitrary number of parts having other type. Let us consider main parameters of different types:

1. Shell is defined by its weight and 3D model in 3DS format. The coordinate system of the shell is a basis for all other parts.
2. Engine and control surface are defined by their weights, 3D model, transformation matrix, describing position and orientation relative to shell. Directions and ranges of engine and control surface allowable movements and rotations are also defined. Thrust vector and its magnitude range could be defined for engine.
3. Sonar, floodlight and camera do not have their own weight and graphic representation. They defined by position and orientation relative to shell. Directions and ranges of allowable movements and rotations are also defined. Light power could be set for floodlight and focus distance, resolution, radial distortion, and lighting dependent errors could be set for camera.

Apart from mentioned parameters, each part has a name that allows the vision system to interact with this part.

4.2 AUV Dynamics

From position of dynamics AUV with all its parts is considered as one rigid body. To calculate its motion Newton–Euler equations are used. The details of motion computation process based on forces and torques acting on the rigid body are well described in [19]. The weight of one part (if part has weight) is uniformly distributed between all the part’s points. The part’s points correspond to vertices of the part’s 3D model. All relative movement of AUV parts leads to redistribution of AUV weight and to recalculation of inertia tensor. Forces that lead to relative movements of parts, as well as forces induced by this movement, are not counted.

Simulator takes into account following forces: gravity force $F_g$, thrust force of $i$-th engine $F_{ti}$, $i = 1, 2, ..., n_e$, where $n_e$ is engines number, pressure force $F_p$
and hydrodynamic force \( F_h \). All forces except \( F_g \) induce torques relative AUV center of mass, that accounted in motion calculation. Let us consider how these forces act.

Gravity force is applied to AUV center of mass and pointed vertically down \( \mathbf{F}_g = (0, 0, -mg) \), where \( m \) is sum of weight of all parts, and \( g \) is acceleration of gravity. Thrust force \( \mathbf{F}_{ti} \) is applied to \( i \)-th engine center of mass and directed along thrust vector, described in AUV text file. If rotation is applied to \( i \)-th engine then its thrust vector is also undergo this rotation. Magnitude of \( \mathbf{F}_{ti} \) is chosen by vision system from the range defined in model text file.

Forces \( \mathbf{F}_p \) and \( \mathbf{F}_h \) are calculated for each face of AUV 3D model if this face has nonzero area \( S \), external unit normal vector \( \mathbf{n} \) and located out of the shell. Forces \( \mathbf{F}_p \) and \( \mathbf{F}_h \) are applied to geometric center of the face \( \mathbf{c} = (c_1, c_2, c_3) \):

\[
\mathbf{F}_p = -\rho g |c_3| S \mathbf{n},
\]

\[
\mathbf{F}_h = F_{hx} \frac{-\mathbf{v}}{|\mathbf{v}|} + F_{hy} \frac{\mathbf{n} \times \mathbf{v} \times \mathbf{v}}{|\mathbf{n} \times \mathbf{v} \times \mathbf{v}|},
\]

where \( \rho \) is a water density, \( \mathbf{v} \) is velocity of \( \mathbf{c} \) relative to environment, \( F_{hx} \) and \( F_{hy} \) are components of \( \mathbf{F}_h \) in direction of \( \mathbf{v} \) and normal to the \( \mathbf{v} \) vector direction:

\[
F_{hx} = |\mathbf{n} \cdot \mathbf{v}| \frac{\rho |\mathbf{v}|}{2},
\]

\[
F_{hy} = \frac{S \rho |\mathbf{v}|}{\sqrt{2}} \mathbf{n} \cdot \mathbf{v} \sqrt{1 - \left( \frac{\mathbf{n} \cdot \mathbf{v}}{|\mathbf{v}|} \right)^2}.
\]

Presented forces are sufficient to describe AUV dynamics accurate enough for technical vision systems testing and debugging.

### 4.3 Interactions with Seabed

Collision of AUV parts with seabed surface is a very dangerous situation that should be necessarily prevented by a technical vision system during visual navigation. Therefore, it is very important to detect such situation and to inform the vision system in the case it occurs. Informing is implemented by passing a message to the standard input stream of the vision system. We consider a process of detecting collision by the simulator and its reaction on this collision.

Collisions are tested on each simulation iteration for parts points. Let us consider collision detection process for point \( \mathbf{x} \). The collision is tested for all seabed model triangles that are closer to \( \mathbf{x} \) than one meter. If there are no such triangles then point \( \mathbf{x} \) is above the seabed. Let \( \mathbf{p}_i \) be the vertices of triangle that is closer to \( \mathbf{x} \) than one meter, \( i = 1, 2, 3 \), and \( \mathbf{n}_i \) be the normals to the model in these vertices. To detect a collision we should perform following steps:

**Step 1:** Computing face normal: \( \mathbf{n} = (\mathbf{p}_2 - \mathbf{p}_1) \times (\mathbf{p}_3 - \mathbf{p}_1) \).

**Step 2:** Computing intersection points between rays from \( \mathbf{p}_i \) in directions \( \mathbf{n}_i \) and plane \( (\mathbf{n}, -\mathbf{n} \cdot \mathbf{x}) \):

\[
\mathbf{p}'_i = \mathbf{p}_i + \frac{\mathbf{n} \cdot \mathbf{x} - \mathbf{n} \cdot \mathbf{p}_i}{\mathbf{n} \cdot \mathbf{n}_i} \mathbf{n}_i.
\]
Step 3: If \( x \) does not lie inside triangle \((p'_1, p'_2, p'_3)\), then no collision detected, else go to step 4.

Step 4: Finding of interpolated unit normal \( \tilde{n} \) in \( x \) point of triangle \((p'_1, p'_2, p'_3)\) with vertex normals \( n_i \), by means of bilinear interpolation.

Step 5: Computing point of triangle \((p_1, p_2, p_3)\), corresponding to \( x \):

\[
\tilde{x} = x + \frac{\bar{n} \cdot p_1 - \bar{n} \cdot x}{\bar{n} \cdot \bar{n}} \tilde{n}.
\]

Step 6: Computing of \( x \) penetration distance into the seabed surface, using (3) and (4):

\[
d = (\tilde{x} + \tilde{n}(N(\tilde{x}, h) + \rho(\tilde{x})) - x) \cdot \tilde{n}.
\]

If \( d > 0 \), then \( x \) point is located inside seabed, therefore collision happened. If value \( d > 0 \) computed for more than one triangle, maximum value should be chosen.

If collision of depth \( d \) for some point \( x \) with normal \( \tilde{n} \) is detected, then following force is applied to \( x \):

\[
F_c = (k_s d - k_d v \cdot \tilde{n})\tilde{n},
\]

where \( k_s \) is an elasticity coefficient, \( k_d \) is a damping coefficient, and \( v \) is a velocity of \( x \) point. Application of \( F_c \) force prevents AUV penetration into the seabed.

Conclusion

Methods and algorithms, suggested in the paper with developed program system “AUV Vision Debugger”, allow testing and debugging of AUV technical vision systems in virtual environment that leads to following advantages:

1. high speed and low cost of testing data acquisition,
2. ability to obtain testing data from different surface types, starting from sand valleys and ending with rocky canyons,
3. ability to evaluate technical vision accuracy since the investigated seabed surface is precisely known,
4. ability to interrupt system working exactly at the moment when error occurs,
5. testing result repeatability.

Currently AUV Vision Debugger allows getting high detailed images of seabed, all parts of which are absolutely unique – Fig. 9. For further researches it is planned to add a possibility of procedural modeling and visualization of marine flora and fauna, including dynamically changing objects, to add more procedural textures and submarine caves network. It is also planned to learn ways of integrating our work with existing general purpose software modeling systems.

It is evident that using of synthetic tests in “AUV Vision Debugger” does not allow us to completely abandon the real experiments, but it significantly reduces their amount. As a result, the time required for AUV technical vision system development is decreased and reliability is increased.
Acknowledgments. This work was supported by Russian Foundation for Basic Research (grant 16-31-00187 mol_a).

References


Methods and Tools of Parallel Programming

Victor Kasyanov and Elena Kasyanova
Institute of Informatics Systems, 630090 Novosibirsk, Russia,
Novosibirsk State University, 630090 Novosibirsk, Russia
kvn@iis.nsk.su

Abstract. Using traditional methods, it is very difficult to develop high quality, portable software for parallel computers. In particular, parallel software cannot be developed on low-cost, sequential computers and then moved to high-performance parallel computers without extensive rewriting and debugging. In this paper, the CSS system being under development at the Institute of Informatics Systems is considered. The CSS is aimed to be an interactive visual environment for supporting cloud parallel programming. The input language of the CSS system is a functional language Cloud Sisal that exposes implicit parallelism through data dependence and guarantees determinate result. The CSS system provides means to write and debug functional programs regardless target architectures on low-cost devices as well as to translate them into optimized parallel programs, appropriate to the target execution platforms, and then execute on high-performance parallel computers in clouds without extensive rewriting and debugging.

Keywords: cloud computing, computer science education, functional programming, hierarchical graph representation, parallel programming.

1 Introduction

Using traditional methods, it is very difficult to develop high-quality, portable software for parallel computers. In particular, parallel software for supporting of enterprise information systems cannot be developed on low-cost, sequential computers and then moved to high-performance parallel computers without extensive rewriting and debugging. Functional programming [1] is a programming paradigm, which is entirely different from the conventional model: a functional program can be recursively defined as a composition of functions where each function can itself be another composition of functions or a primitive operator (such as arithmetic operators, etc.). The programmer need not be concerned with explicit specification of parallel processes since independent functions are activated by the predecessor functions and the data dependencies of the program. This also means that control can be distributed. Further, no central memory system is inherent to the model since data is not “written” in by any instruction but is “passed from” one function to the next. However, scientific world is conservative and the FORTRAN programming language is still quite popular in scientific computations for supercomputers.
Functional language Sisal (Streams and Iterations in a Single Assignment Language) is considered as an alternative to FORTRAN language for supercomputers [2],[3]. Compared with imperative languages (like FORTRAN), functional languages, such as Sisal, simplifies programmer’s work. He has only to specify a result of calculations and it is a compiler that is responsible for mapping an algorithm to certain calculator architecture. In contrast with other functional languages (like Lisp, ML and Haskel), Sisal supports data types and operators typical for scientific calculations such as loops and arrays. At present, there are implementations of the Sisal 1.2 language [4] for many supercomputers (e.g., SGI, Sequent, Encore Multimax, Cray X-MP, Cray 2, etc). Sisal 90 [5] language definitions increase the language’s utility for scientific programming and include language level support for complex values, array and vector operations, higher order functions, rectangular arrays, and an explicit interface to other languages like FORTRAN and C. The Sisal 3.2 language [6],[7] integrates features of Sisal 2.0 [8] and Sisal 90 versions and includes language level support for module design, mixed language programming, and preprocessing.

In this paper, a visual cloud system CSS (Cloud Sisal System) of functional and parallel programming being under development at the Institute of Informatics Systems is considered. The input language of the system is a functional language Cloud Sisal [9] that is a modification of our Sisal 3.2 language which is aimed to increase the language’s utility for supporting of cloud scientific computations and cloud parallel programming. The Cloud Sisal language supports also so-called annotated programming and concretizing transformations [10], [11]. The system presented uses intermediate languages of hierarchical graphs and provides means to write and debug functional programs regardless target architectures on low-cost devices as well as to translate them into optimized parallel programs, appropriate to the target execution platforms, and then execute them on high performance parallel computers in clouds without extensive rewriting and debugging [12], [13]. So, the system can open the world of parallel and functional programming to all students and scientists without requiring a large investment in new, top-end computer systems.

2 The CSS System

The advancement of computer technology and the increasing complexity of research problems are creating the need to teach parallel programming in higher education more effectively. Programming massively-parallel machine is a daunting task for any human programmer and parallelization may even be impossible for any compiler. Instead, the functional programming paradigm may prove to be an ideal solution by providing an implicitly parallel interface to the programmer.

The CSS system is based on hierarchical graph representations of functional and parallel programs and is intended to provide a general-purpose user interface for a wide range of parallel processing platforms (See Fig. 1). In our conception, cloud interface gives transparent ability to execute programs on arbitrary environment. JavaScript client does not demand installation; small educational
programs can be executed on client devices (computers or smart phones). V8 server allows the language parser and some optimizations to be used at both client and server sides.

![Fig. 1. Cloud service structure: 1, 2 and 3 - clients, 4 - cloud access server, 5 - execution environment.]

The CSS system uses a functional language Cloud Sisal and includes five big parts: interface, interpreter, graphic visualization/debugging subsystem, optimizing cross compiler, cluster runtime.

The interpreter is available on web via browser; it translates Cloud-Sisal-program to the first internal representation (IR1) and runs it without making actual low-level code. It is useful because in this case a user can get any debugging information in visual forms of hierarchical graphs [12], [14]. Web interface also contains some usual parts like syntax highlighting, persistent storage for program code, authorization and so on.

The CSS system provides means to write and debug Cloud-Sisal-programs on low-cost devices as well as to translate and execute them in clouds. The CSS system can open the world of parallel and functional programming to all students and scientists without requiring a large investment in new, top-end computer systems.

3 Cloud Sisal Language

The Cloud Sisal language that has been designed as the input language of the CSS system is a modification of our Sisal 3.2 language which is aimed to increase the language’s utility for supporting of cloud scientific computations and cloud parallel programming [6],[7], [9],[15].

3.1 Single Assignment

Cloud Sisal differs from other functional languages and we think that this difference make Cloud Sisal more adapted for computational tasks. First of all, it has some usual functional language benefits like single assignment [16]. This approach requires every variable to be defined only once. Someone would say that it is not an advantage because every imperative program can be converted to SSA-form, and of course at low-level programming it has no difference but imagine some function and the global variable in the language where every variable need to be declared (we use C for example):
int g=0;
void foo(void) { g=1; }

You need to re-declare the global variable when it is modified, but you can’t
make it inside the function. Inside the compiler this program will be converted
quite easy but to write initially single assignment programs is not the same. You
can declare another global variable without setting any value but it can bring
more questions to the rest of the code, we can use more complex example to
withdraw this but we wouldn’t. The idea is that single assignment is something
similar to structural programming where “goto” operator is prohibited.

3.2 Streams and Arrays

Cloud Sisal also uses arrays and loops which is not common for a functional
language, but it is good for computation: you don’t have to worry about the
recognition of the tail recursion or the number of iterations or matrix description
which is simpler with arrays. You can operate with n-th element of the array in
a natural way like in FORTRAN:

for i in 1, N repeat
    R := A[i] * B[k]
    returns array of R
end for

3.3 Verbose Syntax

And the last benefit is more verbose syntax. It makes program source more
readable and as the result - long time development by different people becomes
easier. Many functional languages suffers from the lack of the words in the pro-
gram source, it makes the text hard to understand. The example below is the
famous Haskell quicksort:

qsort [] = []
qsort (x : xs) =
    qsort[y|y < xs, y < x] ++ [x] ++ qsort[y|y <= xs, y >= x]

This kind of code is hard to maintain. The same algorithm implemented in
Cloud Sisal listed below:

function qsort (Data:array[real] returns array[real])
if array_size(Data) > 2 then
    let
        L, Middle, R := for E in Data
        returns
            array of E when E < Data[1]
            array of E when E = Data[1]
            array of E when E > Data[1]
3.4 Loops and Reductions

In functional programming every statement is a function returning the value, the Cloud Sisal loops are the same. Reduction is used to determine the returning value of the loop. Keyword "returns" at the end of the loop is followed by the name of the reduction and its parameters. For example, if we need to summarize the elements in the array or the stream we can use the following function with "for all" loop:

```mathematical
function sum(A: array[real] returns real)
  for r in A
    returns sum of r
  end for
end function
```

Of course, loop construction can be used without any function declaration. Cloud Sisal is pure functional, it has no side effects and any loop contains the reduction call, also user can implement his own reductions.

The reductions are good because its implementation can depend on target system. When the program is executed in single-threaded environment it can be performed sequentially, but when executed on multiple threads it can be performed in parallel. Similar idea can be found in modern library "Threading Building Blocks" by Intel. This library allows usage of reduction mechanism in C++, but user can also use ordinary loops as well. In Cloud Sisal programs reductions can’t be avoided.

In Cloud Sisal we have three kinds of loops: post-conditional, pre-conditional and "for all" (operation is applied to a set). Reductions can be folding or generating (some aggregation function or an array generator). Conditional loops are sequential in general but reduction allows them to be pipelined easier (Fig. 2, Fig. 3). Loops can be divided into parts: initialization, loop body, loop test, loop reduction (ret) and range generator. We think that the part names can briefly describe them, but if you need more information - please check Sisal 3.2 or Cloud Sisal language description [6], [9].

Using reductions matrix multiplication can be implemented meaningfully:

```mathematical
function multiply( A,B : array[array[real]]; M,N,L : integer
```

returns array[array[real]]
for i in 1, M cross j in 1, L
  returns array of
    for k in 1, N repeat
      R := A[i,k] * B[k,j]
      returns sum of R
    end for
  end for
end function

Reduction can be always used in sequential style:

function multiply (A:array[array[real]]; B:array[array[real]]; N:integer
  returns array[array[real]])
for i in 1, N cross j in 1, N
  returns array of
    for
      initial s := 0.0; k := 1
      while (k <= N)
        repeat
          s:=old s + A[i,old k]*B[old k,j];
          k := old k + 1
        returns value of s
      end for
    end for
end function

But imperative languages doesn’t have any reduction mechanism at all.

3.5 Error Handling

Try-catch mechanism is more popular for error handling today but this approach has conflicts with parallel program execution. When the exception occurs all the
execution streams must be stopped, pipeline flushed and so on. Also it is harder
to keep program determinism in the case of the parallel execution and exception
occurs. Check the following Java example:

```java
try {for (int i=0; i<N; i++) {a[i]= a[i]/((i+1) % K);}}
   } catch (Exception e) {
      //display partial results stored in ‘a’
   }
```

In this example loop iterations are independent and can be executed in parallel.
Sequential execution will always give the same result (for the fixed values of N
and K); the result will not depend on the executor properties as far as it remains
to be sequential. While there is no dependence between the iterations, program-
ing language semantics remains to be sequential and parallelism exploration
can break this semantics or demand additional corrections to keep it. Interpreter
or parallelizing compiler needs additional mechanism to differ between the data
before and after the exception.

In Cloud Sisal language we have “always finished computations” semantics,
which means that execution stream will not stop on any error and return result-
ing value even if the error occurs (Fig. 4).

**Fig. 3.** “For all” pipelined structure.

**Fig. 4.** Error value propagation in “always finished computations” semantics.
3.6 Optimizing Annotations

The Cloud Sisal language supports also so-called annotated programming and concretizing transformations [10] and includes pragma statements in the form of formalized comments (annotations) that start with dollar sign ‘$’ and are predicate constraints on admissible properties of program fragments or states of computations. In addition to restricted set of program executions and restricted set of program outputs some suitable criterion of program quality can be defined by annotations, and every concretizing transformation of an annotated program is aimed at improving the program according to the qualitative criterion without disturbing the meaning of the program in the application context defined by annotations.

```cloud-sisal
forward function fact (n: integer
    /*$ assert=n>=1*/
    /*$ assert=_>=n*/
    returns integer)
function fact (n: integer returns integer)
    if n = 1 then 1
    else /*$ assert = _ > 0 */
        fact(n-1)*n
    end if
end function
```

Fig. 5. Cloud-Sisal-program with optimizing annotations.

According to the approach used [10], any source program is considered as a base for constructions of a number of different specialized programs. Every construction starts with the source program and an application context conveyed in annotations. Some program annotations can be formed in parallel with the development of the source program, others are added by users and describe a specific context of source program applications. Then a series of concretizing transformations is applied to the annotated general-purpose program (either automatically or interactively with the user), which results in a correct and qualitative specialized program.

For example, every expression in Cloud-Sisal-program can be prefixed by a pragma “assert = Boolean expression”, that can be checked for truth after the expression evaluation during program debugging and then can be used in program optimizing transformations. The result of the expression can be denoted as the underscore symbol “_” and if the expression is n-ary (where n > 1), then its components can be denoted as an array with the name “_[1]”, “_[n]”. In addition, the pragma “assert = Boolean expression” can be placed before returns keyword in procedure declarations and can be used to control results of this procedure after its invocation. As an example of the assert pragma statement usage please consider factorial function declaration and definition which
are represented in Fig. 5. Another example of optimizing annotations is a pragma “parallel” which can be used before a case expression in Sisal (analogous to a switch expression in C language). This pragma can be specified if it is known that only one test can be true. The pragma of the form “parallel = Boolean expression” means that only one test is true if the specified Boolean expression is true.

4 Internal Representations

The CSS system uses three internal presentations of Cloud-Sisal-programs: IR1, IR2 and IR3.

function sign (N: integer returns integer)
  if N > 0 then 1
  elseif N < 0 then -1 else 0 end if
end function

Fig. 6. A function sign and its IR1-representation.

4.1 Representation IR1

IR1 is a language of hierarchical graphs [12] made up simple and compound computation nodes, edges, ports and types (See Fig. 6). Nodes correspond to computations. Simple nodes are vertices and denote operations such as add or divide. Compound nodes are subgraphs and represent compound constructions such as structured expressions and loops. Ports are vertices that are used for input values and results of compound nodes. Edges show the transmission of data between simple nodes and ports; types are associated with the data transmitted on edges. IR1-program represents data dependencies, with control left implicit; e. g. iteration is represented as a compound node with subgraphs describing generation of index values, the body of the loop, and the packaging of results.
4.2 Representation IR2

IR2 is an extension of IR1. All nodes in the IR2 graph are partially ordered by the \( \preceq_e \) ordering in such a way that if \( N_1 \prec_e N_2 \), then \( N_1 \) must be executed before \( N_2 \), and if \( N_1 =_e N_2 \), then \( N_1 \) and \( N_2 \) can be executed in any order, and a parallel execution is possible. All edges in the IR2 graph are decorated by variables (See Fig. 7) which will be the operands of IR3 operations. Each variable has the following attributes: a unique identifier, a unique name, a type and an additional Boolean variable which defines the “IsError” property. The types in IR2 and IR3 represent the types of the Cloud Sisal language within IR2 and IR3. Each type contains additional low-level information about objects (such as machine representation of the type). IR2 is intended to provide a natural and usable structure for optimizations. During the optimization process, the optimizations can create additional data connected with a node, an edge or a port. The data created by one optimization can be reused by another.

![Fig. 7. IR2-representation of the function sign.](image)

4.3 Representation IR3

IR3 is a classical three-address code representation with hierarchical blocks. For example, function sign can be represented as follows:

```
0 entry "function sign[integer]" (V_1(I32) returns V_3(I32));
{
  1  V_5(I32) = V_1(I32);
  2  V_5(I32) = V_1(I32);
  3  V_9(I32) = 0x0(I32);
  4  V_13(I32) = 0x0(I32);
  5  V_7(BOOL) = (V_9(I32) < V_5(I32));
  6  V_11(BOOL) = (V_5(I32) < V_13(I32));
  7  if (V_7(BOOL) == true(BOOL))
```

{ 
  V_15(I32) = 0x1(I32);
  V_3(I32) = V_15(I32);
}
else
{
  if (V_11(BOOL) == true(BOOL))
  {
    V_19(I32) = 0x1(I32);
    V_17(I32) = - V_19(I32);
    V_3(I32) = V_17(I32);
  }
  else
  {
    V_21(I32) = 0x0(I32);
    V_3(I32) = V_21(I32);
  }

  return;
}

5 Cross-compiler

The optimizing cross-compiler of the CSS system consists of two main parts: front-end and back-end compilers (Fig. 8).

The front-end compiler translates Cloud-Sisal-modules into a monolithic IR1-program which is used also by the interpreter and the graphic visualization/debugging subsystem.

The back-end compiler begins with R2Gen which produces a semantically equivalent program in IR2. Then the IR2Opt subsystem performs some optimizations and concretizations on the annotated program to produce a semantically equivalent, but faster basic program. After completion of the machine-independent optimizations, the IR3Gen subsystem preallocates array storage where compile time analysis or compiler generated expressions executed at run time can calculate the final size of an array. The result of this phase is the production of a semantically equivalent program in IR3. The next phase of compilation (IR3Opt) performs update-in-place analysis and restructuring to help identify at compile time those operations that can execute in-place and to improve chances for in-place operation at run time when analysis fails. It performs also some machine-dependent optimizations and defines the desired granularity of parallelism based on an estimate of computational cost and various parameters that tune analysis. After parallelization, CodeGen generates C++ or C# code, and the compilation can be completed using the target machine’s C++ or C# compiler.

Mathematical and Information Technologies, MIT-2016 — Information technologies

151
6 Related Works

New parallel language development is not popular today; more popular is existing language extension (sometimes it is positioned as a separate language); such approach keeps sequential semantics problems, but considered as the fastest both for the developer and for the final application execution. In this section we will not observe such extensions as related.

6.1 The Pifagor language

This language has been developed as a functional language for creation of architecture-independent parallel programs [17]. The Pifagor language is optimized to dataflow graph description and has a specific syntax which is not easy to understand because it differs from common imperative and functional languages. For example, the following Pifagor function performs vector multiplication by scalar:

```
VecScalMult << funcdef Param
// Argument format: ((x1, x2, : xn), y),
// where ((x1,x2,:xn) is a vector, y is a scalar
{
((Param:1,(Param:2,Param:1:|):dup):#:[]:*) >>return
}
```

It is hard to compare Pifagor syntax and constructions with Sisal because they are completely different. Sisal has loops and arrays; we suppose it is better for science computational tasks. According to the articles of the Pifagor developers [17] the Pifagor language is aimed on the list processing and the conception of unlimited parallelism scheduled as limited at runtime. At present, there are both an experimental compiler and an experimental interpreter for the Pifagor language that are used for scientific proposes such as a development of the new scheduling algorithms and for parallel programming education.
6.2 The F# language

We can’t say that F# [18] is the project in the same direction with Sisal, but Microsoft’s developments in a functional paradigm can’t be avoidable. As the complexity of the systems was increased the complexity of compiler grows and some features of the functional languages formerly considered as ineffective started to implement in imperative languages.

On the one hand F# is functional ML-family language and functional paradigm suits better for parallel computations. On the other hand F# has an ability to create any mutable indexes, non-functional calls or dependencies, external .NET objects and operations. It can’t be considered as single assignment or parallel language. It is hybrid language that can be used for writing both implicitly parallel and sequential programs. Multi-threaded programming on F# is quite similar to C# or C programming.

Not only in case of the F# language but for the most of functional languages developers are trying to make the programming language available for wide range of users but it makes language less pure and less functional. State modification operators such as input and output give the developer familiar ability to process the data but make the semantic sequential or non-deterministic.

7 Conclusion

The project of visual cloud system CSS aimed at supporting of functional and parallel programming teaching and learning was considered.

The CSS system is intended to provide means to write and debug functional programs regardless target architectures on low-cost devices as well as to translate them into optimized parallel programs, appropriate to the target execution platforms, and then execute them in clouds on high performance parallel computers without extensive rewriting and debugging. So, the CSS system can open the world of parallel and functional programming to all students and scientists without requiring a large investment in new, top-end computer systems.

At present, the CSS system consists of experimental versions of Web-interface, interpreter, graphic visualization / debugging subsystem, optimizing cross-compiler. The current target platform for the Cloud-Sisal-compiler is .NET. The compiler performs conventional optimizing transformations and generates the C# code. It allows the users to perform the experimental execution of Cloud-Sisal-programs and examine the effectiveness of optimizing transformations applied by the compiler. We starts some experiments of using our system for teaching and learning of functional and parallel programming as well as of optimizing compilation and high performance computing.

Acknowledgments. Our thanks to all colleagues taking part in the project described. This research is supported by the Russian Foundation for Basic Research under grant RFBR N 15-07-02029.
References

The Comparison of Different SAT Encodings for the Problem of Search for Systems of Orthogonal Latin Squares

Stepan Kochemazov, Oleg Zaikin, and Alexander Semenov
Matrosov Institute for System Dynamics and Control Theory SB RAS, Irkutsk, Russia
{veinamond,zaikin.icc}@gmail.com,biclop.rambler@yandex.ru

Abstract. In this report we present several different propositional encodings for finding systems of mutually orthogonal Latin squares, and evaluate their effectiveness using state-of-the-art parallel and sequential algorithms for solving Boolean satisfiability problem (SAT). We also apply the widely used SMAC tool to study the possibility of improving the effectiveness of LINGELING SAT solver on the considered tests and discuss the results of corresponding computational experiments.

Keywords: latin squares, MOLS, SAT, combinatorial designs.

1 Introduction

A Latin square $A = (a_{ij})$ of order $n$ is an $n \times n$ table filled with symbols from the set $N = \{0, 1, ..., n - 1\}$, in such a way that each symbol occurs precisely once in each row and each column [5]. Diagonal Latin square is a Latin square, in which each symbol from $N$ occurs precisely once in its main diagonal and in its main antidiagonal. Two Latin squares $A = (a_{ij})$ and $B = (b_{ij})$ are orthogonal if all ordered pairs $(a_{ij}, b_{ij})$ are distinct. A set of Latin squares, each two of them orthogonal, is called a set of mutually orthogonal Latin squares (MOLS). A set of diagonal Latin squares, each two of them orthogonal, is called a set of mutually orthogonal diagonal Latin squares (MODLS).

Latin squares represent a well studied combinatorial design with rich history dating several centuries back. They have a number of useful properties that make it possible to use them in various areas, such as experimental design, error correcting codes, etc. Also Latin squares are often used as a basis for mathematical puzzles, such as Sudoku and Magical squares. That being said, there remain several hard combinatorial problems related to Latin squares that have not been solved yet. One of the most complex and widely known problem is the one that consists in answering the question if there exists a triple of MOLS of order 10. Even relatively simple problems, such as finding pairs of orthogonal (diagonal) Latin squares of order 10 are relatively hard and there are no known effective methods for solving them.
In this context it is interesting to apply to this class of problems state-of-the-art combinatorial algorithms. One class of such algorithms is formed by the algorithms for solving Boolean satisfiability problem (SAT) [4]. Boolean satisfiability problem is usually formulated as follows: for an arbitrary Boolean formula to answer the question if it is satisfiable. SAT is historically the first NP-complete problem [6], thus it is possible to reduce many combinatorial problems to SAT. What makes it different from many other NP-complete problems is the fact that in the last twenty years there was achieved a remarkable progress in the effectiveness of SAT solving algorithms. Annual SAT competitions [2] make it possible to quickly evaluate new heuristics and ideas using relatively wide class of tests. Today there exist both sequential and multi-threaded implementations of SAT solving algorithms (the corresponding software complexes are usually referred to as SAT solvers). Interesting feature of these algorithms is that since they incorporate many heuristics, each SAT solver has a number of numeric parameters, that can be easily varied. Given a specific solver, on some classes of tests by carefully choosing proper values of parameters it is possible to achieve quite significant performance gain. In this paper we apply state-of-the-art SAT solving algorithms to the problem of finding pairs and triples of MOLS and evaluate their effectiveness.

Let us present the brief outline of the paper. In the next section we provide basic terms about SAT and consider how the problem of finding MOLS can be reduced to SAT. In the third section we describe state-of-the-art tools for finding effective combinations of SAT solver parameters and describe our experimental setup. In the fourth section we present the results of our computational experiments and their discussion. In the last section we make conclusions and outline our plans for the future.

2 Constructing SAT Encodings for Finding MOLS

Interesting and often forgotten fact about SAT encodings is that usually for the same problem we can construct several different variants of them. For example, it can be done via representing the original problem by some equivalent problem. Latin squares can serve as a good example of such approach, because a Latin square of order \( n \) as a combinatorial design is equivalent to orthogonal array, a set of \( n \) disjoint transversals, a set of 3 orthogonal matrices, etc. It means that in practice to search for Latin squares with specific properties we can search for any of equivalent objects with desired properties adapted to their form. Another way to construct different SAT encodings for the same problem consists in employing different methods for reducing specific constraints to SAT form. A good parallel here can be drawn with sorting methods: in practice we can sort a numerical array via many different methods, and the context defines which one is better. In the present paper we will follow this path. Note, that the problem of finding MOLS with specific properties via SAT was already considered in a number of papers, for example in [14]. However, previous works did not study several significantly different encodings for the same problem and also did not
apply parametrization algorithms to tune solver performance. In the following subsection let us briefly introduce basic notation related to SAT.

2.1 SAT Basics

Boolean formula is constructed from Boolean variables $X = \{x_1, \ldots, x_n\}$, logical operators $\{\neg, \lor, \land\}$ and parentheses. A literal is either a Boolean variable or its negation. A formula is called satisfiable if we can find such an assignment of logical values $\{\text{TRUE, FALSE}\}$ (for simplicity it is usually assumed that TRUE is equivalent to 1 and FALSE is equivalent to 0) to its variables that the result is TRUE. Otherwise the formula is called unsatisfiable. An assignment of logical values that makes formula TRUE is called a satisfying assignment. Boolean satisfiability problem consists in answering the question if a given formula is satisfiable [4]. In practice a SAT solving algorithm needs to either find a satisfying assignment or to prove that formula is unsatisfiable. Usually, state-of-the-art SAT solving algorithms work with Boolean formulas represented in Conjunctive Normal Form (CNF). CNF is essentially a conjunction of clauses, where by clause we mean a disjunction of several literals or a single (unit) literal. An arbitrary formula can be effectively represented in CNF using Tseitin transformations [11]. In the following subsection we will consider the process of constructing different SAT encodings for finding MOLS in more detail.

2.2 Reducing the Problem of Search for MOLS to SAT

Let us first represent the problem of search for MOLS as a problem of satisfying a set of constraints on Boolean variables. Latin square of order $n$ can be represented as an incidence cube [9] — an $n \times n \times n$ 0–1 array where dimensions are identified with the rows, columns and symbols of a Latin square. Cell with coordinates $(i, j, k)$ contains 1 if and only if the cell of the corresponding Latin square with coordinates $(i, j)$ contains $k$. From the definition of Latin square it follows that if we fix two coordinates in incidence cube, then in the remaining ‘line’ of the cube, produced by varying the remaining coordinate, there should be exactly one 1. Thus to represent the Latin square we use $n^3$ Boolean variables $\{x(i, j, k)\}$, $i, j, k = 1, \ldots, n$, encoding the incidence cube. Let us introduce the general form of $EO$ predicate: assume that $X = \{x_1, \ldots, x_m\}$ is a set of Boolean variables. Then $EO(X) = EO(x_1, \ldots, x_m) = 1$ if and only if among $x_1, \ldots, x_m$ exactly one variable takes the value of TRUE and all remaining variables take the value of FALSE. We will consider the ways this predicate can be transformed to CNF below. Then the Latin square can be specified using following constraints:

\[
\begin{align*}
EO(x(i, j, 1), \ldots, x(i, j, n)), i, j = 1, \ldots, n; \\
EO(x(i, 1, k), \ldots, x(i, n, k)), i, k = 1, \ldots, n; \\
EO(x(1, j, k), \ldots, x(n, j, k)), j, k = 1, \ldots, n.
\end{align*}
\]
If we consider a diagonal Latin square, then we add two more constraints on variables corresponding to main diagonal and antidiagonal:

\[
EO(x(1, 1, k), \ldots, x(n, n, k)), k = 1, \ldots, n;
EO(x(1, n, k), \ldots, x(n, 1, k)), k = 1, \ldots, n.
\]

Since we encode the problem of finding MOLS, we also need to specify the orthogonality condition. Let us remind that Latin squares \(A\) and \(B\) are orthogonal if all ordered pairs of the form \((a_{ij}, b_{ij})\) are distinct. Assume that variables \(\{x^A(i, j, k)\}\) correspond to square \(A\) and variables \(\{x^B(i, j, k)\}\) correspond to square \(B\). One way (usually referred to as naïve) to represent orthogonality condition in the form of CNF is to write \(\approx n^6\) clauses of the kind:

\[
\neg x^A(i_1, j_1, k_1) \lor \neg x^A(i_2, j_2, k_2) \lor \neg x^B(i_1, j_1, k_1) \lor \neg x^B(i_2, j_2, k_2),
\]

\(i_1, i_2, j_1, j_2, k_1, k_2 = 1, \ldots, n, i_1 \neq i_2, j_1 \neq j_2.\)

Each of these clauses restricts that the ordered pair \((k_1, k_2)\) appears simultaneously in two different cells with coordinates \((i_1, j_1)\) and \((i_2, j_2)\).

An alternative variant of representing orthogonality condition in CNF relies on the use of auxiliary variables. Assume we use \(n^2\) arrays of \(n^2\) Boolean variables \(OC[u, v] = (oc^{u,v}_{1,1}, \ldots, oc^{u,v}_{n,n})\), \(u, v = 1, \ldots, n\). With each variable from these arrays we associate the following Boolean equation

\[
oc^{u,v}_{i,j} \equiv x^A(i, j, u) \land x^B(i, j, v)
\]

i.e. if the pair \((u, v)\) is formed by elements of Latin squares with coordinates \((i, j)\) then the corresponding Boolean variable \(oc^{u,v}_{i,j}\) takes the value of TRUE. After this the orthogonality condition can be written using the following constraints:

\[
EO(OC[k_1, k_2]), k_1, k_2 = 1, \ldots, n.
\]

It is clear that by employing different encoding schemes for \(EO\) predicate we can produce different SAT encodings for finding MOLS. Let us consider the \(EO\) predicate in more detail.

### 2.3 Encoding \(EO\) predicate

In recent years there had been published several variants of algorithms for encoding it to SAT. The paper [10] contains quite comprehensive and detailed review of them. Below let us briefly cite the encoding variants described in [10] that we will use in our experiments. Assume that \(X = (x_1, \ldots, x_m)\). It is convenient to split the \(EO\) predicate into two:

\[
EO(X) = AMO(X) \land ALO(X)
\]

where \(AMO(X)\) is the predicate encoding that among the variables from the set \(X\) there is at most one with the value of TRUE, and \(ALO(X)\) is a predicate encoding that at least one variable in \(X\) takes the value of TRUE. The encoding
of $ALO$ predicate to CNF is relatively straightforward – it can be written via one big clause:

$$ALO(x_1, \ldots, x_m) = x_1 \lor \ldots \lor x_m.$$ 

It is the $AMO$ predicate that requires a more thorough consideration. The most simple way to transform it to CNF is the so-called Pairwise encoding. It implies the construction of $m \times (m - 1)/2$ clauses of the kind

$$-v_i \lor -v_j, i, j = 1, \ldots, m, i \neq j.$$ 

Pairwise encoding is the only encoding we used that does not employ auxiliary variables for encoding $AMO$ predicate.

When we use Binary encoding to transform $AMO$ predicate to CNF we introduce auxiliary variables $b_1, \ldots, b_{\lfloor \log_2 m \rfloor}$. Then with each variable $x_i \in X, i = 1, \ldots, m$ we associate a set of clauses:

$$\bigwedge_{j=1}^{\lceil \log_2 m \rceil} \neg x_i \lor \psi(i, j),$$

where $\psi(i, j)$ denotes $b_i$ ($\neg b_i$) if the $j$-th bit of binary representation of $i - 1$ is $1(0)$. The idea behind the encoding is to create such situation that when any $x_i$ is assigned the value of TRUE it leads to the assignment of all auxiliary variables $b_1, \ldots, b_{\lfloor \log_2 m \rfloor}$ and thus to assigning the value of FALSE to all other $x_j, j \neq i$.

The Commander encoding can be considered a meta-encoding method as it can employ other encodings. It is based on dividing the set $X$ into $k$ disjoint subsets $G_1, \ldots, G_k$. Then we introduce auxiliary commander variables $c_1, \ldots, c_k$ that act as representatives of corresponding subsets. After this we write the following set of clauses

$$\bigwedge_{i=1}^{k} AMO(\neg c_i \cup G_i) \land \bigwedge_{i=1}^{k} ALO(\neg c_i \cup G_i) \land AMO(c_1, \ldots, c_k).$$

The idea is that once a variable $x_i$ becomes TRUE, the corresponding commander variable also becomes TRUE, and thus all $x_j, j \neq i$ are assigned FALSE.

In Product encoding we introduce $p \times q$ auxiliary variables $U = \{u_1, \ldots, u_p\}$ and $V = \{v_1, \ldots, v_q\}, p \times q \geq m$. Then each variable $x_k, k = 1, \ldots, m$ is mapped to a pair $(u_i, v_j)$, such that $k = (i - 1)q + j$. Then the following set of clauses:

$$AMO(X) = AMO(U) \land AMO(V) \bigwedge_{1 \leq k \leq m, k = (i-1)q+j} (\neg x_k \lor u_i) \land (\neg x_k \lor v_j)$$

guarantees that once $x_k$ is assigned TRUE, the corresponding pair of variables $(u_i, v_j)$ are also assigned TRUE and it leads to assigning FALSE to remaining $x_t, t \neq k$.

The Sequential encoding implements the idea of sequential counter, where we traverse the set of variables in the ordered fashion and track if we already
met a variable with the value of TRUE. It employs $m - 1$ auxiliary variables $s_1, \ldots, s_{n-1}$ in the following set of clauses:

$$(-x_1 \lor s_1) \land (-x_m \lor -s_{n-1}) \land \bigwedge_{1 < i < n} \left((-x_i \lor s_i) \land (-s_{i-1} \lor s_i) \land (-x_i \lor -s_{i-1}) \right).$$

Finally, the Bimander encoding combines the features of Binary and commander encodings in the following way. Similarly to commander encoding the original set $X$ is divided into $k$ disjoint subsets $G_1, \ldots, G_k$, such that each group $G_i$ contains $g = \lceil \frac{m}{k} \rceil$ variables. Then we introduce auxiliary variables $b_1, \ldots, b_{\lceil \log_2 m \rceil}$ similar to binary encoding and write the following set of clauses:

$$\bigwedge_{i=1}^{k} \left(AMO(G_i) \land \bigwedge_{h=1}^{g} \bigwedge_{j=1}^{\lceil \log_2 k \rceil} \neg x_{i,h} \lor \psi(i, j) \right),$$

where $\psi(i, j)$ has the same meaning as in binary encoding.

We applied all considered encoding methods to construct different SAT encodings for several problems of finding pairs and triples of MOLS of various order. The results are presented in Table 1. Since the size difference between SAT encodings for finding MOLS and MODLS is relatively little, we compare only encodings for finding MOLS. The columns NAIVE, PW, BN, CM, PR, SQ, BM correspond to naive, pairwise, binary, commander, product, sequential and bimander encoding schemes, respectively. Note, that we applied them only to encode EO predicates corresponding to orthogonality condition. EO predicates corresponding to constraints on incidence cube were encoded using pairwise encoding in all cases. Below we refer to problem of finding $k$ MOLS (MODLS) of order $n$ as $MOLS_{n,k}$ ($MODLS_{n,k}$). For $n = 8, 9$ we used $p = 3$, $q = 3$ for product encoding and $m = 3$ for bimander and commander encodings, and for $n = 10$ we used $p = 3$, $q = 4$, $m = 3$.

### Table 1. Size of SAT encodings for finding MOLS constructed using different AMO encoding schemes, Kb.

<table>
<thead>
<tr>
<th>Problem</th>
<th>NAIVE</th>
<th>PW</th>
<th>BN</th>
<th>CM</th>
<th>PR</th>
<th>SQ</th>
<th>BM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOLS_8_2</td>
<td>2 371</td>
<td>2 233</td>
<td>691</td>
<td>660</td>
<td>508</td>
<td>512</td>
<td>727</td>
</tr>
<tr>
<td>MOLS_8_3</td>
<td>7 326</td>
<td>6 757</td>
<td>1 952</td>
<td>1 857</td>
<td>1 383</td>
<td>1 420</td>
<td>2 059</td>
</tr>
<tr>
<td>MOLS_9_2</td>
<td>5 164</td>
<td>4 415</td>
<td>1 224</td>
<td>1 126</td>
<td>836</td>
<td>864</td>
<td>1 330</td>
</tr>
<tr>
<td>MOLS_9_3</td>
<td>15 752</td>
<td>13 877</td>
<td>3 551</td>
<td>3 234</td>
<td>2 292</td>
<td>2 336</td>
<td>3 889</td>
</tr>
<tr>
<td>MOLS_10_2</td>
<td>10 196</td>
<td>8 343</td>
<td>1 930</td>
<td>1 856</td>
<td>1 328</td>
<td>1 363</td>
<td>2 169</td>
</tr>
<tr>
<td>MOLS_10_3</td>
<td>30 893</td>
<td>26 185</td>
<td>5 515</td>
<td>5 269</td>
<td>3 566</td>
<td>3 618</td>
<td>6 287</td>
</tr>
</tbody>
</table>

### 3 Parametrization of SAT solving algorithms

State-of-the-art SAT solvers are programs with a large number of various parameters. Meticulous tuning of these parameters for a particular class of problems
may significantly improve average effectiveness of the solver: sometimes it is even possible to achieve more than 100 times better performance.

There are some generally accepted techniques for parametrization of SAT solvers. According to papers [8, 7, 1] one should select from the considered set of SAT instances some subset called training set. On this training set the SAT solver with different combinations of parameter values is launched. For each combination of parameter values a special objective function is computed to measure the effectiveness of SAT solving with these values. The parametrization process itself is essentially a local search in a Cartesian product of domains containing all possible values of considered parameters. So all the solver parameters must be discrete (each non-discrete parameter should be discretized). To jump from local optimums sometimes various metaheuristic procedures are used. When the local search scheme finishes its work either due to reaching the optimality condition or because time limit was exceeded, the resulting combination of parameters is applied to solve all problems from the so-called test set. By test set it is usually meant the set of test instances used to measure the effectiveness of the parametrization procedure. Usually it is assumed that test set and training set do not intersect. In practice, given some test set one employs reasonable heuristics to form training set using problems of the same nature and similar dimension as in test set.

There are three state-of-the-art tools for tuning the parameters of SAT solvers: ParamILS (Iterated Local Search in Parameter Configuration Space [8]), GGA (Gender-based Genetic Algorithm [1]) and SMAC (Sequential Model-based Algorithm Configuration [7]). All mentioned tools are based on some local search metaheuristics. GGA uses the genetic algorithm, ParamILS and SMAC are based on the iterative hill climbing algorithm. Using these tools one can speed up both local search and tree search SAT algorithms. With the help of these tools two Configurable SAT Solver Challenges (CSSC) in 2013 and 2014 were held. According to [7] SMAC shows better efficiency compared to ParamILS and GGA. That is why in our computational experiments we used SMAC.

One of the last stages in the development of a SAT solver consists in fixing default values of its parameters. Usually to find such values developers use one of the mentioned tools. During this process the families of SAT instances from the latest SAT competitions are usually used as a training set. One of the reasons of such choice is that these families are formed by SAT encodings of problems from various areas of science. The set of values that has showed the best performance on the training set is finally utilized in the role of the default set of the parameters values. As a result the tuned solver shows good performance in application to the wide class of problems. Meanwhile, in practice one often faces the situation when it is necessary to solve large amount of SAT instances which encode almost identical combinatorial problems. Often such combinatorial problems are obtained by decomposing an original combinatorial problem. In the next section this situation will be described in application to finding systems of MOLS and MODLS.
We chose the LINGELING SAT solver [3] for tuning, because at CSSC 2014 it won the gold medals in all categories (Random SAT, Random SAT+UNSAT, Industrial SAT+UNSAT, Crafted SAT+UNSAT). There are 323 parameters in LINGELING available for tuning. In the next section we will show how this solver works with the default values of parameters on SAT instances built according to the SAT encodings described in the previous section. Then we will describe the results of its tuning in application to some problems which turned out to be very hard.

4 Computational experiments

To compare the effectiveness of SAT solvers on the SAT encodings described in Section 2 we considered the following combinatorial problems:

- MOLS\(_{8,2}\);
- MOLS\(_{9,2}\);
- MOLS\(_{10,2}\);
- MODLS\(_{8,2}\);
- MODLS\(_{8,3}\);
- MODLS\(_{10,2}\);
- MODLS\(_{10,3}\).

For each of these problems we made all 7 considered SAT encodings. In our experiments besides LINGELING, we used PLINGELING and TREENGELING SAT solvers [3]. We chose them because they won several prizes on the latest SAT competition and SAT Race (they were held in 2014 and 2015 respectively). In particular, we used the versions from SAT Race 2015 for all solvers. Note that PLINGELING and TREENGELING are multi-threaded programs, while LINGELING is a sequential program. We utilized the experimental setup consisting of two 16-core AMD 6276 processors coupled with 64 Gb RAM. Thus, each multi-threaded solver was launched on 32 cores.

On the first stage we launched each of the mentioned solvers with the default values of parameters on every SAT instance (49 SAT instances in total). In this experiment we used the time limit of 5000 seconds for every launch. This limit corresponds to the wall time, so for a sequential solver it is almost identical to CPU time, but in the case of multi-threaded solvers the CPU time can be much greater. It should be noted that this is a standard wall time limit used at SAT competitions. In Tables 2, 3 and 4 the results of the described experiment are shown. Each value in these tables corresopnds to the CPU time in seconds. Here “-” stands for “interrupted due to exceeding the time limit”. The best value for each pair (problem, type of the SAT encoding) is marked with bold.

Let us discuss the obtained results. In every case when a solver could cope with a SAT instance, the answer was SATISFIABLE. Problems MODLS\(_{10,2}\) and MODLS\(_{10,3}\) turned out to be very hard — no pair (SAT solver, SAT encoding) could cope with them. On the considered test instances two variants of SAT encodings (BM and PW) turned out to be quite mediocre. Interesting
Table 2. CPU time for LINGELING with the default parameters values, seconds.

<table>
<thead>
<tr>
<th>Problem</th>
<th>NAIVE</th>
<th>PW</th>
<th>BN</th>
<th>CM</th>
<th>PR</th>
<th>SQ</th>
<th>BM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOLS_8_2</td>
<td>50</td>
<td>151</td>
<td>2489</td>
<td>674</td>
<td>1718</td>
<td>158</td>
<td>276</td>
</tr>
<tr>
<td>MOLS_9_2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MOLS_10_2</td>
<td>2404</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MODLS_8_2</td>
<td>50</td>
<td>200</td>
<td>957</td>
<td>1890</td>
<td>290</td>
<td>847</td>
<td>248</td>
</tr>
<tr>
<td>MODLS_8_3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MODLS_10_2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MODLS_10_3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3. CPU time for PLINGELING, seconds.

<table>
<thead>
<tr>
<th>Problem</th>
<th>NAIVE</th>
<th>PW</th>
<th>BN</th>
<th>CM</th>
<th>PR</th>
<th>SQ</th>
<th>BM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOLS_8_2</td>
<td>13.7</td>
<td>0.4</td>
<td>25.5</td>
<td>0.2</td>
<td>0.1</td>
<td>0.2</td>
<td>620.6</td>
</tr>
<tr>
<td>MOLS_9_2</td>
<td>5581</td>
<td>20135</td>
<td>142346</td>
<td>40396</td>
<td>4516</td>
<td>15141</td>
<td>-</td>
</tr>
<tr>
<td>MOLS_10_2</td>
<td>9169</td>
<td>73653</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>38935</td>
</tr>
<tr>
<td>MODLS_8_2</td>
<td>1400</td>
<td>512</td>
<td>1272</td>
<td>2850</td>
<td>91</td>
<td>851</td>
<td>3617</td>
</tr>
<tr>
<td>MODLS_8_3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MODLS_10_2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MODLS_10_3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4. CPU time for TRENGELING, seconds.

<table>
<thead>
<tr>
<th>Problem</th>
<th>NAIVE</th>
<th>PW</th>
<th>BN</th>
<th>CM</th>
<th>PR</th>
<th>SQ</th>
<th>BM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOLS_8_2</td>
<td>19</td>
<td>779</td>
<td>4976</td>
<td>536</td>
<td>9</td>
<td>643</td>
<td>31</td>
</tr>
<tr>
<td>MOLS_9_2</td>
<td>32569</td>
<td>29876</td>
<td>153762</td>
<td>26940</td>
<td>32175</td>
<td>9290</td>
<td>53727</td>
</tr>
<tr>
<td>MOLS_10_2</td>
<td>18710</td>
<td>102101</td>
<td>-</td>
<td>6111</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MODLS_8_2</td>
<td>437</td>
<td>739</td>
<td>391</td>
<td>1567</td>
<td>1269</td>
<td>2574</td>
<td>698</td>
</tr>
<tr>
<td>MODLS_8_3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>44986</td>
</tr>
<tr>
<td>MODLS_10_2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MODLS_10_3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5. CPU time of parametrized LINGELING on test sets, seconds.

<table>
<thead>
<tr>
<th>Problem</th>
<th>NAIVE</th>
<th>PW</th>
<th>BN</th>
<th>CM</th>
<th>PR</th>
<th>SQ</th>
<th>BM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODLS_10_2, 6 rows known</td>
<td>5.73</td>
<td>9.28</td>
<td>9.78</td>
<td>13.30</td>
<td>8.80</td>
<td>11.79</td>
<td>10.83</td>
</tr>
<tr>
<td>MODLS_10_3, 6 rows known</td>
<td>13.52</td>
<td>29.47</td>
<td>32.02</td>
<td>36.17</td>
<td>35.45</td>
<td>44.34</td>
<td>32.74</td>
</tr>
</tbody>
</table>
The fact is that plingeling showed superlinear speed up (compared to lingeling) on MOLS\_8\_2. On lingeling in all cases the best results were obtained using naive SAT encoding, however plingeling and treengeling in almost all cases displayed the best results with other encodings. Finally, the best CPU times for MOLS\_8\_2 and MOLS\_9\_2 were obtained by plingeling on pr encoding; for MOLS\_10\_2 and MODLS\_8\_2 — by lingeling on naive SAT encoding; for MODLS\_8\_3 — by treengeling on sq SAT encoding.

In [13] we considered weakened problems for MODLS\_10\_3 in which the values of the first 45 cells of the first DLS were added to an original CNF via unit clauses. The corresponding SAT instances turned out to be quite hard. In the present paper we considered another type of weakened problems for MODLS\_10\_3. We constructed 100 SAT instances by adding to an original CNF the values of the first 6 rows of the first DLS via unit clauses. We also made 100 SAT instances for MODLS\_10\_2 in the similar way. Following [13] we took the corresponding values from the first 50 pairs of MODLS of order 10 found in the volunteer computing project SAT@home [12]. As a result we obtained 100 relatively simple (for lingeling with the default parameters values) SAT instances for each of the considered problems. We used SMAC (see Section 3) to find good lingeling parameters values for the obtained families of instances. In every case as the training set we used the first 10 SAT instances from a family, thus the remaining 90 SAT instances formed the test set. In every case SMAC was launched 16 times, each of them for 1 day (with different values of the start seeds). The final performance of the solver on test sets for each type of encoding is presented in Table 5 (the best one out of 16 SMAC results was chosen in every case).

Let us discuss the obtained results. For both considered weakened problems the naive encoding turned out to be better than others. With the help of parametrization the speed-up of 49 \% (40 \%) was achieved for the weakened problem MODLS\_10\_2 (MODLS\_10\_3) on this encoding in comparison with the default parameters values.

5 Conclusion

The results of computational experiments clearly show that it is very important to choose the best pair (SAT solver, SAT encoding) for a particular combinatorial problem. The parametrization of the chosen SAT solver also can provide an additional speed-up. In future we plan to evaluate other types of SAT encodings for the considered problems, such as the ones based on representing Latin squares via orthogonal arrays, sets of transversals, etc. We also plan to utilize other tools for the SAT solvers parameters tuning.

Acknowledgments. The research was funded by Russian Science Foundation (project No. 16-11-10046).
References

Development of Extended Path-based Role Access Control Model for Web Applications

Dmitry Kononov and Sergey Isaev

Institute of Computational Modeling of SB RAS
Akademgorodok 50/44, Krasnoyarsk, Russia
{ddk,si}@icm.krasn.ru
http://icm.krasn.ru

Abstract. Web applications security is a complex problem with several aspects. One aspect is access control according to specified security policy. Access control is accomplished by security model restrictions. This research is dedicated to developing security access control model for web applications. This work describes path-based RBAC model, which improves RBAC and allows flexible access control using request path (URI). Authors created guidelines to apply model’s elements for real-world web applications. Developing web applications with model described allows reducing security risks.

Keywords: security models, access control, web applications.

1 Introduction

Today modern Web applications and services are affected by several security issues. Computer security is becoming increasingly important and actual. According to Symantec security research [1], in spite of security means development there are high security risks for web applications. Web applications security is a complex problem with several aspects. One aspect is access control according to specified security policy. Access control is accomplished by security model restrictions. Choosing and applying appropriate security model is able to reduce risks of successful attacks.

Widely known security models include discretionary, mandatory, and role-based [2, 3]. In our work, we research and develop security model built on Role-based access control model (RBAC) [4]. Role access control represents access rights control from subjects to objects grouped by some characteristics named roles. Original Role-based access control model does not take into account web applications features [5], particularly hierarchic requests. Also assigning permissions is limited to roles only. This work describes path-based RBAC model, which improves RBAC and allows flexible access control using request path (URI).
2 Security models

Currently, there are several security access control models. Some of them include access control accomplished by discretionary matrix, mandatory levels, and role-based.

Discretionary security models are based on access control from subjects to objects by using access control lists or access matrix. This family include security model such as Harrison-Ruzzo-Ulman [6], typed access matrix [7], Take-Grant [8].

Mandatory access control – access control from subjects to objects based on assigned confidentiality label for information contained in the objects and permission entities to access information with such level of confidentiality. An example of the mandatory model is Bell-LaPadula [2]. Classic Bell-LaPadula model analyzes conditions under which the computer system cannot initiate information flows from the objects with a high level of confidentiality to objects with a lower level of confidentiality.

Role-based access control is a further development of the discretionary access control policy: permissions to system objects are grouped according to certain characteristics, forming role. Roles are intended to manage access control rules in a more simple way. These models do not take into account the specifics of web applications, in particular, the hierarchical organization of requests and links. The paper describes the adapted role-based security model that eliminates these problems.

3 Role-based access control

The original role-based access control model [4] defines a set of elements:

\[
\langle U, R, P, S, UA(U), PA(R), user(S), roles(S) \rangle,
\]

where:
- \( U \) – set of users;
- \( R \) – set of roles;
- \( P \) – set of access permissions;
- \( S \) – set of user sessions;
- \( UA: U \rightarrow 2^R \) – function assigning for each user a variety of roles to which he can be authorized;
- \( PA: R \rightarrow 2^P \) – function assigning for each role set of access permissions, while \( \forall p \in P, \exists r \in R \text{ such that } p \in PA(r) \);
- \( user: S \rightarrow U \) – function defining for each user session, on whose behalf it is authorized;
- \( roles: S \rightarrow 2^R \) – function defining for user a variety of roles for which he is authorized with current session; at the same time \( \forall s \in S \text{ satisfies the condition } roles(s) \subseteq UA(user(s)) \).

The model \( RBAC_1 \) is defined as \( RBAC_0 \), at the same time introducing a role hierarchy (RH).
4 Adapting the model for web applications

To existing RBAC\textsubscript{1} model elements ”user”, ”role”, ”permission”, ”session” we added new elements taking into account web application features: ”token”, ”request”.

**Definition 1.** Token (Tk) – set of user attributes that allow him to carry out authentication in a system. Token is a pair <name, password>, or pair <public key, private key>.

**Definition 2.** Request (Rq) – set of information sent by the client to HTTP server. The request contains a set of headers, a unique resource identifier (URI), a set of parameters name/value, and a request payload (body).

A request belongs to the session, one session can handle multiple requests. Request and permission are tied by many-to-many relationship. On top of requests Rq inclusion relation is defined.

**Definition 3.** Request A includes a request B (B ≤ A), if the path of a unique resource identifier (URI) of request A contains the path of the unique identifier of the resource request B with the initial position in the line within the same namespace, with \( \text{len}(B) \leq \text{len}(A) \), where \( \text{len}(x) \) – length of the string x.

\[
\begin{align*}
A / & \text{library} \\
B / & \text{library/category} \\
C / & \text{library/book} \\
D / & \text{library/category} \\
& \cdots
\end{align*}
\]

Fig. 1. An example of the inclusion relations on top of requests

Figure 1 shows the requests A, B, C, and D, which satisfy the following:

A ≤ B, B ≤ C, B ≤ D, A ≤ C, A ≤ D.

Inclusion relation has the following properties:

1. reflexivity:

\( \forall rq \in Rq : rq \leq rq, \)

2. antisymmetry:

\( \forall rq, rq' \in Rq : ((rq \leq rq') \& (rq' \leq rq)) \rightarrow rq = rq', \)

3. transitive:

\( \forall rq_1, rq_2, rq_3 \in Rq : ((rq_1 \leq rq_2) \& (rq_2 \leq rq_3)) \rightarrow rq_1 \leq rq_3. \)
Thus, the inclusion relation on top of requests set $Rq$ defines non-strict partial order.

Next, we define a function $RqA()$ mapping permissions to multiple requests $RqA: P \rightarrow 2^{Rq}$.

**Definition 4.** Requests hierarchy ($RqH$) – inclusion relation defined on top of requests $Rq$. For any $p \in P$ the following condition is true: if $rq, rq' \in Rq$, $rq \in RqA(p)$, and $rq \leq rq'$, then $rq' \in RqA(p)$.

Thus, the definition 4 makes it possible a flexible access control for individual requests, and all children requests.

Figure 2 shows a diagram of adapted security model elements. The model name is path-based role-based access control security model.

![Fig. 2. Elements of path-based RBAC model](image)

## 5 Adapting mandatory access model

In [9] for role-based security model authors describe use of mandatory access control designed to protect against threats to information confidentiality. Within defined terminology, we describe mandatory role-based access model to web applications. In addition to defined above elements, the following elements were added:

$Rq$ – set of requests;

$(L, \leq)$ – confidentiality levels lattice;

$c: U \rightarrow L$ – function of user access levels;

$c: Rq \rightarrow L$ – function of confidentiality levels for requests;
A = \{\text{read, write}\} – access types;
R = \{x \_ \text{read}\, | x \in L\} \cup \{x \_ \text{write}\, | x \in L\} – set of roles;
P = \{(rq, \text{read})\, | rq \in Rq\} \cup \{(rq, \text{write})\, | rq \in Rq\} – set of permissions.

Using the definitions 5.20 and 5.22 [10], according to requirements of liberal mandatory access control for set of requests Rq we define a hierarchy on top of roles R and restriction functions UA(), roles(), and PA().

As a part of the mandatory access control, information flow is defined.

**Definition 5.** We assume that there is an information flow from request rq ∈ Rq to request rq′ ∈ Rq if and only if there are roles r, r′ ∈ R, and session s ∈ S, such that (rq, read) ∈ PA(r), (rq′, write) ∈ PA(r′), and r, r′ ∈ roles(s).

Let’s formulate a proposition about the impossibility of forbidden information flows from the request with a higher confidentiality level to requests with a lower confidentiality level.

**Proposition 1.** If a role-based access model complies with liberal mandatory access control requirements, then for any requests rq, rq′ ∈ Rq, such that c(rq) > c(rq′), it is impossible to initiate information flow from rq to rq′.

The proof is similar to theorem 5.1 [10].

Thus, the model described is safe in terms of information flows for requests with different confidentiality levels.

### 6 Application of models

Security models described can be used in a wide range of applications. To apply these security models the system must meet the following requirements:

– centralized access control – access control is carried out only in a single module without delegating to other units or systems;
– principle of least privilege – provide the user only minimal set of privileges necessary for his work;
– separation of duties support – tasks processed in the system may require multiple users to process one operation;
– possibility of decomposition the system into separate components, which can be accessed using a variety of URIs, which are unique within the system.

When the above requirements are met, the system may be divided into different parts, each of them is uniquely identified by a URI. URI paths are described as access control elements and used to define a set of requests Rq. This set includes all client requests and API calls provided by the system. Using Rq request hierarchy is created that reflects the interaction and dependence between components.

In order to apply security model it is necessary to create a set of roles R. The role defines a set of permissions that a user can perform. Examples of roles: ”user”, ”moderator”, ”registrar”, ”administrator”.

Next, the system should have defined set of permissions P. Permissions define specific action or operation in the system, for example, ”create new user”, ”delete the document,” etc. One role can have many permissions. One permission can be
assigned to many roles. Assigning permissions for roles is performed by function \( PA() \). Permissions are non-overlapping and consistent. Elements from permissions \( P \) map to a subset of requests \( Rq \) using function \( RqA() \). One permission can have multiple requests. One request can be assigned to many permissions.

To maintain users a set of users \( U \) is created, each of which have assigned roles from \( R \) using mapping function \( UA() \). One user can have multiple roles. One role can be assigned to multiple users. To identify a user the model includes a set of tokens \( T_k \). The elements of the set are pairs of \( <\text{username, password}> \), or \( <\text{public key, private key}> \). Each user can have multiple tokens. The token belongs to one user.

Once authenticated, authorized work of users is carried out by sessions. A session is a set of authorized user data, including a set of roles to which the user is authorized. Users can create multiple sessions. A session belongs to one user. Sessions can have additional data related to authorization or specific operation.

As an example, we describe application of extended path-based RBAC for publication system. The system allows users to create articles for public reading. Registered users can create and edit their articles. Editors have the ability to edit articles created by users. The administrator has access to all sections, including user administration. The system has two users, who write articles: Alice and Bob. John is an editor, and Martin is a system administrator. In addition, Martin can work as editor. The system provides a special role for anonymous users for public reading without editing articles.

Users \( U \): \{Anonymous, Alice, Bob, John, Martin\}
Roles \( R \): \{Viewer, User, Editor, Administrator\}
Permissions \( P \): \{"view article", "create article", "edit own articles", "edit all articles", "user management", "access control", "system maintenance"\}
Requests \( Rq \): {
/articles/list,
/articles/view,
/manage/articles/list,
/manage/articles/create,
/manage/articles/edit,
/manage/users/list,
/manage/users/create,
/manage/users/edit,
/manage/permissions/roles,
/manage/permissions/acl,
/manage/system/settings,
/manage/system/maintenance\}
Roles assignment for users:
\( UA(\text{Anonymous}) \rightarrow \{\text{Viewer}\} \)
\( UA(\text{Alice}) \rightarrow \{\text{User}\} \)
\( UA(\text{Bob}) \rightarrow \{\text{User}\} \)
\( UA(\text{John}) \rightarrow \{\text{Editor}\} \)
\( UA(\text{Martin}) \rightarrow \{\text{Editor, Administrator}\} \)
Permissions assignment for roles:
P(A Viewer) → \{"view article"\}
P(A User) → \{"view article", "create article", "edit own article"\}
P(A Editor) → \{"view article", "create article", "edit all articles"\}
P(A Administrator) → \{"user management", "access control", "system maintenance"\}

Requests assignment for permissions:
RQA("view article") → \{/articles/list, /articles/view\}
RQA("create article") → \{/manage/articles/create\}
RQA("edit own article") → \{/manage/articles/edit\}
RQA("edit all article") → \{/manage/articles/edit\}
RQA("user management") → \{/manage/users\}
RQA("access control") → \{/manage/permissions\}
RQA("system maintenance") → \{/manage/system\}

As you can see from the example above, the developed models offer flexibility and simplicity for access control restriction that can be used in real-world web applications.

7 Conclusion

The paper describes extended path-based role access control model, which takes into account web applications features. New elements were defined: token, request, inclusion relation. The model created allows flexible access control for modern web applications.

Also extended path-based mandatory role-based access control model was created with additional sets: requests, security levels lattice, access types, roles, and permissions. Impossibility of forbidden information flows from higher to lower security levels was proven.

Authors created guidelines to apply model’s elements for real-world web applications. Developing web applications according to these guidelines allows reducing security risks. We use the model developed to enhance security in our web applications [11].

References


Model of Neuro-Fuzzy Prediction of Confirmation Timeout in a Mobile Ad Hoc Network

Igor Konstantinov, Kostiantyn Polishchuk, Sergej Lazarev, and Olha Polishchukova
Belgorod State University,
Pobeda Street 85, 308015 Belgorod, Russia
{konstantinov,polshchikov,lazarev,polshchikova}@bsu.edu.ru
http://www.bsu.edu.ru

Abstract. Confirmation timeout (Round Trip Time, RTT) is an important value in data networks. Correct prediction of this characteristic allows us to estimate the network load in order to adequately select packet sending and retransmission parameters. Approximate heuristic models are used in the Transmission Control Protocol (TCP) for Round Trip Time evaluation. The values of the coefficients in these models were obtained experimentally for fixed topology networks. Therefore, the use of these models in a dynamic topology network (mobile ad-hoc network) is inefficient. This article represents an RTT prediction model based on application of fuzzy neural network theory. This model relies upon zero-order Sugeno-Type Fuzzy Inference algorithm. The input values of fuzzy neural network are RTT values measured in the current and two previous cycles. The output value is RTT value expected in the next cycle. The proposed model is set up and examined by means of simulation experiments. In these experiments the functioning of mobile ad-hoc network which is used for communication software while counteracting emergencies was simulated.

Keywords: neuro-fuzzy prediction, round trip time, mobile ad hoc networks, dangerous construction sites.

1 Introduction

Mobile ad hoc networks (MANET) are a promising direction in the development of telecommunication technologies [1]. With its decentralized structure, ad hoc networks provide the ability to transmit information when nodes are moving randomly and under the impact of destructive factors [2,3]. Due to rapid deployment, autonomous power of each node, high survivability and the ability to deliver messages with dynamically changing topology, ad hoc network can be used for communication on dangerous construction sites [4]. Construction of dangerous objects is carried out under the threat of destructive and damaging
natural and man-made factors that can cause explosions, fire, collapse, flooding, radiation, poisoning and other emergencies.

The process of information exchange in a mobile ad hoc network is based on the implementation of the packet data transfer. One of the important characteristics in this case is RTT. Correct prediction of this value allows us to estimate the network load to adequately select the parameters of packet sending and retransmissions. Approximate heuristic models are used in the TCP for its evaluation [5, 6]. The values of the coefficients in these models were obtained experimentally for networks with a fixed topology, that is why their use in the ad hoc networks does not give the desired effect. As a result, time of information delivery increases significantly, which is unacceptable in the construction of dangerous buildings, as the life and health of builders, as well as the extent of damage to constructed facilities, depend on the operational efficiency of messages receiving in emergencies. Therefore, the development of an adequate RTT forecasting model in the mobile ad hoc network is a topical applied science problem, the solution of which is represented in the following researches.

2 Development of Neuro-Fuzzy Model

The neuro-fuzzy model is suggested to predict RTT in an ad hoc network. The following values are used in this model: $M$ is RTT value measured in the current cycle; $M^{pr1}$ is RTT value measured in the previous cycle; $M^{pr2}$ is RTT value measured in the cycle preceding the previous one. The model allows us to calculate the estimated value $\tilde{M}$ of the confirmation timeout for each of the next cycles.

Construction of the model is carried out on the criterion of minimal complexity. The following parameters correspond to this criterion: fuzzy inference algorithm is the zero-order Sugeno [7], the number of membership functions for each input value is 2, the shape of membership functions for each input value is triangular, neuronal learning algorithm is error propagation [8]. The model is using the following fuzzy rulebase:

\[ If(M = X_1) and (M^{pr1} = Y_1) and (M^{pr2} = Z_1), then(\tilde{M} = J_1); \quad (1) \]

\[ If(M = X_1) and (M^{pr1} = Y_1) and (M^{pr2} = Z_2), then(\tilde{M} = J_2); \quad (2) \]

\[ ... \]

\[ If(M = X_2) and (M^{pr1} = Y_2) and (M^{pr2} = Z_2), then(\tilde{M} = J_8); \quad (3) \]

where $X_1, X_2, Y_1, Y_2, Z_1, Z_2$ are terms number 1 and number 2 of the input values $M, M^{pr1}, M^{pr2}$; $J_1 ... J_8$ are individual conclusions of the fuzzy rules.

Type and parameters of the membership functions for each input value are shown in Fig. 1, Fig. 2 and Fig. 3.
Fig. 1. Type and parameters of the membership functions for the value $M$

Fig. 2. Type and parameters of the membership functions for the value $M^{pr1}$

Fig. 3. Type and parameters of the membership functions for the value $M^{pr2}$
The model of forecasting Round Trip Time includes four structural neural layers. Fuzzification procedure is performed by means of the first layer of neurons:

\[
\mu_1(M) = \begin{cases} 
1, & M < a_{x1}; \\
\frac{b_{x1} - M}{b_{x1} - a_{x1}}, & a_{x1} \leq M < b_{x1}; \\
0, & M \geq b_{x1}; 
\end{cases} 
\]  

\[
\mu_2(M) = \begin{cases} 
0, & M < a_{x2}; \\
\frac{M - a_{x2}}{b_{x2} - a_{x2}}, & a_{x2} \leq M < b_{x2}; \\
1, & M \geq b_{x2}; 
\end{cases} 
\]  

\[
\mu_1(M_{pr1}) = \begin{cases} 
1, & M_{pr1} < a_{y1}; \\
\frac{b_{y1} - M_{pr1}}{b_{y1} - a_{y1}}, & a_{y1} \leq M_{pr1} < b_{y1}; \\
0, & M_{pr1} \geq b_{y1}; 
\end{cases} 
\]  

\[
\mu_2(M_{pr1}) = \begin{cases} 
0, & M_{pr1} < a_{y2}; \\
\frac{M_{pr1} - a_{y2}}{b_{y2} - a_{y2}}, & a_{y2} \leq M_{pr1} < b_{y2}; \\
1, & M_{pr1} \geq b_{y2}; 
\end{cases} 
\]  

\[
\mu_1(M_{pr2}) = \begin{cases} 
1, & M_{pr2} < a_{z1}; \\
\frac{b_{z1} - M_{pr2}}{b_{z1} - a_{z1}}, & a_{z1} \leq M_{pr2} < b_{z1}; \\
0, & M_{pr2} \geq b_{z1}; 
\end{cases} 
\]  

\[
\mu_2(M_{pr2}) = \begin{cases} 
0, & M_{pr2} < a_{z2}; \\
\frac{M_{pr2} - a_{z2}}{b_{z2} - a_{z2}}, & a_{z2} \leq M_{pr2} < b_{z2}; \\
1, & M_{pr2} \geq b_{z2}. 
\end{cases} 
\]

Aggregation procedure is performed by the second layer of neurons:

\[
G_1 = \mu_1(M) \land \mu_1(M_{pr1}) \land \mu_1(M_{pr1}); 
\]  

\[
G_2 = \mu_1(M) \land \mu_1(M_{pr1}) \land \mu_2(M_{pr1}); 
\]  

\[
\ldots 
\]  

\[
G_8 = \mu_2(M) \land \mu_2(M_{pr1}) \land \mu_2(M_{pr1}). 
\]

Activation is a part of the defuzzification procedure. Calculation of the amount of aggregated results \(\sum_{r=1}^{8} G_r\) and the weighted sum of the aggregate results \(\sum_{r=1}^{8} J_r G_r\) are performed by means of the third layer of neurons.

The final part of defuzzification procedure is performed by means of the fourth layer:

\[
\tilde{M} = \frac{\sum_{r=1}^{8} J_r G_r}{\sum_{r=1}^{8} G_r}. 
\]

In order to obtain the coefficient values needed to calculate membership function, it is required to set the weights of neurons of the first layer. Training of the neurons of the third layer is needed for evaluating the values of individual fuzzy rules conclusions [9–15]. The receiving of training data for model setup and evaluation of neuro-fuzzy forecasting RTT are carried out on the basis of modeling of various scenarios of ad-hoc network application for communication on dangerous construction sites.
3 Modeling Information Streams Transmission

Let us consider an example in which mobile ad hoc network is used for communication in the construction of underground facilities. Fig. 4 and Fig. 5 show the area of the construction works (limited by bold dotted line).

![Diagram showing network topology and transmission routes]

**Fig. 4.** The routes of transmission of information flows in a fixed network topology

This building belongs to the dangerous construction projects, because works on its construction are carried out in the conditions of a possible collapse of rocks. Works are carried out by a personnel shift which consists of:

1) head of the shift who uses ad hoc node 1;
2) eight workers equipped with ad hoc nodes with numbers 2-9.

Ad hoc units are denoted by small numbered circles, and coverage areas of these units are limited by the corresponding circles of larger radius. The following functions are performed by means of ad hoc nodes:

1) video streams to monitor the status of the facility, conditions and the course of the work;
2) exchange of voice messages to control the construction process and the coordination of countering emergencies;
3) transfer of data on the functional status and current location coordinates of the builders, as well as data of monitoring external conditions on the construction site.

In the given example the transmission of information streams is carried out in an ad hoc network for a period of time of observation which lasts 50 seconds. The characteristics of the streams are represented in Table 1 and Table 2.

Fig. 4 shows the situation where the network topology remains unchanged during the considered time interval. The routes of information streams trans-
**Fig. 5.** The routes of transmission of information flows in a dynamic network topology

**Table 1.** Characteristics of the transmitted information streams

<table>
<thead>
<tr>
<th>Stream number</th>
<th>Type of transferred</th>
<th>Sending node number</th>
<th>Receiving node number</th>
<th>Transmission start time, s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>video</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>data</td>
<td>5</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>acknowledgements</td>
<td>1</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>data</td>
<td>9</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>acknowledgements</td>
<td>1</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>voice</td>
<td>7</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>voice</td>
<td>1</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>video</td>
<td>8</td>
<td>1</td>
<td>22</td>
</tr>
</tbody>
</table>

**Table 2.** Estimated parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Throughput of the radio channel</td>
<td>1000 Kbit/s</td>
</tr>
<tr>
<td>Throughput required to transmit video</td>
<td>256 Kbit/s</td>
</tr>
<tr>
<td>Throughput required to transmit voice</td>
<td>128 Kbit/s</td>
</tr>
<tr>
<td>Size of messages transmitted by data flow</td>
<td>1 MB</td>
</tr>
</tbody>
</table>
mission correspond to the broken lines which connect the nodes-senders and nodes-recipients.

Fig. 5 shows a scenario where an ad hoc network topology changes due to the collapse of rock, which began at time $t_c=4$ s. Collapse zone is highlighted in gray. As a result of emergency workers who used ad hoc nodes 6 and 9 were in the collapse zone and node 6 malfunctioned (corresponding circle in Fig. 5 is crossed).

In response to the collapse, workers with ad hoc nodes 3-5, 8 and 9 have moved. The locations of these nodes at the initial time in Fig. 5 are marked by the dashed circles. In the modified network structure in Fig. 5 routes that transmit information streams (numbered 2, 3, 5–8) differ from the corresponding streams marked in Fig. 4.

Dynamism of network topology had an impact on the radio channels workload and throughput available for transmission of data flows. For example, a radio channel connecting the node 2 to node 1, except for the main streams of 1, 4 and 6, additional streams 2 and 8 started to transmit.

The responsiveness of the node 1 receiving the data file transmitted from the node 9 is of great importance in an emergency. This file contains information about the current parameters of health status and location of the worker who has been exposed to the collapse. On the basis of the data head of the shift can quickly and effectively coordinate the actions of other workers to rescue the injured builder.

For the file to be delivered, streams 4 and 5 need to be transferred. The combination of these interrelated streams is called a controlled flow (CF) [16]. The closed circuit formed by the channels through which CF is transferred is called CF-circuit (Fig. 6).

![Fig. 6. CF-circuit](image)

Duration of data file delivery from node 9 to node 1 is directly dependent on the value $E(t)$, the current CF-circuit throughput available for CF transmission.
To calculate this value, one should use the expression:

\[ E(t) = \min\{E_k(t)\}, \quad (14) \]

where \( E_k(t) \) is the current value of the channel throughput \( k \) of the CF-circuit [17].

The value \( E_k(t) \) can be figured out from the formula:

\[ E_k(t) = \begin{cases} 0, & U_k(t) \geq c; \\ c - \frac{U_k(t)}{D_k(t)}, & U_k(t) < c; \end{cases} \quad (15) \]

where \( c \) is the throughput of the radio channel; \( U_k(t) \) is the current value of the channel throughput \( k \) required to transmit real-time streams; \( D_k(t) \) is the number of data streams, having to be transmitted over the channel by the time \( t \), \( D_k(t) \geq 1 \).

The value \( U_k(t) \) can be determined using the following expression:

\[ U_k(t) = \sum_{l=1}^{L} u_{kl}(t), \quad (16) \]

where \( u_{kl}(t) \) is the current value of the channel \( k \) throughput required for real-time stream transmission \( l \); \( L \) is the number of real-time streams, which need to be transmitted on the CF-circuit channels.

The value of \( u_{kl}(t) \) can be found from the formula:

\[ u_{kl}(t) = \begin{cases} \lambda_t a_{kl}, & x_l^{\text{start}} \leq t < x_l^{\text{stop}}; \\ 0, & t < x_l^{\text{start}} \text{ or } t \geq x_l^{\text{stop}}, \end{cases} \quad (17) \]

where \( \lambda_t \) is the value of the bandwidth of the channel \( k \) required to transmit real-time stream \( l \) [18]; \( a_{kl} \) is the value showing whether the transmission channel is required on real-time stream \( l \) channel \( k \); \( x_l^{\text{start}} \) and \( x_l^{\text{stop}} \) are instants of the beginning and the end of transmission of real-time stream \( l \).

The minimum possible duration of the CF transmission can be determined using the following formula:

\[ \tau_{CF} = \tau_{CF}^{\text{stop}} - \tau_{CF}^{\text{start}}, \quad (18) \]

where \( \tau_{CF}^{\text{start}} \) is starting time of CF transmission; \( \tau_{CF}^{\text{stop}} \) is closure time of CF-stream transmission without packet loss and an ideal correspondence between the intensity of sending data of the stream and the bandwidth of CF-circuit available for the transmission.

The value \( \tau_{CF}^{\text{stop}} \) is calculated on the basis of the obtained values \( E(t) \). To do this, use the formula:

\[ V = \int_{\tau_{CF}^{\text{start}}}^{\tau_{CF}^{\text{stop}}} E(t) \, dt, \quad (19) \]

where \( V \) is the size of the message transmitted by data flow.
Table 3. $\lambda_l, x_{l}^{\text{start}}$ and $x_{l}^{\text{stop}}$ values

<table>
<thead>
<tr>
<th>$l$</th>
<th>$\lambda_l$, bit/s</th>
<th>$x_{l}^{\text{start}}$, s</th>
<th>$x_{l}^{\text{stop}}$, s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>256</td>
<td>0</td>
<td>$&gt;$50</td>
</tr>
<tr>
<td>2</td>
<td>128</td>
<td>16</td>
<td>$&gt;$50</td>
</tr>
<tr>
<td>3</td>
<td>128</td>
<td>16</td>
<td>$&gt;$50</td>
</tr>
<tr>
<td>4</td>
<td>256</td>
<td>22</td>
<td>$&gt;$50</td>
</tr>
</tbody>
</table>

Table 4. $a_{kl}$ values

<table>
<thead>
<tr>
<th>$k$</th>
<th>$l=1$</th>
<th>$l=2$</th>
<th>$l=3$</th>
<th>$l=4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 7. The current values $E(t)$ in a network with a dynamic topology

Fig. 8. The current values $E(t)$ in fixed topology network
To calculate the function $E(t)$ in the case shown in Fig. 4 and Fig. 5, we used data contained in Table 3 and Table 4.

Using these inputs, the function $E(t)$ is calculated and its form is shown in Fig. 7. The fixed network topology function $E(t)$ has the form shown in Fig. 8.

Analysis of Fig. 7 and Fig. 8 shows that the change in the network topology during information exchange leads to a significant deceleration of data file transmission duration. In a network with a dynamic topology the value $\tau_{CF}$ is set to 39.7 s, and in case of a fixed network structure it is $\tau_{CF}=24.2$ s.

4 Setting Parameters of the Model and the Evaluation of the Effectiveness of Its Application

In real operating conditions an ad hoc network overload and packet loss frequently occur, so the actual value of the data file transfer duration can significantly exceed the calculated value $\tau_{CF}$. To evaluate these characteristics a number of simulation experiments were made, in which various scenarios of applying an ad hoc network for providing connectivity on dangerous construction sites were simulated. For this purpose, a simulation model of information streams transmissions in a network with dynamic topology was used. It was developed in MatLab Simulink software environment. The simulation results provided evidence for setting developed neuro-fuzzy model forecasting round trip time. On the basis of these data the training matrix of the following form is made:

$$
\begin{pmatrix}
M_1 & M_2 & M_3 & M_4 \\
M_2 & M_3 & M_4 & M_5 \\
\vdots & \vdots & \vdots & \vdots \\
M_i & M_{(i+1)} & M_{(i+2)} & M_{(i+3)} \\
M_{(I-3)} & M_{(I-2)} & M_{(I-1)} & M_I
\end{pmatrix}
$$

where $M_i$ is a round trip time of confirmation in the loop $i$; $I$ is the number of cycles in each simulation experiment, $I=750$.

Setting neuro-fuzzy model was carried out using software tools Fuzzy Logic Toolbox. Table 5 shows the results of training the neurons of the first layer, while Table 6 contains the results of training the neurons of the third layer.

To assess the efficiency of the developed and customized models, a number of simulations for the transfer of information streams in an ad hoc network were conducted. The selection of retransmission was simulated on the basis of the suggested neuro-fuzzy forecasting RTT and the classical model of evaluation of this quantity used in TCP. The results showed that the use of neuro-fuzzy forecasting RTT in a large network load reduces deviations of timeout retransmission on 5.7-19.2 percents. This contributes to minimizing retransmissions count and average data stream transmission time by 4.2-9.6 percents.
Table 5. Learning outcomes of the first layer of neurons

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{x1}$</td>
<td>3.64</td>
</tr>
<tr>
<td>$a_{x2}$</td>
<td>25.18</td>
</tr>
<tr>
<td>$b_{x1}$</td>
<td>3.62</td>
</tr>
<tr>
<td>$b_{x2}$</td>
<td>27.90</td>
</tr>
<tr>
<td>$a_{y1}$</td>
<td>3.69</td>
</tr>
<tr>
<td>$a_{y2}$</td>
<td>28.10</td>
</tr>
<tr>
<td>$b_{y1}$</td>
<td>3.55</td>
</tr>
<tr>
<td>$b_{y2}$</td>
<td>27.79</td>
</tr>
<tr>
<td>$a_{z1}$</td>
<td>3.61</td>
</tr>
<tr>
<td>$a_{z2}$</td>
<td>28.01</td>
</tr>
<tr>
<td>$b_{z1}$</td>
<td>3.62</td>
</tr>
<tr>
<td>$b_{z2}$</td>
<td>27.81</td>
</tr>
</tbody>
</table>

Table 6. Learning outcomes of the third layer of neurons

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1$</td>
<td>3.91</td>
</tr>
<tr>
<td>$H_2$</td>
<td>-6.02</td>
</tr>
<tr>
<td>$H_3$</td>
<td>7.26</td>
</tr>
<tr>
<td>$H_4$</td>
<td>8.74</td>
</tr>
<tr>
<td>$H_5$</td>
<td>31.51</td>
</tr>
<tr>
<td>$H_6$</td>
<td>20.74</td>
</tr>
<tr>
<td>$H_7$</td>
<td>27.92</td>
</tr>
<tr>
<td>$H_8$</td>
<td>26.49</td>
</tr>
</tbody>
</table>

5 Conclusion

Thus, the model of neuro-fuzzy prediction of confirmation timeout in the mobile ad hoc network is synthesized. The model includes four neuron layers, performing fuzzy inference procedure (fuzzification, aggregation, revitalization and defuzzification). To adjust the weights neurons we used training data, reflecting the dynamics of the RTT in the ad hoc network used for communication on dangerous construction sites. Simulations have shown that the use of the proposed model for selecting timeout retransmission will significantly reduce the duration of the transmission data flows in the mobile ad hoc network.

Acknowledgements. The research have been carried out with the financial support of the Ministry of Education and Science of the Russian Federation (the unique identifier of the project is RFMEFI57815X0138).
References


Implementation of Weibull’s Model for Determination of Aircraft’s Parts Reliability and Spare Parts Forecast

Nataša Kontrec, Milena Petrović, Jelena Vujaković, and Hranislav Milošević

University of Priština, Faculty of Natural Sciences and Mathematics, Kosovo Mitrovica, Serbia
natasa.kontrec@pr.ac.rs

Abstract. Planning of aircraft’s maintenance activities, failure occurrences and necessary spare parts are essential for minimizing downtime, costs and preventing accidents. The aim of this paper is to propose an approach that supports decision making process in planning of aircraft’s maintenance activities and required spare parts. Presented mathematical model is based on Weibull’s model and calculates aircraft’s reliability characteristics by using data on previous failure times of an aircraft part. Further, by capitalizing the random nature of failure time, the number of spare parts and the costs of negative inventory level are determined.

Keywords: aircraft’s spare parts, reliability, forecast, Weibull’s model.

1 Introduction

Optimized maintenance can be used as a key factor in organization’s efficiency and effectiveness. Maintenance in aviation industry requires replacing of parts to assure aircraft availability. Aviation companies are often facing aircraft’s downtime due to spare parts shortage because they simply follow manufacturers’ or suppliers’ recommendation regarding the required number of spare parts to be kept on inventory [1]. Furthermore, that leads to unexpected costs of urgent orders or the passenger accommodation costs in case of flight cancellation, etc. Adequate spare parts management in the aircraft maintenance system improves the aircraft availability and reduces downtime. Spare parts forecasting and provisioning is a complex process and there are numerous paper dealing with this issue [2–6]. In aviation industry some methods described in papers [7–11] found their application but due to stochastic nature of demand they often failed to provide accurate results. In recent times, spare parts forecasting with respect to techno-economical issues (reliability, maintainability, life cycle costs) have been studied [12–14] but not that extensively in aviation industry. In [15] a methodology to forecast the needs for expendable or non-repairable aircraft parts has been presented. That methodology was based on observing total unit time (Tut) provided by manufacturer as stochastic process. In the case when parameter (Tut) is not available, we herewith present a new approach for determination
of spare parts requirements. Described approach relies on historical data of previous failure times of an aircraft part and their stochastic nature. In order to determine the reliability characteristic of each aircraft part, the Weibull’s model has been used. The Weibull’s probability density function (PDF) is given by:

\[ f(w) = \frac{\beta}{\eta} \left(\frac{w}{\eta}\right)^{\beta-1} \exp\left(-\left(\frac{w}{\eta}\right)^\beta\right), \quad f(w) \geq 0, \quad w \geq 0, \quad \beta > 0, \quad \eta > 0, \quad (1) \]

where \( w \) denotes flight hours, \( \beta \) denotes shape parameter or slope, \( \eta \) denotes scale parameter or characteristic life. Based on previous, the cumulative distributive function (CDF) can be determined as given in eq. (2):

\[ F(w) = 1 - \exp\left(-\frac{w}{\eta}\right)^\beta. \quad (2) \]

Further, reliability function of Weibull’s model can be calculated as follows:

\[ R(w) = \exp\left(-\frac{w}{\eta}\right)^\beta. \quad (3) \]

Also, there is a possibility to calculate the conditional reliability i.e. the reliability for the additional period of \( w \) duration for the parts having already accumulated \( W \) flight hours. It can be calculated as given in eq. (4):

\[ R(w|W) = \frac{R(W + w)}{R(W)} = \frac{\exp\left(-\frac{W+w}{\eta}\right)^\beta}{\exp\left(-\frac{W}{\eta}\right)^\beta} = \exp\left(-\frac{(W+w)}{\eta}\beta - \frac{(W)}{\eta}\beta\right). \quad (4) \]

The mean time to failure (MTTF) of Weibull’s PDF can be determined as in eq. (5):

\[ \text{MTTF} = \eta \cdot \Gamma\left(\frac{1}{\beta} + 1\right), \quad (5) \]

where \( \Gamma \) is Gamma function. Failure rate function is given in eq. (6):

\[ \lambda(w) = \frac{f(w)}{R(w)} = \frac{\beta}{\eta} \left(\frac{w}{\eta}\right)^{\beta-1}. \quad (6) \]

In order to calculate reliability characteristic of an aircraft part it is necessary to estimate the parameters of Weibull’s model. There are several ways to achieve that, but in the case when we have limited historical data on previous failures, it is best to perform rank regression on \( Y \) [16]. Rank regression on \( Y \) is a method based on the least squares regression principle, which minimizes the vertical distance between the data points and the straight line fitted to the data as presented in Fig. 1. The idea is to bring our function to linear line. In order to achieve that we are taking natural logarithm of the both sides of the eq. (2).

\[ \ln[1 - F(w)] = \ln[\exp(-w/\eta)^\beta] \]

\[ \ln[-\ln[1 - F(w)]] = \beta \ln(w/\eta) \]

\[ \ln[-\ln[1 - F(w)]] = \beta \ln w - \beta \ln \eta. \]
Then by setting:

\[ y = \ln[-\ln(1 - F(w))] \quad x = \ln w \quad a = \beta \quad \text{and} \quad b = -\beta \ln \eta. \]

the previous equation can be rewritten as \( y = ax + b \). Now, assume that we have sample of failure data set as \( (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \) plotted and \( x \) values are predictor variables. According to least square principle, the straight line that best fit to these data is \( y = \hat{a} + \hat{b}x \), such that:

\[
\sum_{i=1}^{N} (\hat{a} + \hat{b}x_i - y_i)^2 = \min \sum_{i=1}^{N} (\hat{a} + \hat{b}x_i - y_i)^2
\]

where \( \hat{a} \) and \( \hat{b} \) are the least squares estimates of \( a \) and \( b \) and \( N \) is the number of failure data. The equations can be minimized by estimates \( \hat{a} \) and \( \hat{b} \) as in Eqs. (7) and (8)

\[
\hat{b} = \frac{\sum_{i=1}^{N} x_i y_i - \frac{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N}}{\sum_{i=1}^{N} x_i^2 - \left( \frac{\sum_{i=1}^{N} x_i}{N} \right)^2}.
\]

and

\[
\hat{a} = \frac{\sum_{i=1}^{N} y_i - \hat{b} \sum_{i=1}^{N} x_i}{N} = \bar{y} - \hat{b} \bar{x},
\]

The variable \( \bar{y} \) is the mean of all the observed values and \( \bar{x} \) is the mean of all values of the predictor variable at which the observations were taken. Now, according to the previous, we can easily obtain \( y_i \) and \( x_i \)

\[
y_i = \ln[-\ln(1 - F(w_i))], \quad x_i = \ln(w_i).
\]

The \( F(w_i) \) are values determined from the median ranks, and after we calculate \( \hat{a} \) and \( \hat{b} \), we can easily estimate parameters \( \eta \) and \( \beta \).

2 Numerical analysis

According to the previous formulas we can further perform numerical analysis on sample of 14 failure-time data for aircraft part number 302634-2 (Igniter plugs for aircraft Cessna Citation 560XL - provided by Prince Aviation Company, Serbia). Data are sorted by ascending order and presented in Table 1.

First, it was concluded by using Weibull’s probability plotting that data are following Weibull’s distribution, as can be seen in Fig. 1. Since the table provide the sample size less than 15 failed times, rank regression on \( Y \) method, presented in previous section, has been used for parameter estimation. We applied this method since it has been considered as more accurate [16]. It has been calculated
Table 1. Failure time (flight hours for part no. 302634-2 Igniter Plug)

<table>
<thead>
<tr>
<th>No. of part</th>
<th>Failure time (flight hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3258</td>
</tr>
<tr>
<td>2</td>
<td>4321</td>
</tr>
<tr>
<td>3</td>
<td>5183</td>
</tr>
<tr>
<td>4</td>
<td>5223</td>
</tr>
<tr>
<td>5</td>
<td>5786</td>
</tr>
<tr>
<td>6</td>
<td>5920</td>
</tr>
<tr>
<td>7</td>
<td>6004</td>
</tr>
<tr>
<td>8</td>
<td>6321</td>
</tr>
<tr>
<td>9</td>
<td>6550</td>
</tr>
<tr>
<td>10</td>
<td>6893</td>
</tr>
<tr>
<td>11</td>
<td>6906</td>
</tr>
<tr>
<td>12</td>
<td>7221</td>
</tr>
<tr>
<td>13</td>
<td>7305</td>
</tr>
<tr>
<td>14</td>
<td>7400</td>
</tr>
</tbody>
</table>

that shape parameter ($\beta$) is 4.86 and characteristic life ($\eta$) is 6,572.98. According to the previous conclusions and eq. (3), we further determined reliability function of the part Igniter plug. Reliability of the part Igniter plug is given in Fig. 2 and the failure rate is presented in Fig. 3.

According to these figures we can conclude after how many flight hours this part would most likely stop working.
3 Method evaluation

The major contribution of this paper is to determine the number of spare parts that should be kept on stock in interval $[0, w]$. In order to achieve that we are using an approach presented in paper [15] where the number of part exposed...
to failure in certain time frame was calculated. These calculation are based
on Rayleigh’s model in the case when only total unit time (usually provided
by parts manufacturer) is available. Similar approach is applied in this paper
but in the case when data of previous failures are available so the reliability
characteristics of the aircraft parts are determined by using the Weibull’s model.

PDF of Weibull’s distributed failure time is given by eq. (1), while the PDF of
Rayleigh’s distribute failure time is:

\[ f(\mu) = \frac{\mu}{\sigma^2} \exp \left( -\frac{\mu^2}{2\sigma^2} \right). \]  \hspace{1cm} (10)

In the eq. (10), \( \mu \) presents Rayleigh’s random variable, while the PDF of Weibull’s
model has been given by eq. (1). According to the above stated equations it can
be concluded that \( \sigma = \eta/\sqrt{2} \) and \( \mu = \frac{\beta}{2}. \)

In order to create relation between these models we are using the following
transformation:

\[ p_{w\dot{w}}(w, \dot{w}) = p_{\mu\dot{\mu}}(w^{\frac{\beta}{2}}, \dot{w}^{\beta} w^{\frac{\beta}{2} - 1}) |J|, \]  \hspace{1cm} (11)

where \( |J| \) presents Jacobian transformation of random variables given by the
following equation:

\[ |J| = \left| \frac{d\mu}{dw} \frac{d\mu}{d\dot{w}} \frac{d\dot{\mu}}{dw} \right| = \frac{\beta^2}{4} w^{\beta - 2}. \]

So, the eq. (11) further transforms into:

\[ p_{w\dot{w}}(w, \dot{w}) = \frac{\beta^2}{4} w^{\beta - 2} p_{\mu\dot{\mu}}(\mu, \dot{\mu}). \]

Based on the random nature of failure time of an aircraft part, we are observing
the expected number of variations of Rayleigh’s random variable \( \mu \) within an
interval \( (\mu, \mu + d\mu) \), for a given slope \( \dot{\mu} \) within a specified open neighborhood \( d\mu \).
Actually, \( \dot{\mu} \) is a gradient of Rayleigh’s random variable, while \( \dot{w} \) is gradient of
Weibull’s random variable. The number of parts that will be exposed to failure
can be determined as:

\[ n = \int_{0}^{\infty} \mu p_{\mu\dot{\mu}}(\mu, \dot{\mu}) d\dot{\mu} = \int_{0}^{\infty} \dot{\mu} \frac{\mu}{\sigma^2} \exp \left( -\frac{\mu^2}{2\sigma^2} \right) \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left( -\frac{\dot{\mu}^2}{2\sigma^2} \right) d\mu \]

\[ = \int_{0}^{\infty} \dot{w} p_{w\dot{w}}(w, \dot{w}) d\dot{w}. \]

According to the previous equations, the number of spare parts exposed to failure
in time \( w \) can be finally determined as:

\[ n = \frac{4\sqrt{2} w^{\frac{\beta}{2}}}{\eta} exp \left( -\frac{w^{\beta}}{\eta^2} \right). \]
After we calculated average number of parts that are exposed to failure in interval $[0, w]$, we can determine the number of parts that should be on inventory. We are using the approach presented in paper [15] where we observed the expected amount of time when random variable $w$ is below total unit time as quotient of Rayleighs CDF and $n$. Since the characteristic life parameter of Weibull’s distribution $\eta$ is the time at which 63.2% of the units will fail and it is approximately equal to MTTF [17], in this case, we are assessing the amount of time when $w$ is below $\eta$ by dividing CDF function of Weibull’s distributed variable and the average number of parts to fail in time interval $[0, w]$ as:

$$q = \frac{F(w)}{n}.$$  \hfill (12)

As presented in Fig. 4 for the part Igniter plug it can be concluded at what time the spare part should be available. In the case that this part is not available when needed, the underage costs appear. The underage costs are difficult to determine due to their nature. Also, in this paper we are using the well known Newsvendor method [18] in order to calculate these costs. This method gives good results when it is necessary to estimate a stochastic variable. The result of this estimation is a compromise between losses when we decide to order more spare parts than needed and losses when we order less than required. In both cases we have costs, either unnecessary inventory costs or costs of urgent orders. Newsvendor method should provide optimal quantity of spare parts. Since we determined that number in eq. (12), we are using the following formula to

![Fig. 4. Number of spare parts for part Igniter plug.](image)
calculate the underage costs:

\[ q = \Phi^{-1}\left( \frac{c_u}{c_u + c_o} \right), \]

where \( \Phi^{-1} \) presents inverse distribution function (complementary error function), \( c_u \) are underage costs and \( c_o \) are overage costs, which in our case is the spare part price. Fig. 5 presented the underage costs for aircraft part Ignition plug. The overage costs for this part are are $1.925,00 and it can be noticed that the underage costs are growing exponentially in relation to time.

