ISSN 2500-316X (Online)

RUSSIAN TECHNOLOGICAL JOURNAL

РОССИЙСКИЙ ТЕХНОЛОГИЧЕСКИЙ ЖУРНАЛ

Information systems.
Computer sciences.
Issues of information security

Multiple robots (robotic centers) and systems.
Remote sensing and nondestructive testing

Modern radio engineering and telecommunication systems

Micro- and nanoelectronics.
Condensed matter physics

Analytical instrument engineering and technology

Mathematical modeling

Economics of knowledge-intensive and high-tech enterprises and industries.
Management in organizational systems

Product quality management. Standardization

Philosophical foundations of technology and society

12+
www.rtj-mirea.ru

13(2) 2025

ISSN 2500-316X (Online)

RUSSIAN TECHNOLOGICAL JOURNAL

РОССИЙСКИЙ ТЕХНОЛОГИЧЕСКИЙ ЖУРНАЛ

- Information systems. Computer sciences. Issues of information security
- Multiple robots (robotic centers) and systems. Remote sensing and nondestructive testing
- Modern radio engineering and telecommunication systems
- Micro- and nanoelectronics. Condensed matter physics
- Analytical instrument engineering and technology
- Mathematical modeling
- Economics of knowledge-intensive and high-tech enterprises and industries. Management in organizational systems
- Product quality management. Standardization
- Philosophical foundations of technology and society

- Информационные системы. Информатика. Проблемы информационной безопасности
- Роботизированные комплексы и системы. Технологии дистанционного зондирования и неразрушающего контроля
- Современные радиотехнические и телекоммуникационные системы
- Микро- и наноэлектроника. Физика конденсированного состояния
- Аналитическое приборостроение и технологии
- Математическое моделирование
- Экономика наукоемких и высокотехнологичных предприятий и производств. Управление в организационных системах
- Управление качеством продукции. Стандартизация
- Мировоззренческие основы технологии и общества

Russian Technological Journal
2025, Vol. 13, No. 2

Russian Technological Journal
2025, том 13, № 2

https://www.rtj-mirea.ru

https://doi.org/10.32362/2500-316X

ISSN 2500-316X (Online)

Russian Technological Journal
2025, Vol. 13, No. 2

Publication date March 28, 2025.

The peer-reviewed scientific and technical journal highlights the issues of complex development of radio engineering, telecommunication and information systems, electronics and informatics, as well as the results of fundamental and applied interdisciplinary researches, technological and economical developments aimed at the development and improvement of the modern technological base.

Periodicity: bimonthly.
The journal was founded in December 2013. The titles were «Herald of MSTU MIREA» until 2016 (ISSN 2313-5026) and «Rossiiskii tekhnologicheskii zhurnal» from January 2016 until July 2021 (ISSN 2500-316X).

Founder and Publisher:
Federal State Budget
Educational Institution of Higher Education
«MIREA – Russian Technological University»
78, Vernadskogo pr., Moscow, 119454 Russia.

The journal is included into the List of peer-reviewed science press of the State Commission for Academic Degrees and Titles of Russian Federation. The Journal is included in Russian Science Citation Index (RSCI), Russian State Library (RSL), Science Index, eLibrary, Directory of Open Access Journals (DOAJ), Directory of Open Access Scholarly Resources (ROAD), Google Scholar, Ulrich's International Periodicals Directory.

The registration number ПИ № ФС 77 - 81733 was issued in August 19, 2021 by the Federal Service for Supervision of Communications, Information Technology, and Mass Media of Russia.

The subscription index of *Pressa Rossii*: **79641**.

Russian Technological Journal
2025, том 13, № 2

Дата опубликования 28 марта 2025 г.

Научно-технический рецензируемый журнал освещает вопросы комплексного развития радиотехнических, телекоммуникационных и информационных систем, электроники и информатики, а также результаты фундаментальных и прикладных междисциплинарных исследований, технологических и организационно-экономических разработок, направленных на развитие и совершенствование современной технологической базы.

Периодичность: один раз в два месяца.
Журнал основан в декабре 2013 года. До 2016 г. издавался под названием «Вестник МГТУ МИРЭА» (ISSN 2313-5026), а с января 2016 г. по июль 2021 г. под названием «Российский технологический журнал» (ISSN 2500-316X).