![Graph showing Underage costs over Flight hours](image)

**Fig. 5.** Underage cost of the part Igniter plug.

## 4 Conclusion

This paper presents an approach to determine reliability parameters of each aircraft part. This has been achieved by using the observed failure times for certain aircraft part and Weibull’s model. Also, a new methodology for calculation of parts that are exposed to failure in observed period of time is presented. This approach was based on random nature of failure or total unit time of each aircraft’s part. According to the obtained number, we further calculated the quantity of the aircraft spare parts that should be kept on stock in order to avoid necessary costs. Also, the Newsvendor model was used in order to assess the potential underage costs in certain time period. All these calculations aim to support the decision making process in planning of aircraft maintenance activities and spare parts needs. As presented in the paper, we evaluated the method for one specific aircraft part and presented results. Same could be done for any other aircraft part.
part. Also, these analysis could be applied to other industries with no massive production of spare parts such as weapons industry.

References

Abstract. The aim of this work is to create a publicly available database with webometric indicators for research and higher education organizations in Russia updated on monthly schedule accessible through the project's website http://www.webometrix.ru. This paper describes the set-up of the project including initial gathering and actualization of organizations and their web domains list, sources of data, measuring of indicators' values, analytics available on projects website. Starting from 613 institutions of Russian academies of sciences in January 2015 from the middle of 2015 we gather data for 2201 organizations including research and higher education institutions. Continuous data for more than a year allowed us to assess the reliability of indicators used and to draw some conclusions about Russian scientific and educational web space.

Keywords: webometrics, informetrics, websites, rankings.

1 Introduction

In recent years, webometric studies based on the web search engine usage [1][2][3], become a recognized method of measurement of academic institutions websites quality and impact. However, as other informetric studies, this method remains quite controversial due to gaps in research base, quality of measurement instrumentation and weight and meaning of measured indicators. We suppose that there is a lack of nationwide recurring measurements and juxtaposition between webometric and other kinds of scientometric assessments.

This research is based on monthly webometric data collection for websites of over than 2200 Russian research organizations and higher education institutions. Deep analysis of time series of measurements of particular webometric indicators in some cases combined with the parsing of website structure, examining its peculiarities, and correlating with usage statistics [4] allowed us to examine in details the significance of each of indicators, propose some justification methods and to compare different approaches to calculation of webometrics rankings.
In this report, we consider the principles and architecture of the webometric data collecting system. It contains a webometric indicators database with monthly data since January 2015, and web interface (http://www.webometrix.ru), that allows anyone to perform an analysis of trends and evolution of the scientific websites. Institutions can use it to examine the position and dynamics of their website, to compare it with the others and, as a result, to find the ways of its improvement. In the final part of the report, we made an overview of the Russian Academic and Education Web. We also compared webometric rankings with bibliometric data on institutions' academic output and website usage statistics.

2 Area of Study

We began data collection in January 2015 for institutions under the supervision of Russian Federal Agency of Scientific Organizations (so-called academic institutions, since before they were subject to state academies). Official websites and corresponding distinct domains were determined for 613 organizations, most of them being the research institutions. In July 2015 we have added Russian higher and further education institutions, non-academic research organizations and scientific development and production centers to data collection.

The resulting collection contains data for 2201 organization with distinct DNS domains which also include some regional branches with separate websites and corresponding domains. Of these, 1172 organizations are in the research and development segment and 1029 – in the higher education segment.

At the time of the initial information gathering we could not find a consistent and complete list of such organizations and therefore the Scientific Digital Library database combined with Russian Science Citation Index (RSCI) located at http://elibrary.ru was used. Since eLibrary.ru database covers most of the scientific publications of the Russian authors, we assume that it refers almost all organizations with employees that are somehow engaged in scientific activities. Corresponding website addresses were obtained via Google and Yandex search by title with visual verification.

According to Russian State Statistics Agency, there were 2827 research and development organizations in Russia at the end of 2015. However, some of them do not maintain an official website or publish scientific works either due to the restricted field of research or pure technological and construction character of its activities. In higher education segment there are 609 state institutions and universities and 437 private ones with the total of 1046.

Among higher education institutions we allocated well separated classes associated with both the tendencies of development of the Russian higher education in recent years – the federal and national research universities, and with the legacy of Soviet Union – classical, technical, medical, humanitarian, educational, economic, legal and agricultural universities. In R&D segment we allocated the institutions under the control of Federal Agency of Research Organizations (academic institutions) and national research centers, as these classes are funded under government programs in a special way. The rest of the research institu-
Organizations have a substantially different scale and scientific activity. Unfortunately, detailed data on the number of researchers and faculty members at institutions are not available, so we extracted a number of contributing authors for each organization that has publications in 5 recent years. These data do not fully reflect the organization’s research staff, however, these numbers allows us to make an adequate assessment. For large research institutions and universities, the number of contributing authors may exceed the number of actual faculty members and research staff because of temporary employees and students. Thus, the Moscow State University has about 9000 faculty members while eLibrary counts 14211 contributing authors. And vice versa for small organizations and universities the number of authors may be less than faculty members and researchers. The resulting treemap is shown in Fig. 1.

**Fig. 1.** Russian Higher Education and R&D segments treemap by the number of authors were published in 5 recent years (2011-2015), in thousands. Hereinafter: HE Higher Education segment: FU Federal Universities, NRU National research Universities, CU Classic Universities, TU Technical Universities, EU Economics Universities, MS Medical Schools, PU Pedagogical Universities, HU Humanitarian universities, LS Law Schools, AU Agricultural Universities. R&D Research and Development segment: AI Academic Institutions (under the control of Federal Agency of Scientific Organizations), NRC National Research Centers
Also, from the same source data the number of articles published in the last 5 years and registered in Web of Science and Scopus databases have been extracted (Fig. 2). Detailed data are also shown in Table. 1.

Fig. 2. Russian Higher Education and R&D segments treemap by the number of publications published in 5 recent years (2011-2015) and registered in Web of Science or Scopus databases, in thousands.

3 Metrics and Data Sources

As in [2] we consider those main metrics for our study:

1. Domain size that is measured as a number of hits in corresponding search engine for a request limited by domain URL.
2. A number of documents in popular formats ps, pdf, doc(x), ppt(x), xls(x) measured by narrowing previous request by the appropriate filter.
3. A number of scholarly papers indexed by Google Scholar in which can be full-text documents or web pages with the correct publication metadata.
4. A number of external references hyperlinks from other domains to the target domain pages (and a number of such domains).

The first two metrics are collected via Google, Bing and Yandex search engines and the last one with the help of Ahrefs and Majestic SEO search optimization and backlinks tracking services. Additionally, we obtain basic usage statistics
Table 1. Decomposition of Russian Higher Education and R&D segments

<table>
<thead>
<tr>
<th>segment / Class</th>
<th>Number of organizations</th>
<th>Percentage of authors</th>
<th>Percentage of papers in RSCI</th>
<th>Percentage of papers in WoS / Scopus</th>
</tr>
</thead>
<tbody>
<tr>
<td>HE</td>
<td>1029</td>
<td>74.8%</td>
<td>80.0%</td>
<td>50.3%</td>
</tr>
<tr>
<td>FU</td>
<td>10</td>
<td>4.9%</td>
<td>5.2%</td>
<td>5.4%</td>
</tr>
<tr>
<td>NRU</td>
<td>29</td>
<td>11.0%</td>
<td>10.2%</td>
<td>12.0%</td>
</tr>
<tr>
<td>CU</td>
<td>72</td>
<td>14.7%</td>
<td>15.8%</td>
<td>13.7%</td>
</tr>
<tr>
<td>TU</td>
<td>131</td>
<td>13.7%</td>
<td>14.5%</td>
<td>7.1%</td>
</tr>
<tr>
<td>EU</td>
<td>85</td>
<td>5.3%</td>
<td>6.7%</td>
<td>0.7%</td>
</tr>
<tr>
<td>HU</td>
<td>61</td>
<td>2.6%</td>
<td>3.3%</td>
<td>1.0%</td>
</tr>
<tr>
<td>MS</td>
<td>50</td>
<td>6.8%</td>
<td>5.5%</td>
<td>6.0%</td>
</tr>
<tr>
<td>PU</td>
<td>48</td>
<td>4.3%</td>
<td>5.0%</td>
<td>1.3%</td>
</tr>
<tr>
<td>LS</td>
<td>22</td>
<td>1.3%</td>
<td>1.6%</td>
<td>0.1%</td>
</tr>
<tr>
<td>AU</td>
<td>49</td>
<td>4.1%</td>
<td>5.0%</td>
<td>0.5%</td>
</tr>
<tr>
<td>others</td>
<td>472</td>
<td>6.2%</td>
<td>7.2%</td>
<td>2.6%</td>
</tr>
<tr>
<td>R&amp;D</td>
<td>1172</td>
<td>25.2%</td>
<td>20.0%</td>
<td>49.7%</td>
</tr>
<tr>
<td>AI</td>
<td>613</td>
<td>17.7%</td>
<td>15.0%</td>
<td>39.1%</td>
</tr>
<tr>
<td>NRC</td>
<td>35</td>
<td>1.5%</td>
<td>0.8%</td>
<td>3.3%</td>
</tr>
<tr>
<td>others</td>
<td>524</td>
<td>6.1%</td>
<td>4.2%</td>
<td>7.2%</td>
</tr>
<tr>
<td>Total</td>
<td>2201</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

from SimilarWeb service, that uses data extracted from four main sources: 1) a panel of web surfers made of millions of anonymous users equipped with a portfolio of apps, browser plugins, desktop extensions, and software; 2) global and local ISPs; 3) web traffic directly measured from a learning set of selected websites and intended for specialized estimation algorithms; 4) A colony of web crawlers that scan the entire Web. Comparison of this data to the data, gathered from corresponding Google Analytics and Yandex Metrika site counters for several academic websites, participated in our research [4] shows its sufficient accuracy.

4 Data Collection and Processing Pipeline

Data are collected by PowerShell scripts that request Yandex and Bing web services and process web pages obtained from Google, Google Scholar and SimilarWeb using Internet Explorer automation with operators visual control. Ahrefs and Majestic data are collected by their bulk export features. Results are stored in the MongoDB database.

During data collection we met with 2 types of errors and failures mainly in interpreting of web pages received as a response to a query: a) errors caused by wrong response parsing due to unexpected changes in response details, b) changes in search engine’s database caused by reindexing of web sites and global index rebuilding.
The first type of distortions may be detected by low and high pass filter, which compares the measured value with the average of several previous values. Such combination of band filter with moving average allows us to detect single isolated outliers. If errors were detected after the data collection cycle and cannot be corrected by recollection, data may be recovered by linear interpolation of neighbor values. We retain originally collected values also.

The situation is much worse with effects caused by reindexing of some websites and global rebuilding of search engines' indices that occur relatively often [5] and affect measurements dramatically. For example, in July 2015 Bing counted 2.8 millions of pages in the domain of Institute of Astronomy of the Russian Academy of Sciences (inasan.ru), but in subsequent months this indicator falls back to several thousand of pages. We analyzed these effects in details in [6]. These effects may last more than one month and so we need more sophisticated logic to determine and justify them and it is one of directions of future investigations.

For each domain 26 metrics are gathered with 9 main indicators and 17 supplementary. We retain exact timestamp for each value collected. The monthly cycle lasts for more than a week resulting in 57 226 values. The projects website provides the following basic functionality:

- Ranking of organizations web domains by one of the indicators and its change in time in the tabular form.
- Ranking of organizations web domains by every indicator values for the single month.
- Dynamics of totals, means and medians of selected indicators for all or selected part of domains during selected time period.
- Comparison of different series indicator month as a scatter chart.
- Detailed info for a single domain.

5 Analysis of Research and Education Web Space

Data collected for the majority of Russian research and higher education organizations for a period of 10 months allow us to make some brief review of basic characteristics of Russian research and higher education web space in terms of size and quality. By size we mean a number of pages and documents and by quality a number of papers in Google Scholar index, Yandex thematic citation index, and a number of visits per month. As we show in [6] backlinks data is quite controversial and cannot be used in thorough analysis without complex cleaning.

First of all, let us look at the size of Russian scientific and education web and its dynamics (Fig. 3). At first sight, we can deduce that total size of the segment under consideration increases with the exclusion of Bing data, that can be justified by some Bing engine peculiarities. But if we take into account the total size of Russian web space which can be measured as the .ru zone size we can see a more complicated situation (Fig. 4). While the size of overall Russian web in Google index increases in more than 150% and in Bing nearly doubles,
Yandex oscillates near 400 millions of pages. At the same time, academic share in Bing index decreased from 18% to 7%, in Google remains almost the same (7% to 6%) but in Yandex it tripled from 6% to 18%.

![Graph showing number of pages](image)

**Fig. 3.** Total number of pages in Russian research and education web space by month

While considering indicators based on search engines we will use aggregated values calculated as the maximum of values, obtained from Google, Yandex and Bing onwards. The total size of the Education and Scientific web measured by this indicator has grown from 78 million pages in August 2015 to more than 100 million in May 2016. The share of HE decreased by almost 3% from 74.3% to 71.6%. The fastest growing classes were TU, CU and NRU. The share of R&D is divided almost equally between AI and other R&D organizations (14%) with a minor proportion of NRC (0.3%). During the studied period the proportion of AI decreased from 17% with a simultaneous growth of others from 8%. Web space partition in classes is shown on Fig. 5.

HE average domain size increased from 111 to 146 thousand of pages, R&D - from 14 to 19.5 thousand, the most active growth of average domain size from 303 to 525 thousand of pages was in FU and NRU domains grew from 272 to 371 thousand of pages. The average size of domains LS, EU, MS and PU has not changed, and the HU and AU even decreased. In R&D the average size of AI and NRC domains have not changed, the increase was only observed among others.
Fig. 4. Total number of pages in .ru zone by month

Fig. 5. Russian Higher Education and R&D segments treemap by the number of web pages indexed by main search engines, in millions.
The total number of documents increased from 6 to 7.7 million (Fig. 6). The main growth in both absolute terms and in share occurred in the NRU and the CU classes, final distribution is shown in Fig. 7. The average number of documents in FU class increased from 40 to 50 thousand, NRU – from 30 to 42, CU – from 13 to 18. In the R&D average number of documents grows slowly and is slightly more than 1000 documents in a domain. It should be noted that the bulk of documents are on the top domains for each group as mean values significantly higher than the median, and even the upper quartile.

![Graph showing the total number of documents in Russian research and education web space by month](image)

**Fig. 6.** Total number of documents in Russian research and education web space by month

Number of publications that are indexed by Google Scholar has changed slightly from 664 up to 693 thousand. The bulk of publications indexed are in the NRU, CU, and AI domains (Fig. 8) and located on a small number of leading websites. Main growth was observed in FU domains.

Finally, let us take a look at site traffic of research organizations and higher education institutions. The total site traffic has increased from 70 to 97 million sessions per month, most of the growth occurred in the sites of AI and other R&D organizations (in general in the R&D segment we can observe almost two-time growth) as well as in the NRU, FU, PU and CU sites – growth was from 40% to 58%. Final distribution is shown on Fig. 9. The highest average number of 0.5 million of sessions per month was in FU and NRU classes. In R&D NRC shows the highest values of about 153 thousand sessions, slightly less than the
**Fig. 7.** Russian Higher Education and R&D segments treemap by the number of documents indexed by main search engines, in thousands.

**Fig. 8.** Russian Higher Education and R&D segments treemap by the number of publications indexed by Google Scholar, in thousands.
CU (186 thousand sessions). The most effective were NRC sites for which 100 pages indexed by search engines resulted in more than 1800 sessions per month. AI showed the lowest efficiency of all of the classes with 47 sessions per 100 pages indexed. In HE segment NRU were the best (150 sessions) and the other average values were about 100 sessions except MS (70) and the PU (82).

![Fig. 9. Russian Higher Education and R&D segments treemap by the number user sessions per month, in millions.](image)

6 Conclusions

A comparison of the segments and classes shares for different indicators allows identifying of possible points of growth. The greatest growth potential is concentrated in the area of open science – access to full-text and metadata of scientific publications. Most of the publications of more than 7 million indexed by Google Scholar in the Russian segment of the Internet resides on the sites of scientific digital libraries eLibrary.ru (about 4 million) and CyberLeninka.ru (slightly more than 1 million). It should be noted that documents, indexed by Google Scholar are highly rated in general Google search results, leading to an increase in the site traffic and contribute to the promotion of scientific results. Organization’s web site can provide quite a different context with information on current research projects and different kinds of scientific output to a visitor than a digital library and in most cases, it leads to better results. The total number of publications of the organizations in question only for the last 5 years was more than
3 million of which less than 25% are available online. Noting the great progress and the rapid development of Internet resources of federal and national research universities and the weak, and often even a negative trend in other classes, we can conclude about the high and unrealized development potential in some of the classical universities and other types of higher education institutions.

Finally, one can see a clear backlog of R&D segment combined with high scientific potential, which may be partly explained by a more narrow, niche nature of Web resources. However, the leaders of this segment show good results and demonstrate the broad development opportunities for others.

An analysis of the dynamics of webometric indicators allows a better understanding of trends in the development of the studied web space, neutralize weaknesses inherent in measuring instruments and provide a better picture. Source data and tools located on the project site at http://www.webometrix.ru enable researchers and owners of Internet resources to explore trends in the development of scientific and educational web space, determine the position of specific organizations.

We understand that webometric rankings are quite rough because of nature of measurement instrumentation, but we suppose that conclusions drawn from such assessment may give a rise to efforts to improve web representation of educational and scientific activities.

References

LiFlow: A Workflow Automation System for Reproducible Simulation Studies

Evgheniy Kuklin\textsuperscript{1,2,3}, Andrey Sozykin\textsuperscript{1,2,3}, Konstantin Ushenin\textsuperscript{2,3}, and Dmitriy Byordov\textsuperscript{2,3}

\textsuperscript{1} Krasovskii Institute of Mathematics and Mechanics, Ekaterinburg, Russia
\textsuperscript{2} Institute of Immunology and Physiology UrB RAS, Ekaterinburg, Russia
\textsuperscript{3} Ural Federal University, Ekaterinburg, Russia
key@imm.uran.ru
http://www.ipgg.sbras.ru

Abstract. Simulation of living systems often requires numerous computational experiments on the same model for different parameter values. This paper describes the design of a user-friendly workflow automation system LiFlow for simulation of living systems, which is capable of conducting such a large series of computational experiments on supercomputers. The system provides a convenient interface for preparing input experimental data, executing the experiments on a supercomputer, and storing experimental results in a storage system. A distinctive feature of LiFlow is its simplicity and usability—the system is intended to be used by researchers in mathematical biology and biophysics without extensive knowledge in parallel computing. The paper provides examples of the use of the LiFlow system for simulation of the human heart left ventricle.

Keywords: parallel computing systems, supercomputers, living system simulation, computational workflow, computational experiment reproducibility.

1 Introduction

Simulation of living systems requires significant computational resources. Such investigations are certain to be rather time-consuming and, thus, are hard to be conducted in a reasonable time without parallel computing systems and supercomputers. However, the use of parallel computing systems requires a high degree of qualification in computer science, which many researchers involved in living systems modeling do not possess or want. Moreover, the data preparation for computational experiments is routine and time-consuming. The user needs to copy data to a supercomputer, compile the source code if necessary, enqueue the jobs with the supercomputer resource manager, and keep track of their completion. Such routine tasks should be automated.

Living system simulation often demands a large number of computational experiments on the same model but with varying parameter values. Nowadays, researchers have to prepare the configuration, input data, and the desired parameter values, and then separately execute the simulation software for each
experiment. With the number of required computational experiments increasing up to hundreds or even thousands, which is typical for living systems simulation, manual preparation and execution of experiments becomes very labor-intensive and often nearly impossible. In such a case, scientists often execute only a fraction of the required computational experiments, which negatively affects the research results, hence the need to automate the execution of series of computational experiments with varying parameter values.

Another problem that researchers in simulation of living systems often face is non-reproducibility of computational experiments. This problem is directly related to the large number of computational experiments that scientists have to carry out in order to obtain meaningful results. Due to pressure for publishing, scientists devote little time to keeping the records of experimental details, especially in case of hundreds or thousands of experiments. In addition, many other factors can affect computational results, such as a change in the version of the compiler or a required library on the supercomputer. Automated recording of experimental details and storage of simulation results can help to ensure reproducibility of the computational experiments.

We developed LiFlow (LIving system simulation workFLOW), a workflow system that addresses this need for automation. LiFlow provides the scientists with a convenient graphical user interface (GUI) that allows to prepare and execute a series of computational experiments on a parallel computing system with a single click.

One of the important goals of creating the LiFlow system was to make the initial learning process of the workflow tool very simple. Otherwise, busy scientists will not invest their time in studying the capabilities of the new system, and it will be useless.

The LiFlow system is primarily intended for simulation of living systems; we provide some examples of using LiFlow to simulate the human heart left ventricle. However, LiFlow could also be used in other areas that require conducting a large number of computational experiments on parallel computing systems.

2 Related Work

To bridge the gap between researchers and software engineers and reduce experiment preparation time, scientific computation workflow systems [1] are being developed. The most frequently used among them are Taverna, Kepler, and Triana. Taverna [2] is an open source workflow system particularly focused on bioinformatic applications and services; it is based on the XScufl language. Kepler [3] is a scientific workflow system that builds on the PtolemyII system, which is a visual modeling tool written in Java. Triana [4] is a GUI-based workflow system for coordinating and executing a collection of services. All these tools have some visual interfaces that allow graphical composition of operations. The systems provide the ability to integrate distributed computing resources, applications, data sets, and tools for computational experiments. In addition, the systems hide the complexity of distributed computing systems from users,
enabling them to describe the workflow graphically. The existing systems for scientific computing workflows are able to use the computing resources of various types (GRID, supercomputers, distributed systems, etc.), data stores (local, network, cloud), and tools (visualization, statistical processing, etc.); they also include provenance tracking, either as an integral part or as an optional module. As a result, such systems are very complicated and difficult to install, maintain, and use. Their main disadvantage is the fact that creating a new component can require considerable efforts and a detailed knowledge of the workflow system architecture. However, when simplified workflows are sufficient, there is no need for unwieldy options with a lot of settings. On the contrary, computational experiments should preferably launch “in one click.”

Another possible solution to the problem is to use an environment that provides the integration of application software packages with supercomputers. An example of such system is DiVTB [5], which provides user-friendly graphical interface where parameters of a computational experiment can be specified, and the experiment can then be executed on a supercomputer. However, such systems do not provide automation of the tasks that are popular in living systems modeling—such as launching of a series of computational experiments with the same model but varying parameter values; also, they do not support metadata tracking.

To solve the reproducibility problem special software tools can be used. They provide the ability to automatically capture and store for future use all the environment of a computational experiment, such as the simulation software, the input and output data, the hardware and software configuration of the computing system, etc.

There are two basic methods used by the reproducibility improvement tools. One method is based on executing experiments in a virtual environment, such as virtual machines or cloud [6]. After an experiment completes, the snapshot of the virtual machine is saved together with the simulation software, the output data, the experimental log, and so on. Furthermore, the snapshot can be made publicly available; other scientists can use it to reproduce the experiment and cite in their papers. Unfortunately, this approach is not suitable for parallel computing systems because virtualization considerably reduces the performance of such systems. In addition, such approach will require capturing the snapshots of all nodes in the cluster that were used for running the experiment, which is not feasible.

The second method is based on capturing the snapshot not of the entire virtual machine but of the simulation software executable and the output data. This approach is used in the CDE system (Code, Data, and Environment packaging) [7]. However, a package prepared by the CDE system depends on the software configuration of the computational system. Although the configuration of a personal computer or a virtual machine is relatively easy to replicate, it can be very difficult to adjust the configuration of a parallel computing system. Most of such systems are shared among a great number of users; only qualified
administrators can install or configure the software. Hence, such approach is also not suitable for parallel computing systems.

In order to ensure the reproducibility of computational experiments, we aimed at integrating LiFlow with Sumatra [8], which is an open source tool to support reproducible computational research. The Sumatra system [8] aims to capture the information required to recreate the computational experiment environment instead of capturing the experimental context itself. Sumatra uses the source code of the program instead of the binaries, stores the logs of the compilation process, and saves the information about all dependencies and general operating system configuration. Furthermore, Sumatra provides the ability to store the output data for future use in a database. In addition, Sumatra allows to index and search the data about experiments carried out, including additional information provided by scientists. For example, if experimental data was published, scientists can add tags with the name of the paper (and, perhaps, additional information such as the figure or table with the data) to the experiment record in the catalog. This allows researchers to quickly find the information required to reproduce the experiment they are interested in among a large number of experiment records. Unfortunately, Sumatra lacks a convenient desktop user interface. Although Sumatra is a standalone project, it can be used as a library for third-party development and has its own API. LiFlow can use Sumatra for capturing and storing the information of previously conducted experiments in the database.

3 LiFlow system

3.1 Workflow

Workflow in the LiFlow system corresponds the one shown in Fig. 1. During the first stage, researchers prepare the description of the so-called experiment series, which is a set of experiments with the same model and varying parameter values. The preparation includes the selection of simulation software that will implement the required model, generation of the configuration files with the required parameters, and creation of the input data files for each experiment. Next, the experiments are launched on a parallel computing system.

![Fig. 1. LiFlow system workflow](image)

When the experiments are completed, the obtained results are automatically stored in the archive in a form ready for processing (visualization, statistical
processing, etc.). Thus, a user is only required to create a description of the experiment series, all the rest is done automatically. In addition, the user is able to process the results of experiments from the archive manually using third-party tools.

### 3.2 Computational Package

Similarly to the CDE system, LiFlow uses the concept of a computational package that contains all the information required to execute a series of experiments. The LiFlow computational package consists of the following components:

- Source code of the simulation software, which can be loaded from a code repository of a version control system, such as Git.
- Generator of experiment series that describes how to generate the desired parameter values for the experiment series.
- Initial data and parameters to launch the simulation software.

A distinctive feature of the LiFlow computational package is that it describes not one experiment but a whole series of experiments. Each experiment in the series uses the same simulation software but different values of the model parameters. The parameter values for every experiment are produced by the generators, which are a part of the software package that is based on the rules specified by the user.

### 3.3 LiFlow Architecture

The LiFlow system consists of the four main components (Fig. 2). The Computational Package Preparation Tool and the Experiment Execution GUI are installed on the researcher’s personal computer, while the Experiment Execution Engine and the Parallel Computing System Adapter are deployed to the parallel computing system.

A user creates a computational package with the help of the Computational Package Preparation Tool and uses the Experiment Execution GUI to transmit the package to the desired parallel computing system and run the experiment series. Experiment Execution Engine on the computational cluster receives the package, compiles the source code of the simulation software, and executes the generator of the experiment series to produce a set of input data files for simulation software with various parameter values. Next, the set of computational jobs is generated with the same simulation software but different input files. The jobs are queued on the computational cluster using the Parallel Computing System Adapter, which interacts with the resource manager of the cluster.

Once the job is completed, the results of the experiment are automatically recorded to the Experiment Archive on the storage system. After all the jobs in the experiment series are completed, the Experiment Execution Engine sends an email with the report on the experiments’ execution to the user.

The planned Sumatra module would be able to capture the environment of the computational experiment and store it in the Experiment Catalog in order
Fig. 2. LiFlow system architecture

to share the initial data and simulation results among the researchers and to improve the experiments’ reproducibility.

4 Technical Details

The first stage of the LiFlow system implementation has been currently completed. The computational package in the implementation is represented by a directory in a file system that contains the subdirectories with the following components: the source code of the simulation software, the generator of the experiment series, the initial data for the generator, and the script for executing the experiments.

In the current implementation, the generator is a script that creates a series of experiments by varying the parameters in the configuration file of the simulation software. The LiFlow system supports two options for specifying the parameter values:

− The range of the parameter: an initial value, a final value, and an increment. One record in a configuration file of the generator produces the input data for several experiments.
− The explicit parameter values declaration. The parameter values must be specified for each experiment in the series.

The prepared computational package is transferred to the parallel computing system using the SSH or SFTP protocols. Next, the source code of the simulation software is built on the computational cluster. If the build process fails, LiFlow warns the user and sends back to him the build log file. In the case of a successful compilation, the system runs the generator of the experiment series to produce the input data for the experiments.
Currently, only one version of the Parallel Computing System Adapter is implemented, which is based on the SLURM Workload Manager [9]. The experiment startup script from the computational package enqueues the generated tasks for the series of experiments in the SLURM job queue. After the job is complete, the LiFlow system copies the output data to the Experiment Archive using the NFS protocol.

The scripts in the LiFlow system are written in Python. The storage of simulation software source codes is implemented as a Git repository provided by a third-party service.

![LiFlow system GUI](image)

**Fig. 3.** LiFlow system GUI

Users are provided with a simple graphical interface, which allows one to execute a series of experiments on a parallel computing system in one click (Fig. 3). The user needs to select the parallel computing system to perform the computation, specify the credentials (login and password), the path to the folder with the computational package, and the email address (for job completion notifications). When the user clicks the *Launch* button, the LiFlow system starts the workflow process. The text output shows the current stage of the process of setting up the experiment and, if an error occurs, specifies where did it happen. Fig. 3 demonstrates an example of a successfully submitted experiment. The LiFlow GUI is also written in Python using the PyQt4 library and is designed to work both on Windows and Linux.
A disadvantage of the current LiFlow implementation is the lack of a failover mechanism. If an error occurs, the experiment will not be repeated. This approach is chosen because the failure can be caused not only by problems with hardware or system software, but also, more frequently, by an error in the simulation software or a wrong combination of parameters, for which the computation cannot be performed. In such a case, restarting the experiment will not lead to solving the problem, it will only unnecessarily load the computational cluster. Still, a failure in carrying out one experiment does not lead to termination of the entire experiment series.

5 Using LiFlow for Heart Simulation

Nowadays, the LiFlow system is integrated with the URAN supercomputer at the Krasovskii Institute of Mathematics and Mechanics and the computational cluster of the Ural Federal University. The system had been used on these clusters to execute several experiments in simulation of human heart left ventricle (LV) using the LeVen simulation system [10].

The study of the influence of the fiber direction in the LV anatomical model on the speed and consistency of its electrophysiological activation was performed using the LiFlow system [11]. A series of 55 experiments was performed, where two parameters, corresponding to the direction of the fiber course in electrophysiological models, were varied.

The same system can be used to reproduce the results of the research manually conducted before. Two series of experiments were performed in the investigation of the excitation speed of the LV myocardial tissue by using an anatomical model that allows to change the shape of the ventricle and the direction of the fiber course in it [12]. In one of the series of experiments, the area of the initial activation, the fiber direction of the anatomical model, and the ratio of coefficients in the diffusion tensor of the electrophysiological model were varied. In total, the work was based on more than 36 experiments with the parameter values generated by certain rules.

The paper [13] describes the research in the dynamics of the spiral waves in the LV of the human heart model with different geometry and direction of the fiber course. In the research, several series of experiments with the anatomy that approximate normal and pathological anatomy of the LV were carried out. In each series of experiments, the following parameters were varied: the thickness of the top, the value of the diffusion tensor, and the place of the initial start-up wave. In total, the work was based on more than 84 experiments.

6 Discussion

The users of the LiFlow system, researchers in mathematical biology and biophysics from the Institute of Immunology and Physiology UrB RAS, provided a generally positive feedback. Before, they needed approximately 30 minutes to
manually prepare and run one computational experiment on a parallel computing system. With the help of the LiFlow system, they could execute a series of dozens or hundreds of experiments in less than one hour. The users appreciated the convenience of the LiFlow GUI and the ability to obtain the results of simulation from the storage system. As a result, they do not have to deal with the Linux operating system on the computational cluster, which is unfamiliar to them. Overall, LiFlow helped the researchers from the Institute of Immunology and Physiology UrB RAS to conduct computational experiments more efficiently.

As opposed to the popular scientific computation workflow systems such as Taverna, Kepler, and Triana, which provide for building large, complex computational workflows, the LiFlow system provides only one simple workflow. However, this limitation provides an opportunity to make the LiFlow system extremely easy to use. The complicated computational workflow systems are especially useful in domains with standardized data formats and tools, such as bioinformatics (the Taverna system is specifically targeted at bioinformatic applications). However, adding new components into such workflow systems is rather difficult. In contrast, the LiFlow system is more suitable for researchers who write the simulation code by themselves. Unfortunately, due to the beta version of our project, we will be able to publish the source code later.

7 Conclusion and Future Work

The paper presents the LiFlow computational workflow system intended to automate the processing of a large number of computational experiments for living systems simulation on parallel clusters. Distinctive features of LiFlow are the automatic generation of the input data and parameters for carrying out experiment series. The system has been used for simulation of the human heart left ventricle. The use of LiFlow can significantly reduce the preparation time of a series of experiments, as well as make processing of their results more convenient.

Directions for future work include:

– Full integration with the Sumatra tool to ensure the reproducibility of launched experiments.
– Developing the mechanisms of secure integration of several computational clusters from different organizations with a single LiFlow instance in order to share computational resources and simulation results.
– Implementing the Parallel Computing System Adapters for cluster resource managers other than SLURM, as well as for cloud.
– Creating more advanced and flexible generators of experiment series integrated with GUI.

Acknowledgments. The work is supported by the RAS Presidium grant I.33P “Fundamental problems of mathematical modeling”, project no. 0401-2015-0025, and by the Research Program of Ural Branch of RAS, project no. 15-7-1-26. Our study was performed using the Uran supercomputer of the Krasovskii Institute
of Mathematics and Mechanics and computational cluster of the Ural Federal University.

References

Parallel Text Document Clustering Based on Genetic Algorithm

Madina Mansurova¹, Vladimir Barakhnin²,³, Sanzhar Aubakirov¹, Yerzhan Khibatkhanuly¹, and Aigerim Mussina¹

¹ Al-Farabi Kazakh National University, Almaty, Kazakhstan
² Institute of Computational Technologies SB RAS, Novosibirsk, Russia
³ Novosibirsk State University, Novosibirsk, Russia

mansurova01@mail.ru, bar@ict.nsc.ru, {c0rp.aubakirov, x.erzhan, aygerimmusina}@gmail.com

Abstract. This work describes parallel implementation of the text document clustering algorithm. The algorithm is based on evaluation of the similarity between objects in a competitive situation, which leads to the notion of the function of rival similarity. Attributes of bibliographic description of scientific articles were chosen as the scales for determining similarity measure. To find the weighting coefficients which are used in the formula of similarity measure a genetic algorithm is developed. To speed up the performance of the algorithm, parallel computing technologies are used. Parallelization is executed in two stages: in the stage of the genetic algorithm, as well as directly in clustering. The parallel genetic algorithm is implemented with the help of MPJ Express library and the parallel clustering algorithm using the Java 8 Streams library. The results of computational experiments showing benefits of the parallel implementation of the algorithm are presented.

Keywords: clustering algorithm, genetic algorithm, parallel computing.

1 Introduction

The volume of the digital content increases every day. This impedes the process of selection of the most appropriate material, when searching for the necessary information. Clustering is one of the instruments that allows to perceive large volumes of information. Clustering is a process of dividing a set of text documents of the electronic database into classes when the elements united into one class (called a cluster) have a greater similarity than the elements referring to different classes. The process of text document clustering is resource intensive; the problem gets more complicated with the increase in the volume of the data being processed. To solve this problem, researches apply the different technologies of parallel computing.

The aim of this work is development of parallel FRiS-Tax algorithm for clustering of scientific articles. For clustering, the measure of rival similarity was taken as the proximity measure. To automate the search for the most appropriate
weighting coefficients in the formula of similarity measure, a genetic algorithm was developed.

The paper is organized as follows. The relevance of research is substantiated in Section 1. Section 2 describes the problem of text document clustering and the accepted proximity measure. Also, Section 2 presents FRiS-Tax clustering algorithm. Section 3 describes a genetic algorithm to find the most appropriate weighting coefficients in the formula of similarity measure. Section 4 presents parallel versions of genetic and clustering algorithms. Then, the results of computational experiments and the obtained data analysis are given. In conclusion, the results of the performed work are summarized.

2 Clustering algorithm with competitive similarity function

This work deals with the problem of clustering publications from bibliographic databases which allows automating the process of choosing publications for a concrete researcher or a group of working together researchers. In the clustering problem, each cluster is described with the help of one or several identifiers called centroids. These are centers of gravity or central objects of clusters. FRiS-Tax algorithm ([1, 2, 3]) is chosen as a clustering algorithm. Comparison of FRiS-Tax algorithm with the existing analogs is presented in [1] and the results of FRiS-Tax exceed the results of competitors. The experiments with FRiS-Tax algorithm showed its high efficiency when solving the clustering problem, and demonstrated the usefulness of rival similarity functions in different problems of data analysis. To measure similarity, we propose to take the attributes of the bibliographic descriptions of documents as scales.

Let $D$ be a set of documents. Similarity measure $m$ on the $D$ set is defined as follows:

$$m : D \times D \rightarrow [0, 1],$$

and in the case of complete similarity, function $m$ has the value 1, in case of complete difference - 0. Calculation of the similarity measure is performed by the formula of the type:

$$m(d_1, d_2) = \sum_{i=1}^{n} a_i m_i(d_1, d_2),$$

where $i$ is the index of the element (attribute) of the bibliographic description, $a_i$ are weighting coefficients, $m_i(d_1, d_2)$ is the measure of similarity by the $i$-th element (in other words, by $i$-th scale), and $n$ is the number of considered attributes.

The measure of rival similarity is introduced as follows. In the case of the given absolute value of similarity $m(x, y)$ between two objects, the rival similarity of object $a$ with object $b$ on competition with $c$ is calculated by the following formula:
\[
F_{b/c}(a) = \frac{m(a, b) - m(a, c)}{m(a, b) + m(a, c)},
\]
where \( F \) is called a function of rival similarity or FRiS-function. The values of \( F \) change within the range from +1 to −1. This is what we call the function of rival similarity or FRiS-function. Function \( F \) agrees well with the mechanism of perception of similarity and difference which are used by a person when he compares a certain object with two other objects.

Similarity between the object and cluster is assigned by the same principle. In order to evaluate the rival similarity of object \( z \) with the first cluster, the absolute similarity \( m(z, 1) \) of \( z \) with this cluster and similarity \( m(z, 2) \) with the cluster-competitor are taken into account. We use the value of similarity of object \( z \) with the nearest or typical representative of the given cluster as the value of similarity of object \( z \) with the cluster. In this case, the value of the rival similarity is calculated by the formula:

\[
F_{1/2}(z) = \frac{m(z, 1) - m(z, 2)}{m(z, 1) + m(z, 2)}.
\]

The clustering method can be described as follows. Let a set of objects of sampling \( A \) be given. The similarity of objects united into clusters is taken as a rival similarity with the central object of the cluster. Such objects were called pillars of clusters [1]. The peculiarity of the problem of dividing a set into clusters is that at the initial stage the reference of the objects of sampling to this or other cluster in unknown. All the objects of set \( A \) are likely to refer to one cluster. If we fix a set of centroids of this cluster \( S = \{s_1, s_2, \ldots, s_k\} \), then for each object \( a \in A \) it is possible to find the distance \( m(a, s_{a1}) \) (from the object to the nearest centroid from set \( S \)). But the absence of a cluster-competitor does not allow to determine the distance of the object to the nearest pillar of the cluster-competitor. In this regard, in the first stage, a virtual cluster-competitor is introduced the pillar of which is placed from each object of sampling at a fixed distance equal to \( m^* \). Then, the value of rival similarity of object \( a \) with the nearest to it pillar \( s_{a1} \) from \( S \) in comparison with the virtual competitor is written as:

\[
F^*_{s_{a1}}(a) = \frac{m(a, s_{a1}) - m^*}{m(a, s_{a1}) + m^*}.
\]

The number of pillars in the clustering problem will be chosen in such a way so that the value of competitive similarity of each object of sampling \( A \) with the nearest to it pillar from \( S \) is maximum:

\[
\mathcal{F}(S) = \sum_{a \in A} F^*_{s_{a1}}(a) \to \max_S.
\]

The proposed algorithm chooses the number of clusters automatically. The user only assigns the limit number of clusters \( K \), among which he would like to have the best variant of clustering. The algorithm subsequently seeks for solution
of the problem of clustering for all values $k = 1, 2, \ldots, K$, so that to choose the best of them (Fig. 1).

The advantages of using FRiS-Tax algorithm are shown in [4]. Firstly, the use of FRiS-compactness as a criterion of information capability of features at random distributions of images showed a significant advantage in comparison with a widely used criterion of minimum of errors when recognizing the test sampling by Cross Validation or One Leave Out methods. Secondly, at normal distributions, FRiS-algorithm first chooses the pillars located in the area of mathematical expectation, and, if distributions are polymodal and images are linearly inseparable, the pillars will be in the centres of modes. In the process of recognition, the decision is made in the favor of that image the pillar of which is similar to the control object most of all and the value of the function of similarity of the object with the chosen image allows to judge about the reliability of the taken decision. And finally, the use of FRiS-function for solution (at international contest Data Mining Cup 2009) of the problem of predicting the values of variables measured in absolute scale allowed it creator to hold the 4th place among 321 teams. Thus, the efficiency of using FRiS-function in algorithms for solution of problems of predicting quantitative variables is demonstrated.

3 A genetic algorithm for adjustment of coefficients in the formula of similarity measure

In this work, we have chosen:

- the year of issue;
- code UDC;
as attributes of division of articles from bibliographic databases into clusters. To choose weighting coefficients which are used in the formula of similarity measure (2), a genetic algorithm was developed. The genetic algorithm refers to heuristic algorithms of search which is used for solving the problems of optimization and modeling by random selection, combination and variation of the sought-for parameters using the mechanisms similar to natural selection in nature [5]. It should be noted that earlier the weighting coefficients were adjusted manually by an experimental way (see [3]) and the change of the problem domain of documents required a new series of experiments. The use of genetic algorithm allows automating the search for the most acceptable weighting coefficients in the formula of similarity measure.

The genetic algorithm is executed in the following stages ([5]):
1) Creation of initial population.
2) Selection.
3) Choice of parents.
4) Crossover.
5) Mutations.

The description of realization of genetic algorithm stages as applied to the problem of clustering is presented below (Fig. 2).

3.1 Creation of initial population

To create the initial population and its further evolution, it is necessary to have an ordered chain of genes or a genotype. According to [5], in some, usually random, way a set of genotypes of initial population is created. These genotypes are estimated using a "fitness-function" as a result of which each genotype is associated with a definite value ("fitness") that determines how well the genotype described by it solves the set-up task. For this task, a chain of genes has a fixed length equal to 13 and presents a set of parameters made up on the basis of attributes of bibliographic description of documents.

3.2 The structure of a chromosome

In genetic algorithms, the individuals entering the population are presented by ordered subsequent genes or chromosomes with coded in them sets of the problem parameters. Figure 3 presents the structure of a chromosome consisting of 13 genes. Abbreviations in Figure 3 present the first letters of the gene’s name, for example, UseAbstract = UAb.

The values which can be taken by genes are presented in the right column of Table 1. Genes from the given genotype are used as follows. Let us consider
the genes the values of which vary within the range from 0 to 3. If the value of gene is equal to 0, it is not used in creation of population. If the value is more than 0, this value presents the corresponding weight of gene: authorsWeight, keywordsWeight, titleWeight, abstractTextWeight. These weights are used further, when calculating proximity measure $m$ according to formula (2).

Genes from Table 1 which end in the word Equality: AuthorEquality, TitleEquality, KeywordsEquality, AbstractEquality define the way of comparing the attributes of documents and are used in creation of population only if the corresponding values of genes UseAuthors, UseTitle, UseKeyWords, UseAbstract are positive. If the values AuthorEquality, TitleEquality, KeywordsEquality, AbstractEquality are equal to 0, we use usual comparison by the method $Equals$ to compare lists of authors, names of articles and keywords. If the values of AuthorEquality, TitleEquality, KeywordsEquality are equal to 1, we use Levenstein distance for evaluation of proximity measure of attributes [6]. If the value of gene...
### Table 1. A set of genes

<table>
<thead>
<tr>
<th>N</th>
<th>Genes</th>
<th>Possible values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>POSSIBLE-DIFFERENCES</td>
<td>0-3</td>
</tr>
<tr>
<td>2</td>
<td>UseAbstract</td>
<td>0-3</td>
</tr>
<tr>
<td>3</td>
<td>UseUdk</td>
<td>0-1</td>
</tr>
<tr>
<td>4</td>
<td>UseKeyWords</td>
<td>0-3</td>
</tr>
<tr>
<td>5</td>
<td>UseAuthors</td>
<td>0-3</td>
</tr>
<tr>
<td>6</td>
<td>UseJournaSeria</td>
<td>0-1</td>
</tr>
<tr>
<td>7</td>
<td>UseTitle</td>
<td>0-3</td>
</tr>
<tr>
<td>8</td>
<td>UseYear</td>
<td>0-1</td>
</tr>
<tr>
<td>9</td>
<td>AuthorEquality</td>
<td>0-1</td>
</tr>
<tr>
<td>10</td>
<td>TitleEquality</td>
<td>0-1</td>
</tr>
<tr>
<td>11</td>
<td>KeywordsEquality</td>
<td>0-1</td>
</tr>
<tr>
<td>12</td>
<td>AbstractEquality</td>
<td>0-1</td>
</tr>
<tr>
<td>13</td>
<td>K(number of clusters)</td>
<td>2-12</td>
</tr>
</tbody>
</table>

**AbstractEquality** is equal to 1, the algorithm of shingles [7] is used for evaluation of proximity measure of annotations.

The values of genes *UseUdk, UseJournaSeria, UseYear* are binary, i.e. depending on the values the genes are either used or not used, in case of being used, +1 is added to the measure m. The gene *POSSIBLE-DIFFERENCES* is a threshold value, when evaluating proximity by Levenstein distance. The value of this gene varies from 0 to 3. If *POSSIBLE-DIFFERENCES* = 0, the compared names, authors, keywords must completely coincide. If, when comparing, the calculated Levenstein distance is less than the threshold value, the corresponding weight: *AuthorsWeight*, *titleWeight* or *KeywordsWeight* is added to the proximity measure *m*. If Levenstein distance exceeds the threshold value, it is concluded that the attributes are different.

### 3.3 Selection

In genetic algorithm, a set of individuals, each with its own genotype, is a certain solution of the clustering problem. Let us suppose that we have generated an individual, that is a set of weighting coefficients is given to determine the measure of similarity.

At the stage of selection, the parents of the future individual are determined with the help of *Roulette Selection* [8], *Tournament Selection* [9], and *Elitism Selection* [9] methods. Selection by *Roulette Selection* proceeds as follows. The values of fitness of all individuals are summed and we obtain a certain value sum, then choose a random number between 0 and the sum. A cycle is started according to the number of individuals, their fitnesses are summed and as soon as the sum exceeds the random number, we return the index of the individual which was the last to take part in summing. When using *Tournament Selection*, n tournaments are realized to choose n individuals. When using *Elitism Selection*,
the individuals with the greatest fitness securely pass on to a new population. The use of elitarism usually allows to accelerate convergence of the genetic algorithm. The disadvantage of the strategy of elitism is that the probability of getting into the local minimum increases.

3.4 Crossover

The survived individuals take part in reproduction. The crossover operator combines two chromosomes (parents) to produce a new chromosome. The new chromosome may be better than both of the parents if it takes the best characteristics from each of the parents. For this, the following methods are used: One point crossover, Two point crossover, Uniform crossover, and Variable to Variable crossover. Figure 4 presents the stage of crossover of the genetic algorithm.

![Fig. 4. Stage of one point crossover.](image)

3.5 Mutation

The stage of mutation is necessary not to let the solution of the problem get into a local extremum. It is supposed that, after the crossover is completed, part of the new individuals undergo mutations. The essence of mutation operator is as follows. In the chromosome under study, a random number of genes is picked out randomly. The coefficient of mutation determines the intensity of mutations. It determines the fraction of genes subjected to mutation on the current iteration taking into consideration their total amount. In our case, 25% of all individuals are selected which are subjected to mutation (Fig. 5).

Thus, the genetic algorithm for the clustering problem is executed in two stages: a stage of initialization and a stage of iterations.
The stage of initialization:
The first generation is formed.
The stage of iterations:
1) Clustering is performed by FRiS-Tax algorithm.
2) The value of the fitness-function is calculated.
3) The value of the fitness-function is compared with the threshold value
   of quality. For this problem, the threshold value is equal to 0.8. If the pre-
   determined value of clustering quality is reached, the algorithm stops.
4) If not, a new generation is formed: selection of individuals, reproduction,
   and mutations are performed.
5) Transition to step 1 is carried out.

![Diagram of stage of mutation]

**Fig. 5.** Stage of mutation.

### 3.6 Evaluation of the quality of clustering

In the algorithm, a fitness-function is given which allows to determine how well
the clustering problem is solved. In this work, the quality of the obtained clusters
is evaluated using the measures of estimation - *Purity* \([10]\) and *Root mean square
deviation* \([11]\).

The measure *Purity* is an external criterion of the quality of clustering which
is calculated as follows:

\[
purity = \frac{1}{N} \sum_k \max_j |w_k \cap c_j|,
\]

where \(W = w_1, w_2, \ldots, w_k\) is the result of clustering performed by an expert,
\(C = c_1, c_2, \ldots, c_j\) is the result of clustering performed by the program. Then, in
order to determine which of the individuals was not selected and is dying and
which of them survived and will take part in reproduction, we take the threshold
for the values of fitness-function. The individual dies if the function returns the
value which is less than the taken threshold. *Root mean square deviation* is
also used for evaluation of the quality of clustering. The lower the value of this
function, it is the better the quality of clustering.
4 Development of the parallel clustering algorithm

With the increase in the amount of documents to be processed, the time of the clustering process realization increases exponentially, therefore, the aim of developing a parallel algorithm of clustering is justified. Parallelization is carried out in two stages of the algorithm of clustering. Firstly, during selection of individuals in the genetic algorithm when clustering is performed with different sets of weighting coefficients. The program is written in Java, and this stage of the parallel algorithm is performed using MPJ Express, an implementation of an MPI-like API which can execute on a variety of parallel platforms ranging from multicore processors to compute clusters [12]. Secondly, it is directly in the course of performing the clustering algorithm. In FRiS-Tax algorithm, the most complex computing process is traversal of all objects of selection and testing each of them for the role of a pillar. Parallelization of this stage is implemented with the help of technology Java 8 Streams [13].

4.1 Parallelization of the genetic algorithm

The steps of a parallel version of the genetic algorithm are presented below (Fig. 6).

1) $N$ processes are started using MPJ. The number $N$ depends on the number of individuals in the first generation, i.e. if we increase the number of individuals up to 64, then 64 processes are started. Each process is started on a separate computing node.

2) Each process reads-out a file with articles which are to be divided into clusters.

3) The master-process generates $N$ random chromosomes and sends them to the rest processes.

4) Each process takes one chromosome and creates an individual, computes the value of fitness-function and sends it to the master-process.

5) Master-process checks whether there is an individual with the value of fitness-function greater or equal to the given threshold value (0.8). If such individual is found, the master-process informs all the rest processes that the individual is found and can stop the work.

6) If not, the master-process starts selection. Crossover takes place within selection.

7) After the parents are determined, a new individual is born. The old generation does not take part in the further work.

8) After that, mutation is performed, random values are assigned to random genes. The mutation coefficient is 25%. When the master-process performs selection, crossover and mutation, the rest processes wait.

4.2 Parallelization of the clustering algorithm

The load test revealed the two slowest stages in the FRiS-Tax clustering algorithm. They are the methods of finding the first pillar and finding the next pillar,
which are doing $N \times (N - 1)$ and $N \times (N - 1) \times M$ operations, where $N$ is the number of articles and $M$ is the number of already found pillars. To accelerate these methods, the technology *Java 8 Streams* was used. Since repeated $(N - 1)$ and $(N - 1) \times M$ times operations in methods finding first and finding next pillar respectively are simple and their result need to be summarized at the end, it is reasonable to implement here parallel() method of *Java 8*. The *Java* runtime partitions the stream into multiple substreams.

Finding first pillar:
1) Get article $i$ from articles list, $i$ is an iterator over list of articles.
2) Calculate its $F(S)$ by formula (4) with all articles, except itself. Since $F(S)$ is equal to the sum of $F(s_{a1}) \times (a)$ we can implement parallel() method on stream of articles. Each core receive substream and does calculation.
3) When all substreams executed, their result summarize by method sum(). Here at the second step we divide $(N - 1)$ operations between cores in a processor.

Finding next pillar:
1) Get article $i$ from articles list, $i$ is an iterator over list of articles.
2) Calculate its $F(s_{a1}) \times (a)$ with all articles, except itself, and all found pillar as showed in formula (3). We implement parallel() method on stream of articles list and on stream of defined pillars.
3) When all substreams executed, their results summarized by method sum(). Here at the second step we divide $(N - 1) \times M$ operations between cores in a processor.

5 The results of the computing experiment

To study the efficiency of the developed parallel algorithm, we carried out computing experiments in the Laboratory of computer sciences of RIMM at al-Farabi KazNU on the cluster including 16 computing nodes.

For performing analysis and clustering, the journal "Vestnik KazNU" of 2008-2015 was used as initial data. Sampling includes 95 pdf documents. The total number of articles is 2837. The choice of the initial data is conditioned by the fact that all documents were divided into series (mathematics, biology, philosophy, etc.) and further divisions do not cause difficulties, when using measures of similarity based on only bibliographic descriptions or titles of the articles. In order to
evaluate the quality of division of sampling, this body was divided into clusters with the help of an expert into the problem domain. The time of execution was determined as follows. We made measurements of the time of clustering processes for the clusters being formed on one computer node and several computer nodes for parallel realization. Figure 7 presents the dependency of time for realization of the clustering algorithm on the number of processes. Figure 8-9 present acceleration and efficiency of parallel realization. As is seen in the constructed diagrams, with the increase in the number of processes, acceleration increases to a certain value which is related to the expenditure of communication. The optimum number of processes proved to be 8 at which the maximum value of acceleration was observed but the highest value of efficiency was achieved with 4 processes. Figures 10-11 presents distribution of the documents to the resulting clusters.

The second initial sampling consists of 522 scientific articles of the journal "Siberian mathematical journal”. Each article has a code of classifier MSC2010 which was taken as a reference. This allowed to objectively evaluate the quality of clustering when using fitness Purity, the initial sampling was divided in advance by the code of classifier into 8 large clusters. The computing experiment showed the following best gene - [8, 2, 4, 4, 2, 2, 1, 0, 1, 2, 0] at which the value of fitness-function was equal to 80%.
Fig. 8. Speedup of parallel clustering algorithm FRiS-Tax.

Fig. 9. Efficiency of parallel clustering algorithm FRiS-Tax.

Fig. 10. Experimental results of clustering with number of clusters=5.
When using the fitness of Root mean square deviation, the best result was shown by the following gene - $[45, 4, 0, 1, 0, 4, 1, 1, 0, 0, 1]$ and the value of fitness-function was equal to $0.004348930638757773$.

6 Conclusion

The proposed methods for clustering of documents in the electronic form allow to realize processing on the systems consisting of more than one computer node. The attributes of bibliographic description of documents were chosen as scales for determination of the similarity measure. Parallel processes are realized at the stage of preliminary analysis of documents including calculation of similarity measures between the documents as well as directly at the stage of clustering. The use of the genetic algorithm allowed to determine the values of attributes at which clustering of documents gives the best results.

Acknowledgments. This work was supported in part under grant of Foundation of Ministry of Education and Science of the Republic of Kazakhstan ”Development of intellectual high-performance information-analytical search system of processing of semi-structured data” (2015-2017).

References

Using Non-Negative Matrix Factorization for Text Segmentation

Aliya Nugumanova\textsuperscript{1}, Madina Mansurova\textsuperscript{2}, Yerzhan Baiburin\textsuperscript{1}, and Yermek Alimzhano\textsubscript{2}\\
\textsuperscript{1} D. Serikbayev East Kazakhstan State Technical University, Oskemen, Kazakhstan, \textsuperscript{2} Al-Farabi Kazakh National University, Almaty, Kazakhstan\\
yalisha@yandex.kz, mansurova01@mail.ru, \{ebaiburin, aermek81\}@gmail.com

Abstract. The aim of this paper is to investigate whether non-negative matrix factorization (NMF) can be useful for semantic segmentation of large full-text documents. NMF is a universal technique that decomposes the monolithic structure of a massive dataset into different trends. In case of textual data these trends can be interpreted as topics. Thereby NMF can associate each document with topics covered in it, however, without linking topics to the certain parts of that document. In this paper, we complement this traditional NMF technique with a new goal: for a given full-text document we build a semantic map which links document’s parts with topics covered in it.

Keywords: non-negative matrix factorization, text segmentation, topic modeling.

1 Introduction

Text segmentation is a very interesting challenge in the field of natural language processing. It arises in many information retrieval applications providing users with quick access to document repositories. Since full-text documents stored in such repositories are usually large to read and analyze, information retrieval applications should be able to divide them into chunks and deliver the most relevant chunks to users in accordance to their requests [1].

In this paper, we focus on the task of segmentation of full-text documents from a topic modeling perspective. In recent years, topic modeling is gaining momentum in data mining in general [2], and in particular in the text segmentation field [3]. Recent work in this field has shown that using topic distribution over documents instead of term distribution can significantly increase segmentation performance [4,5,6,7,8,9].

The segmentation mechanism drawn from topic modeling is very simple. At first, each document is divided on small segments (e.g., sentences or paragraphs). At second, topics covered in this document are revealed, and each word in each segment is associated with one topic; thereby, for each segment topic occurrences are defined. At last, adjacent document’s segments sharing a certain number of common topics are merged into topical chunks.
One of the most popular approaches to topic modeling is non-negative matrix factorization (NMF). In general, NMF is a well-recognized technique due its ability to extract relevant structures of data and may thus contribute to a deeper understanding of data behavior [10]. This technique, being applied to a collection of full-text documents, maps it into a space of topics. For example, in Fig. 1, NMF is applied to a co-occurrence matrix of a collection, which consists of 5 documents. As we can see from the figure, after matrix factorization, document 1 is represented as a combination of topic 1 and topic 3. Simultaneously, topic 3 is represented as a combination of term 2, term 4 and term 5, and term 5 is the most significant for this topic.

![Fig. 1. A sample of non-negative matrix factorization.](image)

As we can see, NMF has two useful applications. Firstly, for each document it defines the most weighted topics, which we call relevant topics. Secondly, for each topic it finds the most weighted terms, which we call support terms. In this paper, we use support terms for the semantic segmentation of full-text documents. Our contribution is to complement the traditional NMF representation with a new goal: the creation of a semantic map of the given document through using support terms as map’s nodes. We suppose that, by linking these nodes to the corresponding parts of the document, we achieve its smart and comfortable segmentation.

The rest of this paper is organized as follows. In Section 2, we discuss previous work on text segmentation and explain our reasons to use NMF. In Section 3, we present proposed approach. In particular, we demonstrate how we link support terms with the document parts, and present some experimental results. In Section 4, we formulate conclusions and plans for our future work.

2 Related Work

The most simple and intuitive algorithm of text segmentation is TextTiling [11]. It uses a sliding window to move through a document and capture text blocks (tiles). The similarity between consecutive blocks are calculated on the base of cosine metrics. The calculated values are used to draw a similarities curve that
tracks topics changes between consecutive blocks so that the segment boundaries are chosen at the local minima of the curve. The main disadvantage of TextTiling is low accuracy because of the sparsity of text blocks.

Another simple algorithm of text segmentation is C99 [12]. At first, it divides the input document into minimal blocks (sentences) and for each block calculates its rank based on the blocks similarities. Then it performs divisive clustering starting with the whole document and splitting it to parts in accordance with blocks' ranks. In [13] C99 algorithm is improved by applying Latent Semantic Analysis for calculating the blocks' similarity matrix.

C99 algorithm also is used in [1]. This work addresses the issue of providing topic driven access to full-text documents. Authors of this work apply C99 to subdivide documents into smaller thematically homogeneous parts that can be used as link targets. They try to perform segmentation as accurate as possible: document parts should be of such sizes that "shrinking them would cause relevant information to be left, and expanding them would bring in too much non-relevant information" [1]. However, they concentrate only on the segmentation phase without details of designing a whole navigation system.

As we have mentioned in the introduction, a considerable line of research explores text segmentation methods based on topic modeling. The most popular algorithms for topic modeling are Latent Dirichlet Allocation (LDA) [3], [8,9,10] and Non-negative Matrix Factorization (NMF) [10,11]. Although output of LDA is very similar to the output of NMF, these models are fundamentally different in nature: LDA is based on a Bayesian probabilistic model; whereas NMF is based on algorithms of linear algebra that fit root mean squared error. As it’s shown in [14], both LDA and NMF can discover concise and coherent topics and demonstrate similar performance, however NMF learns more incoherent topics than LDA. Authors of [15] also compare LDA and NMF, and conclude that NMF better than LDA "from the perspectives of consistency from multiple runs and early empirical convergence".

We choose NMF here because of its basis sparseness [16]. Basis sparseness means that NMF uses less basis features (terms) than LDA. This makes NMF topics more overlapped, i.e. more semantically related to each other than LDA ones (see an example represented by Table 1). We consider that these relations are essential for the understanding of how the document’s semantic map should be organized.

Fig. 2 gives a visual interpretation of Table 1. The four topics are shown in Fig.2, and some of them are linked with other topics through specific key terms.

3 Proposed Approach

Our method of text segmentation consists of 5 steps. Firstly, we should subdivide a given full-text document into units and build units-by-terms co-occurrence matrix. Secondly, we should define a reasonable number of topics (K) and apply NMF to factorize the co-occurrence matrix and obtain 2 matrices: units-by-topics and topics-by-terms. Thirdly, for each extracted topic we should sort topic terms
Table 1. Number of topics intersected by basis terms in "Geology" text collection

<table>
<thead>
<tr>
<th>#</th>
<th>Basis term in Russian</th>
<th>Number of topics to which this term is assigned through NMF</th>
<th>Number of topics to which this term is assigned through LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>горный</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>порода</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>поверхность</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>склон</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>процесс</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>динамика</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>система</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>эпоха</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>состав</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>кора</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Fig. 2. The visual interpretation of overlapped topics of "Geology" text collection.
by their weights and select only the most weighted terms (no more than 10% of all terms). We call these terms support terms.

In our case we have used "Geology" textbook written in Russian [17], and divided it into 89 units (by number of chapters). Then we have chosen K=5 and extracted 5 topics and 500 support terms. By the way we have proposed information about term distribution over these 5 topics to a geology expert for analysis. Based on the information the expert has concluded that it is best to name extracted topics as "Endogenous", "Solar system", "Oceans", "Exogenous" and "Geochronology". Tables 2-6 represent top 10 support terms for each of 5 extracted topics.

Table 2. Support terms for topic 1 "Endogenous" (top 10)

<table>
<thead>
<tr>
<th>#</th>
<th>Through NMF</th>
<th>Through LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Russian</td>
<td>English</td>
</tr>
<tr>
<td>1</td>
<td>порода</td>
<td>rock</td>
</tr>
<tr>
<td>2</td>
<td>процесс</td>
<td>process</td>
</tr>
<tr>
<td>3</td>
<td>движение</td>
<td>movement</td>
</tr>
<tr>
<td>4</td>
<td>минерал</td>
<td>mineral</td>
</tr>
<tr>
<td>5</td>
<td>землетрясение</td>
<td>earthquake</td>
</tr>
<tr>
<td>6</td>
<td>горный</td>
<td>mining</td>
</tr>
<tr>
<td>7</td>
<td>внутренний</td>
<td>internal</td>
</tr>
<tr>
<td>8</td>
<td>образование</td>
<td>formation</td>
</tr>
<tr>
<td>9</td>
<td>метаморфизм</td>
<td>metamorphism</td>
</tr>
<tr>
<td>10</td>
<td>динамика</td>
<td>dynamics</td>
</tr>
</tbody>
</table>

Table 3. Support terms for topic 2 "Solar system" (top 10)

<table>
<thead>
<tr>
<th>#</th>
<th>Through NMF</th>
<th>Through LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Russian</td>
<td>English</td>
</tr>
<tr>
<td>1</td>
<td>Земля</td>
<td>Earth</td>
</tr>
<tr>
<td>2</td>
<td>магнитный</td>
<td>magnetic</td>
</tr>
<tr>
<td>3</td>
<td>планета</td>
<td>planet</td>
</tr>
<tr>
<td>4</td>
<td>солнечный</td>
<td>solar</td>
</tr>
<tr>
<td>5</td>
<td>поле</td>
<td>field</td>
</tr>
<tr>
<td>6</td>
<td>система</td>
<td>system</td>
</tr>
<tr>
<td>7</td>
<td>ядро</td>
<td>core</td>
</tr>
<tr>
<td>8</td>
<td>поверхность</td>
<td>surface</td>
</tr>
<tr>
<td>9</td>
<td>солнце</td>
<td>Sun</td>
</tr>
<tr>
<td>10</td>
<td>атмосфера</td>
<td>atmosphere</td>
</tr>
</tbody>
</table>
Table 4. Support terms for topic 3 "Oceans" (top 10)

<table>
<thead>
<tr>
<th>#</th>
<th>Through NMF</th>
<th></th>
<th>Through LDA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Russian</td>
<td>English</td>
<td>Russian</td>
<td>English</td>
</tr>
<tr>
<td>1</td>
<td>океан</td>
<td>ocean</td>
<td>океан</td>
<td>ocean</td>
</tr>
<tr>
<td>2</td>
<td>зона</td>
<td>zone</td>
<td>зона</td>
<td>zone</td>
</tr>
<tr>
<td>3</td>
<td>континентальный</td>
<td>continental</td>
<td>губина</td>
<td>depth</td>
</tr>
<tr>
<td>4</td>
<td>континент</td>
<td>continent</td>
<td>континенталь</td>
<td>continental</td>
</tr>
<tr>
<td>5</td>
<td>пояс</td>
<td>belt</td>
<td>континент</td>
<td>continent</td>
</tr>
<tr>
<td>6</td>
<td>глубина</td>
<td>depth</td>
<td>волна</td>
<td>wave</td>
</tr>
<tr>
<td>7</td>
<td>платформа</td>
<td>platform</td>
<td>пояс</td>
<td>belt</td>
</tr>
<tr>
<td>8</td>
<td>разлом</td>
<td>rift</td>
<td>мощность</td>
<td>power</td>
</tr>
<tr>
<td>9</td>
<td>кора</td>
<td>crust</td>
<td>дно</td>
<td>bottom</td>
</tr>
<tr>
<td>10</td>
<td>мощность</td>
<td>power</td>
<td>подводный</td>
<td>underwater</td>
</tr>
</tbody>
</table>

Table 5. Support terms for topic 4 "Exogenous" (top 10)

<table>
<thead>
<tr>
<th>#</th>
<th>Through NMF</th>
<th></th>
<th>Through LDA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Russian</td>
<td>English</td>
<td>Russian</td>
<td>English</td>
</tr>
<tr>
<td>1</td>
<td>вода</td>
<td>water</td>
<td>вода</td>
<td>water</td>
</tr>
<tr>
<td>2</td>
<td>процесс</td>
<td>process</td>
<td>процесс</td>
<td>process</td>
</tr>
<tr>
<td>3</td>
<td>порода</td>
<td>rock</td>
<td>динамика</td>
<td>dynamics</td>
</tr>
<tr>
<td>4</td>
<td>экзогенный</td>
<td>exogenous</td>
<td>внешний</td>
<td>external</td>
</tr>
<tr>
<td>5</td>
<td>внешний</td>
<td>external</td>
<td>экзогенный</td>
<td>exogenous</td>
</tr>
<tr>
<td>6</td>
<td>динамика</td>
<td>dynamics</td>
<td>материал</td>
<td>material</td>
</tr>
<tr>
<td>7</td>
<td>материал</td>
<td>material</td>
<td>склон</td>
<td>slope</td>
</tr>
<tr>
<td>8</td>
<td>склон</td>
<td>slope</td>
<td>поверхность</td>
<td>surface</td>
</tr>
<tr>
<td>9</td>
<td>выветривание</td>
<td>erosion</td>
<td>озеро</td>
<td>lake</td>
</tr>
<tr>
<td>10</td>
<td>поверхность</td>
<td>surface</td>
<td>подземный</td>
<td>underground</td>
</tr>
</tbody>
</table>

Table 6. Support terms for topic 5 "Geochronology" (top 10)

<table>
<thead>
<tr>
<th>#</th>
<th>Through NMF</th>
<th></th>
<th>Through LDA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Russian</td>
<td>English</td>
<td>Russian</td>
<td>English</td>
</tr>
<tr>
<td>1</td>
<td>кора</td>
<td>crust</td>
<td>порода</td>
<td>rock</td>
</tr>
<tr>
<td>2</td>
<td>химический</td>
<td>chemical</td>
<td>минерал</td>
<td>mineral</td>
</tr>
<tr>
<td>3</td>
<td>минерал</td>
<td>mineral</td>
<td>горный</td>
<td>mining</td>
</tr>
<tr>
<td>4</td>
<td>земной</td>
<td>terrestrial</td>
<td>процесс</td>
<td>process</td>
</tr>
<tr>
<td>5</td>
<td>желтый</td>
<td>yellow</td>
<td>химический</td>
<td>chemical</td>
</tr>
<tr>
<td>6</td>
<td>элемент</td>
<td>element</td>
<td>кора</td>
<td>crust</td>
</tr>
<tr>
<td>7</td>
<td>порода</td>
<td>rock</td>
<td>состав</td>
<td>composition</td>
</tr>
<tr>
<td>8</td>
<td>верхний</td>
<td>upper</td>
<td>земной</td>
<td>terrestrial</td>
</tr>
<tr>
<td>9</td>
<td>метод</td>
<td>method</td>
<td>образование</td>
<td>formation</td>
</tr>
<tr>
<td>10</td>
<td>таблица</td>
<td>table</td>
<td>являться</td>
<td>to be</td>
</tr>
</tbody>
</table>
The fourth step is the most important in our method. We should associate our units with topics taking into account topics’ support terms. If we use only the traditional NMF representation we miss opportunities to exploit the distributional power of support terms.

For example, let’s consider the Unit #30 in the given Geology textbook. The unit describes the history of glaciations as well as the impact of glaciers on the Earth’s crust (see Table 7). So the geology expert has associated this unit with the topic "Exogenous" as well as with the topic "Geochronology". In contrast, the traditional NMF algorithm evaluates highly the relation of this unit with the topic "Exogenous" and very lowly the relation with the topic "Geochronology". But if one analyses the support terms used in the Unit #30, one finds that the topic "Geochronology" is well represented in the unit with the help of support terms such as "period", "year", "history", "epoch", "time", "Holocene", "cycle" etc.

Table 7. Topics distribution over the Unit #30 of the Geology textbook [17, p.318-319]

<table>
<thead>
<tr>
<th>Unit #30</th>
<th>История оледенений. Ледниковые эры, периоды, эпохи. В истории Земли неоднократно возникали великие оледенения, при которых площади ледниковых покровов возрастали до десятков миллионов квадратных километров. Интервалы времени длительностью в миллионы лет с характерными для них похолоданием климата и разрастанием оледенений получили название ледниковых периодов. Последний из них, продолжающийся до сих пор и называемый плейстоценовым, или четвертичным, начался 2,5-3 млн лет назад</th>
</tr>
</thead>
<tbody>
<tr>
<td>NMF representation</td>
<td>Expert representation</td>
</tr>
<tr>
<td>Topic</td>
<td>Weight</td>
</tr>
<tr>
<td>Exogenous</td>
<td>125.99</td>
</tr>
<tr>
<td>Solar system</td>
<td>93.70</td>
</tr>
<tr>
<td>Oceans</td>
<td>92.30</td>
</tr>
<tr>
<td>Geochronology</td>
<td>10.56</td>
</tr>
<tr>
<td>Endogenous</td>
<td>0</td>
</tr>
</tbody>
</table>

Therefore, in order to more accurately define topics for each document’s unit we should complement the traditional NMF approach by analyzing support terms distribution in this unit. We should analyze next 3 factors:

1. What support terms related to the certain topic are occurred in this unit?
2. How frequently they are occurred?
3. How important are they for the certain topic (how many their weights in the topic)?

As a result we should decide can we associate this unit with some support terms and with some topics represented via these terms. The decision rule can be summarizes as follows:
\[
\text{Decision} = \begin{cases} 
1, & \sum_{st \in \text{Topic} \cap \text{Unit}} tf(st, Unit) \times \text{weight(st, Topic)} \geq Th \\
0, & \sum_{st \in \text{Topic} \cap \text{Unit}} tf(st, Unit) \times \text{weight(st, Topic)} < Th,
\end{cases}
\]

where \(st\) is a support term related to the topic, \(tf(st, Unit)\) is its frequency in the unit, \(\text{weight(st, Topic)}\) is its weight in the topic, \(Th\) is a threshold value above which the topic is recognized as related to the unit. In this paper we set \(Th = 0.1\).

At the fifth step, we construct a navigation (semantic) map that contains three layers: layers of document units, layers of support terms and layers of topics. In Fig. 3 a part of the Geology textbook's map is illustrated. This map consists of 5 top-level nodes which correspond to textbook topics and 500 middle-level nodes which correspond to support terms. Middle-level nodes can be moved up or down or rolled up and stored away until one activates them again. Active middle-level nodes point out the related units which are bottom-level nodes. If middle level nodes are rolled up, access to bottom-level units is enable directly through top-level nodes.

Fig. 3. A fragment of the Geology textbook semantic map.

4 Discussion

So, the main advantage of NMF in comparison with LDA is a great "naturalness", i.e. the topics in NMF are more widely intersect with each other by a number of representative terms. LDA tries to generate topics in such a way so
that intersect as less as possible, as a result, each topic has its own set of "exclusive" words. In practice, such a result does not look natural because in each document there is a number of general thematic terms which refer to the subject in a whole and can not belong only to one definite topic within the subject. For example, in geology such terms are the words "rock", "process", "dynamics", etc. NMF associates such terms with several topics, while LDA tries to assign each such term to only one topic, as a result its stability suffers. In the new series of experiments, a set of words referred by LDA to one topic can significantly differ from the set of words referred to the same topic in the previous series of experiments. In our experiments, NMF proved to be a more stable method than LDA. The topical dispersion of representative key words does not hinder segmentation of the document, on the contrary, it enhances segmentation not isolating segment from each other but connecting them. The disadvantage of NMF is its computing complexity. But this is the problem of computing technologies, not of the method itself.

5 Conclusion And Future Work

In this paper we considered semantic map as a tool of smart document's segmentation and organization. We presented an approach to automatic creation of semantic maps and showed how this process can benefit from a new interpretation of NMF based on the concept of support terms. Also we performed a little case study to illustrate proposed approach. However, more work should be done to evaluate advantages and disadvantages of this approach, to substantiate choice of the NMF parameters (e.g. the number of reasonable topics, or start number of units, or threshold for topic validation). Efforts must also be devoted to a complete comparison of proposed approach with LDA and NMF.

Acknowledgments. This work was supported in part under grant of Foundation of Ministry of Education and Science of the Republic of Kazakhstan "Development of intellectual high-performance information-analytical search system of processing of semi-structured data" (2015-2017).

References

MATHEMATICAL MODELING
Evgeniya Amelina¹, Sergey Golushko¹,², and Andrey Yurchenko¹

¹ Institute of Computational Technologies
Siberian Branch of the Russian Academy of Sciences,
Academician M.A. Lavrentjev ave. 6, Novosibirsk, Russia
² Novosibirsk National Research State University,
Pirogova str. 2, Novosibirsk, Russia
{amelina.evgenia,s.k.golushko,andrey.yurchenko}@gmail.com
http://www.ict.nsc.ru/en

Abstract. The paper presents a computational technology for optimization of composite overwrapped pressure vessels (COPV). Mathematical modeling and numerical optimization were applied to design COPV. The mathematical models were built using different shell theories and structural models of composites.

The stress-strain state of the vessels was determined and analyzed based on three mathematical models. Several solutions of COPV optimization problem based on different problem statements were obtained. They were analyzed and verified by substituting of the estimated design parameters in a direct problem of stress-strain state determination.

The study demonstrated that using of non-constant design parameters, such as the thickness, the winding angle and the curvature radius of the composite shell gave the possibility for additional reduction of COPV mass, while keeping its strength. In addition, acceptability and convenience of using simpler mathematical models for numerical solving the optimization problems were demonstrated.

Keywords: COPV, mathematical modeling, computational optimization, shell theory, structural model of composite material.

1 Introduction

Composite overwrapped pressure vessels (COPV) are used in the rocket and spacecraft industry due to their high strength and lightweight. Consisting of a thin, non-structural liner wrapped with a structural fiber composite COPV are produced to hold the inner pressure of tens and hundreds of atmospheres. COPV have been one of the most actual and perspective directions of research, supported especially by NASA [1, 2].

Designing of a highly reliable and efficient COPV requires a technology for analysis of its deformation behaviour and strength assessment. This technology should allow one to obtain target COPV parameters through changing vessel’s geometry, structural and mechanical material parameters while keeping its useful load.
Combination of mathematical modeling and numerical optimization makes it possible to reduce the cost and the duration of identifying the best parameters for a COPV. However, this approach is characterized by a number of hurdles. Overcoming these hurdles determine success of an optimum designing of such structures.

So far, there have been two main approaches in optimization of composite structures: analytical and numerical ones.

In the first approach the problems are solved basing on their simplified statement, for example using the momentless (membrane) shell theory and the netting model of composite material (CM) [3–6]. The obtained results may be far from reality, however they are of value for testing of numerical optimization methods.

Application of the numerical approach in designing, on the other hand, produces a number of challenges that must be overcome, e.g. lack of reliable methods for global optimization; nonconvexity and nonlinearity of constraint functions; ill-conditioned boundary value problems; different scaling of optimization criteria represent just some of the obstacles that prevent reliable optimization of COPV.

Numerical analysis is usually a computation-intensive process and takes considerable time. One way to solve this problem is approximation of the objective function using different approaches, such as response surface method [7] and neural network [8]. Some kinds of numerical analyses use a small number of design variables, functions and/or corresponding set of their discrete values (analytical geometry parametrization [9], finite set of feasible winding angles [10]). It leads to reduction in the number of objective function calculations.

Another way is reasonable simplification of the elasticity problem statement, for example by using the membrane theory or other shell theories [9,11,12], that leaves the question of results validity. And this is the approach we have applied in our study. For validation we have used the Timoshenko [13] and Andreev-Nemirovskii [14] shell theories, acounting transverse shears with different degrees of accuracy.

Of course, it should be taken into account that the computed solutions are not optimum in the strict mathematical sense. However, these solutions could provide the considerable economy of the weight while keeping the required strength, and, therefore, they are important from the engineering point of view.

2 The Problem Statement and the Mathematical Models

Let’s consider a multilayer composite pressure vessel at a state of equilibrium under equidistributed inner pressure. We need to determine the parameters of structure and CM meeting the following requirements:

\[ V \geq V_0, \quad P \geq P_0, \quad M \leq M_0, \]

where \( V \) is the volume of the vessel, \( P \) is inner pressure and \( M \) is the vessel’s mass and they are constrained by some preset values \( V_0, P_0, M_0 \).
We define the optimization problems the following way: to find extremum of one functional from (1) under other constraints.

The structures optimization problem statement includes selection of objective functional, formulation of constitutive equations and constraints on performance and design variables.

The mathematical models describing the vessel’s state are based on the following assumptions:

1) the vessel is a multilayer thin-walled structure;
2) the vessel’s layers can have different mechanical characteristics;
3) the reinforced layer’s material is quasi-homogeneous;
4) the vessel’s main loading is high inner pressure, whose alteration happens rather slowly during operation.

These assumptions allow us to reduce dimension of the corresponding mathematical problem and to build the mathematical vessel’s models based on the different theories of multilayer non-isotropic shells.

Let’s consider the vessel as a shell rigidly compressed on the edges. Taking into account a symmetry plane in the middle of the vessel, it is enough to calculate and design only it’s half. The type of loading and boundary conditions allow considering the axisymmetric problem statement.

The half of shell is set by rotation of the generatrix $r = r(\theta)$ around axis $0z$ (fig. 1) where $r$ is the current point of the shell radius, $\theta$ is the angle between the normal to the shell surface and axis $0z$ changing within $[\theta_0; \theta_1 = 90^\circ]$. The full shell is set by reflection the shell’s half about plane $0xy$.

![Shell of rotation geometry](image)

**Fig. 1.** Shell of rotation geometry

The Kirchhoff—Love shell theory [15] (KLST) and the theories with shear terms (Timoshenko [13] (TiST) and Andreev-Nemirovskii [14] (ANST)) are used
to solve the direct calculation problems of multilayer composite vessels, to analyze their behavior and to verify optimization problem solutions. The used coordinate system is \((\theta, \varphi, \zeta)\), where \(\varphi\) denotes polar angle, \(\zeta\) – normal to the surface. The load is equidistributed inner pressure \(q = (0, 0, q_3)\). On the top fixed edge \((\theta = \theta_0)\) all displacement components, angle of normal rotation and additional shear term (ANST) are equal to zero; on the edge \((\theta = 90^\circ)\) we use the symmetry conditions: transverse force, the first displacement component, angle of normal rotation and additional shear term (ANST) are equal to zero. One could find the full systems of equations in the paper [16].

Relations between stresses and strains are defined by the structural models [17]. The main idea of these models is that CM characteristics are calculated through matrix and fibers mechanical characteristics, fibers volume content and winding angles. The stress-strain state of matrix and fibers are evaluated through stresses and strains of the composite shell. A failure criterion is applied for every component of CM. Here we use the Mises criterion to determine the first stage of failure.

The objective function whose minimum is required is the minimum mass:

\[
M = 2\pi \int_{\theta_0}^{\theta_1} r R_1 h d\theta \left[ \rho_m (1 - \omega_r) + \rho_r \omega_r \right] \rightarrow \min
\]

(2)

where \(\rho_m, \rho_r\) are the densities of matrix and reinforcing fibers, \(\omega_r\) is the volume content of reinforcement.

We chose the following design functions: the curvature radius \(R_1(\theta)\) to define the generatrix; the thickness of the shell \(h(\theta)\); the winding angle \(\psi(\theta)\) (fig. 1).

The solution has to satisfy the constraints on the shell’s inner volume:

\[
\pi \int_{\theta_0}^{\theta_1} r^2 R_1 \sin \theta d\theta = V_0,
\]

(3)

and the strength requirement:

\[
\max\{bs_r, bs_m\} \leq 1,
\]

(4)

where \(bs_r, bs_m\) are the normalized von Mises stresses in the matrix and fibers [17]. Note that the factor of safety is widely used while solving engineering problems. It can be considered by correction of the right part of the inequality (4).

We used the following constraints on the design functions:

\[
0 \leq \psi \leq 90, \quad h_0^* \leq h \leq h_1^*, \quad R_0^* \leq R_1 \leq R_1^*.
\]

(5)

The method of the continuous geodesic winding have been widely used in the manufacturing of composite shells of revolutions. In this case the winding angles are defined by the Clairaut’s formula:

\[
r \sin \psi(r) = C,
\]

(6)
where the constant $C$ is defined, as a rule, from the condition at the shell’s equator. The thickness equation is

$$h(r) = h_R \frac{R \cos \psi_R}{r \cos \psi(r)},$$  \hspace{1cm} (7)

which has the singularity at the edge where the winding angle has to be equal to 90$^\circ$. The formula (7) is applied into practice at $r \geq r_0 + r_\omega$, where $r_\omega$ is equal to the width of the reinforcement tape. As a result the equation defining the vessel’s thickness takes the form:

$$h(r) = \begin{cases} h_R \frac{R \cos \psi_R}{r_\omega \cos \psi(r_0 + r_\omega)}, & r \leq r_0 + r_\omega; \\ h_R \frac{R \cos \psi_R}{r \cos \psi(r)}, & r \geq r_0 + r_\omega. \end{cases} \hspace{1cm} (8)$$

We did not consider the problem of fibers slippage. The main goal of the study was to demonstrate the potentials of using CM in one COPV design approach.

3 Direct Problems. Analysis of the Shell Theories

Estimation of composite vessel stress-strain state using the offered models leads to the solution of boundary value problems for stiff systems of differential equations. These problems are ill-conditioned, and their solutions have big gradients near the edges. Numerical analysis was performed by the spline collocation and discrete orthogonalization methods, implemented in the COLSYS [18] and GMDO [19] software. These computing tools have proved to be effective in numerical solving of wide range of composite shell mechanics problems [20].

We investigated the vessel’s deformations by computing of its stress-strain state based on the different shell theories. The vessel’s shape was a part of a toroid: $R_1 = 2.46$ m, $\theta_0 = 0.108^\circ$, $\theta_1 = 90^\circ$ (the computed half), $r(\theta_0) =$0.04 m. The CM parameters were: $E_m = 3 \cdot 10^9$ Pa, $\nu_m = 0.34$, $E_r = 300 \cdot 10^9$ Pa, $\nu_r = 0.3$, $\omega_r = 0.55$, $V_0 =$350 liters where $E_m, E_r$ are the Young’s modulus of the matrix and fibers, $\nu_m, \nu_r$ — their Poisson’s ratio.

Fig. 2 shows the stress-strain state characteristics of the vessel with the thickness $h = 0.6$ cm, reinforced in the circumferential direction ($\psi = 90^\circ$) under the load of 170 atm. On the left are the displacements of the reference surface along the generatrix $u_1(r)$ (dashed curves) and the normal displacement of this surface $w(r)$ (solid curves). On the right are the distribution of normalized von Mises stresses (nVMS) along the thickness in the matrix $bs_m(r)$. The solid curves correspond to a slice at the shell edge, the dashed curves — to a slice at $\theta = 0.1$. The curves without symbols correspond to KLST simulations, the curves marked with $\triangle$ — to those using TiST, and $\square$ — to ANST.

It’s easy to see that the basic kinematic characteristics coincide both qualitatively and quantitatively. Small differences are observed only for the stresses
and deformations near the compressed edge. The maximum results and qualitative difference were obtained for ANST. This is due to accounting for the transverse shears by non-linear distribution in a thickness of a shell. Earlier it was shown [20] that ANST’s based results were closest to the ones of 3D elastic theory in most cases.

The winding angle’s influence on the COPV performance was investigated using parametric analysis.

Dependence of the maximum nVMS in the matrix $b_{sm}$ (dashed curves) and the fibers $b_{sr}$ (dash–dotted curves), and the maximum size of the displacement vector $||v||$ (solid curves) are shown in fig. 3. KLST’s results are drawn without marks, TiST – with symbols $\triangle$, ANST – with $\square$.

**Fig. 2.** The stress-strain state characteristics of the composite vessel computed using different shell theories

**Fig. 3.** The winding angle’s influence on the composite vessel stress-strain state
The calculated values are very close in the area of their minima (fig. 3 left side). The graphs of kinematic function \(|v||\) coincide qualitatively. Some noticeable quantitative difference are revealed only for KLST’s results.

The range \(\psi \in (42; 45)\) corresponds to the zones of minimum values (fig. 3 right side), which practically coincide (\(\text{min}_\psi bs_m \approx 0.65, \text{min}_\psi bs_r \approx 1.05, \text{min}_\psi ||v|| \approx 5 \cdot 10^{-3}\) m), as well as the angles, where these values are obtained (\(\psi \approx 43.2^\circ\) for \(bs_m\) and \(bs_r\), \(\psi \approx 43.8^\circ\) for \(||v||\)).

It was revealed that the winding angles corresponding to minimum stresses values were almost insensitive to the thickness variation. The change of \(h\) from 0.6 to 1.6 cm corresponded to the angle’s change about 0.2°.

Additionally we investigated stress-strain state of the vessel (the thickness \(h = 0.6\) cm, the winding angles \(\psi = \pm 43.2\)), when nVMS in the matrix and the fibers were near their minimum (fig. 4). The adopted notation is the same as in fig. 2.

![Fig. 4. The stress-strain state of the vessel (\(\psi = \pm 43.2\)) computed using the three shell theories](image)

And again the difference is visible only in a very small region near the edge but now this difference is small enough to be neglected. Moreover the displacement values of the reference surface, the efforts and the moments completely coincide for all the theories.

All the theories (KLST, TiST, ANST) provided similar estimated characteristics of stress-strain state. This vessel was characterized not only by essential decrease of the maximal nVMS in the matrix and fibers, but also by their almost uniform distribution along the generatrix. At the same time the values of bending moments significantly reduced bringing vessel’s stress-strain state close to momentless.

The performed analysis showed that the optimizing problem can be solved using rather simple shell theories (KLST, TiST). These theories are characterized
by lower computational complexity of corresponding boundary value problem if compared to ANST. It takes from 10 to 20 times less resources.

One can see that the winding angle as a design parameter gives an opportunity to increase the vessel’s strength significantly. The difference between the “best” and “worst” designs can reach 20 – 35 times comparing their nVMS in the matrix and fibers. The “worst” designs have the winding angle close to 90°. In this case are considerable transverse shears near the compressed edge, and the loading is redistributed to a rather weak matrix while the fibers remain unloaded.

4 Inverse problems. Optimization of the Vessel

Inverse problems involve not only numerical methods for fast and reliable solving of direct boundary value problems, but also require numerical optimization methods for identifying design parameters.

Here we considered conditional optimization problem, including direct constraints on design functions and trajectory constraints on the solution imposed at the end of the interval. The sequential unconstrained optimization is one of the most widespread approaches to solution of such problems. The main idea of the method is terminal functional convolution and multiple solutions of one-criteria problem using different optimization methods [21]. In our study the modified Lagrange function was used for the convolution.

Hence we sought for solution of a nonconvex problem of finite-dimensional optimization [22] by discretization of design functions. The methods implemented in the OPTCON-A software [23] were used to get the corresponding solution.

In our study several vessels with different type of design parameters were investigated. The parameters were either functions or constants. Additionally a design of continuous winding on the geodesic path was considered, where the only design function was its curvature radius.

Uniform mesh for design functions discretization included 7 points, except for the geodesic winding design with nonuniform mesh of 17 points. The distance between points was also the solution of corresponding optimization problem. Approximation of the design functions was carried out using the 3rd degree natural splines.

The masses of these vessels are shown in tab. 1. Numbers after “F” denote the design parameters-functions, after “C” – the design parameters-constants with notations: 1 – ψ, 2 – h, 3 – R1. ”Geod” denotes the design with continuous geodesic winding.

We used the mass of C123 design as a basis for further comparisons. It was about 22 kg (tab. 1).

The considered design with the continuous geodesic winding has been one of COPV widely used in practice [24]. Its mass was 19 kg while the design F3C12 (the constant winding angles and thickness, the variable geometry) was about 16 kg. The main reason was the circumferential winding near the opening, which
Table 1. The different designs comparison

<table>
<thead>
<tr>
<th>Design</th>
<th>Mass, kg</th>
<th>M/M_{C123}</th>
<th>(\psi_{\text{max}} - \psi_{\text{min}})</th>
<th>(h_{\text{max}}/h_{\text{min}})</th>
<th>(R_{1\text{max}}/R_{1\text{min}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>F123</td>
<td>16.04</td>
<td>72.5%</td>
<td>4</td>
<td>1.08</td>
<td>1.62</td>
</tr>
<tr>
<td>F12C3</td>
<td>16.02</td>
<td>72.4%</td>
<td>12</td>
<td>1.55</td>
<td>1.00</td>
</tr>
<tr>
<td>F13C2</td>
<td>16.05</td>
<td>72.5%</td>
<td>2</td>
<td>1.00</td>
<td>1.94</td>
</tr>
<tr>
<td>F23C1</td>
<td>17.01</td>
<td>76.9%</td>
<td>0</td>
<td>4.74</td>
<td>3.14</td>
</tr>
<tr>
<td>F1C23</td>
<td>20.23</td>
<td>91.4%</td>
<td>15</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>F2C13</td>
<td>16.42</td>
<td>74.2%</td>
<td>0</td>
<td>1.83</td>
<td>1.00</td>
</tr>
<tr>
<td>F3C12</td>
<td>16.07</td>
<td>72.6%</td>
<td>0</td>
<td>1.00</td>
<td>1.83</td>
</tr>
<tr>
<td>Geod</td>
<td>19.09</td>
<td>86.2%</td>
<td>85</td>
<td>9.95</td>
<td>5.42</td>
</tr>
<tr>
<td>C123</td>
<td>22.13</td>
<td>100.0%</td>
<td>0</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

did not allow using all the fiber resources. Therefore it was necessary to increase the vessel thickness near the holes.

Comparison of the designs with two constants and one function (F1C23, F2C13, F3C12) shows that the possibility to change the value of parameter along the radius (parameter’s variability) is the most critical for the shell geometry \((R_1)\) and its thickness.

Fig. 5 presents the design parameters and the vessel’s generatrix for several designs.

Important additional design characteristic is its "adaptability in manufacturing". For example, the 5 - 10 times difference of thickness along the meridian would become a serious obstacle for vessels manufacturing. Thus, designs of nearly minimum mass possessing good properties and satisfying to the given technological constraints could be of great value, than optimum without them.

Let’s consider the following characteristics of the obtained solutions (tab. 1): the difference between the maximal and minimum values of angles \((\psi_{\text{max}} - \psi_{\text{min}})\), the maximum thickness ratio (to \(h_{\text{max}}/h_{\text{min}}\)) and the maximum curvature radius ratio \((R_{1\text{max}}/R_{1\text{min}})\), which characterizes deviation of the generatrix from a circle arch.

According to tab. 1 the design with the geodesic continuous winding has the thickness ratio about 10 and large gradient near the edge. We found the design F123, where the parameter variance is small – no more than 10% for the thickness and 4° for the winding angle with low gradients. The designs with the constant thickness (F13C2, F3C12) showed that it was possible to receive the vessels with weight close to minimum, having on hand only such design functions as \(R_1\) and \(\psi\).

5 Inverse Problem. Verification of the Solutions

We verified the solutions of optimization problem by substituting the obtained design parameters into the direct problem. Fig. 6 shows key characteristics of the stress-strain state for the designs F123, F23C1, C123 and Geod.
Fig. 5. The design parameters ($a$, $b$, $c$) and the half vessel's generatrix ($d$)
Fig. 6. The stress-strain state characteristics. (a) – bending moment, (b) – tensile force, (c) — nVMS for fibers, (d) — nVMS for matrix.
It is noteworthy that the stress-strain state of design F123 is almost momentless, and the fibers are equally stressed. The influence of transverse shear is minimum.

Let’s substitute this solution in the direct problem. All the three considered shell theories have yielded close results (fig. 7). The difference is noticeable only for ANST in narrow zones (less than 1% of all the area of calculation) at the edges, where non-linear accounting for transverse shear gives difference of about 5%. At the same time the estimated efforts and bending moments are very close for all the theories, and the bending moments are very small.

Thus, it is possible to use the simplest shell theory to solve such optimization problem and the estimation of stress-strain state will be close to those obtained using more complex theories.

6 Conclusions

A technology of COPV optimization has been developed. It makes possible to obtain high pressure vessel designs that not only meet such requirements as minimum mass, preset volume and strength, but also possess a number of additional valuable engineering characteristics including stress-strain state close to momentless and almost equally stressed fibers.

Non-constant design parameters, such as thickness, winding angles and curvature radius of composite shell give a possibility for additional reduction of COPV mass while keeping its strength. The obtained design with the variable design parameters are up to 27% lighter if compared to the best design with the constant parameters.
The optimization problem solutions have been verified by solving the direct problems with obtained design parameters using the classical shell theory and theories with shear terms.

Our study has demonstrated acceptability and convenience of using simple mathematical models based on the Kirchhoff—Love and Timoshenko shell theories for numerical solving of the optimization problems.

**Acknowledgments.** This study was supported by Integrated program of basic scientific research of SB RAS No. 24, project II.2 "Design of computational technologies for calculation and optimal design of hybrid composite thin-walled structures" and the scientific project RFBR 15-37-20265.

**References**

On the Inverse Problem of a Creeping Motion in Thin Layers

Victor Andreev

Institute Computational Modelling SB RAS, Siberian Federal University, Krasnoyarsk, Russia
andr@icm.krasn.ru

Abstract. The new partially invariant solution of two-dimensional motions of heated viscous liquid equations is considered. For factor-system arised the initial boundary value problem is formulated. This problem is inverse one and describing of common motion of two immiscible liquids in a plane channel under the action of thermocapillary forces. As Marangoni number is small (so-called creeping flow) the problem becomes the linear one. Some a priori estimates are obtained and input data conditions when solution tends to stationary one are found. In Laplace transforms the exact solution is obtained as quadratures and some numerical results of velocities behavior in layers are presented.

Keywords: Thermocapillarity, a priori estimates, conjugate initial-boundary value problem, asymptotic behaviour, numerical simulation

1 Introduction

It is well known that in a non-uniformly heated liquid a motion can arise. In some applications of liquid flows, a joint motion of two or more fluids with surfaces takes place. If the liquids are not soluble in each other, they form a more or less visual interfaces. The petroleum-water system is a typical example of this situation. At the present time modelling of multiphase flows taking into account different physical and chemical factors is needed for designing of cooling systems and power plants, in biomedicine, for studying the growth of crystals and films, in aerospace industry [1-4].

Nowadays, there are exact solutions of the Marangoni convection [5-7]. One of the first solutions was obtained in [8]. This is the Poiseuille stationary flow of two immiscible liquids in an inclined channel. As a rule, all such flows were considered steady and unidirectional. The stability of such flows was investigated in [9, 10]. As for non-stationary thermocapillary flows, studying of them began recently [11, 12].

Thermocapillary convection problem for two incompressible liquids separated by a closed interface in a container was investigated in [13]. Local (in time) unique solvability of the problem was obtained in Holder classes of functions. The problem of thermalcapillary 3D motion of a drop was studied in [14]. Moreover, its unique solvability in Holder spaces with a power-like weight at infinity
was established. Velocity vector field decreases at infinity in the same way as the initial data and mass forces, the temperature diverges to the constant which is the limit of the initial temperature at infinity. The present work is devoted to studying of solutions of a conjugate boundary value problem arising as a result of linearization of the Navier-Stokes system supplemented with temperature equation. The description of the 2D creeping joint motion of two viscous heat conducting fluids in flat layers is also provided here. The motion arises due to thermocapillary forces imposed along two interfaces, after which the unsteady Marangoni convection begins. Such kind of convection can dominate in flows under microgravity conditions or in motions of thin liquid films.

2 Statement of the Problem

The 2D motion of a viscid incompressible heat-conducting liquid in the absence of mass forces is described by the system of equations

\begin{align*}
\frac{\partial u_1}{\partial t} + u_1 \frac{\partial u_1}{\partial x} + u_2 \frac{\partial u_1}{\partial y} + \frac{1}{\rho} \frac{\partial p}{\partial x} &= \nu \left( \frac{\partial^2 u_1}{\partial x^2} + \frac{\partial^2 u_1}{\partial y^2} \right), \quad (2.1) \\
\frac{\partial u_2}{\partial t} + u_1 \frac{\partial u_2}{\partial x} + u_2 \frac{\partial u_2}{\partial y} + \frac{1}{\rho} \frac{\partial p}{\partial y} &= \nu \left( \frac{\partial^2 u_2}{\partial x^2} + \frac{\partial^2 u_2}{\partial y^2} \right), \quad (2.2) \\
\frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y} &= 0, \quad (2.3) \\
\frac{\partial \theta}{\partial t} + u_1 \frac{\partial \theta}{\partial x} + u_2 \frac{\partial \theta}{\partial y} &= \chi \left( \frac{\partial^2 \theta}{\partial x^2} + \frac{\partial^2 \theta}{\partial y^2} \right). \quad (2.4)
\end{align*}

Here $u_1(x, y, t)$ and $u_2(x, y, t)$ are the components of the velocity vector, $p(x, y, t)$ is the pressure, $\theta(x, y, t)$ is the temperature, $\rho > 0$ is the density, $\nu > 0$ is the kinematic viscosity and $\chi > 0$ is the thermal conductivity of the liquid. The quantities $\rho > 0$, $\nu > 0$ and $\chi > 0$ are constant.