Учредитель и издатель:
федеральное государственное бюджетное образовательное учреждение высшего образования «МИРЭА – Российский технологический университет»
119454, РФ, г. Москва, пр-т Вернадского, д. 78.

Журнал входит в Перечень ведущих рецензируемых научных журналов ВАК РФ, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук и доктора наук, входит в RSCI, РГБ, РИНЦ, eLibrary, Directory of Open Access Journals (DOAJ), Directory of Open Access Scholarly Resources (ROAD), Google Scholar, Ulrich's International Periodicals Directory.

Регистрационный номер и дата принятия решения о регистрации СМИ ПИ № ФС 77 - 81733 от 19.08.2021 г. СМИ зарегистрировано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор).

Индекс по объединенному каталогу «Пресса России» **79641**.

https://www.rtj-mirea.ru

Editorial Board

Редакционная коллегия

Contents

Содержание

Information systems. Computer sciences. Issues of information security
Информационные системы. Информатика. Проблемы информационной безопасности

UDC 004.622
https://doi.org/10.32362/2500-316X-2025-13-2-7-17
EDN OQUHWL

RESEARCH ARTICLE

Dataset collection for automatic generation of commit messages

Ivan A. Kosyanenko [@], **Roman G. Bolbakov**

MIREA – Russian Technological University, Moscow, 119454 Russia
[@] *Corresponding author, e-mail: kosyanenko.edu@gmail.com*

Abstract

Objectives. In contemporary software development practice, version control systems are often used to manage the development process. Such systems allow developers to track changes in the codebase and convey the context of these changes through commit messages. The use of such messages to provide relevant and high-quality descriptions of the changes generally requires a high level of competence and time commitment from the developer. However, modern machine learning methods can enable the automation of this task. Therefore, the work sets out to provide a statistical and comparative analysis of the collected data sample with sets of changes in the program code and their descriptions in natural language.

Methods. In this study, a comprehensive approach was used, including data collection from popular GitHub repositories, preliminary data processing and filtering, as well as statistical analysis and natural language processing method (text vectorization). Cosine similarity was used as a means of assessing the semantic proximity between the first sentence and the full text of commit messages.

Results. A comprehensive study of the structure and quality of commit messages encompassed data collection from GitHub repositories and preliminary data cleansing. The research involved text vectorization of commit messages and evaluation of semantic similarity between the first sentences and full texts of messages using cosine similarity. The comparative analysis of message quality in the collected dataset and several analogous datasets used classification based on the CodeBERT model.

Conclusions. The analysis revealed a low level of cosine similarity (0.0969) between the first sentences and full texts of commit messages, indicating a weak semantic relationship between them and refuting the hypothesis that first sentences serve as summaries of message content. The low proportion of empty messages in the collected dataset at 0.0007% was significantly lower than expected, indicating high-quality data collection. The results of classification analysis showed that the proportion of messages categorized as “poor” in the collected dataset was 16.82%, substantially lower than comparable figures in other datasets, where this percentage ranged from 34.75% to 54.26%. This fact underscores the high quality of the collected dataset and its suitability for further application in automatic commit message generation systems.

Keywords: commit message generation, version control systems, description of changes in software code, cosine similarity, data filtering, text vectorization, dataset, machine learning

• **Submitted:** 01.03.2024 • **Revised:** 22.07.2024 • **Accepted:** 06.02.2025

For citation: Kosyanenko I.A., Bolbakov R.G. Dataset collection for automatic generation of commit messages. *Russian Technological Journal.* 2025;13(2):7–17. https://doi.org/10.32362/2500-316X-2025-13-2-7-17, https://elibrary.ru/OQUHWL

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Сбор и анализ датасета для задачи автоматической генерации сообщений коммитов

И.А. Косьяненко [@], Р.Г. Болбаков

МИРЭА – Российский технологический университет, Москва, 119454 Россия
[@] Автор для переписки, e-mail: kosyanenko.edu@gmail.com

Резюме

Цели. Для управления процессом разработки современного программного обеспечения нередко применяются системы контроля версий, которые позволяют фиксировать изменения в программном коде и передавать контекст этих изменений при помощи сообщений коммитов. Релевантное и качественное описание внесенных изменений при помощи таких сообщений требует от разработчика высокой компетенции и времени, но современные методы машинного обучения позволяют решать эту задачу автоматически. Целью работы является статистический и сравнительный анализ собранной выборки данных с наборами изменений в программном коде и их описаниями на естественном языке.