The system of equation (2.1)–(2.4) admits a four-dimensional Lie subalgebra $G_4 = \langle \partial_x, \partial_{u_1} + t \partial_x, \partial_p, \partial_\theta \rangle$. Its invariants are $t$, $y$, $u_2$ and a partially invariant solution of rank 2 and defect 3 should be sought for in the form

$$u_1 = u_1(x, y, t), \quad u_2 = u(y, t), \quad p = p(x, y, t), \quad \theta = \theta(x, y, t).$$

Inserting the exact form of the solution into the equations (2.1)–(2.3) yields

\begin{align*}
u w_x + w^2 &= f(t) + \nu w_y, \quad \frac{1}{\rho} \frac{\partial d}{\partial y} = d(x, y, t) - \frac{f(t)x^2}{2}, \quad (2.5) \\
\frac{\partial d}{\partial y} - v \frac{\partial v}{\partial y} - v_t - vv_y, \quad g_t + vg_y + wg &= 0
\end{align*}

with some function $f(t)$ that is arbitrary so far.

Regarding the temperature field, we assume that equation (2.4) has the solution of the form

$$\theta = a(y, t)x^2 + m(y, t)x + b(y, t). \quad (2.6)$$
As we see below, (2.6) is in good accord with conditions on the interface.

The stationary solution of the Navier-Stokes equations in the form (2.5) for \( g = 0 \) for pure viscous fluid was found for the first time by [15]. It describes the liquid impingement from infinity on the plane \( y = 0 \) under the no slip condition on it. In the paper [16], this solution for the flow between two plates or for the flow in a cylindrical tube (axisymmetric analogue of solution (2.5)) was applied.

It is known that the temperature dependence of the surface tension coefficient is the one of the most important factors leading to the dynamic variety of the interfacial surface. In the papers [17, 18] the stationary solutions in form (2.5), (2.6) was found at \( a(y, t) \equiv 0, b = \text{const} \) for a flat layer with a free boundary \( y = l = \text{const} \) and a solid wall \( y = 0 \). The non-uniqueness of solution depending on the physical parameters of the problem was revealed. A similar problem in the case of half space was investigated in [19].

We assume for simplicity that \( g(y, t) \equiv 0, m(y, t) \equiv 0 \). The latter condition means that the temperature field has an extremum at \( x = 0 \), more exactly, a maximum for \( a(y, t) < 0 \) and a minimum for \( a(y, t) > 0 \).

Let us apply the solution of the form (2.5), (2.6) to described joint motion of two immiscible liquids in the flat layer \( 0 < y < l \) considering that the wall \( y = 0 \) and \( y = h \) are solid and the line \( \Gamma \): \( y = l(x, t) \) is their common interface, see Fig. 1.

![Fig. 1. Geometry of the Marangoni convection problem](image)

Introduction the index \( j = 1, 2 \) for the liquids and using (2.5) and (2.6), we come to the conclusion that the unknowns satisfy the equations

\[
\begin{align*}
    w_{jt} + v_j w_{jy} + w_j^2 &= \nu_j w_{jyy} + f_j(t), \quad w_j + v_j y = 0, \quad (2.7) \\
    \frac{1}{\rho_j} p_j &= d_j(y, t) - \frac{f_j(t)x^2}{2}, \quad d_{jyy} = \nu_j v_{jyy} - v_{jt} - v_j v_{jy}, \quad (2.8) \\
    a_{jt} + 2w_j a_j + v_j a_{jy} &= \chi_j a_{jyy}, \quad b_{jt} + v_j b_{jy} = \chi_j b_{jyy} + 2\chi_j a_j \quad (2.9)
\end{align*}
\]
in domain \( 0 < y < l(x, t) \) for \( j = 1 \) and in domain \( l(x, t) < y < h \) for \( j = 2 \).
At the interface $y = l(x,t)$ the conditions hold [1]
\[ w_1(l(x,t),t) = w_2(l(x,t),t), \quad v_1(l(x,t),t) = v_2(l(x,t),t), \]  
\[ l_t + xw_1(l(x,t),t)l_x = v_1(l(x,t),t), \]  
\[ a_1(l(x,t),t) = a_2(l(x,t),t), \quad k_1 \frac{\partial a_1}{\partial n} = k_2 \frac{\partial a_2}{\partial n}, \]  
\[ b_1(l(x,t),t) = b_2(l(x,t),t), \quad k_1 \frac{\partial b_1}{\partial n} = k_2 \frac{\partial b_2}{\partial n}, \]

$k_1 > 0$, $k_2 > 0$ are the heat conductivity coefficients and $\mathbf{n} = (1 + l_x^2)^{-1/2}(-l_x, 1)$ is the normal to the line $y = l(x,t)$.

The dynamic condition for $\Gamma$ has a vector form [1]
\[ (p_1 - p_2)\mathbf{n} + 2[\mu_2D(\mathbf{u}_2) - \mu_1D(\mathbf{u}_1)]\mathbf{n} = 2\sigma K\mathbf{n} + \nabla_{\Gamma}\sigma, \quad \mu_j = \rho_j\nu_j. \]  

In (2.12) $D(\mathbf{u})$ is the strain-rate tensor, $\sigma(\theta_1)$ is the surface tension coefficient, $K$ is the mean curvature of the interface, whereas $\nabla_{\Gamma} = \nabla - \mathbf{n}(\mathbf{n} \cdot \nabla)$ on the right-hand side designates the surface gradient. For most of real liquid media the dependence $\sigma(\theta_1)$ is approximated well by the linear function
\[ \sigma(\theta_1) = \sigma^0 - \kappa \theta_1, \]  

where $\sigma^0 > 0$ and $\kappa > 0$. They are assumed constant and determined by experimental methods. Projecting condition (2.12) to the tangent direction $\mathbf{r} = (1 + l_x^2)^{-1/2}(1,l_x)$, and using (2.13), (2.6) we obtain
\[ l_x[\mu_2(v_{2y} - w_2) - \mu_1(v_{1y} - w_1)] + \frac{x}{2}(1 - l_x^2)(\mu_2w_{2y} - \mu_1w_{1y}) = -\kappa(\theta_{1x} + l_x\theta_{1y}) = -\kappa[2a_1x + l_x(a_{1y}x^2 + b_{1y})]. \]  

The projection (2.12) to the normal $\mathbf{n}$ yields
\[ \rho_1d_1 - \rho_2d_2 + \frac{[\rho_2f_2(t) - \rho_1f_1(t)]x^2}{2} + 2[\mu_2D(\mathbf{u}_2) - \mu_1D(\mathbf{u}_1)]\mathbf{n} \cdot \mathbf{n} \]  
\[ = [\sigma^0 - \kappa(a_1x^2 + b_1)] \frac{l_{xx}}{(1 + l_x^2)^{3/2}}. \]  

The boundary conditions on the solid walls have the form
\[ w_1(0,t) = 0, \quad v_1(0,t) = 0, \quad w_2(h,t) = 0, \quad v_2(h,t) = 0, \]  
\[ a_1(0,t) = a_{10}(t), \quad a_2(h,t) = a_{20}(t), \]  
\[ b_1(0,t) = b_{10}(t), \quad b_2(h,t) = b_{20}(t), \]

with some given functions $a_{j0}(t)$ and $b_{j0}(t)$.

The initial conditions for the velocities are zero because of we study the properties of the solution of the problem simulating the motion only under the
action of thermocapillary forces $w_j(y,0) = 0$, $v_j(y,0) = 0$. Besides, $l(x,0) = l_0(x)$, $a_j(y,0) = a^0_j(y)$, $b_j(y,0) = b^0_j(y)$.

Note several specific features of the formulated problem. This is a nonlinear and inverse one since the functions $f_j(t)$ are unknowns also. It is easy to understand if we exclude the functions $v_j(y,t)$ from the equations of mass conservation. Then the problem reduces to the conjugate problem for the functions $w_j(y,t)$, $a_j(y,t)$ and $l(x,t)$. The problem for $b_j(y,t)$ given $v_j(y,t)$ and $a_j(y,t)$ can be separated. The functions $d_j(y,t)$ can be recovered by quadrature from the second equation (2.8) up to a function of time. The last condition in (2.10) and the fourth from (2.16) are the additional conditions on $f_j(t)$, $j = 1, 2$.

Let us introduce the characteristic scales of length and time as well as functions $w_j$, $v_j$, $a_j$, $d_j$ and $f_j$, namely, the quantities $l^0$, $l_0^2/\nu_1$, $\kappa a^0 l^0/\mu_1$, $\kappa a^0 l_0^2/\mu_1$, $a^0$, $\kappa a^0 l^0/\rho_1$, $\kappa a^0/(\rho_1 l^0)$, where $l^0 = \text{const} > 0$ is the average value of thickness of the first layer of the liquid at $t = 0$, $a^0 = \max_{t \geq 0} |a_{20}(t) - a_{10}(t)| > 0$, or $a^0 = \max_j \max_y |a_{j0}(y)| > 0$, if $a_{20}(t) = a_{10}(t)$. In the dimensionless variables, some factor appears at the nonlinear terms in (2.7), the Marangoni number

$$M = \kappa a^0 l_0^3/\mu_1\nu_1.$$  

The same applies to the kinematic condition (2.11)

$$\bar{l}_t + \bar{x}M\bar{w}(\bar{l}(\bar{x},\bar{t}),\bar{t})\bar{l}_{\bar{x}} = M\bar{v}_1(\bar{l}(\bar{x},\bar{t}),\bar{t}).$$  

Assume that the $M \ll 1$. The latter holds either in the thin layers or large viscosities. Then the nonlinear terms in the equations can be neglected and the latter become linear. In particular, the kinematic condition (2.20) has the form $\bar{l}_t = 0$, i.e. $\bar{l} = \bar{l}(x)$. Let us turn to (2.15). After transition to the dimensionless variables on the right-hand side the Weber number $\text{We} = \sigma^0/(\kappa a^0 l_0^2)$ appears instead of $\sigma^0$. In the real conditions $\text{We} \gg 1$ for the most of liquid media; for example, for the water–air system $\text{We} \sim 10^6$.

Therefore, for these Weber numbers, (2.14) assume the form $\bar{l}_{\bar{x},\bar{x}} = 0$, i.e. $\bar{l} = \alpha x + l^0$. We assume later that $\alpha = 0$ and the interface is the plane $y = l^0 < h$ parallel to the solid walls $y = 0$ and $y = h$; in what follows, the index 0 for $l^0$ will be omitted.

### 3 A priori Estimates

Let us present the so-obtained linear problem in its entirely in dimensional form

$$w_{jt} = \nu_j w_{jyy} + f_j(t),$$  

$$w_j(y,0) = 0,$$  

$$w_1(0,t) = 0, \quad w_2(h,t) = 0,$$  

$$w_1(l,t) = w_2(l,t),$$

$$\nu_j = \frac{\rho_j \mu_j}{\sigma_j}.$$
\[ \mu_2 w_2(y, t) - \mu_1 w_1(y, t) = -2\kappa a_1(l, t), \quad (3.5) \]
\[ \int_0^l w_1(z, t) \, dz = 0, \quad \int_l^h w_2(z, t) \, dz = 0, \quad (3.6) \]

where \( 0 < y < l \) for \( j = 1 \) and \( l < y < h \) for \( j = 2 \). The first equality in (3.6) follows from (2.10) whereas the last in the no-slip condition \( v_2(h, t) = 0 \).

Let us write the problem for the functions \( a_j(y, t) \)

\[ a_{jt} = \chi_j a_{jyy}, \quad (3.7) \]
\[ a_j(y, 0) = a_j^0(y), \quad (3.8) \]
\[ a_1(0, t) = a_{10}(t), \quad a_2(h, t) = a_{20}(t), \quad (3.9) \]
\[ a_1(l, t) = a_2(l, t), \quad k_1 a_{1y}(l, t) = k_2 a_{2y}(l, t). \quad (3.10) \]

In order to obtain a priori estimates for \( w_j(y, t) \), \( f_j(t) \) of the solution of (3.1)–(3.5), it is necessary firstly to infer the estimates for the solutions of initial-boundary value problem (3.7)–(3.10). We perform the change of variables

\[ a_1(y, t) = \bar{a}_1(y, t) + \frac{a_{10}(t)(y - l)^2}{l^2}, \quad 0 \leq y \leq l^0 \equiv l, \quad (3.11) \]
\[ a_2(y, t) = \bar{a}_2(y, t) + \frac{a_{20}(t)(y - l)^2}{(h - l)^2}, \quad l \leq y \leq h. \]

The functions \( \bar{a}_j(y, t) \) in their domains satisfy the equations

\[ \bar{a}_{1t} = \chi_1 \bar{a}_{1yy} + \frac{2\chi_1 a_{10}(t)}{l^2} - \frac{a'_{10}(t)(y - l)^2}{l^2} \equiv \chi_1 \bar{a}_{1yy} + g_1(y, t), \quad (3.12) \]
\[ \bar{a}_{2t} = \chi_2 \bar{a}_{2yy} + \frac{2\chi_2 a_{20}(t)}{(h - l)^2} - \frac{a'_{20}(t)(y - l)^2}{(h - l)^2} \equiv \chi_2 \bar{a}_{2yy} + g_2(y, t), \quad (3.13) \]

where the prime denotes differentiation with respect to time. Boundary conditions (3.9) for \( \bar{a}_1 \) and \( \bar{a}_2 \) become homogeneous, whereas (3.10) preserve it form. Initial conditions (3.8) for \( \bar{a}_1 \) and \( \bar{a}_2 \) change

\[ \bar{a}_1(y, 0) = a_1^0(y) - \frac{a_{10}(0)(y - l)^2}{l^2} \equiv \bar{a}_1^0(y), \quad (3.14) \]
\[ \bar{a}_2(y, 0) = a_2^0(y) - \frac{a_{20}(0)(y - l)^2}{l^2} \equiv \bar{a}_2^0(y). \]

Let us multiply (3.1), (3.2) by \( \rho_1 c_1 \bar{a}_1 \) and \( \rho_2 c_2 \bar{a}_2 \) \( c_1 \), \( c_2 \) and integrate over the segments \([0, l], [l, h]\) taking into account (3.8) and (3.9). Then add up the result. We infer that

\[ \frac{dA(t)}{dt} + k_1 \int_0^l \bar{a}_{1y}^2 \, dy + k_2 \int_l^h \bar{a}_{2y}^2 \, dy = \rho_1 c_1 \int_0^l g_1 \bar{a}_1 \, dy + \rho_2 c_2 \int_l^h g_2 \bar{a}_2 \, dy, \quad (3.15) \]
\[ A(t) = \frac{\rho_1c_1}{2} \int_0^l \bar{a}_{1t}^2 \, dy + \frac{\rho_2c_2}{2} \int_l^h \bar{a}_{2t}^2 \, dy, \quad (3.16) \]

where \( c_j \) are the coefficients of the specific heat capacity. Along with (3.15) there is another identity

\[ \rho_1c_1 \int_0^l \bar{a}_{1t}^2 \, dy + \rho_2c_2 \int_l^h \bar{a}_{2t}^2 \, dy + \frac{1}{2} \frac{\partial}{\partial t} \left[ k_1 \int_0^l \bar{a}_{1y}^2 \, dy + k_2 \int_l^h \bar{a}_{2y}^2 \, dy \right] \]

\[ = \rho_1c_1 \int_0^l g_1 \bar{a}_{1t} \, dy + \rho_2c_2 \int_l^h g_2 \bar{a}_{2t} \, dy. \quad (3.17) \]

From (3.15) and (3.17) we obtain theinform estimates in \( y \)

\[ |a_j(y, t)| \leq \left( \frac{8\chi_j}{k_j^2} F(t)A(t) \right)^{1/4} + |a_{j0}(t)|, \quad (3.18) \]

where

\[ F(t) = k_1 \int_0^l \bar{a}_{10}^2(y) \, dy + k_2 \int_l^h \bar{a}_{20}^2(y) \, dy + \frac{2k_1}{\chi_1} \left[ \frac{4\chi_1}{l^3} \int_0^t a_{10}^2(\tau) \, d\tau + \frac{l}{5} \int_0^t (a_{10}^2(\tau))^2 \, d\tau \right] + \frac{2k_2}{\chi_2} \left[ \frac{4\chi_2}{(h-l)^3} \int_0^t a_{20}^2(\tau) \, d\tau + \frac{h-l}{5} \int_0^t (a_{20}^2(\tau))^2 \, d\tau \right] \equiv F(t), \quad (3.19) \]

\[ A(t) \leq e^{-2lt} \left[ \sqrt{A(0)} + \sqrt{\frac{k_1}{\chi_1}} \left( \frac{2\chi_1}{l^3} \int_0^t e^{\delta\tau} |a_{10}(\tau)| \, d\tau + \sqrt{\frac{l}{5}} \int_0^t e^{\delta\tau} |a_{10}^2(\tau)| \, d\tau \right) \right. \]

\[ + \left. \sqrt{\frac{k_2}{\chi_2}} \left( \frac{2\chi_2}{(h-l)^3} \int_0^t e^{\delta\tau} |a_{20}(\tau)| \, d\tau + \sqrt{\frac{h-l}{5}} \int_0^t e^{\delta\tau} |a_{20}^2(\tau)| \, d\tau \right) \right]^2. \quad (3.20) \]

As to functions \( w_j(y, t), f_j(t) \) the following estimates hold

\[ |w_1(y, t)| \leq 2 \left[ \frac{E(t)}{\nu_1} \left( F(t) + \frac{4\kappa^2l\alpha_1^2(l, t)}{5\mu_1} \right) \right]^{1/4}, \quad (3.21) \]

\[ |w_2(y, t)| \leq \left( \frac{8}{\nu_2} E(t)F_2(t) \right)^{1/4}, \quad (3.22) \]

\[ |f_1(t)| \leq 2 \left[ \frac{E_1(t)}{\nu_1} \left( F_3(t) + \frac{4\kappa^2l(a_1'(l, t))^2}{5\mu_1} \right) \right]^{1/4} + \frac{12\nu_1}{l^2} \left( \frac{8E(t)}{\nu_1} F_2(t) \right)^{1/4}, \]

\[ 264 \]
\[ |f_2(t)| \leq \left( \frac{8E_1(t)}{V_2} F_3(t) \right)^{1/4} + \frac{12\nu_2}{(h-l)^2} \left( \frac{8E(t)}{V_2} F_2(t) \right)^{1/4}, \quad (3.23) \]

where
\[ E(t) \leq e^{-4\delta_1 t} \int_0^t H(\tau)e^{4\delta_1 \tau} d\tau, \quad (3.24) \]

\[ H(t) = \frac{2\kappa}{\varepsilon} \left[ \left( \frac{8\lambda_1}{k_1^2} F(t)A(t) \right)^{1/2} + a_{10}^2(t) \right]. \quad (3.25) \]

The functions \( E_1(t), F_1(t), F_2(t) \) and \( F_3(t) \) have the same structures as \( E(t), F(t) \).

### 4 Stationary Flow

The problem (3.1)–(3.10) has the stationary solution \( w_j^s(y), a_j^s(y), f_j^s \)

\[ w_1^s(y) = \frac{\kappa(1 - \gamma)Ah(3y^2/h^2 - 2\gamma y/h)}{2\gamma\mu_2[\gamma + \mu(1 - \gamma)]}, \]

\[ w_2^s(y) = \frac{\kappa\gamma Ah(3y^2/h^2 - 2(2 + \gamma)y/h + 1 + 2\gamma)}{2(1 - \gamma)\mu_2[\gamma + \mu(1 - \gamma)]}, \]

\[ a_1^s = \frac{(a_{20}^s - a_{10}^s)}{\gamma + k(1 - \gamma)} \frac{y}{h} + a_{10}, \]

\[ a_2^s = \frac{1}{\gamma + k(1 - \gamma)} \left[ k(a_{20}^s - a_{10}^s) \frac{y}{h} + ka_{10}^s + (1 - k)a_{20}^s \right], \]

\[ f_1^s = -\frac{3\kappa\nu(1 - \gamma)A}{\gamma h\rho_2[\gamma + \mu(1 - \gamma)]}, \quad f_2^s = -\frac{3\kappa\gamma A}{(1 - \gamma)h\rho_2[\gamma + \mu(1 - \gamma)]}, \]

\[ a_1^s(0) = a_{10}^s, \quad a_2^s(h) = a_{20}^s, \quad k = k_1/k_2, \quad \nu = \nu_1/\nu_2, \quad \gamma = l/h < 1, \quad \mu = \mu_1/\mu_2, \]

\[ A = \frac{(a_{20}^s - a_{10}^s)\gamma}{\gamma + k(1 - \gamma)}; \quad (4.2) \]

\[ v_1^s(y) = -\frac{\kappa(1 - \gamma)Ah}{2\gamma\mu_2[\gamma + \mu(1 - \gamma)]} \left( \frac{y^3}{h^3} - \frac{\gamma y^2}{h^2} \right), \]

\[ v_2^s(y) = -\frac{\kappa\gamma Ah^2}{2(1 - \gamma)\mu_2[\gamma + \mu(1 - \gamma)]} \left[ \left( \frac{y^3}{h^3} - \gamma^3 \right) \right. \]

\[ -\left. (2 + \gamma) \left( \frac{y^2}{h^2} - \gamma^2 \right) + (1 + 2\gamma) \left( \frac{y}{h} - \gamma \right) \right]. \quad (4.3) \]

Introducing the differences
\[ N_j(y,t) = a_j^s(y) - a_j(y,t), \quad M_j(y,t) = w_j^s(y) - w_j(y,t) \quad (4.4) \]
and carrying out the calculations analogous to those in Section 2, we can prove that the solution of the nonstationary problem reaches the steady regime \( w_j^*(y), \) \( a_j^*(y) \) and \( f_j^* \) under the conditions of convergence of the integrals

\[
\int_0^\infty e^{\delta \tau} |a_{j0} - a_{j0}(\tau)| \, d\tau, \quad \int_0^\infty e^{\delta \tau} |a_{j0}'(\tau)| \, d\tau, \quad \int_0^\infty e^{\delta \tau} |a_{j0}''(\tau)| \, d\tau. \tag{4.5}
\]

More exactly, \( ||w_j(y, t) - w_j^*(y)|| \leq d_j e^{-\delta_1 t}, \) \( ||a_j(y, t) - a_j^*(y)|| \leq l_j e^{-\delta_2 t}, \) \( ||f_j(t) - f_j^*|| \leq N e^{-\delta_3 t} \) with the positive constant \( d_j, l_j, N, \delta_1, \delta_2, \delta_3 \) depending on physical parameters of liquid and layers thicknesses.

5 Nonstationary Motion and Numerical Results

To describe the nonstationary motion of two viscous thermally conducting liquids the Laplace transform will be applied to problem (3.1)–(3.10). As a result we come to boundary value problem for images \( \hat{a}_j(y, p) \) of functions \( a_j(y, t) \)

\[
\hat{a}_{yy} - \frac{p}{\chi_j} \hat{a}_y = -\frac{a_j^0(y)}{\chi_j}, \tag{5.1}
\]

\[
\hat{a}_1(0, p) = \hat{a}_{10}(p), \quad \hat{a}_2(h, p) = \hat{a}_{20}(p), \tag{5.2}
\]

\[
\hat{a}_1(l, p) = \hat{a}_2(l, p), \quad k_1 \hat{a}_{1y}(l, p) = k_2 \hat{a}_{2y}(l, p), \tag{5.3}
\]

and images \( \hat{w}_j(y, p) \) and \( \hat{f}_j(p) \) of functions \( w_j(y, t), f_j(t) \)

\[
\hat{w}_{yy} - \frac{p}{\nu_j} \hat{w}_y = -\frac{f_j(p)}{\nu_j}, \tag{5.4}
\]

\[
\hat{w}_1(0, p) = 0, \quad \hat{w}_2(h, t) = 0, \tag{5.5}
\]

\[
\hat{w}_1(l, p) = \hat{w}_2(l, p), \tag{5.6}
\]

\[
\mu_2 \hat{w}_{2y}(l, p) - \mu_1 \hat{w}_{1y}(l, p) = -2k \hat{a}_1(l, p), \tag{5.7}
\]

\[
\int_0^l \hat{w}_1(y, p) \, dy = 0, \quad \int_l^h \hat{w}_2(y, p) \, dy = 0. \tag{5.8}
\]

In condition (5.2) and equation (5.4) \( \hat{a}_{j0}(p), \) \( \hat{f}_j(p) \) are images of functions \( a_{j0}(t), f_j(t) \) respectively. The solutions of problem (5.1)–(5.8) can be written as

\[
\hat{a}_j(y, p) = C_j^1 \text{sh} \sqrt{\frac{p}{\chi_j}} y + C_j^2 \text{ch} \sqrt{\frac{p}{\chi_j}} y - \frac{1}{\sqrt{p\chi_j}} \int_y^y a_j^0(z) \text{sh} \sqrt{\frac{p}{\chi_j}} (y - z) \, dz, \tag{5.9}
\]

\[
\hat{w}_j(y, p) = -2k \hat{a}_1(l, p) \left[ D_j^1 \text{sh} \sqrt{\frac{p}{\nu_j}} y + D_j^2 \text{ch} \sqrt{\frac{p}{\nu_j}} y + \frac{L_j(p)}{p} \right], \tag{5.10}
\]

\[
\int_0^\infty e^{\delta \tau} |a_{j0} - a_{j0}(\tau)| \, d\tau, \quad \int_0^\infty e^{\delta \tau} |a_{j0}'(\tau)| \, d\tau, \quad \int_0^\infty e^{\delta \tau} |a_{j0}''(\tau)| \, d\tau. \tag{4.5}
\]
where $\hat{f}_j(p) = -2\kappa \hat{a}_1(l, p)L_j(p)$.

The values $C^1_j, C^2_j, D^1_j, D^2_j$ and $\hat{f}_j(p)$ determined from the boundary conditions (5.2), (5.3), (5.5)–(5.8). Due to the cumbersome type of these values is not given here.

Let us assume that $\lim_{t \to \infty} a_{j0}(t) = a^s_{j0}, j = 1, 2$, using the formulas (5.9), (5.10) and presenting for the values $C^1_j, C^2_j, D^1_j, D^2_j$ and $\hat{f}_j(p)$ we can prove the limit equalities

$$
\lim_{t \to \infty} a_j(y, t) = a^s_j(y), \quad \lim_{t \to \infty} w_j(y, t) = w^s_j(y),
$$

$$
\lim_{t \to \infty} f_j(t) = f^s_j;
$$

where $a^s_j(y), w^s_j(y), f^s_j$ are determined by formulas (4.1), (4.2).

Let us apply the numerical method of inversion of Laplace transformation to obtained formulas (5.9), (5.10). The graphs only for the velocities are given because the have a real physical meanings. All numerical calculations were made for the system of liquid silicon–water. Thickness of the layers is the same and equal to 1 mm. The corresponding values of the defining parameters are given in Table 1.

**Table 1. Physical properties of liquids**

<table>
<thead>
<tr>
<th>Item</th>
<th>liquid silicon</th>
<th>water</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho$, kg/m$^3$</td>
<td>956</td>
<td>998</td>
</tr>
<tr>
<td>$\nu \times 10^{-6}$, m$^2$/s</td>
<td>10.2</td>
<td>1.004</td>
</tr>
<tr>
<td>$k$, kg · m/s$^3$ · K</td>
<td>0.133</td>
<td>0.597</td>
</tr>
<tr>
<td>$\chi \times 10^{-6}$, m$^2$/s</td>
<td>0.0675</td>
<td>0.143</td>
</tr>
<tr>
<td>$\alpha \times 10^{-5}$, kg/s$^2$ · K</td>
<td>6.4</td>
<td>15.14</td>
</tr>
</tbody>
</table>

Figure 2–5 show the profiles of the dimensionless functions $\bar{w}_j(\xi, \tau) = w_j(y, t)\mu_2/(\kappa A)$ ($\xi = y/l, \tau = \nu_1 t/l^2$ are the dimensional variables) and transverse velocity $\bar{v}_j(\xi, \tau) = v_j(y, t)\mu_2/(\kappa A h)$ with $a_{20}(t) = 0$. In particular, the functions $\bar{w}_j$ are negative, so reverse flows arise here. Figure 2, 3 show the results of calculations when $a_{10}(\tau) = \sin \tau, a_{20}(\tau) = 0$. That is the limit of $a_{10}(\tau)$ at $\tau \to \infty$ does not exist and the velocity field does not converge to a stationary one.

Figure 4, 5 show an evolution of the convergence of functions $\bar{w}_j$ and transverse velocities $\bar{v}_j$ to stationary regime for the case $a_{10}(\tau) = 1 + e^{-\tau} \cos(10\tau), a_{20}(\tau) = 0$. These results are good agreement with the a priori estimates were obtained in Section 4.
Fig. 2. Evolution of functions $\bar{w}_j$ for $a_{10}(\tau) = \sin \tau$. Total line is the stationary profiles, $\tau = 4, \ldots, \tau = 7$.

Fig. 3. Evolution of functions $\bar{v}_j$ for $a_{10}(\tau) = \sin \tau$. Total line is the stationary profiles, $\tau = 3, \ldots, \tau = 8$.

Fig. 4. Evolution of functions $\bar{w}_j$ for $a_{10}(\tau) = 1 + e^{-\tau} \cos(10\tau)$. Total line is the stationary profiles, $\tau = 1, \ldots, \tau = 4$. 
Fig. 5. Evolution of functions $\bar{v}_j$ for $a_{10}(\tau) = 1 + e^{-\tau} \cos(10\tau)$. Total line is the stationary profiles, $-\cdots - \tau = 1, -\cdot - \tau = 4, -\cdot - \tau = 2$

6 Conclusion

The two-dimensional horizontal layer is a matter of great importance in connection with the theory of convective stability applications in the design of cooling systems, in studying the growth of crystals and films, or in the aerospace industry. We have presented a theoretical and numerical study of a creeping flow of two immiscible viscous heat conducting liquids in thin layers. The flow arises due to heat exchange with the localized parabolic heating of the borders and through the thermocapillary forces on the interface. The following results are obtained: (1) the exact solution describing the stationary thermocapillary convective flow is found; (2) a priori estimates of the initial boundary value problem are established and sufficient conditions on input data when solution tends to stationary one are obtained; (3) the solution of the non-stationary problem in the form of final analytical formulas in the Laplace representation is found and some numerical results of velocities behaviour in layers are presented.

Acknowledgments. This research was supported by the Russian Foundation for Basic Research (14-01-00067).

References

Tensor Smooth Length for SPH Modelling of High Speed Impact

Roman Cherepanov and Alexander Gerasimov
Institute of Applied mathematics and mechanics, Tomsk State University
634050, Lenina av. 36, Tomsk, Russia
RCherepanov82@gmail.com, Ger@niipmm.tsu.ru
http://niipmm.tsu.ru

Abstract. SPH method with tensor form of smoothing parameter is proposed for high speed impact modelling. Calculation of tensor smoothing parameter is based on deformation of local coordinate system and can be obtained from strain tensor evolution. Strain anisotropy in such an approach does not cause mixing of the particles while maintaining the uniform distribution of particles and the resulting solution is more smooth. Weak variational formulation is used to construct numerical integration of motion equations scheme and procedure of restoring of particle consistency is used for calculation of spatial derivatives. Boundary conditions for free surface and contact surface are realized.

Keywords: smoothed particles, variable smoothing length, high speed impact, SPH

1 Introduction

Smoothed particle hydrodynamics was introduced as a numerical method for study of the motion of compressible gas with self-gravity of arbitrary geometry in three dimensions in astro-physics [1], [2]. Particles are used to represent a sub-set of the fluid elements in the Lagrangian description of a fluid, and spatial derivatives are calculated from analytical interpolation formulae. SPH is well suited to problems in which large deformations can occur, such as the fragmentation of self-gravitating gas clouds. The spatial resolution of SPH is determined by particle density and the smoothing length, $h$. Originally $h$ was taken to be constant, and was a globally defined function of the average density of particles in the system. More recently, however, it has become common for each particle to have its own time dependent smoothing length which corresponds to the local neighbors count. Full advantage is then taken of the Lagrangian nature of SPH, and the dynamic range in spatial resolution of the method is increased. Time-dependent smoothing lengths can, however, lead to problems with energy conservation in certain situations [4]. The problem arises from the fact that the use of variable smoothing lengths means that additional terms should appear in the particle equations of motion. Other important problem is lack of accuracy when variable smoothing length is used. This problem arises from the fact,
that SPH approximation originally has form of analytical interpolation, based on integration, and there is some difference between analytical integration and its approximation via summation over number of particles. Nonuniform particle distribution lead to particle inconsistency problem, as a presence of boundary and variable smoothing length lead to one too. In should be noted, that high speed impact is accompanied by large deformations. When smoothing length is a scalar, smoothing kernel should have spherical symmetry, and large anisotropic deformations in this case can cause non-physical numerical fracture and particle mixing. In this paper we propose conservative SPH with variable time-dependent tensor smooth parameter, and free surface boundary condition algorithm is proposed too.

1.1 SPH basics

Basis of SPH approximation is equation

\[ f^i \approx \int f(x)W(x - x^i, h)dx \]  

(1)

where \( h \) is smoothing parameter, which define a radius of fluency for points, \( x \) – is a space coordinate, \( W \) - smoothing function.

\[ W(r, h) = \frac{C}{h^3} \phi \left( \frac{|r|}{h} \right) \]  

(2)

where \( C \) - integration constant, \( \phi(u) \) - cubical b-spline

Integration constant defined as

\[ C = \int_V \phi(u)dV \]  

(3)

where \( V \) is 1D, 2D or 3D volume. Function \( \phi(u) \) usually is cubical B-spline:

\[ \phi(u) = \begin{cases} 
0, & u \geq 2; \\
0.25 \cdot (2 - u)^3, & u \in (1, 2); \\
0.75 \cdot (1.0 - u^2) (2 - u), & u \in [0, 1]; 
\end{cases} \]  

(4)

Spatial derivatives are defined via:

\[ f^i_{,\alpha} = \frac{\partial f^i}{\partial x_\alpha} \approx \int f(x)W_{,\alpha}(x - x^i, h)dx \]  

(5)

Corresponding to (5) particle approximation is written as:

\[ f^i_{,\alpha} = \frac{\partial f^i}{\partial x_\alpha} \approx \sum_k f^k W_{,\alpha}(x^k - x^i, h) \frac{m^k}{\rho^k} \]  

(6)

where \( x^k, f^k, m^k, \rho^k \) – radius-vector, approximated function value, mass and density at \( k \)-th point.
Equation (6) have $C^0$-consistency [3] near surfaces and boundaries or at non-uniform particle distribution, and for restoring particle consistency special approaches is needed. According [3], a test vector-function is introduced in form:

$$\Delta(\mathbf{x}) = (1, x_0, x_1, x_2)$$ (7)

Approximation of test function (1) and its derivatives (5) can be found with $C^0$-consistency, but due to the fact that this values are known, this approximation can be used to construct correction procedure and restore particle consistency. Let’s define:

$$\Delta(\mathbf{x})_\alpha = \begin{cases} x_\alpha; & \alpha > -1; \\ 1; & \alpha = -1; \end{cases}$$ (8)

Matrix of approximations for this function at $n$-th point, found via uncorrected SPH approximation (6):

$$T^\beta_\alpha(\mathbf{x}^n) = \sum_m \Delta_\alpha(\mathbf{x}^m - \mathbf{x}^n) W_{\beta_\gamma}(\mathbf{x}^m - \mathbf{x}^n, h) \frac{m_k}{\rho_k};$$ (9)

Lets define correction matrix as

$$B^n_{\alpha\beta} = B_{\alpha\beta}(\mathbf{x}^n) = \left[T_{\alpha\beta}(\mathbf{x}^n)\right]^{-1}; \alpha, \beta = -1, 0, 1, 2;$$ (10)

Now, corrected approximation of function $f(x)$ value or its derivatives can be found, if we build full matrix of uncorrected approximations (6) and apply correction matrix (10):

$$f^i_\alpha = \frac{\partial f^i}{\partial x_\alpha} \approx T^{i\beta}_{\alpha\beta} \left\{ \sum_k \frac{m_k}{\rho_k} f^k W_{\beta_\gamma}(\mathbf{x}^k - \mathbf{x}^i, h) \right\};$$ (11)

Smoothing length $h$ defines maximum interparticle distance and therefore defines particle count. From the other hand, particle count is related to interparticle distance and smoothing length- greater particle count allow to use smaller smoothing length. There is some analogy between smoothing length in SPH and space step in FEM/FDM.

For calculation of elastic plastic flow, approximation (11) is applied to Lagrangian of system:

$$L = \sum_k m^k \left( \frac{\mathbf{v}^k \cdot \mathbf{v}^k}{2} - u^k \right)$$ (12)

where $\mathbf{v}^k$- is particle speed, $u^k$- is internal energy (per mass). Internal energy change can be written in form:

$$\frac{du^k}{dt} = \sigma^k : \dot{\varepsilon}^k$$ (13)

where $\sigma^k$ is stress tensor and $\dot{\varepsilon}^k$ is strain rate tensor of particle $k$ respectively.
This implies:

\[ \frac{dL}{dt} = \sum_k m^k \left( v^k a^k - \sigma^k : \varepsilon^k \right) \]  

(14)

Since the relation between a small deformations rate tensor and nodal velocities is linear:

\[ \varepsilon^p_{ij} = \frac{1}{2} \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) = \sum_n \frac{m^n}{\rho^n} W^n_{ik} \left( v^n_i T^n_{jk} + v^n_j T^n_{ik} \right) \]  

(15)

Because of energy conservation (14) and (15) leads to the relation between the accelerations of nodes and stress field:

\[ a^n_i = - \left( \sum_k \frac{m^k}{\rho^k} \sigma^k_{ij} T^k_{ij} W^k_{kn} \right) / \rho^n \]  

(16)

For impact problem we use a condition of zero normal stress at free surface and equal normal stress and equal speeds on contact surface.

In described method this boundary conditions does not need additional operations, such as "ghost particles" or special integration procedures for (5) to treat a boundary.

Smoothing length \( h \) was selected to provide sufficient count of neighbors for each particle for correction matrix to be well defined. Usually neighbor count was from 11 to 16.

Time step is defined via Courant–Friedrichs–Lewy condition: \( \Delta t < \min \left[ \frac{h}{C} \right] \), where \( C \) is the sound speed. At practice stability condition for heat transfer problem is more weak while droplet radius exceed 1 mm.

When smoothing length \( h \) is scalar parameter intensive deformations lead to particle mixing and artificial fragmentation. To avoid this effects we introduced a tensor smoothing parameter \( h_{ij} \):

\[ W(\mathbf{r}, h_{ij}) = \frac{C}{\det|h_{ij}|} \phi \left( \sqrt{\sum_i \left( r_i \cdot [h]^{-1}_{ij} \right)^2} \right) \]  

(17)

Initial value \( h_{ij} = hI \), equation of evolution of \( h_{ij} \) is:

\[ \dot{h}_{ij} = h_{ik} v_{j,k} \]  

(18)

For scalar smoothing parameter spatial derivatives of smoothing function have the form:

\[ \frac{\partial W(\mathbf{r}, h)}{\partial \mathbf{r}} = \frac{C}{h^3} \frac{\partial \phi(|\mathbf{r}/h|)}{\partial \mathbf{r}} = \frac{C}{h^3} \frac{r \cdot h^{-1}}{r^2} \left. \frac{\partial \phi}{\partial u} \right|_{u=(|\mathbf{r}/h|^{-1})} \]  

(19)

for tensor smoothing ones have similar form:

\[ \frac{\partial W(\mathbf{r}, h_{ij})}{\partial \mathbf{r}} = \frac{C}{\det|h_{ij}|} \frac{r \cdot [h_{ij}]^{-1}}{r^2} \left. \frac{\partial \phi}{\partial u} \right|_{u=(|\mathbf{r}/h|^{-1})} \]  

(20)
3 Results

A 38 $\mu$m radius stainless steel cylinder with length of 190 $\mu$m is impacting with a velocity of 200 m/s on stainless steel substrate 64 $\mu$m thin. Parameters of steel: density 7.87 g/cm$^3$, $E=200$ GPa, $G=70$ GPa, $Y_{pl}=1.2$ MPa. As it shown on Fig. 2, when impact speed is relatively small, results are similar.

Fig. 2, when impact speed is relatively small, results are similar. Smoothing tensor remains almost spherical in this condition and smoothing kernel with tensor parameter differs slightly from standard kernel with scalar smoothing length. More different results are obtained for impact speed 800 m/s (Fig. 3). Tensile instability and particle mixing are not observed when tensor smoothing is used, and shape of material ejection is tracked more accurately. At the same time, usage of constant scalar smoothing length lead to particle mixing, numerical fracture (as it seen on bottom part of Fig. 3-I) and intensive particle clustering. Also, symmetry of results obtained with tensor smoothing is much better.

1.2 Conclusion

Introduction of tensor smoothing length in variational SPH-code is quite simple and does not need significant calculation cost, and provide more accurate results for high speed impact modelling.
Fig. 3. Scalar smoothing (I) and tensor smoothing (II) comparison at \( t=0.03 \, \mu s, \, v=800 \, m/s \).

**Acknowledgments.** This work was supported by the Russian Science Foundation (RSF) No. 16-19-10264.

**References**

Multiscale Modeling of Strength Properties of Dispersion-Reinforced Ceramic Composite Materials

Yuriy Dimitrienko, Yulia Zakharova, and Sergey Sborschikov

Computational Mathematics and Mathematical Physics Department,
Bauman Moscow State Technical University
2 Baumanskaya Str., 5, Moscow, 105005, Russia
{dimit.bmstu@gmail.com, shpakovayuliya@bmstu.ru, servasbor@gmail.com}

Abstract. In this work a three-level model of ceramic composites materials based on a reaction bonded silicon carbide is developed. Numerical solution is based on the method of multiscale homogenization along with the finite element method. As a result a series of local problems on the periodical cells of 3 structure levels are solved. The calculations of stress concentration tensors in matrixes and weighing materials are presented. New criteria for matrixes and weighing materials is used to calculate the strength properties in multiaxis stressed condition. This criteria includes essential differences (more than an order of magnitude) of ceramic properties under straining and compression. The model which includes scale effect of strength of ceramic composite materials is proposal. The computational research of sequential micro-destruction processes of ceramic composite until complete destruction is done. The results show that changing of concentration of larger fractions is less significant then content of smaller fractions in the presence of polydisperse structure in ceramics.

Keywords: ceramic composites, reaction bonded SiC, microdestruction, numerical simulation, finite-element method, multiscale homogenization method, strength criterion, stress concentration tensor, scale strength effect

1 Introduction

Composite materials based on reaction bonded silicon carbide matrix (RBSiC) and SiC disperse filler are perspective materials for creation of shockproof protecting systems because of their high strength, stiffness, destruction energy and relatively low cost. However characteristics of this materials significantly depend on manufacturing technological processes and on receipt of composite components. In addition during the hardening details can give strong shrink, giving significant residual stresses. This residual stresses can give deformation and even breakdown in the final product. To select the optimal content of ceramic composite components of SiC system and to calculate strength properties of such
It is demanded to develop a special mathematical model, which can forecast the strength properties of composite materials including variation of content, form and disperse filling. This model should also take into account locked-up stresses appearing in ceramic composite during agglomeration of particles.

Existing analytical and numerical models of composite materials, armed with particles, allow to forecast elastic properties with certain precision; however, numerical calculation of strength properties is a much more complex problem, because it is necessary to build the appropriate model of microcrack emission in a heterogeneous structure. Attempts to build such models using simple concentration of finite-element mesh were not successful because of dramatic increase of non-physical singularity effect of calculation. Widely known commercial software does not always allow to get adequate results of microdestruction modeling of composite.

Nowadays the great attention is paid to the development of numerical finite element methods of microstress modeling in composites [4, 18, 21]. One of the most efficient methods for calculation of microstress in composites is the method of asymptotic averaging (MAA) (or homogenization method) [2, 3, 5, 23, 25], which ensures the high accuracy of calculation of microdestruction in mathematical terms. Possible errors of calculation by this method can be related only to errors of its numerical application as well as to inaccurate specifications of the component characteristics and the microstructure geometry. In [1, 6–8, 13] algorithms of finite element solution of the so-called local problems in periodicity cells, which appear when MAA is used, were developed.

This work continues the development cycle [14, 16, 17, 20] of creating models and numerical methods for modeling microdestruction processes in composite materials. The new 3-level strength model of ceramic composite is presented. This model is based on RBSiC and allows to describe the effect of strengthening of composite material during the changing particles content of SiC including the production technology of its manufacturing.

2 Microstructure of the reaction bonded silicon carbide composite

A composite based on reaction bonded silicon carbide consists of a filler and a silicon carbide matrix. The filler is powder of silicon carbide of the different fractions. The silicon carbide matrix is synthesized by chemical reaction of liquid silicon, carbon and solid carbon, which is produced during the pyrolysis of phenol-formaldehyde resin [22, 24]. The filler is, as a rule, fission fragments which have random character and big difference in fractions. Generally it can be identified large fractions of the size of 20-100 microns and small fractions of 1-10 microns. Photographies of real microstructure of RBSiC are shown in Fig. 1.

We consider a model of reaction bonded silicon carbide composite material which has three structure levels [5, 15, 6, 19, 26] (Fig 2). The first level is formed by the periodicity cells 1 (PC1) consisting of a filler of a coarse fraction and a matrix $m_1$. On the second level the matrix $m_1$ is formed by the periodicity...
cells 2 (PC2), each of them consists of a filler of a fine-grained fraction and a reaction bonded silicon carbide matrix $m_2$. The matrix $m_2$ has defects, for example, high concentration of dissolved, but unreacted component C and Si, microcracks due to technological stresses and mainly pores. So, we introduce the third structural level formed by the periodicity cells 3 (PC3). Each periodicity cell of type 3 is formed by a zero defect silicon carbide matrix $m_3$ and a defect.

### 3 Mathematical formulation of local problems

All structural levels may be considered as independent according to the method of multiscale homogenization [17]. At first we compute the effective elastic and strength properties of the third level, then we calculate the effective characteristics of the second level, considering the composite matrix $m_2$ as a homogeneous
material with effective characteristics of the third level, and then we calculate the characteristics of the first level.

Consider the solution of local problems for the periodicity cell of the second level having the volume \( V_{\xi} \). It includes the matrix \( m_2 \) and fine-disperse filler. We believe that the PC2 has three-axial symmetry, therefore instead of a full volume of PC2 \( V_{\xi} \) we can consider its 1/8th part of volume \( \tilde{V}_{\xi} \). This volume \( \tilde{V}_{\xi} \) consist of the \( N \) components: \( N - 1 \) pieces of fine-disperse particulate of filler of the volume \( V_{\xi \alpha}, \alpha = 1...N - 1 \), and binding matrix \( m_2 \) (\( \alpha = N \)). For calculating microstresses in PC2 by homogenization method [15, 11, 12] we formulate a series of the so-called local problems \( L_{pq} \) of the elasticity theory on the 1/8th part of the periodicity cell

\[
\begin{align*}
\sigma_{ij(pq)}/j &= 0, \quad \tilde{V}_{\xi} \\
\sigma_{ij(pq)} &= C_{ijkl}(\xi_s, z)(\varepsilon_{kl(pq)} - \alpha_{kl}(\theta - \theta^*)), \quad \tilde{V}_{\xi} \cup \Sigma'_s \cup \Sigma_s \\
\varepsilon_{ij(pq)} &= \frac{1}{2}(\bar{U}_{i(pq)}/j + \bar{U}_{j(pq)}/i), \quad \tilde{V}_{\xi}, \\
[U_{i(pq)}] &= 0, \quad [\sigma_{ij(pq)}]n_j = 0, \quad \Sigma_{\xi_{\alpha}N},
\end{align*}
\]

where \( p \) and \( q \) are the indexes of the local problems changing from 1 to 3 (there are a total of nine different problems \( L_{pq} \)); \( U_{i(pq)}(\xi_s) \) are the components of the displacement vectors (the unknown functions) in the problem \( L_{pq} \); \( \sigma_{ij(pq)}, \varepsilon_{ij(pq)} \) are the components of the stress and deformation tensors in \( \tilde{V}_{\xi} \); \( \xi_s \) are the local Cartesian coordinates in the 1/8th PC; \( \partial / \partial \xi_i \) are the derivatives of the local coordinates; \( [U_{i(pq)}] \) are the jumps of functions at the interface \( \Sigma_{\xi_{\alpha}N} \) of the cell components; \( C_{ijkl}(\xi_s, z) \) are the components of the tensors of the elasticity moduli of the composite structural components of PC2 (they are described by the dependencies of the coordinates \( \xi_s \)); \( z \) is the parameter of the component damageability; \( \alpha_{kl}(\theta) \) are the components of the tensor of thermal expansion, which depend on the temperature; \( \theta \) is the current temperature; \( \theta^*(\xi_s) \) is the sintering temperature of the ceramic particles, depending on the local coordinates.

System (1) is supplemented by the special boundary conditions at the surfaces \( \Sigma'_s = \{ \xi_s = 0.5 \} \) of the 1/8th part of PC

\[
\begin{align*}
at \Sigma'_j: & \quad U_{i(pp)} = 1/2\varepsilon_{ip}\delta_{ip}, \quad S_{j(pp)} = 0, \quad S_{k(pp)} = 0, \quad i \neq j \neq k \neq i, \\
at \Sigma'_j: & \quad U_{i(pq)} = (1/4)\varepsilon_{ip}\delta_{ip}, \quad S_{j(pq)} = 0, \quad U_{k(pq)} = 0, \quad i, j = \{ p, q \}, \\
at \Sigma_k: & \quad S_{i(pq)} = 0, \quad S_{j(pq)} = 0, \quad U_{k(pq)} = 0, \quad i \neq j \neq k \neq i,
\end{align*}
\]

where \( \varepsilon_{pq} \) are the components of the averaged deformation tensor for PC, \( S_{i(pq)} \equiv \sigma_{il(pq)}n_l \) are the vectors of forces.

The boundary conditions at the symmetry planes \( \Sigma_s = \{ \xi_s = 0 \} \) are similar to relations (2), where we assume \( \varepsilon_{pq} = 0 \).

4 Effective elastic characteristics of the periodicity cells of the second level structure

Using the numerical solution of problems \( L_{pq} \) (1), (2) we find the fields of displacements \( U_{i(pq)} \) and stresses \( \sigma_{ij(pq)}(\xi_s) \) in the PC2 at given values of average deformations \( \varepsilon_{kl} \).
These fields are used to find the average values of stress:

\[ \bar{\sigma}_{ij} = \langle \sigma_{ij} \rangle = \frac{1}{3} \sum_{p,q} \bar{\sigma}_{ij(pq)}, \]

where

\[ \bar{\sigma}_{ij(pq)} = \langle \sigma_{ij(pq)} \rangle = \int_{V_\xi} \sigma_{ij(pq)}(\xi_s)dV_\xi. \quad (3) \]

Then the components of the tensor of effective elasticity moduli of the composite are calculated by the formulas

\[ \bar{C}_{ijpq} = \bar{\sigma}_{ij(pq)} \bar{\epsilon}_{pq}, \quad (4) \]

where there is no summation over \( p \) and \( q \). After that we calculate the effective tensor of elastic compliances \( \bar{\chi}_{ijpq} \), that is inverse to \( \bar{C}_{ijpq} \), and technical elastic constants of the composite, such as effective Young moduli \( E_\alpha = 1/\bar{\Pi}_{\alpha\alpha\alpha\alpha} \), effective Poisson constants \( v_{\alpha\beta} = -\bar{\Pi}_{\alpha\alpha\beta\alpha} E_\alpha \), and effective shear moduli \( G_{\alpha\beta} = \bar{C}_{\alpha\beta\alpha\beta} \).

The components of the tensor of stress concentrations \( B_{ijkl}^{(\alpha)} \) connect microstresses \( \sigma_{ij}^{(\alpha)}(\xi_s) = \sum_{p,q} \sigma_{ij(pq)}(\xi_s) \) in the matrix and the filler (the fine disperse particles SiC) with average stresses \( \bar{\sigma}_{kl} \) in the PC2 by the formulas

\[ \sigma_{ij}^{(\alpha)}(\xi_s) = B_{ijkl}^{(\alpha)}(\xi_s) \bar{\sigma}_{kl}, \quad \xi_s \in \tilde{V}_{\xi\alpha}, \quad \alpha = 1...N. \quad (5) \]

The components \( B_{ijkl}^{(\alpha)} \) in the matrix and the filler are calculated by the formulas

\[ B_{ijkl}^{(\alpha)} = \sigma_{ij(pq)}(\xi_s) \bar{\Pi}_{pqkl}, \quad \xi_s \in \tilde{V}_{\xi\alpha}, \quad \alpha = 1...N. \quad (6) \]

5 Model of the strength properties of the components

The strength criterion of ceramic materials should take into account the significant differences in their properties in tension and compression. Therefore, we introduce a failure criterion of isotropic matrix \( m_2 \) and filler particles [19] based on Pisarenko-Lebedev criterion:

\[ z = \frac{\sigma_u^{(\alpha)^2}}{3\sigma_S^{(\alpha)^2}(1 + B^{(\alpha)}V(\sigma_\alpha^{(\alpha)}))}, \quad (7) \]

where \( \sigma_\alpha^{(\alpha)} = \frac{1}{2}(|\sigma^{(\alpha)}| - \sigma^{(\alpha)}) \), \( \sigma^{(\alpha)} = \sigma_{11}^{(\alpha)} + \sigma_{22}^{(\alpha)} + \sigma_{33}^{(\alpha)} \) are the invariants of the stress tensor in the matrix and fillers, \( \sigma_u^{(\alpha)} \) are stress intensity [10], \( B^{(\alpha)} = \left( \frac{\sigma_c^{(\alpha)^2}}{3\sigma_S^{(\alpha)^2}} - 1 \right) \frac{1}{\sigma_c^{(\alpha)}} \) is the constant, \( \sigma_c^{(\alpha)}, \sigma_T^{(\alpha)}, \sigma_S^{(\alpha)} \) are the ultimate compression...
strength, ultimate tensile strength and ultimate shear strength. For ultimate
strengths the following relationships should be taken into account: \( \sigma_C > \sqrt{3} \sigma_S \),
\( \sigma_C > 0, \sigma_S > 0 \). In (7) a continuous positive function of the 1st invariant \( V(\sigma^{(a)}_{-}) \)
is introduced
\[
V(\sigma^{(a)}_{-}) = \begin{cases} 
0, & \sigma^{(a)} > 0, \\
-\sigma^{(a)}, & \sigma^{(a)}_C < \sigma^{(a)} < 0, \\
\sigma^{(a)}_C, & \sigma^{(a)} < -\sigma^{(a)}_C.
\end{cases}
\] (8)
The failure criterion \( z \), which is calculated by the formula (7), has the value 0 if the stress is absent in the composite. It is ranged within \( 0 < z(\sigma^{(a)}_{ij}) \leq 1 \) in the loaded condition if there is not damage. And it takes the values \( z(\sigma^{(a)}_{ij}(\xi_s)) \geq 1 \), if the fracture initiation occurs at some point \( \xi_s \). If the failure criterion reaches the value \( z = 1 \), then we obtain strength surface of a component
\[
\sigma^{(a)} u^2 = 3\sigma^{(a)}_S^2 (1 + B^{(a)} V(\sigma^{(a)}_{-})).
\] (9)

In the tensile area \( \sigma^{(a)} > 0 \) the strength surface is the von Mises ellipsoid \( \sigma^{(a)} u^2 = 3\sigma^{(a)}_S^2 \). In the compression area \( -\sigma^{(a)}_C < \sigma^{(a)} < 0 \) the tensile strength is increased. And in the ”supercompression” area \( \sigma^{(a)} < -\sigma^{(a)}_C \) the strength surface again is the von Mises ellipsoid, but with the modified tensile strength: \( \sigma^{(a)} u^2 = \sigma^{(a)}_C^2 \).

If the condition \( z(\sigma^{(a)}_{ij}(\xi_s)) \geq 1 \) is satisfied at the point \( \xi_s \) or in a certain area PC2, there is no complete destruction. This is partial destruction of PC2, hereinafter called microdestruction. Introduce the dependence of the components of the elastic modulus of the failure criterion for accounting microdestruction of components in the model:
\[
C_{ijkl}(\xi_s, z) = (1 - h(z(\sigma^{(a)}_{ij}(\xi_s) - 1)))C^{(a)}_{ijkl}, \quad \xi_s \in \tilde{V}_\alpha, \quad \alpha = 1...N,
\] (10)
where \( C^{(a)}_{ijkl} \) are the components of the tensor of elasticity moduli of the composite components (they are constants). According to the formula (10) if microdestruction occurs at the point \( \xi_s \), elasticity modulus is equal to zero at this point.

To calculate the strength of the composite as a whole, we need to calculate the limit values of average stresses \( \bar{\sigma}_{kl} \). At this stresses an initial microdestruction occurs at least in one of its components (fillers or matrix) in a point \( \xi^* \in \tilde{V}_{\xi} \) at a time \( t^* \), and then complete destruction occurs. For the calculation of limit values of stresses in experimental research usually implement a process of linear load, in which the average stresses are proportional to the time: \( \bar{\sigma}_{kl}(t) = \bar{\sigma}_{kl} t \), where \( \bar{\sigma}_{kl} \) are the components of the stress gradient tensor. Substituting (5) in the strength criterion of the matrix or fillers (7) we obtain the initial failure condition of composite
\[
\max_{\xi_s \in \tilde{V}_{\xi}} \{ z(B^{(a)}_{inkm}(\xi_s)\bar{\sigma}_{km}(t^*)) \} = 1,
\] (11)
where \( \xi_s = \xi^*_s \) are coordinates of the point in the PC2, \( t^* \) is the time point at which the condition (11) is executed first, \( \bar{\sigma}_{km}(t^*) \) are limit stresses.
After appearance of the initial failure elastic moduli are changed in the destroyed areas of the matrix and/or fillers in accordance with the model described above. With further increase in average stress values $\bar{\sigma}_{km}(t)$ failure condition (11) is satisfied in a large number of points of PC2, that is, there is the process of propagation of microdestruction. Some area $V_{\xi'}(t)$ of the partial destruction of the composite is formed in the periodicity cell 2.

For modeling of effective elastic and strength properties of PC3 and PC1 is used a similar method. The stresses occur due to the thermal strain $\varepsilon_{\theta}^{0}_{kl} = \alpha_{kl} (\theta - \theta^*)$ of the ceramic composite during cooling after the laser sintering.

6 Details of numerical simulation

The local tasks (1), (2) are solved by a finite element method which is described in [16, 17, 20]. We use 4-node tetrahedral finite elements, generated by open-source grid generators. The meshes contain different numbers of nodes (from $10^4$ to $10^6$). Meshes with a large number of finite elements are used in the calculation of effective elastic moduli, when micro destruction is not happened. After microdestruction is beginning, the local tasks become nonlinear, because the elastic modulus of the matrix or fillers is changed, so we use iterative method to solve it. The number of iterations to achieve complete destruction is about $10^3$, so for these tasks we use meshes with a smaller number of elements to reduce the time of the numerical experiments. A numerical solution of large systems of linear algebraic equations, preprocessing and postprocessing, including 3D visualization and animation, was implemented in the software package, developed by the scientific and educational center "Supercomputer Engineering Simulation and Development of Software Packages" of the Bauman Moscow State Technical University.

7 Results

7.1 Numerical simulation of microdestruction of ceramics for periodicity cell of the third level

Consider the numerical simulation of microdestruction of ceramics for periodicity cell of the third level with the following properties of SiC matrix $m_3$: elastic modulus $E_m = 320$ GPa, Poisson’s ratio $\nu_m = 0.35$, ultimate strength $\sigma_{0Tm} = 0.07$ GPa; $\sigma_{0Cm} = 4$ GPa; $\sigma_{0Sm} = 0.06$ GPa. We suppose that the pores have a spherical shape. Fig. 3 shows some of the results of microstresses calculations in the PC3. Fig. 3a) shows the distribution of component $B_{1111}^{(1)}$ of the stress concentration tensor in the PC3, where concentration of pore before the start failure is equal to 20%. Fig. 3b) shows the distribution of parameter of damage $z$ in the PC3 under tension in the direction of $Ox_1$.

Fig. 4 shows the process of microdestruction in the PC3 (matrix with defect) under compression. The failure of the periodicity cell starts on the surface of the pore (defect) and at first is spread in a direction perpendicular to load direction,
7.2 Numerical simulation of microdestruction of ceramics for periodicity cell of the second level

Fig. 5 shows the results of numerical solution of the process of microdestruction in the periodicity cell of the second level under compression. These results are
calculated taking into account initial technological stresses, which occurs from the application of laser sintering.

![Fig. 5.](image)

**Fig. 5.** The process of microdestruction of PC2 under compression taking into account initial technological stresses.

Figure 6 shows the tensile strength of ceramic material depending on concentration of coarse-grained fractions of SiC particle taking into account initial technological stresses.

![Fig. 6.](image)

**Fig. 6.** The tensile strength of ceramic material depending on concentration coarse-grained fractions of SiC particle taking into account initial technological stresses.

8 Conclusions

A mathematical model of microdestruction of reaction bonded silicon carbide has been developed. This model is based on the homogenization method and the finite element method for solution of local problems on periodicity cells. The new strength criterion of ceramic materials has been applied. The comparison with
experimental data has shown that this criterion is applicable to solve the problem of microdestruction of the reaction-bonded silicon carbide. It is demonstrated that the developed model allows to simulate the processes of microdestruction of the ceramic composite and can be used as a tool for research and design of new materials with specified properties.

References

Single and Group Impacts of High-speed Elements on Spacecraft

Alexander Gerasimov, Yury Khristenko, and Sergey Pashkov

Tomsk State University,
Lenina. 36, 634050 Tomsk, Russia
ger@mail.tomsknet.ru, hrs@niipmm.tsu.ru, ps@contek.ru
http://www.tsu.ru

Abstract. Creating a reliable system of spacecraft protection from space debris fragments of different shapes and sizes necessitates studying features of interactions of high-speed elongated projectiles with the protected objects. In this paper we consider interactions of single and groups elongated projectiles-rods with a system of spaced layered plates concerning to protection of space and ground facilities by combined barriers. A probabilistic approach to fragmentation of solids under shock loading and proposed numerical technique fully from the physical point of view in three-dimensional formulation enables with sufficient accuracy to reproduce the processes of high-speed elements penetration into multi-layered spaced barriers. In the calculations fragmentation fields were simulated taking into account the interactions of fragments with each other and with the elements of the multi-layered barriers. The results enable to optimize the objects protection by mass-geometric parameters.

Keywords: Numerical simulation, experiment, high-speed projectiles, probability, fragmentation, space debris, protection, high-speed collision, spacecraft, destruction, layered barriers

1 Introduction

The presence in the formed surface layer (ranged from 300 to 2000 kilometers) of a huge number of man-made fragments of various sizes and shapes, formed in the process of destroying the satellites, the last stages of carrier rockets, boosters and other vehicles and equipment, represents a serious threat to the security of automatic and manned space objects. Now the problem of interaction of constructions with high-speed projectiles takes on special significance due to the increasing speed of collisions. It increases the probability of penetration and destruction and violates the normal functioning of the protected objects. In recent years reliable protection of the elements of manned and automatic apparatuses intended to study near-earth and deep space is especially an acute problem due to the increasing duration of flights of these objects. It increases the probability of collisions of these objects with the man-made fragments formed because of the destructions of the orbital constructions. Numerical simulation of high-speed
interaction of solids with the protective systems enables to reproduce typical characteristics of the physical processes occurring in the collision, to consider and select the optimum screen circuits.

The penetration along the normal to the surface of the layered barrier was considered in [1, 2]. Between the layers of metal plates we placed a layer of ceramics. In [3] the process of interaction of group of projectiles with the barrier was numerically simulated using erosion criterion to describe the destruction of the barrier material. The results of the ballistic experiment [4, 5] were compared with the computer simulation data. The comparison was performed on the following parameters: residual speed and the residual rod length after the penetration into the first barrier. Throwing of rods was performed on a light-gas two-stage installation of GU-23. In general in the above papers the authors considered single-layer plates and normal impact. It should be noted that for practical problems spaced barriers and impact at an angle are of great interest. For the numerical solution of these problems a reliable and fairly universal method is required to adequately reproduce the processes of destruction and fragmentation occurring in solids at high-speed interaction.

The use of up-to-date computers and numerical methods in this study to solve the problems of high-speed collisions in a three-dimensional formulation, taking into account fragmentations of projectiles and protective elements of spacecrafts, is theoretically and practically an important task. Accounting for fragmentation and interaction of fragments with each other and with the spacecraft body enables to better understand the processes proceeding at high-speed interaction of space debris with a shell of the space object.

Accounting for fragmentation of solids at intensive dynamic loading enables to use the Lagrange approach to the problem of high-speed impact in a fairly wide range of interaction speeds. This approach is especially useful when considering the multicontact interactions of colliding bodies, especially solving the three-dimensional impact problems. The initial structure heterogeneity of real materials affecting the character of the distribution of physical and mechanical characteristics of the material in body volume is an important factor determining the nature of the fracture. One way to account for this fact is the introduction into the equations of solid mechanics random distribution of deviations of the initial strength properties from the nominal value, that is, simulation, thus the initial structural features of the material, namely: the presence of pores, inclusions, dislocations, etc.

2 Basic relations

In this paper in 3D Lagrangian formulation the process of high-speed interactions of spaced layered plates with elongated fragments is considered.

To describe the processes of deformation and crushing of solids we used a model of an ideal compressible elastic-plastic body. Key equations describing the motion of the medium are based on the laws of mass, momentum and energy conservations and Prandtl-Reuss equations and von Mises yield criterion [6, 8].
The equation of state was taken in the form of Tate and Mi - Grüneisen [6]. It is known that plastic deformations, pressure and temperature affect the yield stress and the shear modulus, so the model was supplemented by the relations, which were proven in [9].

To calculate the elastic-plastic flows we used a technique implemented on the tetrahedral cells and based on the combined use of the Wilkins method [8] to calculate the internal points of the body Johnson’s method and [10, 11] to calculate contact interactions. Dividing three-dimensional region into tetrahedrons occurred sequentially by means of automatic meshing routines.

As the fracture criterion under intensive shear deformation we used achievement of the equivalent plastic deformation of its limit value [12].

The initial structure heterogeneities were simulated distributing limit equivalent plastic strain into cells of the computational domain using a modified random number generator which issued a random variable subjected to the chosen distribution law. The probability density of random variables was taken as a normal Gaussian distribution with the arithmetic mean equal to the tabulated value and variable dispersion. The equations of solid mechanics, used in current papers on dynamic fracture of constructions and materials, do not take into account the probability factor in the problem of solids fragmentation that can significantly distort the real picture of the impact and explosive fractures of the objects under consideration. This is particularly evident in the solution of axisymmetric problems, where all the points on a circumferential coordinate of the calculated element are initially equal due to standard equations of continuum mechanics used at numerical simulation.

However in practice there is a wide range of tasks where fragmentation is largely random process, for example, the explosive destruction of axisymmetric shells, where the nature of fragmentation is not known beforehand, penetration and destruction of the thin barrier by a projectile along the normal to the surface, the so-called "petaling", and so on [13]. Adding of a random distribution of the initial deviations of the strength properties of the nominal value in the physical and mechanical characteristics of the body leads to the fact that in these cases the destruction process becomes a probabilistic process, which is more consistent with the experimental data used in this study. The most complete the ideology and methodology of the probabilistic approach to the problem of solids destruction is given in [14].

3 Test calculations

The problem of extension of a copper shell with a steel ring under the impact of detonation products was considered [14]. A computational grid used in this calculation was approximately 500 thousand tetrahedral cells. To describe destruction we used a method of splitting nodes. When performing failure criterion a splitting of nodes and the formation of the fracture surface occurred. To simulate the initial heterogeneities we used distribution of the limiting value of
equivalent plastic strain in the cells of the computational domain by the normal law with a variance of 10% deviation.

With the extension of the ring one observed localization of deformations on the tops of the radial cracks formed at the initial heterogeneities and the formation of fairly large fragments. The calculated fragmentation spectrum quite satisfactory agreed with the experimental data [15].

In the three-dimensional formulation was considered the problem of penetration of the single-layer barrier (glass fiber ST-NT +alloy D16) by a steel ball SH-15 [1]. The calculations of the ball collision with the barrier were performed at the normal to the surface. Projectile speed was 700 and 900 m/s. Comparison of numerical results with experimental data showed quite satisfactory coincidence.

In the three-dimensional formulation was considered the problem of the penetration of two- and three-layered barriers (steel-ceramics and steel-ceramics-steel) by cylindrical projectile of tungsten alloy [1]. The comparison of the numerical results (n) and the experimental (e) data showed good coincidence of the remaining lengths (ln and le) and speeds (Vn and Ve) of the projectiles for the two- and three-layered barriers. Two-layered barrier: \( l_e = 37 \text{ mm} \), \( V_e = 1120 \text{ m/s} \); \( l_n = 35 \text{ mm} \), \( V_h = 1200 \text{ m/s} \). Three-layered barrier: \( l_e = 11.5 \text{ mm} \), \( V_e = 890 \text{ m/s} \); \( l_n = 10.0 \text{ mm} \), \( V_n = 855 \text{ m/s} \).

4 The calculation results

The process of the cylindrical projectile collision with a flat end with a thin steel plate is shown in Fig. 1.

![Fig. 1](image1)

Fig. 1. Impact at an angle of 60 degrees (projectile: a tungsten alloy, radius 0.2 cm, length 4 cm, \( v = 2764 \text{ m/s} \); barrier: steel, radius 1.6 cm, thickness 0.2 cm): (a) 5 \( \mu \text{s} \); (b) 10 \( \mu \text{s} \); (c) 15 \( \mu \text{s} \).

Fig. 1 shows some moments of the projectile interaction \( (t = 5 \text{ ms}, t = 10 \text{ ms}, t = 15 \text{ ms}) \) with the barrier at an angle of 60 degrees from the normal to its surface obtained by numerical simulation of the collision process. Comparison of the numerical results Fig. 1 with the experimental data Fig. 2 showed good qualitative coincidence of the obtained picture of the barrier penetration by
the projectile as well as of the characteristic features of the fragmentation field formation from the barrier and rod materials. A picture of collision correlates well with the experimental data presented in Fig. 2. One observes the material ejection from the face of the barrier and formation of the original fragment ”bubble” on backside. A part of the material from the front end of the projectile was eroded and contributed to the formation of the fragmentation field of ”projectile-barrier” system.

\[\text{Fig. 2. X-ray pattern of steel screen penetration (thickness 6 mm) by a projectile of tungsten alloy 0.4 cm in diameter and elongation of 10 at an angle of 60 degrees at a speed of 2764 m/s.}\]

\[\text{Fig. 3. Impact at an angle of 60 degrees (projectile: tungsten alloy, radius 0.2 cm, length 4 cm, a barrier: steel, radius 1.6 cm, thickness 0.2 cm, } v = 2764 \text{ m/s).}\]

The results of the experiment is shown Fig. 2 qualitatively prove the pattern of projectile interactions with the barrier presented in Fig. 3. Quantitative estimation of the calculated and experimental data on the remaining rod length after barrier penetration in the case under consideration was complicated because of projectile crushing in the experiment. This happened because of the nutation angle, which was not incorporated in the calculation. This happened because of the nutation angle, which was not built into the calculation.

Fig. 4 shows the characteristic features of the rod collision along the normal with a steel plate — a barrier. Here we observe material ejection from the front
surface and the formation of fragments flow behind the backside of the plate. The formation of the fragments flow at axisymmetric formulation of the problem of the impact along the normal is possible only by using a three-dimensional approach and taking into account the probabilistic nature of the crushing of barrier and projectile materials.

Fig. 4. Normal impact (projectile: tungsten alloy, radius 0.2 cm, length 4 cm; barrier: steel, radius 1.6 cm, thickness 0.2 cm, $v = 2732$ m/s): (a) 1 $\mu$s; (b) 5.5 $\mu$s; (c) 11.3 $\mu$s.

A pattern of barriers and projectile fracture is characterized by the presence of fragments of various sizes with a predominance of a very small fraction, which is typical to high-speed collisions [16].

Fig. 5. Rod interaction with a spaced barrier consisting of three plates (projectile: radius 0.5 cm, length 8 cm; barrier: steel-Al-Ti, radius 3 cm, thickness 0.5 cm, distance between plates 1 cm, $v = 5000$ m/s): (a) 4 $\mu$s; (b) 5.6 $\mu$s.

Capability of the proposed technique illustrates the calculations of rod interaction with a spaced barrier, consisting of three plates is shown Fig. 5, and rod interaction with a spaced-layered barrier is shown Fig. 6. Penetration of the spaced barrier was accompanied by the formation of material ejection from the front side of the first plate and fragments flow from the backside. The next
Fig. 6. Rod interaction with spaced-layered barrier (projectile: radius 0.3 cm, length 6 cm, a barrier: steel-Al + Ti, radius 3 cm, thickness 0.3 cm, distance between plates 0.5 cm, \(v = 2600\) m/s, collision angle 60 degrees): (a) 2.5 \(\mu s\); (b) 5.0 \(\mu s\), (c) 10.0 \(\mu s\).

plate was subjected to the impact the preserved part of the projectile and fragments flow from the previous plate and the destroyed part of the projectile is shown Fig. 5. Qualitatively the penetration of the system of the spaced plates is much similar to the penetration of one plate, but quantitative characteristics differ greatly. The first stage of the rod collision with a spaced - layered plate correlates well with the results presented in Fig. 6. Further interaction with the layered barrier is determined by the physical and mechanical properties of the layers, it changes the time required to their penetration.

Earlier in [13] the interaction of spherical elements with thin barriers was considered at different angles of collision. In this paper we consider the interaction of a group of seven rods of tungsten alloy with a system of steel plates. Radius of the rod 1.5 cm, length 15 cm. Thickness of the first plate 7 cm and the second 3 cm, the distance between them 6 cm, diameter 35 cm. Collision speed 1000 m/s. Rods were arranged in a circle with a variable radius \(R\). One rod was placed in the center and the other six rods were uniformly placed along the circumference. When calculating we varied the distance between the center of the first projectile and the centers of the rest \(R\). The calculation results presented in Fig. 7–11 enabled to determine the best configuration of rods system to penetrate the first barrier and destruct the second one. Fig. 10 shows the location of rods with flat heads on the front side of the barrier system at the initial time.

Fig. 12 shows calculations of the collision of a single projectile with a mass equal to that of seven projectiles, which were discussed above. It can be seen that the size of the mass which was knocked out (a) lighter area in Fig. 12, (b) was much less as compared with that which was knocked out by seven projectiles Fig. 8–11.

In the paper we compared the effectiveness of the impacts of the projectiles group and assessed their effects on the degree of the barrier damage. The increase in radius \(R\) from 5 cm to 9 cm causes the increase in the volume of material knocked out from first barrier at a noticeable speed drop. At \(R = 5\) cm the second barrier is also penetrated and significant flow of the fragments is formed from back and front sides of the barrier, but when \(R = 8\) or 9 cm this is not observed. When \(R = 10\) cm the projectiles only partially penetrate into the first
Fig. 7. Initial configuration of the system ”barriers-projectiles”: (a) three-dimensional picture; (b) 2-D cross-section of a three-dimensional computational domain.

Fig. 8. Interaction of projectiles with a barrier at $R = 5$ cm: (a) three-dimensional picture; (b) 2-D cross-section of a three-dimensional computational domain. $t = 778.76$ microseconds

Fig. 9. Interaction of projectiles with a barrier at $R = 8$ cm: (a) three-dimensional picture; (b) 2-D cross-section of a three-dimensional computational domain. $t = 657.26$ microseconds
Fig. 10. Interaction of projectiles with a barrier at $R = 9$ cm: (a) three-dimensional picture; (b) 2-D cross-section of a three-dimensional computational domain. $t = 715.28$ microseconds

Fig. 11. Interaction of projectiles with a barrier at $R = 10$ cm: (a) three-dimensional picture; (b) 2-D cross-section of a three-dimensional computational domain. $t = 306.08$ microseconds

Fig. 12. The interaction of a single projectile with barriers: (a) initial configuration of the system "barriers-projectile"; (b) 2-D cross-section of a three-dimensional computational domain. (a) 0 microseconds, (b) 80 microseconds
Fig. 13. Interaction of projectiles with a barrier at an angle of 450 from the normal: (a) three-dimensional picture; (b) 2-D cross-section of a three-dimensional computational domain. $t = 100$ microseconds

barrier and no full effect of the collective impact of the group of elements on the barrier is observed. The first barrier bulges toward the second barrier, but it is not punctured and completely destroyed, however one observes the formation of cracks in the circumferential direction. As can be seen from the calculations, there is a certain configuration of the projectiles group, the most dangerous in terms of breaking the barrier and the mass of the ejected material.

The calculation results presented in Fig. 7–12 showed a great danger of the impact of the group of rods on the protected spacecraft as compared with the impact of a single projectile with a mass equal to that of seven projectiles at the same speed. The developed numerical technique enables to simulate the interactions of spacecraft shells with high-speed long rods in a wide range of speeds and collision angles and also to investigate the processes of rods and barriers fragmentations and nature of the forming fragmentation fields.

5 Conclusion

A probabilistic approach and proposed numerical technique enables in full, from the physical point of view, in three-dimensional formulation with sufficient accuracy to reproduce the processes of penetration of multi-layer and spaced barriers with high-speed core elements.

In calculations we fully simulated the fragmentation fields and took into account the interaction of fragments with each other and with the elements of the multi-layer barrier that was extremely important in calculating the protection of spacecrafts, as high-speed particles flows could penetrate the main body of the machine and damage the equipment. Therefore, it is necessary to evaluate the kinetic energy of the fragments and calculate the process of their collision with the main body. The proposed approach enables to calculate the entire process of the projectile interaction with barriers taking into account the formation of fragmentation flows and the collision of the latter with the protected object.

Acknowledgments. The study of A. Gerasimov was carried out by the grant from Russian Science Foundation 16-19-10264.
References

hp-Version of Collocation and Least Residuals Method in Mechanics of Laminated Composite Plates

Sergey Golushko\textsuperscript{1,2}, Semyon Idimeshev\textsuperscript{2}, and Vasiliy Shapeev\textsuperscript{3}

\textsuperscript{1} Institute of Computational Technologies
Siberian Branch of the Russian Academy of Sciences, Academian M.A. Lavrentjev ave. 6, Novosibirsk, Russia
\textsuperscript{2} Design Technological Institute of Digital Techniques
Siberian Branch of the Russian Academy of Sciences, Academician Rzhanova ave. 6, Novosibirsk, Russia
\textsuperscript{3} Institute of Theoretical and Applied Mechanics
Siberian Branch of the Russian Academy of Sciences, Institutskaya Str. 4/1, Novosibirsk, Russia
{s.k.golushko, idimeshev, vshapeev}@gmail.com
http://www.ict.nsc.ru/en

Abstract. A version of collocations and least residuals method (CLS) based on polynomial approximation of high degree ($p$-approach) was proposed and implemented. In rectangular domains collocation points are selected using the roots of Chebyshev polynomials and approximate solution is represented in the form of direct products of Chebyshev polynomials series. It was shown that the use of $p$-approach in the CLS method allows to obtain numerical solutions with high accuracy and to implement complex boundary conditions with no special techniques. The numerical method used to solve a problem of bending of laminated anisotropic rectangular plates within frameworks of classical laminated plate theory, first order shear deformation theory and Grigolyuk-Chulkov’s broken line theory. Several specific example problems are solved, including fixed three-ply laminates with transversely isotropic layers under transverse uniform loading.

Keywords: Collocations and least residuals method, Chebyshev polynomials, plate theory, spectral methods, composite materials

1 Introduction

The collocation and least residuals method (CLS) is an efficient method for numerical solution of boundary value problems both for systems of ordinary and partial differential equations. It is based on the collocation method (CM) [1], with approximate solution is represented as a linear combination of basis functions in some functional space. To determine it unknown coefficients in CM residual of equations $R(x)$ vanishes at given points (collocation points)

$$R(x_{i}^{col}) = 0, \quad \{x^{col}\} - \text{collocation points}. \quad (1)$$
The main difference between CLS method and CM is the minimizing technique of \( R(x) \). In CLS method we minimize some functional of residual in the collocation points \([2, 3]\), instead of the condition (1). The CLS method is used to minimize residual in \( L_2 \) norm

\[
\sum_i \| R(x_{col}^i) \|_2^2 \rightarrow \min .
\]  

(2)

In CLS method the number of equations can exceed the number of unknown coefficients in representation of the solution. The solutions of arising over-determined systems of linear algebraic equations (SLAE) are defined in the sense of (2) (least squares). In comparison with CM (1) the obtained over-determined SLAE is often better conditioned and leads to less non-physical oscillations in numerical solutions. Similar regularization approaches are applied in the finite element method (Least squares finite element methods) \([4]\).

In this paper, an approximate solution is represented as a linear combination of polynomials of high degrees (\( p \)-approach) that is typical for spectral methods. This allows to obtain numerical solutions of high accuracy at low computational cost. Term (2) makes the implementation of \( p \)-approach more convenient in CLS method when compared with CM. This modification based on \( p \)-approach is called \( hp \)-version of CLS method.

We will demonstrate the application of a method to solving problems of solid mechanics – bending of laminated anisotropic rectangular plates. On a practical level, for calculating the stress and displacement fields of such structures the theories of plates are used. They lead to a smaller computational efforts compared to three-dimensional elasticity formulation.

Boundary value problems arising in a plate theories have a number of features that present difficulties for many well-known numerical methods. First, governing equations of plate theories may contain derivatives of high orders. Second, boundary conditions may be quite complicated, for example, in a form of linear combination of functions and their higher order derivatives. Third, equations of the plate theory may contain small parameters in the derivatives. These features cause serious difficulties for widely used finite differences and finite element methods. The use of both \( p \)-approach and term (2) in \( hp \)-version of CLS method may resolve these difficulties and obtain high accuracy solutions at relatively low computational efforts.

2 Formulation of Problem and Governing Equations

Let us consider a static bending of laminate composed of 3 layers of constant thickness (Fig. 1). Layers are transversely isotropic with material symmetry axis in the plate’s plane. Layers orientation scheme is

\[
\theta^1 = \theta^3 = 0, \quad \theta^2 = \pi/2,
\]

where \( \theta^k \) is an angle measured counterclockwise from the \( x \) coordinate axis to the \( k \)-th layer material symmetry axis.
Fig. 1. Rectangular multilayered plate under transverse loading; \(a, b, h\) — plate’s dimensions in the directions \(x, y, z\) respectively, \(z_k\) — \(k\)-th layer lower surface coordinate, \(k = 1, 2, 3\).

Engineering constants of transversely isotropic material are [5]

\[
E_L = 25 \text{ Mpsi}, \quad E_T = 1 \text{ Mpsi}, \quad G_{LT} = 0.5 \text{ Mpsi},
\]

\[
G_{TT} = 0.2 \text{ Mpsi}, \quad \nu_{LT} = \nu_{TT} = 0.25.
\] (3)

Here \(E, G, \nu\) are elasticity and shear modulus, Poisson ratios. \(L\) signifies the material symmetry axe, \(T\) the transverse direction. Layers thicknesses are \(h^1 = h^3 = h/4, \quad h^2 = h/2\).

The upper surface of the plate is under uniform transverse load \(q_0\), the lower surface is free, and a continuity condition of displacements \(u, v, w\) and stresses \(\sigma_{zz}, \sigma_{xz}, \sigma_{yz}\) is used on interface surfaces. The corresponding boundary conditions are defined on the boundary of the plate. The task is to calculate the stress and displacement fields of such plates.

Calculation of thin laminated anisotropic structures within framework of the three-dimensional elasticity is associated with high computational efforts. Therefore, many researchers frequently make use of more robust plate theories, that allows to reduce the dimension of the original problem by excluding the direction of the coordinate \(z\) from consideration. We consider three plate theories, where transverse shear stresses are simulated differently: classical laminated plate theory (CLPT), first order shear deformation theory (FSDT) or Timoshenko’s plate theory [6] and Grigolyuk-Chulkov’s broken line theory (GCT) [7].

**Classical laminated plate theory** uses the classical Kirchhoff assumption which implies the geometric relationships in the form of

\[
e_{xx} = \frac{\partial u_0}{\partial x} - z \frac{\partial^2 w_0}{\partial x^2}, \quad e_{yy} = \frac{\partial v_0}{\partial y} - z \frac{\partial^2 w_0}{\partial y^2}, \quad e_{zz} = 0,
\]

\[
e_{xy} = \frac{1}{2} \left( \frac{\partial u_0}{\partial y} + \frac{\partial v_0}{\partial x} \right) - z \frac{\partial^2 w_0}{\partial x \partial y}, \quad e_{xz} = 0, \quad e_{yz} = 0.
\]
Here \(e_{ij}\) – strains; \(u_0(x, y), v_0(x, y), w_0(x, y)\) – central plane displacements.

Constitutive equations for \(k\)-th layer are expressed by

\[
\begin{pmatrix}
\sigma_{xx}^k \\
\sigma_{yy}^k \\
\sigma_{xy}^k
\end{pmatrix} =
\begin{pmatrix}
Q_{11}^k & Q_{12}^k & Q_{16}^k \\
Q_{12}^k & Q_{22}^k & Q_{26}^k \\
Q_{16}^k & Q_{26}^k & Q_{66}^k
\end{pmatrix}
\begin{pmatrix}
e_{xx} \\
e_{yy} \\
e_{xy}
\end{pmatrix},
\]

\[ (4) \]

where

\[
C_1^k = \begin{pmatrix}
C_{11}^k & C_{12}^k & 0 \\
C_{12}^k & C_{22}^k & 0 \\
0 & 0 & C_{66}^k
\end{pmatrix}, \quad D_1^k = \begin{pmatrix}
\cos^2 \theta_k & \sin^2 \theta_k & -\sin 2\theta_k \\
\sin^2 \theta_k & \cos^2 \theta_k & \sin 2\theta_k \\
(sin 2\theta_k)/2 - (sin 2\theta_k)/2 & \cos 2\theta_k
\end{pmatrix}
\]

Coefficients \(C_{ij}^k\) express in terms of the engineering constant as follows:

\[
C_{11}^k = \frac{E_L^k}{1 - \nu_{LT}^k \nu_{TL}^k}, \quad C_{22}^k = \frac{E_T^k}{1 - \nu_{LT}^k \nu_{TL}^k},
\]

\[
C_{12}^k = \frac{\nu_{LT}^k E_T^k}{1 - \nu_{LT}^k \nu_{TL}^k}, \quad C_{66}^k = G_{LT}.
\]

It is convenient to define the following quantities

\[
A_{ij} = \sum_{k=1}^{3} \int_{z_k}^{z_{k+1}} Q_{ij}^k \, dz = \sum_{k=1}^{3} Q_{ij}^k (z_{k+1} - z_k),
\]

\[
D_{ij} = \sum_{k=1}^{3} \int_{z_k}^{z_{k+1}} Q_{ij}^k z^2 \, dz = \frac{1}{3} \sum_{k=1}^{3} Q_{ij}^k (z_{k+1}^3 - z_k^3),
\]

where \(z_k\) – coordinates of layers lower surface (Fig. 1). Finaly governing equations for considered problem within CLPT framework are given by

\[
(A_{12} + A_{66}) \frac{\partial^2 v_0}{\partial x \partial y} + A_{11} \frac{\partial^2 u_0}{\partial x^2} + A_{66} \frac{\partial^2 u_0}{\partial y^2} = 0,
\]

\[
(A_{12} + A_{66}) \frac{\partial^2 u_0}{\partial x \partial y} + A_{22} \frac{\partial^2 v_0}{\partial y^2} + A_{66} \frac{\partial^2 v_0}{\partial x^2} = 0,
\]

\[
(2D_{12} + 4D_{66}) \frac{\partial^4 w_0}{\partial x^2 \partial y^2} + D_{11} \frac{\partial^4 w_0}{\partial x^4} + D_{22} \frac{\partial^4 w_0}{\partial y^4} = q_0.
\]

We consider two kinds of boundary conditions:
- clamped

\[
\begin{align*}
  x &= 0 : \\
  &\quad u_0 = 0, \quad v_0 = 0, \quad w_0 = 0, \quad \frac{\partial w_0}{\partial x} = 0; \\
  x &= a : \\
  &\quad u_0 = 0, \quad v_0 = 0, \quad w_0 = 0, \quad \frac{\partial w_0}{\partial x} = 0; \\
  y &= 0 : \\
  &\quad u_0 = 0, \quad v_0 = 0, \quad w_0 = 0, \quad \frac{\partial w_0}{\partial y} = 0; \\
  y &= a : \\
  &\quad u_0 = 0, \quad v_0 = 0, \quad w_0 = 0, \quad \frac{\partial w_0}{\partial y} = 0;
\end{align*}
\]

- simply-supported

\[
\begin{align*}
  x &= 0 : \\
  &\quad \frac{\partial u_0}{\partial x} = 0, \quad v_0 = 0, \quad w_0 = 0, \quad \frac{\partial^2 w_0}{\partial x^2} = 0; \\
  x &= a : \\
  &\quad \frac{\partial u_0}{\partial x} = 0, \quad v_0 = 0, \quad w_0 = 0, \quad \frac{\partial^2 w_0}{\partial x^2} = 0; \\
  y &= 0 : \\
  &\quad u_0 = 0, \quad \frac{\partial v_0}{\partial y} = 0, \quad w_0 = 0, \quad \frac{\partial^2 w_0}{\partial y^2} = 0; \\
  y &= a : \\
  &\quad u_0 = 0, \quad \frac{\partial v_0}{\partial y} = 0, \quad w_0 = 0, \quad \frac{\partial^2 w_0}{\partial y^2} = 0.
\end{align*}
\]

**First order shear deformation theory** allows transverse shear in a first approximation by defining independent function of rotation of the transverse normal about central surface: \( \phi_x(x, y) \) and \( \phi_y(x, y) \). The strains are obtained by

\[
\begin{align*}
  e_{xx} &= \frac{\partial u_0}{\partial x} + z \frac{\partial \phi_x}{\partial x}, \\
  e_{yy} &= \frac{\partial v_0}{\partial y} + z \frac{\partial \phi_y}{\partial y}, \\
  e_{xy} &= \left( \frac{\partial u_0}{\partial y} + \frac{\partial v_0}{\partial x} \right) + z \left( \frac{\partial \phi_x}{\partial y} + \frac{\partial \phi_y}{\partial x} \right), \\
  e_{xz} &= \frac{\partial w_0}{\partial x} + \phi_x, \\
  e_{yz} &= \frac{\partial w_0}{\partial y} + \phi_y, \\
  e_{zz} &= 0.
\end{align*}
\]

Constitutive equations of FSDT are obtained by adding to (4) expressions for the shear stresses

\[
\begin{pmatrix}
\sigma_{yz}^k \\
\sigma_{xz}^k 
\end{pmatrix} = \begin{pmatrix}
Q_{44}^k & Q_{45}^k \\
Q_{45}^k & Q_{55}^k 
\end{pmatrix} \begin{pmatrix}
e_{yz} \\
e_{xz}
\end{pmatrix},
\]

\[
\begin{pmatrix}
Q_{44}^k & Q_{45}^k \\
Q_{45}^k & Q_{55}^k 
\end{pmatrix} = D_2^k C_2^k (D_2^k)^T, \quad C_2^k = \begin{pmatrix}
C_{44}^k & 0 \\
0 & C_{55}^k 
\end{pmatrix}, \quad D_2^k = \begin{pmatrix}
\cos \theta^k & \sin \theta^k \\
-\sin \theta^k & \cos \theta^k 
\end{pmatrix},
\]

where stiffness coefficients are expressed by engineering constants

\[
C_{44}^k = G_{TT}, \quad C_{55}^k = G_{LT}.
\]
Governing equations for FSDT can be written as

\begin{align*}
(A_{12} + A_{66}) \frac{\partial^2 v_0}{\partial x \partial y} + A_{11} \frac{\partial^2 u_0}{\partial x^2} + A_{66} \frac{\partial^2 u_0}{\partial y^2} &= 0, \\
(A_{12} + A_{66}) \frac{\partial^2 u_0}{\partial x \partial y} + A_{22} \frac{\partial^2 v_0}{\partial y^2} + A_{66} \frac{\partial^2 v_0}{\partial x^2} &= 0, \\
-A_{44} \frac{\partial \phi_y}{\partial y} - A_{44} \frac{\partial^2 w_0}{\partial y^2} - A_{55} \frac{\partial \phi_x}{\partial x} - A_{55} \frac{\partial^2 w_0}{\partial x^2} &= q_0,
\end{align*}

where coefficients are defined by (5).

Boundary conditions for clamped edges in FSDT are written as follows:

\begin{align*}
u_0 &= 0, \quad v_0 = 0, \quad w_0 = 0, \quad \phi_x = 0, \quad \phi_y = 0, \quad (x, y) \in \partial \Omega.
\end{align*}

More details of CLPT and FSDT plate theories are described in [6].

Grigolyuk-Chulkov’s theory is layerwise theory, where mechanical properties of each layer are considered separately. For this purpose, in each layer the rotations of transverse normal about central surface \( \phi^k_x(x, y) \) and \( \phi^k_y(x, y) \) are defined. It can be assumed that the GCT is a generalization of the FSDT that takes into account transverse shear stresses in each layer separately.

Expressions for geometrical equations in GCT have the form

\begin{align*}
\epsilon^k_{xx} &= \frac{\partial u_0}{\partial x} + \sum_{i=1}^{3} P_{ki} \frac{\partial \phi^i_x}{\partial x} + (z - z_{k-1}) \frac{\partial \phi^k_x}{\partial x}, \\
\epsilon^k_{yy} &= \frac{\partial v_0}{\partial y} + \sum_{i=1}^{3} P_{ki} \frac{\partial \phi^i_y}{\partial y} + (z - z_{k-1}) \frac{\partial \phi^k_y}{\partial y}, \\
\epsilon^k_{xy} &= \left( \frac{\partial v_0}{\partial x} + \frac{\partial u_0}{\partial y} \right) + \sum_{i=1}^{3} P_{ki} \left( \frac{\partial \phi^i_y}{\partial x} + \frac{\partial \phi^i_x}{\partial y} \right) + (z - z_{k-1}) \left( \frac{\partial \phi^k_y}{\partial x} + \frac{\partial \phi^k_x}{\partial y} \right), \\
\epsilon^k_{xz} &= \phi^k_x + \frac{\partial w}{\partial x}, \quad \epsilon^k_{yz} = \phi^k_y + \frac{\partial w}{\partial y}, \quad \epsilon^k_{zz} = 0,
\end{align*}

where

\[ P = \begin{pmatrix} 0 & 0 & 0 \\ h^1 & 0 & 0 \\ h^1 & h^2 & 0 \end{pmatrix}. \]

Boundary conditions for clamped edges in GCT are \((k = 1, 2, 3)\)

\begin{align*}
u_0 &= 0, \quad v_0 = 0, \quad w_0 = 0, \quad \phi^k_x = 0, \quad \phi^k_y = 0, \quad (x, y) \in \partial \Omega.
\end{align*}
Because of awkward form, governing equations of GCT are not presented. Detailed description of the theory can be found in [7]. It is necessary to note, that plate theories are approximations to elasticity theory and bring about their own errors which are important to estimate.

3 \textit{hp} - Version of CLS Method

Implementation aspects of CLS method are similar to the CM. Consider a general boundary value problem for a linear elliptic system in a rectangular domain \( \Omega = [0, a] \times [0, b] \):

\[
Lu(x, y) = f(x, y), \quad (x, y) \in \Omega,
\]

\[
L_{bnd} u(x, y) = g(x, y), \quad (x, y) \in \partial \Omega.
\]

Let us define a grid with non-overlapping rectangular cells \( \Omega^k \ (k = 1, \ldots K) \). In each cell \( \Omega^k \) we introduce local variables \((\alpha_1^k, \alpha_2^k)\), which are associated with global variables \((x, y)\) in the Cartesian coordinate system by

\[
\alpha_1^k = \frac{x - x^*_k}{d_1^k}, \quad \alpha_2^k = \frac{y - y^*_k}{d_2^k},
\]

where \(2d_1^k, 2d_2^k\) — sizes of cell in \(x\) and \(y\) directions, \((x^*_k, y^*_k)\) — the coordinates of the cell centers. Local variables are varying in canonical interval \(\alpha_1^k \in [-1, 1]\), \(\alpha_2^k \in [-1, 1]\). The upper index \(k\), that indicates the cell’s number, will be omitted further.

In this version of CLS method approximate solution in the cell is represented in form of direct product of single variable basis functions:

\[
\sum_{i_1=0}^{N_1-1} \sum_{i_2=0}^{N_2-1} c_{i_1i_2} \phi_{i_1}(\alpha_1) \phi_{i_2}(\alpha_2).
\]

Functions \(\phi_i\) are chosen as the Chebyshev polynomials of the first kind \(T_n\)

\[
\phi_{i_1}(\alpha_1) = T_{i_1}(\alpha_1), \quad \phi_{i_2}(\alpha_2) = T_{i_2}(\alpha_2).
\]

In previous paper [8] we used cardinal functions in Lagrange-like form

\[
\phi_{i_1}(\alpha_1) = \prod_{m=0, m \neq i_1}^{N_1-1} (\alpha_1 - \alpha_1^{col}_m) / (\alpha_1^{col}_i - \alpha_1^{col}_m), \quad \phi_{i_2}(\alpha_2) = \prod_{l=0, l \neq i_2}^{N_2-1} (\alpha_2 - \alpha_2^{col}_l) / (\alpha_2^{col}_{i_2} - \alpha_2^{col}_l),
\]

where \((\alpha_1^{col}_m, \alpha_2^{col}_m)\) are local coordinates of collocation points. But in practice polynomials (8) require a large number of arithmetic operations and lead to complex expressions when differentiating. In this sense, Chebyshev polynomials are more convenient choice.

To determine the unknown coefficients in representation (7) for each cell let us write down the equation of three types
collocation equations at the collocation point \((\alpha_1^{col}, \alpha_2^{col})\)

\[ Lu(\alpha_1^{col}, \alpha_2^{col}) = f(\alpha_1^{col}, \alpha_2^{col}); \quad (9) \]

boundary equations at given point \((\alpha_1^{bnd}, \alpha_2^{bnd})\), on boundary \(\partial \Omega^k\), adjacent to \(\partial \Omega^k\)

\[ L_{bnd}u(\alpha_1^{bnd}, \alpha_2^{bnd}) = g(\alpha_1^{bnd}, \alpha_2^{bnd}); \quad (10) \]

matching conditions on interface between neighbour cells at given points \((\alpha_1^{mat}, \alpha_2^{mat})\)

\[ L_{mat}u(\alpha_1^{mat}, \alpha_2^{mat}) = L_{adj}u^{adj}(\alpha_1^{mat}, \alpha_2^{mat}), \quad (11) \]

Here \(u^{adj}\) – solution defined in neighbour cell \(\Omega^{adj}\).

Matching conditions \(L_{mat}\) usually require the continuity of the solutions and the necessary number of its derivatives along the normal to the boundary of the cell.

In this version of CLS method the local coordinates of collocation points are roots of Chebyshev polynomials \((i_1 = 1, \ldots, N_1, i_2 = 1, \ldots, N_2)\)

\[ ((\alpha_1)^{col}_{i_1}, (\alpha_2)^{col}_{i_2}) = (t^{i_1}_1, t^{i_2}_2), \]

where \(t^{i_1}_1\) and \(t^{i_2}_2\) – roots of Chebyshev polynomials of \(N_1\) and \(N_2\) degree respectively. By the same way we define \((\alpha_1^{bnd}, \alpha_2^{bnd})\) and \((\alpha_1^{mat}, \alpha_2^{mat})\) on cell boundaries \((i_1 = 1, \ldots, N_1, i_2 = 1, \ldots, N_2)\)

\((-1, t^{i_2}_2), \quad (1, t^{i_2}_2), \quad (t^{i_1}_1, -1), \quad (t^{i_1}_1, 1)\).

Thus, for \(N_1N_2\) unknown coefficients in cell we use \(N_1N_2\) collocation equations appended by equations on cell boundary. Thus, in \(hp\)-version of CLS method corresponding SLAE becomes overdetermined. In this version of method approximate solution form does not satisfy to boundary and matching conditions identically, so residual \(R(x)(2)\) must contain not only collocation equations, but boundary conditions (10) and matching conditions (11) too. In particular, this allows us to consider the boundary conditions in complex form.

To solve the overdetermined linear systems in the least squares sense (2) we use \(QR\) factorization of its matrix, implemented by Householder method. In the case of large linear systems we use domain decomposition method [9]. This allows to reduce the solution of the problem in whole region to iterative process through subdomains with computational complexity is much smaller than the original problem for the region. For linear systems in a subdomain we use Householder method again. In this case special matching conditions between subdomains are used. For example, the continuity of function and its first derivative at the boundary of the cell \(\Omega^k\) can be written as

\[ u + p_1 \frac{\partial u}{\partial n} = u^{adj} + p_1 \frac{\partial u^{adj}}{\partial n}, \]

where \(u\) is solution in the cell at the current iteration; \(u^{adj}\) — solution in the neighbor cell; \(n\) — the outer normal to the boundary \(\Omega^k\). For plate theories,
that may contain derivatives up to the 4th order, we can additionally require
the continuity of a linear combination of the second and third derivatives of the
approximate solution:
\[
\frac{\partial^2 u}{\partial n^2} + p_2 \frac{\partial^3 u}{\partial n^3} = \frac{\partial^2 u^{adj}}{\partial n^2} + p_2 \frac{\partial^3 u^{adj}}{\partial n^3}.
\]
The choice of weights \(p_1, p_2\) can affect the properties of the numerical solutions
and speed of convergence of the iterative process.

4 Numerical Experiments

In all numerical experiments a single cell that coincides with the entire domain
is used. Further we will consider only square plates \(a = b = 1\) m.

To demonstrate the capabilities of \(hp\)-version of CLS method let us consider
the problem with the known exact solution. The last equation of the system (6) is
similar to the Kirchhoff-Love plate theory equation for bending of a homogeneous
orthotropic plates
\[
D_{11} \frac{\partial^4 w_0}{\partial x^4} + (2D_{12} + 4D_{66}) \frac{\partial^4 w_0}{\partial x^2 \partial y^2} + D_{22} \frac{\partial^4 w_0}{\partial y^4} = q_0.
\]
Consider 3-ply simply-supported square laminate under uniform transverse
load \(q_0\) with stiffness coefficients defined by (3). This problem can be solved by
Fourier method [10]. In this case maximum deflection is observed in the center of
the plate and if \(h = 0.01\) m, then the deflection value for Fourier method solution
(sum of first 2500 members) is
\[
w^* = \frac{w_0(0.5, 0.5)}{q_0} \cdot 10^8 = 9.8577127.
\]
Deflections in the center of the plate are calculated by \(hp\)-version of CLS method
are shown in Table 1.

<table>
<thead>
<tr>
<th>(N_1 \times N_2)</th>
<th>(w_0(0.5, 0.5)/q_0 \cdot 10^8)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 (\times) 10</td>
<td>(9.8579)</td>
</tr>
<tr>
<td>20 (\times) 20</td>
<td>(9.85773)</td>
</tr>
<tr>
<td>30 (\times) 30</td>
<td>(9.857715)</td>
</tr>
</tbody>
</table>

Table 1 demonstrates high accuracy of numerical results, obtained by \(hp\)-version of CLS method even for the differential equation with derivatives of
4-th order of the unknown functions. Thus, \( hp \) – version of CLS method has no difficultnes when working with differential equations containing derivatives of high orders like (3).

Consider another formulation of the problem. Let us use the free edge conditions on the one of the edges \( (x = 1) \) in the previous problem.

\[
D_{11} \frac{\partial^2 w_0}{\partial x^2} + D_{12} \frac{\partial^2 w_0}{\partial y^2} = 0,
\]

\[
D_{11} \frac{\partial^3 w_0}{\partial x^3} + (D_{12} + 2D_{66}) \frac{\partial^3 w_0}{\partial x^2 \partial y} = 0.
\]

These boundary conditions by the use of term (2) implements with no additional effort in the CLS method. Fig. 2 shows the deformed shape of the plate with the free edge.

**Fig. 2.** Simply-supported plate with free edge.

Now let us consider the bending of clamped 3-ply laminates with different relative thickness \( h/a \), as described in Section 2. Stress and displacement fields calculation for such plates will be carried out within framework of three theories described above.

**Fig. 3.** Stress fields for absolute value of \( \sigma_{xx} \) and \( \sigma_{yy} \) in 3-ply laminate for \( h/a=0.02 \).
Brief analysis of solution shows that $\sigma_{xx}$ prevails in the stress state of outer layers (Fig. 3). And maximum absolute values are observed in the vicinity of the clamped edges ($x = 0, 1$). Similar conclusion are true for the middle layer of the plate and component $\sigma_{yy}$, which is associated with an orientation of transversely isotropic material.

Further we will use the following normalized quantities

$$
\bar{\sigma}^k_{xx} = \frac{\sigma^k_{xx}}{q_0 S^2}, \quad \bar{\sigma}^k_{yy} = \frac{\sigma^k_{yy}}{q_0 S^2}, \quad \bar{\sigma}^k_{xy} = \frac{\sigma^k_{xx}}{q_0 S^2} 10^{-2},
$$

$$
\bar{\omega}_0 = \frac{\sigma^k_{xx}}{q_0} 10^{-9}, \quad \bar{z} = \frac{z}{h}.
$$

Table 2. Stresses and deflection in 3-ply laminate. The results of calculations carried out in the framework of the CLPT, FSDT and GCT plate theories. Sign (%) is used for relative percentage deviation from GCT.

<table>
<thead>
<tr>
<th>$h/a$</th>
<th>GCT</th>
<th>FSDT</th>
<th>CLPT</th>
<th>FSDT (%)</th>
<th>CLPT (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{\sigma}^3_{xx}(a, 0, h/2)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.975</td>
<td>0.475</td>
<td>0.579</td>
<td>51.2</td>
<td>40.6</td>
</tr>
<tr>
<td>0.05</td>
<td>0.569</td>
<td>0.545</td>
<td>0.579</td>
<td>4.28</td>
<td>1.61</td>
</tr>
<tr>
<td>0.02</td>
<td>0.576</td>
<td>0.574</td>
<td>0.578</td>
<td>0.30</td>
<td>0.42</td>
</tr>
<tr>
<td>0.01</td>
<td>0.578</td>
<td>0.578</td>
<td>0.578</td>
<td>0.05</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>$\bar{\sigma}^2_{yy}(0, a, h/4)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.636</td>
<td>0.662</td>
<td>0.456</td>
<td>4.21</td>
<td>28.2</td>
</tr>
<tr>
<td>0.05</td>
<td>0.605</td>
<td>0.536</td>
<td>0.456</td>
<td>11.5</td>
<td>24.6</td>
</tr>
<tr>
<td>0.02</td>
<td>0.483</td>
<td>0.471</td>
<td>0.456</td>
<td>2.65</td>
<td>5.72</td>
</tr>
<tr>
<td>0.01</td>
<td>0.463</td>
<td>0.460</td>
<td>0.455</td>
<td>0.64</td>
<td>1.82</td>
</tr>
<tr>
<td></td>
<td>$\bar{\sigma}^3_{xy}(3/4a, 3/4a, h/2)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.983</td>
<td>0.675</td>
<td>0.006</td>
<td>31.3</td>
<td>38.4</td>
</tr>
<tr>
<td>0.05</td>
<td>0.632</td>
<td>0.647</td>
<td>0.006</td>
<td>2.28</td>
<td>4.20</td>
</tr>
<tr>
<td>0.02</td>
<td>0.616</td>
<td>0.616</td>
<td>0.006</td>
<td>0.04</td>
<td>1.74</td>
</tr>
<tr>
<td>0.01</td>
<td>0.609</td>
<td>0.609</td>
<td>0.006</td>
<td>0.02</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>$\bar{\omega}(a/2, a/2, h/2)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>−0.751</td>
<td>−0.603</td>
<td>−0.208</td>
<td>31.30</td>
<td>38.36</td>
</tr>
<tr>
<td>0.05</td>
<td>−2.650</td>
<td>−2.538</td>
<td>−1.067</td>
<td>2.28</td>
<td>4.20</td>
</tr>
<tr>
<td>0.02</td>
<td>−28.523</td>
<td>−28.335</td>
<td>−26.044</td>
<td>0.04</td>
<td>1.74</td>
</tr>
<tr>
<td>0.01</td>
<td>−213.312</td>
<td>−212.974</td>
<td>−208.354</td>
<td>0.02</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 2 shows calculated deflections and stresses in vicinities of their maximum absolute values. We assume GCT as the most accurate among considered theory, because its hypothesis is the most suitable in given structure [8, 7]. Therefore, we will treat it as a reference.
Despite the simplicity, the CLPT may be used for very thin laminated plates. If the 5% deviations from GCL are admissible, CLPT can be applied for $h/a < 0.02$ case. More accurate results are obtained within the framework of FSDT. It can be used for plates with aspect ratio $h/a < 0.05$. For thicker plates GCT theory differs significantly.

In presented calculations all functions are approximated with $N_1 = N_2 = 17$ that allowed to get three guaranteed digits for the maximum deflection. In this case, the CLPT is required to determine $17 \cdot 17 \cdot 3 = 867$ unknown coefficients, FSDT $- 17 \cdot 17 \cdot 5 = 1445$, and GCT $- 2601$. It is interesting that for $h/a < 0.02$ CLPT results do not differ significantly from those of FSDT. In particular, it means that for plates with $h/a < 0.02$ CLPT is preferable, because of lower computational efforts.

![Stress distribution along z coordinate](image)

**Fig. 4.** Stresses distribution along $z$ coordinate $\sigma_{xx}(a, 0, \bar{z})$, $\sigma_{xx}(a/2, a/2, \bar{z})$ and $\sigma_{yy}(0, b, \bar{z})$, $\sigma_{yy}(a/2, a/2, \bar{z})$ in 3-ply laminate for $h/a = 0.02$.

Fig. 4 shows the distribution of stresses $\sigma_{xx}$ and $\sigma_{yy}$ along the $z$ coordinate (thickness) for different points $(x, y)$: in the center of the plate and on the border where they reach the maximum values. The absolute maximum stresses values are observed at the outer surfaces of layers. And the absolute values at the edges of the plate exceed values at the center of the plate a few times, that is true for both the stress tensor components.

Presented numerical results shows that $hp$-version of CLS method can be successfully applied to problems of mechanics of laminates anisotropic rectangular plates within framework the various plates theories. Term (2) allows us to
consider a wide class of boundary value problems including complex boundary conditions. Moreover $p$-approach allows to obtain high accuracy of numerical solutions at low computational efforts.

References

Computational Modelling of Metal Vapor Influence on the Electric Arc Welding

Amanbek Jainakov, Rena Sultangazieva, and Bubusara Medralieva

Kyrgyz State Technical University after I.Razzakov,
Mira ave. 66, 720044 Bishkek, Kyrgyzstan
{jainakov-41,renasultangazieva,medralieva}@mail.ru
http://www.kstu.kg

Abstract. The mathematical model, which takes into account the effects of the Fe vapours in the arc welding of stainless steel workpieces in closed volume are proposed. The physical phenomena in arc plasma and molten pool are considered in coupled unified MHD model. The system is solved in the variables vorticity-stream function for five variables: stream function, vorticity, current function, enthalpy and metal vapour concentration. The system of equations solved by the finite difference method on a rectangular non-uniform orthogonal grid using the five-point difference scheme. The effect of metal vapour from weld pool on the characteristics of the arc column was numerically investigated. Distribution of electric field, current density and temperature field in the arc column and weld pool with and without consideration of the Fe vapor are shown.

Keywords: electric arc plasma, MHD equations, metal vapours, vorticity, stream function, weld pool, Marangoni effect

1 Introduction

Arc welding is characterized by high values of molten metal temperature gradients, with a significant portion of the surface of the weld pool metal is at a temperature close to the boiling temperature and generates a modest amount of metal vapor arc zone, which has a significant impact on the basic physical properties of the arc, energy efficiency, impact the size and shape of the weld pool. The atoms of metal have a lower ionisation energy compared with inert gases such as argon and helium. For example, the ionization energy of argon is 15,755 eV and the ionization energy of iron is 7.8 eV. This increases the radiation and electric conductivity of the plasma and causes a change in composition and properties of the plasma arc in the anode region and a portion of the arc column. In turn, evaporation of workpieces impurities changes the composition of the molten pool, which can cause changes in the microstructure of the metal and mechanical properties of the alloys.
2 Governing equations

In this paper, we propose a mathematical model of the joint consideration of the electric plasma arc and the workpiece where their mutual influence on each other, taking into account the influence of metal vapor evaporated anode. Physical processes in electric arc column and interacting with the discharge of the liquid metal are described by a single system of magnetohydrodynamics equations \[1\]. When recording MHD equations in the simplest form it is assumed the following conditions: the plasma column is assumed to be in local thermodynamic equilibrium (LTE), the plasma is a Newtonian fluid, flows are the steady and laminar. MHD system of equations in cylindrical coordinates is as follows \[2\]:

The mass continuity equation:

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \rho u \right) + \frac{\partial}{\partial z} \left( \rho v \right) = 0 \quad (1)$$

The radial momentum conservation equation:

$$\rho v \frac{\partial v}{\partial r} + \rho u \frac{\partial v}{\partial z} = -\frac{\partial P}{\partial r} - j_z B_\phi + \frac{2}{3} \frac{\partial}{\partial r} \left( \mu \frac{\partial v}{\partial r} + \frac{\partial v}{\partial z} \right) - \frac{2}{3} \frac{\mu}{r} \left( \frac{1}{r} \frac{\partial (rv)}{\partial r} + \frac{\partial u}{\partial z} \right) \quad (2)$$

The axial momentum conservation equation:

$$\rho v \frac{\partial u}{\partial r} + \rho u \frac{\partial u}{\partial z} = -\frac{\partial P}{\partial z} + j_r B_\phi + \frac{1}{r} \frac{\partial}{\partial r} \left( \mu \left( \frac{\partial u}{\partial r} + \frac{\partial v}{\partial z} \right) \right) - \frac{2}{3} \mu \left( \frac{1}{r} \frac{\partial (ru)}{\partial r} + \frac{\partial u}{\partial z} \right) + S_u \quad (3)$$

The energy conservation equation:

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \rho v h - \frac{\lambda}{c_p} \frac{\partial h}{\partial r} \right) + \frac{\partial}{\partial z} \left( \rho u h - \frac{\lambda}{c_p} \frac{\partial h}{\partial z} \right) = \frac{1}{\sigma} \left( j_r^2 + j_z^2 \right) - q + S_{C_1} \quad (4)$$

Maxwell’s equations:

$$\frac{\partial E_r}{\partial z} - \frac{\partial E_z}{\partial r} = 0, \text{ } \frac{1}{r} \frac{\partial r H_\phi}{\partial r} = j_z, \text{ } -\frac{\partial H_\phi}{\partial z} = j_r \quad (5)$$

Ohm’s law:

$$j_r = \sigma E_z, \text{ } j_z = \sigma E_r \quad (6)$$

The system is supplemented by the equation of convective diffusion of metal vapor \[3\]:

$$\frac{1}{r} \frac{\partial}{\partial r} \left( r \rho v C_1 \right) + \frac{\partial}{\partial z} \left( \rho u C_1 \right) = \frac{1}{r} \frac{\partial}{\partial r} \left( r \rho D \frac{\partial C_1}{\partial r} \right) + \frac{\partial}{\partial z} \left( \rho D \frac{\partial C_1}{\partial z} \right) \quad (7)$$
where $u$, $v$ are axial and radial flow velocity; $P$ is pressure; $T$ is temperature; $j$ is current density, $E$ is intensity of electric field, $H$ is intensity of the magnetic field, $B$ is magnetic induction, $C_1$ is mass concentration of metal vapor, $D$ is diffusion coefficient, $\rho$ is density of the plasma, $c_p$ is specific heat, $\mu$ is viscosity, $\lambda$ is thermal conductivity, $q$ is radiation, $\sigma$ is electrical conductivity, $h$ is enthalpy.

In the momentum conservation equation:

$$S_u = \begin{cases} 0 & \text{for arc plasma} \\ \rho g - \rho g \beta (T - T_0) & \text{for weld pool} \end{cases}$$

where $\beta$ is coefficient of thermal expansion, $g$ is acceleration of gravity.

In the energy equation for the weld pool the effective heat capacity is used:

$$c_p^{ef} = c_p + \Delta H_f \frac{\partial f_l}{\partial T}$$

where $\Delta H_f$ is specific heat of melting of the anode material.

The weld pool liquid fraction $f_l$ varies linearly with temperature:

$$f_l = \begin{cases} 1 & T > T_l \\ \frac{T - T_s}{T_l - T_s} & T_s < T < T_l \\ 0 & T < T_s \end{cases}$$

where $T_s$ is solid phase temperature, $T_l$ is liquid phase temperature of the metal anode. Term

$$S_{C_1} = \frac{\partial}{\partial z} \left[ \left( \rho D - \frac{\lambda}{c_p} \right) (h_m - h) \frac{\partial C_1}{\partial z} \right] + \frac{\partial}{\partial r} \left[ \left( \rho D - \frac{\lambda}{c_p} \right) (h_m - h) \frac{\partial C_1}{\partial r} \right]$$

on the right side of the law of conservation of energy determines the enthalpy change due to the mixing of the metal vapor and the plasma gas, $h_m$ is enthalpy of the metal vapor.

The interaction between the plasma and metal vapor, their mutual influence on each other is determined by the thermal properties of the medium as a function of temperature and concentration of metal vapor in the plasma:

$$\sigma = \sigma(T, C_1), \quad \lambda = \lambda(T, C_1), \quad \mu = \mu(T, C_1), \quad \rho = \rho(T, C_1)$$

$$q = q(T, C_1), \quad h = h(T, C_1), \quad c_p = c_p(T, C_1)$$

For determine the diffusion coefficient used the approximation of viscous approximation [4]. Diffusion coefficient in the approximation is calculated by the formula:

$$D_{Ar-Fe} = \frac{2\sqrt{2} \left( \frac{1}{M_1} + \frac{1}{M_2} \right)^{0.5}}{\left( \left( \frac{\rho_1^2 \eta_1^2 M_1}{M_2} \right)^{0.25} + \left( \frac{\rho_2^2 \eta_2^2 M_2}{M_2} \right)^{0.25} \right)^{2}}$$

(8)
where \( M_1, M_2 \) are molar weight of the metal and the plasma gas, \( \rho_1, \rho_2, \mu_1, \mu_1 \) are density and viscosity of the metal and gas, respectively; \( \beta_1 = \beta_2 = 1.385 \) based on experimental data.

MHD system of equations is solved in the variables "vorticity-stream function", the introduction of the following variables: \( \omega \) is the intensity of the vortex, \( \psi \) is stream function, \( \chi \) is the function of the electric current, which in the case of a cylindrical coordinate system defined by the relations with axial symmetry:

\[
\omega = \frac{1}{r} \left( \frac{\partial v}{\partial z} - \frac{\partial u}{\partial r} \right); \quad \frac{\partial \psi}{\partial r} = \rho ur; \quad -\frac{\partial \psi}{\partial z} = \rho vr;
\]

\[
\frac{\partial \chi}{\partial r} = r j_z; \quad -\frac{\partial \chi}{\partial z} = r j_r;
\]

Then the original system can be written in the following canonical form:

\[
a \left[ \frac{\partial}{\partial z} \left( \varphi \frac{\partial \psi}{\partial r} \right) - \frac{\partial}{\partial r} \left( \varphi \frac{\partial \psi}{\partial z} \right) \right] - \frac{\partial}{\partial r} \left[ b \frac{\partial}{\partial r} (c \varphi) \right] - \frac{\partial}{\partial z} \left[ b \frac{\partial}{\partial z} (c \varphi) \right] + er = 0 \quad (9)
\]

where \( \varphi \) is desired function, taking values \( \omega, \psi, h, \chi \) and \( C_1 \); \( a, b, c, e \) are nonlinear coefficients corresponding to each of the equations. The values of these ratios are presented in the Table 1.

**Table 1.** Values of the coefficients of the canonical equations

<table>
<thead>
<tr>
<th>( \varphi )</th>
<th>( a )</th>
<th>( b )</th>
<th>( c )</th>
<th>( e )</th>
<th>( \omega )</th>
<th>( \rho_r )</th>
<th>( 1 )</th>
<th>( \omega r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \omega )</td>
<td>( r^2 )</td>
<td>( r^3 )</td>
<td>( \mu )</td>
<td>-2</td>
<td>( \frac{u^2 + v^2}{2} )</td>
<td>( \frac{\partial \rho}{\partial r} )</td>
<td>( -\frac{u^2 + v^2}{2} )</td>
<td>( \frac{\partial \rho}{\partial z} )</td>
</tr>
<tr>
<td>( \psi )</td>
<td>0</td>
<td>( \frac{1}{\rho_r} )</td>
<td>1</td>
<td>1</td>
<td>( \frac{1}{\rho_r} )</td>
<td>( (\frac{\partial \chi}{\partial r})^2 )</td>
<td>( (\frac{\partial \chi}{\partial z})^2 )</td>
<td>- ( qr )</td>
</tr>
<tr>
<td>( h )</td>
<td>1</td>
<td>( \frac{\Delta \rho r}{\rho r} )</td>
<td>1</td>
<td>1</td>
<td>( \frac{1}{\rho_r} )</td>
<td>( (\frac{\partial \chi}{\partial r})^2 )</td>
<td>( (\frac{\partial \chi}{\partial z})^2 )</td>
<td>- ( qr )</td>
</tr>
<tr>
<td>( \chi )</td>
<td>0</td>
<td>( \frac{1}{\sigma_r} )</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( C_1 )</td>
<td>1</td>
<td>( \rho r D )</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Single entry form allows for solving the system of equations to use the same calculation algorithm. To solve the resulting system of differential equations is necessary to set the boundary conditions for these functions. Since the system equations are of elliptic type, the boundary conditions must be given around the contour surrounding the computational domain. The computational domain is shown in Fig. 1. Real non-consumable cathode plasma torch is a cylinder with a flat end, as the anode serves a workpiece, the system is in a confined space, limited by side walls at a distance of \( R \).
3 Boundary conditions

1) Boundary conditions for the all solid walls are set as follows: Condition of impermeability for the stream function $\psi = 0$. The function $\omega$ is determined from the condition of sticking. The temperature is assumed to be $T_0 = 300$, thus determined $h = h(T_0)$. Electric current function is defined as $\chi = \frac{I}{2\pi}$. Metal vapor concentration equal zero $C_1 = 0$.

2) The boundary conditions at the cathode are defined as follows:

\[
\psi = 0;
\frac{\partial \psi}{\partial z} = 0;
T_k(r) = (T_b - T_0) \left( 1 - \frac{r}{R_k} \right)^m \left( 1 + \frac{r}{R_k} m \right) + T_0;
\chi_k = \frac{I}{2\pi \int_0^r \sigma r \, dr};
C_1 = 0.
\]

Fig. 1. Scheme of the computational domain
Here $T_b$ is the boiling point of the cathode, $m$ is the degree of filling of the temperature profile.

3) In the arc column axis of symmetry conditions are implied:
\[
\psi = 0; \quad \frac{\partial \omega}{\partial r} = 0; \quad \frac{\partial h}{\partial r} = 0; \quad \chi = 0; \quad \frac{\partial C_1}{\partial r} = 0;
\]

4) At the weld pool surface the boundary conditions are defined as follows:
\[
\psi = 0; \quad \frac{\mu_p \omega_p r}{\mu_0 \omega_0 r} = \frac{\partial \alpha}{\partial T} \frac{T}{\partial r}; \quad \lambda_a \frac{\partial T_a}{\partial z} = \lambda_a \frac{\partial T}{\partial z} - \sigma \varepsilon (T_a^4 - T_0) - W_v h_{fg}; \quad \frac{\partial \chi_p}{\partial z} = \frac{\partial \chi_a}{\partial z}; \quad C_1 = \frac{P_{vap} M_1}{P_{vap} M_1 + (P_{atm} - P_{vap}) M_2}. \quad \text{Here the index } "p" \text{ refers to the plasma arc, the index } "a" \text{ refers to the material of the anode; } \sigma \varepsilon \text{ to Stefan-Boltzmann coefficient; } \varepsilon \text{ to the emissivity of the anode; } h_{fg} \text{ to latent heat of evaporation; } W_v \text{ to evaporation rate, which is obtained from the following approximation [5]:}\]
\[
\log W_v = A_v + \log P_{atm} - 0.5T, \quad A_v-\text{constant depending on the workpiece material; } P_{vap} = P_{atm} \exp \left( \frac{-H_{vap}}{RT} \left( \frac{1}{T} - \frac{1}{T_{boi}} \right) \right) \text{- the partial vapor pressure of the metal, which is a function of the molten metal weld pool temperature.}
\]

5) At the lower boundary of the workpiece conditions are stated:
\[
\psi = 0; \quad \omega = 0; \quad h = h(T_0); \quad \frac{\partial \chi}{\partial z} = 0;
\]

In the area of anode, the equation of convective diffusion of metal vapor is not solved. The boundary conditions for the vorticity were set at a point at one step from the solid boundaries, thus avoiding the ambiguity of the boundary conditions at the corners to ensure sustainable convergence and bridge solutions on a rectangular grid for the boundary of any shape.

### 4 Numerical methods

The canonical equation was solved using integro-interpolation method based on finite difference approach [2]. Computational domain is covered by a rectangular orthogonal non-uniform grid. We are integrating the equation (9) for the area which is bounded by the dotted line (Fig. 2):

\[
\int_{z_{i-\frac{1}{2}}}^{z_{i+\frac{1}{2}}} \int_{r_{j-\frac{1}{2}}}^{r_{j+\frac{1}{2}}} a \left[ \frac{\partial}{\partial z} \left( \varphi \frac{\partial \psi}{\partial r} \right) - \frac{\partial}{\partial r} \left( \varphi \frac{\partial \psi}{\partial z} \right) \right] dr dz - \int_{z_{i-\frac{1}{2}}}^{z_{i+\frac{1}{2}}} \int_{r_{j-\frac{1}{2}}}^{r_{j+\frac{1}{2}}} \left[ \frac{\partial}{\partial z} \left( b \frac{\partial c \varphi}{\partial z} \right) + \frac{\partial}{\partial r} \left( b \frac{\partial c \varphi}{\partial r} \right) \right] dr dz + \int_{z_{i-\frac{1}{2}}}^{z_{i+\frac{1}{2}}} \int_{r_{j-\frac{1}{2}}}^{r_{j+\frac{1}{2}}} r dr dz = 0 \quad (10)
\]

We will call the first term as the convective term $I_c$, second as the diffusion term $I_d$, the third - source term $I_s$. The first-order derivatives are approximated...
Fig. 2. Difference grid fragment

by backward difference, the second-order derivatives are approximated by central difference scheme. After integration of the convective term we have:

$$I_c = \sum A_{k,l} (\varphi_{i,j} - \varphi_{k,l})$$

where

$$A_{k,l} = \frac{a_{i,j}}{8} (F_{k,l} + |F_{k,l}|)$$

$$F_{i-1,j} = \psi_{i-1,j+1} + \psi_{i,j+1} - \psi_{i-1,j-1} - \psi_{i,j-1}$$

$$F_{i+1,j} = \psi_{i+1,j-1} + \psi_{i,j-1} - \psi_{i+1,j+1} - \psi_{i,j+1}$$

$$F_{i,j-1} = \psi_{i+1,j+1} + \psi_{i+1,j} - \psi_{i-1,j+1} - \psi_{i-1,j}$$

After integration of the diffusion term we have:

$$I_d = \sum B_{k,l} (c_{k,l} \varphi_{k,l} - c_{i,j} \varphi_{i,j})$$

where

$$B_{i-1,j} = \frac{b_{i-\frac{1}{2},j}}{2} \frac{r_{j+1} - r_{j-1}}{z_{i-1} - z_i}, B_{i+1,j} = \frac{b_{i+\frac{1}{2},j}}{2} \frac{r_{j+1} - r_{j-1}}{z_{i-1} - z_i}$$

$$B_{i,j-1} = \frac{b_{i,j-\frac{1}{2}}}{2} \frac{z_{i+1} - z_{i-1}}{r_j - r_{j-1}}, B_{i,j+1} = \frac{b_{i,j+\frac{1}{2}}}{2} \frac{z_{i+1} - z_{i-1}}{r_j - r_{j+1}}$$

After integration of the source term, we have:

$$I_s = e_{i,j} r_j \frac{r_{j+1} - r_{j-1}}{2} \frac{z_{i+1} - z_{i-1}}{2}$$

Thus, the differential equation is transformed into a system of nonlinear algebraic equations:

$$\varphi_{i,j} = \frac{\sum [(A_{k,l} + c_{k,l} B_{k,l}) \varphi_{k,l}] - e_{i,j} r_j \frac{(r_{j+1} - r_{j-1})(z_{i+1} - z_{i-1})}{4}}{\sum (A_{k,l} + c_{i,j} B_{k,l})}$$

This system of nonlinear algebraic equations solved by an iterative method of Gauss-Seidel:

$$\varphi_{i,j}^\nu = \alpha (S_{i-1,j} \varphi_{i-1,j}^\nu + S_{i+1,j} \varphi_{i+1,j}^\nu + S_{i,j-1} \varphi_{i,j-1}^\nu + S_{i,j+1} \varphi_{i,j+1}^\nu + D_{i,j}) + (1-\alpha) \varphi_{i,j}^{\nu-1}$$

where

$$S_{k,l} = \frac{A_{k,l} + c_{k,l} B_{k,l}}{\sum (A_{k,l} + c_{i,j} B_{k,l})}$$
\[ D_{i,j} = \frac{1}{4} e_{i,j} r_j (r_{j+1} - r_{j-1}) (z_{i+1} - z_{i-1}) \]

Over relaxation method was used to improve the convergence of the iterative process and stopping criterion was:

\[
\max_{i,j} \frac{\left| \phi_{i,j}^\nu - \phi_{i,j}^{\nu-1} \right|}{\max_{i,j} \left| \phi_{i,j}^{\nu-1} \right|} < \varepsilon \approx 10^{-3}
\]

5 Results and discussion

Based on the properties of the pure components, with the help of software ASTRA and TERRA transfer coefficients for mixtures of Ar + 1% Fe, Ar + 3% Fe were calculated. The data are in good agreement with the data given in [6]. When the content of iron vapors is about 1% electrical conductivity and radiation have a noticeable difference in the temperature range from 5000 to 10000 K. In this area isotherm of 8000 K lies, which usually take the visible border of arc.

The calculations used the following data:
- The melting point of steel \( T_{\text{plav}} = 1773 \) K;
- The boiling point of steel \( T_{\text{boi}} = 3133 \) K;
- The specific heat of fusion \( \Delta H_f = 2.47 \times 10^5 \) J/kg;
- The molar weight of argon \( M_1 = 55 \times 10^{-3} \) kg/mol;
- Molar mass of steel \( M_2 = 27 \times 10^{-3} \) kg/mol;
- Molar heat of vaporization of steel \( H_{\text{vap}} = 340 \times 10^3 \) J/mol;
- Specific heat of vaporization of steel \( h_{fg} = 6.2 \times 10^6 \) J/kg;
- Steel surface tension is determined according to the data given in Fig. 3.

The calculations were performed for the current \( I = 150 \) A and 200 A. To current \( I=150 \) A maximum concentration of iron vapor on the surface of the weld pool on the basis of the boundary conditions was 0.6%, which does not affect the transport coefficients argon arc.

Fig. 4 shows graphs of the distribution of iron vapor concentration on the anode surface and within the scope of the electric arc at a current \( I = 200 \) A. The maximum concentration of iron vapor with a current of 200 A is 1.05% on the axis of the arc. The distribution of the concentration of metal vapor is determined by convective and diffusive fluxes. Axial gas flow rate directed to the anode 5 times the radial velocity of the anode surface, so the metal vapor in the axial part concentrated mainly near the surface of the anode, and the metal vapor expansion region occurs outside of the arc axis. Also, this is due to the nature of the diffusion coefficient, the maximum value of which falls on the periphery of the nucleus of the arc where the metal atoms easily diffuse into the arcing region. Thus, the axial part of the convection is predominant, so the metal vapor are drawn into a radial motion of the gas flow and flow over the anode surface.
Fig. 3. Surface tension gradient of steel

Fig. 4. Distribution of Fe vapors concentration a) on the anode surface b) in a column of the electric arc.

Fig. 5 shows the temperature fields with and without taking into account the metal vapor in the argon plasma at a current $I = 200$ A. The presence of metal vapor in the anode part narrow arc in radial direction, cooling the arc column at the edges, and heating the arc core. This is because the emissivity of a mixture of argon with significantly higher metal vapor in a temperature range of 5000 to 13000 K, which leads to an increase in radiation loss in the given interval and a narrowing of the arc. Another cooling mechanism of the arc on the periphery is to increase the thermal conductivity at temperatures below 8000 K, caused by greater diffusion of heat in the vicinity of the plasma arc. This arc cooling effect in the presence of metal vapor is consistent with the experimental and theoretical results of [7].

Fig. 6 shows graphs of current density in the arc column. The current density at the anode surface in the presence of iron vapor is reduced, it is because the
Fig. 5. The temperature distribution of arc, I = 200 A, a) for mixture Ar+1%Fe, b) for pure Ar

Fig. 6. Current density in the electric arc column, I = 200 A, a) for mixture Ar+1%Fe, b) for pure Ar

presence of vapor increases the electrical conductivity at temperatures below 10000 K, and electric current flows in the colder regions of the arc. Changing the conductivity a mixture of argon and metal vapor in the anode part is formed by two mechanisms. On the one hand, the presence of iron should increase plasma vapor conductivity. On the other hand, cooling of the arc due to the higher radiation losses and increase the thermal conductivity in the peripheral portion decreases the overall electrical conductivity of the mixture. As a result, the contribution of the electromagnetic component on the penetrating ability of the arc is reduced.

Fig. 7 illustrates the heat flux from arc column to the anode. Despite the fact that the core temperature of the arc to above metal vapor with argon, the heat flow toward the anode member to pure argon, due to the higher thermal
Fig. 7. Heat flux to the anode. Fig. 8. The surface temperature of the anode.

conductivity coefficient in this temperature range. Thus, the temperature of the weld pool surface in the presence of metal vapor is reduced (Fig. 8).

Fig. 9. The temperature distribution of weld pool, $I = 200$ A, a) for mixture Ar+1%Fe, b) for pure Ar

The properties of the workpiece produce a noticeable effect on the hydrodynamic conditions in the weld pool. The steel has a relatively low coefficient of thermal conductivity and high heat capacity ratio, which should lead to a shallow depth and radius of the weld pool (Fig. 9).

The plasma flow spreads radially from the surface of the molten metal and involves radial movement of the upper layers of the liquid metal due to shear stress of plasma convective flow and Marangoni convection and causes the formation of vortex in the weld pool volume. At the edges of the weld pool Marangoni force generated an additional reverse vortex involving in motion the same amount of
Fig. 10. The distribution of fluid flow and vector fields in the weld pool, I = 200 A a) for mixture Ar+1%Fe, b) for pure Ar

metal, as in the main vortex (Fig. 10). Since the intensity of mixing of metal in a small vortex is very high, this strong vortex flow carries heat deep into the pool, which leads to additional melting of the base metal at the edges of the bath. The above phenomena are formed similar to the form of the weld pool.

References

Supercomputer Modeling of Stochastic Dynamics of the Mercury Ion Array in an Optical Lattice

Leonid Kamenshchikov and Igor Krasnov

Institute of Computational Modeling of SB RAS, 660036 Krasnoyarsk, Russia
{lpk,krasn}@icm.krasn.ru
http://icm.krasn.ru

Abstract. Simulations of the resonant ions stochastic dynamics in the polychromatic optical field are presented. We prove the possibility of long-term four- and nine-particle ionic Coulomb planar clusters (crystals) by all-optical method. An estimate of lifetime of a single particle in an optical lattice is also carried out. Our analysis is based on the numerical solution of the stochastic differential equations with multiplicative noise using MVS-100K and MVS-10P supercomputers.

Keywords: Stochastic differential equations, Dissipative optical lattice, Coulomb clusters, Parallel Monte Carlo method, Distributed computing.

1 Introduction

Electromagnetic ion traps have many important applications in quantum informatics, high resolution spectroscopy of ions, metrology, physics of cold collisions and many-body physics [1–4]. A new and interesting trend in this field of research is the so-called all-optical confinement of ions, i.e. optical ion trapping without applying additional radiofrequency or electrostatic and magnetic fields [5–9]. In particular, it is assumed that development of all-optical methods of ion trapping can be useful for creation of ionic clock with better characteristics [10].

In our previous papers [7–9] the solution of all-optical ion trapping problem was proposed, based on using the rectified gradient forces that act on ions in the polychromatic field [11–15]. We demonstrated, by the numerical simulations of stochastic ion motion in the 3D polychromatic optical super-lattice (OSL), the long-term all-optical trapping of two- and three-ion ytterbium clusters in OSL. In the present work, we carried out numerical simulations of dynamics of four- and nine mercury ions in OSL and demonstrated the long-term all-optical trapping of ordered ion array (planar Coulomb cluster) in OSL. Note that a large array of trapped cold ions has attracted special interest from researchers because of its very useful applications [1, 3]. Now there is a broader interest in the form of arrays of ion traps in the context of quantum computings [4]. Stochastic dynamics of a single particle in an optical lattice is also considered.
The mathematical model is a system of stochastic differential equations (SDEs) for positions and velocities of each ionic particle. We take into account four acting forces in the model: the trapping, friction, Coulomb, and stochastic forces. The last force arises due to quantum fluctuations of the optical forces [11]. The Monte Carlo method with parallelization among computing cores is used to evaluate different average characteristics of this physical problem.

2 Equations of Stochastic Motion

Our study is based on the system of stochastic differential equations with multiplicative noise which can be written in the following dimensionless form

$$
\begin{align*}
    dr_\alpha(t,\omega) &= \zeta v_\alpha(t,\omega)dt, \quad t \in \mathbb{T} = [0, t_f] , \\
    dv_\alpha(t,\omega) &= F_\alpha(r_1,\ldots,r_N, v_\alpha)dt + \sqrt{2D(v_\alpha)} \circ dW_\alpha(t,\omega) , \\
    F_\alpha(r_1,\ldots,r_N, v_\alpha) &= F^{fr}_\alpha(r_\alpha, v_\alpha) + F^{fr}_\alpha(r_\alpha, v_\alpha) + F^C_\alpha(r_1,\ldots,r_N) , \\
    r_\alpha(0,\omega) &= r^0_\alpha(\omega), \quad v_\alpha(0,\omega) = v^0_\alpha(\omega), \quad \alpha = 1,\ldots,N ,
\end{align*}
$$

where $r_\alpha, v_\alpha \in \mathbb{R}^3$ are the position $r_\alpha$ and velocity $v_\alpha$ of the center-of-mass of the $\alpha$-th ion; $r_\alpha = (r_{\alpha x}, r_{\alpha y}, r_{\alpha z})^T$, $v_\alpha = (v_{\alpha x}, v_{\alpha y}, v_{\alpha z})^T$; $t$ is time; $N$ is the number of ions; $F^{fr}_\alpha, F^{fr}_\alpha, F^C_\alpha$ are the trapping, friction and Coulomb forces acting on the $\alpha$-th ion, respectively. The symbol $\circ$ means that this SDEs are interpreted in the Stratonovich sense [16].

$W_\alpha(t,\omega)$ is a standard three-dimensional vector Wiener process [17]. Recall some of its properties: 1) $W_\alpha(0,\omega) \equiv 0$; 2) for fixed $\omega \in \Omega$ the vector-function $W_\alpha(t,\omega)$ is continuous on $\mathbb{T}$; 3) for $\forall m > 1$ and $\forall \{t_k\} \in \mathbb{T}$ $(k = 1,\ldots,m)$ such that $0 < t_1 < t_2 < \cdots < t_m$, the random vectors $W_\alpha(t_1,\omega), W_\alpha(t_2,\omega) - W_\alpha(t_1,\omega), \ldots, W_\alpha(t_m,\omega) - W_\alpha(t_{m-1},\omega)$ are independent; 4) for $\forall t_1, t_2 \in \mathbb{T}, (t_1 < t_2)$, the random vector $W_\alpha(t_2,\omega) - W_\alpha(t_1,\omega)$ has a Gaussian distribution with mean 0 and dispersion matrix $(t_2 - t_1)I_3$, where $I_3$ is an unit matrix of the 3-th order. Notice also that for $\forall \alpha_1, \alpha_2, (1 \leq \alpha_1 < \alpha_2 \leq N)$, the random vectors $W_{\alpha_1}(t,\omega)$ and $W_{\alpha_2}(t,\omega)$ are independent.

The argument $\omega \in \Omega$ emphasizes that $r_\alpha, v_\alpha$, and $W_\alpha$ are the random vector functions in corresponding probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Below the letter $\omega$ will be omitted.

We use the dimensionless variables measuring the positions in units of $L = L_x$, time $t$ in units of $\omega_R^{-1}$ (where $\omega_R = \hbar k^2/m$ is the photon recoil frequency, $m$ is the ion mass), $L_i$ is a period of the OSL cell along $i$-axis ($i = x, y, z$), $k$ is the wave number, and velocities in units of $s_0 = \sqrt{T_D/\hbar}$ where $T_D = h\gamma/2$ is the characteristic temperature determining the so-called Doppler cooling temperature limit [11]. The dimensional time $t = t/\omega_R$ is also calculated for describing the simulation results; $\zeta = s_0/\omega_R L \ll 1$ is an analog of the Knudsen number (a small parameter) for the problem under consideration.
Components of the trapping force vector $\mathbf{F}_{tr}^\alpha = (F_{tr\alpha x}, F_{tr\alpha y}, F_{tr\alpha z})^T$ are [7]

$$F_{tr\alpha i} = -\zeta \frac{\partial U_i(r_{\alpha i}, v_{\alpha i})}{\partial r_{\alpha i}}, \quad i = x, y, z,$$

$$U_i(r, v) = -\frac{\cos(2\pi p_i r)}{p_i} \left( \frac{W_0 \mathcal{L}(v_{i c0})}{v_{i c0}} + \frac{W_1 \mathcal{L}(\frac{v}{v_{c1}})}{v_{c1}} \right),$$

where $W_0 = 4a_1 GL / 9\lambda(b + 1)(4\chi + 3)$; $W_1 = 2\chi W_0 / (4\chi + 3)$; $p_i = L/L_i$; $\mathcal{L}(u) = 1/(1 + u^2)$ is the Lorentzian function; $v_{i c0}^2 = (\gamma/6\omega_R)^2$ and $v_{c1}^2 = (\gamma/6\omega_R)^2(4\chi + 3)^2$ are the squares of the so-called capture velocities. The physical constants $a_1, G, b, \chi, p_i, a_1, \gamma, \lambda$ determine various ion and force characteristics of the three-dimensional OSL; it is supposed that $L \gg \lambda$, where $\lambda$ is the light wavelength.

Components of the friction force vector $\mathbf{F}_{fr}^\alpha$ are defined as

$$F_{fr\alpha i} = \left( -\kappa_i(r_{\alpha i}, v_{\alpha i}) + \frac{\partial D(v_{\alpha i})}{2v_{\alpha i}\partial v_{\alpha i}} \right) v_{\alpha i}, \quad i = x, y, z,$$

where

$$\kappa_i(r, v) = \kappa(v)[b + \cos(2\pi p_i r)]/(1 + b)$$

are the friction coefficients, $\kappa(v) = \kappa_0 \mathcal{L}(v/v_{c0}) + \kappa_1 \mathcal{L}(v/v_{c1})$, $\kappa_0 = 2a_1 G / 3\chi(4\chi + 3)$, $\kappa_1 = [2\chi^2/(4\chi + 3)^2]\kappa_0$.

The velocity diffusion coefficient can be written in the form [15]

$$D(v) = D_s + D_R(v),$$

where $D_s = 2\chi(\chi + 1)/3(4\chi + 3)$, $D_R(v) = D_0 \mathcal{L}(v/v_{c0}) + D_1 \mathcal{L}(v/v_{c1})$, $D_0 = (2G^2/9\chi)(16\chi^3 + 40\chi^2 + 33\chi + 9)/(4\chi + 3)^3$, $D_1 = (16G^2/9)(\chi + 1)/(4\chi + 3)^3$. An amplitude of the noise $\sqrt{2D(\mathbf{v}_\alpha)}$ in Eq. (2) is calculated on the basis of diffusion coefficients (9).

The long-range Coulombic interaction $\mathbf{F}_\alpha^C$ can be expressed via the dimensionless Coulomb energy $U^C(r) = e^2/(4\pi\varepsilon_0 rLT_D)$ of the ions separated by the distance $r$ in the following way

$$\mathbf{F}_\alpha^C(r_1, \ldots, r_N) = -\zeta \sum_{\alpha' = 1}^{N} \frac{\partial U^C(|r_\alpha - r_{\alpha'}|)}{\partial r_\alpha}.$$ 

Pay attention, the phases of optical fields forming OSL are set so that friction coefficients (8) reach maximum at the center of OSL cell (unlike the case of our previous articles [7, 8]).

3 Numerical Algorithm

For solution of Eqs. (1)–(4), we developed the numerical algorithm which is a combination of two other computational approaches: 1) the velocity Verlet
method [18] for integrating Newton’s ordinary differential equations of particle motion, and 2) the numerical scheme (so called an “integrator”) published by R. Mannella et al. [19–21] for solution of the Langevin stochastic equation. The integrator of Mannella has the following advantages: a) ability to reproduce the equilibrium behaviour and properties of dynamical system with a high accuracy; b) ability to well reproduce long-time dynamics of phenomena, such as large rare fluctuations [20], and hence correctly to describe decay of metastable states.

Recall the velocity form of the Verlet algorithm:

\[ \mathbf{r}_{\alpha}^{n+1} = \mathbf{r}_{\alpha}^{n} + \mathbf{v}_{\alpha}^{n} h + \mathbf{F}_{\alpha}(\mathbf{r}^{n}, \mathbf{v}^{n}) \frac{h^2}{2} + \mathbf{O}(h^3), \]
\[ \mathbf{F}_{\alpha}^{n+1} = \mathbf{F}_{\alpha}(\{\mathbf{r}^{n+1}\}), \quad \alpha = 1, \ldots, N, \]
\[ \mathbf{v}_{\alpha}^{n+1} = \mathbf{v}_{\alpha}^{n} + (\mathbf{F}_{\alpha}^{n+1} + \mathbf{F}_{\alpha}^{n}) \frac{h}{2} + \mathbf{O}(h^2). \]

However our numerical algorithm for solving stochastic system of differential equations (1)–(4) turns out considerably more complicated:

\[ \mathbf{r}_{\alpha}^{n+1} = \mathbf{r}_{\alpha}^{n} + \zeta \mathbf{v}_{\alpha}^{n} h + \zeta \mathbf{F}_{\alpha}(\mathbf{r}^{n}, \mathbf{v}^{n}) \frac{h^2}{2} + \zeta \mathbf{g}_{\alpha}^{n} \mathbf{Z}_{2,\alpha}^{n} + \zeta \mathbf{S}_{1} \mathbf{Z}_{3,\alpha}^{n} + \mathbf{O}(h^{5/2}), \quad \alpha = 1, \ldots, N, \quad (11) \]
\[ \mathbf{v}_{\alpha}^{n+1} = \mathbf{v}_{\alpha}^{n} + \left[ \mathbf{F}_{\alpha}(\mathbf{r}^{n}, \mathbf{v}^{n}) + \mathbf{F}_{\alpha}(\mathbf{r}^{n+1}, \mathbf{v}^{n}) \right] \frac{h}{2} + \mathbf{g}_{\alpha}^{n} \mathbf{Z}_{1,\alpha}^{n} + \mathbf{S}_{2} \mathbf{Z}_{2,\alpha}^{n} + \mathbf{S}_{1} (\mathbf{Z}_{1,\alpha}^{n})^2 + \mathbf{S}_{3} \mathbf{Z}_{1,\alpha}^{n} + \mathbf{S}_{4} (\mathbf{Z}_{1,\alpha}^{n})^3 + \mathbf{O}(h^2), \quad (12) \]

where

\[ \mathbf{r}^{n} = \{\mathbf{r}_{\alpha}^{n}\}; \quad \mathbf{v}^{n} = \{\mathbf{v}_{\alpha}^{n}\}; \quad \mathbf{g}_{\alpha}^{n} = \sqrt{2D(\mathbf{v}_{\alpha}^{n})}; \quad \mathbf{S}_{1} = \mathbf{g}_{\alpha}^{n}(\mathbf{g}_{\alpha}')^{n}/2; \]
\[ \mathbf{S}_{2} = -\kappa(\mathbf{r}^{n}, \mathbf{v}^{n})\mathbf{g}_{\alpha}^{n} + \mathbf{f}_{\alpha}\mathbf{g}_{\alpha}^{n} - (\mathbf{g}_{\alpha}')^{n}\mathbf{F}_{\alpha}(\mathbf{r}^{n}, \mathbf{v}^{n}); \]
\[ \mathbf{S}_{3} = (\mathbf{g}_{\alpha}')^{n}\mathbf{F}_{\alpha}(\mathbf{r}^{n}, \mathbf{v}^{n}); \quad \mathbf{S}_{4} = \{(\mathbf{g}_{\alpha}''')^{n}(\mathbf{g}_{\alpha}')^{n} + (\mathbf{g}_{\alpha}')^{2n}\mathbf{g}_{\alpha}^{n}\}/6; \]

with [20]

\[ \mathbf{Z}_{1,\alpha}(h) = \int_{0}^{h} dW(t) = \mathbf{Y}_{1,\alpha} h^{1/2}; \]
\[ \mathbf{Z}_{2,\alpha}(h) = \int_{0}^{h} \mathbf{Z}_{1,\alpha}(t) dt = \left[ \mathbf{Y}_{1,\alpha} + \mathbf{Y}_{2,\alpha} \frac{1}{\sqrt{3}} \right] \frac{h^{3/2}}{2}; \]
\[ \mathbf{Z}_{3,\alpha}(h) = \int_{0}^{h} \mathbf{Z}_{2,\alpha}(t) dt \approx \left[ (\mathbf{Y}_{1,\alpha})^2 + \mathbf{Y}_{3,\alpha} + \frac{1}{2} \right] \frac{h^2}{3}. \]

\( \mathbf{Y}_{1,\alpha}, \mathbf{Y}_{2,\alpha}, \mathbf{Y}_{3,\alpha} \in \mathbb{R}^3 \) are three uncorrelated random vectors with normal distribution \( N(0,1) \) (mean zero and standard deviation one) [19]; \( \mathbf{f}_{\alpha} \) is a given function of physical parameters; \( h \) is a time step. By definition, we assume here that for any vectors \( \mathbf{A}, \mathbf{B} \) the record \( \mathbf{AB} \) gives the vector \( (A_x B_z, A_y B_y, A_z B_z)^T \), and

\[ \sqrt{2D(\mathbf{v}_{\alpha}^{n})} \equiv \left( \sqrt{2D(\mathbf{v}_{\alpha x}^{n})}, \sqrt{2D(\mathbf{v}_{\alpha y}^{n})}, \sqrt{2D(\mathbf{v}_{\alpha z}^{n})} \right)^T. \]
The Monte Carlo method is used to evaluate both the average of the solution and the average for different functions from the solution (the size of the cluster, the cluster lifetime, kinetic energy, temperature, etc.). Due to the slow convergence of the Monte Carlo method, the volume of independent samples can be very large. We set different values (from $2^{14}$ to $2^{16}$) in different variants. The number of time steps reached $\sim 6 \cdot 10^7$. Use was made of 128–256 processing cores and the run time reached 12 hours. We used the uniform random number generator with period length $\approx 10^{38}$ from [22, 23].

To implement parallel computing the DVM-system developed in Keldysh Institute of Applied Mathematics of RAS was used. The calculations were carried out using the MVS-100K and MVS-10P supercomputers at the Joint Supercomputer Center of RAS.

Besides, we tested the algorithm by comparing our simulation result (for two-ion cluster [7]) with the analytical predictions of so-called renormalized model of the metastable (cluster) state of ions in the dissipative optical superlattice [9]. As a result, the very good agreement between the results of a numerical simulations and analytical results of renormalized model was obtained.

4 Results of Computation

In all computations we set $\gamma = 1.46 \cdot 10^8 \text{s}^{-1}$, $m = 199 \text{amu}$, $\lambda = 194 \text{nm}$, $\chi = 0.3$, $G = 2.2$, $b = 1.2$, $v_{c0}^2 = 6.542$, $v_{c1}^2 = 358.1$, $a_1^2 = 0.2$, $\zeta = 2.63 \cdot 10^{-3}$, $\omega_R = 3.35 \cdot 10^5 \text{s}^{-1}$, $p_x = p_y = 1$, $p_z = 0.5$ (i.e. the OSL cell is a cuboid). The number of ions $N$ takes values 1, 4, and 9.

In Fig. 1 a general view of the computational domain and initial positions of particles at $N=9$ is shown. Fig. 2 also shows initial position of ions in more

![Fig. 1. General view of the computational domain in dimensionless coordinates for the case of nine particles (N=9). Bold points are the initial positions of ions in the XOY-plane at Z=0. Here there are nine OSL cells in all domain.](image-url)
The example of nine-ion Coulomb cluster formation is shown in Fig. 3. Here the positions of ions (averaged over $2^{14}$ independent samples) in 0.8 seconds for two values of the parameter $L$ is presented. Average coordinates of the central ion coincide with its initial values.

Fig. 4 shows a behavior of a particle at one partial solution of the basic equations (without averagings).

**Fig. 2.** The initial positions of the ions correspond to the local minimum values of the potential functions, Eq. (6), for trapping forces.

**Fig. 3.** Formation of the nine-ions metastable Coulomb clusters: projections of the mean ions positions on the plane XOY (the bold points for $L = 0.5$ mm and the circles for $L = 0.6$ mm).

**Fig. 4.** An example of sample stochastic motion of a single particle in any OSL cell. The point 1 is the initial position, the initial velocity is zero. The particle tends to a point $(0, 0, 0)$ where there is a minimum of potential, and then it makes chaotic fluctuations there.

In Fig. 5 dependence of the clusters lifetime and a single particle on parameter $L$ is shown. The present result, i.e. almost linear dependence of $\ln \tau$ on $L$, are in a very good agreement with theories of the metastable states of stochastic dynamical system [26]. For the case of small noises they predict the exponential
dependence of the metastable state lifetime on the relative height of the energetic barrier $\Delta W/T$ (Arrhenius law). Indeed, the $\Delta W \sim L$ (at fixed parameters $G$ and $\chi$), therefore $\ln \tau \sim A + BL$, where $A$ and $B$ are almost independent on $L$ at fixed $G$ and $\chi$. The small deviation from the linear relation probably is caused by influence of a Coulomb interaction of ions on height of an energy barrier.

The relative root-mean-square interionic separation $\delta$ is defined as [24, 25]

$$\delta = \frac{2}{N(N-1)} \sum_{i<j} \sqrt{\langle r_{ij}^2 \rangle - \langle r_{ij} \rangle^2} / \langle r_{ij} \rangle,$$  \hspace{1cm} (13)

where $r_{ij}$ is a distance between ions $i$ and $j$. And Coulomb coupling parameter $\Gamma$ is defined as [3]

$$\Gamma = \frac{e^2}{4\pi \varepsilon_0 k_B L T},$$  \hspace{1cm} (14)

where $T$ is kinetic temperature, $L$ is an OSL period.

---

**Fig. 5.** The lifetime of nine-ions, four-ions clusters, and a single ion (solid line) versus parameter $L$. The left axis is scaled logarithmically but digits on the axis indicate the actual lifetime (in seconds).

The ordered crystal-like states (Coulomb cluster states) of ion array correspond to small values of $\delta$, $\delta \ll 1$, and large magnitude of large $\Gamma \gg 1$ [7]. It means that relative fluctuations of ion positions around their mean are small compared to the cluster size and interionic distance but ions are strongly coupled by the Coulomb interactions.

We see from Fig. 6 that for the long-time ion array the both necessary conditions (of Coulomb ion cluster formation) can be satisfied simultaneously. Note, that the case $\delta \approx 1$ corresponds to breakup of the cluster.
Note also that at $\delta \geq 0.1$ the lifetime of cluster is small (Figs. 5, 6). Cluster states quickly break up in view of a Coulomb interaction and quantum fluctuations of the optical forces. Pay attention that in the theory of clusters [24, 25] the value $\delta = 0.1$ usually defines a point of the cluster melting according to Lindemann’s criterion.

5 Conclusion

So, our numerical experiments prove that dissipative optical superlattices are able to form a long-term (up to $\sim 1000$ seconds) many-particle Coulomb cluster, which is the highly ordered array of mercury ions. Such cluster is characterized by the small values of the relative root-mean-square interionic distance, $\delta \ll 1$, and by the large magnitude of the Coulomb coupling parameter $\Gamma \gg 1$. Dependences of basic parameters of a Coulomb cluster on the period of a dissipative optical superlattice are investigated. They sufficiently well correspond to the known theories of metastable states of stochastic dynamical systems and clusters [25, 26]. Comparison of the obtained numerical results with results of theoretical paper [9] shows very good agreement. For numerical solution of stochastic equations, we give generalization of the well-known velocity Verlet scheme for accounting of a random force.

Our algorithm, Eqs. (11)–(12), allows to consider correctly key features of our stochastic model: a metastability of Coulomb clusters in OSL, non-conservatism of optical trapping forces, nonlinearity of friction coefficients (8), and a multiplicity of stochastic noise. Parallel realization of this algorithm on supercomputers was performed. In future works, it is planned to increase the number of ions up to several dozens.
References

Abstract. The article is devoted to optimal covering and packing problems for a bounded set in a two-dimensional metric space with a given amount of congruous circles. Such problems are of both theoretical interest and practical relevance. For instance, such statements appear in logistics when one needs to locate a given number of commercial or social facilities. A numerical algorithm based on fundamental physical principles due to Fermat and Huygens is suggested and implemented. It allows us to solve the problems for the cases of non-convex sets and non-Euclidean metrics. The results of numerical experiments are presented and discussed. Calculations show the applicability of the proposed approach its high efficiency for covering of a convex set in the Euclidean space by a sufficiently large amount of circles.

Keywords: circles covering problem, circles packing problem, non-Euclidean metric, optical-geometric approach, logistics, facilities, numerical algorithm, computational experiment.

1 Introduction

The facility location problem is a branch of mathematical modeling concerned with the optimal placement of facilities to minimize various negative factors. The Supply Chain Management Terms and Glossary defines facilities as “An installation, contrivance, or other thing which facilitates something; a place for doing something: Commercial or institutional buildings, including offices, plants and warehouses”. There is a number of papers devoted to the problem, see e.g. [11,12,37]. However, the known publications are usually concerned with certain particular cases. At the same time, we are aimed at a systemic solution of this problem at the level of regional, national and international transport and logistics systems (TLS).
In connection with the above, we develop a multi-stage technology for studying complex systems. On the first stage, we solve the problem of optimal placement of infrastructure logistic facilities, assumed the absence of these objects in the considered region. For example, there may be cellular towers of a certain operator, ATMs of particular bank, etc. This problem is reduced to special modifications of two well-known mathematical problems: covering of a bounded set in a two-dimensional space with non-Euclidean metric by equal circles. On the second stage, we solve the problem of optimal placement of additional logistics facilities in terms of cooperation and competition. The third stage assumes designing of a proper communication system for the above defined objects. On the final stage, we treat the problem of communications’ support in order to keep them in satisfactory conditions. Note that, though the developed approach operates with a variety of mathematical models, the key part is played by covering and packing problems.

The covering problem is to locate congruent geometric objects in a metric space so that its given area lies entirely within their union. This theoretical problem is widely used in solving practical tasks in various fields of human activity. Examples of such tasks are placement of cell towers, rescue points, police stations, ATMs, hospitals, schools [4,7,12,15], designing energy-efficient monitoring of distributed objects by wireless sensor networks [2,8,13,14] etc. Algorithms for covering of simply connected sets by congruent circles employing quasi-differentiability of the objective function are presented in [18], heuristic and metaheuristic methods can be found in [1,3,28,38], algorithms of integer and continuous optimization are proposed by [27,29,30]. A modification of feasible directions’ method appears in [32], where optimal coverings are given for different $n \leq 100$.

The optimal circle packing problem is to place objects of a prescribed form in a given container. Apparently, the most popular statement here is optimal packing two-dimensional spheres (circles, discs) in a convex set. For example, the authors of [9,26,33] consider the following problem: maximize the radius of a given number $n$ of congruent circles packed in a unit square. The number of circles varies between 1 and 200. Papers [16,17,25] address the problem of packing a family of equal circles of unit radius in a great circle. The results for number of packing elements up to 81 are obtained. Birgin and Gentil [5] consider the problem of packing equal circles of unit radius in a variety of containers (circles, squares, rectangles, equilateral triangles and strips of fixed height) in order to minimize the size of the latter.

Note that the most of known results are obtained for the case when covered areas or containers are subsets of the Euclidean plane or a multi-dimensional Euclidean space. In the case of a non-Euclidean metric, covering and packing problems are relatively poorly studied. Here we could mention the works by Coxeter [10] and Boroczky [6], which deal with congruent circles packing problems for multidimensional spaces of a constant curvature (elliptic and hyperbolic cases) and assess the maximum packing density. Besides above, this problem was studied in a series of papers by Szirmai. In [34,35] we find a method to deter-
mine the data and the density of certain optimal ball and horoball packings with Coxeter tiling for hyperbolic 3-, 4- and 5-D spaces, based on a projective interpretation of hyperbolic geometry. The goal of [36] is to extend the problem of finding the densest geodesic ball (or sphere) packing to different 3-D homogeneous geometries.

A detailed description of the proposed research technology for transport and logistics systems (including the developed software) is the subject of an extra publication. In the present paper, we are focused on mathematical modeling and their numerical implementation. The study follows our previous works on mathematical apparatus for problems of domestic logistics. In particular, in [19–23] we elaborate numerical algorithms based on optical-geometric analogy.

2 Mathematical models

Assume we are given a bounded domain containing a collection of disjoint “prohibited” areas, i.e. subdomains, where any activity (including passing through them) is banned. Suppose, the number of consumers is large enough, so that we can regard them as continuously distributed over the domain.

It is required to locate a given number of logistic centers so that, at first, they can serve the maximum possible proportion of the domain, at second, their service areas do not overlap, and, finally, the maximum time of delivery to the mostly distant consumer coincides for all logistics centers.

A mathematical model of the logistic problem is as follows.

Assume we are given a metric space $X$, a bounded domain $D \subset X$, compact sets $B_k \subset D, k = 1, \ldots, m$ (prohibited domains), and $n$ of logistic centers $S_n = \{s_i\}$ with coordinates $s_i = (x_i, y_i), i = 1, \ldots, n$. Let $0 \leq f(x, y) \leq \beta$ be a continuous function, which makes sense of the instantaneous speed of movement at every point of $D$. Note that $f(x, y) = 0 \Leftrightarrow (x, y) \in B_k, k = 1, \ldots, m$. Then, instead of $D$, we can consider closed multiply-connected set $P$:

$$P = \text{cl} \left( D \setminus \bigcup_{k=1}^{m} B_k \right) \subset X \subseteq \mathbb{R}^2. \quad (1)$$

Here $\text{cl}$ is the closure operator.

The distance in space $X$ is determined as follows:

$$\rho(a, b) = \min_{\Gamma \in G(a, b)} \frac{\int_{\Gamma} f(x, y) \, d\Gamma}{\sqrt{\int_{\Gamma} f(x, y) \, d\Gamma}}, \quad (2)$$

where $G(a, b)$ is the set of all continuous curves, which belong to $X$ and connect the points $a$ and $b$. In other words, the shortest route between two points is a curve, that requires to spend the least time.

We are to find a location $S_n^* = \{s_i^*\}$, which brings maximum to the expression

$$R^* = \min_{i=1,n} \min \left\{ \rho \left( s_i, (S_n \setminus \{s_i\}) \right), \rho(s_i, \partial P) \right\}. \quad (3)$$
Here, \( \partial P \) is the boundary of the set \( P \) and \( \rho(s_i, \partial P) \) is the distance from a point to a closed set,
\[
\rho(s_i, \partial P) = \min_{x \in \partial P} \rho(s_i, x). \tag{4}
\]

One can reformulate (3) as follows: maximize the radius of equal circles, which can be located so that they overlap each other and the boundary of the set \( P \) only at their boundary points. In other words, a solution to the logistic problem is equivalent to an optimal packing circles of equal radius in a multiply-connected set with metric (2).

Along with (3), we consider another problem: place a predetermined number of logistic centers so that, at first, all consumers are serviced, secondly, the maximum time of delivery to the mostly distant consumer is the same for all logistic centers, and the time of delivery is minimal.

Let \( M \subset X \) be a given bounded set with continuous boundary, \( m \) be an amount of logistic centers \( P_m = \{O_k\} \), and \( O_k = (x_k, y_k) \), \( k = 1, ..., m \), be their coordinates.

Our second goal is to find a partition of \( M \) on \( m \) segments \( M_k \), \( k = 1, ..., m \), and the location of the centers \( P_m^* = \{O_k^*\} \), which provide minimum for
\[
R_* = \max_{k=1,m} \rho(O_k, \partial M_k). \tag{5}
\]

The formulated logistic problem is equivalent to optimal covering of the set \( M \) with the metric (2) by circles of equal radius.

3 Numerical methods

In this section, the authors propose methods for solving problems (2),(3) and (2),(5), based on the analogy between the propagation of the light wave and finding the minimum of the functional integral (2). This analogy is a consequence of physical laws of Fermat and Huygens. The first principle says that the light in its movement chooses the route that requires to spend a minimum of time. The second one states that each point reached by the light wave, becomes a secondary light source. This approach is described more detail in [19, 20, 22, 23].

The essence of the algorithm is as follows. We consistently divide the given set \( P \) into segments with respect to the randomly generated initial set of circles centers based on Voronoi diagrams; then find the best center covering or packaging circle for each segment; finally construct segmentation for the found centers.

An algorithm for circles covering constructing

1. Randomly generate an initial coordinates of the circles centers \( O_k, k = 1, m \). Coordinate coincidences are not allowed.
2. From \( O_k, k = 1, m \) we initiate the light waves using the algorithm [19]. It allows us to divide set \( M \) on \( m \) segments \( M_k \) and to find their boundaries \( \partial M_k, k = 1, m \).
3. Boundary $\partial M_k$ of segment $M_k$ is approximated by the closed polygonal line with nodes at the points $A_i, i = 1, q$.

4. From $A_i, i = 1, q$ we initiate the light waves using the algorithm [19] as well.

5. Every point $(x, y) \in M_k$, first reached by one of the light waves is marked (here and further we assume using an analytical grid for $x$ and $y$). We memorize time $T(x, y)$ which is required to reach $(x, y)$.

6. Find $\bar{O}_k = \arg \max_{(x,y) \in M_k} T(x, y)$. Then, the minimum radius of circle which covers $M_k$, is given by

$$R_{k \min} = \max_{i = 1, q} \rho(\bar{O}_k, A_i).$$

Steps 3–6 are carried out independently for each segment $M_k, k = 1, m$.

7. Find $R_{\min} = \max_{k = 1, \ldots, m} R_{k \min}$. Then go to step 2 with $O_k = \bar{O}_k, k = 1, m$.

Steps 2–7 are being carried out while $R_{\min}$ is decreasing, then the current covering

$$P_m = \bigcup_{k = 1}^{m} C_k(\bar{O}_k, R_{\min})$$

is memorized as a solution.

8. The counter of an initial coordinates generations $Iter$ is incremented. If $Iter$ becomes equal some preassigned value, then the algorithm is terminated. Otherwise, go to step 1.

An algorithm for circles packing constructing

1. Randomly generate an initial coordinates of the circles centers $s_i, s_i \in P, i = 1, n$. Radius $R$ is assumed to be zero.

2. Domain $P$ is divided on segments $P_i, i = 1, \ldots, n$, as well as in the algorithm above.

3. We initiate the light waves propagating from the boundary of $\partial P_i$ of every segment $P_i$ in the inner area and construct the wave fronts until until they converge at a point. Denote this point by $\bar{s}_i$ and calculate $r_i = \rho(\bar{s}_i, \partial P_i)$ by (4), $i = 1, n$.

4. Calculate $R = \min_{i = 1, \ldots, n} r_i$.

Steps 2-4 are being carried out until $R$ is increasing, then the current vector $\bar{S}_n = \{\bar{s}_i\}$ is memorized as a solution.

5. The counter of an initial coordinates generations $Iter$ is incremented. If $Iter$ becomes equal some preassigned value, then the algorithm is terminated. Otherwise, go to step 1.

4 Computational experiment

Example 1. This example illustrates algorithm for circles covering constructing in the case of the Euclidean metric $f(x, y) \equiv 1$. We solve the equal circle covering problem in unit square. The number of circles is given and we maximize the
radius. The results are presented in table 1. Here $R_{\text{min}}$ is the best radius of covering obtained by the presented algorithm for circles packing constructing, $\Delta R = R_{\text{Known}} - R_{\text{min}}$, $t$ is time of calculation, $\text{Iter} = 25$.

Note, that the $R_{\text{Known}}$ results were obtained from [31].

**Table 1.** Comparison of the results of covering equal of circles in the unit square

<table>
<thead>
<tr>
<th>$n$</th>
<th>$R_{\text{Known}}$</th>
<th>$R_{\text{min}}$</th>
<th>$\Delta R$</th>
<th>$t$(sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.218233512793</td>
<td>0.218233693441</td>
<td>0.000000180648</td>
<td>43.477</td>
</tr>
<tr>
<td>15</td>
<td>0.179661759933</td>
<td>0.180281054179</td>
<td>0.000619294246</td>
<td>47.611</td>
</tr>
<tr>
<td>20</td>
<td>0.152246811233</td>
<td>0.152426892598</td>
<td>0.000180081365</td>
<td>51.683</td>
</tr>
<tr>
<td>25</td>
<td>0.133548706561</td>
<td>0.134470667521</td>
<td>0.000921960960</td>
<td>54.835</td>
</tr>
<tr>
<td>30</td>
<td>0.122036868819</td>
<td>0.123001449585</td>
<td>0.000964580766</td>
<td>58.984</td>
</tr>
<tr>
<td>40</td>
<td></td>
<td>0.108376286825</td>
<td></td>
<td>67.221</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>0.095904051463</td>
<td></td>
<td>73.508</td>
</tr>
<tr>
<td>75</td>
<td></td>
<td>0.078877824148</td>
<td></td>
<td>88.141</td>
</tr>
<tr>
<td>100</td>
<td></td>
<td>0.068332659403</td>
<td></td>
<td>128.342</td>
</tr>
<tr>
<td>150</td>
<td></td>
<td>0.05554107666</td>
<td></td>
<td>144.831</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.03082452774</td>
<td></td>
<td>408.707</td>
</tr>
</tbody>
</table>

Blank lines in table 1 means no known results for the corresponding $n$. It is easily seen that in comparison with known results, the results obtained by the authors, a little bit worse, but the deviation of circles radius does not exceed 0.1%. The total time for solving the problem is relatively small even for $n = 500$. It can be concluded that the proposed algorithm, despite the fact that it is, strictly speaking, not directly suitable for the covering problem in the Euclidean metric, shows reasonably good results here.

**Example 2.** This example shows a comparison of the results of the authors with the results from [24] and [31] for the packing of equal of circles in the unit square in the Euclidean metric $f(x, y) = 1$ (table 2).

It is easy to see that, as in example 1, the results obtained by the authors, is slightly worse, but the deviation of radius of packed circles from the optimal is low. Furthermore, when $n = 1,500$ we found a solution which improves the known one. At the same time, the total time for solving the problem is relatively small even for $n = 3000$ (for example, compared with the FSS-algorithm [24]).

**Example 3.** Let now $M = \{(x,y): (x - 6)^2 + (y - 6)^2 \leq 4^2 \}$ and

$$f(x, y) = \frac{(x - 4.5)^2 + (y - 6)^2}{(x - 4.5)^2 + (y - 6)^2 + 1} + 0.5.$$  

It is required to find the covering $P_n^*$ with minimal radius $R$ and $n = 8$.

The resulting approximation of coordinate of covering circles centers is following

$$S_8 \approx \{(3.610, 4.375), (3.725, 7.750), (5.603, 9.748), (6.0, 8.745) , \ldots \}.$$
Table 2. Comparison of the results of packing equal of circles in the unit square

<table>
<thead>
<tr>
<th>n</th>
<th>Packomania</th>
<th>FSS-Algorithm</th>
<th>Proposed Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R_{K_{max}}$</td>
<td>$R_{max}$</td>
<td>$\Delta R$</td>
</tr>
<tr>
<td>50</td>
<td>0.071377104</td>
<td>0.071376623</td>
<td>0.000000481</td>
</tr>
<tr>
<td>75</td>
<td>0.058494535</td>
<td>0.058091304</td>
<td>0.000403232</td>
</tr>
<tr>
<td>100</td>
<td>0.051401072</td>
<td>0.051227263</td>
<td>0.00128308</td>
</tr>
<tr>
<td>150</td>
<td>0.036612799</td>
<td>0.035722283</td>
<td>0.010890516</td>
</tr>
<tr>
<td>200</td>
<td>0.030219556</td>
<td>0.029447787</td>
<td>0.000771768</td>
</tr>
<tr>
<td>250</td>
<td>0.023455498</td>
<td>0.022480566</td>
<td>0.000974943</td>
</tr>
<tr>
<td>300</td>
<td>0.013157896</td>
<td>0.013163195</td>
<td>0.000876538</td>
</tr>
</tbody>
</table>

(6.115, 7.125), (6.918, 3.375), (7.628, 8.875), (9.156, 6.0) \}

Radius $R_{min} \approx 1.8134$. Set $M$ (bold line), covering $P^*_8$ (thin line) and coordinates of circles centers $S_8$ (dots) are shown at fig. 1.

Example 4. Here set $M$ and function $f(x, y)$ are the same as in the example 3. It is required to find the packing $U^*_n$ with maximum radius $R$ and $n = 8$.

The resulting approximation of circles centers coordinate is following

$$O_8 \approx \{(3.7997, 5.852), (6.547, 6.4971), (4.802, 6.2009), (8.5778, 5.132), (3.7201, 7.6607), (6.7970, 3.3827), (4.9825, 4.8146), (5.7846, 8.7356)\}.$$

Radius $R_{max} \approx 0.8787$. Set $M$ (bold line), packing $U^*_8$ (thin line) and coordinates of circles centers $O_8$ (dots) are shown at fig. 2.

5 Conclusions

We raised two classical problems of continuous optimization: optimal covering and optimal packing with equal 2-D spheres (circles, discs) for a bounded subset of a metric space. Practically, the addressed issues appear in logistics (so-called “facility location problems”), communication, security, energy management etc. The developed numerical algorithms, based on fundamental physical principles by Fermat and Huygens (an “optical-geometrical approach”), prove themselves rather efficient for covering and packing problems with non-convex sets, even if the metric is non-Euclidean: numerical simulation confirms that the designed approach is relevant. At the same time, in the case of Euclidean metric and a sufficiently large number of covering circles, our algorithms are shown to be competitive, compared to known approaches. The latter observation was pleasantly unexpected.

The obtained results could be further extended to multiply covering problems. This would be a subject of our future study.
Acknowledgments. The study is partially supported by the Russian Foundation for Basic Research, projects nos 16-31-00356, 16-06-00464.

References

Some Exact Solutions of a Heat Wave Type of a Nonlinear Heat Equation

Alexander Kazakov¹, Anna Lempert¹, Sergey Orlov², and Svyatoslav Orlov¹

¹ Matrosov Institute for System Dynamics and Control Theory, SB RAS, Lermontov st. 134, 664033 Irkutsk, Russia
{kazakov,lempert}@icc.ru, orlov_svyatoslav@list.ru
http://www.icc.irk.ru

² Institute of Mathematics, Economics and Informatics of Irkutsk State University, K. Marx st. 1, 664015 Irkutsk, Russia
orlov_sergey@inbox.ru
http://isu.ru

Abstract. The exact solutions of the nonlinear heat (porous medium) equation are constructed. We obtain a new class of the heat wave type solutions the construction of which is reduced to the Cauchy problems for nonlinear second order differential equations with a singularity. For these problems we prove a new existence and uniqueness theorem in the class of analytic functions. A special case of the heat wave front is considered in details. The results of numerical experiments are presented and discussed.

Keywords: nonlinear heat equation, porous medium equation, exact solutions, heat waves, Cauchy problem, existence and uniqueness theorem.

1 Introduction

The heat equation [1,2] is one of the well-known objects of classical mathematical physics. If a thermal conductivity does not depend on temperature, we have the linear equation. This case is well studied and we do not consider it. In this paper we deal with the nonlinear heat equation when the coefficient of thermal conductivity has a power-law dependence on the temperature. Besides heat conduction this equation also describes the ideal polytropic gas filtration in a porous medium. Therefore, in the literature it is also called “the porous medium equation” [2,3].

Solutions of a heat wave type are an important and interesting class of nonlinear heat equation solutions. Description of the process of the heat wave spread across the cold background at a finite speed, and the first examples of heat wave type solution were given by Ya.B. Zel’dovich in [4]. In the class of analytical functions the boundary-value problem with degeneration (Sakharov’s problem of the initiation of the heat wave) was first considered by A.F. Sidorov in [5]. The inverse problem, where for a given edge of the heat wave solution is recovered,
including the boundary regime, was studied by S.P. Bautin in [6]. There are certain papers of the scientific Sidorov’s school members, which are devoted to this problem [7–9]. The numerical methods for the construction of a heat wave are proposed in [10, 11].

In this paper we construct exact solutions of the heat wave type for the nonlinear one-dimensional heat equation. The construction reduces to the Cauchy problem for nonlinear ordinary differential equations of second order with a singularity at the highest derivative. In the literature such solutions of nonlinear partial differential equations are called “the exact solutions” [12, 13]. The obtained exact solutions allow us to find some of global properties of heat waves.

2 Problem Statement

We consider the nonlinear parabolic equation

\[ T_t = \text{div}(k \nabla T) , \]

in the case of \( k = T^\sigma, \sigma \in \mathbb{R}^{>0} \) (the porous medium equation) [2,3], i.e.

\[ T_t = \text{div}(T^\sigma \nabla T) . \] (1)

Here \( T \) is a function (temperature), depending on the time \( t \geq 0 \) and \( x \) be a vector of spatial variables. Operators \( \text{div} \) and \( \nabla \) act on \( x \).

If there are the symmetries, Eq. (1) can be converted to the form of one-dimensional heat equation

\[ u_\tau = uu_{\rho\rho} + \frac{1}{\sigma}u_{\rho}^2 + \frac{\nu}{\rho}uu_{\rho} , \quad \nu \in \{0, 1, 2\} , \] (2)

where \( u : D \rightarrow \mathbb{R} \) is a unknown function, defined on a set \( D \subset \mathbb{R}^2 \). It depends on the time variable \( \tau \geq 0 \) and the space variable

\[ \rho \overset{\text{def}}{=} ||x|| = \left( \sum_{k=1}^{\nu+1} x_k^2 \right)^{\frac{1}{2}} . \]

If \( \nu \neq 0 \), it should be noted that \( \rho \neq 0 \).

The values of the parameter \( \nu \) correspond to the heat propagation on the line, on the plane and in the space of symmetrically with regard to the origin.

In this paper we construct and study the exact heat wave-type solutions of Eq. (2), which satisfy the condition

\[ u|_{\rho=f(\tau)} = 0 , \] (3)

where \( \rho = f(\tau) \) is a front of the heat wave, defined in the plane of the variables \( (\tau, \rho) \). We have found that the boundary problem (2), (3), besides the trivial solution \( u(\tau, \rho) = 0 \), which is obvious, has some nontrivial classes of exact solutions.
Similar one-dimensional nonlinear heat conduction problems with the heat flux at the origin specified in the form of an exponential time dependence are considered in paper [14]. We construct exact (automodel) and approximate solutions of this problem.

3 Construction of Exact Solutions

This section is dedicated to finding non-trivial heat wave-type solutions of Eq. (2), the construction of which is associated with the solution of ordinary differential equations.

We assume that

\[ u(\tau, \rho) = \psi(\tau, \rho)w(\xi) , \quad \xi \overset{\text{def}}{=} \xi(\tau, \rho) , \quad (4) \]

where \( \psi, \xi \) and \( w \) are twice continuously differentiable functions of their variables, such that \( \psi_\tau \xi_\tau \xi_\rho \neq 0 \). Now we substitute (4) in (2) and find the acceptable expression for \( \psi(\tau, \rho) \) and \( \xi(\tau, \rho) \).

After dividing the resulting equation by \( \psi_\rho^2 \xi_\rho^2 \neq 0 \), we have

\[
ww'' + \frac{1}{\sigma}(w')^2 + \left[ \frac{2}{\sigma}(1 + 1) \frac{\psi_{\rho\rho}}{\psi_\xi^2} + \frac{\xi_{\rho\rho}}{\xi_\rho^2} + \nu \frac{1}{\rho \xi_\rho} \right] ww' + 
+ \left( \frac{1}{\sigma} \frac{\psi_{\rho\rho}^2}{\psi_\xi^2} + \frac{\psi_{\rho\rho}}{\psi_\xi^2 \xi_\rho^2} + \nu \frac{\psi_{\rho\rho}}{\rho \psi_\xi^2 \xi_\rho^2} \right) w^2 - \frac{\xi_\tau}{\psi_\xi^2} w' - \frac{\psi_\tau}{\psi_\xi^2 \xi_\rho^2} w = 0 .
\]

In order that the obtained expression becomes an ODE we should solve an overdetermined system of partial differential equations

\[
\frac{\psi_\rho}{\psi_\xi} = a_1 , \quad \frac{\xi_{\rho\rho}}{\xi_\rho^2} = a_2 , \quad \frac{\psi_{\rho\rho}}{\psi_\xi^2} = a_3 , \quad \frac{\xi_\tau}{\psi_\xi^2} = a_4 , \quad \frac{\psi_\tau}{\psi_\xi^2 \xi_\rho^2} = a_5 , \quad \frac{\psi_\rho}{\rho \psi_\xi^2 \xi_\rho^2} = a_6 , \quad \frac{1}{\rho \xi_\rho} = a_7 , \quad (5)
\]

where \( a_l \in \mathbb{R}, l = 1, 7 \).

Proposition 1. Let \( \nu \neq 0 \), then the system (5) is solvable if

\[
2a_2 = -a_1 = -2a_7 , \quad 2a_2^2 = a_3 = a_6 , \quad a_2 \neq 0 .
\]

Proof. Let \( a_4 \neq 0 \). We have \( \psi(\tau, \rho) = \xi_\tau/(a_4 \xi_\rho^2) \) from the fourth equation of system (5). We can find \( \xi(\tau, \rho) \) from the second and seventh equations. These two equations are solvable, only if \( a_2 = -a_7 \neq 0 \). In this case \( \xi(\tau, \rho) = \ln|f(\tau)\rho|^{-1/a_2} \). Substituting \( \psi(\tau, \rho) \) and \( \xi(\tau, \rho) \) in (5), we obtain the relations \( 2a_2 = -a_1 = -2a_7, 2a_2^2 = a_3 = a_6 \) and solvable by quadratures ODE

\[
f f'' + \left( \frac{a_5}{a_2 a_4} - 1 \right) (f')^2 = 0 ,
\]

346
which determines
\[ f(\tau) = \begin{cases} 
C_2 e^{C_1 \tau}, & \text{if } a_5 = 0, \\
(C_1 \tau + C_2)^{\frac{a_3 a_4}{a_5}}, & \text{if } a_5 \neq 0.
\end{cases} \]

Consequently, the system (5) is solvable.

2°. Let \( a_4 = 0 \). From the fourth equation of system (5) we have \( \xi(\tau, \rho) = \xi(\rho) \).
Then second and seventh equations provide that \( a_2 = -a_7 \neq 0 \) and we have \( \xi(\rho) = \ln[c\rho]^{-1/a_2} \). Substituting \( \xi(\rho) \) in (5), we obtain the system of equations
\[ \frac{\rho \psi }{\psi } = -\frac{a_1}{a_2}, \quad \frac{\rho^2 \psi }{\psi} = \frac{a_3}{a_2^2}, \quad \frac{\rho^2 \psi }{\psi^2} = \frac{a_5}{a_2^2}, \quad \frac{\rho \psi }{\psi} = \frac{a_6}{a_2^2}. \] (6)

Equations (6) have solutions
\[ \psi(\tau, \rho) = f_1(\tau)\rho^{-\frac{a_1}{a_2}}, \quad \psi(\tau, \rho) = f_2(\tau)\rho^{\frac{a_6}{a_2}}. \]
\[ \psi(\tau, \rho) = f_3(\tau)\rho^{\frac{a_2 + \sqrt{a_2^2 + 4a_3}}{2a_2}} + f_4(\tau)\rho^{\frac{a_2 - \sqrt{a_2^2 + 4a_3}}{2a_2}}, \quad \psi(\tau, \rho) = \frac{a_2^2 \rho^2}{a_2 f_5(\rho)\rho^2 - a_5 \tau}. \]

Thus, it is obvious that for the compatibility of (6) and, as a consequence, the system (5) as well, it is required that
\[ -\frac{a_1}{a_2} = \frac{a_6}{a_2^2} = \frac{a_2 \pm \sqrt{a_2^2 + 4a_3}}{2a_2} = 2 \iff 2a_2 = -a_1, \ 2a_2^2 = a_6 = a_3. \]

The proposition is proved. \( \square \)

The case \( \nu = 0 \) deserves a special attention. Here the system for \( \xi(\tau, \rho) \) and \( \psi(\tau, \rho) \) consists of five equations:
\[ \frac{\psi }{\psi \xi } = a_1, \quad \frac{\xi }{\xi ^2} = a_2, \quad \frac{\psi }{\psi \xi } = a_3, \quad \frac{\xi }{\xi ^2} = a_4, \quad \frac{\psi }{\psi ^2 \xi } = a_5, \] (7)
where \( a_l \in \mathbb{R}, \ l = 1, 5 \).

**Proposition 2.** Let \( \nu = 0 \), then (7) is solvable if
\[ 2a_2 = -a_1, \quad 2a_2^2 = a_3. \]

**Proof.** 1°. a) Let \( a_2, a_4 \neq 0 \). We have \( \psi(\tau, \rho) = \xi_{\tau}/(a_4 \xi^2) \) from the fourth equation of system (7) and \( \xi(\tau, \rho) = \ln[f(\tau)\rho + g(\tau)]^{-1/a_2} \) from the second one. Thus, substituting the expression for \( \psi(\tau, \rho) \) and \( \xi(\tau, \rho) \) in (7), we get the system of ODE’s for \( f(\tau) \) and \( g(\tau) \):
\[ \frac{(2\rho f + g)f' + f g'}{f(\rho f' + g')} = -\frac{a_1}{a_2}, \quad \frac{(\rho f + g)f'}{f(\rho f' + g')} = \frac{a_3}{2a_2^2}, \quad \rho f(\rho f + g)f'' + f(\rho f + g)g'' - \rho(\rho f + 2g)(f')^2 + f(g')^2 - 2gf'g' = 0 \]
\[ f(\rho f' + g')^2 \]
\[ = -\frac{a_5}{a_2a_4}. \]
In order to get rid of the variable $\rho$ in (8) we demand that $g(\tau) \equiv 0$. Then from the first and second equations we have $a_1 = -2a_2$ and $a_3 = 2a_2^2$, respectively, and the third one is converted to the exactly solvable ODE
\[
f f'' + \left( \frac{a_5}{a_2 a_4} - 1 \right) (f')^2 = 0 ,
\]
which determines
\[
f(\tau) = \begin{cases} C_2 e^{C_1 \tau}, & \text{if } a_5 = 0 , \\
(C_1 \tau + C_2) \frac{a_2 a_4}{a_5}, & \text{if } a_5 \neq 0 .
\end{cases}
\]
Consequently, the system (7) is solvable.

**b)** Let $a_2 \neq 0$, $a_4 = 0$. We have $\xi(\tau, \rho) = \xi(\rho)$ from the fourth equation of system (7). With this in mind we obtain $\xi(\rho) = \ln[c_1 \rho + c_2]^{-1/a_2}$ from the second equation of system (7). Substituting $\xi(\rho)$ in (7), we obtain the system of equations
\[
\frac{(\rho + c)\psi_\rho}{\psi} = -\frac{a_1}{a_2}, \quad \frac{(\rho + c)^2\psi_{\rho\rho}}{\psi} = \frac{a_3}{a_2^2}, \quad \frac{(\rho + c)^2\psi_\tau}{\psi^2} = \frac{a_5}{a_2^2} , \quad (9)
\]
where $c = c_2/c_1$. Equations (9) have solutions
\[
\psi(\tau, \rho) = f_1(\tau)(\rho + c)^{-\frac{a_1}{a_2}} , \\
\psi(\tau, \rho) = f_2(\tau)(\rho + c)^{-\frac{a_2 + \sqrt{a_2^2 + 4a_3}}{2a_2}} + f_3(\tau)(\rho + c)^{-\frac{a_2 - \sqrt{a_2^2 + 4a_3}}{2a_2}} , \\
\psi(\tau, \rho) = \frac{a_2^2(\rho + c)^2}{a_2^2 f_4(\rho)(\rho + c)^2 - a_5 \tau} .
\]
Thus, it is obvious that for the compatibility of (9) and, as a consequence, of the system (7) it is required that
\[
-\frac{a_1}{a_2} = \frac{a_2 \pm \sqrt{a_2^2 + 4a_3}}{2a_2} = 2 \iff 2a_2 = -a_1, 2a_2^2 = a_3 .
\]

**2. a)** Let $a_2 = 0$, $a_4 \neq 0$. We have $\psi(\tau, \rho) = \xi(\tau)/(a_4 \xi_0^2)$ from the fourth equation of system (7) and $\xi(\tau, \rho) = f(\tau)\rho + g(\tau)$ from the second one. Thus, substituting the expression for $\psi(\tau, \rho)$ and $\xi(\tau, \rho)$ in (7), we get $a_3 = 0$ and the system of ODE’s for $f(\tau)$ and $g(\tau)$:
\[
\frac{f'}{f(\rho f' + g')} = a_1 , \quad \frac{\rho f'' + fg'' - 2f(f')^2 - 2f'g'}{f(\rho f' + g')^2} = \frac{a_5}{a_4} . \quad (10)
\]
To eliminate the variable $\rho$ in (10) we have to demand that $f(\tau) \equiv \text{const}$. Then from the first equation we have $a_1 = 0$, and the second one is converted to the exactly solvable ODE
\[
g'' - \frac{a_5}{a_4} (g')^2 = 0 ,
\]
348
which determines

\[ g(\tau) = \begin{cases} 
C_1 \tau + C_2, & \text{if } a_5 = 0, \\
\ln(C_1 \tau + C_2)^{-\frac{a_4}{a_5}}, & \text{if } a_5 \neq 0.
\end{cases} \]

Consequently, the system (7) is solvable.

b) Let \( a_2, a_4 = 0 \). We have \( \xi(\tau, \rho) = \xi(\rho) \) from the fourth equation of system (7). Given this, we obtain \( \xi(\rho) = c_1 \rho + c_2 \) from the second equation of system (7). Substituting \( \xi(\rho) \) in (7), we obtain the system of equations

\[ \frac{\psi_\rho}{\psi} = c_1 a_1, \quad \frac{\psi_\rho^2}{\psi^2} = c_1^2 a_3, \quad \frac{\psi_\tau}{\psi^2} = c_1^2 a_5. \]  

Equations (11) have solutions

\[ \psi(\tau, \rho) = f_1(\tau)e^{c_1 a_1 \rho}, \quad \psi(\tau, \rho) = f_2(\tau)e^{c_1 \sqrt{a_3} \rho} + f_3(\tau)e^{-c_1 \sqrt{a_3} \rho}, \]

\[ \psi(\tau, \rho) = \frac{1}{f_4(\rho) - c_1^2 a_5 \tau}. \]

Thus, it is obvious that for the compatibility of the system of equations (11) and, as a consequence, of the system (7) it is required to \( a_1 = a_3 = 0 \).

The proposition is proved. \( \square \)

Using the obtained results we can present the following non-trivial exact solution of the equation (2):

\[ u(\tau, \rho) = f'(\tau)w(\xi), \quad \xi = \rho - f(\tau), \quad f(\tau) = \begin{cases} 
C_1 \tau + C_2, \\
\ln(C_1 \tau + C_2)^\alpha;
\end{cases} \]

\[ u(\tau, \rho) = f'(\tau)\rho^2 w(\xi), \quad \xi = \ln[\rho/f(\tau)], \quad f(\tau) = \begin{cases} 
C_2 e^{C_1 \tau}, \\
(C_1 \tau + C_2)^\alpha,
\end{cases} \]

where \( \alpha \neq 0, \ |C_1| + |C_2| > 0 \), and (12) takes place only when \( \nu = 0 \). Note that \( w(\xi) \) in (12) satisfies the ODE

\[ ww'' + \frac{1}{\sigma}(w')^2 + w + K(\alpha)w = 0, \]  

where \( K(\alpha) \) is equal to zero or \( \alpha^{-1} \) if \( f \) is a linear or logarithmic function, respectively. \( (\xi) \) in (13) satisfies the ODE

\[ ww'' + \frac{1}{\sigma}(w')^2 + \left( \nu + 3 + \frac{4}{\sigma} \right) ww' + w' + \left( 2\nu + 2 + \frac{4}{\sigma} \right) w^2 + K(\sigma)w = 0, \]  

where \( K(\alpha) \) is equal to zero or \( \alpha^{-1} \) if \( f \) is exponential or power-law function, respectively.
It is obvious that the solutions of (12) and (13) are of heat wave type solutions and satisfy the boundary condition (3) if and only if the solutions \( w(\xi) \) of (14) and (15) satisfy the initial conditions

\[
  w|_{\xi=0} = 0, \quad w'|_{\xi=0} = -\sigma, \quad (16)
\]

Thus, in this section we obtain exact solutions of the heat wave type (12) and (13), the procedure of construction is reduced to the solution of the Cauchy problem (14), (16) and (15), (16) respectively. Next the important question concerning the solvability of these problems will be investigated.

**Remark 1.** If \( f(\tau) = C_1 \tau + C_2 \) we have the known linear heat wave type solution

\[
  u(\tau, \rho) = \sigma C_1 (C_1 \tau - \rho + C_2).
\]

Indeed, in this case \( K(\sigma) = 0 \) and the Cauchy problem (14), (16) have a unique solution \( w(\xi) = -\sigma \xi \). Then from (12) we obtain a linear function.

**Remark 2.** If \( f(\tau) = (C_1 \tau + C_2)^\alpha, \quad \alpha = 1, \quad \nu = 0 \) we have a linear heat wave as well.

### 4 The Existence and Uniqueness of Solutions

The Cauchy problem for ordinary differential equations, which in the previous section was reduced to the construction of exact solutions of the equation (2), have a singularity, since \( \xi = 0 \) degenerates the order of the equations. Therefore, the existence of their solutions requires additional study, which will be done in this section. Consider the general form of the problem

\[
  w w'' + \frac{1}{\sigma} (w')^2 + w' + K_1 w w' + K_2 w^2 + K_3 w = 0, \\
  w|_{\xi=0} = 0, \quad w'|_{\xi=0} = -\sigma, \quad (17)
\]

where \( K_i \in \mathbb{R}, \ i = 1, 3 \). We have the following theorem.

**Theorem 1.** The Cauchy problem (17) has a unique nontrivial analytic solution in a neighborhood of \( \xi = 0 \).

**Proof.** The proof is presented briefly because it is carried out by standard procedure of the majorants method.

The solution of the Cauchy problem (17) is constructed in the form of a power series

\[
  w(\xi) = \sum_{n=0}^{+\infty} a_n \xi^n, \quad a_n \overset{\text{def}}{=} \frac{w^{(n)}(\xi)|_{\xi=0}}{n!}, \quad (18)
\]

In this case, \( a_0 \equiv 0, \ a_1 \equiv -\sigma \), and the remaining coefficients of the series (18) are uniquely determined according to the recurrence formula
\[ a_{n+1} = \frac{1}{\sigma(\sigma n + 1)(n + 1)} \left[ \sigma \sum_{k=0}^{n-2} (k+1)(k+2)a_{k+2}a_{n-k} + \sum_{k=1}^{n-1} (k+1)(n-k+1)a_{k+1}a_{n-k+1} + K_1 \sum_{k=0}^{n-1} (k+1)a_{k+1}a_{n-k} + K_2 \sum_{k=1}^{n-1} a_ka_{n-k} + K_3a_n \right], \quad n \in \mathbb{N}. \]

Next, we move to a new function \( v \) defined by the formula
\[ w(\xi) = -\sigma \xi + \xi^2 v(\xi). \]

Thus, we have the Cauchy problem
\[ Av + B\xi v' + C\xi^2 v'' = D + \xi g_1(\xi,v) + \xi^2 g_2(\xi,v,v') + \xi^3 g_3(\xi,v,v',v''), \quad v|_{\xi=0} = v_0, \quad v'|_{\xi=0} = v_1, \]
where \( A, B, C \in \mathbb{R}_+, \ D \in \mathbb{R}, \) and \( g_{1,2,3} \) are analytic functions of their arguments (a specific type of these constants and functions is irrelevant for the proof).

Majorant Cauchy problem for (19) has the form
\[ V'' = E[(G_1)\xi + (G_1)V' + G_2 + \xi G_3] , \quad V|_{\xi=0} = V_0, \quad V'|_{\xi=0} = V_1, \]
where
\[ E = \max_{n \in \mathbb{Z}_{\geq 0}} \left[ \frac{(n-1)n+1}{A + nB + (n-1)nC} \right], \]
\[ G_1 \overset{\text{def}}{=} G_1(\xi,V), \quad G_2 \overset{\text{def}}{=} G_2(\xi,V,V'), \quad G_3 \overset{\text{def}}{=} G_3(\xi,V,V',V''), \]
\[ V_0 > v_0, \quad V_1 > v_1, \quad G_i > g_i, \quad i = 1,3. \]

It is easy to show that the Cauchy problem (20) in a neighborhood of \( \xi = 0 \) has a unique analytic solution majorizing zero. Consequently, the functions \( v \) and \( w \) are also analytical.

The theorem is proved. \( \square \)

Therefore, the local solvability of the Cauchy problem (17) in the class of analytic functions is proved.

5 Particular Case

5.1 Evaluation of the Interval of Existence of a Solution

Theorem 1 provides local solvability of the Cauchy problem in the class of analytic functions. However, it does not allow to evaluate the interval of convergence...
of the series. In this section, this complex and substantive problem is investigated for a particular case. Consider the Cauchy problem

\[ w'' + \frac{1}{\sigma}(w')^2 + \frac{1}{\alpha}w = 0, \]
\[ w|_{\xi=0} = 0, \quad w'|_{\xi=0} = -\sigma. \]  

(21)

Construction of the solution of (2) is reduced to the problem (21), when \( \nu = 0 \) and the heat front has the form \( f(\tau) = \ln(C_1\tau + C_2)^{\alpha} \).

We construct the solution of (21) in the form of a power series

\[ w(\xi) = \sum_{n=0}^{+\infty} a_n \xi^n, \quad a_n \overset{\text{def}}{=} \frac{w^{(n)}(\xi)}{n!} \bigg|_{\xi=0}. \]  

(22)

In this case \( a_0 \equiv 0, \ a_1 \equiv -\sigma \). The remaining coefficients of the series (22) are determined from the recurrence formula

\[ a_{n+1} = \frac{1}{(\sigma n + 1)(n + 1)} \left[ \sum_{k=0}^{n-2} \left( k + 1 + \frac{n-k}{\sigma} \right) (k+2)a_{k+2}a_{n-k} + \frac{a_n}{\alpha} \right], \quad n \in \mathbb{N}. \]  

(23)

**Proposition 3.** Power series (22) is convergent if \( |\xi| \leq |\alpha|\sigma, \ \sigma \geq 1 \).

The proof is cumbersome and is not given here. However, the idea of the proof is simple and consists in the construction of the estimates for (23) with two well-known inequalities

\[ \sum_{k=1}^{n-3} \frac{1}{k+1} \leq \ln \left( \frac{2n-1}{5} \right), \quad \sum_{k=1}^{n-3} \frac{1}{(k+1)^2} < \frac{\pi^2}{6} - \frac{5}{4}. \]

Proposition 3 allows us to specify the area of existence of analytical solutions of the Cauchy problem (21). This result means that the analytical solution exists and is unique in the segment \( \xi \in [-|\alpha|\sigma, |\alpha|\sigma] \). Let us see what conclusions this fact leads to for the original problem.