Методы. В исследовании использован комплексный подход, включающий сбор данных с популярных репозиториев на GitHub, предварительную обработку и фильтрацию данных, а также статистический анализ и метод обработки естественного языка (векторизация текста). Для оценки семантической близости между первым предложением и полным текстом сообщений коммитов было использовано косинусное сходство.

Результаты. Проведено исследование структуры и качества сообщений коммитов, включающее сбор данных из репозиториев GitHub и их предварительную очистку. Осуществлена векторизация текста сообщений коммитов и оценка семантической близости между первыми предложениями и полными текстами сообщений с использованием косинусного сходства. Выполнен сравнительный анализ качества сообщений в собранном датасете и в нескольких аналогичных наборах данных с помощью классификации при помощи модели CodeBERT.

Выводы. Проведенный анализ выявил низкий уровень косинусного сходства между первыми предложениями и полными текстами сообщений коммитов (0.0969), что свидетельствует о слабой семантической связи между ними и опровергает гипотезу о том, что первые предложения выступают в качестве обобщения содержания сообщений. Процентная доля пустых сообщений в собранном наборе данных составила лишь 0.0007%, что существенно ниже ожидаемого значения и указывает на высокое качество собранных данных. Классификационный анализ показал, что доля сообщений, отнесенных к категории «плохих», в собранном датасете составляет 16.82%, что значительно ниже аналогичных показателей в других сопоставимых наборах данных, где этот процент варьируется от 34.75% до 54.26%. Данный факт подчеркивает высокое качество собранного набора данных и его адекватность для дальнейшего применения в системах автоматической генерации сообщений коммитов.

Ключевые слова: генерация сообщений коммитов, системы контроля версий, описание изменений в программном коде, косинусное сходство, фильтрация данных, векторизация текста, датасет, машинное обучение

• **Поступила:** 01.03.2024 • **Доработана:** 22.07.2024 • **Принята к опубликованию:** 06.02.2025

Для цитирования: Косьяненко И.А., Болбаков Р.Г. Сбор и анализ датасета для задачи автоматической генерации сообщений коммитов. *Russian Technological Journal*. 2025;13(2):7–17. https://doi.org/10.32362/2500-316X-2025-13-2-7-17, https://elibrary.ru/OQUHWL

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

Glossary

Dataset—collection (set) of data that is usually processed and analyzed as a whole. In the context of machine learning, datasets usually contain examples used to train a model.

Commit (in the context of version control systems)—record of a specific set of changes to code. Each commit is usually accompanied by a message that describes the changes made.

INTRODUCTION

Version control systems, which play a key role in modern software development, allow developers to track and manage changes to code as a means of ensuring effective collaboration and improving product quality. One of the main elements of version control systems are commits, i.e., records of every significant change made to the code base. The important role played by commit messages consists in the context they provide for each change, helping other developers to understand what was done and why. For example, commit messages can be used to detect vulnerabilities in a software product.[1]

However, writing effective commit messages represents a difficult task requiring time and effort. As well as describing changes in program code, a good message should explain the reason for the change [1]; moreover, according to many recommendations, it should not exceed 30 words (lexemes, tokens).

Although several approaches have been proposed for the automatic generation of commit messages [2], studies show that about 50% of messages generated by automatic generation tools turn out to be irrelevant or incorrect [3].

The present work focuses on the collection and analysis of training data for an automatic commit message algorithm. Representing one of the most important factors affecting the quality and efficiency of machine learning models [4], a training dataset comprises a collection of examples that are used to train and test the model, as well as to evaluate its generalization ability. The quality of a training dataset depends on its size, diversity, representativeness, purity, and relevance to the machine learning task. A poor quality dataset can lead to a number of problems during model training, including overtraining [5], undertraining, high bias, and variance. This, in turn, can reduce the accuracy and completeness of the model predictions. Hence, the generation and optimization of the training dataset are critical steps in the machine learning process, requiring detailed analysis, training and data cleaning to ensure the quality and efficiency of the model.

1. THEORETICAL BACKGROUND AND REVIEW OF EXISTING DATASETS

1.1. Importance of dataset in machine learning tasks

In 2001, researchers from Microsoft noted [6] that, despite the availability of extensive text corpora on the Internet, natural language processing experts continued to use relatively small datasets (up to 1 million words) to train models, which were mainly focused on optimizing algorithms. Their work emphasizes that different algorithms, including simple ones, demonstrate similar high performance in the task of eliminating language ambiguity given sufficient data. In the experiments, four algorithms were trained on data with a one-word context window, gradually increasing the sample sizes. The results of the study are shown in Fig. 1.