Let us recall that in the present case

\[ u(\tau, \rho) = \frac{\alpha C_1}{C_1\tau + C_2}w(\xi), \quad \xi = \rho - \ln(C_1\tau + C_2)^{\alpha}. \]

We assume that the heat wave starts from the origin. For this purpose the heat wave front \( f(\tau) = \ln(C_1\tau + C_2)^{\alpha} \) must satisfy the condition \( f|_{\tau=0} = 0 \). Thus, \( C_2 = 1 \), therefore, \( f(\tau) = \ln(C_1\tau + 1)^{\alpha} \). Since \( \tau \geq 0 \), then \( f(\tau) \) is analytical for \( 0 \leq \tau \leq 1/C_1, \ C_1 > 0 \).

It should be noted that depending on the sign of the parameter \( \alpha \) the heat wave may have two directions of motion. Let \( \alpha > 0 \), then \( u(\tau, \rho) \geq 0 \) if and only if \( \xi \leq 0 \). Since we are interested in the analytical solution, we assume \( -\alpha\sigma \leq \xi \leq 0 \). In this case, the heat wave moves to the right, and the area of the existence of an analytic solution is \( 0 \leq \tau \leq 1/C_1, \ 0 \leq \rho \leq \alpha \ln 2 \). In the case of \( \alpha < 0 \) heat wave moves to the left.
Remark 3. The constraint $\sigma \geq 1$, violates the generality, however, is it physically motivated, because in filtration problems $\sigma$ is a measure of the gas adiabatic, which is known [15] to be greater than one.

Remark 4. For the coefficients $a_{n+1}$ of (22) we have

$$a_{n+1} = \frac{(-1)^n}{(n+1)!} \alpha^n (\sigma + 1)^n \prod_{m=2}^{n} (m\sigma + 1) \left\lceil \frac{n}{m} \right\rceil$$

where $[x] \overset{\text{def}}{=} \max \{ n \in \mathbb{Z} \mid n \leq x \}$, $b_1 \equiv -1$, and the remaining coefficients $b_{n+1}$ are determined from the recurrence formula

$$b_{n+1} = \frac{1}{\sigma} \sum_{k=0}^{n-2} \binom{n}{k} \left( \sigma + \frac{n-k}{k+1} \right) \prod_{m=2}^{n-1} (m\sigma + 1) \left\lceil \frac{n-1}{m} \right\rceil \left\lceil \frac{k+1}{m} \right\rceil \left\lceil \frac{n-k-1}{m} \right\rceil b_{k+2} b_{n-k-1} - \prod_{m=1}^{n-1} (m\sigma + 1) \left\lceil \frac{n-1}{m} \right\rceil b_n \ , \ n \in \mathbb{N} .$$

Note that $b_{n+1} \in \mathbb{Z}[\sigma]$, and

$$\text{deg}(b_{n+1}) = \sum_{k=2}^{n-1} \left\lfloor \frac{n}{k} \right\rfloor .$$

For example,

$$b_2 = 1 \ , \ b_3 = 1 \ , \ b_4 = 3\sigma + 5 \ , \ b_5 = 36\sigma^3 + 132\sigma^2 + 143\sigma + 41 \ , \ b_6 = 360\sigma^4 + 1824\sigma^3 + 3203\sigma^2 + 2232\sigma + 469 \ , \ ...$$

The leading coefficients of the polynomials $b_{n+1}$ are calculated according to the formula

$$\frac{(n-1)!}{2^{n-1}} \prod_{m=2}^{n} m \left\lfloor \frac{n}{m} \right\rfloor .$$

Using the representation (24) it can be assumed that the interval of convergence of the series (22) is $|\xi| < 2|\alpha|N(\sigma)$, where $N(\sigma) \sim \sigma$.

5.2 Numerical Research

Finally, we present the results of numerical research of the problem (21). Using the fourth order Runge-Kutta method in increments of $h = 10^{-4}$ the numerical solution of problem (21) is constructed. Calculations show that the solution $w(\xi)$ has a singular point: in the case of $\alpha > 0$ it can’t be extended to the left of some point $-\xi_0$ (fig. 1 (a)), and in the case $\alpha < 0$ it can’t be extended to the right of the point $\xi_0$ (fig. 1 (b)).
Fig. 1. The behavior of the numerical solution of the Cauchy problem (21): (a) if $\alpha > 0$, (b) if $\alpha < 0$.

In table 1 we present calculations, illustrating the behavior of the studied numerical solutions of the Cauchy problem near the point $\xi_0$ for some values of the parameters $\alpha$ and $\sigma$. Here $\xi_*$ is some point close to $\xi_0$.

From the results of numerical calculations which are presented in table 1 it can be assumed that the position of the singular point $\xi_0$ on the $\xi$–axis is defined as $\xi_0 = 2\alpha N(\sigma)$, where $N(\sigma) \sim \sigma \to +\infty$.

The presented in this section results are easy to interpret in terms of the original problem (2), (3). In this case, we have a heat wave (in assumption that its movement starts from the origin and $\alpha > 0$) with the front of $f(\tau) = \ln(C_1\tau + 1)^\alpha$.

The behavior of this wave is shown schematically in figure 2. It should be noted here that in this case we observe an effect of the heat wave separation.

6 Conclusion

The authors obtained exact heat wave type solutions of the nonlinear heat equation (2), satisfying the boundary condition (3). The procedure for constructing
Table 1. Numerical calculations

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\xi$</th>
<th>$w(\xi)$</th>
<th>$w'(\xi)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.68609</td>
<td>9.53136 · 10$^{-5}$</td>
<td>1.16480</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1.88372 · 10$^6$</td>
<td>7.43198 · 10$^4$</td>
</tr>
</tbody>
</table>

these solutions is reduced to the Cauchy problem for nonlinear ordinary differential equations of second order with a singularity. We establish the solvability of the obtained problems in the class of analytic functions (theorem 1).

Unlike solutions in form of power series [8–11], obtained exact solutions have several advantages. For example, it is possible to get comprehensive information on the properties of the heat waves. In this paper we have obtained an estimates for the area of analyticity (proposition 3) of heat wave type solution with front $f(\tau) = \ln(C_1\tau + 1)^\alpha$. It’s behavior have been studied by numerical methods.

Note that the heat wave type solutions of the nonlinear heat equation are important both from a theoretical point of view and in connection with applications. For instance, heat waves propagating with a finite rate, can be used to describe high-temperature processes in plasma [4].

Acknowledgments. The study is partially supported by the Russian Foundation for Basic Research, projects 16-01-00608, 16-31-00291.

References


About Verification of Calculation Methods of the Shock Waves

Valentin Kuropatenko\textsuperscript{1,2} and Elena Shestakovskaya\textsuperscript{2}

\textsuperscript{1} Russian Federal Nuclear Center-Zababakhin All-Russia Research Institute of Technical Physics, 456770, Snezhinsk, Russian Federation
\textsuperscript{2} South Ural State University (National Research University), 454080, Chelyabinsk, Russian Federation
v.f.kuropatenko@yandex.ru, leshest@list.ru

Abstract. Mathematical modeling is now a key tool of research into dynamic processes in continuum mechanics. Each particular problem is solved with already existing or newly developed models and methods, whose properties are determined from a priori study into stability, approximation, monotonicity etc within linear approaches. The accuracy of difference schemes is mainly evaluated through comparison between calculated results and reference solutions. The paper discusses some problems which have analytical solutions. These are shock convergence, the dynamic compression of a gas sphere, and some problems with stationary shocks.

Keywords: shock, analytical solution, ideal gas, spherical symmetry, stationary shock

1 Introduction

The properties of difference schemes that approximate the conservation laws are often evaluated with a priori methods which involve studies into stability, approximation, monotonicity, conservatism, distraction etc. It should however be noted that most of these methods are developed for acoustic approximations and simple equations of state. In continuum mechanics, the properties of a mathematical model may notably differ from what linear theory predicts due to nonlinearities induced by real-world equations of state, shocks, plasticity and other material properties.

Linear theory loses its rigor when applied to nonlinear equations. The importance of the convergence theorem [1] is strongly exaggerated because it is still proved for linear equations and not for nonlinear ones, and real calculations are done not for vanishing but finite $\Delta x$ and $\Delta t$. A very vivid discussion of stability and convergence criteria and their rigidity can be found in [2].

The calculation of shock waves strong discontinuities in all material properties requires special attention. On the shock surface the conservation laws take the form of nonlinear algebraic equations which relate the values of quantities across the shock. Entropy jumps as all the other functions do. This is the fundamental difference between a shock and a wave where the quantities vary continuously. Flows with shocks are often
simulated with homogeneous methods which treat the strong shock as a layer of a finite width comparable with the cell size. This ability of difference schemes is called distraction [3]. Since the states across the shock are related by the Hugoniot, there must be a mechanism which allows entropy to grow in the shock distraction region. Physical viscosity and heat conduction in continuum mechanics equations cannot give a distraction width of several cell sizes. The proposal by Neumann and Richtmyer to use a mathematical 'viscosity' [4] seems to resolve the problem. Their method has gained wide acceptance. Pseudo-viscosity is taken in different forms linear, quadratic, or linear-quadratic [4-7]. But the method does not ensure convergence to the exact solution if the form of pseudo-viscosity changes. So, the author of [8] gives an example where different schemes with different viscosities converge to different solutions in the limit. In [9], there is an example of spherical convergence where energy dissipation defined by pseudo-viscosity is shown to be several times higher than energy dissipation due to plasticity.

Advantages and disadvantages of a mathematical model can be seen from comparison between its predictions and analytical solutions. The paper discusses some problems which have exact analytical solutions. It gives their statements (initial and boundary conditions, equations of state, and physical parameters) and solutions in the form of formulas or tables. In order to verify performance of a difference scheme, one needs to solve the problem numerically and compare the result with the exact solution.

The problems are broken into two groups for stationary and non-stationary shocks. In problems with stationary shocks, derivatives in the exact solution are zero everywhere beyond the distraction zone and hence approximation errors are also zero. In the distraction zone, the derivatives and approximation errors reach high values. Here the strong shock is smeared over several cells where entropy differs. These problems help verify real shock distraction, monotonicity, entropy variation, and the dependence of calculated results on the relation between steps in space and time, and on cell size (the number of points in the mesh). All these properties reveal themselves differently for strong and weak shocks. Shock strength is characterized by the difference between pressures behind and before the shock.

In problems with non-stationary shocks, the derivatives and derivative-dependent approximation errors are high beyond the distraction zone. This group includes shock convergence and spherical shell convergence problems. In the last problem, the boundary conditions and released energy are adjusted so as to keep material density constant despite large pressure and velocity gradients.

Some analytical solutions are used for comparison with results obtained with the difference schemes which are based on the energy dissipation method described in [10-13].

2 Stationary shock

Consider a material with parameters $P_0$, $V_0$, $E_0$, $U_0$ which do not change with time. At a time $t_0$ its left boundary instantaneously starts moving at a constant positive velocity, producing a shock wave which propagates into the material. The equations
\[ \begin{align*}
WV_+ + U_+ &= WV_- + U_-, \\
WU_+ - P_+ &= WU_- - P_-, \\
W\varepsilon_+ - P_+U_+ &= W\varepsilon_- - P_-U_-
\end{align*} \]

relate the material states \( P_- = P_0, \ V_- = V_0, \ E_- = E_0, \ U_- = U_0 \) with the state after the discontinuity \( P = P_+ , \ V = V_+, \ E = E_+, \ U = U_+ \) before and behind the shock, and the shock velocity \( W \). The number of quantities is larger than the number of equations (1) - (3) + equation of state. To solve this system of equations requires that one of the quantities be taken as parameter. Let it be \( U \). Take the equation of state (EOS) for ideal gas in the form

\[ PV = (\gamma - 1)E \]

and transform equations (1) - (3) to the dependence of \( P \) on \( U \) and other zero-subscripted quantities

\[ P = P_0 + \frac{\gamma + 1}{4} \frac{\Delta U^2}{V_0} + \sqrt{\frac{(\gamma + 1)\Delta U^2}{V_0}} + \frac{\gamma P_0}{V_0} \Delta U^2, \]

where \( \Delta U = U - U_0 \). From equations (1) - (3) and (5) we find \( P, W, V \) and \( E \):

\[ W = (P - P_0)/(U - U_0), \ V = V_0 - (U - U_0)/W, \ E = E_0 + 0.5(P + P_0)(V - V_0). \]

For condense matter, a simple equation of state has the form

\[ P = (n - 1)\rho E + C_{0k}^2(\rho - \rho_{0k}), \]

where \( \rho = 1/V \) - is material density, \( \rho_{0k} \) - is its density at a point with coordinates \( T = 0, P = 0 \) and \( C_{0k} \) - is sound velocity at this point. For EOS (7) equations (1) - (3) transform to the Hugoniot equation

\[ P = P_0 + \frac{n + 1}{4} \rho_0 \Delta U^2 + \sqrt{\left(\frac{\gamma + 1}{4} \rho_0 \Delta U^2\right)^2 + \rho_0 \Delta U^2 (nP_0 + \rho_{0k} C_{0k}^2)}. \]

Equations (6) for \( W, V, \) and \( E \) remain the same.

For convenience, we treat all quantities in equation (5) - (8) as dimensionless. That is why both in gas and in condense matter, the initial dimensionless density is unity. Conversion to density in g/cm³ is done through multiplying by the constant used for conversion to dimensionless density. All the other quantities are treated similarly.

**Problem 1.** Strong shock in monatomic gas. At \( t = 0 \), a region \( 0 \leq x_0 \leq 1 \) is occupied with gas described by EOS (4) with parameters \( \gamma = 5/3, \ P_0 = 0, \ \rho_0 = 1, \ E_0 = 0, \ U_0 = 0 \). Here \( x_0 \) - is the Eulerian coordinate at \( t = 0 \). In Lagrangian difference schemes \( x_0 \) is a Lagrangian coordinate. At \( t > 0, \ U = 1 \) is specified on the left boundary \((x_0 = 0)\) and \( U = 0 \) is on the right one \((x_0 = 1)\).

The quantities behind the shock front and front velocity \( W \) are determined from equations (1) - (3): \( \rho = 4, \ E = 0.5, \ P = 4/3, \ W = 4/3 \). At \( t = 0.375 \), the shock is at a point \( x_0 = 0.5 \) and the analytical solution is determined by
\[ P = 1.33333, \quad \rho = 4.0, \quad E = 0.5, \quad U = 1.0, \quad \text{for} \ x_0 \leq 0.5, \ \text{and} \\
\text{\quad} P = 0, \quad \rho = 1.0, \quad E = 0, \quad U = 0, \quad \text{for} \ x_0 > 0.5 \]

Figures 1 and 2 depict \( P(x_0) \) and \( U(x_0) \) at \( t = 0.575 \). The solid lines show analytical solutions and the marked ones show calculations with the difference scheme from [12]. The calculations were done with a uniform mesh of \( N = 100 \) points in \( x_0 \) and Courant number 0.5.

**Problem 2.** Strong shock in monatomic gas described by EOS (4) and \( \gamma = 1.25 \) (ethylene). All the other parameters are the same as in Problem 1: \( \rho_0 = 1, \ P_0 = 0, \ E_0 = 0, \ U_0 = 0 \). The boundary conditions are also the same: \( U(x_0 = 0, t) = 1, \ U(x_0 = 1, t) = 0. \)

The quantities behind the shock front and front velocity are determined from equations (1) - (3): \( \rho = 9, \ E = 0.5, \ P = 1.125, \ W = 1.125. \) At \( t = 0.44444 \) the shock is at a point \( x_0 = 0.5 \), and the analytical solution is determined by

\[
\text{\quad} P = 1.125, \quad \rho = 9.0, \quad E = 0.5, \quad U = 1 \quad \text{for} \ x_0 \leq 0.5, \ \text{and} \\
\text{\quad} P = 0, \quad \rho = 1.0, \quad E = 0, \quad U = 0, \quad \text{for} \ x_0 > 0.5. \\
\]

Figures 3 and 4 depict \( P(x_0) \) and \( U(x_0) \) at \( t = 0.44444 \). The solid lines show analytical solutions and the marked ones show calculations with the difference scheme from [12]. The calculations were done with a uniform mesh of \( N = 100 \) points in \( x_0 \) and Courant number 0.5.

**Problem 3.** The weak shock wave in monatomic gas. At \( t = 0 \), a region \( 0 \leq x_0 \leq 1 \) is occupied with monatomic ideal gas described by EOS (4) with parameters \( \gamma = 5/3, \rho_0 = 1, \ P_0 = 1, \rho_0 = 1, \ E_0 = 1.5, \ U_0 = 0 \). At \( t > 0 \), \( U = 0.5 \) is specified on the left boundary \( (x_0 = 0) \) and \( U = 0 \) is on the right one \( (x_0 = 1) \). Behind the shock, \( \rho = 1.428573, \ E = 1.925, \ P = 1.833333, \) and \( W = 1.666666. \) The analytical solution at \( t = 0.3 \) is determined by
Fig. 3. Problem 2. $P(x_0)$ at $t = 0.44444$

Fig. 4. Problem 2. $U(x_0)$ at $t = 0.44444$

\[ P = 1.83333, \quad \rho = 1.42857, \quad E = 1.925, \quad U = 0.5 \quad \text{for} \quad x_0 \leq 0.5, \text{and} \]
\[ P = 1.0, \quad \rho = 1.0, \quad E = 1.50, \quad U = 0 \quad \text{for} \quad x_0 > 0.5. \]

Figures 5 and 6 depict $P(x_0)$ and $U(x_0)$ at $t = 0.3$. The solid lines show analytical solutions and the marked ones show calculations with the difference scheme from [13]. The calculations were done with a uniform mesh of $N = 100$ points in $x_0$ and Courant number 0.5.

Fig. 5. Problem 3. $P(x_0)$ at $t = 0.3$

Fig. 6. Problem 3. $U(x_0)$ at $t = 0.3$
Problem 4. Strong shock in condense matter. At \( t = 0 \), a region \( 0 \leq x_0 \leq 1 \) is occupied with condense matter described by EOS (7) which at \( \rho_0 k = 1 \) and \( C_0 k = 1 \) takes the form

\[
P = (n - 1) \rho E + \rho - 1,
\]

At \( t = 0 \), the parameters are \( \rho_0 = 1, E_0 = 0, P_0 = 0, U_0 = 0, n = 3 \). For \( t > 0 \), \( U = 2 \) is specified on the left boundary \( (x_0 = 0) \) and \( U = 0 \) is on the right one \( (x_0 = 1) \). The equation for \( P \) is obtained from (8):

\[
P = \frac{n + 1}{4} \rho_0 U^2 + \sqrt{\left(\frac{n + 1}{4} \rho_0 U^2\right)^2 + \rho_0 U^2}.
\]

![Fig. 7. Problem 4. \( P(x_0) \) at \( t = 0.118034 \)](image)

![Fig. 8. Problem 4. \( U(x_0) \) at \( t = 0.118034 \)](image)

The analytical solution at \( t = 0.118034 \) is determined by

\[
\begin{align*}
P &= 8.47214, & \rho &= 1.89443, & E &= 2.0, & U &= 2.0 \quad \text{for} \ x \leq 0.5, \text{and} \\
P &= 0, & \rho &= 1.0, & E &= 0, & U &= 0 \quad \text{for} \ x > 0.5.
\end{align*}
\]

Figures 7 and 8 depict \( P(x_0) \) and \( U(x_0) \) at \( t = 0.118034 \). The solid lines show analytical solutions and the marked ones show calculations with the difference scheme from [13]. The calculations were done with a uniform mesh of \( N = 100 \) points in \( x_0 \) and Courant number 0.5.

Problem 5. Weak shock in condense matter. At \( t = 0 \), a region \( 0 \leq x_0 \leq 1 \) is occupied with a material described by EOS (9) with parameters: \( \rho_0 = 1, E_0 = 2, P_0 = 4, U_0 = 0, n = 3 \). For \( t > 0 \), \( U = 1 \) on the left boundary and \( U = 0 \) on the right one.

The analytical solution at \( t = 0.105448 \) is determined by

\[
\begin{align*}
P &= 8.74166, & \rho &= 1.26726, & E &= 3.34359, & U &= 1.0 \quad \text{for} \ x_0 \leq 0.5, \text{and}
\end{align*}
\]
Figures 9 and 10 depict $P(x_0)$ and $U(x_0)$ at $t = 0.105448$. The solid lines show analytical solutions and the marked ones show calculations with the difference scheme from [13]. The calculations were done with a uniform mesh of $N = 100$ points in $x_0$ and Courant number 0.5.

3 The motion of a spherical layer of compressible ideal fluid

The problem of bubble collapse in fluid, or spherical shell convergence, arises in connection with cavitation corrosion of propellers. Solutions to the problem can be found in [14-16]. The full continuum mechanics model for 1D spherically symmetric flow of ideal compressible continua includes mass conservation, motion and internal energy equations:

\[
\frac{\partial V}{\partial t} - 4\pi \frac{\partial r^2 U}{\partial M} = 0, \tag{10}
\]

\[
\frac{\partial U}{\partial t} + 4\pi r^2 \frac{\partial P}{\partial M} = 0, \tag{11}
\]

\[
\frac{\partial E}{\partial t} + P \frac{\partial V}{\partial t} = 0. \tag{12}
\]

Equations (10)–(12) are written in Lagrangian coordinates. Here the partial derivatives with respect to time are substantial derivatives.

There no incompressible matter in nature. Mechanics simply considers a wide class of flows where density remains constant in time. Density constancy is often understood as incompressibility, i.e., $\beta_s = 0$ and $C^2 = \infty$. It is not true. The property of flow is not the property of matter.
As a rule [16], the models of ‘incompressible’ fluid do not include the energy conservation law and the equation of state. This makes them internally contradictive. As follows from (12), in ideal compressible fluid, \( E = \text{const} \) at \( V = \text{const} \). But in this case from the equation of state \( P = P(V, E) \) we obtain that \( P = \text{const} \), too. So, on the one hand, \( P \) varies, and on the other hand, \( P \) is constant. This contradiction can be removed if assume that fluid is not adiabatic, i.e., there is a source of energy in it. Then equation (12) is written as

\[
\frac{\partial E}{\partial t} = -P \frac{\partial V}{\partial t} + \frac{\partial q}{\partial t}.
\] (13)

Equations (10), (11) and (13) allow solutions where density is constant. To keep the density of fluid constant requires energy. As follows from the theory of equations of state [17], in fluid, thermal pressure and energy, \( P_T \) and \( E_T \), are related by the equation \( P_T V = \Gamma(V) E_T \). Hereafter for \( V = \text{const} \) the equation is taken in the form

\[
PV_0 = \Gamma E,
\] (14)

where \( \Gamma = \text{const} \), \( P = P_T \), \( E = E_T \). Since \( P(t, M) \) is a solution to equations (11) and (12), then the dependence \( E(t, M) \) which follows from (13) is quite specific in each flow. It is defined by the necessity of meeting the condition \( V = \text{const} \).

Here we limit ourselves to flows where specific volume is independent of either \( M \), or \( t \), i.e., \( V = \text{const} \). Also, we assume that fluid is compressible, i.e., its compressibility \( \beta_S \) is nonzero.

For \( V = V_0 \), the equation that relates the Eulearian coordinate \( r \) and the Lagrangian coordinate \( dM = (4\pi r^2/V_0) \, dr \) can be integrated from \( M = 0 \) at \( r = r_B \) to an arbitrary finite \( M \)

\[
r = \left( r_B^3 + 3V_0M/4\pi \right)^{1/3}.
\] (15)

Here \( r_B \) is the time dependent coordinate of the bubble boundary. At \( V = V_0 \), equation (10) has the solution

\[
r^2U = f(t).
\] (16)

Since \( f \) is independent of \( M \), equations (15) and (16) are valid for arbitrary \( M \). On the bubble boundary where \( M = 0 \), equation (16) takes the form

\[
r_B^2U_B = f(t), \quad U = U_B r_B^2.
\] (17)

where \( U_B \) is bubble boundary velocity.

Find \( U_B \) from (17) and substitute in the bubble boundary motion equation

\[
\left( \frac{dr_B}{dt} \right)_M = U_B.
\] (18)

Integrate (18) together with (15), to obtain the dependence of \( r_B \) on \( t \).
It is seen from (17) and (19) that the motion of the bubble boundary is completely defined by \( f(t) \). If \( f(t) < 0 \), then \( U_B < 0 \) too, i.e., the bubble collapses. The boundary convergence time \( t_f \) is found from (19) at \( r_B = 0 \)

\[
\frac{r_B^3}{B_0} + \int_{t_0}^{t_f} 3f(t) \, dt = 0.
\]  

(20)

Following [16], consider shell convergence with zero pressure on the inner boundary and \( f(t) = U_{B0}r_{B0}^2 = \text{const.} \) At time \( t_0 \), the radius and velocity of the inner boundary, \( r_{B0} \) and \( U_{B0} < 0 \), are defined. In accord with (17) at time \( t_0 \), velocity depends on radius \( U = U_{B0}(r_{B0}/r)^2 \).

Let all quantities on the outer boundary be subscripted "a". Assume that the shell mass is equal to \( M_a \). The coordinate of the outer boundary, \( r_a \), relate to that of the inner boundary \( r_B \) as \( r_a = (r_B^3 + b)^{1/3} \), where \( b = \frac{3V_0}{4\pi} M_a = r_{a0}^3 - r_{B0}^3 \). Pressure and velocity on the outer boundary are

\[
U_a = U_{B0} \left( \frac{r_{a0}}{r_{B0}} \right)^3 \frac{t - t_0}{t_f - t_0}^{-2/3},
\]  

(22)

\[
P_a = \frac{U_{B0}^2}{2V_0} \left( \frac{t_f - t}{t_f - t_0} \right)^{-4/3} - \left( \frac{r_{a0}}{r_{B0}} \right)^3 - \left( \frac{t - t_0}{t_f - t_0} \right)^{-4/3}. \]  

(23)

The values \( U_{a0} \) and \( P_{a0} \) are found from (22) and (23) at \( t = t_0 \). In Lagrangian coordinates the pressure, velocity and released energy are defined by

\[
P = \frac{U_{B0}^2}{2V_0} \left( \frac{t_f - t}{t_f - t_0} \right)^{-4/3} - \left( \frac{t_f - t}{t_f - t_0} + \frac{3V_0 M_a}{4\pi r_{B0}^3} \right)^{-4/3},
\]  

(24)

\[
U = U_{B0} \left( \frac{t_f - t}{t_f - t_0} + \frac{3V_0 M_a}{4\pi r_{B0}^3} \right)^{-2/3},
\]

\[
\frac{\partial q}{\partial t} = \frac{2U_{B0}^3}{Fr_{B0}} \left( \frac{t_f - t}{t_f - t_0} + \frac{3V_0 M_a}{4\pi r_{B0}^3} \right)^{-7/3} - \left( \frac{t_f - t}{t_f - t_0} \right)^{-7/3}.
\]

As the reference problem we consider the motion of 10% shell.

**Problem 6.** The motion of a 10%-shell. At \( t_0 = 0 \), \( r_{B0} = 1 \), \( r_{a0} = 1.1 \), \( V_0 = 1 \), and \( M_a = 1.38649 \). The velocity of the inner boundary is \( U_{B0} = 1 \). The EOS of shell material with parameters \( \rho_{0k} = 1 \), \( C_{0k} = 1 \) and \( \Gamma = 2 \) has the form \( P = 2\rho E + \rho - 1 \). At the initial time, pressure, specific internal energy and velocity in the shell are defined by

\[
\]


\[ P(M) = 1 - \left(1 + \frac{3M}{4\pi}\right)^{-4/3}, \quad E(M) = \frac{1}{4} \left(1 - \left(1 + \frac{3}{4\pi}\right)^{-4/3}\right), \]

\[ U = U_B \left(\frac{4\pi r_B^3 \rho_0}{3M}\right)^{2/3}. \]

Boundary conditions: at \( t \geq 0 \), \( P_B = 0 \), \( M_B = 0 \), \( U_a = -\left(1, 1^3 - 3t\right)^{-2/3} \) at \( t \geq 0 \), \( M = M_a \). From (20) we found \( t_f = 1/3 \). At \( t \geq 0 \), energy release as a function of \( t \) and \( M \) is defined by

\[ \frac{dq}{dt} = \left(1 - 3t + \frac{3M}{4\pi}\right)^{-7/3} - (1 - 3t)^{-7/3}. \]

For \( t \geq 0 \) and \( M_a \geq M \geq 0 \) the solution has the form

\[ E(t, M) = \frac{1}{4} \left[(1 - 3t)^{-4/3} - \left(1 - 3t + \frac{3M}{4\pi}\right)^{-4/3}\right], \quad P = 2E, \quad \rho = 1, \]

\[ U(t, M) = -\left(1 - 3t + \frac{3M}{4\pi}\right)^{-2/3}. \]

Figures 11 and 12 show pressure and velocity as functions of \( m = M/M_a \) at \( t_1 = 0, 30 \), \( t_2 = 0, 32 \).
4 Shock in a gas sphere

In different years, there were published a number of papers [18-22] with self-similar solutions to shock convergence in infinite ideal gas. Shock convergence in a gas sphere of finite radius is considered in [23,24]. At \( t = t_0 \), pressure in gas is \( P_0 = 0 \), density \( \rho_0 = \text{const} \), velocity \( U_0 = 0 \), and specific internal energy \( E_0 = 0 \). The boundary of the sphere is at a point \((r_0, t_0)\). Velocity on the boundary is \( U_{g0} < 0 \). In other words, velocity jumps on the boundary, producing a shock wave which moves into the sphere. At the time when the shock converges, \( t_f \), its coordinate \( r_w \) is zero. The equation of motion which satisfies all these conditions is

\[
\frac{r_w}{r_0} = (t_f - t) - t_f - t_0
\]

for \( n > 0 \) Its differentiation gives shock velocity

\[
D = D_0 \left(\frac{t_f - t}{t_f - t_0}\right)^{n-1}
\]

where

\[
D_0 = -r_0 n (t_f - t_0).
\]

Flow parameters are defined by

\[
\frac{\partial \rho}{\partial t} + U \frac{\partial \rho}{\partial r} + \rho \frac{\partial U}{\partial r} + \frac{2 \rho U}{r} = 0,
\]

\[
\frac{\partial U}{\partial t} + U \frac{\partial U}{\partial r} + \frac{1}{\rho} \frac{\partial P}{\partial r} = 0,
\]

\[
\frac{\partial P}{\partial t} + U \frac{\partial P}{\partial r} + \gamma P \left( \frac{\partial U}{\partial r} + \frac{2 U}{r} \right) = 0.
\]

For solving the problem we change from the variables \( t \) and \( r \) to variables \( t \) and \( \xi(t, r) \). The function \( \xi(t, r) \) is taken such as to remain constant on the shock. Its simplest form reads as

\[
\xi = \frac{r}{r_0} \left(\frac{t_f - t_0}{t_f - t}\right)^n.
\]

Now express \( P, \rho \) and \( U \) as functions of time multiplied by functions of \( \xi \):

\[
P = \alpha_p(t) \Pi(\xi), \quad \rho = \alpha_\rho(t) \delta(\xi), \quad U = \alpha_u(t) M(\xi).
\]

Choose \( \Pi(\xi), \delta(\xi), M(\xi) \) such that to allow them at \( \xi = 1 \) take the values

\[
\delta_w = \frac{\gamma + 1}{\gamma - 1}, \quad M_w = \frac{2}{\gamma + 1}, \quad \Pi_w = \frac{2}{\gamma + 1}.
\]

With these \( \delta_w, \Pi_w, M_w \) the function \( \alpha_\rho, \alpha_u \) and \( \alpha_p \) take the forms
\[ \alpha_p = \rho_0, \quad \alpha_u = D_0 \left( \frac{t_f - t_0}{t_f - t} \right)^{1-n}, \quad \alpha_p = \rho_0 D_0^2 \left( \frac{t_f - t_0}{t_f - t} \right)^{2(1-n)}. \quad (32) \]

After appropriate manipulation for conversion to the functions \( \Pi, \delta, M \) and variables \( t, \xi \), we obtain equations for functions which only depend on \( \xi \):

\[
(M - \xi) \delta' + \delta M' + \frac{2M\delta}{\xi} = 0, \quad \frac{n-1}{n} \delta M + M' (M - \xi) + \Pi' = 0, \quad (33)
\]

\[
\frac{2(n-1)}{n} \Pi + \Pi' (M - \xi) + \gamma \Pi M' + \frac{2\gamma \Pi I}{\xi} = 0. \quad (34)
\]

For \( M', \delta', \Pi' \), these equations give a system of linear homogeneous equations. If its determinant

\[
Z = (M - \xi) \left( \gamma \Pi - \delta (M - \xi)^2 \right) \quad (35)
\]

is nonzero, the system has a unique solution. At the point \( \xi_* \) where \( Z = 0 \), the matrix of coefficients and the augmented matrix of coefficients should be considered. At this point their ranks are identical and equal to 2, and all third-order minors are zero, hence the system of equations (33) and (34) has a unique solution. It is easy to show that with the zero third-order minors we come to

\[
(n-1) \xi_* (2 (M_* - \xi_*) - \gamma M_* (M_* - \xi_*)) = 0. \quad (36)
\]

The value of \( n \) is found from the condition that equation (36) holds simultaneously with

\[
Z_* = (M_* - \xi_*) \left( \gamma \Pi_* - \delta_* (M_* - \xi_*)^2 \right) = 0. \quad (37)
\]

From equations (36) and (37) we find the appropriate values of \( n \) for each \( \gamma \). This solution was used to evaluate the accuracy of some shock calculation methods.

**Problem 7.** A cold gas sphere of radius \( r_{g0} = 1 \) with parameters \( P_0 = 0, \rho_0 = 1, U_0 = 0, U_{g0} = 1, \gamma = 5/3 \). The boundary condition is defined through reverse transition from \( t, \xi \) and \( \Pi, \delta, M \) to \( t, M \) and \( P, \rho, U \). Pressure and boundary velocity as functions of time are presented in Table 1. Pressure, density and velocity profiles at \( t = 0.4, 0.45, 0.5 \) (marked 1, 2, 3) are shown in Figs. 13-15. The solid lines show the analytical solution derived in this work, the lines with circles show calculations by the VOLNA code [25] with no shock smearing, and the dashed lines show VOLNA calculations with shock smearing. The calculations were done on a uniform mesh of 100 points in \( r \) at \( t = t_0 \). In Fig.14, entropy traces are seen in the dashed density profiles, which are a result of shock smearing on the boundary. Figure 16 depicts \( M(\xi), (\xi) \) and \( \delta(\xi) \) for \( 1 \leq \xi \leq 5 \).

**Table 1.**

<table>
<thead>
<tr>
<th>No.</th>
<th>t</th>
<th>U</th>
<th>P</th>
<th>No.</th>
<th>t</th>
<th>U</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.4</td>
<td>-1.000000</td>
<td>1.333333</td>
<td>17</td>
<td>0.30</td>
<td>-1.045146</td>
<td>2.445445</td>
</tr>
</tbody>
</table>
Table 1. Continuation.

<table>
<thead>
<tr>
<th>No.</th>
<th>t</th>
<th>U</th>
<th>P</th>
<th>No.</th>
<th>t</th>
<th>U</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.04</td>
<td>-1.008791</td>
<td>1.415094</td>
<td>18</td>
<td>0.31</td>
<td>-1.044921</td>
<td>2.517083</td>
</tr>
<tr>
<td>3</td>
<td>0.07</td>
<td>-1.015161</td>
<td>1.484529</td>
<td>19</td>
<td>0.32</td>
<td>-1.04473</td>
<td>2.592849</td>
</tr>
<tr>
<td>4</td>
<td>0.10</td>
<td>-1.021256</td>
<td>1.562232</td>
<td>20</td>
<td>0.33</td>
<td>-1.043787</td>
<td>2.673060</td>
</tr>
<tr>
<td>5</td>
<td>0.13</td>
<td>-1.026982</td>
<td>1.649664</td>
<td>21</td>
<td>0.34</td>
<td>-1.042842</td>
<td>2.758063</td>
</tr>
<tr>
<td>6</td>
<td>0.16</td>
<td>-1.032224</td>
<td>1.748622</td>
<td>22</td>
<td>0.35</td>
<td>-1.041619</td>
<td>2.848236</td>
</tr>
<tr>
<td>7</td>
<td>0.18</td>
<td>-1.035380</td>
<td>1.822074</td>
<td>23</td>
<td>0.36</td>
<td>-1.040097</td>
<td>2.943996</td>
</tr>
<tr>
<td>8</td>
<td>0.20</td>
<td>-1.038211</td>
<td>1.902413</td>
<td>24</td>
<td>0.37</td>
<td>-1.038251</td>
<td>3.045798</td>
</tr>
<tr>
<td>9</td>
<td>0.22</td>
<td>-1.040657</td>
<td>1.990554</td>
<td>25</td>
<td>0.38</td>
<td>-1.036058</td>
<td>3.154141</td>
</tr>
<tr>
<td>10</td>
<td>0.23</td>
<td>-1.041717</td>
<td>2.037877</td>
<td>26</td>
<td>0.39</td>
<td>-1.033488</td>
<td>3.269573</td>
</tr>
<tr>
<td>11</td>
<td>0.24</td>
<td>-1.042655</td>
<td>2.087565</td>
<td>27</td>
<td>0.40</td>
<td>-1.030514</td>
<td>3.392698</td>
</tr>
<tr>
<td>12</td>
<td>0.25</td>
<td>-1.043463</td>
<td>2.139778</td>
<td>28</td>
<td>0.42</td>
<td>-1.023220</td>
<td>3.664748</td>
</tr>
<tr>
<td>13</td>
<td>0.26</td>
<td>-1.044129</td>
<td>2.194694</td>
<td>29</td>
<td>0.45</td>
<td>-1.008354</td>
<td>4.149469</td>
</tr>
<tr>
<td>14</td>
<td>0.27</td>
<td>-1.044643</td>
<td>2.252502</td>
<td>30</td>
<td>0.50</td>
<td>-0.969947</td>
<td>5.233492</td>
</tr>
<tr>
<td>15</td>
<td>0.28</td>
<td>-1.044992</td>
<td>2.313410</td>
<td>31</td>
<td>0.55</td>
<td>-0.899013</td>
<td>6.144230</td>
</tr>
<tr>
<td>16</td>
<td>0.29</td>
<td>-1.045165</td>
<td>2.377642</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 13. Problem 7. pressure profiles at times 0.4 (1), 0.45 (2), 0.5 (3)

Fig. 14. Problem 7. density profiles at times 0.4 (1), 0.45 (2), 0.5 (3)
Fig. 15. Problem 7. velocity profiles at times 0.4 (1), 0.45 (2), 0.5 (3)

Fig. 16. Problem 7. profiles of dimensionless $M(\xi)$, $\Pi(\xi)$, $\delta(\xi)$

References


Applications of Regularly Varying Functions in Study of Cosmological Parameters

Žarko Mijajlović

Faculty of Mathematics, University of Belgrade
Studentski trg 16, 11000 Belgrade, Serbia
State University of Novi Pazar
Vuka Karadžića bb, 36300 Novi Pazar, Serbia
zarkom@matf.bg.ac.rs

Abstract. Most of the cosmological parameters, such as the scale factor $a(t)$, the energy density $\rho(t)$ and the pressure of the material in the universe $p(t)$ under usual circumstances satisfy asymptotically the power law. On the other hand the quantities that satisfy the power law are best modeled by regularly varying functions. The aim of this paper is to apply the theory of regularly varying functions to study Friedmann equations and their solutions which are in fact mentioned cosmological parameters. In particular we shall consider possible formulas for cosmological parameters of the dual universe.

Keywords: regular variation, cosmological parameter, Friedmann equations, dual universe.

1 Introduction

Theory of regularly varying functions was started by J. Karamata in [13] and sometimes it is also called Karamata theory of regular variation. Many other mathematicians further developed this theory, see Bingham et al. [2] and Seneta [22]. At the present time this theory is used in many areas, including asymptotic analysis of functions, Tauberian theory, probability, differential equations and analytic number theory. There were several attempts to use this theory in cosmology, particularly in the study of asymptotic behavior of cosmological parameters, eg Mijajlovic et al., [18], [19], but also by Molchanov [20] and Stern [23]. Barrow in [3] and Barrow and Show in [4] used a theory of Hardy and Fowler which preceded the theory of regular variation in studies of asymptotic behavior of solutions to the Einstein equations describing expanding universes.

In [18] we introduced a new constant $\Gamma$ related to Friedmann equations. Determining the values of $\Gamma$ one can obtain the asymptotical behavior of the solutions, i.e. of the expansion scale factor $a(t)$ and terms $\rho(t)$ and $p(t)$. It appears that the instance $\Gamma < 1/4$ is appropriate for both cases, a spatially flat and an open universe, and gives a sufficient and necessary condition for the solutions to be regularly varying. In describing cosmological parameters we used the theory of regularly varying solutions of linear second order differential equations, see
Marić [16], which gives necessary and sufficient conditions for the existence of such solutions. From the theory of regular variation it follows that the solutions under usual assumptions include a multiplicative term which is a slowly varying function. We also present a set of formulas that can be assigned to cosmological parameters of the dual universe. These formulas correspond to the second fundamental solution of the acceleration equation.

We shall shortly review definitions and properties of regularly varying functions. In particular we shall use some theorems on regularly varying solutions of the second order differential equation

$$\ddot{y} + f(t)y = 0, \quad f(t) \text{ is continuous on } [\alpha, \infty].$$

(1)

The notion of regular variation is related to the power law distribution represented by the following relationship between some quantities $F$ and $t$:

$$F(t) = t^r(\alpha + o(1)), \quad \alpha, r \in \mathbb{R}.$$  

(2)

This definition of power law is in a close relation to the notion of a slowly varying function. A real positive continuous function $1/L(t)$ defined for $x > x_0$ which satisfies

$$\frac{L(\lambda t)}{L(t)} \to 1 \quad \text{as} \quad t \to \infty, \quad \text{for each real } \lambda > 0.$$  

(3)

is called a slowly varying (SV) function.

**Definition 1.** A function $F(t)$ is said to satisfy a generalized power law if

$$F(t) = t^rL(t)$$

(4)

where $L(t)$ is a slowly varying function and $r$ is a real constant.

Logarithmic function $\ln(x)$ and iterated logarithmic functions $\ln(\ldots \ln(x) \ldots)$ are examples of slowly varying functions. More complicated examples are provided in [2], [22] and [16].

A positive continuous function $F$ defined for $t > t_0$, is a regularly varying (RV) function of an index $r$, if and only if it satisfies

$$\frac{F(\lambda t)}{F(t)} \to \lambda^r \quad \text{as} \quad t \to \infty, \quad \text{for each } \lambda > 0.$$  

(5)

It immediately follows that a regularly varying function $F(t)$ has the form (4). Therefore $F(t)$ is regularly varying if and only if it satisfies the generalized power law. By $\mathcal{R}_\alpha$ we denote the class of regularly varying functions of an index $\alpha$. Hence $\mathcal{R}_0$ is the class of all slowly varying functions. By $\mathcal{Z}_0$ we shall denote the class of zero functions at $\infty$, i.e. $\varepsilon \in \mathcal{Z}_0$ if and only if $\lim_{t \to +\infty} \varepsilon(t) = 0$. The following theorem [13] describes the fundamental property of these functions.

---

1. Continuing the works of G.H. Hardy, J.L. Littlewood and E. Landau, Karamata [13] originally defined and studied this notion for continuous functions. Later this theory was extended to measurable functions. Due to physical constraints, we are dealing here only with continuous functions.
Theorem 1. (Representation theorem) $L \in \mathcal{R}_0$ if and only if there are measurable functions $h(x)$, $\varepsilon \in \mathcal{Z}_0$ and $b \in \mathbb{R}$ so that
\[
L(x) = h(x)e^{\int_b^x \frac{\varepsilon(t)}{t} dt}, \quad x \geq b,
\]
and $h(x) \to h_0$ as $x \to \infty$, $h_0$ is a positive constant.

If $h(x)$ is a constant function, then $L(x)$ is called normalized. Let $\mathcal{N}$ denote the class of normalized slowly varying functions. The next fact on $\mathcal{N}$-functions will be useful for our later discussion. If $L \in \mathcal{N}$ and there is $\ddot{L}$, then $\varepsilon$ in (6) has the first order derivative $\dot{\varepsilon}$. This follows from the identity $\varepsilon(t) = t\dot{L}(t)/L(t)$.

For our study of Friedmann equations we need the next result [9], [16] on solutions of equation (1). This theorem gives necessary and sufficient conditions for equation $\ddot{y} + f(t)y = 0$ to have regularly varying solutions.

Theorem 2. (Howard-Marić) Let $-\infty < \Gamma < 1/4$, and let $\alpha_1 < \alpha_2$ be two roots of the equation
\[
x^2 - x + \Gamma = 0.
\]
Further let $L_i$, $i=1,2$ denote two normalized slowly varying functions. Then there are two linearly independent regularly varying solutions of $\ddot{y} + f(t)y = 0$ of the form
\[
y_i(t) = t^{\alpha_i} L_i(t), \quad i = 1, 2,
\]
if and only if $\lim_{x \to \infty} x \int_x^\infty f(t) dt = \Gamma$. Moreover, $L_2(t) \sim \frac{1}{(1 - 2\alpha_1)L_1(t)}$. \hfill \square

The limit integral in the theorem is not easy to compute. As $\lim_{t \to \infty} t^2 f(t) = \Gamma$ implies $\lim_{x \to \infty} x \int_x^\infty f(t) dt = \Gamma$, we see that
\[
\lim_{t \to \infty} t^2 f(t) = \Gamma
\]
gives a useful sufficient condition for the existence of regular solutions of the equation $\ddot{y} + f(t)y = 0$ as described in the previous theorem.

2 Cosmological parameters

Cosmological parameters are usually defined as some general physical quantities related to the Universe. Such approach for Lambda cold dark matter model of Universe ($\Lambda$CDM model) is presented in the standard literature, for example in [12], [10] and [21]. Here our approach is somewhat formalistic. For cosmological parameters we take primarily solutions of Fiedmann equations [7]:
\[
\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3} \rho - \frac{kc^2}{a^2}, \quad \text{Friedmann equation},
\]
\[
\frac{\ddot{a}}{a} = -\frac{4\pi G}{3} \left(\rho + \frac{3p}{c^2}\right), \quad \text{Acceleration equation},
\]
\[
\dot{\rho} + 3\frac{\dot{a}}{a} \left(\rho + \frac{p}{c^2}\right) = 0, \quad \text{Fluid equation}.
\]
and any functions derived from these solutions. Therefore, the scale factor $a(t)$, the energy density $ρ(t)$ and the pressure of the material in the universe $p(t)$ are basic cosmological parameters. We remind that Friedmann equations are derived from the Einstein field equations. These three equations are not independent. For example, the fluid equation can be inferred from the other two equations. Therefore, for solving of these system which consists essentially of two equations and three unknowns some additional condition is needed. Usually equation of state $p = wρc^2$ is assumed.

Suppose $\bar{a}(t)$, $\bar{ρ}(t)$ and $\bar{p}(t)$ are some definite solutions of Friedmann equations. Taking

$$μ(t) = \frac{4πG}{3}t^2\left(\bar{ρ}(t) + \frac{3\bar{p}(t)}{c^2}\right),$$

we see that then $\bar{a}(t)$ is a solution of the second order linear differential equation:

$$\ddot{a} + \frac{μ(t)}{t^2}a = 0. \quad (11)$$

It is easy to check that in fact any solution $b(t)$ of (11) jointly with $\bar{ρ}(t)$ and $\bar{p}(t)$ is a solution of all three Friedmann equations. Therefore, in search for RV solutions of the acceleration equation and so of the Friedmann equations, we can use the Howard-Marić theorem 2. We just did this in our previous work [18]. We review some results from there we need in our further discussion.

First observe that the integral limit in the Howard-Marić theorem for the equation (11) is given by:

$$M(μ) = \lim_{x→∞} x \int_x^∞ \frac{μ(t)}{t^2}dt. \quad (12)$$

The functions for which this integral limit converges define so called Marić class of functions $ℳ$. Then $M$ is a real functional defined on $ℳ$. Also, in view of (9) we have

$$\text{If } \lim_{t→∞} μ(t) = Γ \text{ then } M(μ) = Γ. \quad (13)$$

We note that the opposite of (13) does not hold, see [18], [19]. There RV solutions of Friedmann equations are found (theorems 3.2 and 3.3) and appropriate cosmological parameters for non-oscillatory universe are determined. Assuming that the integral limit $M(μ)$ is convergent, say $M(μ) = Γ$, there is proved:

- If $Γ < 1/4$ then the universe is non-oscillatory.
- The converse is almost true, namely, if the universe is non-oscillatory then $Γ ≤ 1/4$.
- If $Γ < 1/4$ and in some special cases for $Γ = 1/4$, the scale factor $a(t)$, a solution of Friedmann equations, is an RV function.

In view of these properties it is justified to call the constant $Γ$ a threshold constant. Assume that $α$ is a root of the polynomial $x^2 - x + Γ$. Then

$$Γ = α(1 - α) \quad (14)$$
In this case cosmological parameters are represented as follows:

**Scale factor** $a(t)$: $a(t) = t^\alpha L(t)$, $\alpha \neq 0$ and $L$ is an RV function. In other words, $a(t)$ is a regularly varying function of an index $\alpha$.

**Hubble parameter** $H(t) = \dot{a}(t)/a(t)$:

$$H(t) = \frac{\alpha}{t} + \frac{\varepsilon}{t}, \quad \varepsilon \in \mathbb{Z}_0.$$  \hspace{1cm} (15)

**Deceleration parameter** $q(t)$:

$$q(t) = \frac{\mu(t)}{\alpha^2} (1 + \eta) = \frac{1}{\alpha} - \frac{t \dot{\varepsilon}}{\alpha^2} (1 + \eta) + \tau, \quad \varepsilon, \eta, \tau \in \mathbb{Z}_0.$$  \hspace{1cm} (16)

Assuming that the scale factor $a(t)$ satisfies the generalized power law one can introduce a new constant $w$. It will appear that $w$ is in fact the equation of state parameter. Assuming $a(t) = t^\alpha L(t)$, $L \in \mathcal{N}$ and $\alpha \neq 0$, we define $w$ by

$$w \equiv w_\alpha = \frac{2}{3\alpha} - 1.$$  \hspace{1cm} (17)

Then the cosmological parameters can be put in the following form:

$$\alpha = \frac{2}{3(1 + w)}, \quad a(t) = a_0 t^{\frac{2}{3(1 + w)}} L(t)$$  
$$H(t) \sim \frac{2}{3(1 + w)t}, \quad M(q) = \frac{1 + 3w}{2}.$$  \hspace{1cm} (18)

Formulas for the exponent $\alpha$ and the Hubble parameter $H(t)$ are widely found in the literature. Formulas for $a(t)$ and $q(t)$ are also reduced to the standard form if $L(t)$ and $q(t)$ are constant at infinity, or if the equation of state $p = w\rho c^2$ is assumed, or $\lim_{t \to \infty} t \dot{\varepsilon}(t) = 0$. We did not assumed in derivation of (18) any of these assumptions. In fact, we found asymptotics for solutions of Friedmann equations only assuming $M(\mu) = \Gamma < 1/4$, and in certain cases for $\Gamma = 1/4$. As far as we know, it is implicitly widely assumed that the limit $\lim_{t \to \infty} \mu(t)$ exists and is finite, what is much stronger assumption than that the integral limit $M(\mu)$ is convergent.

We note, if basic cosmological parameters satisfy power law under definition (17), then for the universe with the flat curvature the following weak form of the equation of state holds:

There are functions $\xi, \zeta \in \mathbb{Z}_0$ such that $p = \dot{w}\rho c^2$, where $\dot{w}(t) = w - t\dot{\xi} + \zeta$.

Therefore, if $t\dot{\xi} \to 0$ as $t \to \infty$, then $\dot{w}(t) \approx w$, what leads to $p = w\rho c^2$, the standard equation of state and classical asymptotics for cosmological parameters. In [18] is also found

$$M(\mu) = \Gamma = \frac{2}{9} \cdot \frac{1 + 3w}{(1 + w)^2}.$$  \hspace{1cm} (19)

As the Friedmann equations are invariant under translation transformation, the above formulas also hold for the expanding universe with the cosmological constant $\Lambda$. 

376
3 Cosmological parameters for dual universe

In the previous section we have seen that \( q(t) \) and \( p(t) \) may vary, depending on the limit of the hidden parameter \( t \dot{\varepsilon}(t) \) as \( t \to \infty \). As indicated in [19] one can speculate that this variation is an effect of the existence of the dual universe. We remind that one of the concepts of string theory and hence M-theory is that the big bang was a collision between two membranes. The outcome was the creation of two universes, one in the surface of each membrane. Using the Large Hadron Collider (LHC) located in CERN, some data are collected that might lead to the conclusion that the parallel universe exist. Specifically, if the LHC detects the presence of miniature black holes at certain energy levels, then it is believed [6] that these would be the fingerprints of multiple universes. Collected data are still analyzed.

We will not enter here into a full discussion on the existence of the multiverse. But if the existence the parallel universe is assumed, we can explicitly find a set of formulas that might represent cosmological parameters of the dual universe. We obtain them using the second fundamental solution \( L_2(t) \) in Howard - Marić theorem applied to the acceleration equation. To find the second fundamental solution and therefore the dual set of these formulas we take the second root \( \beta = 1 - \alpha \) of the quadratic equation \( x^2 - x + \Gamma = 0 \) appearing in this theorem. To avoid singularities, we assume \( \alpha, \beta \neq 0 \). Now we use \( \beta \) instead of \( \alpha \) for the index of RV solution \( a(t) \) - scale factor and for determination of other constants and cosmological parameters. As in (17) we introduce \( w_\beta = \frac{2}{3\beta} - 1 \). Then we have the following symmetric identity for the equation of state parameters:

\[
 w_\alpha + w_\beta + 3w_\alpha w_\beta = 1
\]  

(20)

For our universe we have \( w = w_\alpha \), while for the dual universe the corresponding equation of state parameter is \( w_\beta \). Then the dual formulas are obtained by replacing \( \alpha \) with \( \beta \) and \( w_\alpha \) with \( w_\beta \) in (15), (16) and (18). If one wants to give any physical meaning to the so obtained dual set of functions, it is rather natural to interpret them as the cosmological parameters of the dual universe.

As we shall see these two universes are isomorphic in the sense that there is an isomorphism which maps cosmological parameters into their dual forms. In this derivation we shall use some elements of the Galois theory. For the basics of this theory the reader may consult for example [11].

Our assumption that \( \Gamma < \frac{1}{4} \) and that the solutions \( \alpha \) and \( \beta \) of the equation (7) differ, say \( \alpha < \beta \), introduces the following kind of symmetry. Let \( F = \mathbb{R}(t, \Gamma) \) be the extension algebraic field where \( \mathbb{R} \) is the field of real numbers and \( t \) and \( \Gamma \) are letters (variables). It is easy to see that for such \( \Gamma \) the polynomial \( x^2 - x + \Gamma \) is irreducible over the field \( F \). Hence, the Galois group \( G \) of the equation (7) is of the order 2 and has a nontrivial automorphism \( \sigma \). Let \( \alpha \) and \( \beta \) be the roots of the polynomial \( x^2 - x + \Gamma \). Then \( \sigma(\alpha) = \beta \) and \( \sigma(\beta) = \alpha \). Further, let \( \Gamma = \frac{2}{9} \cdot \frac{1+3w}{(1+w)} \cdot (1+3w)^2 \) where \( w \) is a parameter. Then we can take \( \alpha = \frac{2}{3(1+w)} \) and \( \beta = \frac{1+3w}{3(1+w)} \). Let \( w_\alpha \equiv w \) and \( w_\beta \equiv \frac{1-w}{1+3w} \). Then \( \sigma(w_\alpha) = w_\beta \) since \( w_\alpha \) and \( w_\beta \) are rational expressions respectively in \( \alpha \) and \( \beta \). Further, the time \( t \) and the
constant \( \Gamma \) are invariant under \( \sigma \) i.e. \( \sigma(t) = t \) and \( \sigma(\Gamma) = \Gamma \) since \( t \) and \( \Gamma \) are the elements of the ground field \( F \). The cosmological parameters (15), (16) and (18) are rational expressions of \( w \) so if \( P_{\alpha} \) is the corresponding parameter to the solution \( \alpha \), then \( \sigma(P_{\alpha}) = P_{\beta} \). For example, for the Hubble parameters we have \( \sigma(H_{\alpha}) = H_{\beta} \). Hence, not only solutions (isomorphic via \( \sigma \)) come into the pairs but the sets of all cosmological parameters come as well. At this point one may speculate about two dual universes having the same time \( t \) and the constant \( \Gamma \) and the conjugated parameters \( w_{\alpha} \) and \( w_{\beta} \) connected by the relation (20).

Of course, there is a question what are the values of the constants appearing in cosmological parameters, for example of \( w = w_{\alpha} \). Most results in the literature see e.g. [25], are consistent with the \( w = -1 \) cosmological constant case. Results from experimental cosmology, such as the Baryon Oscillation Spectroscopic Survey (BOSS) of Luminous Red Galaxies (LRGs) in the Sloan Digital Sky Survey (SDSS) are consistent with \( w = -1 \), the dark energy equation of state, [1]. However, the value \( w = -1 \) yields singularity in (18). For such \( w \) there is no corresponding \( \alpha \) neither \( \Gamma \). Equation of state is \( p = -\rho c^2 \) and then by fluid equation we have \( \dot{\rho} = 0 \), i.e \( \rho \) is a constant. This case corresponds to the cosmological constant, so \( \rho = \rho_\Lambda = \frac{\Lambda}{8\pi G} \). In the absences of \( \alpha \) and \( \beta \) for dual \( w_{\beta} \) of \( w = w_{\alpha} \) we may take (20) for defining relation . Putting \( w_{\alpha} = -1 \) in this identity we obtain \( w_{\beta} = -1 \). Hence, dual universe is also equipped with a cosmological constant and its expansion is governed with the dark energy.

The other values of \( w \) are also considered. For example if \( w = 1/3 \) then \( \alpha = \beta = 1/2, \Gamma = 1/4 \) and in this case Howard-Marić theorem cannot be applied since functions \( L_1(t) \) and \( L_2(t) \) from this theorem are not fundamental solutions. But there is a variant of this theorem appropriate for this case [16], and applying it one can show that \( a(t) \) is regularly varying of index \( \frac{1}{2} \) if and only if \( w \sim \frac{1}{3} \) as \( t \to \infty \), i.e. \( p \sim \frac{1}{3} c^2 \rho \) holds asymptotically. This is the second classic cosmological solution. For more details one can consult [18].

4 Conclusion

A detailed analysis of Friedmann equations and cosmological parameters from the point of view of regular variation is presented. The central role in this analysis has the acceleration equation since it can be considered as a linear second order differential equation and that the theory of regularly varying solutions of such equations is well developed [16]. We introduced in a formal way certain constants such as the threshold constant \( \Gamma \) and the equation of state parameter \( w \). Both constants have the fundamental role in describing asymptotics of cosmological parameters and evolution of the Universe. We also inferred formulas that might represent the cosmological parameters of the dual universe.

References

Mathematical Modeling of Artificial Mitral Heart Valve

Hranislav Miloshevich¹, Yury Zakharov²,*, Yurii Shokin³, Dmitry Dolgov², and Irene Grigorieva²

¹ University of Pristina, Serbia, ² Kemerovo State University, Kemerovo, Russia ³ Institute of Computational Technologies SB RAS, Novosibirsk, Russia

*zaxarovyn@yandex.ru
http://www.kemsu.ru

Abstract. The research shows the mathematical model, describing the dynamics of the artificial aortic heart valve and the model of blood thrombus moving in large vessels, as well as the method of numerical calculation of these models. There are represented numerical modelling results of the tricuspid valve operation and the blood thrombus moving in large vessels.

Keywords: Mathematical modeling, artificial mitral heart valve, aneurysm, immersed boundary method

1 Introduction

The research of heart and blood-vascular system diseases is a task that has extremely high socio-economic importance and a long history. The knowledge of the human cardiovascular system is actual as never before and during the last few decades the methods of mathematical modeling are widely used in their accumulation.

In the world 80-90% of non-traumatic subarachnoid hemorrhages are occurred due to the bursting of intracranial aneurysm [10]. The aneurysm rupture leads to neurological deficits related to brain tissue damage or even death. Approximately 50% of patients with aortic aneurysm bursting die before the hospitalization [11]. Each year, about 250 thousand heart valve restoration or surrogation operations are carried out in the world and the number of these operations and their necessity is increasing year by year [9].

The cardiovascular system is extremely complex. The heart is a complex multi-valve muscular organ. Blood vessels are multilayered structure that substantially differs from each other, depending on the type of vessel and its position. Blood has a heterogeneous structure including formed elements. Some authors consider the blood as an incompressible viscous Newtonian fluid [13] including formed blood elements [8]. Sometimes the blood circulation is presented as non-Newtonian fluid flow [3].
This research presents a mathematical model of blood circulation in large blood vessels, suitable for modelling of vascular malformations and operation of artificial heart valves. Artificial heart valves are one of the most complex prosthetic devices in cardio surgery. They enable you to deal effectively with diseases and injuries of natural valves, but their operation time is much less than a human life, which means that the patients need to be re-prosthesis every few years. Mechanical valves have high reliability and durability, but can lead to a serious deformation of the blood circulation, the formation of the blood cell clot and as a result to the thrombus formation. Biomaterial valves don’t have this drawback, but they are less durable, and their production is a difficult technical problem, which is not completely resolved at the present time.

Also in this research we investigate the mechanism of the formation of blood aneurysms in the large blood vessels. The research [10], [16] shows that the formation of blood aneurysms is caused by swelling one of the layers of its wall, of intima, which is the most subtle and least durable. It defects the vessel shape and the average blood circulation in it.

The mathematical models, written in the form of quite complex differential equations that can be solved by various numerical methods, are often used as a modeling tool of the blood circulation. One of the most commonly used methods is the finite element method (FEM) [6], [17]. FEM is widely tested on the problems of elasticity theory and hydrodynamics. Furthermore there is a sufficient number of sets, realizing this method. FEM enables to take into account the complex shape of the solution field for the deflection of the blood vessel walls and valve leaflets, but the need to take into account the interaction between the fluid and flexible walls leads to a constant reconfiguration of the analysis grid to comply the changing shape of the object, consequently the finite element method has a high temporal and spatial complexity. There is another approach to solving problem of blood circulation in vessels based on the applying of the lattice Boltzmann method [15], [1]. This approach uses the methods of statistical mechanics, numerically solving Boltzmann discrete equation, the direction of the fluid flow is defined in the lattice sites and the fluid flow is possible only in the directions of the lattice. Another common method for researching of the hemodynamics, vascular structures and heart valves problems is the method of the immersed boundary. The immersed boundary method is a relatively new technique that was proposed for the modeling of the heart valves operation [7], [12]. It enables you to simulate the deformation of arbitrarily thin valve leaflets.

In this research we consider the blood circulation in the large elastic blood vessels and artificial heart valve as a three-dimensional variable flow of the incompressible fluid with variable density and viscosity [4], [5]. The resulting system of the differential equations with appropriate boundary and initial conditions describes the artificial tricuspid aortic valve operation, the formation and growth of the blood aneurysm in large blood vessels. The immersed boundary method is used in conjunction with the method of nets to solve the differential problems.
2 Problem definition

The blood consists of plasma and measured formed elements, which account for about 45% of the total volume. The size of the formed elements is small compared to the size of a large vessel: for example, aortic diameter is $3 \times 10^{-2} \text{m}$, and the diameter of an erythrocyte is $6 \times 10^{-9} \text{m}$. The research [14] shows that the blood plasma behaves like a Newtonian fluid, which makes the blood an incompressible inhomogeneous two-component liquid with variable viscosity and density. The blood vessel walls and valve leaflets we will consider as fluid-tight surfaces that have certain rigidity. Also the blood vessel walls and valve leaflets can be deformed under the influence of the fluid pressure.

The source of the fluid flow in vessels and heart valves is the pressure generated by contraction of the heart muscle. Therefore, we will describe its flow in the blood vessels and heart valves by Navier-Stokes equations system with variable density and viscosity and the pressure drop is set up at the input-output of the solution field, depending on the time:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} = -\frac{1}{\rho} \nabla p + \nabla \sigma + \mathbf{f}$$  

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0$$

with the initial conditions and the boundary conditions

$$\begin{align*}
\mathbf{u}(\bar{x}, 0) &= \mathbf{u}_0 \\
\mathbf{u}|_{\Gamma_1, \Gamma_4} &= \mathbf{u}_b \\
v, w|_{\Gamma_2, \Gamma_3} &= 0
\end{align*}$$

$$p_{\Gamma_2} = p_{\text{in}} \quad p_{\Gamma_3} = p_{\text{out}}$$

$x(x, y, z) \in \tilde{\Omega}$ are the points of the solution field, $\mathbf{u}(u, v, w)$ is the vector of the velocity field, $p(x, t)$ is the pressure, $\sigma = \mu(\nabla \mathbf{u} + (\nabla \mathbf{u})^T)$ is the viscous stress tensor, $\mu(x, t)$ is the fluid viscosity, $f(x, t)$ is the vector of the body forces. The fluid density is defined by the formula

$$\rho = c(\rho_2 - \rho_1) + \rho_1$$
and the viscosity is a function of the flow rate, its density and their gradients. In this research we use a simple linear relationship between fluid density and viscosity:

$$\mu = c(\mu_2 - \mu_1) + \mu_1$$  \hspace{1cm} (6)

Here, $\rho_1$, $\mu_1$ is the density and viscosity of the carrier fluid (blood plasma in this case), $\rho_2$, $\mu_2$ is the density and viscosity of the admixtures (formed elements), $c(x, t)$ is an admixtures concentration determined from the transfer equation

$$\frac{\partial c}{\partial t} + \mathbf{u} \cdot \nabla c = 0$$  \hspace{1cm} (7)

with the initial conditions

$$c(x, 0) = c_b(x), \ x \in \hat{Q}$$  \hspace{1cm} (8)

and the boundary conditions on the boundary of the inflow:

$$c(\bar{x}, t)|_{r_2} = c_s(\bar{x}, t)$$  \hspace{1cm} (9)

In this approach, the displacement of the vessel walls and valve leaflets occurs under the influence of fluid flow in consideration of the walls rigidity, but the tissues tension and deformation forces give the contribution to the sum of the forces in the motion equations

$$\mathbf{f}(x, t) = \int_{\Gamma} \mathbf{F}(\bar{q}, t) \cdot \delta(x - X(\bar{q}, t)) \ d\bar{q}$$  \hspace{1cm} (10)

$$\frac{\partial X(\bar{q}, t)}{\partial t} = \int_{\Omega} \mathbf{u}(x, t) \cdot \delta(x - X(\bar{q}, t)) \ dx$$  \hspace{1cm} (11)

is the vector of the body forces, is the deformation resistance force. The equations (10) and (11) use an integral transformation that uses three-dimensional Dirac function for the transition from Lagrangian to Eulerian coordinates. The equation (11) shows that the immersed boundary moves according to the local velocity of the ambient fluid, which is an analog of the adhesion condition:

$$\frac{\partial X(q, r, s, t)}{\partial t} = \mathbf{u}(X(q, r, s, t), t)$$  \hspace{1cm} (12)

Adhesion condition is used to determine the movement of the immersed elastic boundary bounded by fluid flow. The formula from [7] is used to describe the deformation forces

$$F = \frac{\partial}{\partial s} k \tau + \frac{\partial^2}{\partial s^2} \left(k_s \left( \frac{\partial^2 X^0}{\partial s^2} - \frac{\partial^2 X}{\partial s^2} \right) \right),$$  \hspace{1cm} (13)

where $k$ is the elastic coefficient, $\tau = \frac{\partial X}{\partial s} / |\frac{\partial X}{\partial s}|$, $k_s$ is the flexural rigidity coefficient. If we assume that the Lagrangian coordinates $(q, r, s)$ are chosen in such a way that coordinate pair $(q, r)$ defines the single fiber for the fixture $s$, and the formula $X(q_0, r_0, s)$ defines the single fiber for the fixture $s$, and the formula $X(q, r, s)$ determines the parametric representation of the fiber $X^0(q, r, s) = X(q, r, s, t_0)$.
3 Solution method

In this research we will use immersed boundary method [12] to determine the movement of the valve leaflets and deformation of the vessel walls. In accordance with this method, we will calculate the fluid flow in the $\tilde{\Omega}$ parallelepiped, which includes $\Omega$ (see Fig. 1). The adhesion condition is required at the $\tilde{\Omega}$ boundaries. We will use a rectangular, actually non-uniform staggered grid $\tilde{\Omega}_h$ with $h_{xi}$, $h_{yj}$, $h_{zk}$ pitches and staggered nodes, where the pressure, the velocity divergence and the concentration are determined at the center of the socket and the components of the velocity vector and external forces are determined at the socket boundaries to determine the fluid flow. The splitting scheme on physical factors is used to solve the equations (1) - (4):

$$\frac{u^* - u^n}{\Delta t} = - (u^n \cdot \nabla) u^* - \frac{1}{\rho} \nabla \sigma + f^n$$ (14)

$$\rho \Delta p^{n+1} - \nabla \rho \cdot p^{n+1} = \frac{\rho^2 \nabla u^*}{\Delta t}$$ (15)

$$\frac{u^{n+1} - u^*}{\Delta t} = - \frac{1}{\rho} \Delta p^{n+1}$$ (16)

The numerical scheme consists of the following steps:

- The intermediate velocity field $u^*$ is determined according to the known velocity values from the preceding time layer. For this purpose the equation (14) is solved by the stabilizing correction method.
- $p^{n+1}$ is calculated from the formula (15) by the stabilizing correction method.
- The new velocity field $u^{n+1}$ is determined according to the explicit formulas (16).
- The new concentration values are calculated from the transfer equation (7).
- The new values of the density and viscosity are defined according to the formulas (5) and (6).
- The velocity values at the points of the immersed boundary are approximated.
- The displacement of the free boundary is calculated.
- The deformation resistant forces are determined and the projections of the body forces vector are calculated in the grid nodes.

4 Results

The results of the numerical modelling of the artificial heart valve "UniLine" operation as well as the formation and growth of the blood aneurysm in large blood vessels are presented in this paragraph. The calculations were performed for the case of constant and variable density and viscosity in non-dimensional variables.
4.1 The calculation of the artificial heart valve "UniLine" operation

The artificial heart valve "UniLine" (see Fig. 2) is a stented tricuspid biological prosthetic device wherein inanimate, especially treated biological tissues are fixed on the support frame (stent) covered with a synthetic fabric. The high-precision leaflets cutting by the laser system, which enables to avoid the collagen fibers separation on the edge of the cut [2] is used in the production of this valve. We placed the valve inside a circular cylinder with length of 1 and radius $r = 0.11$. The coefficient of the vessel walls elasticity is $k = 1 \times 10^3$, for the valve leaflets is $k = 5 \times 10^3$, the flexural rigidity coefficient is $k_s = 2 \times 10^3$. The pressure drop $p_{in} - p_{out}$ varies continuously from 0 to 6 at times. The pitch on the spatial grid is 0.01 and in time it is the same (0.01).

![Fig. 2. The artificial heart valve "UniLine".](image)

The Figure 3 shows the movement of the valve leaflets and fluid flow through it at periodic increasing and decreasing of the pressure drop.

![Fig. 3. The dynamics of the valve leaflets and the line of the certain particles. The direction of flow is indicated by the arrow. $t = 0, 0.7, 1.5$.](image)
As you can see in the Figure 3, the valve leaflets are opened when the pressure difference is changing, and then reset at pressure balancing.

Figure 4 by a dotted line shows a graph of fluid flow within the valve according to the time for the first three cycles. A leap fluid flow corresponds to each pulse, and the swings of the valve leaflets are reflected in the very slight oscillations of the graph when closing.

We use physiological conditions for the pressure presented at Figure 5. In the picture aorta pressure is input, ventricular pressure is the output one. Extension strength coefficient $k_s = 2 \cdot 10^3$, $k_b = 1 \cdot 10^2$ while valve opening and $k_b = 20 \cdot 10^3$ when valve closing. Viscosity is $\mu = 0.25 \cdot 10^{-2}$ Pa/s, density $\rho = 1 \cdot 10^3$ kg/m$^3$. Step on space is $h_x = h_y = h_z = 2 \cdot 10^{-3}$ m, step on time $t = 1 \cdot 10^{-4}$ s, vessel length $l = 0.1$ m, radius $r = 0.03$ m. Figure 4 also shows a comparison with the same calculation of the work [7]. The comparison shows that the rise and flow rate at the time of full valve opening match quite well. However, in our calculations, the valve closes a little more slowly, due to the differences in the parameters for rigidity. You can also note that in [7] the first flow peak is much smaller than the other. This is due to the fact that in this paper describe the initial conditions already strained leaflets, whereas in our calculation they rested without tension.

For the pressures analysis during "UniLine" valve operation we monitored their change at two points of this valve. The valve leaflets were moved under the influence of the fluid with constant, and then variable viscosity and density. These two points are marked in the Figure 6 and hereafter we will call them "active point" and the "mid-point". The "active point" is located on the one of the valve spindles at the point of the two adjacent leaflets fastening and the "mid-point" is located in the center of the leaflet.
**Fig. 5.** The graph of physiological pressure.

**Fig. 6.** The arrangement of the points spacing on the tissue annulus.
The Figure 7 shows graphs of the surface traction dependence on the time for these two points in the various parts of the valve for the cases of constant density and viscosity \((\rho_1 = \rho_2 = 1, \mu_1 = \mu_2 = 1 \times 10^{-2}, c=0\%)\) and variable density and viscosity \((\rho_1 = 1, \rho_2 = 2, \mu_1 = 1 \times 10^{-2}, \mu_2 = 2 \times 10^{-2})\) for the two concentration values \(c = 20\%\) and \(c = 40\%\).

![Graphs showing surface traction dependence on time](image)

**Fig. 7.** The dependence of the traction on the time for two points on the valve for the cases without admixtures (continuous line) and with admixtures where \(c=20\%\) (dashed line) and with admixtures where \(c = 40\%\) (dotted line).

### 4.2 The formation and growth of blood aneurysms in the large blood vessels

The presented in this research model can be successfully used to research the formation and growth of aneurysms in the large blood vessels. In this paragraph we will consider the growth process of the aneurysm in the large blood vessel under the periodically changing pressure, occurred during the blood circulation in the vessel. We considered the vessel of a constant cross section, the set of the cross-section centers is defined by the spline, the vessel radius is \(r = 0.11\) and the walls rigidity is of \(k = 2.5 \times 10^3\). It is assumed that the aneurysm grows in areas with high pressure on the walls therefore the vessel rigidity is decreased in areas with high pressure. The Figure 8 shows the shape of the vessel during the formation of the aneurysm and also it shows the particle paths.

When the blood aneurysm is formed, the blood clots, which can be washed out by the flow and can get into the mainstream of the vessel, are formed in the area of aneurysm. Therefore we take into account the importance to investigate how a fairly dense and viscous area of the fluid will be spread in the vessel, which has the narrowing (angiostenosis). Two stenosis scenarios are under consideration: 1) insignificant vasoconstriction (15%), that happens gradually, 2) significant vasoconstriction (72.5%) at a small vessel area. The vessel wall stiffness is \(4.5 \times 10^3\), that is quite high, clot density is \(\rho_2 = 2\), viscosity is \(\mu_2 = 3 \times 10^2\). In the latter case vessel wall stiffness decreases to \(1 \times 10^3\) due to linear law (before stenosis) and increases to the initial value at the stenosis area. Fig. 9, 10 show...
modeling results. In case the vessel has insignificant stenosis its part with strong concentration and, as a result, high density and viscosity is washed out by fluid flow (from vessel axis), the clot passes vessel constriction area by changing its form and leaves the computational domain. In case the stenosis is more significant the scenario takes more time. The clot reaches vessel constriction area, but only some part of it can passes through stenosis, the clot is washed out gradually and it never passes the stenosis. This situation leads to slight deformation of vessel walls before stenosis. The clot does not completely block the vessel and there is still flow inside, due to the insignificant density and viscosity of the clot, though its concentration changes more slowly compared with the first case.

5 Conclusion

We develop the model of the aneurysm and artificial heart valve operation, considering the blood circulation with variable density and viscosity. The immersed boundary method used for the realization of this model enables to get the movement patterns of the valve leaflets for various shapes, to analyze the stress rates, occurred during their movement and to predict the formation and growth of blood aneurysm in the blood vessels.

The work was carried out with support of state task of Ministry of Science and Education, project 1.630.2014/K.
1) \( t = 2.5 \)  

2) \( t = 10.5 \)  

3) \( t = 19.5 \)  

4) \( t = 34.0 \)

Fig. 9. The passage of the blood clot through the insignificant vessel angiostenosis, the coloring of the flow area is according to the concentration.

References

Fig. 10. The passage of the blood clot through the significant vessel angiostenosis, the coloring of the flow area is according to the concentration.

Forecast of Thermokarst Lakes Dynamics in Permafrost Based on Geo-Simulation Modeling and Remote Sensing Data

Yury Polishchuk$^{1,2,*}$ and Vladimir Polishchuk$^{3,4}$

$^1$ Ugra Research Institute of Information Technologies, Mira, 151, 628011, Khanty-Mansiysk, Russia
$^2$ Institute of Petroleum Chemistry of SB RAS, Akademichesky, 4, 634021, Tomsk, Russia
$^3$ Institute of Monitoring of Climatic and Ecological Systems of SB RAS, Akademichesky, 8/3, 634021, Tomsk, Russia
$^4$ Tomsk Politechnic University, Lenina Str., 30, 634004, Tomsk, Russia

* yupolishchuk@gmail.com

Abstract. The information technology of forecast of the dynamics of lake’s fields in permafrost was developed using geo-simulation approach to modeling. The model properties were determined on base of analysis of data on climatic changes and satellite images. The program complex for predicting geocryological changes under global warming using computer experiments with the model is presented. A new forecast assessments of changes of thermokarst processes on the territory of West-Siberian permafrost were obtained on base of computer experiments model. It is shown that the gradual increase in temperature to $2 - 3 \, ^{\circ}C$ in future decades will cause a reduction in the area of thermokarst lakes, what is an indicator of the continuing degradation of permafrost by the end of the century. The developed information technology can be used for solving problems of predicting the dynamics of greenhouse gas emissions from thermokarst ponds in Western Siberia under the impact of global warming.

Keywords: modeling, geo-simulation, forecast, permafrost, thermokarst lakes, information technology, satellite images, climate changes

1 Introduction

The global warming leads to an increase of accidents on pipelines and other oil and gas facilities due to lower strength of permafrost. Moreover permafrost, as a repository of carbon conserved in the vast frozen peat bogs of northern Eurasia and America may cause even more warming if greenhouse gas release. The development of measures to reduce the damage of oil and gas companies require the use of forward-looking assessments of the dynamics of thermokarst processes. Due to considerable bogging and inaccessibility of the territory of Western Siberia, where is located the main oil and gas complex of Russia the research of
these processes is impossible without the use of remote sensing data. The information technology of modeling and forecasting of the dynamics of thermokarst lakes fields was developed using satellite images for the period 1973-2010 years. An important issue is the creation of a mathematical model. The complexity of modeling the field of thermokarst lakes has led to the need to use geo-simulation approach to modeling of natural objects with a spatial structure.

We know that global warming leads to the northern territories to the growth of accidents on pipelines and other oil and gas facilities. Reducing the strength of permafrost caused by the acceleration of thermokarst processes under the influence of warming, is accompanied by the growth of economic and environmental damages on the domestic oil and gas companies, located in the permafrost zone. The development of measures to reduce the damage of oil and gas companies is impossible without predictive estimates of the dynamics of the morphological structure of thermokarst lake fields, the preparation of which requires the use of mathematical modeling of the dynamics of thermokarst processes on the territory of permafrost under a global warming.

Due to the high degree of waterlogging and remote areas of permafrost, these studies both in our country and abroad are carried out with the use of remote sensing data. At the same time as the most suitable geomorphological indicator of changes in permafrost is used thermokarst lakes which is well seen on satellite images [1]. So important is the question of forecasting the dynamics of fields of thermokarst lakes.

Thermokarst processes can be modeled mathematically with analytical models based on theory. Matt has shown [2] that such models are efficient for studying processes in a single thermokarst lake, but unsuitable for modeling spatio-temporal changes of thermokarst lake fields. The methods of mathematical morphology developed by Victorov [3] are of great importance here as they are designed to use analytical models for territory dynamics modeling. These methods enable long-term dynamics of the state of a territory to be predicted; but they are not designed for the study of the spatio-temporal changeability of fields of thermokarst lakes. A new approach to modeling the dynamics of spatio-temporal systems proposed in [4,5] allowed to develop geo-simulation model of thermokarst lake’s dynamics. This model allows to take into account important regularity of dynamics of thermokarst fields - reducing areas of thermokarst lakes in the permafrost in last decades, confirmed in a large number of remote studies, for example, [6,7]. On the basis of this model authors [8] carried out a forecast of changes in the permafrost zone of Western Siberia thermokarst lakes area up to 2030 using data on temperature and presipitation obtained by linear extrapolation of the reanalysis data.

However, modern forecasts of temperature changes for the north of Western Siberia [9] shown that the trend of average annual temperatures in the long term is differ from the linear type. At present the forecast estimates of dynamics of thermokarst lakes under climate changes in coming decades after 2030 year are not available. Therefore, it is interesting to carry out predicting dynamics of thermokarst processes in West-Siberian permafrost on base of the forecast
estimates of climate changes obtained Klimenko et al [9], what is the aim of the present work.

2 Geo-simulation model of lake field dynamics based on experimental data from satellite images

Simulation modelling is one of the most important mathematical modelling types. According to Moiseev and Svirezhev [10], simulation modelling is a research method which can build an approximate model of a studied object; the simulation model describes a real object with accuracy sufficient for current research. Kosolapova and Kovrov [11], and Low and Kelton [12] claim that simulation modelling is used to construct models in cases where, firstly, there is no analytical solution or this solution is very complex and requires huge computer capacity and, secondly, the amount of experimental data about a modelled object is insufficient for statistical method. In such a case a mathematical model is developed in simulation modelling. For modelling spatial objects Polishchuk and Tokareva [13] and later Zhao and Murayama [14] have introduced a special term ”geo-simulation modeling”. Problems of creating a geo-simulation model of thermokarst lake fields will be considered further.

Creation of a geo-simulation model of thermokarst lakes fields requires knowledge of the basic properties of these fields, which can be obtained experimentally. Because of the inaccessibility of the northern territories of Siberia, thermokarst experimental studies were carried out by remote sensing. For remote study twenty-nine test sites were chosen in different zones of the West-Siberian permafrost (sporadic, discontinuous and continuous). Remote study of the shape of thermokarst lakes boundaries was carried out via satellite images in our research [15]. Research conducted in test sites in sporadic, discontinuous and continuous permafrost showed that the error in estimating lakes areas while replacing their real lakes boundaries by a circle is comparatively small (about 5%). It may serve as a reason to choose a circle as a model for a lake in geo-simulation modelling thermokarst lake fields. In addition, the formation of geo-simulation model of thermokarst lakes fields in the form of a population of random circles requires experimental knowledge about the distribution of coordinates of lakes centres and the distribution of lakes sizes (areas).

To state the regularities for distribution of random coordinates of lakes and size-distribution of them, satellite images Landsat obtained in period 1984-2014 years were used. All space images are selected from the public archive - Global Land Cover Facility and these images are georeferenced in the UTM projection. Processing of space images was carried out by using the software ENVI 4.7 and ArcGIS 9.3. Lakes classification on the Landsat images was carried out by the method of a binary coding (encoding binary classification algorithm in the software ENVI 4.7). Lakes areas are defined by using ArcGIS 9.3. At each test site were identified from hundreds to thousands of lakes. Received data about lake areas were used to determine the average area of lakes and to build both
distribution histograms of coordinates centers and of lakes areas for each test site.

Analysis of histograms of distribution of latitude and longitude values of location of lakes centers given in [4,16] showed that experimental regularities of distribution of coordinates of lakes centers correspond to the law of uniform density according to criterion with a probability of 95% [17]. Histograms of size-distribution of lakes were built for all the test sites, located in different permafrost zones. Examples of the histograms are represented in [4]. Comparison of the histograms shows that they have, in general, an exponential character by means of the experimental law of distribution, which makes it possible for thermokarst lake fields to be modelled easily. We may choose a one-parameter exponential law to describe thermokarst lake area distribution in the following form:

\[ y = \lambda \times \exp^{-\lambda S} \]  

where \( \lambda \) - a parameter of distribution law.

The value of parameter \( \lambda \) can be determined with the help of experimental data according to the formula:

\[ \lambda = \frac{1}{\bar{S}} \]  

where

\[ \bar{S} = \frac{1}{n} \sum_{i=1}^{n} S_i, \quad i = 1, n \]

\( S_i \) is area of \( i \)-th lake in test site; \( n \) - number of lakes in this test site.

Testing correspondence of exponential law of lake area distribution given by Eq. (1) to experimental histograms shows that in all researched test sites this law corresponds to experimental data in accordance with criterion \( \chi^2 \) with average probability 90%. Consequently, the stated law of distribution of lake area in form Eq. (1) does not contradict the experimental data. The analysis of the experimental distribution of lakes according to their areas shows that \( \lambda \) in all test sites varies in the range of 0.034 – 0.086 with average values 0.06.

Accordingly, the following fundamental principles determining substantial properties of a model of spatial-temporal structure of thermokarst lake field can be formulated:

1. Lake coastline shapes can be represented by a circle equation with centres coordinates \( x_i, y_i \), and area \( S_i \) (\( i \)-lake serial number).
2. Spatial changes in the position of centres of circles and their areas are statistically independent.
3. Random distribution of circle centres coordinates \( x_i, y_i(i = 1, n) \) is governed by a uniform law.
4. Random distribution of number of circles over their areas conforms to the exponential law of distribution as in (1) with \( \lambda \) as a parameter.
5. Time changes in statistical properties of population of random circles and their dependence on climatic changes are determined by dependency of parameter $\lambda$ on time and climatic characteristics in the following equation:

$$\lambda = f(T, P, t)$$  \hspace{1cm} (3)

where $T$ - temperature, $P$ - level of precipitation and $t$ - time.

Model of field of thermokarst lakes is a collection of random circles, the statistical properties of which correspond to the above principles (1-5). Consequently, major elements in the model description are characteristics of lake shapes, parameters of their random location on surface and random distribution of lakes over their size (areas).

It is necessary to discuss questions of study of interrelation of climate and geo-cryological changes in permafrost and its accounting in the model. To analyse a correlation of area change of thermokarst lakes and climatic indices (average annual temperature and precipitation level) an alternative approach was taken to obtain data on air temperature and precipitation. The approach is based on re-analysis of meteorological data \[18\] which makes it possible to estimate the value of climatic characteristics in test sites. On base of the re-analysis approach tables of temporal series of annual average value of air temperature and annual sum of precipitation for each test site were obtained.

To study the interrelation between the changes of thermokarst lake areas and changes of air temperature and precipitation level we shall compare coefficients of a linear trend of time changes of average values of the lakes’ areas and climate characteristics. It is the analysis of the data obtained for developing a model of thermokarst lake fields suitable for prediction that is of most interest. It is necessary to study temperature dependence of parameter $\lambda$, which determines the kind of law for thermokarst lakes’ distribution in accordance with their areas, discussed in \[8\]. The data exist only for the years when cloudless images were taken, which made it possible to calculate the value of parameter $\lambda$.

Previously the equation of dependence of parameter $\lambda$ on time and climate features was introduced in implicit form (3). To develop a model of actual thermokarst lake dynamics it is necessary to define this dependence in explicit form. This was the reason for doing multidimensional regression analysis \[19\] of time series of the values of parameter $\lambda$ and climate features in the West Siberian territory under study represented in \[16\].

The results of multidimensional regression analysis of the data on parameter $\lambda$ and climate features can be presented as an equation of multiple regression in the form:

$$\lambda = c_0 + c_1 \times x_1 + c_2 \times x_2 + c_3 \times x_3$$  \hspace{1cm} (4)

where $x_1$ - average annual air temperature, $x_2$ - precipitation level, $x_3$ - time, $c_i$ - coefficients of regression equation $i = 0, \ldots, 3$.

In the result of the regression analysis of time series of the values of parameter and climate features, the following values of regression equation coefficients were obtained (4):
\[ c_0 = -0.585ha^{-1}; \ c_1 = 0.00062ha^{-1}/^\circ{C}; \ c_2 = 0.00014ha^{-1}/mm; \ c_3 = 0.00032ha^{-1}/year. \]

The stated regression dependence of parameter \( \lambda \) on time and climate changes is a basis for developing algorithms for modelling random thermokarst lake fields, discussed in the next section.

3 Methodic questions of predicting of lake’s field based on geo-simulation modeling

In a general case, mutual density of probabilities of random coordinates of centres and areas of circles imitating lakes in a mathematical model of random thermokarst lake fields can be presented in the form:

\[
f(x, y, s) \quad (5)
\]

where \( x \) and \( y \) - coordinates of circle center in a model; \( s \) - area of a circle imitating a lake.

Consequently, the totality of circles in the model of lake fields will be presented as a totality of groups of three random values \((x, y, s)\). To develop an algorithm for modelling thermokarst lake fields, it is necessary to take into consideration statistical connections between changes in lakes coordinates and their areas.

Further to equation (5), the random-number sequence determining characteristics of location of circles centers \((x, y)\) is generated using the antenna of pseudo-random numbers distributed in accordance with the law of even distribution. And to form circles of random size whose areas are distributed according to the law conforming to equation (1) it is necessary to generate random-number sequences distributed in accordance with the exponential law. Consequently, together with using software generators for even distribution of pseudo-random numbers, the software realization of an imitation model of thermokarst lake fields includes creating a generator for pseudo-random number sequences, distributed in accordance with the exponential law.

We should consider a geo-simulation model of spatial structure of thermokarst lake field \( M_{Sp}(t) \) that is a totality of circles and reflects the state of a thermokarst lake field at the moment of time \( t \). To model the dynamics of thermokarst lake fields, we should consider a general model of spatio-temporal structure of a thermokarst lake field in the form:

\[
M_{SpTM} = \{M_{Sp}(t_1), \ldots, M_{Sp}(t_j), \ldots, M_{Sp}(t_n)\}, j = 1, \ldots, n \quad (6)
\]

which is a time sequence of geo-simulation models of a thermokarst lake field \( M_{Sp}(t_j), j = 1, \ldots, n \) where each model relates to a particular moment of time.

Fig. 1 gives a visual presentation of the general model for spatio-temporal structure of thermokarst lake fields in the form of geo-information system (GIS) layers that relate to given time moments \( t_1, t_2, \ldots, t_n \in (t_1, t_n) \).
When modelling spatio-temporal structure of thermokarst lake fields it is important to take into consideration both time dependence and climate features (temperature, precipitation level). Accordingly, the dependence of parameter $\lambda$ on time and climate features is determined by the equation of multiple regression in the form (4). This is the reason why equation (4) was used to develop an algorithm for numerical modelling dynamics of thermokarst lake fields.

The developed algorithm for modelling dynamics of thermokarst lake fields can be presented as follows:

step 1 — the year of modelling is specified $t_j, j = 1, \ldots, m$;
step 2 — the areas ($S_{MA}$) of the model area (MA) under study are specified;
step 3 — lake density ($\sigma_{MA}$) in MA is specified;
step 4 — the number of circles within MA is determined in accordance with formula: $N_{MA} = S_{MA} \times \sigma_{MA}$;
step 5 — the centre of MA location in the map is specified;
step 6 — parameter $\lambda$ is determined in accordance with formula (4) for given values of temperature and time (year of modelling);
step 7 — pseudo-random number is generated, distributed in accordance with uniform law;
step 8 — using the number obtained at the previous step, a pseudo-random number is calculated to characterize the value of circle area according with formula:

$$s_i = -\frac{1}{\lambda} \ln z_j$$

where $z_j$ - pseudo-random number distributed in accordance with uniform law in interval $(0, 1), j = 1, \ldots, m$;

step 9 — two pseudo-random numbers are generated, distributed in accordance with uniform law, determining the coordinates for circle centre location on the screen;
**step 10** — using the values of a number triple \((x, y, s)\) obtained at previous steps 8 and 9, in accordance with equations

\[ x_{ki} = R_i \times \cos \theta_i + x_i \] (8)

and

\[ y_{ki} = R_i \times \sin \theta_i + y_i \] (9)

a circle is formed on the screen;

**step 11** — if the number of circles obtained is less than \(N_{MA}\), determined at step 4, the algorithm repeats beginning with step 7, otherwise it is completed.

The given algorithm allows formation of a model of spatial structure for a given time moment \(M_{SP}(t_j)\), where \(j = 1, \ldots, m\). To make a general model of dynamics of a thermokarst lake field by means of forming a time sequence of models \(M_{SP}(t_j)\) for a given set of moments \(t_j(i = 1, \ldots, m)\) the algorithm repeats for the number of times \((m)\) needed.

Accuracy of modelling dynamics of thermokarst lake fields was studied in the form of computer experiment on the model. Values of parameter \(\lambda\) in this case are calculated according to the multiple regression formula (4) using the data about average annual temperature and precipitation level determined for each \(S\) by re-analysis. Then a model field is formed in accordance with the algorithm described above. Estimation showed that the error of determination of average values of lakes areas on base of modelling with use of experimental data is 17%. This may well be regarded as a suitable result of modelling thermokarst lake fields for predicting thermokarst lake fields dynamics.

### 4 Software of predicting of lake’s field dynamics

Geo-simulation models are considered as more promising and make it possible to study the dynamics of the thermokarst lakes fields in today’s global warming. Recently, within the framework of the ideology of the simulation formed one of the new areas of computer modeling, which is called geo-simulation modeling. It is a simulation of complex objects with a spatial structure and realized with the use of methods and means of geoinformatics. The geo-simulation model of thermokarst lakes fields in the form of the random circles set is described in [16]. This model takes into account the properties of the main characteristics of the real thermokarst lakes fields, which are identified by the experimental data of remote sensing. However, issues of information technology of the geo-simulation thermokarst lakes fields, in particular, software implementation, is now not enough considered and this fact defined the purpose of the present work.

The implementation of a mathematical model of thermokarst lakes fields considered in [20] involves the creation of the pseudo-random number sequences - triples: the first two pseudo-random numbers are distributed over a uniform density of the law, and the third - exponentially. The software package is designed, and its structural diagram is shown in fig. 2.
The structure of the developed software package includes the following main components:

1. **subsystem of geo-simulation modeling thermokarst lakes fields**;
2. **the subsystem of the displaying model results**;
3. **database (DB)**;
4. **the subsystem of dataset formation**.

The following describes components of the software package. The subsystem of geo-simulation modeling thermokarst lakes fields, developed by the author, is a set of software modules that provide the input parameters of the model, the formation of pseudo-random number sequences and output of simulation results. The structure of the subsystem of geo-simulation modeling thermokarst lakes fields includes the following main blocks:

- **input module** is designed to provide the pseudo-random number sequences values of model parameters;
- **generator of the pseudo-random number sequences** is a major component of the subsystem of geo-simulation modeling thermokarst lakes field and it is designed to generate random number sequences in the algorithms implementation for numerical simulation of thermokarst lakes fields. The numerical simulation algorithm is described in details in [16];
- **output module** is designed to convert the simulation results in one of the following formats: Microsoft Excel (* .xls), a vector format (* .shp), bitmap format (* .jpeg)).

**The subsystem of the displaying model results** allows showing the output given either on a digital map by means of geographic information system (ArcGIS), or in the form of electron tables and graphics in MSExcel. The database is a part of the software package. It is a store spatial and attributes information on the study sites obtained during the field experiment. Description of the database structure and capacity is given in [16]. **The subsystem of dataset formation** allows extracting from the database the information about the object of research and forming data sets for model experiments.
5 Sample of predicting of lake’s field dynamics in West-Siberian permafrost

To generate the forecast of dynamics of the thermokarst lakes fields in permafrost of Western Siberia for future decades, it is necessary to have projections on climate changes in the study area. Temperature forecast for the north of Western Siberia to 2300 are presented in [9,21]. Data on precipitation forecasts in the coming decades are not existed in the literature. A comparison of the coefficients of the regression equation (4) shows, that in predicting the dynamics of thermokarst processes in the permafrost zone of Western Siberia can neglect the effect of precipitation and take into account only the temperature changes. Therefore, predicting the dynamics of thermokarst lakes fields may be carried out using the temperature forecast data [21] for the north of Western Siberia by the computer experiments with the model in accordance with the below considered scenario.

**Scenario of computer simulation experiment:** Predicting the dynamics of thermokarst lake fields on the basis of the geo-simulation model using predictive estimates of temperature changes in the north of Western Siberia [21] for the period up to 2050 year.

As shown in [21], an increase in temperature, which began after 1970, will continue in the coming decades. According to [21], the maximum warming could reach by the end of the forecast period almost 1°C compared to the present time. The result of forecasting the dynamics of thermokarst-lake fields in Western Siberia, is shown in Fig. 3 as plot of time dependence of the mean value of thermokarst lake area.

Because of the current lack of prognosis of precipitation changes in the prediction period, at obtaining of forecast assessments of lakes area changes are used data only about temperature changes. To substantiate the prediction possibility without precipitation we conducted a comparison of the regression equation coefficients (4), the values of which are given on p. 6. Comparison of these coefficients shows that $c_2 \ll c_1$ and $c_2 \ll c_3$. This allows us to ignore the contribution of term $c_2 \times x_2$ in the equation (4) in the value of the parameter $\lambda$. Comparison of the results of the lakes dynamics prediction by using the developed algorithm for the period up to 2030 [16] showed the unessential difference between the forecast estimates for both cases with and without precipitation, that proves our statement.

The most important result of the analysis of forecast estimates presented in Fig. 3, is the conclusion about continuation of the reduction of the average area of thermokarst lakes in the West Siberian permafrost. The graph in Fig. 3 shows that at the end of the forecast period, the average area of the lakes can be reduced to 14.5 hectares, i.e. approximately 20% compared to 2010 year. Thus, the continued increase in the coming decades the average annual temperature of surface atmosphere will be accompanied by a reduction in the average area of thermokarst lakes in the permafrost zone of Western Siberia, what is the result of permafrost degradation and reduce its strength.
6 Conclusion

Designed geo-simulation model of the dynamics of thermokarst lake fields, taking into account the relationship between regional geocryological and climate changes, allowed to carry out forecasting changes of lake sizes in conditions of continuing global warming. Long-term prognosis of the dynamics of lake thermokarst-fields using this model showed that with the growth of the air temperature in West-Siberian permafrost lake areas will be reduced on average by approximately 20% by the end of the forecast period. A new forecast assessments of changes of thermokarst processes on the territory of West-Siberian permafrost were obtained using computer experiments with the model. It is shown that the gradual increase in temperature to 1°C by 2050 year will cause a reduction in the area of thermokarst lakes, what is an indicator of the continuing degradation of permafrost in coming decades.

The information technology of modeling and forecasting the dynamics of thermokarst lakes fields can be used for solving the problems of reducing the accident rate on the infrastructure facilities in the permafrost territories and predicting the dynamics of greenhouse gas emissions from thermokarst ponds in Western Siberia under the impact of global warming.

This research was supported by grant RFBR Project No. 15-45-00075.

References


Modification of Fourier Approximation for Solving Boundary Value Problems Having Singularities of Boundary Layer Type

Boris Semisalov and Georgy Kuzmin
Institute of Computational Technologies, SB RAS,
Academic Lavrentiev av. 6, 630090, Novosibirsk, Russia,
Novosibirsk State University
Pirogova str. 1, 630090, Novosibirsk, Russia
ViBiS@ngs.ru, kubgeom@mail.ru

Abstract. A method for approximating smooth functions has been developed using non-polynomial basis obtained by mapping of Fourier series domain to the segment $[-1, 1]$. High rate of convergence and stability of the method is justified theoretically for four types of coordinate mappings, the dependencies of approximation error on values of derivatives of approximated functions are obtained. Algorithms for expanding of functions into series with coupled basis composed of Chebyshev polynomials and designed non-polynomial functions are implemented. It was shown that for functions having high order of smoothness and extremely steep gradients in the vicinity of bounds of segment the accuracy of proposed method cardinally exceeds that of Chebyshev's approximation. For such functions method allows to reach an acceptable accuracy using only $N = 10$ basis elements (relative error does not exceed 1 per cent)

Keywords: singular perturbation, small parameter, coordinate mapping, boundary value problem, Fourier series, Chebyshev polynomial, non-polynomial basis, estimate of convergence rate, collocation method

1 Introduction

By now, a huge amount of urgent scientific and technological problems reduces to boundary value problems for differential equations having pronounced singularities of boundary layer type. The most popular approaches to solving them are based on construction of computational grids with piecewise linear/polynomial approximation of the unknown function in each cell of grid [1]. Such approaches provide relatively low rate of convergence and lead to essential refinement of grid in the vicinity of boundary layer and consequently to growth of computational costs and errors. In [2,3] methods of coordinate transformations were developed which allow to decrease the influence of mentioned effect due to application of special coordinate mappings eliminating the singularity. Nevertheless, the analysis of key issue concerning the smoothness of such transformations and its influence on the quality of approximation method is absent. Frequently, authors
restrict their self by transforming uniform or more special grid and performing numerical experiments using finite difference or spectral methods [2]– [5].

In the present paper a step aside from the traditional grid approaches is made and the approximations based on mapping of Fourier series domain to the segment $[-1, 1]$ are used to approximate functions having singularities of boundary layer type. One of such mappings assigned by function $\cos(x)$ transforms Fourier basis to basis composed of Chebysev polynomials. A special feature inherent to these bases is the absence of saturation of corresponding approximations [6]. It means that using Fourier and Chebyshev bases allows to obtain the asymptotic of error of best polynomial approximation while approximating functions having any order of smoothness or singularities in complex plain. The loss of effectiveness of the mentioned approximations while solving problems having singularities of boundary layer type is caused by degradation of asymptotic properties of best polynomial approximations with fast increase of gradients of approximated function. In order to eliminate this problem, trigonometric Fourier basis can be transformed into non-polynomial algebraic one retaining all its good properties of high convergence rate and computational stability, but specially adapted to approximation of smooth functions having singularities of boundary layer type.

2 Preliminary information. Problem description

In framework of approximation theory of continuous and smooth functions $f(x)$ ($x \in D \subset \mathbb{R}$) that are elements of spaces $C(D)$, $W^r_p (M, D) = \{ f \in C^r (D) : |f^{(r)}| \leq M = M(r) \}$ the following notions are often used (see for example [6]).

1. **Norms of function** $\| f \|_D = \max_{x \in D} |f|$, $\| f \|_p = \left( \int_D f^p (x) dx \right)^{\frac{1}{p}}$.

2. **Finite-dimensional approximating space** $K_n$ with elements used for approximation of $f(x)$ that usually are series or polynomials including $n$ summands or monomials.

3. **Operator of approximation (or simply approximation) of function** $f(x)$ is a continues mapping $P_n$ performing a projection of functional space on approximating one ($P_n : C(D) \rightarrow K_n$ or $P_n : W^r_p (M, D) \rightarrow K_n$).

4. **Best approximation of function** $f \in C(D)$ is element $e_n (f, K_n) \in K_n$ providing the lower bound to be reached $\varepsilon_n^*(f, K_n) = \inf_{g \in K_n} \| f - g \|$. 

5. **Method without saturation** (loose definition) is a method of approximation of function $f(x)$ having asymptotic of error of the best polynomial approximation for any order of smoothness of $f(x)$. Rigorous definition based on the analysis of asymptotic of Alexandrov’s diameters is given in [6].

The results obtained in works by Lebegue, Faber, Jackson, Bernstein allow to separate three classes of smooth functions with fundamentally different asymptotical behavior of errors of best approximations in space $K_n$ of algebraic polynomials of $n$th power. The similar asymptotical behavior is valid for periodic functions and trigonometrical polynomials.
I. If \( f(x) \in W^r_p(M,D) \) is \( r \)-times continuously differentiable function on segment \( D \) and all its derivatives up to order \( r \) are bounder by value \( M(r) \), then

\[
\sup_{f \in W^r_p(M,D)} \varepsilon_n^*(f, K_n) \leq M(r) C_r n^{-r}, \tag{1}
\]

where \( C_r \) depends on \( r \) only, [7].

II. If \( f(x) \in C^\infty(D) \) is infinite differentiable function with a singularity (like pole) on complex plain, then one can find a number \( q (0 < q < 1) \) and a sequence of polynomials \( P_n(x) \), such that

\[
\|f(x) - P_n(x)\| \leq C q^n, \ x \in D, \tag{2}
\]

here \( q < 1 \) is defined by location of singularity in complex plain, \( C \) is constant, [8].

III. If \( f(x) \in \text{Ent} \) is entire function, then

\[
\varepsilon_n^*(f, K_n) \leq \frac{M(n)(\text{diam}D)^n 2^{1-2n}}{n!}, \tag{3}
\]

where \( M(n) = \|f^{(n)}\| = o(n!) \). It follows from estimates of error of polynomial interpolation (see [9]) and Cauchy–Hadamard inequality.

**Remark 1.** For smooth functions of I and III classes the accuracy of best approximations depends on maximal values of derivatives of function on \( D \) (values of \( M(r) \) and \( M(n) \) in (1), (3)). For functions of class II the error can be defined through the values of function itself beyond the segment \( D \) on complex plain.

In order to implement the properties of best approximations in this work the Fourier and Chebyshev series are used. Note that such approximations are equal in a specific sense. Indeed, let \( f \in C([-1,1]) \), then performing the change of variable \( x = \cos \theta, \ \theta \in [0,2\pi] \) one can obtain 2\( \pi \)-periodic even function \( \tilde{f}(\theta) = f(\cos \theta) \). Fourier decomposition of it is \( \tilde{f}(\theta) = \sum_{k=0}^{\infty} a_k \cos(k\theta) \). Hence

\[
f(x) = \sum_{k=0}^{\infty} a_k \cos(k \arccos(x)) = \sum_{k=0}^{\infty} a_k T_k(x). \tag{4}
\]

In other words Chebyshev polynomials \( T_k(x) \) can be obtained by mapping Fourier series domain to the segment \([-1,1]\), see [10]. In [6] is proved that such approximations presents the methods without saturation and therefore they are extremely efficient for solving problems with smooth solutions. However, if values \( M(r) \), \( C \), \( M(n) \) in (1)–(3) are large, then it can be wrong.

A simple example is a boundary-value problem for differential equation of second order with small factor \( \varepsilon \) of second derivative

\[
\varepsilon \frac{d^2 f}{dx^2} - f = \varepsilon g''(x) - g(x), \ f(-1) = 1, f(1) = -1, \tag{5}
\]

here \( x \in [-1,1], g(x) \) is a given smooth function. Solution to the problem is

\[
f(x) = \xi(x) + g(x), \ \xi(x) = C_1 e^{A(0.5x+0.5)} + C_2 e^{-A(0.5x+0.5)},
\]

where \( A > 0 \).
Here $\xi(x)$ is an exponential boundary value component of $f(x)$, $C_1, C_2, A = 1/ \sqrt{1/\varepsilon}$ are constants. Table 1 shows the values of dimension of $K_n$ ensuring that the relative errors of approximation is never higher than 1 per cent. These results were obtained in accordance with the estimates of best polynomial approximations (1), (3), while function $g(x)$ has different orders of smoothness.

Thus, it can be observed that the higher order of smoothness is, the less data is necessary for recovering solution with a desired accuracy. However, even in the case of infinitely differentiable function a space $K_n$ of dimension of many thousands can be required to reach considerably low accuracy of 1 per cent. This effect is shown on Fig 1 where a graph of solution to a problem (5) is given with $g(x) \equiv 0$ (i.e. a graph of function $\xi(x)$) and logarithm of error of approximation of $\xi(x)$ in space of Chebyshev polynomials

$$\nu = \max_{x \in [-1,1]} |\xi(x) - \sum_{m=0}^{n-1} a_m T_m(x)|.$$ 

![Fig. 1. Solution to the problem (5) with $\varepsilon = 10^{-3}$ (solid line), $\varepsilon = 10^{-4}$ (dash line), $\varepsilon = 10^{-6}$ (dote-and-dash line): a – graph of $\xi(x)$; b – dependance of $\log_{10} \nu$ on $n$](image-url)
in the vicinity of boundary layers, see. [10], [11]. Considering a problem with
boundary layer of size \( \rho \) having solution without singularities in inner part of
the segment \([-1,1] \), to reach an accuracy of 1 per cent one should use a basis of

\[
\begin{align*}
n &\approx 3/\sqrt{\rho} \\
\end{align*}
\]  

(6)

Chebyshev polynomials for approximation of unknown function. In the consid-
ered case \( \rho \approx 5.6/A \) (here condition \( \xi(\rho) = 0.1 \) is used). Appearance of expres-
sion \( \sqrt{\rho} \) in the denominator of fraction is caused by concentration of Cheby-
shev nodes. Indeed, uniformly distributed zeroes of trigonometrical monomials
\( \cos(k\theta) \) are concentrated in the vicinity of points \( \pm 1 \) under the map \( x = \cos(\theta) \)
(see fig. 2 a). Moreover as \( \cos(\theta) \approx 1 - \theta^2/2 \) when \( \theta \to 0 \) and the first Chebyshev
node \( x_1 = \cos(\theta_1) = \cos \pi/2n \), then \( |1 - x_1| \approx \pi^2/8n^2 \). Further, the empiri-
cal requirement that even three nodes should lay on the boundary layer gives

\[
3|1 - x_1| \leq \rho \text{ or } n \approx \frac{3\pi}{2\sqrt{2}\sqrt{\rho}} \text{ that corresponds to (6).}
\]

3 Description of a method

For efficient approximation of function having boundary layer component a mod-
ification of map \( x = \cos(\theta) \) that transformed Fourier basis to Chebyshev one (4)
is proposed. According to (6) a natural requirement is to use stronger concen-
tration of zeroes of basis functions in the vicinity of segment borders (see 2 b).

Let us assume that \( D = [-1,1] \). Define a function \( x = \varphi(y) : [-1,1] \to [-1,1] \)
satisfying the following basic requirements:

1) function \( \varphi(y) \) is bijective mapping, \( \varphi(1) = 1, \varphi(-1) = -1 \);
2) \( \varphi(y) \) is an infinite differentiable or even entire function;
3) an inverse function \( y = \varphi^{-1}(x) \) can be expressed in analytical form or easily
computed;
4) a derivative \( \varphi'(y) \) in the vicinity of points \( \pm 1 \) is close to zero.

Note that concentration of zeroes of basis functions in the vicinity of segment
borders can be obtained due to the last requirement. One of possible forms of
function \( \varphi(y) \) is given on fig 2 b.

For approximation of function \( f(x), x \in [-1,1] \) let us consider the expansion
of \( 2\pi \)-periodic even function \( f(\varphi(\cos \theta)) \, (\theta \in \mathbb{R}) \) into Fourier series:

\[
f(\varphi(y)) = f(\varphi(\cos \theta)) \approx \sum_{k=0}^{n-1} a_k \cos(k\theta). \quad (7)
\]

As a result one obtains

\[
f(x) \approx P_n(x) = \sum_{k=0}^{n-1} a_k \cos[k \arccos(\varphi^{-1}(x))]. \quad (8)
\]
Fig. 2. Concentration of nodes of trigonometrical monomials $\cos(n \theta)$ as $n = 11$ (indicated by arrows): a – using mapping $y = \cos(\theta)$, b – using mapping $x = \varphi(y)$

Here we denote $y = \cos \theta$. Thus, the basis of approximating space $K_n$ can be specified as $B_k(x) = \cos[k \arccos(\varphi^{-1}(x))]$, where zeroes of $B_k(x)$ are $x_{km}^k = \varphi\left(\cos\left(\frac{(2m + 1)\pi}{2k}\right)\right)$, $k, m = 0, \ldots, n - 1$. Note, that under the properties 1, 3, $\varphi(y)$, $B_k(x)$ are bijective easily computed functions.

Lemma 1. Approximation (8) is equivalent to expansion of function $f(\varphi(y))$ into series with basis consists of Chebyshev polynomials $T_k(y) = \cos(k \arccos(y))$, that is why

1) for approximations (8) error estimations of best approximations (1)–(3) hold;
2) elements of matrices $B_n - b_{km} = B_k(x_{mn}^m)$, $k, m = 0, \ldots, n - 1$ do not depend on $\varphi$.

The proof of Lemma 1 is obvious taking into account (8) and that Chebyshev and Fourier expansions are approximations without saturation. The second condition of Lemma 1 concerns numerical approximation of $f$ by collocation method with nodes $x_{mn}^m$. Lemma declares that such a method is universal, its properties do not depend on the choice of function $\varphi(y)$.

To settle a key question on rate of growth of coefficients $M(r), M(n)$ in estimates (1), (3) the following property was proved.
Lemma 2. For derivative of composed function the following equality is valid
\[
(f(\alpha(y)))^{(n)} = \sum_{k=1}^{n} f^{(k)}(\alpha(y)) \left[ \sum_{\alpha_k + \ldots + \alpha_l = n} C_{kl}^n \prod_{j=1}^{k} (\alpha(y))^{(\alpha_j)} \right],
\]
where the second summation goes over all possible integer partitions of number \(n - A_{kl}, \) that consist of \(k\) positive integer numbers \(\alpha_1, \ldots, \alpha_k, \) \(l = 1, \ldots, L; C_{kl}^n\) are constants. Moreover \(\forall n\) as \(k = 1\) and \(k = n\) one has: \(L = 1, C_{11}^n = 1, C_{n1}^n = 1.\)

This Lemma corresponds to the well-known Fa di Bruno’s formula and can be proved using mathematical induction technique.

4 Analysis of four types of function \(\alpha(y).\)

Now let us propose and investigate four types of function \(\alpha(y).\) To this end the following notations are necessary. Let \(D_0 \subset D, D_1 \subset D\) be neighborhoods of segment \(D\) representing boundary layers, \(D_0\) be a certain neighborhood of central point of \(D.\) Let us denote by \(\|\cdot\|_L, \|\cdot\|_D\) the supremum norms of continues function on \(D_0 \cup D_1\) and on \(D_0\) correspondingly. Further, we assume \(a = -1, b = 1, \forall s < r, A_s f^{(s)} || = \| f^{(s+1)} \|, \) where \(A_s > 1\) is constant, \(r\) is order of smoothness of \(f(x).\) Typically for problems with boundary layer one has \(A_s >> 1.\) Let \(A = \max_{s < r} A_s, \rho(A)\) be a size of boundary layer (as it follows from the example with exponential boundary layer, \(\rho\) depends on \(A,\) see comments to formula (6)), \(o(f^{(s)})\) be such value that \(\lim_{A_1, \ldots, A_{r-1} \to \infty} \frac{o(f^{(s)})}{\| f^{(s)} \|_L} = 0.\)

I. Trigonometric function
\[
\alpha(y) = \sin \left( \frac{\pi y}{2} \right).
\]

Theorem 1. Let \(\rho^2(A)A \to 0\) as \(A \to \infty.\) For the approximation (8) with function \(\alpha(y)\) of form (10) the estimate (1) is valid, but constant
\[
M(r) = C_{r/2}^r \left( \frac{\pi}{2} \right)^r \| f^{(r/2)} \| + o(f^{(r/2)}) \text{ (if } r \text{ is even),}
\]
\[
M(r) = C_{r/2}^r B(A) \left( \frac{\pi}{2} \right)^r \| f^{(r/2)} \| + o(f^{(r/2)}) \text{ (if } r \text{ is odd),}
\]
where \([s]\) denotes integer part of number \(s, C_{r/2}^r\) \(t, C_{r/2}^r t\) are coefficients of (9), corresponding to integer partitions \(A_{r/2} t = (2, \ldots, 2), A_{r/2} t = (2, \ldots, 2, 1).\)

II. Polynomial of third power
\[
\alpha(y) = ay^3 + by^2 + cy + d.
\]

After taking into account properties 1)–4) of \(\alpha(y)-function\) one obtains \(b = d = 0, a = 1 - c, 1 \leq c \leq 1.5.\) Assuming \(p = c,\) one gets
\[
\alpha(y) = (1 - p)y^3 + py,
\]
where $1 \leq p \leq 1.5$ is a free parameter equal to the value of derivative $\varphi'(0)$ and determining the value $\varphi'(\pm 1)$: as $p \rightarrow 1.5 \varphi'(\pm 1) \rightarrow 0$.

**Theorem 2.** For approximation (8) with function $\varphi(y)$ of form (12) as $p = 1.5$ and $\rho^2(A)A \nrightarrow 0$ as $A \rightarrow \infty$ the estimate (1) is valid, but constant

$$M(r) = C_{r/2} i 3^{r/2} ||f^{(r/2)}|| + o(f^{(r/2)}) \text{ (if } r \text{ is even)},$$

$$M(r) = C_{r/2} i 3^{r/2} + 1 ||f^{(r/2)}|| + o(f^{(r/2)}) \text{ (if } r \text{ is odd),}$$

where $C_{r/2} i$, $C_{r/2} i'$ are the same as in theorem 1.

Theorems 1, 2 can be proved using Lemmas 1, 2 and taking into account that $C_{r/2} i$ provides maximal rate of decrease of right part of (3) with function $\varphi(y)$ has form (10), (12).

Remark 2. By changing in given theorems index ”$r$” on index ”$n$”, one can obtain similar results for case of entire $f(x)$, namely the similar equalities for $M(n)$ from estimate (3) when $\varphi(y)$ has form (10), (12).

Let us consider another approach to construction of $\varphi(y)$-function, it does not require first derivative of $\varphi(y)$ to be equal to zero in points $\pm 1$, but requires all the derivatives to be very small in vicinity of $\pm 1$.

III. Function

$$\varphi(y) = \arctan (by) / \tilde{b}, \quad \tilde{b} = \arctan b. \quad (13)$$

**Theorem 3.** If in expression (8) function $\varphi(y)$ of form (13) has parameter

$$b << n^{k-2} \sqrt{\frac{f^{(k)}}{L} ||f^{(k)}||_D} = b_k, \quad \forall k = 1, 2, ..., n, \quad (14)$$

then (8) satisfies the estimate of accuracy of best approximation (3) with $M(n)$ equals for large $n$ and $A$ to the following

$$M(n) = \max \left( \frac{||f^{(n)}||_L}{b(bb)n^{-1}}, \frac{||f'||_L}{b} n! \right) + o(n!) + o(f^{(n)}). \quad (15)$$

Values

$$b \approx b_* = \frac{2}{\pi} n^{-1} \sqrt{\frac{||f^{(n)}||_L}{n!} \frac{1}{n!}} \quad (16)$$

under condition $1.56 < b << \min_{k=1,...,n} b_k$ provide maximal rate of decrease of right part of (3) with growth of $n$ and $b(bb)n^{-1}$-time profit in accuracy in comparison with the best polynomial approximation.

IV. Function

$$\varphi(y) = \tilde{\mu} \left( \frac{2}{1 + \exp(-\mu y)} - 1 \right), \quad \tilde{\mu} = \frac{1 + \exp(-\mu)}{1 + \exp(-\mu)}. \quad (17)$$
Theorem 4. Let \( f(x) \in W_r^p(M, D) \), \( r > 1 \) and in expression (8) function \( \varphi(y) \) of form (17) has parameter, satisfying the inequalities

\[
\mu << \frac{1}{\alpha_s} W(\alpha_s \beta_s) = \mu_s, \ \forall s = 1, \ldots, r - 1, \quad \mu << \pi \sqrt{\frac{\|f^{(r)}\|_L}{\|f^{(r)}\|_D}} = \mu_0,
\]

where \( W(x) \) is the Lambert \( W \)-function (the inverse function to \( x = w \exp(w) \)),

\[
\alpha_s = \frac{s}{r - s}, \quad \beta_s = 4^{\alpha_s} \pi^{r-s} \sqrt{\frac{\|f^{(s+1)}\|_L}{\|f^{(s+1)}\|_D}}.
\]

In this case (8) satisfies the estimate of accuracy of best approximation (1) with \( M(r) \) equals for large \( r \) and \( A \) to the following

\[
M(r) = \max \left( \|f^{(r)}\|_L (2\mu \tilde{\mu} \exp(-\mu))^{r-1}, \|f'\|_L \tilde{\mu} r! \right) + o(r!) + o(f^{(r)}). \tag{19}
\]

Values

\[
\mu \approx \mu_* = -W \left( \frac{1}{2}^{r-1} \sqrt{\frac{\|f'\|_L r!}{\|f^{(r)}\|_L}} \right) \tag{20}
\]

under condition \( 1.57 < \mu << \min_{s=0,1,\ldots,r-1} \mu_s \) provide maximal rate of decrease of right part of (1) with growth of \( n \) and \( (\exp(\mu)/(2\mu \tilde{\mu}))^{r-1} \)-time profit in accuracy in comparison with the best polynomial approximation. Here it is considered the branch of function \( W(x) \) such that \( W(x) \to -\infty \) as \( x \uparrow 0 \).

Theorems 3, 4 can be proved using Lemmas 1, 2 and asymptotical analysis of growth of \( n \)th derivative of \( f[\varphi(y)] \) with grows of \( n \). These results can be extended to case of estimates (1), (3).

5 Computation of approximate values of functions with boundary value components

In this section one can find an experimental confirmation of results of theorems 1–4 considering approximation of solution to the boundary value problem (5) with exponential boundary layer. Detailed comments on correspondence of theoretical and experimental data given on fig. 3–5 and in Tables 2–5, are absent. Such correspondence becomes obvious if in theorems 1–4 one supposes the values \( A_1, \ldots, A_{n-1} \) to be equal to the value of \( A \) from formula of solution \( f(x) \) to the problem (5). This identification is correct because the values of derivatives \( f^{(n)}(\pm 1) \) grow with a rate of \( A^n \).

To implement the approximation according to (8) four considered types of function \( \varphi(y) \) were used, expressions for inverse functions \( \varphi^{-1}(x) \) were obtained and the values of zeroes of basis functions of approximations \( x^n_k \) were specified.

1. Sinus \((\sin)\), \( \varphi_s(y) = \sin \left( \frac{\pi y}{2} \right) \):

\[
\varphi^{-1}_s(x) = \frac{2 \arcsin x \pi}{\pi}, \quad x^n_k = \sin \left[ \frac{\pi \cos \left( \frac{(2k+1)\pi}{2n} \right)}{2} \right], \quad k = 0, \ldots, n - 1.
\]
2. Polynomial \( \text{pol} \), \( \varphi_p(y) = (1-p)y^3 + py \):

\[
\varphi_p^{-1}(x) = R \left[ -\cos \frac{\arccos z}{3} + \sqrt{3} \sin \frac{\arccos z}{3} \right],
\]

\[
R = \sqrt{\frac{p}{3(p-1)}}, \quad z = -3\sqrt{3x\sqrt{p-1}} \frac{1}{2p^{3/2}},
\]

\[
x_k^n = (1-p)\cos^3\left[ \frac{(2k+1)\pi}{2n} \right] + p\cos \left[ \frac{(2k+1)\pi}{2n} \right], \quad k = 0, ..., n-1.
\]

These expressions were obtained using Cardano’s method and analysis of three branches of inverse function \( \varphi_p^{-1}(x) \).

3. Function inverse to tangents \( \text{tg} \), \( \varphi_t(y) = \arctan(\frac{by}{\arctan(b)}) \):

\[
\varphi_t^{-1}(x) = \frac{\tan(x \arctan b)}{b}, \quad x_k^n = \frac{\arctan \left[ b \cos \left( \frac{(2k+1)\pi}{2n} \right) \right]}{\arctan(b)}, \quad k = 0, ..., n-1.
\]

4. Function including exponent \( \text{exp} \), \( \varphi_e(y) = \tilde{\mu} \left[ \frac{2}{1 + \exp(-\mu y)} - 1 \right] \):

\[
\varphi_e^{-1}(x) = -\frac{1}{\mu} \ln \left[ \frac{2}{\tilde{\mu}x + 1} - 1 \right], \quad x_k^n = \tilde{\mu} \left[ \frac{2}{\exp \left\{ -\mu \cos \left( \frac{(2k+1)\pi}{2n} \right) \right\} + 1} - 1 \right],
\]

\( k = 0, ..., n-1 \).

5.1 Approximations of boundary layer component \( \xi(x) \)

In this section results on numerical approximation of the solution \( f(x) \) to the boundary-value problem (5) are given for different values of small parameter \( \varepsilon = 10^{-3}, ..., 10^{-10} \). Let us consider first the case \( g(x) \equiv 0 \), then \( f(x) = \xi(x) \) is an exponential boundary value component.

For searching the coefficients \( a_k \) of the expansion (8), where \( k = 0, ..., n-1 \) a system of \( n \) equations should be composed:

\[
\sum_{j=0}^{n-1} a_j B_j(x_k^n) = f(x_k^n).
\]

Matrix of this system with elements \( b_{jk} = B_j(x_k^n) \) is \( C = (b_{jk}) \). Denoting column vectors

\[
\alpha = (a_0, a_1, ..., a_N)^T, \quad \beta = (f(x_0), f(x_1), ..., f(x_N))^T,
\]

one obtains a system of linear algebraic equations (SLAE) \( C\alpha = \beta \). Finally the coefficients of expansion (8) can be expressed in form \( \alpha = C^{-1}\beta \).
Remark 3. It was established by computations that the growth of number of basis function \( n \) does not affect the condition number of matrix \( C \). The first five digits of it \( 1.4142 \) remain invariable while \( n \) grows from 1 to 200. This means that using the orthogonal transformations matrix \( C \) can be inverted on computer with high precision.

The algorithm of searching coefficients \( a_k, \ k = 0, \ldots, n - 1 \) was implemented on MATLAB programming language for the following values of small parameter \( \varepsilon = 10^{-3}, 10^{-4}, \ldots, 10^{-10} \).

For each of considered types of \( \alpha(y) \) a grid of points \( (z_1, z_2, \ldots, z_K) \) was generated on the segment \([-1; 1]\). It was used for searching the maximal deviation of a given function \( f(x) \) form its approximation \( P_n(x) \) by enumeration over all points. Due to the fact that function has boundary layer in the vicinity of points \(-1\) and \(1\) a grid for enumeration was composed of large amount of Chebyshev nodes providing significant concentration of grid near the bounds of segment. The number of nodes \( z_1, z_2, \ldots, z_K \) was chosen so that the doubling of it does not affect first 2–3 digits of error \( \nu = \max_{i=1,\ldots,K} |f(z_i) - P_n(z_i)| \).

For example if \( \varepsilon = 10^{-5} \) and function \( \alpha_t(y) \) is used, the experiments show that for \( n = 48 \)

\[
\text{as } K = 100000 \text{ point} \quad \text{the error } \nu = 2.7737 \times 10^{-13}, \quad (21)
\]

\[
\text{as } K = 200000 \text{ points} \quad \text{the error } \nu = 2.7734 \times 10^{-13}. \quad (22)
\]

Consequently, we conclude that number \( K = 100000 \) is sufficient for estimating the error \( \nu \) with desired accuracy.

Now let us use a single coordinate plane to represent dependencies of \( \log_{10} \nu \) on the number of basis functions \( n \) of approximation \( P_n \) for different types of function \( \alpha(y) \) together with well known Chebyshev approximations. (see. fig. 3, 4). Number of points for enumeration \( K \) here is equal to a maximal one of all \( K \) obtained for each of considered types. Note that the values of parameters \( p, b, \mu \) in these results are taken from the vicinity of their ”optimal” values (see Table 2). Further method for searching these values is explained.

<table>
<thead>
<tr>
<th>Parameter of approximation</th>
<th>Value of small parameter ( \varepsilon )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( 10^{-3} )</td>
</tr>
<tr>
<td>( p )</td>
<td>1.1</td>
</tr>
<tr>
<td>( b )</td>
<td>1.2</td>
</tr>
<tr>
<td>( \mu )</td>
<td>1.3</td>
</tr>
</tbody>
</table>

On given figures one can observe that the decrease of values of small parameter \( \varepsilon \) results in fast decrease of convergence rate of Chebyshev expansions. Whereas, the methods designed using functions \( \alpha_t(y), \alpha_e(y) \) do not significantly
Fig. 3. Dependencies of $\log_{10} \nu$ on number $n$ for values $\varepsilon = 10^{-4}$ (b), $\varepsilon = 10^{-6}$ (b)

Fig. 4. Dependencies of $\log_{10} \nu$ on number $n$ for values $\varepsilon = 10^{-8}$ (a), $\varepsilon = 10^{-10}$ (b)
change their convergence rate while decreasing small parameter. As it was already proved, increase of $b, \mu$ allows one to reduce the influence of steep gradient.

Now let us estimate "optimal" values of parameters $p, b, \mu$ for approximations based on $ae_p, ae_t, ae_e$. To this end the dependencies of logarithm of error $\log_{10} \nu$ on $n$ and value of parameter of function $ae(y)$ were represented in Cartesian coordinate system, see those for $ae_t$ on fig. 5, 6.

![Fig. 5. Dependencies of value $\log_{10} \nu$ on the number $n$ and the value of parameter $b$ of method with function $ae_t$ as $\varepsilon = 10^{-4}$ (a), $\varepsilon = 10^{-6}$ (b), $\varepsilon = 10^{-8}$ (c), $\varepsilon = 10^{-10}$ (d)](image)

The error was computed in a similar way as before, e.g. for $ae_t$ and $\varepsilon = 10^{-8}$ parameter $b$ goes over all integer values from 0 to 100. Maximal convergence rate is provided by such a value $b = b^*$ that specifies maximal slope of a curve laying in section of the represented surface by a plain $b = b^* = \text{const}(or p = \text{const}, or \mu = \text{const})$. For $ae_t$ and $\varepsilon = 10^{-8} b^* \in [60; 80]$. This segment is marked with circle on the graph of fig. 6.

5.2 Approximation of solution of inhomogeneous problem (5)

Let us consider a function $f(x)$ that is solution to (5) with $g(x) = \sin(\pi x)$. Here $g(x)$ provides a perturbation in inner points of domain of problem. A question is how such perturbations can affect convergence of proposed methods.

\[
f(x) = C_1 e^{A(\frac{1}{2}x + \frac{1}{2})} + C_2 e^{-A(\frac{1}{2}x + \frac{1}{2})} + \sin(\pi x),
\]

\[
C_1 = \frac{1 + e^{-A}}{e^{-A} - e^A}, \quad C_2 = \frac{1 + e^A}{e^A - e^{-A}}, \quad A = \sqrt{1/\varepsilon}.
\]
Let $\nu$ be maximal values of deviation of approximation (8) from the values of function (23). $\nu$ was computed in numerical experiments by enumeration over points $z_1, \ldots, z_K$ as it was described, for different types of $\alpha(y)$ and various values of $n$ and $\varepsilon$: $\varepsilon = 10^{-3}, \ldots, 10^{-10}$. The values of parameters $p, b, \mu$ were taken from Table 2 excluding cases of small $n$. Numerical experiments showed that the smaller $n$ is, the stronger appears the dependence of convergence rate on $n$.

While using functions $\alpha_s, \alpha_p, \alpha_e$ approximation errors for (23) agree with those obtained for $f(x) = \xi(x)$ (see fig 3, 4). "Optimal" values of parameters $p$ and $\mu$ in these cases retain too. However the approximation based on $\alpha_t$ loose its convergence rate much faster. This effect is explained by inequality for $b$ (see estimate (14)) that turns out to be more essential than, for example, (18). Even for moderate values of $\|f^{(s+1)}\|_D$ (in our case $\forall s \|f^{(s+1)}\|_D \approx 1$) one obtains that value of $b$ should be considerably less than those given in Table 2. Then, according to (15), $M(n)$ becomes large and the convergence rate decreases.

In order to obtain efficient approximation with function $\alpha_t$ a coupled basis $B_k(x) \cup T_m(x)$ ($k = 0, \ldots, K$, $m = 0, \ldots, M$, $K + M = n - 2$) should be used. It is composed of designed functions $B_k(x)$ and Chebyshev polynomials $T_m(x)$:

$$f(x) \approx \sum_{k=0}^{K} a_k B_k(x) + \sum_{m=0}^{M} c_m T_m(x),$$

$$T_m(x) = \cos(m \arccos(x)), \ B_k(x) = \cos \left(k \arccos \left[ \tan(\frac{x-b}{b}) \right] \right).$$
The idea of a method consists in two steps: first function $f(x)$ should be expand in basis $T_m(x)$ and then difference $f(x) - \sum_{m=0}^{M} c_m T_m(x)$ should be approximated using $B_k(x)$. Let us compose matrices

$$
\mathcal{T} = (\gamma_{mi}), \quad \gamma_{mi} = T_m(x_i), \quad \text{where } x_i = \cos \left[ \frac{(2i + 1)\pi}{2(M + 1)} \right], \quad m, i = 0, \ldots, M;
$$

$$
\mathfrak{B} = (b_{kj}), \quad b_{kj} = B_k(x_j), \quad \text{where } x_j = \left( \arctan \left[ b \cos \left( \frac{(2j + 1)\pi}{2(K + 1)} \right) \right] \right) / b,
$$

$k, j = 0, \ldots, K$.

Introducing the notations of vectors $\alpha = (a_0, \ldots, a_K)$, $\gamma = (c_0, \ldots, c_M)$, $f_\alpha = (f(x_0), \ldots, f(x_K))$, $f_\gamma = (f(x_0), \ldots, f(x_M))$ one can found the coefficients of expansion in Chebysev basis $\gamma = \mathbf{T}^{-1}f_\gamma$. Further the following SLAE was composed

$$
\mathfrak{B}\alpha = f_\alpha - C\gamma,
$$

$$
C_{jm} = T_m(x_j) = \cos \left( m \arccos \left[ \arctan \left[ b \cos \left( \frac{(2j + 1)\pi}{2(K + 1)} \right) \right] \right] / b \right).
$$

and the values of coefficients $a_0, \ldots, a_K$ were obtained

$$
\alpha = \mathfrak{B}^{-1} [f_\alpha - C\mathbf{T}^{-1}f_\gamma]. \quad (25)
$$

The computations by formula (25) can be modified. So, for small parameters $\varepsilon = 10^{-4}, \varepsilon = 10^{-6}$ the domain of distribution of Chebyshev nodes in (24) was narrowed from segment $[-1; 1]$ to $[-0.85; 0.85]$ in case $\varepsilon = 10^{-4}$ and to $[-0.95; 0.95]$ in case $\varepsilon = 10^{-6}$.

Tables 3–5 show the value of approximation error $\nu$ obtained for function (23) using Chebyshev basis and the designed ones $B_k$ with functions $\varphi(y)$ of four considered types. If value of parameters of $\varphi(y)$ differs from those given in Table 2 than it is explicitly indicated in brackets. So does a number of basis functions $B_k(x)$ plus number of basis polynomials $T_m(x)$ for approximations obtained using $\varphi_t$, see (24).

### Table 3. Values of $\nu$ obtained for function (23) with $\varepsilon = 10^{-6}$

<table>
<thead>
<tr>
<th>$n$</th>
<th>Chebyshev</th>
<th>sin($y$)</th>
<th>pol($y$)</th>
<th>exp($y$)</th>
<th>tan($y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.997</td>
<td>0.168</td>
<td>0.208</td>
<td>0.069(5.5)</td>
<td>0.036(60, 7 + 3)</td>
</tr>
<tr>
<td>20</td>
<td>0.727</td>
<td>0.047</td>
<td>0.064(1.5)</td>
<td>0.005</td>
<td>5.21 $\times 10^{-4}(80, 14 + 6)$</td>
</tr>
<tr>
<td>30</td>
<td>0.359</td>
<td>0.077</td>
<td>0.025</td>
<td>2.4418 $\times 10^{-5}$</td>
<td>7.935 $\times 10^{-6}(50, 23 + 7)$</td>
</tr>
<tr>
<td>40</td>
<td>0.147</td>
<td>7.659 $\times 10^{-4}$</td>
<td>0.001</td>
<td>3.679 $\times 10^{-7}$</td>
<td>4.283 $\times 10^{-8}(14, 30 + 10)$</td>
</tr>
<tr>
<td>50</td>
<td>0.051</td>
<td>5.419 $\times 10^{-5}$</td>
<td>2.216 $\times 10^{-5}$</td>
<td>1.936 $\times 10^{-9}$</td>
<td>2.637 $\times 10^{-10}(14, 38 + 12)$</td>
</tr>
<tr>
<td>60</td>
<td>0.015</td>
<td>8.019 $\times 10^{-6}$</td>
<td>5.348 $\times 10^{-7}$</td>
<td>3.738 $\times 10^{-11}$</td>
<td>1.436 $\times 10^{-12}(9, 47 + 13)$</td>
</tr>
<tr>
<td>70</td>
<td>0.004</td>
<td>7.348 $\times 10^{-7}$</td>
<td>2.533 $\times 10^{-8}$</td>
<td>5.473 $\times 10^{-13}$</td>
<td>2.625 $\times 10^{-13}(9, 57 + 13)$</td>
</tr>
<tr>
<td>80</td>
<td>7 $\times 10^{-4}$</td>
<td>3.346 $\times 10^{-8}$</td>
<td>7.458 $\times 10^{-10}$</td>
<td>3.321 $\times 10^{-13}$</td>
<td>2.26 $\times 10^{-13}(9, 67 + 13)$</td>
</tr>
<tr>
<td>90</td>
<td>1.2 $\times 10^{-4}$</td>
<td>2.941 $\times 10^{-9}$</td>
<td>1.791 $\times 10^{-11}$</td>
<td>3.375 $\times 10^{-13}$</td>
<td>2.597 $\times 10^{-13}(9, 77 + 13)$</td>
</tr>
</tbody>
</table>
Table 4. Values of $\nu$ obtained for function (23) with $\epsilon = 10^{-8}$

<table>
<thead>
<tr>
<th>$n$</th>
<th>Chebyshev</th>
<th>sin($y$)</th>
<th>pol($y$)</th>
<th>exp($y$)</th>
<th>tan($y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.0000</td>
<td>0.5024</td>
<td>0.5903(1.5)</td>
<td>0.2040(7.8)</td>
<td>0.0463(650, 7 + 3)</td>
</tr>
<tr>
<td>20</td>
<td>1.0000</td>
<td>0.1972</td>
<td>0.6521</td>
<td>0.0366</td>
<td>6.0843 $\times 10^{-4}$(180, 14 + 6)</td>
</tr>
<tr>
<td>30</td>
<td>0.9987</td>
<td>0.1119</td>
<td>0.2743</td>
<td>3.5359 $\times 10^{-4}$</td>
<td>1.2506 $\times 10^{-5}$(140, 23 + 7)</td>
</tr>
<tr>
<td>40</td>
<td>0.9730</td>
<td>0.0208</td>
<td>0.0873</td>
<td>1.0721 $\times 10^{-5}$</td>
<td>5.3592 $\times 10^{-8}$(110, 30 + 10)</td>
</tr>
<tr>
<td>50</td>
<td>0.8920</td>
<td>0.0130</td>
<td>0.0186</td>
<td>3.8726 $\times 10^{-7}$</td>
<td>4.9031 $\times 10^{-10}$(90, 39 + 11)</td>
</tr>
<tr>
<td>60</td>
<td>0.7705</td>
<td>0.0032</td>
<td>0.0020</td>
<td>8.8276 $\times 10^{-9}$</td>
<td>3.4193 $\times 10^{-12}$(90, 47 + 13)</td>
</tr>
<tr>
<td>70</td>
<td>0.6384</td>
<td>0.0011</td>
<td>4.28 $\times 10^{-4}$</td>
<td>4.0243 $\times 10^{-10}$</td>
<td>1.8532 $\times 10^{-12}$(90, 57 + 13)</td>
</tr>
<tr>
<td>80</td>
<td>0.5144</td>
<td>3.7 $\times 10^{-4}$</td>
<td>1.42 $\times 10^{-4}$</td>
<td>9.3578 $\times 10^{-12}$</td>
<td>1.9960 $\times 10^{-12}$(90, 67 + 13)</td>
</tr>
<tr>
<td>90</td>
<td>0.4059</td>
<td>5.7 $\times 10^{-5}$</td>
<td>1.27 $\times 10^{-5}$</td>
<td>4.3484 $\times 10^{-12}$</td>
<td>2.3292 $\times 10^{-12}$(90, 77 + 13)</td>
</tr>
</tbody>
</table>

Table 5. Values of $\nu$ obtained for function (23) with $\epsilon = 10^{-10}$

<table>
<thead>
<tr>
<th>$n$</th>
<th>Chebyshev</th>
<th>sin($y$)</th>
<th>pol($y$)</th>
<th>exp($y$)</th>
<th>tan($y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.0000</td>
<td>1.0005</td>
<td>1.0003</td>
<td>0.3875(10.4)</td>
<td>0.0529(6500, 7 + 3)</td>
</tr>
<tr>
<td>20</td>
<td>1.0000</td>
<td>0.3498</td>
<td>0.9986</td>
<td>0.0391</td>
<td>4.7344 $\times 10^{-4}$(4000, 13 + 7)</td>
</tr>
<tr>
<td>30</td>
<td>1.0000</td>
<td>0.2231</td>
<td>0.9282</td>
<td>0.0027</td>
<td>1.0203 $\times 10^{-5}$(1200, 22 + 8)</td>
</tr>
<tr>
<td>40</td>
<td>1.0000</td>
<td>0.2041</td>
<td>0.7389</td>
<td>2.1276 $\times 10^{-4}$</td>
<td>9.2063 $\times 10^{-8}$(1100, 31 + 9)</td>
</tr>
<tr>
<td>50</td>
<td>1.0000</td>
<td>0.1418</td>
<td>0.5336</td>
<td>1.5681 $\times 10^{-5}$</td>
<td>3.5014 $\times 10^{-10}$(1100, 38 + 12)</td>
</tr>
<tr>
<td>60</td>
<td>1.0000</td>
<td>0.0700</td>
<td>0.3642</td>
<td>1.0989 $\times 10^{-6}$</td>
<td>2.1094 $\times 10^{-11}$(1100, 47 + 13)</td>
</tr>
<tr>
<td>70</td>
<td>1.0000</td>
<td>0.0226</td>
<td>0.2371</td>
<td>7.3964 $\times 10^{-8}$</td>
<td>2.2963 $\times 10^{-11}$(1100, 57 + 13)</td>
</tr>
<tr>
<td>80</td>
<td>0.9999</td>
<td>0.0223</td>
<td>0.1470</td>
<td>4.8155 $\times 10^{-9}$</td>
<td>1.8534 $\times 10^{-11}$(1100, 67 + 13)</td>
</tr>
<tr>
<td>90</td>
<td>0.9994</td>
<td>0.0124</td>
<td>0.0865</td>
<td>3.0489 $\times 10^{-10}$</td>
<td>2.1706 $\times 10^{-11}$(1100, 77 + 13)</td>
</tr>
<tr>
<td>100</td>
<td>0.9973</td>
<td>0.0041</td>
<td>0.0481</td>
<td>4.0388 $\times 10^{-11}$</td>
<td>2.1034 $\times 10^{-11}$(1100, 87 + 13)</td>
</tr>
</tbody>
</table>

6 Conclusion

In this work a method for approximating smooth functions having boundary layer components was developed, justified and implemented. It is based on expansion of function into series with basis consisting of non-polynomial functions obtained from trigonometric Fourier one by special mapping operation. Such representation ensures the estimations of accuracy of best polynomial approximations, but essentially reduces the values of coefficients in them. That is why convergence can be observed starting from small number of basis elements. Moreover the proposed approximations have good properties of numerical stability inherent to Fourier expansions.

Further development and successful application of the method is connected with analysis of different forms of $\alpha$-function. Proposed method can be modified for approximation of functions having singularity in inner point of domain, or even for problems with unknown position of singularity. From the other side, combination of these approximations with collocation methods will allow to design efficient algorithms for solving singularly-perturbed boundary value problems for differential equations.
Acknowledgments. This work was supported by Integrated program of basic scientific research of SB RAS No. 24, project II.2 "Design of computational technologies for calculation and optimal design of hybrid composite thin-walled structures".

References

Methods for Optimal Control of Hydraulic Fracturing Process

Yurii Shokin, Sergey Cherny, Vasily Lapin, Denis Esipov, Dmitriy Kuranakov, and Anna Astrakova

Institute of Computational Technologies
Siberian Branch of Russian Academy of Sciences
Academician Lavrentjev ave., 6, 630090, Novosibirsk, Russian Federation
Tel.: +7-383-330-73-73, cher@ict.nsc.ru

Abstract. The problem of optimal control of the hydraulic fracturing process is set as the optimization problem of finding the input parameters vector for the fracture propagation model that provides the minimization of objective functions. Various objective functions based on the output data of the fracture propagation model are considered. The problem is formulated within the framework of multivariate and multiobjective optimization method, which is based on the combined features of fracture propagation model and Genetic Algorithm. Plane radial model of fracture propagation caused by Herschel-Bulkley fluid injection is used to establish relationships between input parameters and fracture growth. The potential of the proposed method for control of hydraulic fracturing process is demonstrated by application to hydrocarbon reservoirs of shallow bedding. Results show that the proposed methods for optimal control of hydraulic fracturing process play the important role in maximization of the volume of mined hydrocarbon with significant decrease of the costs for hydraulic fracturing execution.

1 Introduction

The goal of the optimization of the hydraulic fracturing process is the maximization of gas and oil production by maximization of the fractured reservoir rock volume. The optimization of hydraulic fracturing is based on the model of fracture propagation in elastic media caused by viscous fluid injection. Input parameters of fracture propagation model are: the surface of the cavity in infinite elastic media; pumping pressure of fluid that causes fracture initiation and propagation (or pumping schedule for fluid of given rheology); elastic media parameters. Output characteristics of the model are: the fracture surface; fracture width distribution; speed of fracture front propagation. Calculation of output characteristics using the input parameters is the direct problem of fracture propagation. By solving it one can predict the geometry of forming fracture, volume of hydrocarbon mined from the fracture, calculate costs of this process, etc. The inverse problem is to find the vector of input parameters of fracture propagation model that satisfies the given performance criteria of fracture formation and propagation processes. Optimal control of hydraulic fracturing process
consists in the solution of the inverse problem. In this problem it is required to find the parameters of rheological laws for fluid, pumping schedule (e.g. time dependence of fluid injection rate), conditions of fracture initiation (geometry of the cavity, its orientation against in-situ stresses of elastic media) that satisfy the needed location of incipient fracture, linearity of fracture propagation trajectory, uniformity of fracture width distribution along the trajectory, no sharp bends along the fracture trajectory, minimal costs for hydraulic fracturing, maximal volume of mined hydrocarbon.

In [1] the optimization of hydraulic fracturing process is carried out for the PKN model. The model describes straight fracture propagation from the linear source. It is supposed that the fracture is of constant height much lesser than the fracture length. On this assumption the change of the parameters along fracture height is negligible and the rock deformation can be considered as plain strain state independently in each vertical section. In PKN model fracture toughness is not taken into account. It is considered that the hydraulic fracturing fluid fills the whole of the fracture, fracture tip mechanics is not considered. Filtration leakoff to the rock is taken into account. Fracture geometry is considered as a function of the following parameters that influence the hydraulic fracturing process: hydraulic fracturing fluid viscosity $\mu$, injection rate of fracturing fluid $Q_{in}(t)$, injection time $T$, proppant concentration $\delta$ and fracture half-length $R_{frac}$. Fracture height $h_{frac}$ and width $W$ are calculated solving coupling relationships based on fracture geometry and material balance. Optimization problem for hydraulic fracturing process consists in maximization of total production over 10 years with bound and design constraints. To solve the optimization problem the algorithm INTEMOB (INTElligent Moving Object) is used. The main concept of INTEMOB is based on the concepts of Genetic Algorithm, simplex-method and EVOP-algorithm (EVolutionary OPeration).

In [2] the optimization of hydraulic fracturing process is carried out using Pseudo-3D (P-3D) model of fracture propagation. P-3D model differs from PKN model by taking into account variability of the parameters along fracture height that influences fracture width. Problem statement and solution method for optimization problem are the same as in [1]. The results of [1] and [2] show that the suggested method for solving of the stated optimization problems plays the important role in hydraulic fracturing process improvement. Only a 12% compromise with the production over 10 years saves about 44% of the treatment cost.

In [3] the optimization of hydraulic fracturing process using 3D model is presented. Barnett Shale gas field is considered as the object of investigation. It is characterized by 7 rock layers and 4 sets of joints per rock layer. The goal of the optimization of the hydraulic fracturing procedure is maximizing the fractured reservoir rock volume which results in the maximization of gas production. For three dimensional modeling, analysis and post processing FEM simulator ANSYS is used. For fracturing process modeling the program multiPlas is used. Optimization tool optiSLang is used for calibration of more than 200 parameters of the reservoir production simulator. The main result of modeling is 3D
fissured reservoir rock. In the calibration process, the physical parameters are updated until time and location of seismic fracture measurements show reasonable agreement with the simulation results. Correlation analysis of optiSLang identifies the main reservoir parameter to additionally calibrate the mechanism of how the fracturing design parameter effects the fracture growth. Among the wide range of the varied parameters the most important are layer thickness and its location, elastic properties of the rock, hydraulic parameters, bedding plane of the shale, the location and frequency of the fractures in this plane. These parameters are compared (correlated) with reservoir and well test data. The advanced functionality of the software supports taking into account some additional physical effects such as thermal effects. Calibrated model was used to predict the gas production rate. The predicted gas production rate from the calibrated model showed very good agreement to the real production rate and much better agreement than the estimated production rates with the help of seismic fracture measurements only. With help of the optimization design an increase of gas production of 25 % was possible with just an optimized well position in the reservoir.

In [4] optimization pump flow rates scenarios has been performed for the particular wellbore situated in Iranian Sand Stone Reservoir. Pump flow rates scenario is characterized by two parameters in this study: pump rate that is assumed to be constant and time of pumping that is replaced by the fracture half length. These two parameters are varied to maximize wellbore productivity over a period of one year. The parameters are varied independently and only one parametric problem of minimization arises. So no optimization algorithm is needed. But the feature of the paper is the fact that a few different models have been applied: pseudo three dimensional model, PKN, KGD and radial ones.

In [5] 2-D hydraulic fracturing model GEOS-2D, is used to simulate dynamic fracture propagation within a pre-existing facture network. Instead of integrating physical models and economic models to maximize net present value as the objective function, or estimating the total production of the wellbore over any time period the authors of [5] focus on physical criteria, that is, the optimal hydraulic fracture propagation under uncertain natural conditions. The fractal dimension of the connected fractures can be derived from the post-fracturing network simulated by GEOS-2D to represent the network density and connectivity. Therefore, the fractal dimension is chosen as the objective function to optimize the hydraulic fracturing well design. BOBYQA, a derivative-free nonlinear optimization algorithm, is applied in [5] to drive a global search on the modeled response surface.

In the present paper the problem of optimal control of the hydraulic fracturing process is formulated as the optimization problem of finding the input parameters vector for the fracture propagation model that provides the minimization of several objective functions. Various objective functions based on the output data of the fracture propagation model are considered. The problem is formulated within the framework of multivariate and multiobjective optimization method, which is based on the combined features of fracture propagation
model and Genetic Algorithm. Plane radial model of fracture propagation caused by Herschel-Bulkley fluid injection is used to establish relationships between input parameters and fracture growth. The capability of the proposed method for control of hydraulic fracturing process is demonstrated by application to hydrocarbon reservoirs of shallow bedding. Results show that the proposed methods for optimal control of hydraulic fracturing process play the important role in maximization of the volume of mined hydrocarbon with significant decrease of the costs for hydraulic fracturing execution.

2 Direct problem of hydraulic fracture propagation

The construction of the methods for optimal control of hydraulic fracturing is carried out using the plane radial model of fracture propagation caused by Herschel-Bulkley fluid injection.

2.1 Radial or penny-shaped model for fluid-proppant slurry

Hydraulic fracture propagation is simulated by the classical penny shaped model [6,7,8] improved by more complex model of slurry flow inside the fracture. Newtonian fluid model is replaced by more general Herschel-Bulkley model [9] and the variation of proppant concentration is taken into account by convection equation added into the model. The geometrical concept of the penny-shaped fracture model is presented in Fig. 1. Rock deformation under axisymmetric pressure load in fracture is given by the integral relation

\[
W(r) = \frac{8}{\pi E'} \int_{R_{frac}}^{R_{frac}} \left( \int_0^\zeta \frac{p_{net}(\xi)\xi}{\sqrt{\xi^2 - \xi^2 \sqrt{\xi^2 - r^2}}} d\xi \right) d\zeta, \quad E' = \frac{E}{1 - \nu^2}, \quad p_{net} = p - \sigma_{min},
\]

where \( p \) is absolute pressure of the slurry of fracturing fluid and proppant flow, \( R_{frac} = R_{frac}(t) \) is fracture front position defined from the well-known criterion of brittle fracture propagation

\[
K_I = \frac{2}{\sqrt{\pi R_{frac}^2}} \int_0^{R_{frac}} \frac{p_{net}(\xi)\xi}{\sqrt{R_{frac}^2 - \xi^2}} d\xi = K_{Ic}.
\]

Slurry flow in the fracture is described by continuity equation of fluid phase

\[
\frac{\partial (rW\alpha)}{\partial t} + \frac{1}{2\pi} \frac{\partial (Q\alpha)}{\partial r} + \frac{1}{2\pi} \alpha Q_L(r,t) = 0,
\]

where

\[
Q = 2\pi r Wu, \quad Q_L(r,t) = \frac{4\pi r C_L}{\sqrt{t - t_{exp}(r)}},
\]
continuity equation of proppant phase
\[
\frac{\partial (r W \delta)}{\partial t} + \frac{1}{2\pi} \frac{\partial (Q \delta)}{\partial r} = 0,
\] (5)
and the momentum equation of slurry flow
\[
\frac{\partial p_{\text{net}}}{\partial r} = -2K \left( \frac{2n+1}{\pi n} \right)^n \frac{Q^n}{W^{2n+1} r^n} + \left( \frac{4n+2}{n+1} \right) \frac{\tau_0}{W},
\] (6)
where $K$, $n$ and $\tau_0$ are slurry parameters. In the considered model the rheology of Herschel-Bulkley fluid is taken into account. In Herschel-Bulkley fluid shear stress $\tau$ is expressed from yield stress $\tau_0$, power law fluid rheology consistency coefficient $K$, power law fluid rheology behavior index $n$ and the shear rate $\dot{\gamma}$ by the formula
\[
\tau = \tau_0 + K \dot{\gamma}^n.
\] (7)
Volume concentrations of the fluid $\alpha$ and the proppant $\delta$ for all possible $r$ and $t$ are connected by the relation
\[
\alpha(r,t) + \delta(r,t) = 1.
\] (8)
The slurry viscosity is determined from the proppant concentration using Maron-Pierce [10] relationship
\[
K(\delta) = K(0) \left( 1 - \frac{\delta}{\delta^*} \right)^{-2},
\] (9)
where $\delta^*$ is the critical concentration of proppant. According to the experimental study performed by Mueller [11], this relationship can be used to calculate viscosity of suspensions of solid particles. Yield stress $\tau_0$ and behaviour index $n$ correspond to the used hydraulic fracture fluid.
On the wellbore the boundary condition for the slurry pumping rate
\[
Q(R_w,t) = Q_{\text{in}}(t)
\] (10)
and for one of the two concentrations, e.g. for proppant,
\[
\delta(R_w,t) = \delta_{\text{in}}(t)
\] (11)
is set. The fluid front $R_{\text{fluid}}$ is supposed to lag the fracture front $R_{\text{frac}}$
\[
R_{\text{frac}} - R_{\text{fluid}} > 0.
\] (12)
On fluid front $R_{\text{fluid}}$ Stefan condition
\[
\frac{\partial R_{\text{fluid}}}{\partial t} = u(R_{\text{fluid}},t) = \frac{Q(R_{\text{fluid}})}{2\pi R_{\text{fluid}} W(R_{\text{fluid}})}
\] (13)
and the condition of zero absolute pressure $p$
\[
p_{\text{net}}(R_{\text{fluid}},t) = -\sigma_{\text{min}}
\] (14)
are set. Net pressure $p_{\text{net}}$ on the interval from fluid front $R_{\text{fluid}}$ to fracture front $R_{\text{frac}}$ is also supposed to be $-\sigma_{\text{min}}$. From (3) taking into account (10) one can derive the equation of slurry balance at any time moment $T$

$$
\int_{0}^{T} Q_{\text{in}}(t) dt = 2\pi \int_{R_{\text{w}}}^{R_{\text{fluid}}(T)} rW(r,T) dr + 2\pi \int_{0}^{T} \int_{R_{\text{w}}}^{R_{\text{fluid}}(T)} rQ_{L}(r,t) dr dt. \hspace{1cm} (15)
$$

Initial data is set

$$
R_{\text{frac}}(0) = R_{0}, \quad R_{\text{fluid}}(0) = R_{0}, \quad W(r,0) = 0, \quad R_{w} \leq r \leq R_{0}. \hspace{1cm} (16)
$$

2.2 Direct problem solution

The hydraulic fracturing process lasts $T$ seconds. In Fig. 2 the flowchart of numerical solution of incorporated to the model subproblems is presented. Let it be the fracture with front $R_{n-\text{frac}}^{n}$ at the timestep $n$. Fracture front at $(n+1)$ timestep is defined by the increment $\Delta R_{\text{frac}}$ to fracture front $R_{\text{frac}}^{n}$. Fracture increment $\Delta R_{\text{frac}}$ has the fixed value for all timesteps of the problem. The fluid front position is set $R_{n-\text{fluid}}^{n} < R_{n-\text{frac}}^{n}$. For the given fluid front position the equations (1), (2), (3), (4), (5), (10), (11), (13) of the “hydrodynamics-elasticity” problem are solved numerically. After the convergence of pressure $p_{\text{net}}$ distribution is achieved fulfillment of the condition (14) is checked. If this condition is not fulfilled the fluid front position is corrected and the “hydrodynamics-elasticity” problem is solved again. The process continues until the pressure on fluid front reaches the value $-\sigma_{\text{min}}$. Note that the time needed for the given fracture increment $\Delta R_{\text{frac}}$ is calculated from (13).
Fig. 2: The flowchart of direct problem solution.
2.3 Production model

The primary goal of hydraulic fracturing is to increase the productivity of a well by superimposing a highly conductive structure onto the formation. One of the objective functions that are maximized in present paper is productivity index \( \text{PI} \) \( J \). This index comes from the linear relation between the production rate \( Q_{\text{oil}} \) and the driving force (pressure drawdown) \( \Delta p \) [12]

\[
Q_{\text{oil}} = J \Delta p.
\] (17)

To find PI the problem of non-stationary oil filtration from the reservoir to the fracture and through the fracture to the wellbore is solved. The scheme of this problem is presented in Fig. 3.

**Filtration of the fluid in the fracture filled by proppant**

The process of non-stationary oil filtration in plane radial fracture with faces closed on the proppant at the moment \( t_p \) and then having the fracture opening distribution

\[
W_p(r) = \delta(r, t_p)W(r, t_p),
\] (18)

is described by the convection-diffusion partial differential equation

\[
\frac{\partial p_p}{\partial t} - \frac{k_p}{\beta^* W_p \mu r} \frac{\partial}{\partial r} \left( r W_p \frac{\partial p_p}{\partial r} \right) + \frac{2w_{\text{res}}}{\beta^* W_p} = 0
\] (19)

and the equation of Darcy’s law

\[
w_p = -\frac{k_p}{\mu} \frac{\partial p_p}{\partial r}.
\] (20)

In (19), (20) \( p_p(r, t) \) is pressure in the fracture; \( k_p \) is fracture permeability; \( w_p \) is Darcy flux in the fracture; \( w_{\text{res}} \) is Darcy flux of the oil from the reservoir to the fracture; total system compressibility \( \beta^* \) is defined as

\[
\beta^* = m_p (\beta_{\text{oil}} + \beta_p),
\] (21)

where \( m_p \) is the porosity of the proppant in the fracture, \( \beta_{\text{oil}} \) and \( \beta_p \) are the coefficients of oil and proppant compressibility correspondingly. It is supposed that oil is a newtonian fluid with the coefficient of dynamic viscosity \( \mu \).

The following boundary conditions are set on the wellbore

\[
2\pi R_w \frac{k_p}{\mu} \frac{\partial p_p}{\partial r} \bigg|_{r=R_w} = Q_{\text{oil}}(t),
\] (22)

and at the fracture front

\[
p_p(R_{\text{frac}}, t) = 0.
\] (23)

Here \( Q_{\text{oil}}(t) \) is debit of the wellbore. The initial condition for (19) is established

\[
p_p(r, 0) = 0, \quad R_w \leq r \leq R_{\text{frac}}.
\] (24)
Oil filtration from the reservoir to the fracture

The process of non-stationary oil filtration from the reservoir to the fracture is described by the convection-diffusion equation [13], [14]

$$\frac{\partial p_{\text{res}}}{\partial t} - \frac{k_{\text{res}}}{\beta^* \mu} \frac{\partial^2 p_{\text{res}}}{\partial z^2} = 0$$ (25)

and the equation of Darcy’s law

$$w_{\text{res}} = -\frac{k_{\text{res}}}{\mu} \frac{\partial p_{\text{res}}}{\partial z}.$$ (26)

In (25), (26) $p_{\text{res}}(r,z,t)$ is reservoir pressure on the cylindrical surface of radius $r$ (see. Fig. 3); $k_{\text{res}}$ is permeability of the reservoir; $w_{\text{res}}$ is Darcy flux of the oil along the cylinder surface; total compressibility $\beta^*$ is defined by the formula

$$\beta^* = m_{\text{res}} (\beta_{\text{oil}} + \beta_{\text{res}}),$$ (27)

where $m_{\text{res}}$ is reservoir porosity, $\beta_{\text{res}}$ is reservoir compressibility.

At the distance $z_{\text{res}}$ from the fracture the condition is set

$$p_{\text{res}}(r,z_{\text{res}},t) = 0,$$ (28)

and at the fracture the following condition is set

$$p_{\text{res}}(r,0,t) = p_p(r,t).$$ (29)

Initial data for (25) is established

$$p_{\text{res}}(r,z,0) = 0, \quad 0 \leq z \leq z_{\text{res}}.$$ (30)
Coupled solution of the production model equations

The stated two subproblems of the production model are solved iteratively by means of interchange of parameters $w_{\text{res}}(r,0,t)$ and $p_p(r,t)$. Differential equations (19) and (25) are solved numerically using the finite difference methods.

3 Inverse problem of hydraulic fracture propagation

3.1 Problem statement

Variables

*Penny-shaped fracture model parameters*

The vector of free design variables $x$ (which have been independently varied in this study to find an optimum design) includes injection rate of fracturing fluid $Q_{\text{in}}(t)$, injection proppant concentration (volume fraction) $\delta_{\text{in}}(t)$, fracturing fluid parameters $K$, $n$, $\tau_0$, the Young’s modulus $E$ and the Poisson’s ratio $\nu$ of the rock, the Carter’s leak-off coefficient $C_L$. In that way

$$x = (Q_{\text{in}}(t), \delta_{\text{in}}(t), K, n, \tau_0, E, \nu, C_L). \quad (31)$$

*Production model parameters*

The vector of free variables $y$ includes debit of the wellbore $Q_{\text{oil}}(t)$, oil viscosity $\mu$, fracture $k_p$ and reservoir $k_{\text{res}}$ permeability, fracture $m_p$ and reservoir $m_{\text{res}}$ porosity, compressibility coefficients of proppant $\beta_p$, oil $\beta_{\text{oil}}$ and reservoir $\beta_{\text{res}}$. In that way

$$y = (Q_{\text{oil}}(t), \mu, k_p, k_{\text{res}}, m_p, m_{\text{res}}, \beta_p, \beta_{\text{oil}}, \beta_{\text{res}}). \quad (32)$$
**Constraints**

**Bound constraints**

The design variables are constrained within lower and upper bounds as follows:

\[ x^l_i \leq x_i \leq x^u_i, \quad i = 1, \ldots, N. \] (33)

**Design constraints**

Design constraints are formulated to prevent uncontrolled fracture growth, multiple secondary fracture initiation, excessive fluid loss, to ensure that the designed treatment program can be executed in the field using specified surface equipment and downhole tubing, to ensure adequate fracture width, fluid efficiency and desired geometric proportions. Design constraints are stated as the inequalities:

\[ \phi_k(x) \leq 0, \quad k = 1, \ldots, K. \] (34)

For example, to constrain below the minimal fluid discharge to the fracture in (34) the following function is assigned

\[ \phi_1(x) = Q_{in}^{\min} - \min_t Q_{in}(t). \] (35)

**Objective functions**

The objective of the hydraulic fracturing is the increase of reservoir productivity. Therefore in optimization problem it is reasonable to maximize the cumulative production over time \( T_{prod} \) years by means of minimization of the major design objective

\[ F_1(x, y) = -\int_0^{T_{prod}} Q_{oil}(t) dt. \] (36)

To build the functional (36) one should solve all the problems formulated above: fracture propagation during the time \( T \) caused by the pumping of the fluid-proppant mixture; non-stationary oil filtration from the reservoir to the fracture closed on the proppant and through the fracture to the wellbore during the period of time \( 0 \leq t \leq T_{prod} \). It is quite complicated even if the considered 1D problem statement is used. Therefore some simplifications are implemented. E.g., in [1], [2] instead of solving the filtration problem numerically its approximate closed-form solutions are used for transient period [13], [14] and pseudo-steady-state [15], derived with the assumption of the constant pressure in the wellbore and using the empirical relations between flow rate and pressure gradient in the near-wellbore zone. In practice even simpler formulation is often used. E.g., according to [12] in terms of productivity the optimal fracture closed on the proppant can be considered the one for which the following relation is fulfilled

\[ \frac{W_p(0)}{R_{frac}(T)} = 1.6 \frac{k_{res}}{k_p}. \] (37)
Then the maximization of cumulative production during the time $T_{\text{prod}}$ can be carried out by means of the minimization of the functional

$$F_2(x) = \left| \frac{W_p(0)}{R_{\text{frac}}(T)} - 1.6 \frac{k_{\text{res}}}{k_p} \right|,$$

for which one need to solve just the problem of fracture propagation during the time $T$ caused by the pumping of the fluid-proppant mixture.

As far as treatment cost of hydraulic fracturing is quite high, it also should be considered during design work. One of the ways of reduction of the treatment cost is the increase of fluid efficiency, i.e. the relation between the final fracture volume and the volume of the fluid pumped into the fracture. The maximization of fluid efficiency corresponds to minimization of the fluid leakoff total volume that is achieved by using the functional

$$F_3(x) = \int_0^T \int_{R_w} R_{\text{fluid}}(t) Q_L(r,t) dr dt.$$  (39)

For low-permeable reservoirs with low value of $k_{\text{res}}$ in (37), such as shale, it is urgent to create the fractures of large radius. The problem of radius maximization can be reformulated as the problem of minimization of fracture width after the pumping stops

$$F_4(x) = W(R_w, T),$$  (40)

and the problem of average fracture front velocity maximization can be reformulated as the problem of averaged fracture width opening minimization.

$$F_5(x) = \frac{1}{T} \int_0^T W(R_w, t) dt.$$  (41)

To demonstrate the possibilities developed methods for multiobjective design optimization of hydraulic fracturing in this study these functionals are considered.

**General mathematical problem statement**

It is necessary to find the parameter vector

$$x = (x_1, \ldots, x_N) \in X,$$  (42)

providing the minimums for the functionals

$$\min_{x \in X} F_1(x), \ldots, \min_{x \in X} F_M(x)$$  (43)

in the presence of bound (33) and design (34) constraints. In (43) the numeration of the functionals is not linked with the numeration of the particular functionals defined above. It is supposed to consider $M \geq 1$ abstract functionals.
3.2 Solution method

Since in general case it is impossible to find the vector \( \mathbf{x} \) minimizing two or more functionals at the same time, the solution of the problem is Pareto front. In [1], [2] building Pareto front is carried out by solving the series of one-objective optimization problems with freezing one of the functionals by means of the design constraint. The technique used in our paper allows building Pareto front directly. In case of multi-objective optimization Pareto front allows choosing compromise between several performance criteria. Optimization problem (42), (43), (33), (34) was solved by Genetic Algorithm that was used earlier by the authors for multi-objective shape optimization of hydraulic turbines [16].

![Graph](image.png)

Fig. 4: Fracturing fluid discharge laws for \( T = 300 \) s: \( Q_{\text{in}}(t) = 0.1 \text{ m}^3 / \text{s} \) (1); (44), (45) with \( Q_{\text{in}}^0 = 0.022 \text{ m}^3 / \text{s} \) and \( Q_{\text{in}}^T = 0.350 \text{ m}^3 / \text{s} \), obtained from the solution of 4.1 (2).

3.3 Fracturing fluid discharge law

For fracturing fluid discharge law variation during the time interval from 0 to \( T \) we used its representation in the form of second order polynomial

\[
Q_{\text{in}}(t) = at^2 + bt + c, \quad 0 \leq t \leq T,
\]

which coefficients are found from the conditions

\[
Q_{\text{in}}(0) = Q_{\text{in}}^0, \quad Q_{\text{in}}(T) = Q_{\text{in}}^T, \quad \int_0^T Q_{\text{in}}(t) \, dt = Q_{\text{in}}^* T
\]

and have the following representation

\[
a = (-6Q_{\text{in}}^* + 3Q_{\text{in}}^T + 3Q_{\text{in}}^0) / T^2, \quad b = (6Q_{\text{in}}^* - 2Q_{\text{in}}^T - 4Q_{\text{in}}^0) / T, \quad c = Q_{\text{in}}^0.
\]
The parameter $Q_{\text{in}}^*$ in (45) defines the volume of the pumped hydraulic fracturing fluid at constant injection rate $Q_{\text{in}}(t)$ and is set as non-varied parameter. Parameters $Q_{\text{in}}^0$ and $Q_{\text{in}}^T$ are the first components of the vector $x$: $x_1$ and $x_2$ correspondingly. In Fig. 4 two fracturing fluid discharge laws at $T = 300\,s$ are presented: $Q_{\text{in}}(t) = Q_{\text{in}}^*$ with $Q_{\text{in}}^* = 0.1\,m^3/s$ and (44), (45) with $Q_{\text{in}}^0 = 0.022$ and $Q_{\text{in}}^T = 0.35$, obtained from the solution of the problem 4.2 with bound constraints $0.02 \leq x_1 \leq 0.1$, $0.1 \leq x_1 \leq 0.5$.

4 Results

4.1 One-objective optimization

In is necessary to find vector

$$x = (x_1, x_2) = (Q_{\text{in}}^0, Q_{\text{in}}^T),$$

providing the minimum for the functional

$$\min_{x \in X} F_3(x)$$

with bound

$$0.02 \leq x_1 \leq 0.1, \quad 0.1 \leq x_2 \leq 0.5$$

and design

$$\phi_1(x) = 0.002 - \min_t Q_{\text{in}}(t) \leq 0$$

constraints. The values of the rest parameters are shown in Tab. 1.

<table>
<thead>
<tr>
<th>Name</th>
<th>Nomenclature</th>
<th>Value</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Young modulus</td>
<td>$E$</td>
<td>20</td>
<td>GPa</td>
</tr>
<tr>
<td>Poisson ratio</td>
<td>$\nu$</td>
<td>0.2</td>
<td>–</td>
</tr>
<tr>
<td>In-situ stress</td>
<td>$\sigma_{\text{min}}$</td>
<td>10</td>
<td>MPa</td>
</tr>
<tr>
<td>Fracture toughness</td>
<td>$K_{Ic}$</td>
<td>1</td>
<td>MPa $\cdot \sqrt{m}$</td>
</tr>
<tr>
<td>Consistency coefficient</td>
<td>$K$</td>
<td>1</td>
<td>Pa $\cdot s^n$</td>
</tr>
<tr>
<td>Flow index</td>
<td>$n$</td>
<td>1</td>
<td>–</td>
</tr>
<tr>
<td>Yield stress</td>
<td>$\tau$</td>
<td>0</td>
<td>MPa</td>
</tr>
<tr>
<td>Carter leakoff coeff.</td>
<td>$C_L$</td>
<td>$10^{-4}$</td>
<td>m $/ \sqrt{s}$</td>
</tr>
<tr>
<td>Wellbore radius</td>
<td>$R_w$</td>
<td>0.1</td>
<td>m</td>
</tr>
<tr>
<td>Pumping period</td>
<td>$T$</td>
<td>300</td>
<td>s</td>
</tr>
<tr>
<td>Average discharge</td>
<td>$Q_{\text{in}}^*$</td>
<td>0.1</td>
<td>m$^3/s$</td>
</tr>
</tbody>
</table>
The solution of the problem is the vector

$$\mathbf{x} = (0.022, 0.35),$$

(51)

providing minimum $F_3$ equal to 9.5 m$^3$ and corresponding to fracturing fluid discharge law shown in Fig. 4 under number 2. If one uses four parameters

$$\mathbf{x} = (x_1, x_2, x_3, x_4) = (Q_{in}^0, Q_{in}^T, K, n)$$

(52)

instead of two (47) with bound constraints

$$0.02 \leq x_1 \leq 0.1, \quad 0.1 \leq x_2 \leq 0.5, \quad 0.5 \leq x_3 \leq 2, \quad 0.8 \leq x_4 \leq 1$$

(53)

then the better minimal value of $F_3$ equal to 8.4 m$^3$ for the solution vector

$$\mathbf{x} = (0.022, 0.36, 2, 1)$$

(54)

is obtained. It should be noted that the fracturing fluid discharge law is the same for solutions (51) and 54. Also there is insignificant difference between the dependences $R_{frac}(t)$ and $W(r, T)$ in the solutions of two- and four-parameter problems shown in Fig. 5.

If one minimizes $F_4$ instead of $F_3$ in (48)

$$\min_{\mathbf{x} \in \mathbf{X}} F_4(\mathbf{x})$$

(55)

with the same constraints then two-parameter problem statement gives $\mathbf{x} = (0.1, 0.1)$ with minimal value of $F_4$ equal to 8.4 mm, and four-parameter statement gives $\mathbf{x} = (0.1, 0.1, 0.5, 0.8)$ with minimal value of $F_4$ equal to 5.4 mm. The dependencies of the direct problem solution for this vectors $\mathbf{x}$ are shown in Fig. 6. Finally, the solutions of minimization problems for $F_5$ are the vectors $\mathbf{x} = (0.022, 0.35)$ and $\mathbf{x} = (0.02, 0.348, 0.53, 0.81)$ in cases of two (47) and four (52) parameter optimization. In the first case the minimum is 6.1 mm, in the second case – 3.9 mm. Corresponding dependencies of the direct problem solution for this $\mathbf{x}$ are presented in Fig. 7.

### 4.2 Two-objective optimization

Let us find the vector (47) providing the minimum for the functionals

$$\min_{\mathbf{x} \in \mathbf{X}} F_3(\mathbf{x}), \quad \min_{\mathbf{x} \in \mathbf{X}} F_4(\mathbf{x})$$

(56)

with bound (49) and design (50) constraints. The solution of this problem is Pareto front presented in Fig. 8 with extreme points 1 and 2 marked. This points correspond to the vectors $\mathbf{x}_1^1 = (0.1, 0.1)$ and $\mathbf{x}_2^2 = (0.022, 0.35)$. In Fig. 9 the solutions of direct problem for this vectors are shown.

Using four parameters (52) instead of two (47) with bound constraints (53) gives Pareto front presented in Fig. 10. Points 1 and 2 on the front correspond to
Fig. 5: Minimization of $F_3$: two-parameter (curves 2) and four-parameter (curves 4).

Fig. 6: Minimization of $F_4$: two-parameter (curves 2) and four-parameter (curves 4).
Fig. 7: Minimization of $F_5$: two-parameter (curves 2) and four-parameter (curves 4)

Fig. 8: Pareto front in two-parameter optimization problem 4.2 with marked points for analysis.
Fig. 9: Distributions $Q_{in}(t)$ (a), $R_{frac}(t)$ (b), $W(R_w, t)$ (c), $W(r, T)$ (d), obtained from the solution of direct problem for $\mathbf{x}^1$ (curves 1) and $\mathbf{x}^2$ (curves 2) solutions of two-parameter optimization problem 4.2.

Fig. 10: Pareto front in four-parameter optimization problem 4.2 with marked points for analysis.
Fig. 11: Distributions $Q_{in}(t)$ (a), $R_{frac}(t)$ (b), $W(R_w, t)$ (c), $W(r, T)$ (d), obtained from the solution of direct problem for $x^1$ (curves 1) and $x^2$ (curves 2) solutions of four-parameter optimization problem 4.2.

Fig. 12: Pareto fronts in two-parameter (solid) and four-parameter (dashed) two-objective optimization problems with extreme points 1, 1’, 2, 2’ correspondingly.
the solution vectors \( \mathbf{x}^1 = (0.1, 0.1, 0.5, 0.8) \) and \( \mathbf{x}^2 = (0.022, 0.36, 2, 1) \). In Fig. 11 the solutions of direct problem obtained with this two vectors are shown.

In Fig. 12 Pareto fronts obtained from the solutions of two- and four-parameter two-objective optimization problems are compared. In tab. 2 the summary data for two- and four-parameter optimization problems is brought together. The parameter vectors \( \mathbf{x} \) and minimal values for the functionals \( F \) from the solutions of one-objective optimization problems and from the extreme points of Pareto fronts are compared.

Note that the extreme points 1 and 2 on Pareto front of the absolute minimum for functionals \( F_4 \) and \( F_3 \) correspondingly give the values of this functionals close to the ones obtained in one-objective optimization problem in which only one functional is minimized while the other is not taken into account. Hence the solutions obtained in 4.1 are the particular cases of the solution of two-objective optimization problem or the extreme points of Pareto front.

### Table 2: Summary data for optimization problems.

<table>
<thead>
<tr>
<th></th>
<th>( \min_{\mathbf{x}} F_3(\mathbf{x}) )</th>
<th>( \min_{\mathbf{x}} F_4(\mathbf{x}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( F )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-obj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-par</td>
<td>9.5</td>
<td>8.4</td>
</tr>
<tr>
<td>4-par</td>
<td>8.4</td>
<td>5.4</td>
</tr>
<tr>
<td>2-obj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-par</td>
<td>9.5</td>
<td>8.4</td>
</tr>
<tr>
<td>4-par</td>
<td>8.4</td>
<td>5.4</td>
</tr>
<tr>
<td>( \mathbf{x} )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-obj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-par</td>
<td>((0.022, 0.35))</td>
<td>((0.1, 0.1))</td>
</tr>
<tr>
<td>4-par</td>
<td>((0.022, 0.36, 2, 1))</td>
<td>((0.1, 0.1, 0.5, 0.8))</td>
</tr>
<tr>
<td>2-obj</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-par</td>
<td>((0.022, 0.35))</td>
<td>((0.1, 0.1))</td>
</tr>
<tr>
<td>4-par</td>
<td>((0.022, 0.36, 2, 1))</td>
<td>((0.1, 0.1, 0.5, 0.8))</td>
</tr>
</tbody>
</table>

### 5 Conclusion

The methods for optimal control of hydraulic fracture are proposed. The procedure consists of the following parts.

- The simulation of hydraulic fracture propagation caused by the pumping of Herschel-Bulkley fluid and proppant slurry.
- The computation of productivity of the fracture filled with proppant based on the combination of the models of plane-parallel and plane-radial filtration for the simultaneous description of fluid filtration in the reservoir and the fracture.
- Genetic algorithm for solution of multi-objective optimization problem and building of Pareto front.
The method capabilities are demonstrated on the proposed functionals that have practical value and are used in non-automatic hydraulic fracturing design. By means of choosing the discharge law of the fluid pumping and its rheological parameters the following problems have been solved.

- Leakoff minimization.
- Fracture width on the wellbore minimization (or maximization of its radius).
- Time-averaged fracture width minimization (or maximization of fracture propagation velocity).

It has been shown that while solving multi-objective optimization problem the obtained Pareto front includes the solutions of one-objective optimization problems as particular cases.

It has been shown that if only fluid discharge law is varied the it is impossible to reduce the fracture width and fluid leakoff simultaneously. But if the variation of fluid rheological parameters is allowed then it is possible to decrease fracture width and leakoff volume simultaneously opposing to the case of fixed fluid rheology.

Acknowledgements

Authors gratefully acknowledge the financial support of this research by the Russian Scientific Fund under Grant number 14-11-00234.

References


Abstract. This paper discusses discrete-continuous (hybrid) systems and corresponding simulation tools. Modern hybrid systems (HS) formalism can be effectively used in problem-oriented environments of computer analysis. One of the many HS applications is the study of transients in electrical power systems (EPS). A special module is developed in ISMA (translated from Russian "Instrumental Facilities of Machine Analysis") simulation environment to support the research of transient processes by Park-Gorev’s equations. Discrete behavior of EPS associated with nonlinear characteristics of generator speed regulators. Also, the EPS operating mode can be changed upon the occurrence of certain events: switching, short circuit, breakage of power lines, etc. Therefore HS methodology is adequate for description and study of transient processes in EPS. The solver of ISMA uses the library of classical and original numerical methods intended for solving systems of differential-algebraic equations with discontinuities. The original algorithm of correct event detection is developed for processing gaps, which is an integral part of numerical analysis.

Keywords: Hybrid systems, modal behavior, transient processes, numerical analysis, library of numerical methods, event detection, principle circuits.

1 Introduction

Hybrid systems (HS) theory is a modern and versatile apparatus for mathematical description of the complex dynamic processes in systems with different physical nature (mechanical, electrical, chemical, biological, etc.). Such systems are characterized by points of discontinuity in the first derivative of the phase variables. HS behavior can be conveniently described as sequential changes of
continuous modes [1], [2], [3], [4]. Each mode is given by a set of differential-algebraic equations with the following constraints:

\[ y' = f(x, y, t), \quad x = \varphi(x, y, t), \]
\[ pr : g(y, t) < 0, \]
\[ t \in [t_0, t_k], x(t_0) = x_0, y(t_0) = y_0, \]

where \( x \in R^{N_x}, y \in R^{N_y}, t \in R, f : R^{N_x} \times R^{N_y} \times R \rightarrow R^{N_y}, g : R^{N_y} \times R \rightarrow R \).

A scalar function \( g(y, t) \) is called event function or guard [1], [4], [5], [6]. The inequality \( g(y, t) < 0 \) means that the phase trajectory in the current state should not cross the border \( g(y, t) = 0 \). Therefore HS state is determined by a predicate \( pr \). The system is in the appropriate state when \( pr = true \). Events occurring in violation of this condition and leading to a transition to a different state without crossing the border, called the one-sided [4], [6]. Such events are of particular practical interest. For example, due to changes in the electrical power systems (EPS) configuration, the operating mode cannot be determined at the time the event occurred.

Models of EPS based on the Park-Gorev equations [7], [8], [9] are traditionally used to describe the electromagnetic and electromechanical processes when studying synchronous operation of generators and solving many other problems in the analysis of electric power systems. Generator mode parameters are written in the rotating coordinate system \( d – q \) associated with respective rotors of electrical machines. Mode parameters of other elements belong to the synchronously rotating coordinate system. The equations for the currents and voltages are written by the laws of Ohm’s and Kirchhoff’s for each of the axes of the coordinate system in accordance with the circuit topology. Thus, equations of electric circuit and its elements completely match the class (1). Therefore, the developed mathematical and instrumental tools for HS simulation can be unified to the electricity problems.

New formalism and methodology for the analysis of complex dynamic systems should be implemented in a problem-oriented environment with plenty of services and techniques for computational experiments. Leading native (RastrWin, ANARES) and foreign (EUROSTAG, DIgSILENT PowerFactory, PSS®E) software systems for the calculation of steady-state and transient processes in EPS implement the traditional models and methods of analysis [9], [10], [11]. Experts in the power industry almost never use modern methodology of hybrid systems. Therefore, the task of developing custom tools with object-oriented interface and input language, a new formalism and original interpretation mechanisms is new and topical.

2 Specification Languages in ISMA

Software for instrumental analysis of HS in ISMA[4], [12], [13] is unified to the problems of different nature: the study of simple dynamic processes, automatic control, chemical kinetics, electrical engineering. Unification to computer analysis problems of transient processes in power systems requires the development
of tools for specification of EPS program models. Fig. 1 shows the architecture of instrumental environment ISMA. Designed architecture allows customizing the environment to a new application with minimal modifications in the organization of interaction of available modules and libraries with object-oriented graphics editor and model interpreter. Modules discussed in this paper are grayed out. The tools provide five different input languages for computer analysis using the methodology of hybrid systems: multi-purpose symbolic (LISMA_PDE)[14] and graphic (Harel charts) languages, as well as thematic language of block diagrams of automatic control systems (ACS) and the equations of chemical kinetics (LISMA+). The language of principal EPS circuits (LISMA_EPS) is also thematic. In the proposed architecture, subject-oriented interfaces interact with the computing core of the system through a universal internal representation of hybrid models. This ensures the continuity of the developed software for new applications with its own characteristics.

Fig. 1. Architecture of ISMA.
3 Test Model of EPS

A test circuit of institute "Energosetproject" [9] is given as an example of solving the electricity problems. Schematic diagram of the closed energy system with two voltage levels and six synchronous machines of different types and powers is presented in Fig. 2. A mathematical model for the calculation of electromechanical transients is built. Synchronous machines are described by Park-Gorev equations in normal form [7], [8], [9]. Generator $G_1$ is a powerful hydroelectric plant. Generators $G_2$ and $G_3$ simulate a thermal power plant with a small power which aggregates operate on power line with different rated voltage 500 and 220 kV. Generators $G_4$ and $G_5$ simulate a thermal power plant with a high power, which aggregates also operate on power lines of different voltage. $G_6$ helps to simulate synchronous compensators mounted to the hub substation.

Here are the equations of main network elements. The equations of synchronous machine $G_i$, $i = 1, \ldots, 6$ have the differential-algebraic form (2) and (3).

\[
\frac{d\Psi_i}{dt} = -U_i - \omega_i \cdot \gamma \cdot \Psi_i,
\]

\[
\frac{d\Psi_{fi}}{dt} = E_{qe i} \cdot \frac{r_{fi}}{x_{adi}} - r_{fi} \cdot i_{fi},
\]

\[
\frac{d\omega_i}{dt} = \frac{1}{T_{ji}} \cdot (M_{Ti} - M_i),
\]
\[
\Psi_{di} = \frac{1}{\omega_{nom}} \cdot x_{di} \cdot i_{di} + \frac{1}{\omega_{nom}} \cdot x_{adi} \cdot i_{fi}, \\
\Psi_{qi} = \frac{1}{\omega_{nom}} \cdot x_{qi} \cdot i_{qi}, \\
\Psi_{fi} = \frac{1}{\omega_{nom}} \cdot x_{fi} \cdot i_{fi} + \frac{1}{\omega_{nom}} \cdot x_{adi} \cdot i_{di}, \\
0 = -M_i + i_{qi} \cdot \Psi_{di} - i_{di} \cdot \Psi_{qi},
\]
(3)

where \( \Psi_i = (\Psi_{di}, \Psi_{qi})^T \) are projections of interlinkage of the stator windings on the axis \( d \) and \( q \), \( U_i = (U_{di}, U_{qi})^T \) are projections of the stator windings voltages, \( I_i = (i_{di}, i_{qi})^T \) are projections of the stator current, \( \omega_i \) is rotor rotation frequency, \( \omega_{nom} \) is a rated frequency, \( \gamma = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \), \( \Psi_{fi} \) is an interlinkage of excitation winding, \( E_{qe1} \) is a electromotive force, \( x_{di} \) and \( x_{qi} \) are synchronous inductances of longitudinal and transverse axis, \( x_{fi} \) is inductive reactance of the excitation winding, \( r_{fi} \) is resistance of the field winding, \( x_{adi} \) and \( x_{aqi} \) are inductances of longitudinal and transverse stator reaction, \( i_{fi} \) is a field winding current, \( T_{fi} \) is inertia constant, \( M_{fi} \) is turbine moment, \( M_i \) is electromagnetic torque. The mutual angles of the generator rotors are defined relative to the generator \( G_1 \) by the formula (4).

\[
\frac{d\delta_{i1}}{dt} = \omega_i - \omega_1, i = 2, \ldots, 6.
\]
(4)

The equation for the block transformer of the generator \( G_1 \) has of the form:

\[
\frac{1}{\omega_{nom}} \cdot \frac{dI_1}{dt} = \frac{U_1}{x_{m1}} - \frac{U_8}{x_{m1}} - \frac{\omega_1}{\omega_{nom}} \cdot \gamma \cdot I_1,
\]
(5)

where \( x_{m1} \) is inductive resistance of the transformer.

The equations for generators 2–6 are written likewise considering numbering and using voltages \( U_{i(1)} \) and currents \( I_{i(1)}, i = 2, \ldots, 6, \) given to coordinate system of the base machine \( G_1 \). The equations of the autotransformers \( T_1–T_3 \) are similar to (5).

The equations for the longitudinal elements of the power line \( P_3 \):

\[
\frac{1}{\omega_{nom}} \cdot \frac{dI_{10}}{dt} = \frac{U_9}{x_{p8-9}} - \frac{U_8}{x_{p8-9}} - \frac{r_{p8-9}}{x_{p8-9}} \cdot \frac{U_8}{\omega_{nom}} \cdot \gamma \cdot I_{10},
\]
(6)

where \( x_{p8-9} \) are inductive resistance of a branch, \( r_{p8-9} \) is active resistance.

Inductive conductivities of power line are equated to conductivities on the start and on the end node. Thus, the node 8 gets the equations (7). For the other power lines program models are written in form (6) and (7).

\[
\frac{1}{\omega_{nom}} \cdot \frac{dU_8}{dt} = x_{c8} \cdot I_{20} - \frac{\omega_1}{\omega_{nom}} \cdot \gamma \cdot U_8.
\]
(7)

449
Loads are given by active-inductive elements and reactor is given only by inductance. For example, the load at the node 7 is given by equation (8).

\[
\frac{1}{\omega_{nom}} \cdot \frac{dI_{15}}{dt} = \frac{U_7}{x_{17}} - \frac{r_{17}}{x_{17}} \cdot I_{15} - \frac{\omega_1}{\omega_{nom}} \cdot \gamma \cdot I_{15}.
\]  

(8)

The complete mathematical model of the EPS contains equations of excitation systems and speed regulators of generators as well as the current balance equations at the nodes and coordinate transformation. Coordinate transformation (9) is necessary for communication of machines currents and voltages at the nodes connecting generators to the coordinate system of the base machine \(G_1\).

\[
I_{i(1)} = (1 \cdot \cos \delta_{i1} + \gamma \cdot \sin \delta_{i1}) \cdot I_i, \\
U_{i(1)} = (1 \cdot \cos \delta_{i1} + \gamma \cdot \sin \delta_{i1}) \cdot U_i, i = 2, \ldots, 6.
\]  

(9)

Synchronous machines \(G_1, G_4, G_5\) and \(G_6\) has installed automatic excitation regulators (AER) of strong action. Generators \(G_2\) and \(G_3\) are equipped with AER of proportional action. As models of turbine regulators applied the ones used in the software package MUSTANG [9]. Speed controller model is described by no more than two differential equations. Speed regulators of turbine of generators \(G_2–G_5\) have a deadband

\[
A_i = \begin{cases} 
0, & \text{for } |\alpha_i| < Z_{Hi}, \\
(|\alpha_i| - Z_{Hi}) \cdot \text{sign}(\alpha_i), & \text{for } |\alpha_i| > Z_{Hi}, 
\end{cases} \quad i = 2, \ldots, 6.
\]  

(10)

Here \(\alpha_i\) is signal to input of the speed control system (in relative units), \(Z_{Hi}\) is deadband, \(A_i\) is movement of the clutch of the centrifugal pendulum. All values are given in relative units.

General mathematical model of the analyzed EPS contains 279 nonlinear differential-algebraic equations. The hybrid behavior of the system under study is due to the equations (10).

4 Graphical Specification

The program model of the test circuit [9] contains 173 lines in LISMA_PDE. This form of presentation is useful when checking the correctness of the mathematical notation and experimenting with various mathematical models of components and the network as a whole. However, for the specialist the great importance has the ability to quickly change the scheme, adding and removing elements and connections, edit the properties of the elements. Editing a text model in this case becomes very time-consuming and increases the likelihood of errors. In such cases, the representation of the problem as a circuit diagram of the EPS is preferred. To do this a graphical editor of principal schemes is developed in ISMA. The interface of the editor is shown in Figures 2 and 3. Fig. 3 shows a power supply circuit of the neighborhood in Novosibirsk [15]. Presentation of computer models in the form of concepts of EPS is more compact and familiar
to user. Tools converting graphical models to universal internal representation of HS hide from the user the process of constructing a mathematical model and allow to quickly getting the results of a computational experiment in the treated and convenient way.

5 Library of Numerical Methods in ISMA

This section is devoted to the integration algorithms of variable order and step based on explicit methods of Runge-Kutta type. The algorithms are applied to numerically solve the Cauchy problem for ODE systems of the following form:

\[ y' = f(y), y(t_0) = y_0, t_0 \leq t \leq t_k. \] (11)

Consideration of autonomous problem does not reduce the generality because non-autonomous problem always can be cast to autonomous by introducing an additional variable. Particular attention should be paid to the choice of the integration method. Fully implicit methods cannot be used because they require the calculation of \( f(y) \) at a potentially dangerous area, where the model is not defined. Therefore here we will use explicit methods with solution: \( y_{n+1} = y_n + h_{n+1} \varphi, n = 0, 1, 2, \ldots \). As a result we obtain the dependence of the predicted integration step \( h_{n+1} \).

Considering that explicit methods are known by poor stability this paper examines integration methods with accuracy and stability control. Generally accuracy and stability control are used to limit the size of the integration step. As a result projected step \( h_{n+1} \) is calculated as follows.

The choice of the next integration step size is based on the proved theorem [16] and can be written as follows:

\[ h_{n+1} = \max[h_n, \min(h^{ac}, h^{st})], \]

where \( h^{ac} \) and \( h^{st} \) are step sizes obtained as a result of accuracy control and stability control respectively. This formula allows to stabilize the step behavior in the area of solution establishing where stability plays a decisive role. Because the presence of this area severely limits the use of explicit methods for solving stiff problems. Suppose that for numerical solution of problem (11) the following implicit methods of Runge-Kutta type is used:

\[ y_{n+1} = y_n + \sum_{i=1}^{m} p_i k_i, k_i = h_n f(y_n + \sum_{j=1}^{i-1} \beta_{ij} k_j), \]

where \( y \) and \( f \) are real \( N \)-dimensional vector-functions, \( h_n \) is an integration step, \( k_i \) are the method stages, \( p_i \) and \( \beta_{ij} \) are numerical coefficients.

Peculiarities of numerical analysis are defined by the configuration and implementation of the solver in the scheme interpreter. Solver is configured to numerical analysis not only of smooth dynamical systems but also systems with ordinary discontinuity and stiff systems [4]. For the analysis of the stiff modes
Fig. 3. Scheme of a power supply of neighborhood.

**Table 1. Library of Numerical Methods**

<table>
<thead>
<tr>
<th>Method (p, m)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISPFF (5, 6)</td>
<td>Stability control, systems of medium and low stiffness</td>
</tr>
<tr>
<td>RADAU5 (3, 3)</td>
<td>Stiff systems</td>
</tr>
<tr>
<td>DISPFF1_RADAU</td>
<td>Adaptive method DISPFF in combination with RADAU5 with stiffness control, essentially stiff systems</td>
</tr>
<tr>
<td>DP78ST (8, 13)</td>
<td>Stability control, variable order and step, systems of medium stiffness and high precision</td>
</tr>
<tr>
<td>RKF78ST (7, 13)</td>
<td>Stability control, variable order and step, systems of medium stiffness and high precision, based on Runge-Kutta-Feldberg method[17]</td>
</tr>
<tr>
<td>RK2ST (2, 2),</td>
<td>Explicit methods with stability control for analysis of non-stiff systems</td>
</tr>
<tr>
<td>RK3ST (2, 3)</td>
<td></td>
</tr>
<tr>
<td>DISPS1</td>
<td>Algorithm of variable order with adaptive stability region</td>
</tr>
<tr>
<td>MK22 (2, 2), MK21 (2, 2)</td>
<td>Freezing of Jacobean matrix, stiff systems</td>
</tr>
<tr>
<td>MK11F</td>
<td>Algorithm of analysis of implicit problems</td>
</tr>
</tbody>
</table>
new original \( m \)-phasic methods of \( p \)-order (Table 1), developed by the authors, are included in the solver library.

DISPF(5, 6), DP78ST(8, 13), RKF78ST(7, 13), RK2ST(2, 2), RK3ST(2, 3), and DISPS1 are integration algorithms based on explicit methods of the Runge-Kutta type [4][16][20]. All of these algorithms have accuracy and stability control and are applied to solving multi-mode problems. RADAU5(3, 3), MK22(2, 2), MK21 (2, 2) are based on fully-implicit schemes which are aimed at solving single-mode systems. The first one is the implicit fifth order three stage Runge-Kutta method, the second and third schemes belong the \((m, k)\)-class of numerical methods [4][21][22]. DISPF1_RADAU is the variable structure algorithm of alternating order and step based on explicit Runge-Kutta numerical formulas of the first, second, and fifth order and the implicit Runge-Kutta type method of the fifth order. This algorithm has accuracy and stability control and can be applied to single-mode as well as multi-mode systems [4]. Finally, MK11F is the algorithm based on the L-stable Rosenbrock method aimed at solving implicit problems [19][23].

Libraries of standard blocks and numerical methods are implemented as independent application modules that are loaded at run time. This approach allows to allocate in the application programming interface (API) a set of functions and classes required for the implementation of element libraries and numerical methods. API is a public interface of the computing module consisted of public classes and interfaces used by other components to interact with the implemented solvers and to create new ones. API classes describes the subject area and declares the type of systems and problems recognized by the solver. Using the API any user with basic knowledge of object-oriented programming able to develop and built in the system new typical elements and numerical methods without recompiling the entire system.

6 Event Detection in Hybrid Systems

The correct analysis of hybrid models is significantly depends on the accuracy of detection [6][18] of the change of the local states of the HS. Therefore, the numerical analysis is necessary to control not only the accuracy and stability of the calculation, but also the dynamics of the event-function. The degree of approximation by the time the event occurred is defined by the behavior of event driven function.

Analyze the behavior of the event function \( g(y, t) \). Let the method of the form \( y_{n+1} = y_n + h_n \phi_n \), where function \( \phi_n \) is calculated in point \( t_n \), is used for calculations. Then the event-function \( g(y, t) \) at point \( t_{n+1} \) has a form \( g_{n+1} = g(y_n + h_n \phi_n, t_n + h_n) \). Decomposing the \( g_{n+1} \) in a Taylor series and taking into account the linearity of \( g_{n+1} \), we obtain the dependence of \( g_{n+1} \) of the projected step \( h_n \):

\[
g_{n+1} = g_n + h_n \left( \frac{\partial g_n}{\partial y} \cdot \phi_n + \frac{\partial g_n}{\partial t} \right). \tag{12}
\]
**Theorem.** The choice of the step according to the formula

\[ h_{n+1} = (\gamma - 1) \frac{g_n}{\partial g_n \partial y \cdot \varphi_n + \frac{\partial g_n}{\partial t}}, \gamma \in (0, 1), \]  

(13)

provides the event-dynamics behavior as a stable linear system, the solution of which is asymptotically approaching to the surface \( g(y, t) = 0 \).

**Proof.** Substituting (13) in (12), we have \( g_{n+1} = \gamma g_n \), \( n = 0, 1, 2, \ldots \). Converting recurrently this expression we get \( g_{n+1} = \gamma^{n+1} g_0 \). Given that \( \gamma < 1 \), then \( g_n \to 0 \) takes place when \( n \to \infty \). In addition, condition \( \gamma > 0 \) implies that function \( g_n \) does not change sign. Therefore, when \( g_0 < 0 \), \( g_n < 0 \) will be valid for all \( n \). Then the guard condition will never cross the potentially dangerous area \( g(y_n, t_n) = 0 \), which completes the proof.

### 6.1 Control of Event Function in the Integration Algorithm

We complete the implicit problem’s integration algorithm by the algorithm of the step control that takes into account the event function dynamics. Let the solution \( y_n \) at the point \( t_n \) is calculated with the step \( h_n \). In addition, the new accuracy step \( h_{n+1}^{ac} \) is computed by the formula (13). Then the approximate solution at the point \( t_{n+1} \) is calculated as follows:

**Step 1.** Calculate the functions

\[ g_n = g(y_n, t_n), \frac{\partial g_n}{\partial y} = \frac{\partial g(y_n, t_n)}{\partial y}, \frac{\partial g_n}{\partial t} = \frac{\partial g(y_n, t_n)}{\partial t}. \]

**Step 2.** Calculate

\[ g'_n = \frac{\partial g_n}{\partial y} \cdot \varphi_n + \frac{\partial g_n}{\partial t}, \]

where \( \varphi_n = y_n \).