In the experiments, the authors studied:

- a memory-based algorithm that stores and uses previous information for decision making;
- a winnow algorithm that applies techniques to discard unnecessary data or noise to improve model quality;
- a perceptron, representing the simplest model of a neural network used for binary classification;
- a naive Bayesian classifier based on the application of Bayes' theorem with the assumption of feature independence.

[1] Wan L. *Automated vulnerability detection system based on commit messages*: Master's Thesis. Singapore: Nanyang Technological University, 2019. 123 p.

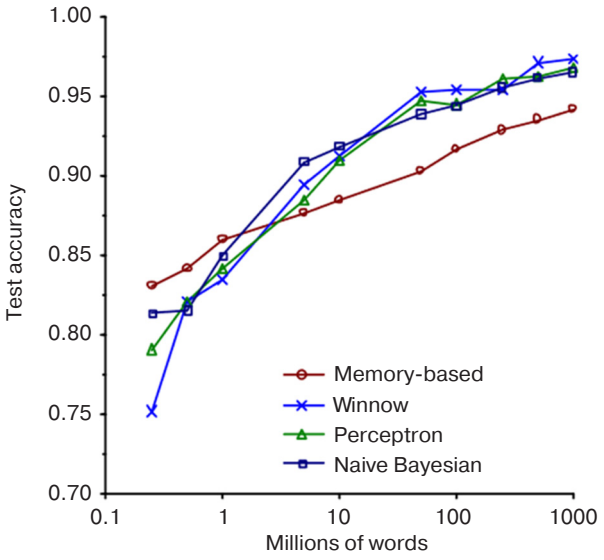


Fig. 1. Experimental results for increasing the size of the training corpus of text [6]

The results of the experiments show that the accuracy of the considered machine learning algorithms increases significantly as the data set increases. Notably, while the algorithms demonstrate varying accuracy on small amounts of data (less than 1 million words), as the amount of data increases, the differences between them become less significant. The authors noted that their results suggest that the trade-off between spending resources on algorithm development and on the aggregate development of the text corpus (dataset for model training) should be reconsidered.

The idea that data is more important than algorithms for complex tasks was popularized by Peter Norvig and other researchers from Google [7]. Conversely, expert judgment for data partitioning is often difficult and slow due to inconsistency in the estimates. The authors of the article conclude that useful semantic relations can be obtained by statistical methods relying on large volumes of unlabeled data used for text processing.

In the task of automatic generation of commit messages, the training sample plays a key role, since the quality and volume of data directly affect the model's ability to interpret and describe changes in the code. A large and diverse dataset containing examples of commits from different projects and written by different developers can improve a model's ability to generalize and adapt to new data.

Thus, when solving the problem of automatic commit message generation, it is necessary to pay due attention to the collection and preparation of the training data set, which will increase the efficiency and accuracy of the model, as well as its practical value for developers.

1.2. Overview of the existing datasets

Let us proceed to review existing datasets with commit messages [8]:

- **CommitGen** is one of the first commit datasets [9] collected from thousands of the most popular Java projects. Commits that do not carry meaningful load for message generation (e.g., rollback and merge) are excluded from the dataset. We also applied the Verb-Direct Object (V-DO) filter based on morphological analysis, which showed that messages often start with a verb followed by a direct complement [10]. As a result, the dataset contains 537000 labeled diff files.
- **NNGen** [3] improves the CommitGen by removing "noisy" data (16% of the original set).
- **CoDiSum** [11] is based on CommitGen, restricted to .java files and cleaned of special characters.
- **PtrGen** [12] includes 32663 <code:message> pairs from 2081 highly-valued Java projects, with special characters replaced by tokens.
- **MultiLang** [13] is a multi-language set of three popular repositories for Python, Java, JavaScript, and C++.
- **ATOM** [14] contains 197968 records after filtering noisy messages and commits without code changes from 56 most popular Java projects.
- **CommitChronicle** [15] is the largest commit dataset (as of July 2024), containing 10.7 mln commits for 20 different programming languages.

A key problem with many datasets is the focus on Java, which limits the applicability of a potential tool for automatically generating commit messages. To solve this problem, a dataset should be assembled that enables the generation tool to work with as many programming languages as possible and contain information about the relationships between code changes and the messages to them.