**Step 3.** If \( g'_n < 0 \), then \( h_{n+1} = h_{n+1}^{ac} \) and go to the Step 6.

**Step 4.** Calculate the new “Event” step \( h_{n+1}^{ev} \) by the formula

\[ h_{n+1}^{ev} = (\gamma - 1) \frac{g_n}{g'_n}. \]

**Step 5.** Calculate the new step \( h_{n+1} \) by the formula \( h_{n+1} = \min(h_{n+1}^{ev}, h_{n+1}^{ac}) \).

**Step 6.** Go to the next integration step.

In the Step 3, unlike the previously presented algorithm [4], we determine the direction of event-function change. Near the boundary regime denominator (13) will be positive, and away from the boundary \( g(y, t) = 0 \) it becomes negative. Then, defining the direction of event-function change, we do not impose any further restrictions on the integration step if the event-function is removed from the state boundary.
7 Simulation Results

At time $t=0$ s a far electrical load decreases by 10% initiating an electromagnetic transient. Numerical experiment results are shown in Fig. 4. The calculations are performed by RK3ST algorithm with initial step 0.000001 s. Results from [9] obtained by implicit Euler method with integration step 0.00001 s are shown in the same figure.

![Fig. 4. The moment of generator $G_4$ turbine.](image)

The results of six-machine EPS simulation correspond with those in the original source [9] and do not contradict the theoretical concepts that confirms the correctness of the method used. Discrepancy between the results is based not only on different numerical methods used by [9] and the authors. In [9] the system is considered as a continuous one, so the problem of correct hybrid system event detection arisen from regulators’ deadzone is not dealt with [10]. It contributes to the accumulation of a global calculation error. Using a correct event detection algorithm let obtain better simulation results, as shown by the authors in [20]. It should be noted that the sensitivity to chosen formalisms (continuous, hybrid) varies from one phase variable to another.

8 Conclusion

Computer-aided analysis of discrete-continuous systems has been an actual scientific research area for many years. Modern formalisms and methods for analysis of complex systems can be effectively used by domain specialists only if domain-specific software tools have been developed. Such software frees end-users from routine translation from a mathematical model to its program implementation and helps to carry out a computer experiment. This software solves problems of the optimal representation of a model in the computer memory, a model preparation for computation including choice of the effective step size and numerical method, start and control of the computational experiment process. Expanding
a modeling and simulation environment must demand of minimal changes of existing components as well as development of model specification tools.

For a description and analysis of transients in power systems and their components the use of the methodology of hybrid systems surrounded by the tools of computer analysis is proposed. Approaches to specification of EPS in ISMA instrumental environment are presented. Textual specification is fully consistent with the mathematical notation and can be used for experiments with different models of network elements. Graphical specification convenient for specialists in the field of electricity when the network configuration of the circuit is frequently changing. It should be noted that ISMA has tools for translation from graphical to textual model. Thus, it is possible to verify the program models. The main advantage is the dramatic increase in the efficiency of research when a specialist using an instrumental service deals exclusively with the analysis of the results of their design decisions. While in the traditional practice energy specialist to get the results requires additional expertise in the field of computational mathematics and programming.

The architecture of the solver for the continuous behavior of hybrid systems is proposed. The library of numerical methods can be easily extended by new methods of the Runge-Kutta type as well as other one-step integration methods for ODE systems. The API also provides the mechanism to add implementations of algorithms that deal with another type of systems. The new original method of switching point’s localization is proposed. The algorithm easily complements the existing numerical solvers based on explicit and semi-explicit schemes.

The presented results of the test problem (six-machine EPS) calculation are obtained using the considered approaches and methods. Thus, the correctness of theoretical assumptions, mathematical and algorithmic software is constructively proved.

Acknowledgments. This work was supported by the grant of the Russian Foundation for Basic Research (RFBR grant 17-07-01513) and by the grant of the Education, Audiovisual and Culture Agence (EU), programme ERASMUS + Capacity building in higher education, project 573751-EPP-1-2016-1-DE-EPPKA2-CBHE-JP, Innovative teaching and learning strategies in open modelling and simulation environment for student-centered engineering education.

References

Using of the Mosaic-skeleton Method for Numerical Solution of Three-Dimensional Scalar Diffraction Problems

Sergei Smagin, Aleksei Kashirin, and Mariia Taltykina

Computing Center of FEB RAS, Khabarovsk, Russia

smagin@as.khb.ru, elomer@mail.ru, taltykina@yandex.ru

http://www.ccfebras.ru/en/general

Abstract. In the paper three-dimensional stationary scalar diffraction problems are considered. They are formulated in the form of boundary weakly singular Fredholm integral equations of the first kind with a single function, each of which is equivalent to the original problem. These equations are approximated by systems of linear algebraic equations, which are then solved numerically by an iterative method. In order to reduce the computational complexity, the mosaic-skeleton method is used at the stage of numerical solution of the systems.

Keywords: diffraction problem, Helmholtz equation, integral equation, mosaic-skeleton method, GMRES

1 Introduction

Three-dimensional stationary diffraction problems of acoustic waves are widely used in science and technology. Most often such problems are solved numerically, since their analytical solutions can be constructed only in simplest cases. The method of numerical solution of the initial problems should take into consideration that the solutions are found in unbounded domains, must satisfy the radiation condition at infinity, and are rapidly oscillating functions at large wave numbers. Keeping these limitations in mind, it is preferable to use methods of the potential theory for creating the algorithms for numerical solution. In this case, the three-dimensional problem in an unbounded domain can be reduced to a two-dimensional problem formulated on a closed surface of inclusion.

Using the methods of potential theory, two equivalent weakly singular boundary Fredholm integral equations of the first kind with one unknown function (density) [1–3] are constructed for the problem of diffraction. A special method of averaging kernels of weakly singular integral operators [4, 5] is used to build their discrete analogs. In this case, the unknown density is sought as a linear combination of continuously differentiable finite functions forming a partition of unity on the surface of inclusion. In the process of discretization of the integral equation, the surface integrals are approximated by expressions containing integrals over space $\mathbb{R}^3$, which are then calculated analytically. This allows calculation of...
coefficients of systems of linear algebraic equations (SLAE), approximating corresponding integral equations, by very simple formulas. Once the approximate solution of integral equation is found, the required approximate solution of the diffraction problem by using integral representations is restored in any point of space.

Matrices of the obtained SLAE are dense. The complexity of solving such SLAE by direct methods is \( O(n^3) \), where \( n \) is the order of the system. It was established earlier [1] that the use of the generalized minimal residual method (GMRES) [6] allows to lower the complexity to \( O(mn^2) \), where \( m \) is the number of iterations, which is much less than \( n \) and barely depends on it. The most time-consuming part of the algorithm of GMRES is the use of multiple matrix-vector multiplication. The time expenses on multiplication can be lowered using the "quick" method, the complexity of which is \( o(n^2) \), when \( n \to \infty \). The mosaic-skeleton method [7] – [11] can be used for already developed software for solving problems of mathematical physics. Its basic idea is to approximate the large blocks of dense matrices by low-rank matrices. As a result, the SLAE matrix is not fully computed and stored, but its approximation is used in the procedure of matrix-vector multiplication. In this case, the complexity of multiplication of this approximation by a vector is almost linear.

To implement the method in an already established program for numerical solution of diffraction problems, it is only necessary to add the procedure of creating and storing the approximate matrix and to change the module of matrix-vector multiplication, whereas the method of sampling and the calculation procedure of the elements of the original matrix remain the same. An additional advantage of this method is that, due to the cost effective storage of the approximate matrix, the requirements for computer resources are reduced.

The present work outlines the numerical method for solving three-dimensional problems of diffraction formulated as boundary Fredholm integral equations of the first kind and describes the use of the mosaic-skeleton method for the numerical solution of such problems. It also presents the results of numerical experiments, which allow us to judge the effectiveness of the approach.

2 The initial problem and its equivalent integral equations

Let us consider a three-dimensional Euclidean space \( \mathbb{R}^3 \) with orthogonal coordinate system \( o x_1 x_2 x_3 \), filled with a homogeneous isotropic medium with density \( \rho_e \), with speed of propagation of acoustic oscillations \( c_e \) and absorption coefficient \( \gamma_e \), which has a homogeneous isotropic inclusion, limited by the arbitrary closed surface \( \Gamma \), with density \( \rho_i \), speed of sound \( c_i \) and absorption coefficient \( \gamma_i \). Let’s denote domains \( \mathbb{R}^3 \), occupied by the inclusion and containing medium, by \( \Omega_i \) and \( \Omega_e (\Omega_e = \mathbb{R}^3 \setminus \Omega_i) \).

Suppose there are harmonic sound sources in space \( \Omega_e \), stimulating initial pressure wave field \( u_0 \) in the containing medium. Sound waves come through space and scatter, reaching the inclusion. As a result, there appear reflected
waves in domain $\Omega_e$, and transmitted waves in domain $\Omega_i$. Therefore, the complex amplitude of the complete field of pressures $u$ can be presented as:

$$u = \begin{cases} u_i, & x \in \Omega_i, \\ u_e + u_0, & x \in \Omega_e, \end{cases}$$

where $u_i$, $u_e$ are complex amplitudes of the pressure field of transmitted and reflected wave fields.

The inclusion, initial, reflected and transmitted pressure wave fields are shown in the Figure 1.

Fig. 1. Wave propagation in the inclusion and in the containing medium

Let’s formulate the initial problem.

**Problem 1.** In bounded domain $\Omega_i$ of three-dimensional Euclidean space $\mathbb{R}^3$ and unbounded domain $\Omega_e = \mathbb{R}^3 \setminus \overline{\Omega}_i$, separated by closed surface $\Gamma \in C^{r+\beta}$, $r + \beta > 1$, find complex-valued functions $u_{i(e)} \in H^1(\Omega_{i(e)}, \Delta)$, which satisfy integral identities

$$\int_{\Omega_{i(e)}} \nabla u_{i(e)} \nabla v^*_{i(e)} \, dx - k^2_{i(e)} \int_{\Omega_{i(e)}} u_{i(e)} v^*_{i(e)} \, dx = 0 \quad \forall v_{i(e)} \in H^1_0(\Omega_{i(e)}),$$

conjugation conditions on the material interface between $\Omega_i$ and $\Omega_e$

$$\langle u^-_i - u^+_e, \mu \rangle_\Gamma = \langle f_0, \mu \rangle_\Gamma \quad \forall \mu \in H^{-1/2}(\Gamma),$$

$$\langle \eta, p_i N^- u_i - p_e N^+ u_e \rangle_\Gamma = \langle \eta, p_e f_1 \rangle_\Gamma \quad \forall \eta \in H^{1/2}(\Gamma),$$

as well as the radiation condition at infinity

$$\frac{\partial u_e}{\partial |x|} - i k_e u_e = o \left( |x|^{-1} \right), \quad |x| \to \infty,$$

if functions $f_0 \in H^{1/2}(\Gamma)$ and $f_1 \in H^{-1/2}(\Gamma)$ are set on the boundary $\Gamma$. 

460
Here \( v^* \) is a complex conjugate function to \( v \), \( \langle \cdot, \cdot \rangle \Gamma \) is duality ratio on \( H^{1/2}(\Gamma) \times H^{-1/2}(\Gamma) \), which generalizes the inner product in \( H^0(\Gamma) \), \( u^\pm = \gamma^\pm u \), \( \gamma^- : H^1(\Omega_i) \to H^{1/2}(\Gamma) \), \( \gamma^+: H^1(\Omega_e) \to H^{1/2}(\Gamma) \) are trace operators, \( N^- : H^1(\Omega_i, \Delta) \to H^{-1/2}(\Gamma) \), \( N^+ : H^1(\Omega_e, \Delta) \to H^{-1/2}(\Gamma) \) are operators of normal derivatives [12], \( f_0 = u^+_0 \), \( f_1 = N^+ u_0 \),

\[
k^2_{i(e)} = \omega (\omega + i\gamma_i(e)) / c^2_{i(e)}, \quad \text{Im}(k_{i(e)}) \geq 0, \quad p_{i(e)} = c^2_{i(e)} k^{-2}_{i(e)} \rho^{-1}_{i(e)},
\]

\( \omega \) is a wave circular frequency, \( c_{i(e)} > 0 \), \( \rho_{i(e)} > 0 \), \( \gamma_i(e) \geq 0 \). The definitions of functional spaces used hereinafter are available in [12].

**Remark 1.** If \( \text{Im}(k_e) = 0 \), then \( u_e \in H^1_{\text{loc}}(\Omega_e, \Delta) \).

In works [1], [2] the next theorem was proven.

**Theorem 1.** Problem 1 has no more than one solution.

Let’s introduce the following notation

\[
(A_{i(e)} q)(x) \equiv \langle G_{i(e)}(x, \cdot), q \rangle \Gamma, \quad (B_{i(e)} q)(x) \equiv \langle N_x G_{i(e)}(x, \cdot), q \rangle \Gamma,
\]

\( (B^*_{i(e)} q)(x) \equiv \langle N_{\cdot} G_{i(e)}(x, \cdot), q \rangle \Gamma, G_{i(e)}(x, y) = \exp (i k_{i(e)} |x - y|) / (4\pi |x - y|) \).

The solution to problem 1 will be sought in the form of potentials

\[
u_e(x) = (A_e q)(x), \quad x \in \Omega_e, \quad (5)
\]

\[u_i(x) = (p_{ei} A_i (N^+ u_e + f_1) - B^*_i (u^+_e + f_0))(x), \quad x \in \Omega_i,
\]

where \( q \in H^{-1/2}(\Gamma) \) is an unknown density, \( f_0 \in H^{1/2}(\Gamma) \), \( f_1 \in H^{-1/2}(\Gamma) \), \( p_{ei} = p_e / p_1 \).

The kernels of these integral operators are fundamental solutions of the Helmholtz equations and their normal derivatives. Therefore, as shown in [2], they satisfy identities (1) and radiation condition at infinity (3). In addition, the fulfillment of the first of the matching conditions for them (2) automatically entails the fulfillment of the second matching condition. Substituting potentials (5) in the first matching condition, we obtain a weakly singular Fredholm integral equation of the first kind for defining unknown density \( q \):

\[
\langle C q , \mu \rangle \Gamma = \langle f_2, \mu \rangle \Gamma \quad \forall \mu \in H^{-1/2}(\Gamma),
\]

where

\[
C = (0.5 + B^*_i) A_e + p_{ei} A_i (0.5 - B_e), \quad f_2 = - (0.5 + B^*_i) f_0 + p_{ei} A_i f_1.
\]

Problem 1 allows another equivalent formulation in the form of Fredholm integral equation of the first kind with a weak singularity in the kernel [1]. We shall seek its solution in the form

\[
u_i(x) = (A_i q)(x), \quad x \in \Omega_i,
\]

\(461\)
\[ u_e(x) = \left( A_e \left( f_1 - p_{ie} N^{-} u_i \right) - B_e^* \left( f_0 - u_i^{-} \right) \right)(x), \quad x \in \Omega_e, \]

where \( q \in H^{-1/2}(\Gamma) \) is an unknown density, \( f_0 \in H^{1/2}(\Gamma), f_1 \in H^{-1/2}(\Gamma), \) \( p_{ie} = p_i/p_e. \)

In this case problem 1 is reduced to equation

\[
\langle Dq, \mu \rangle_{\Gamma} = \langle f_0, \mu \rangle_{\Gamma} \quad \forall \mu \in H^{-1/2}(\Gamma), \tag{8}
\]

\[ D = (0.5 - B_e^*) A_i + p_{ie} A_e (0.5 + B_i). \]

The next theorem is proved [1].

**Theorem 2.** Let \( f_0 \in H^{1/2}(\Gamma), f_1 \in H^{-1/2}(\Gamma), \gamma_e > 0 \) or \( \omega \) is not the eigen frequency of the problem

\[
\Delta u + k_e^2 u = 0, \quad x \in \Omega_t, \quad u^{-} = 0.
\]

Then equations (6) and (8) are correctly solvable in the class of densities \( q \in H^{-1/2}(\Gamma) \) and formulae (5) and (7) provide the solution to problem 1.

**Remark 2.** In cases where we are more interested in the wave field in domain \( \Omega_e, \) it is preferable to use equation (6), which allows the calculation of the reflected field by a simpler formulae. For a similar reason, if we are interested in the transmitted wave field in domain \( \Omega_t, \) it is preferable to use equation (8).

### 3 Numerical method

The applied method of numerical solution is presented in work [3] and appears to be the development of the technics proposed and tested for the first time in [4]. Let us briefly describe the general scheme of its implementation. Let us construct a surface \( \Gamma \) coating by system \( \{ \Gamma_m \}_{m=1}^{M} \) of neighborhoods of nodes \( x'_m \in \Gamma, \) lying within the spheres of radii \( h_m \) centered at \( x'_m, \) and denote its subordinate partition of unity by \( \{ \varphi_m \}. \) In place of \( \varphi_m \) let us take functions

\[ \varphi_m(x) = \varphi'_m(x) \left( \sum_{k=1}^{M} \varphi'_k(x) \right)^{-1}, \quad \varphi'_m(x) = \begin{cases} \left(1 - r_m^2/h_m^2\right)^{3}, & r_m < h_m, \\ 0, & r_m \geq h_m, \end{cases} \]

where \( x \in \Gamma, r_m = |x - x'_m|, \varphi_m \in C^1(\Gamma) \) when \( \Gamma \in C^{r+\beta}, r + \beta > 1. \)

Approximate solutions of equations (6) and (8) will be sought on grid \( \{ x_m \}, \)

\[ x_m = \frac{1}{\varphi_m} \int_{\Gamma} x \varphi_m d\Gamma, \quad \varphi_m = \int_{\Gamma} \varphi_m d\Gamma, \]

the nodes are the centers of gravity of functions \( \varphi_m. \) We assume that for all \( m = 1, 2, \ldots, M \) inequalities are satisfied

\[ 0 < h' \leq |x_m - x_n|, \quad m \neq n, \quad n = 1, 2, \ldots, M, \]
\[ h' \leq h_m \leq h, \quad h/h' \leq q_0 < \infty, \]

where \( h, h' \) are positive numbers depending on \( M \), \( q_0 \) does not depend on \( M \).

Instead of unknown function \( q \) set on \( \Gamma \) we shall seek generalized function \( q\delta_\Gamma \), acting according to the rule

\[ \langle q\delta_\Gamma, \eta \rangle_{\mathbb{R}^3} = \langle q, \eta \rangle_{\Gamma} \quad \forall \eta \in H^1(\mathbb{R}^3). \]

We shall approximate this function by expression

\[ q(x)\delta_\Gamma(x) \approx \sum_{n=1}^{M} q_n \varphi_n \psi_n(x), \]

where \( q_n \) are unknown coefficients,

\[ \psi_n(x) = (\pi \sigma_n^2)^{-3/2} \exp \left(-\frac{(x-x_n)^2}{\sigma_n^2}\right), \quad \sigma_n^2 = 0.5 \varphi_n. \]

In paper [3] it is shown that for any functions \( \eta \in H^1(\mathbb{R}^3) \) and \( q \in H^1(\Gamma) \) equality

\[ \langle q, \eta \rangle_{\Gamma} = \sum_{n=1}^{M} q_n \varphi_n (\psi_n, \eta)_{\mathbb{R}^3} + O(h^2). \]

Approximation of the density of a single layer potential by a volume density allows us to obtain simple formulae for approximating integral operator \( A_{i(e)} \) from (4). The theoretical justification of the presented approach is given in [4].

The integral operators from (4) on \( \Gamma \) are approximated by expressions [1, 5].

\[ \langle A_{i(e)}q, \varphi_m \rangle_{\Gamma} \approx \sum_{n=1}^{M} A_{i(e)}^{mn} q_n, \quad m = 1, 2, ..., M, \quad (9) \]

\[ A_{i(e)}^{mn} \equiv A_{mn} \left(k_{i(e)}\right), \]

\[ A_{mn}(k) = \frac{\varphi_m \varphi_n}{8\pi r_{mn}} \exp \left(\mu_{mn}^2 - \gamma_{mn}^2\right) \left(w(z^-_{mn}) - w(z^+_{mn})\right), \quad n \neq m, \]

\[ A_{mm}(k) = \frac{\varphi_m^2}{4\pi} \exp \left(\mu_{mm}^2\right) \left(ikw(\mu_{mm}) + \sqrt{2\pi} \frac{\varphi_m}{\pi \sigma_m} \left(\frac{\varphi_m}{\pi \sigma_m} + 2\sigma_m - \frac{k^2 \sigma_m^3}{3}\right)\right), \]

\[ \sigma_{mn}^2 = \sigma_m^2 + \sigma_n^2, \quad \mu_{mn} = 0.5k \sigma_{mn}, \quad z_{mn}^\pm = \mu_{mn} \pm i\gamma_{mn}, \]

\[ \gamma_{mn} = r_{mn}/\sigma_{mn}, \quad i^2 = -1, \]

\[ w(z) = -\frac{2i}{\sqrt{\pi}} \exp \left(-z^2\right) \int_z^\infty \exp \left(t^2\right) dt, \]

\[ \langle aq + B_{i(e)}q, \varphi_m \rangle_{\Gamma} \approx \sum_{n=1}^{M} B_{i(e)}^{mn} q_n, \quad m = 1, 2, ..., M, \quad a = \pm 0.5, \quad (10) \]
\[
\langle aq + B^*_i(e)q, \varphi_m \rangle _\Gamma \approx \sum_{n=1}^{M} B_{i(n)}^{nm} q_n, \ m = 1, 2, ..., M, \quad (11)
\]

\[
B_{i(n)}^{nm} = \frac{\eta_{mn}}{4\pi r_{mn}^2} \exp \left(i k_{i(e)} r_{mn} \right) \left(i k_{i(e)} r_{mn} - 1 \right) \bar{\varphi}_m \bar{\varphi}_n, \quad n \neq m,
\]

\[
B_{i(e)}^{nm} = (-|a| + a + G_{s_m}) \bar{\varphi}_m, \quad \eta_{mn} = \sum_{l=1}^{3} n_{ml} \frac{x_{ml} - x_{nl}}{r_{mn}}, \quad G_{s_m} = - \sum_{n \neq m} \frac{\eta_{nm} \bar{\varphi}_n}{4\pi r_{mn}^2},
\]

\(n_{ml}\) are components of the unit vector of the external normal to the surface at point \(x_m\).

The operators on the left sides of equations (6) and (8) are approximated by the composition of operators (9)–(11):

\[
\langle Cq, \varphi_m \rangle _\Gamma \approx \sum_{n=1}^{M} C_{i(e)}^{nm} q_n, \quad \langle Dq, \varphi_m \rangle _\Gamma \approx - \sum_{n=1}^{M} C_{i(e)}^{nm} q_n, \ m = 1, 2, ..., M, \quad (12)
\]

\[
C_{i(e)}^{nm} = B_{i(n)}^{nm} A_{ie}^{mn} - p_{ei} A_{i(n)}^{mn} B_{i(e)}^{mn},
\]

and the right sides of equations (6) and (8) by formulæ

\[
\langle f_2, \varphi_m \rangle _\Gamma \approx \sum_{n=1}^{M} (p_{ei} A_{i(n)}^{mn} f_{1n} - B_{i(e)}^{nm} f_{0n}), \quad \langle f_0, \varphi_m \rangle _\Gamma = \bar{\varphi}_m f_{0m}, \quad (13)
\]

\[
f_{lm} = \langle f_l, \varphi_m / \bar{\varphi}_m \rangle _\Gamma, \quad l = 0, 1, \quad m = 1, 2, ..., M.
\]

Solving the corresponding SLAE, we find the approximate values of the density of integral equations at discretization points. After that, an approximate solution of the diffraction problem using integral representations can be calculated at any point in space.

## 4 Mosaic-skeleton method

Let us introduce the definitions necessary for describing the mosaic-skeleton method [7]. Let \(A_k\) be block matrix \(A_{n \times m}\), and \(\Pi(A_k)\) be matrix of size \(n \times m\), resulting from \(A_k\) by adding zeros up to \(A\).

**Definition 1.** The systems of blocks \(\{A_k\}\) will be called covering \(A\), if

\[
A = \sum_k \Pi(A_k),
\]

and mosaic partitioning \(A\), if, in addition, \(\bigcap_k A_k = \emptyset\).

**Definition 2.** The mosaic rank of matrix \(A \in \mathbb{C}^{n \times m}\), corresponding to some covering, is number

\[
\mr A = \sum_k \mr A_k/(n + m), \quad (14)
\]

where the sum is taken over all blocks of covering \(A_k \in \mathbb{C}^{n_k \times m_k}\), and \(\mr A_k = \min\{n_k m_k, (n_k + m_k) \ \mr A_k\}\).
The mosaic rank determines the memory requirements for storing the mosaic partitioning of matrix $A$, and also the computational complexity of matrix-vector multiplication.

An important indicator of the effectiveness of this method is the compression factor $I$ [7]. It is defined by the following formulae

$$I = \frac{\text{mem} A_p}{\text{mem} A} \times 100,$$

where $\text{mem} A_p$ is the amount of memory required to store a matrix in a low-rank format, $\text{mem} A$ is the amount of memory for storing the original matrix.

**Definition 3.** We shall call a matrix of $uv^*$ kind a skeleton, where $u$ and $v$ are column vectors, $v^*$ denotes the row-vector Hermite-conjugate to vector $v$.

From definition 3 it follows that rank $(uv^*) = 1$.

The mosaic-skeleton method consists of three phases: constructing a cluster tree, creating a list of blocks and finding a low-rank approximation. For the first phase, the input data is a grid, on which discretization is carried out. The set of all points of the grid is called a zero-level cluster (or a root of the tree). At each step, the original cluster is split into several disjoint subclusters. This continues until the level of the tree reaches the prescribed one. As a zero-level cluster, we can take a cube, and plunge the domain on the border of which the grid is built there. When moving to the next level of the tree, each edge of this cluster bisects, resulting in 8 subcubes. Then, all the points belonging to the original cube are distributed across 8 subcubes, forming 8 subclusters (some of them might be empty). On reaching the maximum level of $k$, the number of clusters equals $8^k$.

The next step is creating the list of blocks. Any two clusters determine a block in a matrix. If the points of the clusters are geometrically separated from each other, the block enters a far zone and an approximation will be constructed for it, or, otherwise, it enters a near zone, and the elements of the blocks will be calculated by formulae (12).

At the final stage, the blocks of the far zone are approximated by low-rank matrices. Such matrices can be sought in different ways [10, 11]. In paper [13] the cross of the matrix row and column is chosen as a template. A current approximation is built on the cross at each step. This algorithm is called incomplete cross approximation [8, 14].

**Algorithm 1. Incomplete cross approximation.**

Approximation of matrix $A$ of size $n \times m$ by matrix $A^{(r)}$, being the sum of $r$ skeletons.

1. Let $p$ be the number of the calculated skeleton. At this step, we assume that $p = 1$ and select an arbitrary number $j_p$ from the column of matrix $A$.
2. In the column with number $j_p$ we calculate all the elements of matrix $A$, then, subtract the elements of all previously obtained skeletons in these positions from them. In the obtained column $u_p$ we find the maximum element in modulus. Let it be located in the row with number $i_p$. 

465
3. We calculate the row of matrix $A$ with number $i_p$ and subtract from it the elements of all already found skeletons in their respective positions. In the obtained row $v^*_p$ we find the maximum element in modulus. It should be noted that an element from the column with number $j_p$ cannot be selected again. The number of the column in which the maximum element is found will be denoted by $j_{p+1}$.

4. Along the cross with centre $(i_p, j_p)$ we build a skeleton so that the coefficients of matrix $A(p) = \sum_{\alpha=1}^{p} u_\alpha v^*_\alpha A_{i_\alpha j_\alpha}$ (16) coincide with the coefficients of the original matrix in positions $p$ of computed columns $j_1, ..., j_p$ and $p$ of calculated rows $i_1, ..., i_p$.

5. We check the stopping criterion. If it is satisfied, the calculation stops. If it is not satisfied, we set $p := p + 1$ and repeat the algorithm from step 2.

An approximation (16) is considered to be sufficiently accurate if inequality (stopping criterion) $||A - A(p)||_F \leq \varepsilon||A(p)||_F$, where $\varepsilon$ is a relative approximation error, $|| \cdot ||_F$ is a Frobenius norm [15], is satisfied.

To verify this criterion, it is necessary to calculate $nm$ of the elements of matrix $A$, which is very time-consuming. The number of operations can be reduced by using the upper estimate of $||A - A(p)||_F$, obtained in [8]

$$||A - A(p)||_F \leq (l - p)||u_p v_p^*||_F/|A_{i_p j_p}| = (l - p)||u_p||_F||v_p||_F/|A_{i_p j_p}|,$$

where $l = \min(n, m)$.

Here the stopping criterion takes the form of

$$(l - p)||u_p||_F||v_p||_F/|A_{i_p j_p}| \leq \varepsilon||A(p)||_F.$$

Using recurrence formulae we can obtain next result

$$||A(p)||_F^2 = ||A(p-1)||_F^2 + 2Re\left(\sum_{\alpha=1}^{p-1} \frac{(u_\alpha^* v_\alpha)(v^*_\alpha v_p)}{A_{i_\alpha j_\alpha}}\right) + \frac{||u_p||_F^2||v_p||_F^2}{|A_{i_p j_p}|^2},$$

$$||A(1)||_F = \frac{||u_1||_F||v_1||_F}{|A_{i_1 j_1}|}, \quad p \geq 2.$$

5 Numerical results

The program for the numerical solution of diffraction problems is written in Fortran 90 and has the form of a console application, designed to run on multiprocessor computing systems. Intel Fortran Compiler performs as a compiler,
Intel Math Kernel Library (Intel MKL) and implementation of the OpenMP standard for the compiler are also used. The above mentioned software is used in the computing cluster of CC FEB RAS (http://hpc.febras.net/). The cluster consists of one subcontrol and thirteen compute nodes. The compute node with four Six Core AMD Opteron™ 8431 processors, with clock frequency of 2.40 GHz and 96 GB of RAM has been involved in testing. To read more information about the program, see [16].

**Example 1.** The diffraction problem of a plane acoustic wave on a triaxial ellipsoid with semi-axes (0.75, 1, 0.5) centered at the origin of coordinates and having three different sets of parameters of the host medium and inclusion. The complex amplitude of the initial wave field of pressures has the form $u_0(x) = \exp(ik_es_3)$, $f_0 = u_0^+$, $f_1 = N^+u_0$, parameters of the media: $I$ $k_i = 8, \rho_i = 3, k_e = 5.5, \rho_e = 1$; $II$ $k_i = 15.5, \rho_i = 5, k_e = 9, \rho_e = 4$; $III$ $k_i = 21, \rho_i = 7, k_e = 30.5, \rho_e = 9.5$.

In all the examples, the diffraction problem was solved twice: using the mosaic-skeleton method in GMRES and without it. In the process discretization of equations (6) and (8) was carried out by formulas (12) and (13), the order of matrix $M$ ranged from 1032 to 64139. Hereinafter the accuracy of GMRES amounted to $10^{-7}$, and the accuracy of the low-rank approximations was $10^{-5}$. The level of the cluster tree varied from 4 to 5. Further, squares denote the first set of media parameters, circles stand for the second set and triangles mark the third set in all the graphs.

In this case, the solutions found approximately were compared with the solutions found on a denser grid, i.e. with 64139 sampling points. This is because the analytical solution of Example 1 is unknown. The relative errors of solution of the problem in Example 1 using equation (6) and (8) are shown in Figures 2. The solid lines indicate errors $u_i$, and the dashed lines represent errors $u_e$. According to the numerical experiments, the method has the second order of accuracy, as the errors fall twofold when the order of matrix is doubled ($O(M^{-1}) = O(h^2)$). The errors were calculated in the norm of grid functions $H_0^h(\Omega_{i(e)})$.

The time spent on the solution of SLAE depending on the order of matrix $M$ with the mosaic-skeleton method and without it is shown in Figures 3 for equation (6). The results for equation (8) are similar. Hereinafter, the time $t$ is presented in seconds. The dashed lines show the time for solving SLAE using the mosaic-skeleton method, and the solid lines indicate the time for solving SLAE without it. When using the mosaic-skeleton method in GMRES, the solution time is 115 times less (at $M = 64139$ for the third set of parameters for equation (6)) than without the fast method. When the number of nodes is doubled, the time for solving SLAE using the mosaic-skeleton method increases 2.5 times on average.

Tables 1 and 2 show the values of the mosaic rank and compression factor for SLAE, approximating the integral equation (6), depending on the number of grid points. For equation (8) almost the same results were obtained. This is because the kernels of equations (6) and (8) have similar properties as they are...
compositions of similar integral operators. The mosaic rank was calculated using formula (14), and the compression factor by formula (15). The numerical experiments show that the mosaic rank increases like $O(\ln^3(M))$, and the compression factor, from a certain $M$, decreases like $O(\ln^3(M)/M)$.

**Table 1.** Mosaic ranks for the problem in Example 1.

<table>
<thead>
<tr>
<th>$k$</th>
<th>1032</th>
<th>2096</th>
<th>4010</th>
<th>8022</th>
<th>16033</th>
<th>32120</th>
<th>64139</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>493.1</td>
<td>751.9</td>
<td>991.0</td>
<td>1244.1</td>
<td>1517.1</td>
<td>1845.1</td>
<td>2241.0</td>
</tr>
<tr>
<td>II</td>
<td>509.9</td>
<td>818.5</td>
<td>1103.0</td>
<td>1376.8</td>
<td>1666.3</td>
<td>2004.4</td>
<td>2410.3</td>
</tr>
<tr>
<td>III</td>
<td>518.0</td>
<td>948.5</td>
<td>1327.9</td>
<td>1677.7</td>
<td>2011.6</td>
<td>2380.7</td>
<td>2810.7</td>
</tr>
</tbody>
</table>

**Table 2.** Compression factors for the problem in Example 1.

<table>
<thead>
<tr>
<th>$k$</th>
<th>1032</th>
<th>2096</th>
<th>4010</th>
<th>8022</th>
<th>16033</th>
<th>32120</th>
<th>64139</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>95.6</td>
<td>71.7</td>
<td>49.4</td>
<td>31.0</td>
<td>18.9</td>
<td>11.5</td>
<td>7.0</td>
</tr>
<tr>
<td>II</td>
<td>98.8</td>
<td>78.1</td>
<td>55.0</td>
<td>34.3</td>
<td>20.8</td>
<td>12.5</td>
<td>7.5</td>
</tr>
<tr>
<td>III</td>
<td>100.0</td>
<td>90.5</td>
<td>66.2</td>
<td>41.8</td>
<td>25.1</td>
<td>14.8</td>
<td>8.8</td>
</tr>
</tbody>
</table>

**Example 2.** The problem of diffraction on a triaxial ellipsoid with semi-axes $(0.75, 1, 0.5)$ centered at the origin is considered. The incident field is generated by a point source

$$u_0(x) = \exp(ik|x-x_0|)/|x-x_0|,$$
where \( x_0 = (1.125, 1, 0.75) \). The sets of parameters of the containing medium and inclusion are the same as in Example 1.

The relative errors of solutions \( u_{i(e)} \) for the problem in Example 2 formulated in the form of equation (6) and (8) are shown in Figures 4. The solid lines indicate errors \( u_i \), and the dashed lines represent errors \( u_e \). All the errors are calculated in the norm of grid functions \( H^0_h(O_{i(e)}) \). When we increase the order of the matrices twofold, they also halve, as in the above example.

The values of mosaic rank and compression factor for the problem in Example 2 formulated in the form of equation (6) are shown in Tables 3 and 4, respectively. The mosaic rank, just as in the previous example, grows at the rate of \( O(\ln^3(M)) \), and the compression factor falls like \( O(\ln^3(M)/M) \). Almost the same results were obtained for equation (8).

<table>
<thead>
<tr>
<th>( k ) ( M )</th>
<th>1032</th>
<th>2096</th>
<th>4010</th>
<th>8022</th>
<th>16033</th>
<th>32120</th>
<th>64139</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>493.1</td>
<td>751.9</td>
<td>991.0</td>
<td>1244.1</td>
<td>1517.1</td>
<td>1845.1</td>
<td>2241.0</td>
</tr>
<tr>
<td>II</td>
<td>509.9</td>
<td>818.5</td>
<td>1103.0</td>
<td>1376.8</td>
<td>1666.3</td>
<td>2004.4</td>
<td>2410.3</td>
</tr>
<tr>
<td>III</td>
<td>518.0</td>
<td>948.5</td>
<td>1327.9</td>
<td>1677.7</td>
<td>2011.6</td>
<td>2380.7</td>
<td>2810.7</td>
</tr>
</tbody>
</table>

Figures 5 show the real and imaginary parts of solution \( u_e \) for the first set of media parameters in Example 2 on the interval \(-5 \leq x_1 \leq 5, \ x_2 = 0, \ x_3 = 1\) and on the interval \( x_1 = 0, \ -5 \leq x_2 \leq 5, \ x_3 = 1\), respectively. It is evident that solution \( u_e \) oscillates rapidly and slowly decreases while distancing from the inclusion.
Fig. 4. Errors of solutions \( u_i \) (solid lines) and \( u_e \) (dashed lines) for the problem in Example 2, obtained with the help of equations (6) and (8), respectively.

Fig. 5. Real and imaginary parts of solution \( u_e \) in Example 2.
Table 4. Compression factors for the problem in Example 2

<table>
<thead>
<tr>
<th>k ∖ M</th>
<th>1032</th>
<th>2096</th>
<th>4010</th>
<th>8022</th>
<th>16033</th>
<th>32120</th>
<th>64139</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>95.6</td>
<td>71.7</td>
<td>49.4</td>
<td>31.0</td>
<td>18.9</td>
<td>11.5</td>
<td>7.0</td>
</tr>
<tr>
<td>II</td>
<td>98.8</td>
<td>78.1</td>
<td>55.0</td>
<td>34.3</td>
<td>20.8</td>
<td>12.5</td>
<td>7.5</td>
</tr>
<tr>
<td>III</td>
<td>100.0</td>
<td>90.5</td>
<td>66.2</td>
<td>41.8</td>
<td>25.1</td>
<td>14.8</td>
<td>8.8</td>
</tr>
</tbody>
</table>

6 Conclusion

We have studied the possibilities of using the mosaic-skeleton method for numerical solution of three-dimensional diffraction problems of acoustic waves, for which two new formulations have been offered in the form of boundary Fredholm integral equations of the first kind with one unknown function. The method has been implemented on the stage of solving linear systems. All of the most labor-intensive computing processes have been parallelized using OpenMP. The obvious advantage of this approach compared to the others, for example, to a fast multipole method, is that the mosaic-skeleton method can be used in already existing software for the numerical solution of integral equations of diffraction theory. Thereat only the algorithms for calculating the coefficients of linear systems and matrix-vector multiplication are changed. Using the multipole method and other similar methods involves creating virtually new computing algorithms and programs.

Numerical experiments have shown that the modification of numerical algorithms for solving diffraction problems with the mosaic-skeleton method leads to a significant acceleration of their work while maintaining the same accuracy of calculations.

References

Numerical Analysis of Acoustic Waves in a Liquid Crystal Taking Into Account Couple-Stress Interaction

Irina Smolekho, Oxana Sadovskaya, and Vladimir Sadovskii

Institute of Computational Modeling SB RAS, Akademgorodok 50/44, 660036 Krasnoyarsk, Russia
{ismol, o_sadov, sadov}@icm.krasn.ru
http://icm.krasn.ru

Abstract. Based on the mathematical model describing the thermo-mechanical behavior of a liquid crystal, which takes into account the couple-stress interactions, the system of two differential equations for tangential stress and angular velocity was obtained in two-dimensional case. Computational algorithm for numerical solution of this system of equations of the second order under given initial data and boundary conditions is worked out. The algorithm is implemented as a parallel program in the C language using the CUDA technology for computer systems with graphic accelerators. A series of numerical computations of acoustical waves in a liquid crystal was carried out to demonstrate the efficiency of proposed parallel program.

Keywords: liquid crystal, couple-stress medium, dynamics, finite-difference scheme, parallel computational algorithm, CUDA technology

1 Introduction

Liquid crystal – it is an intermediate state of matter, which appears at the same time the properties of elasticity and fluidity. The liquid crystal phase is formed during melting of a number of organic substances. It exists in a range from the melting temperature to a higher temperature, when heated to a substance which becomes an ordinary liquid. Below this range the substance is a solid crystal. In the liquid crystal state may be some organic compounds consisting of molecules of an elongated shape (in the form of elongated rods or plates) having parallel folding of such molecules. Liquid crystal as a liquid can take the container shape in which it is placed. However, apart from this property, combining it with a liquid, it has the property, characteristic of crystals, – the presence of the order of the spatial orientation of the molecules. The liquid crystal molecules are oriented in a direction which is determined by the unit vector, called “director”. Depending on the type of ordering axes of liquid crystal molecules, they are divided into three types: nematic, smectic and cholesteric. Nematic and smectic liquid crystals are characterized by a parallel arrangement
of the molecules. Cholesteric liquid crystals are a kind of nematic liquid crystals, but they lack long-range order. In this paper we consider a nematic liquid crystal. Nematic liquid crystals are characterized by the orientation of the longitudinal axes of the molecules along a certain direction (long-range orientational order is characteristic for them). Molecules continuously slide in the direction of their long axes, revolving around them, but at the same time retain orientational order: the long axes are directed along a preferred direction. In a nematic state, not all molecules have the same orientation. Because of the lack of fixing of molecules the orientation of the director changes. Since the director at different sections is oriented differently, the regions with different directions of director – the domains – appear in a liquid crystal. At the interfaces of domains the light refractive index changes, so liquid crystals become hazy. The main properties of liquid crystals are described in [1].

Liquid crystals find many applications because of strong dependence of their properties on external influences. Due to their ability to reflect light, the liquid crystals are widely used in laboratories and in the art as convenient means of visualizing the thermal fields and temperature changes. Due to the high resolution of the liquid crystals, they can be used in microelectronics technology to detect defects in chips, that improves reliability of the circuit. With the help of liquid crystals in medicine can directly observe the distribution of the human body surface temperature, and it is important to identify foci of inflammation hidden under the skin. Liquid crystals found the most widely usage in digital technology. Currently, color LCD screens have even greater range of applications: mobile phones, personal computers and televisions, which have a small thickness, low power, high resolution and brightness.

One of the approaches to the construction of a mathematical model to describe the behavior of liquid crystals is based on the representation of a liquid crystal medium as a fine-dispersed continuum. At each point of this continuum, the domains of a liquid crystal can move in accordance with laws of the dynamics of viscous or inviscid liquid and can rotate relative to a liquid, encountering resistance to rotation. The models of liquid crystals have been proposed by Eriksen [2], Leslie [3], Aero [4] and other authors. This paper is devoted to the numerical solution of differential equations of the second order for tangential stress and angular velocity, obtained from the system of equations describing the thermomechanical behavior of a liquid crystal in the two-dimensional case.

2 Governing Equations

In the framework of acoustic approximation, the mathematical model of a liquid crystal without taking into account the couple stresses is described in [5, 6]. The system of equations of this model includes the equations of translational and rotational motion, the equation for the angle of rotation, the constitutive equations for pressure and tangential stress, as well as the equation of anisotropic heat conduction with variable coefficients. Parallel computational algorithm for the solution of this system is represented in [7, 8].
In two-dimensional case the complete system of equations describing the behavior of a liquid crystal under weak acoustic perturbations taking into account the couple stresses is as follows:

\[\begin{align*}
\rho u_{,t} &= -p_{,x} - q_{,y}, & \rho v_{,t} &= q_{,x} - p_{,y}, \\
\dot{\omega}_{,t} &= 2q + \mu_{x,x} + \mu_{y,y}, & \varphi_{,t} &= \omega, \\
p_{,t} &= -k(u_{,x} + v_{,y}) + \beta T_{,t}, & q_{,t} &= \alpha(v_{,x} - u_{,y}) - 2\alpha(\omega + q/\eta), \\
\mu_{x,t} &= \gamma \omega_{,x}, & \mu_{y,t} &= \gamma \omega_{,y}, \\
c T_{,t} &= (a_{11} T_{,x} + a_{12} T_{,y})_{,x} + (a_{12} T_{,x} + a_{22} T_{,y})_{,y} - \\
&\quad - \beta T(u_{,x} + v_{,y}) + 2q^2/\eta.
\end{align*}\]  

(1)

Here \(u\) and \(v\) are the projections of the velocity vector on the coordinate axes, \(\omega\) and \(\varphi\) are the angular velocity and the rotation angle, \(p\) is the hydrostatic pressure, \(q\) is the tangential stress, \(\mu_x\) and \(\mu_y\) are the couple stresses, \(T\) is the absolute temperature, \(\rho\) is the density, \(j\) is the moment of inertia, \(k\) is the bulk compression modulus, \(\alpha\) is the modulus of elastic resistance to rotation, \(\eta\) is the viscosity coefficient, \(c\) and \(\beta\) are the coefficients of heat capacity and thermal expansion, \(a_{11}, a_{12}\) and \(a_{22}\) are the components of the thermal conductivity tensor: \(a_{11} = a_1 \cos^2 \varphi + a_2 \sin^2 \varphi, \ a_{12} = (a_1 - a_2) \sin \varphi \cos \varphi, \ a_{22} = a_1 \sin^2 \varphi + a_2 \cos^2 \varphi, \ \) (\(a_1\) and \(a_2\) are the thermal conductivity coefficients of a liquid crystal in the direction of molecular orientation and in the transverse direction). Subscripts after a comma denote the partial derivatives with respect to time \(t\) and spatial variables \(x\) and \(y\).

The system (1) includes the equations of translational and rotational motion, the equation for the angle of rotation, the equations for state for pressure and tangential stress, the equations for couple stresses, and the equation of anisotropic heat conduction with variable coefficients.

Let’s consider how to obtain the system of equations of the second order for tangential stress and angular velocity. Differentiating the first equation of (1) by \(x\), the second equation by \(y\) and subtracting the second from the first, we find:

\[\rho(u_{,y} - v_{,x})_{,t} = -\Delta q,\]

where \(\Delta\) is the Laplace operator. In view of this expression and also expressions for \(\mu_{x,t}\) and \(\mu_{y,t}\), after differentiation of corresponding equations of the system (1) by \(t\), we obtain a separate subsystem for the tangential stress \(q\) and the angular velocity \(\omega\):

\[\begin{align*}
q_{,tt} + \frac{2\alpha}{\eta} q_{,t} + 2\alpha \omega_{,t} &= \frac{\alpha}{\rho} \Delta q, \\
\omega_{,tt} - \frac{2}{j} q_{,t} &= \frac{\gamma}{j} \Delta \omega.
\end{align*}\]  

(2)
Initial data for the system (2) have the following form:
\[
q|_{t=0} = q^0, \quad q_t|_{t=0} = \alpha(v^0_y - u^0_y) - 2\alpha\left(\omega^0 + \frac{q^0}{\eta}\right) = -2\alpha\left(\omega^0 + \frac{q^0}{\eta}\right),
\]
\[
\omega|_{t=0} = \omega^0, \quad \omega_t|_{t=0} = \frac{1}{j}(2q^0 + \mu^0_{x,x} + \mu^0_{y,y}) = \frac{2q^0}{j},
\]
where \(u^0, v^0, \omega^0, q^0, \mu^0_{x,x}, \mu^0_{y,y}\) are given constants at the initial time moment.

The fourth-order equation for \(q\) can be derived from the subsystem (2). So, expressing \(\omega_t\) from the first equation of (2) and substituting it into the second equation, differentiated by \(t\), we obtain the next chain of equations:
\[
\omega_t = \frac{1}{2\alpha}\left(\frac{\alpha}{\rho}\Delta q - q_{tt} - \frac{2\alpha}{\eta}q,t\right), \quad \omega_{ttt} = \frac{2}{j}q_{tt} + \frac{\gamma}{j}\Delta\omega_t,
\]
\[
q_{ttt} + 2\frac{\alpha}{\eta}q_{tt} + \frac{4\alpha}{j}q_{tt} - \left(\frac{\alpha}{\rho} + \frac{\gamma}{j}\right)\Delta q_{tt} - \frac{2\alpha\gamma}{j\eta}\Delta q_t = -\frac{\alpha}{\rho}\frac{\gamma}{j}\Delta^2 q.
\]
Initial data for the Cauchy problem are as follows:
\[
q|_{t=0} = q^0, \quad q_t|_{t=0} = \alpha(v^0_y - u^0_y) - 2\alpha\left(\omega^0 + \frac{q^0}{\eta}\right) = -2\alpha\left(\omega^0 + \frac{q^0}{\eta}\right),
\]
\[
q_{tt}|_{t=0} = \frac{\alpha}{\rho}\Delta q^0 - \frac{2\alpha}{j}(2q^0 + \mu^0_{x,x} + \mu^0_{y,y}) + \frac{2\alpha}{\eta}q_{tt}^0 = 4\alpha\left[\frac{\alpha}{\eta}\omega^0 + \left(\frac{\alpha}{\eta^2} - \frac{1}{j}\right)q^0\right],
\]
\[
q_{ttt}|_{t=0} = \frac{\alpha}{\rho}\Delta q^0 - \frac{2\alpha}{j}(2q^0 + \gamma \Delta\omega^0) - \frac{2\alpha}{\eta}q_{tt}^0 = 8\alpha^2\left[\left(\frac{1}{j} - \frac{\alpha}{\eta^2}\right)\omega^0 + \frac{1}{\eta}\left(\frac{1}{j} + \frac{1}{j\eta} - \frac{\alpha}{\eta^2}\right)q^0\right].
\]

3 Finite-Difference Scheme

Computational algorithm is developed for numerical solution of the system of two second-order equations (2) with the initial data (3). The unknown variables are the tangential stress \(q\) and the angular velocity \(\omega\) within computational domain. Boundary conditions are defined in terms of \(q, \omega\) and also \(q_x, \omega_x, q_y, \omega_y\). The explicit finite-difference scheme “cross” of the second order approximation by \(x, y\) and \(t\) is used [9].

Equations of the system (2) at each time step are approximated by replacing the derivatives with respect to time and spatial variables by the finite differences:
\[
\frac{q^n_{j1,j2} - 2q^n_{j1,j2} + q^{n-1}_{j1,j2}}{(\Delta t)^2} + \frac{\alpha}{\eta} \left(\frac{q^{n+1}_{j1,j2} - q^{n-1}_{j1,j2}}{\Delta t}\right) + \frac{\alpha}{\rho} \left(\frac{q^n_{j1,j2} - q^n_{j1,j2}}{\Delta t}\right) = 8\nu^2\left(\frac{q^n_{j1,j2} - q^n_{j1,j2}}{(\Delta x)^2}\right) + \frac{\alpha}{\rho} \left(\frac{q^n_{j1,j2} - q^n_{j1,j2}}{(\Delta y)^2}\right),
\]
\[
\omega_{j_1,j_2}^{n+1} - 2 \omega_{j_1,j_2}^n + \omega_{j_1,j_2}^{n-1} = \frac{1}{\Delta t} \left( q_{j_1,j_2}^{n+1} - q_{j_1,j_2}^{n-1} \right) \\
\frac{\omega_{j_1,j_2}^{n+1} - 2 \omega_{j_1,j_2}^n + \omega_{j_1,j_2}^{n-1}}{(\Delta x)^2} - \frac{\omega_{j_1,j_2}^{n+1} - 2 \omega_{j_1,j_2}^n + \omega_{j_1,j_2}^{n-1}}{(\Delta y)^2} = \\
\frac{\gamma}{j} \left( \omega_{j_1+1,j_2}^n - 2 \omega_{j_1,j_2}^n + \omega_{j_1-1,j_2}^n \right)
\]

where \( j_1 = 2, N_1 - 1, j_2 = 2, N_2 - 1 \). Next, \( \omega_{j_1,j_2}^{n+1} \) can be expressed from the second equation:

\[
\omega_{j_1,j_2}^{n+1} = 2 \omega_{j_1,j_2}^n - \omega_{j_1,j_2}^{n-1} + \frac{\Delta t}{j} \left( q_{j_1,j_2}^{n+1} - q_{j_1,j_2}^{n-1} \right) + \\
\frac{\gamma (\Delta t)^2}{j} \left( \omega_{j_1+1,j_2}^n - 2 \omega_{j_1,j_2}^n + \omega_{j_1-1,j_2}^n \right) + \frac{\omega_{j_1,j_2+1}^n - 2 \omega_{j_1,j_2}^n + \omega_{j_1,j_2-1}^n}{(\Delta y)^2} \equiv \hat{\omega}.
\]

Substituting (4) into the first equation, we obtain the formula for \( q_{j_1,j_2}^{n+1} \):

\[
\left( \frac{\alpha}{j} + \frac{\alpha}{\eta \Delta t} + \frac{1}{(\Delta t)^2} \right) q_{j_1,j_2}^{n+1} = \frac{2}{(\Delta t)^2} q_{j_1,j_2}^n + \\
\frac{\alpha}{j} \left( q_{j_1+1,j_2}^n - 2 q_{j_1,j_2}^n + q_{j_1-1,j_2}^n \right) + \frac{\alpha}{\eta \Delta t} \left( \omega_{j_1,j_2}^{n-1} - \omega_{j_1,j_2}^n \right) + \\
\frac{\alpha}{\rho} \left( q_{j_1,j_2+1}^n - 2 q_{j_1,j_2}^n + q_{j_1,j_2-1}^n \right) - \frac{\alpha \gamma \Delta t}{j} \left( \omega_{j_1+1,j_2}^n - 2 \omega_{j_1,j_2}^n + \omega_{j_1-1,j_2}^n \right) + \frac{\omega_{j_1,j_2+1}^n - 2 \omega_{j_1,j_2}^n + \omega_{j_1,j_2-1}^n}{(\Delta y)^2} \equiv \hat{q}.
\]

Calculating tangential stress by the formula (5) and then angular velocity by the formula (4) at each time step, one can find numerical solution of the problem.

Finite-difference scheme has the second-order approximation by time and spatial variables. According to the Lax theorem, the sequence of approximate solutions converges to the exact solution with the second order, too.

### 4 Stability of the Scheme

Under analysis of the stability of the finite-difference scheme for simplicity let’s neglect the viscous term tending \( \eta \to \infty \). This simplification is based on the assumption that viscosity increases the reserve of stability of the scheme. According to the Fourier method, let

\[
q_{j_1,j_2}^n = \lambda^n \hat{q} e^{i(j_1 \alpha_1 + j_2 \alpha_2)}, \quad \omega_{j_1,j_2}^n = \lambda^n \hat{\omega} e^{i(j_1 \alpha_1 + j_2 \alpha_2)}.
\]

Substituting these values into the first equation of the system (2) and dividing both sides of the equation by \( \lambda^n e^{i(j_1 \alpha_1 + j_2 \alpha_2)} \), we get:

\[
\frac{\lambda - 1 + 1/\lambda}{(\Delta t)^2} \hat{q} + \alpha \frac{\lambda - 1/\lambda}{\Delta t} \hat{\omega} = \frac{\alpha}{\rho} \left( e^{i \alpha_1} - 2 + e^{-i \alpha_1} \right) + \frac{e^{i \alpha_2} - 2 + e^{-i \alpha_2}}{(\Delta y)^2} \hat{q}.
\]
Consequently,
\[
\left( \frac{\lambda^2 - 2 \lambda + 1}{(\Delta t)^2} + \frac{4 \alpha \rho}{\lambda} \left( \frac{\sin^2(\alpha_1/2)}{(\Delta x)^2} + \frac{\sin^2(\alpha_2/2)}{(\Delta y)^2} \right) \right) \dot{q} + \alpha \frac{\lambda^2 - 1}{\Delta t} \dot{\omega} = 0.
\]

After similar calculations for the second equation of (2) we find:
\[
\left( \frac{\lambda^2 - 2 \lambda + 1}{(\Delta t)^2} + \frac{4 \gamma j}{\lambda} \left( \frac{\sin^2(\alpha_1/2)}{(\Delta x)^2} + \frac{\sin^2(\alpha_2/2)}{(\Delta y)^2} \right) \right) \dot{\omega} - \frac{1}{j} \frac{\lambda^2 - 1}{\Delta t} \dot{q} = 0.
\]

To obtain the characteristic equation, we form the matrix of coefficients under $\dot{q}$ and $\dot{\omega}$:
\[
\begin{vmatrix}
\frac{(\lambda - 1)^2}{(\Delta t)^2} + \frac{4 \alpha}{\rho} \lambda A & \alpha \frac{(\lambda - 1)(\lambda + 1)}{\Delta t} \\
\frac{\lambda^2 - 1}{\Delta t} & \frac{(\lambda^2 - 1)^2}{(\Delta t)^2} + \frac{4 \gamma j}{\lambda} A
\end{vmatrix} = 0,
\]

where $A = \frac{\sin^2(\alpha_1/2)}{(\Delta x)^2} + \frac{\sin^2(\alpha_2/2)}{(\Delta y)^2}$. Introducing the notations
\[
a = \frac{\alpha}{\rho} A (\Delta t)^2, \quad b = \frac{\gamma j}{\lambda} A (\Delta t)^2, \quad c = \frac{\alpha}{j} (\Delta t)^2,
\]
one can calculate the determinant:
\[
(\lambda - 1)^4 + 4 \lambda (\lambda - 1)^2 (a + b) + 16 \lambda^2 a b + (\lambda^2 - 1)^2 c = 0,
\]
\[
(1 + c)(\lambda^2 - 1)^2 + 4 \lambda (\lambda - 1)^2 (a + b - 1) + 16 \lambda^2 a b = 0.
\]

So, let us consider three cases with different values $a$, $b$ and $c$:

1) If $b = 0$, then $(1 + c)(\lambda + 1)^2 + 4 \lambda (a - 1) = 0$, $\lambda^2 + 2 \lambda \left( 1 - 2 \frac{1 - a}{1 + c} \right) + 1 = 0$,
\[
\lambda_1 = \bar{\lambda}_2, \quad |\lambda_1| = |\lambda_2| = 1, \quad \left( 1 - 2 \frac{1 - a}{1 + c} \right)^2 - 1 \leq 0,
\]
\[
-1 \leq 1 - 2 \frac{1 - a}{1 + c} \leq 1, \quad 2 \geq 2 \frac{1 - a}{1 + c} \geq 0, \quad a \leq 1.
\]
Therefore, in this case we obtain the stability condition $\forall \alpha_1, \alpha_2$:
\[
\frac{\alpha}{\rho} (\Delta t)^2 \leq \left( \frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2} \right)^{-1}.
\]

2) If $a = 0$, then $(1 + c)(\lambda + 1)^2 + 4 \lambda (b - 1) = 0$ and, similarly to the previous case, one can find that $b \leq 1$. The stability condition is as follows:
\[
\frac{\gamma}{j} (\Delta t)^2 \leq \left( \frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2} \right)^{-1}.
\]
3) If $a + b = 1$, then $(1 + c)(\lambda^2 - 1)^2 + 16 \lambda^2 a b = 0$. Making the substitution $z = \lambda^2$, we obtain $z^2 - 2 z \left(1 - 8 \frac{a b}{1 + c}\right) + 1 = 0$,

$$|z_1| = |z_2| = 1, \quad \left(1 - 8 \frac{a b}{1 + c}\right)^2 - 1 \leq 0, \quad -1 \leq 1 - 8 \frac{a b}{1 + c} \leq 1,$$

$$4 a b \leq 1 + c, \quad 4 a b \leq (a + b)^2 + c, \quad 0 \leq (a - b)^2 + c.$$

The latter condition is satisfied automatically. The condition $a + b = 1$ means that

$$\left(\frac{\alpha}{\rho} + \frac{\gamma}{j}\right)(\Delta t)^2 \leq \left(\frac{\sin^2(\alpha_1/2)}{(\Delta x)^2} + \frac{\sin^2(\alpha_2/2)}{(\Delta y)^2}\right)^{-1}.$$

Under such choice of time step, $|\lambda| = 1$ for given values $\alpha_1$ and $\alpha_2$.

Thus, in the general case, the following stability condition takes place:

$$\left(\frac{\alpha}{\rho} + \frac{\gamma}{j}\right)(\Delta t)^2 \leq \left(\frac{1}{(\Delta x)^2} + \frac{1}{(\Delta y)^2}\right)^{-1}.$$

5 Parallel Program

The described algorithm for numerical solution of the system (2) by the formulas (4), (5) is implemented as a parallel program in the C language using the CUDA technology for computer systems with graphic accelerators, which allows to significantly increase the computing performance [10]. GPU (Graphics Processing Unit) is focused on the implementation of programs with a large amount of computation. Due to the large number of parallel working cores, it turns an ordinary computer into a supercomputer with the computing speed of hundreds of times higher than the PC, using only the computing power of the CPU. All computations are performed on the GPU, which is a coprocessor to the CPU. The computational domain is divided into square blocks containing the same number of threads. Each block is an independent set of interacting threads, threads of different blocks can not communicate with each other. Due to the identifiers available in the CUDA, each thread is associated with the mesh of finite-difference grid. In parallel mode, the threads of a graphic device perform operations of the same type in the meshes of grid on the calculation of solution at each time step.

Parallel program has the following structure:

1. Setting the dimensions of finite-difference grid and all the constants used (on the CPU).
2. Description of one-dimensional arrays for tangential stress and angular velocity (on the CPU).
3. Setting the initial data for these variables at the nodes of the finite-difference grid (on the CPU).
4. Description of the events *start* and *stop* measuring the program execution time on the GPU, beginning of the measuring time, beginning of the parallel part of the program.

5. Copy of the constants needed for computations from the CPU to the GPU.

6. Description of the arrays for angular velocity and tangential stress, and also for all necessary auxiliary quantities, allocation of memory for them (on the GPU).

7. Copy of the data from the CPU to the GPU (arrays of the unknown quantities).

8. Setting the variables of the dim3 type for the number of blocks in the grid and the number of threads in each of these blocks (on GPU).

9. The main computational cycle with respect to time, in which the procedures-kernels are executed sequentially (on GPU):
   (a) setting the boundary conditions in the $x$ direction;
   (b) setting the boundary conditions in the $y$ direction;
   (c) solving the system of equations for tangential stress and angular velocity by means of the finite-difference scheme “cross”; after performing of cores, the barrier synchronization is necessary, to ensure completion of the computations by each thread before starting the next computations;
   (d) copy of computational results from the GPU to the CPU (arrays of angular velocity and tangential stress) at the control points (in certain time steps).

10. Free of memory of the variables on the GPU.

11. Ending of the measuring time on the GPU, print of this time, destruction of the events *start* and *stop*, completion of the parallel part of the program.

12. Free of memory of variables on the CPU, completion of work of the program.

Here you can see the part of the program code for computation of tangential stress and angular velocity at the internal nodes of finite-difference grid by each thread of the GPU:

```c
__global__ void syst_qw(int it, double *q, double *q1, \
                         double *q2, double *w, double *w1, double *w2) 
{
    int ix, iy, id, idm_x1, idp_x1, idm_x2, idp_x2;
    double c1, c2, c3, c4, c5, c6, c7, c8, Delta_q, Delta_w;
    ix=threadIdx.x+blockIdx.x*blockDim.x;
    iy=threadIdx.y+blockIdx.y*blockDim.y;
    if ((ix > 0) && (ix < Nx1Dev-1) && (iy > 0) && (iy < Nx2Dev-1))
    {
        id=IDX1X2(ix,iy,Nx1Dev);
```
\begin{verbatim}

\text{idm}_x1=IDX1X2(ix-1,iy,Nx1Dev); \text{idp}_x1=IDX1X2(ix+1,iy,Nx1Dev);
\text{idm}_x2=IDX1X2(ix,iy-1,Nx1Dev); \text{idp}_x2=IDX1X2(ix,iy+1,Nx1Dev);
c1=alfaDev/jiDev; \quad c2=gamDev/jiDev;
c3=alfaDev/roDev; \quad c4=1./tauDev/tauDev;
c5=2.*alfaDev/tauDev; \quad c6=-c1*gamDev*tauDev;
c7=tauDev/jiDev; \quad c8=alfaDev/etaDev/tauDev;
\text{Delta}_q=(q1[idp}_x1\text{-}2.0*q1[id]+q1[idm}_x1)/h1Dev/h1Dev; \quad \text{Delta}_q+=\text{(q1[idp}_x2\text{-}2.0*q1[id]+q1[idm}_x2)/h2Dev/h2Dev;}
\text{Delta}_w=(w1[idp}_x1\text{-}2.0*w1[id]+w1[idm}_x1)/h1Dev/h1Dev; \quad \text{Delta}_w+=(w1[idp}_x2\text{-}2.0*w1[id]+w1[idm}_x2)/h2Dev/h2Dev;
q2[id]=((c1-c4+c8)*q[id]+2.*c4*q1[id]+c5*(w[id]-w1[id])
+c6*Delta_w+c3*Delta_q)/(c1+c4+c8);
w2[id]=2.*w1[id]-w[id]+c7*(q2[id]-q[id])+c2*Delta_w/c4;
\__syncthreads();
\}

\__host\__ void system_qw(int it, dim3 blocks, dim3 threads, \
dim3 blocks1, dim3 threads1, dim3 blocks2, \
dim3 threads2, double *t, double *q, double *q1, \
double *q2, double *w, double *w1, double *w2)
{
... \quad \text{syst}_qw <<blocks,threads>>>(it,q,q1,q2,w,w1,w2);
\quad \text{cudaThreadSynchronize();}
\}

Previously, in solving the problem without taking into account the couple-stress interactions, the efficiency of the parallel program for complete system of equations of a model was analyzed [7]. To evaluate the efficiency of parallelization, a large number of computations was performed at different grid dimensions. The computation time for parallel program and sequential program was compared. Fig. 1 shows a graph of the dependence of acceleration of the parallel

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig1.png}
\caption{Acceleration of the program on GPU as compared with CPU}
\end{figure}
\end{verbatim}
program on the dimension $N \times N$ of a finite-difference grid. Here $N$ takes the values: 10, 100, 200, 400, 600, ..., 3000, 3200. The acceleration $t_{CPU}/t_{GPU}$ of the parallel program as compared with the corresponding sequential program is about 25 times on the grids of dimension $1000 \times 1000$ and above.

6 Exact Solution

For one-dimensional problem on the action of tangential stress $q = \hat{q} e^{i(ft-ky)}$ at one of the boundaries of computational domain, the comparison of the numerical solution by described parallel program and the exact solution was carried out.

Substituting $q = \hat{q} e^{i(ft-ky)}$, $\omega = \hat{\omega} e^{i(ft-ky)}$ in the equations of system (2) after simplification we obtain:

$$
(-f^2 + 2i\frac{\alpha f}{\eta} + \frac{\alpha k^2}{\rho})\hat{q} + 2i\frac{f}{j} \hat{q} + \left(\frac{\gamma k^2}{j} - f^2\right)\hat{\omega} = 0,
$$

$$
-2i\frac{f}{j} \hat{q} + \left(\frac{\gamma k^2}{j} - f^2\right)\hat{\omega} = 0.
$$

Calculating the determinant of this system, one can find the expression for $k^\pm$:

$$
k^\pm = \sqrt{\frac{\rho j f}{2\alpha\gamma} \left(d \pm \sqrt{d^2 - 4\frac{\alpha\gamma}{\rho j} \left(f^2 - 2i\frac{\alpha f}{\eta} - 4\frac{\alpha}{\eta j}\right)}\right)}, \quad d = \left(\frac{\alpha}{\rho} + \frac{\gamma}{j}\right)f - 2i\frac{\alpha}{\eta j}.
$$

Here $f$ is the frequency, $k^\pm = k^\pm_1 + i k^\pm_2$ are the wave numbers.

The characteristic dispersion curves are represented in Figs. 2 and 3: dependence of the phase velocity $c^\pm = \frac{f}{\text{Re} k^\pm}$ on the frequency $\nu = \frac{f}{2\pi}$ and dependence of the damping decrement $\lambda^\pm = -\frac{1}{\text{Im} k^\pm}$ on the frequency $\nu$ (for $k^+$ and $k^-$ – left and right, respectively). The dashed line corresponds to the eigenfrequency of rotational motion of the particles of a crystal: $\nu_\star = \frac{1}{\pi} \sqrt{\frac{\alpha}{j}}$.

Computations were performed for the liquid crystal 5CB with the next parameters [11]: $\rho = 1022 \text{ kg/m}^3$, $j = 1.33 \cdot 10^{-10} \text{ kg/m}$, $\alpha = 0.161 \text{ GPa}$, $\gamma = 10 \mu\text{N}$, $\eta = 10 \text{ Pa} \cdot \text{s}$. For this crystal $\nu_\star = 350 \text{ MHz}$. The size of a domain is $4 \mu\text{m}$.

Initial data and boundary conditions for one-dimensional problem are determined from the next equations:

$$
\text{Re} \hat{q} = e^{k2y}(\hat{q}_1 \cos(ft - k_1y) - \hat{q}_2 \sin(ft - k_1y)),
$$

$$
\text{Re} \hat{\omega} = e^{k2y}(\hat{\omega}_1 \cos(ft - k_1y) - \hat{\omega}_2 \sin(ft - k_1y)).
$$

Figure 4 shows results of numerical solution for described above parameters: dependence of $\text{Re} \omega$ on $y$ for $k^+$ and $k^-$ at one of the instants of time. The dimension of a finite-difference grid is 1000 meshes. The relative error is $3 \cdot 10^{-3}$ in calculations for $k^+$ and $5 \cdot 10^{-4}$ in calculations for $k^-$. 

482
Fig. 2. Dependence of phase velocity on frequency: a) for $k^+$, b) for $k^−$.

Fig. 3. Dependence of damping decrement on frequency: a) for $k^+$, b) for $k^−$.

Fig. 4. Dependence of angular velocity on coordinate: a) for $k^+$, b) for $k^−$. 
7 Results of Computations

A series of numerical calculations was carried out on the high-performance computational server Flagman with eight graphic solvers Tesla C2050 (448 CUDA cores on each GPU) of the Institute of Computational Modeling SB RAS to demonstrate the efficiency of proposed parallel program.

In Fig. 5 one can see the results of computations for the problem on the action of three $\Pi$-shaped impulses of tangential stress on the parts of lateral boundaries of computational domain. Duration of each impulse and the interval between them are 8 ns. Initial data are zero. The boundary conditions at the left and right boundaries: $q = \bar{q}$, if $|y - y_c| \leq l$, and $q = 0$, if $|y - y_c| > l$; $\omega_x = 0$. Here $y_c$ is the center of zone, where the load acts, $l$ is the radius of this zone. In computations $y_c = 20 \mu m$, $l = 10 \mu m$. At the upper and lower boundaries the periodicity conditions are given. The size of rectangular computational domain is $100 \mu m \times 40 \mu m$, the dimension of a finite-difference grid is $2560 \times 1024$ meshes. Computations were performed for the liquid crystal 5CB with the parameters: $\rho = 1022 \text{kg/m}^3$, $j = 1.33 \cdot 10^{-7} \text{kg/m}$, $\alpha = 0.161 \text{GPa}$, $\gamma = 1 \text{mN}$, $\eta = 100 \text{Pa \cdot s}$.

![Fig. 5](image1.png)

Fig. 5. The action of three $\Pi$-shaped impulses of tangential stress at the lateral boundaries: level curves of tangential stress at different instants of time.

Figure 6 shows the results of computations for the problem on the action of a concentrated tangential stress $q = \bar{q} \delta(x) \delta(t)$ at the upper boundary. So, at this boundary: $q = \bar{q}$, if $|x - x_c| \leq \Delta x$ ($x_c = 7 \mu m$ is the point of loading) and $t \leq \Delta t$, and $q = 0$, otherwise; $\omega_y = 0$. The lower boundary is fixed: $q = 0$, $\omega = 0$. At the left and right boundaries the periodicity conditions are given. The size of computational domain is ten times less than in the previous problem: $10 \mu m \times 4 \mu m$. The dimension of a grid is the same: $2560 \times 1024$ meshes. Computations were performed for the crystal 5CB with parameters as in the previous case, but $j = 1.33 \cdot 10^{-10} \text{kg/m}$, $\gamma = 10 \mu N$. In Fig. 6 a the level curves of tangential stress, and in Fig. 6 b the level curves of angular velocity at different instants of time are represented.

![Fig. 6](image2.png)
Fig. 6. Lamb’s problem for the action of concentrated tangential stress at the upper boundary: level curves of tangential stress (a) and angular velocity (b) at different instants of time.

Figure 7 demonstrates numerical results for the problem on periodic action of a tangential stress on the part of upper boundary. The boundary conditions at the upper boundary: \( q = \bar{q} \sin(2 \pi \nu t) \), if \( |x - x_c| \leq l \), and \( q = 0 \), if \( |x - x_c| > l \) (\( x_c = 5 \mu m \), \( l = 2.5 \mu m \)); \( \omega_{y} = 0 \). As before, the lower boundary is fixed, at the left and right boundaries the periodicity conditions are given. The frequency \( \nu \) is equal to \( \nu_s = 350 \) MHz. Computations were performed with parameters as in the previous case. The fields of angular velocity are shown in Fig. 7.

Fig. 7. Periodic action of tangential stress at the upper boundary: fields of angular velocity at different instants of time.
8 Conclusions

Within the framework of acoustic approximation of the model of a liquid crystal, the system of second-order equations for tangential stress and angular velocity taking into account the couple-stress interactions was derived. The parallel computational algorithm for numerical solution of this system was developed. A comparison of exact and numerical solutions for one-dimensional problem was fulfilled, and good correspondence of results was obtained. Some computations were performed by means of the parallel program using the CUDA technology for computer systems with graphic accelerators.

Acknowledgments. This work was partially supported by the Complex Fundamental Research Program no. II.2P “Integration and Development” of SB RAS (project no. 0356–2016–0728), the Russian Foundation for Basic Research (grants no. 14–01–00130, 16–31–00078).

References

Biochemical Processes of Self-Purification Model in Small Rivers

Olga Taseiko¹, Tatiana Spitsina¹, Hranislav Milosevic², Dragan Radovanovic², and Aleksandar Valjarevic²

¹Siberian State Aerospace University, Krasnoyarsky Rabochy Av., 31, Krasnoyarsk, Russia
²Faculty of Science and Mathematics, Universityof Pristina, Lole Ribara b.b. 38220 Kosovska Mitrovica, Serbia
taseiko@gmail.com

Abstract. The studying of water self-purification ability is a part of the regional limits developing problem that has the important role for the water quality management. This work shows that a simple model structure can be set up to describe the water quality in small river basins in terms of carbon, nitrogen and phosphorus compounds, when it is unfeasible to use complex models. In this article we used both mathematical modeling and natural sampling of surface water in small river of Central Siberia for the control parameters assessment. The obtained results have allowed analyzing the annual variations of the nitrification and denitrification rates, the mineralization rate of total phosphorus and organic nitrogen. The contribution of main biochemical processes in self-purification of the small river under conditions of Central Siberia climate is estimated numerically.

Keywords: self-purification processes, one-dimensional advective-diffusive equation, upwind approximation scheme.

1 Introduction

Water quality management is an essential problem for preserving water resources and facilitating sustainable socio-economic development in watershed systems. However, this task is usually affected by a variety of uncertainties raising from the stochasticity in hydrodynamic conditions, the variability in the pollutant transport, the physicochemical processes, the indeterminacy of available water and wastewater, etc. [1]. The studying of water self-purification ability is a part of the problem of the regional limits developing of water quality that has the important role for the water quality management. Mechanisms of self-purification processes are strongly influenced by local characteristics of stream. A model developed for a certain stream type and region is in many cases not applicable to other stream types or regions. Therefore, local stream characteristics should be included if the model should be more generally applicable [2].

Self-purification of natural water systems is a complex process that often involves physical, chemical, and biological processes working simultaneously. The
water is purified in the sense that the concentration of waste material has been reduced mostly by means of biodegradation processes. Therefore, this process is very closely tied with the dissolved oxygen content and indeed with all the sources and sinks of oxygen in a river. So dissolved oxygen (DO) and biochemical oxygen demand (BOD) are critical water quality parameters. [3].

Research on the modeling of the BOD-DO interaction in the river has been dominated by the classical model of Streeter and Phelps, which first appeared in 1925, and has been improved by Dobbins and Camp, Peter Young and Bruce Beck, C.J.Harris, etc. [3]. These models have been widely used not only in to assess water quality, but also to predict damage resulting from the implementation of water resources management measures [4]. However, each water quality model has its own limiting conditions. Therefore, these models still need to be further studied to overcome these shortcomings.

Eutrophication of surface water is closely connected to the self-purification processes. Eutrophication is enrichment in nutrients, principally phosphorus and nitrogen, leading to an increase in algae and higher plant growth and a disturbance of the ecological balance of the aquatic ecosystem. In contrast to standing waters, the effects of eutrophication and enhanced organic load on running water ecosystems have not been given much attention [5], [6]. Models that include all these factors are not yet available but there are many models which focus on only a part of these processes. Some of these parts can be used to fill in a complete stream eutrophication model. However, because of differences between stream types and regions these parts should first be tested for their applicability to the stream type and region of interest [2].

The present study is based on and continues the project aimed at the studying of self-purification processes under strong anthropogenic exposure in small rivers of Krasnoyarsk region that was started by authors in 2013. The principal equations of the developed self-purification model are specified in [7]. The developed model generally has shown a satisfactory capability in reproducing the measured values of nitrogen and phosphorus concentrations. The main goal of this paper is to study the contribution of biochemical transformation of biogens in self-purification processes of small rivers. For this goal it was developed mathematical approach for estimation of some biochemical parameters such as reaeration and biodegradation rates, transformation rates of phosphorus and nitrogen compounds. Then it were calculated the amounts of these rates for small river of Central Siberia and finally, were described the regional features of ones varieties depends on hydrological conditions.

2 Object description

The Kacha river considering in this study is the river in the basin of Central Enisey. Hydraulically the river is subjected to spring flood, but water level reduces significantly during the summer months, when the water flow quality becomes critical and the self-purification processes are almost stopped. Sharp continental climate of Central Siberia, basin geology and vegetation define hy-
drological conditions of river flow. So, river flow rate and flow velocity differ significantly in various hydrological stages. For example, flow rate varies in the range 0.1 - 41 m3/s, maximum level reaches in spring flood. The maximal value of water temperature above 20 C is observed in June. All factors define regional features of the eutrophication processes.

To verify the developed model it was used the data from state monitoring network for the period since 1985 to 2014 in Kacha river. The hydrochemical parameters are measured one time in month (7-12 times in year). In this work we use the concentrations of oxygen, nitrogen, phosphorus and their compounds. The hydrological parameters such as river flow rate, temperature, stream velocity are measured every day. All measurements have carried out in three hydrological posts of Kacha river whose basin length is about 100 km.

Moreover, some complex parameters were measured in Kacha river during 2013 - 2016: pH, dissolved oxygen, BOD, redox potential and conductivity. An area of sampling lies near one of hydrological posts in estuary. These parameters measured two times in week during free ice cover period (since May to October).

3 Model structure

One of the difficult problems of self-purification studying - the variety of ecological structure from source to mouth of river [8]. The self-purification processes in the river are complex and can be described by a series of bio-chemical and hydrological parameters. Biochemical oxidation process through which organic wastes are consumed leaving behind end products such as carbons, phosphates and nitrates [3].

Inorganic carbon availability is determined by levels of dissolved carbon dioxide. Concentration of carbon dioxide is less significantly than dissolved oxygen concentration. So, carbon dioxide isn’t key element in self-purification processes in the Siberian water ecosystems as this can be in equatorial and subequatorial ecosystems. In continental climate conditions carbon dioxide in water streams has influence mainly on redox processes. Organic carbon is included in the model indirectly via biochemical oxygen demand (BOD). The transformation processes of phosphates and nitrates are coupled in the model directly.

In general the one-dimensional advective-diffusive dynamic for reactive pollutant neglecting the diffusion term can be written as a differential equation [9], [10], [11]:

$$\frac{d(\omega \cdot C_j)}{dt} + \frac{d(Q(t) \cdot C_j)}{dx} = K_{C_j}(t) \cdot C_j \cdot \omega + G_j(t) \cdot \omega$$

(1)

where $K_{C_j}(t)$ is a function of decay of $j$ pollutant concentration, that characterizes transformation velocity defined by the influence of chemical and biological processes, $Q(t)$ is a function of river flow rate, $\omega$ is cross-sectional area of river ($m^2$) and $G_j(t)$ is a runoff of $j$ pollutant.

The equations system based on this (1) includes the equations for concentration of phosphate $C_{PO4}$, total phosphorus $C_{DOP}$, ammonium nitrogen $C_{NH4}$,
4 An intensity estimation of biochemical transformation processes

All model’s rates vary with temperature, microbial metabolism, the composition and concentration of the biogens from the pollution source.

The estimation of $K_{RO}$ and $K_{BOD}$ parameters is important for selecting a solution curve that best represents a real system. However, there is no method available to determine values that fit precisely to the reality of a given water body. Reaeration coefficients vary widely due to their dependence on air-water interface turbulence making them complex and difficult to accurately measure. High nutrient levels result a high biomass of algae and plants and an increased biodegradation rate. If algae and plants produce oxygen during the day, they consume this during the night and an increased biomass means an increase uptake of oxygen at night. It also means an increase in organic matter when the organisms die. Decomposition of organic matter is increased in presence of high nutrient level, there is more organic matter to decompose and decomposition consumes oxygen. All these processes can lead to oxygen depletion [4], [13].

The highest $K_{RO}$ values may be observed during summer due to the higher concentration of organic matter while maintaining a contribution of organic matter in the stream. A high temperature of water river promotes this process (greater 20$^\circ$C in July). The classical model of Streeter and Phelps defines the ratio of $K_{RO}/K_{BOD}$ as the self-purification constant and it is equal 0.50–5.0. Several studies give methods to estimate $K_{RO}$ and $K_{BOD}$ that provide reasonable approximations within predefined limits. However, due to the non-linearity nature of these coefficients, there is no formula for generic cases.

The nitrification and denitrification velocity depends on temperature and pH value of surface water. For example, the denitrification process reaches maximal activity, when pH value is in the range 7.0–8.2. This process is stopping when the pH value is lower than 6.1 or higher than 9.6. Nitrogen mineralization is considered as transformation of organic nitrogen to inorganic.
Primarily, decay rates were included in the self-purification model as constants. However, the results of numerical calculations were differing from natural measurements considerably. This can was connected with significantly seasonal variability of these rates.

The algorithm of calculation of destruction rates includes next steps. Firstly, the equation (1) is written as the system including seven equations for next variables: concentrations of phosphorus compounds $C_{DOP}$ and $C_{PO_4}$, ammonium compounds $C_{NH_4}$, $C_{NO_3}$ and $C_{NO_3}$, biochemical oxygen demand $C_{org}$, and dissolved oxygen $C_{O_2}$. These equations were specified in [7]. Next the equations for biochemical destruction rates are obtained by algebraic manipulations with previous system. Finally, the rate’s equations are written according numerical scheme. As a result of algorithm’s applying it was obtained a system of six numerical equations:

1) Mineralization rate of total phosphorus:

\[
K(t)_{PO_4} = \frac{1}{C_{DOP}} \cdot \left( -\frac{d(C_{PO_4})}{dt} - \frac{1}{\omega} \cdot \frac{d(Q(t) \cdot C_{PO_4})}{dx} + G(t)_{PO_4} \right)
\]  

(2)

2) Nitrogen mineralization rate:

\[
K_{NH_4}(t) = \frac{1}{C_{DON}} \cdot \left( \frac{d(C_{DON})}{dt} + \frac{1}{\omega} \cdot \frac{d(Q(t) \cdot C_{DON})}{dx} - G(t)_{DON} \right)
\]  

(3)

3) Nitrification rate:

\[
K(t)_{12} = \frac{1}{C_{NH_4}} \cdot \left( -\frac{d(C_{NH_4})}{dt} - \frac{1}{\omega} \cdot \frac{d(Q(t) \cdot C_{NH_4})}{dx} + G(t)_{NH_4} + K(t)_{NH_4} \cdot C_{DON} \right)
\]  

(4)

4) Denitrification rate:

\[
K(t)_{NO_3} = \frac{1}{C_{NO_3}} \cdot \left( -\frac{d(C_{NO_3})}{dt} - \frac{1}{\omega} \cdot \frac{d(Q(t) \cdot C_{NO_3})}{dx} + G(t)_{NO_3} + K(t)_{12} \cdot C_{NH_4} \right)
\]  

(5)

5) Biochemical degradation rate:

\[
K(t)_{BOD} = \frac{1}{C_{org}} \cdot \left( -\frac{d(C_{org})}{dt} - \frac{1}{\omega} \cdot \frac{d(Q(t) \cdot C_{org})}{dx} + K(t)_{NO_3} \cdot C_{NO_3} \cdot \beta_{O_2/DN} \right)
\]  

(6)

6) Reaeration rate:

\[
K(t)_{RO} = \frac{1}{C_{O_2}} \cdot \left( -\frac{d(C_{O_2})}{dt} - \frac{1}{\omega} \cdot \frac{d(Q(t) \cdot C_{O_2})}{dx} + K(t)_{12} \cdot C_{NH_4} \cdot \beta_{O_2/NT} + K(t)_{BOD} \cdot C_{org} \right)
\]  

(7)
where $\beta_{O_2/DN}$ is the yield factor describing the amount of oxygen used for denitrification ($gO_2/gN$), $\beta_{O_2/NT}$ is the yield factor describing the amount of oxygen used for nitrification ($gO_2/gN$).

Time-space grid for the equations (2)-(7) is the same as for equation (1).

Next step connects to defining run-off values. Low-water small rivers of Eastern Siberia are recharged mostly by groundwater in winter. This occurs because all precipitation falls in a solid phase and there is no their thawing. The water hardness value is the indicator of increase in a share of an underground water. In spring and summer this parameters is much lower because of influence of liquid precipitation. The volume of groundwater run-off is calculated with using the hydrochemical analysis of surface water quality of Kacha river. The greatest values characterize the groundwater run-off of nitrate nitrogen.

5 Results and discussion

The observation results received on three hydrological posts of the state monitoring network during 29 years were used as input data to calculate the biochemical parameters. The calculation results for whole time period were averaged monthly. In this study the length of computational domain was 100 km, space step was 0.5 km and time step was 1 day.

![Fig. 1. Seasonal variations of nitrogen destruction rates for self-purification model](image)

Some results of calculations by the system (2) - (7) are shown on Figure 1 - 2. Receiving numerical functions for describing the transformation rates allows considering an influence of all factors without construction of functional dependences. In general, the values of transformation rates demonstrate inhibition of self-purification processes in river.

Seasonal factor has influence on variations of all modeling parameters. The maximal values of reaeration rate are obtained for period from end of April to early in June. This can be explained by flood peak in Kacha river when flow rate and water level are highest. A decreasing of the nitrogen mineralization rate is induced by some reasons having regional features such as low flow rate, low concentration of dissolved oxygen and high level of chemical pollution. The
biodegradation rate depends mostly on phytoplankton activity that is minimal in summer low water (during July). The value of this rate is increasing significantly in the period of spring and autumn high water seasons. The obtained dependencies of biochemical rates characterize both common seasonal variations features and regional specialty, for instance, high water level during flood, short vegetation period and low water temperature.

Variability ranges of the transformation rates and parameters of developing model are presented in Table 1. The variations of biochemical purification rates are studying not often, so ones have shown as variability ranges to compare with the data that were given for other rivers. These values agree in general with values obtained in other rivers and regions. However, exact comparison is incorrect because literature data were calculated for various water bodies differing both in climatic and hydrological conditions.

All factors influencing on the variability of biochemical parameters are generalized in Table 1. A water temperature is common factor for the most of calculated rates. Numerical estimations of temperature influencing on studied rates haven’t obtained in this study. Overall on the base of water quality monitoring data it can conclude that the main processes of biogens transformation and water self-purification are observed in the period from April to October, when water temperature is higher than 0°C. Winter period is characterized the presence of ice cover on water surface, so the sampling in studied river aren’t executed. And to estimate the values of biochemical parameters in this period it’s impossible.

Table 1 presents a coefficient that was named in this study the contribution to the model accuracy. To define one the calculations with using equation (1) are executed for rates $K_{RO}$, $K_{BOD}$, $K_{12}$, $K_{NO3}$, $K_{PO4}$, $K_{NH4}$ as constant values and as functions of time (as shown in Fig. 1). The contribution to the model accuracy was obtained on the base of comparison of above mentioned calculations with monitoring data (Table 1). This coefficient characterizes numerically an influence of considering the biochemical rates like time functions in the developed model. This allows taking into account the influence of seasonal variability of all factors.

All processes in small rivers have different velocity, but the velocity of phosphorus transformation is a slowest. It can be explain that phosphorus concentra-
### Table 1. The principle parameters of self-purification model

<table>
<thead>
<tr>
<th>Rates and parameters</th>
<th>Variability range</th>
<th>Given in the literature data</th>
<th>Influencing factors to the model accuracy, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reaeration rate ( K_{RO}, \text{day}^{-1} )</td>
<td>0.09...0.55</td>
<td>0.1...0.25 [14] 0.05...0.5 [15]</td>
<td>temperature, BOD, dissolved oxygen</td>
</tr>
<tr>
<td>Biodegradation rate ( K_{BOD}, \text{day}^{-1} )</td>
<td>0.05...0.35</td>
<td>0.06 [14] 0.4...1.5 [15]</td>
<td>temperature, nitrate, nitrogen, BOD, dissolved oxygen</td>
</tr>
<tr>
<td>Nitrification rate ( K_{12}, \text{day}^{-1} )</td>
<td>0.06...0.6</td>
<td>0.027...0.76 [16] 0.04 [11]</td>
<td>temperature, pH value, nitrate and ammonium nitrogen concentration</td>
</tr>
<tr>
<td>Denitrification rate ( K_{NO_3}, \text{day}^{-1} )</td>
<td>0.7...1.1</td>
<td>0.1 [11]</td>
<td>temperature, pH value, nitrate and ammonium nitrogen concentration</td>
</tr>
<tr>
<td>Phosphorus mineralization rate ( K_{PO_4}, \text{day}^{-1} )</td>
<td>0.05...0.2</td>
<td>0.14 [11]</td>
<td>total phosphorus, phosphates</td>
</tr>
<tr>
<td>Nitrogen mineralization rate ( K_{NH_4}, \text{day}^{-1} )</td>
<td>0.01...0.45</td>
<td>0.06 [11]</td>
<td>ammonium nitrogen, total nitrogen</td>
</tr>
</tbody>
</table>

... is lower significantly than nitrogen concentration in small rivers of Central Siberia. Also, nitrogen transformation processes demonstrate the greatest contribution to model accuracy.

### 6 Conclusions

This work presents numerical algorithm for estimation of biochemical coefficients of self-purification processes based on the developed mathematical model. The given algorithm was used to define the values of reaeration and biodegradation rates, transformation rates of phosphorus and nitrogen compounds for small river in Central Siberia. It was analyzed the influence of regional hydrological and meteorological factors on temporal variations of biochemical coefficients of self-purification processes. It was estimated the contribution of developed approach to accuracy on numerical calculating of studying processes in comparison with experimental measurements.

To improve reproducing the measured values via model calculation the daily variations of the rates giving the greatest contribution to model accuracy will be studied further. The developed model can be useful to solve some problems, for instance an optimization of environmental monitoring, a forecasting of ecosystem productivity, a developing of regional water quality limitations and management of water quality.
The reported study was funded by RFBR according to the research project 15-07-06982 and III 44006 MNRS

References

Modeling an Initial Tsunami Waveform by Inverting Remote Sea Level Records Through the r - Solution Method

Tatyana Voronina

Institute of Computational Mathematics and Mathematical Geophysics of SB RAS, Academician M.A. Lavrentiev ave. 6, 630090, Novosibirsk, Russia

Abstract. Modeling the initial water displacement in the tsunami source area based on r-solution method is presented. This approach is independent of the earthquake parameters, because there are used only observed tsunami waveforms and a roughly estimated tsunami source area. The method proposed suppresses the negative effect of the ill-posedness of the problem determining the inevitable instability of the numerical solution. Furthermore, this approach allows one to obtain a more reasonable strategy for deploying the tsunami monitoring system. In this paper, the tsunami source of the 2013 Solomon Island tsunami event was reconstructed by the method proposed, and the deployment of DART Buoy monitoring system was examined for the efficiency to infer the tsunami source.

1 Introduction

The mega thrust earthquakes quite often result in large tsunamis that may inflict a severe loss and pain to the population of coastal communities. Based on the experience gained from the large Indian Ocean 2004 and the Tohoku 2011 Tsunamis the humanity recognized the importance of the real time monitoring of such severe events. Since then, concerted efforts of the scientific community have been devoted to developing the numerical simulation tools and other related technologies of the tsunami waves to mitigate the adverse impacts of plausible tsunamis.

Tsunamis are very long gravity waves, with wavelengths of tens to hundreds kilometers, which exceed the ocean depth. Under such conditions, in the deep ocean their propagation can be described by the shallow-water theory. Tsunamis can be triggered by a variety of geophysical phenomena. In the first and more common case, an earthquake occurred near the sea floor, may produce a co-seismic deformation that can cause a displacement in the sea floor that can, in turn, cause an initial sea surface deformation that may result in a tsunami wave. This sea surface deformation will be called an initial tsunami waveform or, simply, a tsunami source. To accurately forecast the inundation and run up in the near-field coast, where a warning should be issued no more than 20 minutes, it is necessary to gain the insight into a tsunami source at the early stages of tsunami propagation.
The tsunami waveform inversion has the advantage in determining a tsunami source, as compared with seismic waveform inversion because seismic data are often imprecisely translated into tsunami data. Furthermore, the tsunami wave propagation can be more accurately simulated than seismic waves due to the fact that bathymetry is better known than subsurface seismic velocity structure. Numerous studies deal with application of tsunami waveforms inversion for determining the tsunami source characteristics [e.g., Satake, 1989; Tinti et al., 1996; Piatanezi et al., 2001; C.Pires and P.M.A.Miranda, 2001; Wei et al., 2003; Titov et al., 2005; Baba et al., 2009; Percival et al., 2011; Saito et al., 2010; Tsushima et al., 2012; Mulia et al., 2016].

In this paper, the inverse problem in question is treated as an ill-posed problem of the hydrodynamic inversion with tsunami waveforms, so, it imposes some restrictions on using the mathematical techniques. In other words, any attempt to solve this inverse problem numerically must be followed by a regularization procedure. To this end, the technique based on the least-squares inversion using the truncated Singular Value Decomposition (SVD) and r-solution methods ([13]) has been proposed and was first described in its fundamentals in [12], [14]. As a result of the numerical process, the so-called r-solution is a projection of the exact solution onto a linear span of the r first right singular vectors corresponding to the largest singular values of a compact operator of the direct problem. The properties of r-solution obtained are defined to a large extent by the properties of the inverse operator which were numerically investigated in [15], [16].

The direct problem of tsunami wave propagation is considered within the scope of the linear shallow-water theory. The computer simulation is based on a finite difference algorithm. The method proposed does not require any a priori information of a source, but only its general spatial localization assumed to be known from seismological data. Furthermore, this method allows one to control the instability of the numerical solution and to obtain an acceptable result in spite of the ill-posedness of the problem.

Presently, a lot of offshore tsunami monitoring systems using submarine cabled seafloor observatory technology have been deployed in the deep ocean. The Deep-ocean Assessment and Reporting of Tsunamis (DART) buoy system, developed by the Pacific Marine Environmental Laboratory (PMEL) of the National Oceanic Atmospheric Administration (NOAA), is one of the deep-ocean tsunami observational systems [8]. The tsunami waveforms acquired by cabled offshore ocean bottom tsunami meters are more available, free of the tide gauge response functions as well as the coastal and the harbor effects. Hence, the inversion approaches based on the deep-ocean observations can be used for a rapid estimation of a tsunami source, which, in turn, can be used as direct input for the real-time forecast of the tsunami impact.

Although a considerable attention has been given to developing the inversion methods to infer the initial tsunami waveform, a lesser number of studies has been devoted to revealing the influence of such characteristics of the monitoring system as the number and spatial distribution of the recording devices on the inversion results. In order to correlate these notions, a series of numerical ex-
periments with synthetic data and different computational domains have been carried out using r-solution method ([15], [16]). As it was shown, the number r is tightly bounded with the parameters of the observational system.

The focus of this research is on the attempt of applying the regularities obtained for a more reasonable strategy of deployment of a tsunami monitoring system in reality. Results of numerical experiments are presented in the case study of Solomon Islands tsunami of 6th February 2013.

2 Model

The tsunami wave is assumed to be triggered by a sudden vertical displacement of the sea floor. The tsunami propagation can be considered within of shallow-water theory. The tsunami source area is assumed to be known from the seismological data as a rectangle $\Omega$, $\Omega \subset \Phi \subseteq \Pi$, where a rectangular domain $\Pi$ is a calculation domain and $\Phi$ is the aquatic part of $\Pi$ with the piecewise-linear solid boundaries $\Gamma$ and straight-line sea boundaries. The problem is considered in an orthogonal coordinate system. The plane $\{z = 0\}$ corresponds to the undisturbed water surface. The curvature of the Earth is neglected. The wave run up is not considered.

Let $\eta(x, y, t)$ be a function of the water surface elevation relative to the mean sea level which is considered to be a solution of the linear shallow-water equations:

$$
\begin{align*}
\eta_t + g\nabla \cdot (hV) &= 0 \\
V_t + g\nabla \eta &= 0
\end{align*}
$$

completed by the following initial conditions:

$$
\eta|_{t=0} = \varphi(x, y), \quad V|_{t=0} = 0; \tag{2}
$$

and the boundary condition on the solid boundary:

$$
V \cdot \mathbf{n} = 0 \tag{3}
$$

as well as absorbing boundary conditions (ABC) of second order accuracy are implied at the sea boundaries on the sides of the rectangle $\Pi$:

$$
\begin{align*}
c_\eta \eta_{tt} - \eta_{tt} + \frac{c^2}{2} \eta_{xx}|_{y=0} &= 0; \\
-c_\eta \eta_{tt} - \eta_{tt} + \frac{c^2}{2} \eta_{xx}|_{y=Y} &= 0; \\
-c_\eta \eta_{xt} - \eta_{tt} + \frac{c^2}{2} \eta_{yy}|_{x=X} &= 0; \\
-c_\eta \eta_{xt} - \eta_{tt} + \frac{c^2}{2} \eta_{yy}|_{x=0} &= 0; \tag{4}
\end{align*}
$$

In the above equations, the vector $V = (v_x, v_y)$ is the horizontal fluid velocity vector whose $x$- and $y$-components are, respectively, $v_x$ and $v_y$, $h(x, y)$ is the water depth relative to the mean sea level, $g$ is the gravity acceleration, $c(x, y) = \sqrt{gh(x, y)}$ is the wave phase velocity and $\mathbf{n}$ is the unit vector, outwardly directed, normal to the boundary, $\varphi(x, y)$ is the initial water displacement defined in a tsunami source area $\Omega$. 
3 Inversion method

The inverse problem at hand is to infer the unknown initial water displacement \( \varphi(x, y) \) as output while the observed tsunami waveforms as data input are assumed to be known on a set of points \( R = \{(x_i, y_i), i = 1, \ldots, P\} \) (below called as receivers):

\[
\eta(x_i, y_i, t) = \eta_0(x_i, y_i, t), \quad (x_i, y_i) \in R. \tag{5}
\]

This inverse problem is treated as an ill-posed problem of the hydrodynamic inversion with tsunami sea-level records, so it imposes some restrictions on the use of mathematical techniques. In the approach applied, regularization is performed by means of the truncated SVD that brings about the notion of \( r \)-solution (see [13]). This solution will be sought for in a least squares formulation. The application of this approach to tsunami waveforms inversion was detail described in [15], [16].

The unknown function of the water surface displacement \( \varphi(x, y) \) in the source area \( \Omega \) was sought for as a series of spatial harmonics

\[
\varphi(x, y) = \sum_{m=1}^{M} \sum_{n=1}^{N} c_{mn} \sin \frac{m\pi}{l_1} x \cdot \sin \frac{n\pi}{l_2} y \tag{6}
\]

for \((x, y) \in [0, l_1] \times [0, l_2]\), with unknown coefficients \( c = \{c_{mn}\} \).

In our case, the inverse problem data are the observed waveforms (marigrams) \( \eta = (\eta_{11}, \eta_{12}, \ldots, \eta_{1N_1}, \eta_{21}, \ldots, \eta_{2N_2}, \ldots, \eta_{PN_P})^T \), \( \eta_{pj} = \eta(x_p, y_p, t_j) \) on the set of points \((x_p, y_p)\), \( p = 1, \ldots, P \) and at time instants \( t_j \), \( j = 1, \ldots, N_t \). Then the vector \( \eta \) containing the observed tsunami waveforms can be expressed as follows:

\[
\eta = Ac, \tag{7}
\]

where \( A \) is a matrix which columns consist of computed waveforms for every spatial harmonic \( \varphi_{mn}(x, y) = \sin \frac{m\pi}{l_1} x \cdot \sin \frac{n\pi}{l_2} y \) used as initial condition to the direct problem (1)-(4). The coefficients \( \alpha_k \) of decomposition of vector \( c \) to the right singular vectors \( c = \sum_{j=1}^{MN} \alpha_j e_j \) are expressed as follows \( \alpha_j = \frac{(\eta l_j)}{s_j} \), where \( l_j \) and \( e_j \) are the left and the right singular vectors of the matrix \( A \) and \( s_j \) are its singular values. Then, the \( r \)-solution of Eq.(7) is represented as \( c^{[r]} = \sum_{j=1}^{r} \alpha_j e_j \)

and, finally, the desired function \( \varphi(x, y) \) takes the form

\[
\varphi^{[r]}(x, y) = \sum_{j=1}^{r} \alpha_j \sum_{m=1}^{M} \sum_{n=1}^{N} \beta_{mn}^j \varphi_{mn}(x, y), \tag{8}
\]

where \( e_j = (\beta_{11}^j, \beta_{12}^j, \ldots, \beta_{MN}^j)^T \). The solution obtained is stable for any fixed \( r \) with respect to perturbations of the right-hand side. The relationship between \( r \) and the singular values of matrix \( A \) as well as the conditioning number (noted as \( \text{cond} \)) of the matrix obtained by projection of the operator \( A \) in Eq. 7 onto a linear span of its \( r \) first right singular vectors can be expressed as \( r = \max\{k :
Thus, the value of $r$ is determined by the singular spectrum of the matrix $A$, and it is still significantly smaller than the dimension of the matrix obtained. A sharp decrease in the singular values, when their numbers increase, is typical of all the calculations, due to the ill-posedness of the problem. Increasing the value $r$ leads therewith to a higher instability. On the other hand, parameter $r$ should be large enough to provide a suitable spatial approximation of the function $\varphi(x, y)$. It is clear that properties of matrix $A$ and, consequently, the quality of the obtained solution are determined by the location and extent of the tsunamigenic area, the configuration of an observation system and the temporal extent of the signal. Some properties of the inverting operator in the context of retrieving a tsunami source were studied numerically in [15].

4 The influence of a tsunami monitoring system location on the inversion results

A series of calculations have been carried out by the method proposed to clarify the dependence of the efficiency of the inversion on certain characteristics of the observation system such as the number of receivers and their location. The inversion method described above was applied to the 2013 Solomon Islands event. The 6 February 2013 magnitude 8.0 Mw Santa Cruz Islands, Solomon Islands earthquake ($10.738^\circ S, 165.138^\circ E$), depth 29 km, generated a tsunami that was observed all over the Pacific region and caused deaths and damage locally. In Fig.1 the domain $\Pi = \{(x; y) : 140^\circ E \leq x \leq 185^\circ E; 17^\circ S \leq y \leq 13^\circ N\}$ of tsunami propagation calculated with GEBCO bathymetry (1-min resolution; available at http://www.gebco.net/) is presented. We consider a Cartesian coordinate system with the origin at the point $(140^\circ E, 17^\circ S)$. Let the $Ox$-axis and the $Oy$-axis are directed along the longitude and latitude accordingly. The tsunami source area is a rectangle $\Omega = \{164.638^\circ E \leq x \leq 165.638^\circ E; 11.238^\circ S \leq y \leq 10.238^\circ S\}$. The sea levels were recorded by the system of six (P = 6) DART® marked by the white color (○) and enumerated clockwise in Fig.1: 1-55012; 2-55023; 3-52403; 4-52402; 5-52406; 6-51425. The time interval was long enough for the tsunami wave to reach all the receivers, specifically, the time step equaled 4 sec, the number of time steps was defined as $N_t = 2000$. In these calculations the values of parameters $M$ and $N$ are empirically established as $M = 15$; $N = 15$. The matrix $A$ is about $(225 \times (2000 \times p))$, where $p$ is equal to the number of tsunami waveforms used in the inversion. Numerical simulation is based on a finite difference algorithm and the method of staggered grids. A rectangular grid of $2700 \times 1800$ nodes was placed over the domain $\Pi$ while a rectangular grid of $61 \times 61$ nodes was placed over the domain $\Omega$, respectively. The epicenter of the tsunami source is assumed to be at the node $(1509, 376)$. The matrix $A$ is computed with MOST (Method of Splitting Tsunami) package [http://nctr.pmel.noaa.gov/model.html] adopted to NVIDIA GPU ([17]). Further, standard SVD- procedure was applied to matrix $A$. The analysis of singular spectrum of matrix $A$ allows one to define the number $r$ and to compute the coefficients $\{c_{mn}\}$ as an $r$-solution of Eq.(7). After this, the function $\varphi^{[r]}(x, y)$ was computed in the form (8).
The series of the numerical experiments with real data were aimed to highlight the way of varying the observation system on improving inversion. One of the main factors which contribute to the difficulties is the complexity of the bottom relief with a plenty of submarine rocks and chains in the considered domain. Furthermore, it was interesting to obtain an acceptable results of the inversion using a minimum number of observed waveforms. As is known, increasing the number of receivers does not often lead to a good inversion if there is no optimal azimuthal coverage with respect to the source and, on the contrary, in real cases it turns out that the noisiness of data is raised resulting in lowering the efficiency of inversion.

First of all, the singular spectrum of matrix $A$ was analyzed in every case. In Fig. 2, left, common logarithms of singular values of $A$ are shown for different subsets of the receivers. In Fig. 2, right, zoomed plots from Fig. 2, left, are presented. Given a fixed bound on the conditioning number one can define value of $r$ as an x-coordinate of the intersection point for the corresponding horizontal line and the singular value plot. Obviously, if the singular value plot decreases more or less smoothly up to some point, there is an opportunity to use larger $r$ and, hence, to get the more informative solution. For the below considered receiver subsets using $r > 21$ appears to be impracticable due to the high level of the noisiness of the observed data which leads to the solution instability (see such example in Fig. 3 (a), (b)).
Fig. 2. Plots of singular values in the common logarithmic scale of matrix $A$ with respect to their numbers are marked by the different line styles relative to the receivers used in the inversion: $\{4, 5, 6\}$ (the dashed-dotted line), $\{1, 3, 5\}$ (the dashed line), $\{1, 5, 6\}$ (the thin solid line), five $\{1, 2, 3, 5, 6\}$ (the dotted line) and six $\{1, 2, 3, 4, 5, 6\}$ (the black line). The receivers are enumerated clockwise according to Figure 1.

Analysis of the singular spectra plays the key role to understand the relationship between the improvement of inversion and a change in the configuration of the observation system. It is clear that modifying in the subset of receivers results in changing the corresponding singular spectrum. Indeed in Fig. 2, the dashed line for the subset consisting of Receivers 1, 3, 5 and the dashed-dotted line for the subset consisting of Receivers 4, 5, 6 significantly differ, that is a consequence of the replacement of Receiver 1 by Receiver 4. This is in good agreement with a change in the inversion results presented in Fig. 4 (b), (c).

Analysis of the plots in Fig. 2 makes possible to expect that the worst inversion results would be obtained by using the subset consisting of Receivers 4, 5,
Indeed, comparison of the results presented in Fig. 4 confirms this assumption.

As it is shown in Fig. 2, the singular spectra of the monitoring systems involving Receivers 1, 3, 5 (the dashed line), 1, 2, 3, 5, 6 (the dotted line) and 1, 2, 3, 4, 5, 6 (the solid line) are similar in appearance. It is possible to expect similar results of the inversion in these cases. The results presented in Fig. 3 (b) and in Fig. 4 (b), (f) confirm this idea. Such sets of receivers provide sufficiently

Fig. 4. Recovered initial functions in the target domain inverted by using different subsets of receivers. There are the cases study when the conditioning number of the matrix obtained is equal to 2.5. Numbers of the receivers used are in parentheses.
plausible results, as evidenced by comparing the marigrams from this recovered source with observed ones that are presented in Fig. 5.

The importance of azimuthal coverage with respect to the tsunami source and bathymetry features are illustrated by the inversion results for different receiver sets such as 1, 2, 5 (Fig. 4 (a)), 1, 3, 5 (Fig. 4 (b)) and 4, 5, 6 (Fig. 4 (c)). As is clear from the plots (a), (b), (c) in Fig. 4, the inversion results for the equipotent subsets and common the cond of the matrices obtained are different. The results obtained by Receivers 1, 3, 5 are much better than for those subsets 1, 2, 5 and 4, 5, 6. The usage of Receiver 3 and Receiver 1 have significantly improved both the shape and the amplitude of the source (the plots in Fig. 4 (a) and (b) as well as in Fig. 4 (e) and (f) the plots in Fig. 4 (c) and (d)). The latter can be due to its perfect location in the direction of reflections from the submarine rock trail. On the contrary, a remote Receiver 4, surrounded by the islands, does not have any impact on the solution and, probably, only introduces additive noise. The same conclusion can be made for Receivers 6 and 2 from the comparison the results presented in Fig. 4 (a) and (e), as well as from the comparison the results presented in Fig. 4 (a) and (b). The fact is in our case, a decrease in the length of the records used in the inversion does not make any evidence of a positive effect.

The approach proposed provides a way to balance the number of the receivers and the quality of the inversion. Based on the analysis of a singular spectrum for each specific observation system one can define a maximal r which allows one to avoid the numerical instability.

Indeed, the results of numerical experiments presented in Fig.4 substantiate our assumption based on analyzing singular spectra.

Based on the carried out the numerical experiments, it is possible to conclude that the subsets including Receivers 1, 3, 5 are the most efficient to reconstruct the tsunami source by the method proposed.

After the inversion by tsunami waveforms from the Receivers 1, 2, 3, 5, 6 was completed, the direct problem was once again solved with the recovered function \( \varphi(x, y) \) as initial condition (2) and the marigrams were calculated at the same six points where DARTs Buoys were assumed. As is clear from Fig. 5, the marigrams computed with the recovered tsunami source have a sufficient matching with the real data. This result can be improved by special filtration of the observed data.

5 Conclusion

The instability of a numerical solution of the ill-posed inverse problem in question in many instances is due to the noise in real marigrams that is a common feature in any real applications. An approach based on r-solution method allows one to control the instability of a numerical solution and to obtain an acceptable result in spite of ill-posedness of the problem. The method seems attractive from the computational point of view since the main efforts are required for calculating
Fig. 5. Comparison of the observed tsunami waveforms (dash line) and calculated ones by using the records of Receivers 1, 2, 3, 5, 6 (solid line) for the 2013 Solomon Islands event. Numbers below ID of DARTs indicate the time (in minutes) after the earthquake origin time. Location of the DARTs® are shown in Figure 1.

matrix A. If an observation system is fixed and tsunami-prone areas are defined, one can compute the matrix only once as a preliminary stage.

It is possible to make a preliminary evaluation of the efficiency of the inversion with a given set of recording stations by analyzing the singular spectrum of a relevant matrix. The results obtained allow to find the way to improve the inversion by selecting the most informative set of available recording stations. Since tsunami sources often have a dipolar shape, the location of receivers on direct and reflected rays corresponding to the direction of the strongest variability of the dipole source have the greatest effect for the inversion result. In addition, one should keep in mind that increasing the number of marigrams used for inversion does not always lead to an improved accuracy of the numerical solution. The receiver location effects the choice of number r by such a way: the better is the configuration of the observation system, the longer is a weakly decreasing part of the spectrum. Thus, the rate of the singular values descent which is
most directly correlated with the receiver location should be considered as main parameter of the efficiency of the inversion.

The function recovered by the method proposed can find practical use both as an initial condition for various optimization approaches and for computer calculation of the tsunami wave propagation. It may be useful to designing future observation systems for regions of perceived tsunami risk by providing a well-aimed precomputation with varying locations of potential sea level recorders.

Acknowledgments. The author thanks Gorge Shevchenko, Ph.D., from IMGG FEB RAS for his assistance the preparation of data and Alexei Romanenko, Ph.D., from NSU for helping in calculating.

References

10. Tsunami Forecast by Joint Inversion of Real-Time Tsunami Waveforms and Seismic or GPS Data: Application to the Tohoku 2011 Tsunami


Effect of Dielectric Permittivity Distribution in Invaded Zone on Induction Log Data

Igor Yeltsov\textsuperscript{1,2,3}, Timofey Eltsov\textsuperscript{1,2,3}, Aleksandr Makarov\textsuperscript{4}, and Marina Nikitenko\textsuperscript{1,4}

\textsuperscript{1} Institute of Petroleum Geology and Geophysics of SB RAS, Novosibirsk, Russia
\textsuperscript{2} Novosibirsk State University, Novosibirsk, Russia
\textsuperscript{3} Novosibirsk State Technical University, Novosibirsk, Russia
\textsuperscript{4} Baker Hughes, Novosibirsk, Russia

\texttt{yeltsovin@ipgg.sbras.ru}

Abstract. The objective of this study is to demonstrate the effect of changes in dielectric permittivity on induction logging data during invasion evolution. As an investigation method we use invasion simulation and electromagnetic (EM) modeling of induction tool responses. In this study we simulate signals of EM tools to analyze the influence of different types of dielectric permittivity distribution and to estimate the influence of dielectric permittivity on induction logging signals. Analysis of the computed induction tool signals shows that dielectric permittivity influence on magnetic field attenuation is higher than on magnetic field phase difference. Induction measurements (transmitter frequency $> 1$ MHz) are significantly influenced by dielectric permittivity distribution in the invaded zone. The highest influence of dielectric permittivity on induction logging signals is observed in the case of low formation water saturation and high resistivity of drilling mud.

Keywords: Mud invasion, near wellbore dielectric permittivity distribution, near wellbore dielectric resistivity distribution, invaded zone, induction logging tool, electromagnetic modeling

Introduction

The mud invasion process causes formation fluid displacement that results in changes of resistivity and dielectric permittivity distribution (electro-physical properties) in the near wellbore zone. For conducting a correct interpretation of electromagnetic logging data, it is necessary to take into account varying resistivity and dielectric permittivity distribution in the invaded area, as far as these parameters affect induction logging measurements.

Near wellbore water saturation and salinity distributions can be described by numerical modeling of the mud filtrate penetration into the formation. The Buckley-Leverett equations for two-phase flow in porous media are used to carry out mud invasion simulation. These distributions are utilized to calculate resistivity and dielectric permittivity profiles using Archies equation and the complex refractive index method (CRIM) respectively. To estimate the influence of
resistivity and dielectric permittivity on induction logging signals we use electromagnetic (EM) modeling. Signals of induction logging tools are computed using an axisymmetric cylindrically layered earth model.

1 Dielectric permittivity mixing laws used for an oil and water saturated formation

There are many different mixing laws described in literature, all of them having empirical type. One of the most well-known equations is the Bruggeman mixing law:

\[
\left( \frac{\epsilon_m - \epsilon_{\text{eff}}}{\epsilon_m - \epsilon_w} \right) \left( \frac{\epsilon_w}{\epsilon_{\text{eff}}} \right)^{\frac{1}{3}} = \phi ,
\]

where \( \epsilon_m \) — complex permittivity of the matrix grains, \( \epsilon_w \) — complex permittivity of the saturating brine, \( \epsilon_{\text{eff}} \) — effective permittivity of the rock, \( \phi \) — volume fraction of the saturating brine.

The main merit of the Bruggeman mixing law is taking into account interaction between fractions of the rock. There are only two different types of fraction: saturating brine in the separated spheres inside the matrix. There is an option to use the Bruggeman mixing law for a three component system (water, oil and matrix) proposed by [13]. In the proposed geometry each matrix grain is coated layer-by-layer with oil and water, but that could be applied only to oil-wet rocks.

There is an algorithm of dielectric permittivity mixing, proposed by Shelukhin and Terentev [14]. Actually it is the most complex and modern way of modeling the dielectric permittivity of a multicomponent system, but it is too complicated for usage and it is not suitable for our purposes, because it was not verified via laboratory experiment.

Further come the most extensively applicable mixing laws [13] used for multicomponent systems. The \( \hat{j} \)-scaled mixing law is:

\[
\epsilon_{\text{eff}}^{\frac{1}{\hat{j}}} = \sum_{n=1}^{N} \phi_n \epsilon_n^{\frac{1}{\hat{j}}} ,
\]

where \( n \) — number of a single component, \( \epsilon_n \) — dielectric permittivity of a sample component, \( \phi_n \) — fraction of a sample component, \( \hat{j} \) — empirical constant estimated from experimental measurements, \( N \) — a number of components.

The volumetric mixing law is defined as:

\[
\epsilon_{\text{eff}} = \sum_{n=1}^{N} \phi_n \epsilon_n .
\]

The Birchak mixing law is:

\[
\epsilon_{\text{eff}}^{\frac{1}{\hat{j}}} = \sum_{n=1}^{N} \phi_n \epsilon_n^{\frac{1}{\hat{j}}} .
\]
The complex refractive index method (CRIM) or Birchaks mixing law proposed by Birchak and others [2] was checked by [11] and [13], who experimentally proofed that the CRIM formula is the most reliable in the case of a three-component system (measurements of electro physical parameters in a laboratory for a wide frequency range). The CRIM model is based on the optical path length of a single electromagnetic ray. It is equivalent to a volumetric average of the complex refractive index, and it assumes that total transit time of a propagating pulse is equal to the sum of transit times of the constituents [13].

The Looyenga-Landau-Lifshiz mixing model:

\[
\frac{1}{\varepsilon_{\text{eff}}} = \sum_{n=1}^{N} \phi_n \frac{1}{\varepsilon_n}.
\] (5)

The Lichtenecker’s mixing law is as follows:

\[
\ln \varepsilon_{\text{eff}} = \sum_{n=1}^{N} \phi_n \ln \varepsilon_n.
\] (6)

The CRIM formula was chosen because it was verified by well-known specialists and it is applicable for multicomponent systems.

For modeling we chose typical resistivities of water-based drilling mud. Formation water saturations were selected in such a way as to cover both water and oil saturated reservoirs. Dielectric permittivity of mineralized water was set according to Fig. 1 [8]. Water saturations were selected to cover both water and oil saturated reservoirs. Analysis of current publications allowed us to select typical examples of petrophysical properties of oil and water saturated sandstones.

Fig. 1. Dependence of dielectric permittivity of NaCl solution on mineralization at temperature of +20 °C [8]
2 Dielectric Permittivity Distribution Modelling

Fig. 2 shows the formation earth model utilized for invasion and electromagnetic modeling, which is a radial axisymmetric layered medium. The layers are: borehole, formation and a number of layers between them characterizing the distribution of dielectric permittivity ($\epsilon_{iz}(r)$) and resistivity ($\rho_{iz}(r)$) in the invaded zone. $\epsilon_f$ and $\rho_f$ are correspondingly dielectric permittivity and resistivity of undisturbed formation. The modeling is divided into two parts. The first part is invasion zone modeling that provides water saturation and salinity distributions in near wellbore area.

2.1 Invasion Zone Simulation

Reservoir drilling is accompanied by mud invasion into a permeable formation. Mud solids plug the near-wellbore area and form a low-permeability mudcake at the borehole wall. The mudcake reduces the rate of mud filtrate invasion and slows the invaded zone extension. The formation properties, such as resistivity, density and others, change in the near-wellbore zone, depending on mud filtrate penetration and displacement of formation fluids. Mud invasion structure can be estimated via inversion of electromagnetic log data measurements [5]. In this study, the aim of invasion modeling is to compute water saturation and salinity distributions in the invaded zone.

Water-based mud filtration is simulated using mud reports, drilling regime data and available a-priori information, such as formation and mudcake properties [15, 3, 7]. We consider the Buckley-Leverett model without capillary forces
to describe the invasion of water-based mud into a reservoir saturated by oil and water \cite{9, 4}. The transport equations for the 1D radial axi-symmetric case are the following:

\[
\frac{\partial}{\partial t} (r \phi f S_w) = \frac{\partial}{\partial r} \left( r k_0 k_w \frac{\partial p}{\partial r} \right), \quad R_{wb} < r < L ,
\]

\[
\frac{\partial}{\partial t} (r \phi f S_{oil}) = \frac{\partial}{\partial r} \left( r k_0 k_{oil} \frac{\partial p}{\partial r} \right), \quad R_{wb} < r < L ,
\]

\[
\phi_f = \phi_0 + \lambda p ,
\]

\[
S_{oil} + S_w = 1 ,
\]

where \( r \) is the distance from the center of the wellbore to the remote part of the formation. The region of modeling is \( R_{wb} < r < L \), where \( R_{wb} \) is the wellbore radius and \( L \) is the boundary of the modeling area. \( L \) is the distance at which pressure in the near-wellbore zone equals the formation pressure, \( L \gg R_{wb} \); at distances greater than \( L \), pressure does not change; \( t \) is the modeling time that varies from 0 (the beginning of reservoir drilling) up to \( T_d \) (time since the reservoir drilling); \( p \) is the pressure difference between the well pressure and formation pressure; \( S = S_w, S_{oil} \) are the saturations of water and oil fractions, respectively; \( \phi_f \) is the formation porosity; \( \phi_0 \) is the porosity of the undisturbed formation; \( \lambda \) is the compressibility of the formation; \( k_w(S_w) = S_{nw}^w \), \( k_{oil}(S_{oil}) = S_{noil}^{oil} \) are the functions of phase permeabilities; \( n_w, n_{oil} \) are the empirical exponents found from petrophysical analysis (Corey-Brooks constants); \( \mu_w, \mu_{oil} \) are the viscosities of water and oil, respectively; \( k_0 \) is the absolute permeability of the formation.

The boundary and initial conditions are added to the equations (7) and (8), and mud filtration through mudcake is determined by the expression:

\[
p|_L = 0, \quad p|_{t=0} = 0, \quad Q(t) = R_{wb} \frac{(P_{wb} - p|_{r=R_{wb}})}{(b_0^{-1} + d \mu_w/k_c)}, \quad S|_{t=0} = S_f, \quad S_{R_{wb}-d} = 1 ,
\]

\( b_0^{-1} \) is the filtration drag of the plugging zone; \( d \) is the mudcake thickness; \( k_c \) is the mudcake permeability; \( P_{wb} \) is the pressure drop from the wellbore to the formation; \( Q \) is the mud filtrate flow rate and \( S_f \) is the initial formation water saturation. The mudcake thickness, permeability and porosity are assumed to be constant, which significantly simplifies the filtration model. To solve the system of the equations (7)–(11) the numerical solution was implemented.

Combining the equations (7) and (8), and taking into account the equations (9)–(11), we obtain the following piezo-conductivity equation \cite{1}:

\[
\lambda \cdot r \cdot \frac{\partial p}{\partial t} = \frac{\partial}{\partial r} \left[ \lambda \cdot k_0 \cdot \left( \frac{k_w}{\mu_w} + \frac{k_{oil}}{\mu_{oil}} \right) \frac{\partial p}{\partial r} \right], \quad R_{wb} < r < L .
\]

Salt transport during invasion occurs because of formation fluid and mud filtrate mixing. The water phase transport is caused by pressure overbalance.
Salinity transport modeling is described using the following transport equation and boundary conditions:

\[
\frac{\partial}{\partial t} (r\phi_f S C) = \frac{\partial}{\partial r} \left( r k_w \frac{\partial p}{\partial r} C \right),
\]

(13)

\[
C|_{t=0} = 1, \quad C_{R_{wb}} = 0,
\]

(14)

where \( C \) is the relative salinity that is equal to \( C_{wb} \) in mud filtrate and to \( C_f \) in formation brine. To evaluate the true salinity profile, the relative salinity distribution should be normalized by the true salinity values. The finite-difference scheme is developed and realized for the solution of the system of equations (7) - (14). Firstly, the pressure overbalance distribution is computed via implicit finite-difference method for equation (12). Secondly, for the computation of water saturation and salinity distribution the explicit finite-difference method is used, after that, the computation goes to the next time step, and the scheme is realized again until the simulation time \( T_d \). The parameters used for invasion simulation are shown in the Table 1.

### Table 1. Invasion simulation parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formation permeability, D</td>
<td>( k_0 ) 0.1</td>
</tr>
<tr>
<td>Initial formation water saturation, fraction</td>
<td>( S_f ) 0.2, 0.4, 0.6</td>
</tr>
<tr>
<td>Formation compressibility, atm (^{-1})</td>
<td>( \lambda ) 0</td>
</tr>
<tr>
<td>Formation porosity, p.u.</td>
<td>( \phi_f ) 25</td>
</tr>
<tr>
<td>Borehole radius, m</td>
<td>( R_{wb} ) 0.108</td>
</tr>
<tr>
<td>Pressure overbalance, atm</td>
<td>( P_{wb} ) 10</td>
</tr>
<tr>
<td>Time since reservoir drilling, day</td>
<td>( T_d ) 0.2</td>
</tr>
<tr>
<td>Mudcake permeability, D</td>
<td>( k_c ) 0.00001</td>
</tr>
<tr>
<td>Mudcake thickness, m</td>
<td>( d ) 0.001</td>
</tr>
<tr>
<td>Corey-Brooks constant for oil phase permeability, unitless</td>
<td>( n_{oil} ) 2.5</td>
</tr>
<tr>
<td>Corey-Brooks constant for water phase permeability, unitless</td>
<td>( n_w ) 3.0</td>
</tr>
<tr>
<td>Oil viscosity, cP</td>
<td>( \mu_{oil} ) 10.0</td>
</tr>
<tr>
<td>Water viscosity, cP</td>
<td>( \mu_w ) 1.0</td>
</tr>
</tbody>
</table>

#### 2.2 Calculation of Resistivity and Dielectric Permittivity Profiles

These distributions are used to calculate the resistivity and dielectric permittivity profiles via Archie’s equation and the CRIM model. The following Archie’s equation is used:

\[
\rho_{iz}(r) = \rho_{w}(r)\phi_f^{-m} S_{w}^{-n}(r),
\]

(15)
here \( n = 2, \, m = 1.5 \) - Archie’s exponents, \( \rho_{iz}(r) \) — invasion zone resistivity; \( \rho_w(r) \) — water fraction resistivity; \( \phi_f \) — formation porosity; \( S_w(r) \) — water saturation distribution.

Calculation of dielectric permittivity distribution using the CRIM model comprises the following steps. Firstly, it is necessary to calculate the volume fractions of each constituent. Knowing water saturation distribution \( \phi_w(r) \) in the near borehole zone it is easily to calculate the volume fractions of matrix \( \phi_m(r) \) and oil \( \phi_{oil}(r) \) in the whole space of interest.

\[
\phi_w(r) = \phi_f S_f(r), \tag{16}
\]
\[
\phi_{oil}(r) = \phi_f - \phi_w(r), \tag{17}
\]
\[
\phi_m(r) = 1 - \phi_f. \tag{18}
\]

Secondly, after volume fraction calculation, dielectric permittivity distribution is calculated in each point in the near borehole zone using the CRIM formula:

\[
\varepsilon_{eff}(r) = \left( \phi_m \varepsilon_m^{\frac{1}{2}} + \phi_w(r) \varepsilon_w^{\frac{1}{2}} + \phi_{oil}(r) \varepsilon_{oil}^{\frac{1}{2}} \right)^2, \tag{19}
\]

where \( \varepsilon_w \) — dielectric permittivity of water fraction, \( \varepsilon_{oil} \) — dielectric permittivity of oil fraction, \( \varepsilon_m \) — dielectric permittivity of rock matrix.

3 Discussion

It is necessary to consider the difference between the radial profiles of electro physical parameters for different geological conditions. Pictures Fig. 3–5 show the distribution of electro-physical parameters of sandstones.

![Fig. 3. Water saturation profile. Sandstone, porosity 25 p.u. Different water saturation of collector](image_url)
Fig. 4. Resistivity distribution. Sandstone, porosity 25 p.u. Different conductivities of brine.

Fig. 5. Dielectric permittivity profile. Sandstone, porosity 25 p.u. Different water saturation of collector.
Fig. 3 shows the dielectric permittivity distribution for sandstone, porosity 25 for different initial formation water saturations ($S_f = 0.2, 0.4, 0.6$).

The shape of water saturation profile (Fig. 3) looks like corresponding dielectric permittivity profile (Fig. 5). The reason of this phenomenon is due to the CRIM formula usage. The value of dielectric permittivity falls down from 15 to 6.7 (Fig. 5). The most significant change of dielectric permittivity along the invaded zone is for the lowest water saturation.

4 Signal Modeling

We simulate a vertical magnetic field component ($H_z$). The transmitter is assumed to be a vertical magnetic dipole. To simulate induction log synthetic responses, we use forward modeling software for the case of 1D coaxial cylindrical geometry.

Modeling is performed using the parameters of LWD resistivity tool. This LWD resistivity tool measures formation resistivity in real-time for indication of oil and gas saturated reservoirs [10]. It consists of three-coiled induction short-spaced (23 in.) and long-spaced (35 in.) probes, and operates on two frequencies: 400 kHz and 2 MHz. The measured signals are the magnetic field attenuation and magnetic field phase difference (between two receiver coils). Usually these signals are converted into apparent resistivities for further interpretation.

Also we simulate responses of the induction logging tool described in a patent application published by Nikitenko M.N. in 2009 [12]. In our paper this tool is called “Induction tool”. Its frequency range is 1.75–24.5 MHz, with spacings varying from 0.18 m to 1 m. The three-coil array consists of a transmitter, a main receiver and a bucking coil. The moments are chosen in such a way as to compensate direct field. Seven different three-coil probes are considered. The parameters of the tools under consideration the distances between the coils and operating frequencies are shown in Table 2.

5 Influence of Dielectric Permittivity Distribution on Induction Logging Signals

The resistivity and dielectric permittivity profiles shown in Fig. 3–5 were selected as input earth models for the EM modeling of the induction log synthetic responses. Signals were modeled taking into account dielectric permittivity distribution and compared with signals modeled without dielectric permittivity distribution.

Influence of dielectric permittivity distribution on the LWD resistivity tool and high-frequency induction probes is noticeable. Before comparison of the signals it is necessary to estimate their level. The common value of the noise level is 0.01 dB. The common value of noise for the phase shift signal is 0.1 degrees. The signal was considered to be high if it was two times bigger than
Table 2. Induction tool and LWD resistivity tool modeling parameters

<table>
<thead>
<tr>
<th>3 coil tool number</th>
<th>Lengths, m</th>
<th>Frequencies, MHz</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Induction tool</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.18, 0.25</td>
<td>12.25, 24.50</td>
</tr>
<tr>
<td>2</td>
<td>0.22, 0.32</td>
<td>10.50, 21.00</td>
</tr>
<tr>
<td>3</td>
<td>0.28, 0.40</td>
<td>8.75, 17.50</td>
</tr>
<tr>
<td>4</td>
<td>0.35, 0.50</td>
<td>7.00, 14.00</td>
</tr>
<tr>
<td>5</td>
<td>0.45, 0.64</td>
<td>5.25, 10.50</td>
</tr>
<tr>
<td>6</td>
<td>0.56, 0.80</td>
<td>3.50, 7.00</td>
</tr>
<tr>
<td>7</td>
<td>0.70, 1.00</td>
<td>1.75, 3.50</td>
</tr>
<tr>
<td><strong>LWD resistivity tool</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.47, 0.67</td>
<td>0.4</td>
</tr>
<tr>
<td>2</td>
<td>0.47, 0.67</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0.80, 1.0</td>
<td>0.4</td>
</tr>
<tr>
<td>4</td>
<td>0.80, 1.0</td>
<td>2</td>
</tr>
</tbody>
</table>

the noise. The deviation between the signals is shown in errors. The error is calculated using the formula:

\[
\delta = \frac{|S_{eps} - S_0|}{\gamma},
\]

where \( \gamma \) is value of noise (0.1 degree for phase difference and 0.01 dB for attenuation). \( S_{eps} \) — signal in case of media with dielectric permittivity distribution, \( S_0 \) — signal in media without dielectric permittivity distribution.

For most of the sandstone models the LWD resistivity tool signals are high enough. The exception is the most resistive models.

Fig. 6–8 exemplify the induction tool signals. Fig. 11–13 shows an example of the signals of LWD resistivity tool. The influence of dielectric permittivity distribution on induction tool signals is shown on Fig. 9–15.

The difference between the signals is shown for sandstone, porosity 25 p.u., mud conductivity 1 S/m, brine conductivity 0.5 S/m. The tool numbers are numbers of the three-coil tools shown in Table 3. Fig. 9 and Fig. 10 corresponds to the induction tool. The difference between the signals for all the frequencies and the longest array (0.7 m and 1 m) is shown in Fig. 9. The difference between the signals for the highest frequency and all the arrays at the highest frequency of the induction tool is shown in Fig. 10. The longer the tool spacing and the higher the frequency, the greater the permittivity influence is. The influence of dielectric permittivity distribution on the LWD resistivity tool signals is shown in Fig. 14 and Fig. 15. When considering the LWD resistivity tool signals we may note that the influence of dielectric permittivity is less than the noise value.

The overall results are shown in Table 3, where the differences between the signals are calculated using the equation (20).
Fig. 6. The induction tool, longest array, all frequencies. Porosity 25 p.u., $S_t = 0.6$, $C_b = 0.5$ S/m, $C_f = 1$ S/m. With (+) and without (×) dielectric permittivity distribution

Fig. 7. The induction tool, longest array, all frequencies. Porosity 25 p.u., $S_t = 0.4$, $C_b = 0.5$ S/m, $C_f = 0.1$ S/m. Symbols are the same as in Fig. 6

Fig. 8. The induction tool, longest array, all frequencies. Porosity 25 p.u., $S_t = 0.2$, $C_b = 0.5$ S/m, $C_f = 1$ S/m. Symbols are the same as in Fig. 6
Fig. 9. The induction tool, longest array, all frequencies. Porosity 25 p.u., $C_b = 0.5 \text{ S/m}$, $C_f = 1 \text{ S/m}$

Fig. 10. The induction tool, highest frequency, all arrays. Porosity 25 p.u., $C_b = 0.5 \text{ S/m}$, $C_f = 1 \text{ S/m}$

Fig. 11. LWD resistivity tool computed signals, longest array, all frequencies. Porosity 25 p.u., $S_f = 0.6$, $C_b = 0.5 \text{ S/m}$, $C_f = 1 \text{ S/m}$. Symbols are the same as in Fig. 6
Fig. 12. LWD resistivity tool computed signals, longest array, all frequencies. Porosity 25 p.u., $S_l = 0.4$, $C_b = 0.5$ S/m, $C_f = 1$ S/m. Symbols are the same as in Fig. 6.

Fig. 13. LWD resistivity tool computed signals, longest array, all frequencies. Porosity 25 p.u., $S_l = 0.2$, $C_b = 0.5$ S/m, $C_f = 1$ S/m. Symbols are the same as in Fig. 6.

Fig. 14. LWD resistivity tool, difference of computed attenuations. Porosity 25 p.u., $C_b = 0.5$ S/m, $C_f = 1$ S/m.
Fig. 15. LWD resistivity tool, phase discrepancy of computed phase differences. Porosity 25 p.u., $C_b = 0.5 \, \text{S/m}$, $C_f = 1 \, \text{S/m}$

Table 3. Difference between induction signals

<table>
<thead>
<tr>
<th>Tool number</th>
<th>$S_f = 0.6$</th>
<th>$S_f = 0.4$</th>
<th>$S_f = 0.2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference, $\delta$</td>
<td>+</td>
<td>$\times$</td>
<td>$\circ$</td>
</tr>
<tr>
<td>Tool number</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

**Induction tool**

| Normal phase difference signal level, degrees | 8–31.3 | 5–25 | 2–18 |
| Phase signal difference, noise level | 1–50 | 1–56 | 1–65 |
| Normal attenuation signal level, dB | 0.4–2.7 | 0.1–0.55 | $< 0.1$ |
| Attenuation signal difference, noise level | 5–80 | 5–90 | $< 0.01$ |

**LWD resistivity tool**

| Normal phase difference signal level, degrees | 0.6–3.7 | 0.13–2.6 | 0.1–1 |
| Phase signal difference, noise level | $< 0.1$ | $< 0.1$ | $< 0.1$ |
| Normal attenuation signal level, dB | 0.01–0.2 | $< 0.01$ | $< 0.01$ |
| Attenuation signal difference, noise level | $< 0.01$ | $< 0.01$ | $< 0.01$ |
Summary

In this study, we propose a workflow of dielectric permittivity distribution computation in the near wellbore area, using invasion simulation and the complex refractive index method. The analysis of dielectric permittivity distribution computation shows that in case of low mud and brine conductivity ($C_b = 0.5 \text{ S/m}, C_f = 1 \text{ S/m}$) water saturation distribution is the main factor affecting dielectric permittivity distribution in the invaded zone. It is observed that dielectric permittivity distribution is almost insensitive to water salinity in case of low conductivity both mud and brine.

The numerical modeling of induction logging synthetic responses (LWD resistivity tool) shows that influence of dielectric permittivity distribution on the LWD resistivity tool phase difference of magnetic field is less than the value of the noise level (< 0.1 degree) for all the considered initial formation water saturations. Modeling of the induction tool signals shows that the influence of dielectric permittivity distribution on the phase difference of magnetic field is considerable in the case of high values of water saturation ($S_f = 0.6$) and high values of fluid and mud and brine conductivity ($C_b = 0.5 \text{ S/m}, C_f = 1 \text{ S/m}$). The effect of dielectric permittivity distribution on the phase difference can be 2 times lower than that on the attenuation.

Dielectric permittivity distribution influence on the signals of the induction tool has been much higher than that on the LWD resistivity tool signals (up to 100 times of the noise values in the case of sandstone with low mineralization of brine and mud). It means that low frequency (less than 2 MHz) induction tools are more reliable for estimation of the resistivity profile through the LWD resistivity log data processing, as opposed to the induction tool measurements. We recommend using the joint inversion scheme based on invasion simulation and electromagnetic modeling.

Acknowledgments. This work was supported by RFBR grants 16-05-00830 and 16-35-00395.

References

Computational Investigation of Turbulent Flow Impact on Non-cohesive Soil Erosion near Foundations of Gravity Type Oil Platforms

Alexander Yushkov1, Igor Nudner2,3, Konstantin Semenov2,4, Konstantin Ivanov1, Nazim Geidarov1, Sergey Stukolov1, and Yury Zakharov1

1Kemerovo State University, Krasnaya, 6, Kemerovo, Russia
223 State Marine Design Institute Branch of ”31 State Design Institute of Special Construction”, St.-Petersburg, Russia
3Baltic State Technical University, St.-Petersburg, Russia
4Saint-Petersburg State Polytechnical University, St.-Petersburg, Russia
zaxarovyn@rambler.ru

Abstract. The flow turbulence impact on the formation of erosion areas near gravity type oil platforms is studied. The SST (shear-stress transport) turbulence model describing large-scale structures in the internal area and small-scale turbulence in the external area is used for computing turbulent fluid flow. The model grounded on estimation of turbulent behavior of the fluid flow in the bottom flow region where the soil particles transfer is influenced by fluid flow and sea-bed irregularities is applied for estimation of soil erosion. Three sets of numerical studies referred to increase of hydrodynamic values and flow turbulent transition are given.

Keywords: Viscous incompressible fluid, Navier-Stokes equations, non-cohesive soil erosion, three-dimensional flow, turbulence, gravity type oil platforms, numerical and laboratory-based experiments

1 Introduction

The application of gravity type oil platforms at shallow water marine coastal areas is one of the most current means for oil extraction. Processes of sea floor erosion near foundations of such oil platforms and its stability issues are of great interest. In the recent years different investigations of those issues were actively undertaken, by means of laboratory-based and seminatural experiments as well as by means of mathematical simulation [1], [2]. The papers [2], [3] contain results of a great number of experimental and numerical studies of non-cohesive soil erosion near the foundation of the Prirazlommaya platform, comparison charts of laboratory and simulation experiments, analysis of the impact of different wave conditions of fluid flow on the process of particles shift of seabed material. In those papers the laminar model of fluid flow was applied for numerical evaluation of hydrodynamic quantities. The results of the studies [2] show that when the fluid flow is slow and there are no surface waves, the laminar model provides a good match with laboratory results (up to 10-15% accuracy) on the one
hand, and a significant economy of computational resources, on the other hand. The paper [3] illustrates the waves impact on the structures of soil erosion near the platform foundation. The estimations executed at different wave conditions attest the significant change in the pattern of fluid flow and in the structure of soil erosion when internal flow velocity and wave amplitude are increased. Under such modes the applied in [3] laminar model becomes inoperative. As the results of the former studies show, the increase of the hydrodynamic behavior (that corresponds to small surface waves in a natural experiment) leads to flow laminar-turbulent transition. It is evident that the analysis of such flows demands application of the valid turbulence models. The current paper studies the impact of the developed flow turbulence on the formation of outwashes and inwashes (accretion) near the gravity type oil platforms, both numerically and experimentally. The $k - \omega$, SST (shear-stress transport) turbulence model is applied for estimation of bottom fluid flow velocity that directly constitutes the certain structures of soil erosion. For evaluation of soil erosion we apply the model, which is grounded on the estimation of turbulent properties of fluid flow in the bottom region of the flow. The results of three sets of numerical and experimental studies referred to the increase of the hydrodynamic behavior and flow turbulent transition are given.

2 Flow Model

Variable flow of viscous incompressible fluid with constant properties is described by three-dimensional system of Reynolds-averaged Navier-Stokes in accordance with Boussinesq hypothesis on Newtonian turbulent fluid flow where turbulent viscous stresses are related to average flow properties with the same correspondence as molecular resilient friction is related to velocity field, yet turbulent viscosity needs to be defined instead of 6 components of Reynolds stress symmetric tensor ($\nu_t$):

$$\begin{cases} \frac{\partial U_i}{\partial t} = 0, \\ \frac{\partial U_i}{\partial t} + U_j \frac{\partial U_i}{\partial x_j} = -\frac{1}{\rho} \frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j} \left( \nu + \nu_t \right) \frac{\partial U_i}{\partial x_j}, \end{cases}$$

(1)

where $U_i$ - components of velocity vector, $t$ - time, $x_j$ - Cartesian coordinates, $\rho$ - fluid density, $p$ - pressure, $\nu$ - kinematic fluid viscosity, $\nu_t$ - turbulent eddy viscosity estimated on the ground of the applied turbulence model.

The present paper applies $k - \omega$, SST (shear-stress transport) turbulence model [4], where $k$ - turbulent kinetic energy, $\omega$ - specific dissipation rates referred to isotropic dissipation as follows as: $\varepsilon = \beta^* k \omega$, where $\beta^* = 0.09$.

The classic $k - \omega$ model [5] has issues while calculating stream flows due to extreme sensitivity to boundary conditions in the external flow. The model $k - \omega$, SST, suggested by Menter, eliminates the said imperfection and combines the benefits of $k - \omega$ and $k - \varepsilon$ turbulence models. Menter’s model applies the modified $k - \omega$ model designed for description of large-scale structures in the internal area, and in the external area - $k - \varepsilon$, aimed to solution of small-scale
turbulence. In Menter’s model, the model $k - \varepsilon$ is re-formulated in terms $k$ and $\omega$, and a special transition function is added in the received equations in order to transfer from one model to another. The model $k - \omega$, SST is grounded on two equations, the one is for turbulent kinetic energy and the second is for specific dissipation:

$$\frac{\partial k}{\partial t} + U_j \frac{\partial k}{\partial x_j} = P_k - \beta^* k \omega + \frac{\partial}{\partial x_j} \left( (\nu + \sigma_k \nu_t) \frac{\partial k}{\partial x_j} \right),$$

(2)

$$\frac{\partial \omega}{\partial t} + U_j \frac{\partial \omega}{\partial x_j} = \frac{\alpha}{\rho \nu_t} \tau_{ij} \frac{\partial U_i}{\partial x_j} - \beta \omega^2 + \frac{\partial}{\partial x_j} \left( (\nu + \sigma_\omega \nu_t) \frac{\partial \omega}{\partial x_j} \right) + 2 (1 - F_1) \sigma_\omega \frac{1}{\omega} \frac{\partial k}{\partial x_i} \frac{\partial \omega}{\partial x_i},$$

(3)

where

$$\tau_{ij} = \rho \nu_t \left( 2 S_{ij} - \frac{2}{3} \frac{\partial U_n}{\partial x_n} \delta_{ij} \right) - \frac{2}{3} \rho k \delta_{ij},$$

$$S_{ij} = \frac{1}{2} \left( \frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i} \right).$$

The transition function $F_1$ is determined as follows as:

$$F_1 = \tanh \left[ \left( \min \left[ \max \left( \frac{\sqrt{k}}{\beta^* \omega y}, \frac{500 \nu}{y^2 \omega} \right), \frac{4 \rho \sigma_\omega k^2}{CD_{k\omega} y^2} \right) \right] \right],$$

(4)

where $CD_{k\omega} = \max \left( \frac{2 \rho \sigma_\omega^2}{\omega} \frac{\partial k}{\partial x_i} \frac{\partial \omega}{\partial x_i}, 10^{-10} \right)$, $y$ - distance to the surface. Turbulent eddy viscosity is determined by the following formula:

$$\nu_t = \frac{a_1 k}{\max \left( a_1 \omega, SF_2 \right)},$$

(5)
where

\[ S = \sqrt{2S_{ij}S_{ij}}, \]

\[ F_2 = \tanh \left[ \max \left( 2\sqrt{k/\beta^*\omega^2} \right)^2 \right], \]

\[ P_k = \min \left( \rho \nu \frac{\partial U_i}{\partial x_j} \left( \frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i} \right), 10\beta^* \rho \omega \right), \]

\[ \sigma_k = F_1 \sigma_{k1} + (1 - F_1) \sigma_{k2}, \]

\[ \sigma_\omega = F_1 \sigma_{\omega1} + (1 - F_1) \sigma_{\omega2}, \]

\[ \alpha = F_1 \alpha_1 + (1 - F_1) \alpha_2, \]

\[ \beta = F_1 \beta_1 + (1 - F_1) \beta_2. \]

The models invariables are: \( \alpha_1 = \frac{5}{9}, \alpha_2 = 0.44, \beta_1 = \frac{3}{40}, \beta_2 = 0.0828, \sigma_{k1} = 0.85, \sigma_{k2} = 1, \sigma_{\omega1} = 0.5, \sigma_{\omega2} = 0.856. \)

For completion of the objective it is necessary to postulate the boundary conditions for the components of velocity vector \( U_i \), pressure \( p \), turbulent kinetic energy \( k \) and specific dissipation \( \omega \).

In the inlet (Fig. 1): \( U_1 = u_0, U_2 = 0, U_3 = 0, \frac{\partial P}{\partial n} = 0, k = k_0, \omega = \omega_0 \), where \( u_0, k_0, \omega_0 \) are selected on the basis of comparison of velocity profile at the entrance with the experimental data. In the outlet: \( \frac{\partial U_i}{\partial n} = 0, P = 0, \frac{\partial k}{\partial n} = 0, \frac{\partial \omega}{\partial n} = 0 \), On the top surface: \( \frac{\partial}{\partial n} = 0 \) for all variables. On the side surface, pile and bottom: \( U_i = 0, \frac{\partial P}{\partial n} = 0, k = 0, \omega = 10 \frac{\nu^2}{y^2}, \) where \( y \) - distance from the surface to the center of the closest mesh.

### 3 Scour Model

The paper [6] applies the model of soil erosion [7], where computations are based on the values of shear stress on the bottom surface. Here bottom is represented by a certain surface divided into meshes by the grid. At first, the vector of soil particles transport through each mesh is estimated, and then the height of the bottom is determined on the ground of mass balance equation.

Within the frames of the present numerical model, particle motion is influenced by the impact of fluid flow and irregularities in bottom surface. For each design moment, bottom shear stress is estimated at first, and then the calculation of the particle motion on the bottom surface is done and mass balance equation is solved.

Two-dimensional coordinate system \( (x_1, x_2) \) is implemented for the bottom surface. In order to estimate the stress vector \( \tau_b \) it is necessary to find the product of stress tensor \( T \) and surface normal \( \hat{n} \), and then the received vector is projected to the bottom surface.

To determine the vector component \( \tau_b \) it is necessary to know the value of turbulent eddy viscosity \( \nu_t \). In case of application of the laminar model of the
flow while computing velocity field, this parameter was assigned manually to be equal to the average value \( \nu_t \) at the turbulent flow mode with analogous parameters (size of order \( 10^{-5} - 10^{-4} \)).

After computing the component of shear stress vector, we determine vector \( \bar{q} = (q_1, q_2) \) of bottom seabed material in unit time per unit of length, \( i \)-th component of which is evaluated by the following formula

\[
q_i = q_0 \cdot \frac{\tau_i}{|\bar{\tau}_b|} - C' \cdot q_0 \cdot \frac{\partial h}{\partial x_i}, \quad i = 1, 2
\]  

(6)

Here, the first term represents the component of soil particles transport induced by flow in the basin, and the second addend soil shift due to bottom surface irregularities.

The accumulation factor is

\[
q_0 = \begin{cases} 
12 \cdot \sqrt{g} \cdot (s-1) \cdot d^3 \cdot (\theta - \theta_c) \cdot \sqrt{\theta}, & \theta > \theta_c \\
0, & \text{else} \end{cases}
\]  

(7)

the value of soil transport on the horizontal bottom which is equal to zero in case when Shields’ parameter does not exceed the critical value of \( \theta_c \). Herewith \( \theta_c = \theta_0 \frac{\sin(\alpha + \varphi)}{\sin(\varphi)} \), where \( \theta_0 \) - Shields’ rejection number for horizontal bottom, \( \alpha \) - tilting angle of bottom, and \( \varphi \) - angle of repose equal to 23°. In accordance with this formula, Shields rejection number compared with this parameter for horizontal bottom increases while moving up the slope and decreases when moving down. According to calculations, the value \( \theta_0 \) is in the range of 0, 01 < \( \theta_c \) < 0, 06.

After receiving the bed load transfer vector for computing the bottom changes formed due to soil transfer, the mass balance equation is solved

\[
\frac{\partial h}{\partial t} = \frac{1}{1 - p} \cdot \sum_i \frac{\partial q_i}{\partial x_i}, \quad i = 1, 2
\]  

(8)

written relating to function \( h \) of the elevation of the formed bottom profile over its initial level; here \( p \) bed load porosity.

4 Results of Numerical and Experimental Studies

The present paper represents several sets of numerical experiments compared with the laboratory tests laid out in [1] in detail.

Numerical Domain (Fig. 1) corresponds to the operating area in a model tank with a gravity type oil platform model. The operating area has the size of 12x6 meters. Left bottom corner is considered to be the origin of coordinates. At a distance of 4 meters from the Inlet there is an oil platform model with the size of 2x2 meters, the model has chamfered corners (45 degrees) with the sides of 0.2x0.2 meters. The depth of fluid in the tank is 0.3 meters.

The patterned finite-volume grid consisting of octagons is used for discretization of a numerical domain. The partition along axes Ox and Oy amounts for
20 points per 1 meter, and along axis Oz - 20 points per 0.3 meter with the concentration to the bottom of a numerical domain (proportion of the vertical dimension of a near-bottom mesh to the top one 1:60). The vertical dimension of a near-bottom mesh is 1 mm. Such partition is prescribed by 556 664 points and it contains 531 200 meshes.

The first set of computation refers to the use of a laminar fluid flow model studied in [3] in detail. The velocity of inlet flow corresponds to the average by depth of fluid flow velocity in a laboratory test and is equal to 0.25 m/s. Fig. 2, Fig. 3 contain soil discontinuity structures received in numerical and laboratory studies accordingly.

<table>
<thead>
<tr>
<th>Number of point</th>
<th>Numerical result, mm</th>
<th>Measurement result, mm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-77.75</td>
<td>-75.34</td>
</tr>
<tr>
<td>2</td>
<td>74.64</td>
<td>71.81</td>
</tr>
<tr>
<td>3</td>
<td>31.39</td>
<td>33.09</td>
</tr>
<tr>
<td>4</td>
<td>-34.42</td>
<td>-31.19</td>
</tr>
<tr>
<td>5</td>
<td>10.21</td>
<td>11.17</td>
</tr>
</tbody>
</table>

Table 1. Comparison between experimental and numerical studies

The detailed numerical studies of soil erosion undertaken in [3] with the use of finite difference methods and reproduced in this study with the help of finite volume approaches show that in case of low velocity (up to 0.3 m/s) the use of laminar fluid flow model allows to receive good quantitative agreement with laboratory tests data with the significant economy of computational resources.
Table 1 contains the bottom height function value resulted from numeric computation and experiment in the points shown in Fig. 2.

The second set of computation refers to the use of a turbulent fluid flow model represented in this study, with the same value (0.25 m/s) of inlet flow velocity. These computations applied structured computational grid. The segmentation along axes OX and OY is 40 points per 1 meter and along the axis OZ is 40 points per 0.3 meter, with the concentration to the bottom of the computation domain. The vertical dimension of a near-bottom mesh is 0.01 mm which corresponds to dimensionless distance to the wall $y^+ = 1$.

Fig. 4 shows streamlines drawn up in horizontal plane with $z=0.1$ for different time points demonstrating eddy formation (a), its development (b), separation (c) and cycle mode (d, e). These figures allow suggesting the periodic nature of flow. Though the chart (represented in Fig. 5) of correspondence of velocity vector magnitude to time (0 sec $< t < 1000$ sec) in the point beside the streamlined platform with the following coordinates (7, 4.5, 0.15) allows to conclude that the flow nature is close to periodic, but it is not such.

Fig. 6 contains space patterns of fluid flow demonstrating formation and development of complex eddies.

Fig. 7 represents the structure of soil erosion received by a numerical experiment.

The results of computation correspond qualitatively and quantitatively with the data of laboratory studies and results of numerical computation received with the use of a laminar fluid flow model.

The third set of computation refers to the increase of inlet fluid flow velocity. Fig. 8 demonstrates the pattern of soil erosion when the inlet flow velocity is equal to 0.5 m/s.

In this case the flow structure acquires significantly turbulent nature that stipulates essentially different arrangement of erosion areas and accretion areas.

Fig. 3. Structure of soil erosion. Laboratory experiments data (mm).
Fig. 4. Streamlines drawn up in horizontal plane z=10.
Fig. 5. Flow time variations in the point (7, 4.5, 0.15)

Fig. 6. Formation and development of complex eddies.
5 Conclusion

The present paper represents the numerical studies of the process of soil erosion near foundations of gravity type oil platforms under conditions of developing turbulence of fluid flow. The calculation data show that when the flow hydrodynamic behavior is of low values, the small-scale turbulence has little impact on erosion structure, and in this case the laminar flow model for computing bottom velocity is more preferable. If the input flow velocity is increased, the turbulence mode of fluid flow obtains the paramount importance for outwash...
and inwash (accretion) areas, and the application of the turbulent model for estimating hydrodynamic values becomes principal in this case.

Acknowledgments. The study has been executed within the frames of Federal Assignment # 1.630.2014/K "Simulation of Fluid Flow with Variable Density and Viscosity for Applied Issues".

References

Numerical Simulation of Surface Waves Arising from Underwater Landslide Movement

Yury Zakharov and Anton Zimin

Kemerovo State University,
Krasnaya Street, 6, 650043 Kemerovo, Russia
zaxarovyn@rambler.ru, sliii@mail.ru

Abstract. The main aim of this paper is to construct a model of simultaneous movement of the landslide, internal currents and surface waves that can come ashore. The idea of a multicomponent fluid movement is used in this paper. We consider soil, liquid and gas as components of non-homogeneous fluid. Movement of such fluid is described by the Navier-Stokes equations with variable density and viscosity and the convection-diffusion equations. Special ratios are used to calculate the density and the viscosity of the medium. The results of test calculations for two-dimensional problems of the wave generation are presented.

Keywords: Navier-Stokes equations, surface wave propagation, landslide movement, inhomogeneous fluid, multicomponent fluid

1 Introduction

Tsunami waves generated by underwater and above-water landslides can be dangerous for buildings located on the shore and settled lands. Under natural conditions the underwater landslide is a movement of some soil mass along the slope of the bottom. Large volumes of moving mass generate surface waves that are close in their characteristics to the waves generated by tsunamigenic earthquake. The overview of historical landslides and tsunamis that they generated can be found in [1,2].

Construction of tsunami wave model generated by landslide movement can be divided into two tasks: construction of a model of wave propagation on a free surface and a model of landslide movement on the bottom of water basin.

Free surface of fluid means the border between fluid and gas that is above it. Due to the fact that fluid density is several times greater than gas density, influence of gas on the movement of fluid is often neglected, and it is considered that fluid moves independently of gas movement or, in other words, “freely”.

Models of wave fluid dynamics are the examples of such approach. They include shallow water theory equations, ideal fluid movement equations, etc [3]. In case of complex wave movements, big splashes and active interaction of two phases, multicomponent non-homogeneous medium is considered, where fluid and gas are separate components with their own values of density and viscosity [4].
According to the manner of discretization, mathematical models of multicomponent fluid movement can be divided into Lagrangian and Eulerian. Lagrangian methods are based on recording the equations of the medium movements in Lagrangian coordinates connected with particles of the moving medium. A set of nodal points moving together with the medium can be used in order to get the discrete analogues of such equations. It can be grid points (grid Lagrangian methods [5,6]) or point particles that are not connected with each other by grid lines (meshless Lagrangian methods [7-9]). In this approach the position of the interphase boundary is tracked automatically.

For discretization of medium movement equations in Euler approaches a fixed computational grid is usually used. At this, the interphase boundary moves on the grid, and special methods are applied in order to determine its position. These methods include MAC [10,11], VOF [12,13], Level Set [14,15].

In laboratory studies the underwater landslide can be imitated either by movement along the slope of a fully submerged solid body [16], or by some granulated soil slipping down the basin [17].

Several approaches are identified for computational modelling of landslide movement. It can be a model of movement of an absolutely solid body [18] or an ensemble of such bodies [19]; or a model of fluid flow that has different density, viscosity etc. [20].

Until recently, the majority of models applied for modelling of the tsunami of landslide type, relied on nonlinear shallow water wave theory [21,22]. In many cases the Boussinesq equations are applied [23]. And also the attempts are made to apply three-dimensional non-homogeneous models based on Navier-Stokes equations and the concentration transfer equations [24].

The aim of this paper is to apply the model of three-component viscous incompressible fluid with variable viscosity and density and with the presence of mass diffusion between the components for the problems of occurrence of surface waves generated by landslide movement. Previously two-component model has been used in problems of substance diffusion in the branched channel [25], cohesive soil erosion [26] and surface wave propagation [27].

2 Mathematical model

The movement of the medium consisting of three incompressible interfusing fluids with constant density $\rho_1$, $\rho_2$, $\rho_3$ and viscosities $\mu_1$, $\mu_2$, $\mu_3$ is considered. Let the particle of mixture be the solution $\bar{x} = \bar{x}(t)$ of the Cauchy problem $\frac{d\bar{x}}{dt} = \vec{V}(\bar{x}, t)$, $\bar{x}(0) = \bar{x}_0$, where $\vec{V}(\bar{x}, t) = (v_1, v_2, v_3)$ is a velocity vector of the mixture in point $\bar{x} = (x_1, x_2, x_3)$ and $t$ is a time point. Let $C_1(\bar{x}, t)$, $C_2(\bar{x}, t)$, $C_3(\bar{x}, t)$, $\mu$ and $\rho$ be volume concentrations of the components, dynamic viscosity and mixture density correspondingly. Components concentrations are interconnected in the following way:

$$C_1 + C_2 + C_3 = 1, \quad 0 \leq C_i \leq 1.$$  

(1)
In order to find out the values of viscosity and density of the mixture we have the following dependence on the concentrations of the components:

\[
\begin{align*}
\mu &= \frac{\mu_1 \mu_2 \mu_3}{C_1 \mu_2 \mu_3 + C_2 \mu_1 \mu_3 + C_3 \mu_1 \mu_2}, \\
\rho &= C_1 \rho_1 + C_2 \rho_2 + C_3 \rho_3. 
\end{align*}
\]  

(2)

Mass diffusion occurs between the particles of the mixture according to the law:

\[ q_n = -D \frac{\partial \rho}{\partial n}, \]

(3)

where \( D \) is the diffusion coefficient.

We consider the mixture components to possess the incompressibility property and its interfusion to form an incompressible medium, which density depends only on the concentrations. Let \( \omega_t \) be a control moving volume of such medium. Value of \( \omega_t \) remains constant due to the incompressibility:

\[ \int_{\omega_t} d\tau = const. \]

From known formula for the time differentiation of the integral taken over the moving volume [28]

\[ \frac{d}{dt} \int_{\omega_t} f d\tau = \int_{\omega_t} \left[ \frac{df}{dt} + f \text{div} \mathbf{V} \right] d\tau \]  

(4)

we obtain the condition of incompressibility:

\[ \text{div} \mathbf{V} = 0, \]  

(5)

where \( \frac{d}{dt} \) is total time derivative.

The equations of mass balance for fluid volume \( \omega_t \) considering the presence of mass diffusion, take the following form:

\[ \frac{d}{dt} \int_{\omega_t} \rho d\tau = -\int_{\partial \omega_t} q_n d\sigma, \]  

(6)

where \( q_n \) is defined in (3).

From (6) taking into account (3) and (5) we get convection-diffusion equation for density:

\[ \frac{d\rho}{dt} = D \Delta \rho, \]  

(7)

where \( \Delta \) is Laplace operator.

(5) and (7) together provide the condition of mass balance conservation in the medium for three-component incompressible fluid.
For our objectives we consider three-layered fluid, i.e. we suppose that the first and the third component do not directly interact in the solution:

\[
\begin{aligned}
C_3 &= 0, \quad C_1 \neq 0, \\
C_1 &= 0, \quad C_3 \neq 0.
\end{aligned}
\]

Then the density equation (2) will be the following:

\[
\rho = \begin{cases} 
C_1 \rho_1 + C_2 \rho_2, & C_3 = 0, \\
C_2 \rho_2 + C_3 \rho_3, & C_1 = 0.
\end{cases}
\]

Or considering (1)

\[
\rho = \begin{cases} 
C_1 \rho_1 + (1 - C_1) \rho_2, & C_3 = 0, \\
C_3 \rho_3 + (1 - C_3) \rho_2, & C_1 = 0.
\end{cases}
\] (8)

The diffusion coefficient \(D\) can be also expressed as:

\[
\begin{cases} 
D = D_1, & C_3 = 0, \\
D = D_3, & C_1 = 0.
\end{cases}
\] (9)

Then (7) can be transformed into the following equations for the component concentrations:

\[
\begin{aligned}
\frac{dC_1}{dt} &= D_1 \Delta C_1, \quad C_3 = 0, \\
\frac{dC_3}{dt} &= D_3 \Delta C_3, \quad C_1 = 0, \\
C_2 &= 1 - C_1 - C_3.
\end{aligned}
\] (10)

From the integral momentum equation

\[
\frac{d}{dt} \int_{\omega_t} \rho V \, dx = \int_{\partial \omega_t} P_n \, d\sigma + \int_{\omega_t} \rho F \, dx
\] (11)

considering (4) and (5) we get:

\[
\frac{d}{dt} \left( \rho \bar{V} \right) = div P + \rho \bar{F},
\] (12)

where \(P\) is the stress tensor in the mixture, \(\bar{F} = (f_1, f_2, f_3)\) is the vector of mass forces.

Then, considering variable viscosity, we obtain a system of equations for the motion of the mixture of three viscous incompressible interfusing fluids:
\[
\frac{d (\rho \mathbf{V})}{dt} = -\nabla p + \text{div} \mathbf{I} + \rho \mathbf{F},
\]
\[
d\text{div}\mathbf{V} = 0,
\]
\[
\frac{dC_1}{dt} = D_1 \Delta C_1,
\]
\[
\frac{dC_3}{dt} = D_3 \Delta C_3,
\]
\[
C_2 = 1 - C_1 - C_3,
\]
\[
\mu = \frac{\mu_1 \mu_2}{C_1 \mu_2 \mu_3 + C_2 \mu_1 \mu_3 + C_3 \mu_1 \mu_2},
\]
\[
\rho = \rho_1 C_1 + \rho_2 C_2 + \rho_3 C_3.
\]

where \( p \) is pressure in the mixture, \( \mathbf{I} = \mu \Delta \mathbf{V} + (\nabla \mu \cdot \nabla) \mathbf{V} + (\nabla \mu \cdot J_{\nabla}) \) is viscous part of stress tensor, \( J_{\nabla} \) is Jacobian matrix, arranged as follows:

\[
J_{\nabla} = \begin{pmatrix}
\frac{\partial v_1}{\partial x_1} & \frac{\partial v_1}{\partial x_2} & \frac{\partial v_1}{\partial x_3} \\
\frac{\partial v_2}{\partial x_1} & \frac{\partial v_2}{\partial x_2} & \frac{\partial v_2}{\partial x_3} \\
\frac{\partial v_3}{\partial x_1} & \frac{\partial v_3}{\partial x_2} & \frac{\partial v_3}{\partial x_3}
\end{pmatrix}.
\]

Thus, the given model consists of the convection-diffusion equations for concentration of the components, correlations for the determination of density and viscosity, and also hydrodynamic Navier-Stokes equations for incompressible viscous fluid.

We use a no-slip condition on the solid wall and boundary conditions of the second kind for concentration equations. Some initial distribution for concentrations is also given.

### 3 Solution scheme

For discretization of the system (13) in the spacial variables is used a finite-difference method on a rectangular uniform grid with staggered arrangement of nodes [29]: pressure, velocity divergence and component concentration are determined in the centers of cells; velocity vector components are determined on the borders of cells.

Application of the staggered grid allows to link speed and pressure values in the adjacent nodes and avoid the appearance of oscillations in solution, which arise when using central differences on a combined grid. Also, a staggered arrangement of the nodes automatically allows to satisfy the discrete representation of the continuity equation.

Time motion algorithm consists of the following stages:

1. By taking into account the known velocity vectors and concentration distribution (and thus density and viscosity values), a time step for the Navier-Stokes equations system is made.
2. Using the received values of velocity component, a time step for the “soil” convection-diffusion equation is made.
3. Using the received values of velocity component, a time step for the “air” convection-diffusion equation is made.
4. Knowing the distribution for “air” and “soil” concentrations, a value of “water” concentration is calculated according to the formula (1).
5. Recalculation of density and viscosity values in the medium is carried out according to the formulae (2). After that follows a transition to the first step of the next time iteration.

To solve the Navier-Stokes equations system there is used a splitting scheme on physical factors [30] with regard to variable density. It consists of three steps. At the first step a momentum transfer is carried out due to convection and diffusion; intermediate velocity field is calculated according to an implicit scheme:

\[
\frac{\tilde{V} - V^n}{\Delta t} = - (V^n \cdot \nabla) \tilde{V} + \frac{1}{\rho} \left( -V^n D \Delta \rho + \mu \Delta \tilde{V} + (\nabla \mu \cdot \nabla) \tilde{V} + (\nabla \mu \cdot J) \tilde{V} \right) + F,
\]  

(14)

In order to solve the system (14) a prediction-correction method is used [31]. The obtained system of algebraic equations is solved by sweep method.

At this, despite the fact that the obtained intermediate velocity field \(\tilde{V}\) does not satisfy continuity equation, it has a physical meaning because it preserves vortex characteristics in internal points.

At the second step, with regard to (5) and variable density, the pressure field is calculated according to the obtained intermediate velocity field \(\tilde{V}\):

\[
\sum_i \frac{\partial}{\partial x_i} \left( \frac{1}{\rho} \frac{\partial p^{n+1}}{\partial x_i} \right) = \frac{\nabla \tilde{V}}{\Delta t}.
\]

(15)

Solution of the system of equations obtained as a result of discretization of the equation in order to find pressure in (15) is one of the most important and dominant aspects of computational procedure from the viewpoint of machine resources expenses. Operator of this system can have a complex structure, which complicates the task significantly. To solve this stage of computational process a gradient iterative method BiCGStab [32] is used.

At the third step the transfer of momentum is carried out only due to pressure gradient:

\[
\frac{V^{n+1} - \tilde{V}}{\Delta t} = - \frac{1}{\rho} \nabla p^{n+1}
\]

(16)

The equation (15) obtained by taking the divergence of both sides of equation (16) with regard to \(\nabla V^{n+1} = 0\).
To solve the convection-diffusion equations (10) a prediction-correction scheme with approximation of convective components against the stream is used [31]. The obtained system of algebraic equations is solved by sweep method. The numerical scheme has first-order temporal and spatial approximations.

4 Results

4.1 Collapse of viscous soil

There was considered a problem of collapse of viscous and stiff soil on the bottom of reservoir that generates waves on the surface of fluid. Here one of the components (more stiff and viscous) models the behavior of bottom soil, another one – liquid, and the third one – aerial environment. Fig. 1 shows the geometry of area and the scheme of component arrangement.

![Fig. 1. Geometry and initial distribution of components.](image)

At the initial time the half circle of the wet soil is located at the center of the area. As time passes, it caves under the influence of gravity $\mathbf{F} = (0, -9.8)$ m/s² and causes the movement of the entire medium. Fig. 2 shows the results of the calculation for various time points. Parts of the bottom soil spread in the opposite directions, then reflected from the side walls and connected again in the center of the area. The liquid phase surface followed the bottom soil movements.

![Fig. 2. Picture of medium motion for various time points $t = 0.9$ s, 1.9 s, 2.7 s, 5.4 s.](images)
The following hydrodynamic parameters were chosen: $\mu_1 = 10$ Pa·s, $\mu_2 = 10^{-3}$ Pa·s, $\mu_3 = 10^{-5}$ Pa·s and $\rho_1 = 3000$ kg/m$^3$, $\rho_2 = 1000$ kg/m$^3$, $\rho_3 = 1$ kg/m$^3$ for the soil, liquid and gas phases. The following grid parameters and time step were used: $hx = 5 \cdot 10^{-2}$ m, $hy = 5 \cdot 10^{-2}$ m, $\tau = 10^{-2}$ s. All the area walls are solid except the upper one. The atmosphere pressure $P_{atm} = 101325$ Pa is indicated at the top. We consider the boundary between components to take place at $C = 0.4$.

The calculation demonstrates possibility of the model to simulate direct interaction between the bottom soil and the surface waves without distinguishing characteristics at the phases boundaries.

### 4.2 Soil movement on the inclined bottom

A problem of the soil slip movement on the inclined bottom that generates waves on the surface of fluid was considered. The scheme of area was taken from [33] (see Fig. 3). The result of calculation was compared with one of the numerical model presented in [33] and with laboratory experiment carried out in [34].

![Fig. 3. Geometry and initial condition.](image)

At the initial time the wet soil is located on the left side of the inclined bottom. As time passes, it rolls down by gravity $\vec{F} = (0, -9.8) \text{ m/s}^2$ and causes the movement of the entire structure, simulating the soil slip movement. The hydrodynamic parameters were used the same as in [33]: $\mu_1 = 10$ kg/m·s, $\mu_2 = 10^{-3}$ kg/m·s, $\mu_3 = 10^{-5}$ kg/m·s and $\rho_1 = 1950$ kg/m$^3$, $\rho_2 = 1000$ kg/m$^3$, $\rho_3 = 1$ kg/m$^3$ for the soil, liquid and gas phases. The following grid parameters and time step were used: $hx = 5 \cdot 10^{-2}$ m, $hy = 5 \cdot 10^{-2}$ m, $\tau = 10^{-2}$ s. All the area walls are solid except the upper one. The atmosphere pressure $P_{atm} = 101325$ Pa is indicated at the top. We consider the boundary between components to take place at $C = 0.4$. Fig. 4 shows the results of the calculation for various time points.

Fig. 5 shows the comparison of water surface elevations with results obtained in [33] and [34].

Calculation demonstrates good agreement with MM3 [33]. This approach produces similar waveforms and slide deformation geometries. However, there
**Fig. 4.** Picture of medium motion for various time points $t = 0.0 \, \text{s}, 0.4 \, \text{s}, 0.8 \, \text{s}, 1.0 \, \text{s}$. 

**Fig. 5.** A comparison of water surface elevations for various time points $t = 0.4 \, \text{s}, 0.8 \, \text{s}$. Given model (solid red), MM3 (dashed blue, [33]) and experiment (dotted green, [34]).
is a discrepancy in the fields of high gradients, what can be explained by the form of equation (2) for $\mu$. Differences do not exceed 12% on the whole surface for $t = 0.4\ s$; 20% for $t = 0.8\ s$. The surface shape in numerical simulation is qualitatively similar to that observed in the laboratory experiment. Also it has a slightly better agreement with experiment surface than MM3.

5 Conclusion

Presented model of three-component viscous incompressible fluid was used for modeling simultaneous movement of the landslide, internal currents and surface waves. The advantage of this approach is that it allows one to simulate the complex phenomenon of the interaction of waves and bottom soil using a uniform numerical algorithm for a number of different problems without distinguishing characteristics at the phases boundaries.

Test calculations for two-dimensional problems of the wave emergence and propagation on the free surface were carried out. The results obtained show the possibility of modeling such a phenomenon. Agreement with other model and experiment was demonstrated.

Acknowledgments. The work was carried out with support of state task of Ministry of Science and Education, project number 1.630.2014/K.

References


22. Petrukhin, N. S., Pelinovsky, E. N.: Riemann waves in the dynamics of the landslides above plane slope. Sovremennye problemy nauki i obrazovaniya, No. 6 (2011)


Abstract. Experimental and simulation results of the motion and heat exchange of the beryllia slurry in the annular cavity are presented. The slurry is a highly concentrated structured system where the mineral phase is beryllia powder and the liquid phase is an organic binder (paraffin, oleic acid and beeswax). Mathematical model describes motion and heat exchange of the liquid thermoplastic slurry of beryllia including the aggregate state change. The results of experiments and calculations show the process of molding of the slurry in the annular cavity. The obtained temperature field determines the transition of the slurry from liquid (viscous–plastic) to solid–plastic state. Calculated data of the isotherms of solidification zone of the thermoplastic slurry are in agreement with the experimental results.

Keywords: Thermoplastic slurry, Molding, Solidification Process, Heat Transfer

1 Introduction

The technology of hot pressure molding [1, 2] remains the basis for obtaining long-length, multi-channel, and complex shaped ceramics from non-plastic powders, in spite of using isostatic pressing. Nowadays technology of slurry casting (extrusion) is very relevant in connection with intensive development of MIM technology [3–8], where similar physical processes take place.

While a lot of attention has been paid to improve the technology and to create new equipment in previous years, obtaining fault free products by this method remains an unsolved problem up to now. As a result, the desired quality of moldings and obtaining of acceptable products often can not be achieved, which makes in practice this process low protable. Obtaining ceramic fabrications by hot molding from dispersion materials with anomalous physical properties, such as beryllia (BeO), is particularly complicated. In this case, the difficulties of obtaining the high quality products are caused firstly by thermal properties of BeO, in particular, its unique thermal conductivity. Beryllia ceramic exhibits the
highest thermal conductivity among all ceramic materials used in contemporary electronics, new fields of technology and special instrument building [9, 10].

Clearly, it is impossible to eliminate technological limitations and problems without the development based on all experience and knowledge of theoretical representations about regularities and mechanisms of regulation of the thermal conditions on the forming process of molding.

The results of the experiment and the generalizations by calculations of mathematical model of molding process of the beryllia thermoplastic slurry are presented in this paper.

2 Experimental research of the solidification process of molding

Experimental research of the effect of casting conditions on the temperature field in the zone of solidification of the molding was made on the experimental bushing (Fig. 1), by measuring the temperature using a thermocouple installed on the different levels by the height of crystallizer. Experimental bushing is designed for casting of circular tube with the outer diameter 0.02 m and the inner diameter 0.012 m. Material of mandrel and crystallizer is steel of grade X18H10T. The total height of the cylindrical part of the annular cavity is $H = 0.028$ m, the height of the hot zone of the annular cavity is $h_1 = 0.008$ m, the height of the cold zone of the annular cavity is $h_2 = 0.020$ m. Water with the temperature $t_1 = 80^\circ$C is fed to the upper contour of the crystallizer. Water with the temperature $t_2 = 15-20^\circ$C is fed to the bottom cold contour. The maximum water flow capacity of the crystallizer is 1500 l/hour.

Fig. 1. Scheme of molding solidification in the bushing. (1 – molding, 2 – mandrel, 3 – thermocouple, 4 – bushing, 5 – liquid slurry, 6 – zone of solidification).

Conical input part of the bushing is connected with the working tank of casting installation where BeO slurry is kept. Slurry flows to the conical inlet
of the annular cavity with an initial temperature of \( t_0 = 80^\circ C \). Slurry changes aggregate state and solidifies in result of heat exchange with the walls of mandrel and bushing.

According to the experimental data, dependence of position of the boundary of solidification slurry on different parameters of casting was constructed. The shape of the curve of surface solidification, which is dependent on the casting parameter, is defined by taking into account that the temperature changes linearly by the height and radius of crystallizer on the short segments.

The influence of molding speed on the thermal regime of the casting control was determined in the first series of experiments. The flow rate and temperature values of hot and cold water in the cooling contours are presented in Table 1. Fig. 2 shows positions of solidification zones depending on the molding speed. Isotherm of the AB corresponds to the temperature 54\(^\circ\)C and isotherm of the CD to 40\(^\circ\)C (Fig. 2).

### Table 1. Regimes of experiments as a function of the molding speed

<table>
<thead>
<tr>
<th>Number of the diagram</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molding regime</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hot water flow rate, l/hour</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Cold water flow rate, l/hour</td>
<td>1500</td>
<td>1500</td>
<td>1500</td>
<td>1500</td>
<td>1500</td>
</tr>
<tr>
<td>Molding speed, mm/min</td>
<td>20</td>
<td>40</td>
<td>60</td>
<td>80</td>
<td>100</td>
</tr>
<tr>
<td>Hot water temperature, (^\circ)C</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>Cold water temperature, (^\circ)C</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

The increase of the molding speed leads to expansion of solidification zone and its movement to the area of heat extraction of cold contour (Fig. 2). It explains that with increase of molding speed, heat extraction on the walls of the annular cavity does not have time to cool the slurry, and solidification zone extends, and it moves downstream.

In the second series of experiments the influence of cold water temperature and flow rate on the thermal regime of molding solidification was investigated (Table 2).

### Table 2. Regimes of experiments as a function of the cold water flow rate and temperature

<table>
<thead>
<tr>
<th>Number of the diagram</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molding regime</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hot water flow rate, l/hour</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Cold water flow rate, l/hour</td>
<td>1000</td>
<td>500</td>
<td>2500</td>
<td>1500</td>
<td>250</td>
</tr>
<tr>
<td>Molding speed, mm/min</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Hot water temperature, (^\circ)C</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>Cold water temperature, (^\circ)C</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>
The results of the second series of experiments are shown in Fig. 3. Solidification zones of molding are located in the area of cold contour, and move downstream with reducing cold water flow rate. Reducing of cold water flow rate leads to reduction of the heat extraction on the wall of bushing (Fig. 3).

In the fourth experiment, the beginning of solidification zone is located nearer to the hot contour area. Increase of cold water temperature leads to reduction of temperature difference of the hot slurry and cold water. In the fifth experiment parallel with the increase of cold water temperature reduction of its flow rate takes place (Table 2). Therefore, heat extraction reduces and solidification zone of molding shifts down (Fig. 3).

Accordingly, the experimental data show that the thermal regime of molding of the beryllia thermoplastic slurry is sensitive to change of the molding speed and to heat extraction conditions on the walls of the annular cavity.

3 Mathematical model of the solidification process

Motion and heat exchange of the beryllia thermoplastic slurry in the annular cavity are considered. The slurry flows into cavity with initial temperature of $t_0=80^\circ C$ (Fig. 1). As it moves the slurry is cooled and solidified, and on the output from the cavity it acquires structural form of the tube. The slurry flows in the laminar regime. Due to high viscosity of thermoplastic slurry, Prandtl number is much higher than one. Density of the slurry mass is a function of temperature and it increases while solidification process.

The problem is considered in Cartesian coordinate system with axis $z$ and $r$. OZ axis coincides with the cavity axis direction, and OR axis is radially directed to it ($r_1$, radius of the mandrel; $r_2$, radius of the bushing; $r_2-r_1$, thickness of the annular cavity). Molding speed is directed vertically downward along the
Fig. 3. The position of solidification zone depending on the flow rate and the temperature of cold water. (AB – the solidus line (54 °C), CD – the solidus line (40 °C), BS – the surface of the bushing, MS – the surface of the mandrel).

OZ axis. Radial component of the slurry velocity originates due to the heat exchange of the slurry with the walls of annular cavity.

Rheological properties of the slurry change with temperature. The heat of phase transition is released during the change of state. Cooling of the slurry may lead to the irregularity of the temperature profile and rheological properties of the pressing molding. Solidification begins at the walls of the annular cavity, while in the central part slurry may be in liquid state. As a result, in-feeding of slurry for the compensation of internal shrinkage of volume in the cooling zone of the cavity may occur.

According to the experimental data, the slurry solidification occurs at the temperature range from 54 to 40°C. Binder of the slurry is in the amorphous state and passes from the liquid amorphous state to the solid–plastic amorphous state in the zone of solidification [11–14]. The total amount of heat released per unit mass of the slurry mass is determined by the change of enthalpy $H$ at the phase transition zone.

Heat capacity of the slurry changes in the transition zone. Increase of the enthalpy during the phase transition can be determined by the apparent heat capacity method [15–19]. In this method, the latent heat is taken into account by increasing the heat capacity in the phase transition zone. Changing of heat capacity can be represented as [15, 16]:

$$c_p = \begin{cases} c_s, & t < t_s \\ c_{in}, & t_s \leq t \leq t_l \\ c_l, & t > t_l \end{cases}$$

where $A_{in} = \{ \int_{t_s}^{t_l} c(t) dt + H_{1\rightarrow 2} \} / (t_l - t_s)$, $H_{1\rightarrow 2}$ – the phase transition specific enthalpy of beryllia slurry with binder mass fraction of $\omega = 0.117$ is determined by experimental data and is equal to $H_{1\rightarrow 2} = 7800$ J/kg [20].
In [17, 18] it is believed that the temperature in the transition zone changes linearly, the expression of the specific heat is defined as:

\[
c_p = c(t) + H_{1\rightarrow 2} \frac{\partial f_{sl}}{\partial t}
\]  

(2)

\[
f_{sl} = \begin{cases} 
0, & t < t_s \\
\frac{t-t_s}{t_l-t_s}, & t_s \leq t \leq t_l \\
1, & t > t_l
\end{cases} 
\]

solid phase
transition zone
liquid phase

In [19] phase transition function \(\alpha(\bar{t})\) is introduced to the transition zone to consider the latent heat, and changing of the slurry heat capacity is expressed by:

\[
c_p = c_s \cdot (1 - \alpha(\bar{t})) + c_l \cdot \alpha(\bar{t}) + H_{1\rightarrow 2} \frac{d\alpha}{dt}
\]  

(3)

where \(c_s\) – specific heat of the slurry in the solid state, \(c_l\) – specific heat of the slurry in the liquid state, \(\alpha(\bar{t})= 0\) for the pure solid slurry and \(\alpha(\bar{t})= 1\) for the pure liquid slurry, \(\bar{t}\) – dimensionless temperature of slurry.

According to the experimental data of beryllia slurry with binder mass fraction of \(\omega= 0.117\) function \(\alpha(\bar{t})\) takes a form \(\alpha(\bar{t}) = 5.712 \cdot \bar{t} - 2.8544\).

The equations (1) – (3) of the method of apparent heat capacity include the latent heat of the phase transition, and are convenient for calculations. For convenience position of the transition zone is not known in advance and is determined as a result of the calculations.

The rheological properties of the slurry for \(\omega = 0.117\) depend on temperature [21]:

\[
\mu(t) = 293.6259 \cdot \exp(-0.05816 \cdot t), \quad (Pa \cdot s)
\]  

(4)

\[
\tau_0(t) = 11.4 + 11.41 \cdot \exp \left(\frac{-(t-70.05)}{5.47}\right), \quad (Pa)
\]  

(5)

Density of the thermoplastic slurry is defined by the concentrations of the beryllium oxide powder and the binder:

\[
\rho = \frac{\rho_{BeO} \cdot \rho_{bin}}{((1-\omega) \rho_{bin} + \omega \cdot \rho_{BeO})}, \quad (g/cm^3)
\]  

(6)

where \(\rho_{BeO}\) is the density of the BeO powder, \(\rho_{bin}\) is the density of the binder, \(\omega\) is relative mass content of the binder in the fractions.

The density of the binder where \(\omega = 0.117\) is determined by Eq. (7):

\[
\rho_{bin}(t) = 0.852 + 0.0725 \cdot \cos(0.0561 \cdot (t + 273.15) - 16.7361), \quad (g/cm^3)
\]  

(7)

The density of the beryllium oxide is \(\rho_{BeO} = 3.02\) g/cm\(^3\). The density of the binder \(\rho_{bin}\) in the range of temperature from \(t = 80 - 40\) °C changed within
0.7797 to 0.9010 g/cm³ and the density of the thermoplastic slurry during solidification increases from 2.2457 to 2.3553 g/cm³ for the fraction \( \omega = 0.117 \).

Thermal conductivity of the slurry the binder mass fraction \( \omega = 0.117 \) has the following form [20]:

\[
\lambda = 1.6 + 4.8 \cdot \exp (-0.017 \cdot t), \quad W/(m \cdot ^\circ C)
\]  

(8)

In the experiments [11, 14] beryllia slurry shows thixotropic properties of non-Newtonian fluid, and is described by Shvedov-Bingham rheological model [22]. The motion of the slurry in annulus is considered to be steady-state and the system of equations in the narrow channel is used for its study [23, 24]:

\[
\rho u \frac{\partial u}{\partial z} + \rho v \frac{\partial u}{\partial r} = -\frac{dp}{dz} + \frac{1}{r} \frac{\partial}{\partial r} \left( r \mu \frac{\partial u}{\partial r} \right) - \frac{1}{r} \frac{\partial}{\partial r} \left( r \tau_0 \right)
\]  

(9)

\[
\frac{\partial \rho u}{\partial z} + \frac{1}{r} \frac{\partial}{\partial r} \left( r \rho u v \right) = 0
\]  

(10)

In the limit of solid-plastic state of the slurry the motion equation (9) expresses the squeezing-out of the molding from the cavity and takes the form:

\[
-\frac{dp}{dz} = \frac{1}{r} \frac{\partial}{\partial r} \left( r \tau_0 \right)
\]

In contrast to the motion equation (9) conduction heat transfer along the OZ axis is substantially due to solidification of the slurry, and heat of phase transition is determined by the apparent heat capacity method (3). In the steady-state solidification process of slurry energy equation can be written as [17–19]:

\[
\rho u c_p \frac{\partial t}{\partial z} + \rho v c_p \frac{\partial t}{\partial r} = \frac{\partial}{\partial z} \left( \lambda \frac{\partial t}{\partial z} \right) + \frac{1}{r} \frac{\partial}{\partial r} \left( r \lambda \frac{\partial t}{\partial r} \right)
\]  

(11)

The following notes are used in the Eqs. (9)–(11): \( z, r \) – axial and radial coordinates; \( u, v \) – components of the velocity vector; \( p, \rho, T, \tau_0, c_p, \mu, \lambda \) – pressure, density, temperature, shear stress, thermal capacity, viscosity and thermal conductivity of the slurry, respectively.

Condition of the mass flow rate conservation determines the pressure gradient for thermoplastic slurry extrude from the annular cavity [23]:

\[
2\pi \int_{r_1}^{r_2} r u dr = \pi \left( r_2^2 - r_1^2 \right) \rho_0 u_0
\]  

(12)

where \( r_1, r_2 \) – the radius of the mandrel and bushing, respectively.

Distributions of the velocity and the temperature at the inlet of the cylindrical portion are constant along the cross section of the annular cavity; respectively, all the thermo-physical properties of the slurry are constant:

\[
u = u_0, \quad v = 0, \quad t = t_0,
\]  

if \( z = 0 \).
On the cavity walls in the area of the liquid slurry state for velocity are put conditions of sticking:

\[ u_i = v_1 = 0, \quad i = 1, 2 \]  

if \( z > 0, \quad r = r_i \).

In solid–plastic state they are conditions impermeability and slip:

\[ v_1 = 0, \quad \left( \frac{\partial u}{\partial r} \right)_{r_i} = 0 \]  

if \( z > 0, \quad r = r_i \).

The assumption is that the heat from the hot slurry is transferred to the walls of bushing and mandrel. Then, condition of heat exchange can be applied to the surface of the mandrel [23]:

\[ \lambda \frac{\partial t}{\partial r} = \alpha_d (t_c - t_d), \]  

if \( z > 0, \quad r = r_1 \).

where \( \alpha_d \) – coefficient of heat exchange between the slurry and the wall of the mandrel, \( t_d \) – the temperature of the wall of mandrel, \( t_c \) – the average temperature of slurry in cross-section of the annular cavity.

If we mark temperature of the water in the hot and cold contours as \( t_1, t_2 \), we can put the boundary conditions on the wall of bushing as:

\[ -\lambda \frac{\partial t}{\partial r} = k (t - t_i), \quad i = 1, 2 \]  

if \( z > 0, \quad r = r_2 \).

where \( k \) is the coefficient of heat transfer on the wall of bushing.

At the outlet section of the cavity for temperature are put the following condition:

\[ \frac{\partial t}{\partial z} = 0 \]  

if \( z = l \).

Eqs. (9)–(12) and boundary conditions (13)–(18) are presented in the dimensionless form for convenience. Coordinates \( z, r \) are divided by \( r_1 \), velocity components \( u \) and \( v \) by \( u_0 \), pressure \( p \) by the value of dynamic head \( \rho_0 u_0^2 \), temperature \( t \) by \( t_0 \), density, yield point, coefficients of thermal capacity, viscosity, and thermal conductivity by their values at the temperature \( t_0 \).

Set of Eqs. (3)–(12) is solved numerically at boundary conditions of Eqs. (13)–(18) [25]. The considered zone is divided into elementary cells with sides \( \Delta z_i, \Delta r_j \). Different analogues of the motion equation (9) and energy (11) were obtained by the Crank–Nicolson method of the second order precision, and different analogue of Eq. (10) was obtained by two layer scheme of the second order precision [23]. Pressure gradient is defined by the splitting method [23] from the condition of conservation of mass flow rate (12).
Coefficients of heat exchange $\alpha_d$ and heat transfer $k$ on the walls of the annular cavity depend on conductive and convective heat flows, molding velocity, temperature and flow rate of water in the cooling circuits. Heat exchange on the mandrel wall $\alpha_d$ and heat transfer on the bushing wall $k$ were determined by comparison of experimental and calculated data. The inverse problem of an energy equation with unknown heat transfer coefficients at the walls has been solved so that the slurry temperature calculations coincide with the experimental temperature distributions of the slurry. Generalization of experimental and calculated data allows us to determine the dependence of the coefficients of heat transfer and heat exchange from the molding velocity.

4 Results of calculations

The calculation is performed under the same regime parameters and conditions as of the experiments. Calculated data by distribution of temperature in an annular cavity according to the conditions of the first series of experiments are demonstrated in Fig. 4 (Table. 1).

At the inlet of the cylindrical part of the annular cavity the temperature of the slurry is constant and equal to $t_0=80^\circ$C. In the area of hot contour isolines (isotherms) of the temperature show the zones of the constant values of the temperature and parameters of the slurry mass is in liquid state. In this part the temperature of the slurry and the hot water is the same, heat transfer practically does not occur on the bushing wall.

Cooling of the slurry mass starts in the area of cold contour. Difference of the temperature of the slurry mass and cold water leads to an intensive heat transfer in the second cooling contour, and to reduction of the temperature and to change of rheological properties of the slurry.

The slurry temperature field is variable and changes from 80 to 54$^\circ$C in the beginning of the second contour.

Isotherm with the temperature 54$^\circ$C expresses upper bound of solidification zone of the slurry mass, and isotherm 40$^\circ$C expresses the lower bound of the solidification zone. In the area of solidification between isotherms 54 and 40$^\circ$C the slurry passes from the liquid (viscous–plastic) state to solid–plastic state. The experimental data of isotherms "solidus" AB (54$^\circ$C) and "solidus" CD (40$^\circ$C) are shown in Fig. 2 and 3. It may be noticed an agreement between the calculated data and experiments of positions of isotherms AB and CD.

At the value of the molding velocity of $u_0=20$ mm/min position of the transition zone of the slurry from the liquid (viscous-plastic) state to solid-plastic state is located closer to the beginning of the cold contour of cooling. With increasing molding velocity from 40 to 100 mm/min the position of the transition zone begins to pull down towards movement of the molding and takes extensive areas. It explains that with increasing of molding velocity convective component of heat flow of the slurry mass increases. The position of the transition zone increases and it covers all length of the mold cavity (Fig. 4).
The effect of flow rate and temperature of cold water on the position of the transition zone of the molding by conditions of the experimental researches were determined in the second series of calculations (Table. 2). Calculation data of the temperature distribution and the position of the transition zone of molding, limiting by isotherms "solidus" AB and "solidus" CD are shown in Fig. 5. In the first three cases, the temperature of cold water is 15°C, and its flow rate reduces from 1000 to 250 l/hour, respectively. Reducing the temperature of cold water increases the intensity of heat extraction and reducing of its flow rate and vice versa, it slows down the process of heat extraction. Increasing the temperature of cold water till 20°C, as well as reducing of its flow rate leads to a reduction of heat extraction.

The calculated temperature data are in agreement with the experiment results (Fig. 5).

The experimental and calculated data show that beryllia thermoplastic slurry solidification process can be controlled by adjusting the flow rate and temperature of the cold water.

5 Conclusion

During the series of experiments of the research on the effect of casting regimes on the temperature field in the solidification zone of the molding were identified the followings:

- the position of solidification zone of the slurry mass when molding velocity changes from 20 to 100 mm/min;
Fig. 5. Comparison of the calculated and experimental data of distribution of the temperature depending on the flow rate and the temperature of cold water (Table 2).

- the position of solidification zone of the slurry mass in the form-building cavity depending on the water flow rate and the temperature in the cold contour of cooling.

Temperature distribution, estimated during the experiments, in the form-building cavity of bushing depending on the molding velocity and heat extraction conditions on the walls of form-building of annular cavity lets us determine the transition from liquid (viscous-plastic) state to solid-plastic one.

The experiment results were analyzed and generalized using mathematical model of the thermoplastic slurry molding process. The latent heat of the phase change has been accounted by the apparent heat capacity.

The mathematical model includes the equations of the law of conservation of mass, momentum and energy of non-Newtonian fluid with the Shvedov-Bingham’s rheological model. Rheological and thermo-physical properties of the slurry were found on the basis of experimental data and express dependence on the temperature. The coefficients of heat exchange and heat transfer on the walls of the annular cavity were determined by comparison of experimental and calculation data. The temperature field of the slurry in liquid (viscous-plastic) and solid-plastic states were obtained in the calculations. The positions of isotherm "solidus" (54 °C) and "solidus" (40 °C), expressing the upper and lower boundaries of the solidification zone position were determined.

The results of calculation are in agreement with the experimental data, and they show physical validity of the proposed mathematical model of the molding process of the beryllia thermoplastic slurry.
References

List of Authors

Alimzhanov, Yermek, 233
Amelina, Evgeniya, 244
Andreev, Victor, 258
Assilbekov, Bakhytzhan, 547
Astrakova, Anna, 423
Aubakirov, Sanzhar, 218

Baiburin, Yerzhan, 233
Barakhnin, Vladimir, 2, 218
Baranovskiy, Nikolay, 2
Bodyakin, Evgeniy, 16
Bychkov, Igor, 26
Bykhovtsev, Egor, 196
Byordov, Dmitriy, 208

Cherepanov, Roman, 271
Chernoskutov, Mikhail, 33
Cherny, Sergey, 423

Denic, Nebojsa, 52
Dimitrienko, Yuriy, 277
Dolgov, Dmitriy, 380
Dorofeev, Nikolay, 41, 74, 84
Dostovalov, Dmitry, 445

Eremenko, Vladimir, 74
Esipov, Denis, 423

Filatov, Yuri, 61

Gachenko, Andrey, 26
Gavrilovic, Snezana, 52
Geidarov, Nazim, 524
Gerasimov, Alexander, 271, 288
Golushko, Sergey, 244, 299

Goudov, Alexander, 61
Grecheneva, Anastasia, 41, 74, 84
Gribanova, Irina, 98
Gridin, Vladimir, 114
Grigorieva, Irina, 61, 380
Guskov, Andrey, 196

Hmelnov, Alexei, 26

Idimeshev, Semyon, 299
Isaev, Sergey, 166
Ivanov, Konstantin, 524

Jainakov, Amanbek, 312

Kamaev, Aleksandr, 126
Kamenshchikov, Leonid, 324
Karmanov, Dmitriy, 126
Kashirin, Aleksei, 458
Kasyanov, Victor, 141
Kasyanova, Elena, 141
Kazakov, Alexander, 334, 344
Khbatkhamuly, Yerzhan, 218
Khristenko, Yury, 288
Kochemazov, Stepan, 98, 155
Kononov, Dmitriy, 166
Konstantinov, Igor, 84, 174
Kontrec, Nataša, 187
Korac, Vukman, 52
Kosyakov, Denis, 196
Krasnov, Igor, 324
Kuklin, Evgeniy, 208
Kuranakov, Dmitriy, 423
Kuropatenko, Valentin, 357
Kuzichkin, Oleg, 41, 74, 84
Kuzmin, Georgy, 406

Lapin, Vasily, 423
Lazarev, Sergei, 174
Lebedev, Pavel, 334
Lee Hee Un, 61
Lempert, Anna, 334, 344

Makarov, Aleksandr, 508
Mansurova, Madina, 218, 233
Medralieva, Bubusara, 312
Mijajlovic, Zarko, 372
Milosevic, Hrannislav, 187, 380, 487
Mussina, Aigerim, 218

Nasyrova, Maria, 445
Nikitenko, Marina, 508
Nudner, Igor, 524
Nugumanova, Aliya, 233

Orlov, Sergey, 344
Orlov, Svyatoslav, 344
Otpuschenkov, Ilya, 98

Pashkov, Sergey, 288
Peretokin, Sergey, 16
Perminov, Valery, 61
Petrovic, Milena, 187
Petrovic, Vera, 52
Podmasteriev, Konstantin, 41
Polishchuk, Vladimir, 393
Polishchuk, Yuri, 393
Polshchikov, Kostiantyn, 174
Polshchikova, Olha, 174

Radovanovic, Dragan, 487
Ramazanova, Gaukhar, 547
Rugnikov, Gennady, 26

Sadovskaya, Oxana, 473
Sadovskii, Vladimir, 473

Sattinova, Zamira, 547
Sborschikov, Sergey, 277
Semenov, Alexander, 98, 155
Semenov, Konstantin, 524
Semisalov, Boris, 406
Shapeev, Vasily, 299
Shestakovskaya, Elena, 357
Shokin, Yuri, 380, 423
Shornikov, Yuri, 445
Simonov, Konstantin, 16
Smagin, Sergei, 126, 458
Smolekho, Irina, 473
Solodovnikov, Vladimir, 114
Sotnikov, Igor, 61
Sozykin, Andrey, 208
Spitsina, Tatiana, 487
Stukalov, Sergey, 524
Sukhenko, Viacheslav, 126
Sultangazieva, Rena, 312

Taltykina, Mariia, 458
Taseiko, Olga, 487

Ushenin, Konstantin, 208

Valjarivic, Aleksandar, 487
Voronina, Tatyana, 496
Vujakovic, Jelena, 187

Yankovich, Elena, 2
Yeltsov, Igor, 508
Yeltsov, Timofey, 508
Yurchenko, Andrey, 244
Yushkov, Alexander, 524

Zaikin, Oleg, 98, 155
Zakharov, Yuri, 380, 524, 535
Zakharova, Yulia, 277
Zavozkin, Sergey, 61
Zhapbasbayev, Uzak, 547
Zimin, Anton, 535
The book of proceedings is devoted to the recent advances in the fields of mathematical modeling and combining information technologies. The areas of revealed topics vary from the mathematical, computational and information methods to the modeling and simulation of challenging applications. In addition, many engineering and industrial applications of these technologies are presented. The book will be useful for researchers, practitioners and advanced students in the fields.