Each dataset has its own characteristics and limitations, while the selection of an appropriate dataset depends on the specific objectives and requirements of the study.

In the study [9], the authors note that approximately 14% of commit messages are empty. If true, such a fact serves as one of the justifications for the need of a commit message generation tool. Verifying this information is thus one of the research questions of the present work.

B1: What share of commit messages in the sample is empty?

Hypothesis B1: approximately 14% of the commit messages in the dataset under study are empty messages.

The result of hypothesis testing is presented in Paragraph 3.2.

2. METHODOLOGICAL BASIS

2.1. Dataset collection planning

In order to overcome the limitations on the number of programming languages in the datasets, it is necessary to define a list of relevant languages before the data collection process. The generated dataset should contain changes in program code written in languages from the selected set. Thus, a model trained on such a dataset will be able to generalize syntactic and semantic features of the selected languages and synthesize descriptions of changes in natural language on this basis. The relevance of program languages can be assessed on the basis of statistical studies.[2]

Since the choice of data sources directly affects the quality of the final dataset, the selection of donor repositories was done manually. Sources were selected based on the popularity of the repository as assessed by approval ratings from users on the GitHub[3] platform. However, a significant portion of popular repositories consisted of tutorials, which were excluded from the list of sources. The final list of repositories can be found in the online appendix to this article.[4]

Based on the specifics of the task, it was decided to extract the following features from each donor repository (Table 1):

Table 1. Attribute description

Attribute	Description
hash	Commit hash. It is generated by the version control system and serves as a change identifier
author	The identifier of the author of the message. May be required to form a more generalized sample, where the style of a particular author will not be predominant
commiter_date	Date and time of the commit
timezone	Author's timezone
parents	List of parent commits
message	Commit message. Description of changes in natural language
language	Programming language. The main language of the repository
changes	List of changes made in the commit

[2] Most used programming languages among developers worldwide as of 2023. Statista. https://www.statista.com/statistics/793628/worldwide-developer-survey-most-used-languages/. Accessed October 10, 2023.

[3] https://github.com/. Accessed October 10, 2023.

[4] Online appendix to the article. https://gist.github.com/Malomalsky/a243e43c00adb56fd11c19242a239275. Accessed February 06, 2025.

Two attributes—changes and message (natural language description)—are directly fed to the input of the commit message generation model. The other attributes, which are of a service nature, will be applied at the stage of filtering the collected data.

2.2. Methods and procedure for cleaning the dataset

In order for a dataset to be suitable for use in machine learning algorithms, it must be prepared. Preparation means filtering the data, cleaning it, and developing new features if required. The quality of the dataset—and consequently of the models that are trained on it—will depend on the above step.

The very first datasets of commit messages contained a large number of messages generated by "bots"—utilities that capture changes in the code repository and add trivial natural language descriptions to them [9]. Such messages could describe the changes made (answering the question "what was changed?"), but fail to provide enough context for making these changes (in other words, did not answer the question "why were the changes made?") [1]. Consequently, the number of such messages should be minimized in a high-quality dataset.

Formally speaking, one of the necessary procedures for cleaning the collected data should be the classification of messages into "natural" and "generated" and filtering of the latter. This task can be solved using the pattern matching method [16]—the author's name of an automatically generated message contains the token "[bot]". Thus, in most cases, a record with such a template in the *author* attribute can be classified as generated and deleted within the filtering process.

In addition to generated messages, trivial messages must also be filtered. According to the taxonomy of commit messages [1], these include messages generated by the git version control system itself (messages about merging branches in the repository) and duplicate messages—messages describing the content of changes that can be easily deduced from differences in the code (Update readme.md, Add <file name>). To classify and filter such messages, we can use regular expressions [17] by composing templates of trivial messages using the regular expression syntax.

One popular filter for datasets for the task of automatically generating commit messages is the V-DO filter, which emerged from the observation that about half of the commit messages correspond to the structure 'verb followed by an object' [10]. We can formally describe the V-DO filter in the form of a function:

$$f(m) = \begin{cases} 1, \exists i : (w_i, \text{verb and } w_{i+1}, \text{object}), \\ 0, \text{ in any other case.} \end{cases} \quad (1)$$

where m is the commit message, w_i is the ith word in the message m, and $f(m)$ is the result of the filtering. If $f(m) = 1$, the commit message passes through the filter, otherwise it does not.

Suppose we have a commit message *Minor changes to the database schema.* The V-DO filter will not pass this message because it does not start with a verb followed by a direct object. However, the commit message *Modified database schema* will be skipped by the V-DO filter because it matches the following pattern.

One important filter is the limit on the number of tokens (length) in a commit message [8]. An informative and relevant message should be neither too short nor too long. Most of the researchers [10–12] filtered the dataset by a maximum length of 30 tokens, although more recent work [15] focused on the range of 5 to 600 tokens in a message. In most cases, five tokens are insufficient to describe the changes made in a relevant way; conversely, leaving excessively long messages (more than 600 tokens) in the set may increase the disk space occupied by the set and the computational complexity to process it.

Some studies suggest leaving only the first sentence from a commit message [13] on the basis that the first sentence is often a generalization of the whole message. Such a filter would reduce the amount of disk space occupied by the dataset, but also potentially lead to the loss of important semantic information. Testing this hypothesis is another of the research questions of the work.

B2: do the first sentences in commit messages summarize the whole message?

Hypothesis B2: the first sentences in commit messages are a summary of the whole message.

The methods for verifying this hypothesis are outlined in paragraph 1.3; the result of the test is presented in paragraph 2.3.

2.3. Methods for checking semantic proximity of two text sequences

In the field of natural language processing, the cosine similarity measure [18, 19] is used to test the semantic proximity of two text sequences. Formally, cosine similarity reflects the cosine of the angle between vectors of the pre-Hilbert space and can be expressed as the formula:

$$similarity = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}}, \quad (2)$$

where a_i and b_i are the corresponding elements of the vectors **a** and **b**.

The range of the measure is from 0 to 1; if the measure is 0, then the vectors of the two sequences are orthogonal and far apart semantically. If the measure is 1, then the two text sequences have close semantic meaning.

It is also worth mentioning the methods of mapping symbolic sequences into vector (pre-Hilbert) space. In order to compute a measure on the collected data and test hypothesis B2, commit messages and their first sentences must be transformed into numerical vectors of the same dimensionality. Such a text vectorization process involves the conversion of text into numerical vectors that not only can be used by machine algorithms, but also reflect the semantics of the text due to the principle of distributive semantics.

One method of such vectorization is vectorization by hashing [20]. This tool applies a hash function to words and converts them into numeric indices in vector space, while preserving semantic relations between words.

3. RESULTS

3.1. Data collection process

The data was collected during the period from 15/10/2023 to 25/11/2023. The data source was the online GitHub service. The raw dataset contains 3141212 records and occupies 73 Gb of disk space. The distribution of records by programming languages is presented in Table 2.

Table 2. Distribution of the records by programming language

Language	Number of records in the collected dataset
C	932003
Ruby	253331
TypeScript	251127
C++	251125
Rust	211660
Python	209374
Java	141024
Go	140211
JavaScript	121610
C#	107960
Scala	86929
Dart	86532
Kotlin	81183
Lua	61622
PHP	61399
Groovy	50647
Shell	45687
R	22190
Swift	14396
Objective-C	11200

It is worth noting that the number of records for the different programming languages was not evenly distributed. If required, it would be appropriate to sample evenly from this set.

3.2. Analyzing and cleaning the collected data

The number of records comprised of empty messages in the collected dataset is 22, which is approximately 0.0007% of the total number of all records. The records with empty messages were removed from the dataset. In this connection, it is important to clarify that the data was collected from the most popular repositories, which are often owned by technology companies. Thus, this figure may not reflect the actual statistics on empty messages.

Hypothesis B1, that "approximately 14% of the commit messages in the dataset under study are empty messages," was not confirmed.

In the collected dataset, 2952077 messages do not match the V-DO filter, i.e., approximately 94%. Since its application would have resulted in a significant reduction in the size of the dataset, and along with it the useful dependencies between code changes and natural language, the decision was taken not to apply this filter.

In the unfiltered collected dataset, the maximum commit message length is 68529. This is a clear statistical outlier that can negatively affect the algorithm [21].

After removing empty messages, the following statistics were calculated for the number of tokens in commit messages (Table 3):

Table 3. Statistical indicators on the number of tokens in commit messages

Statistical indicators	Value
Number of records (count)	3141186
Mean value (mean)	52.19
Standard deviation (std)	129.61
Minimum (min)	1
25th percentile	10
Median (50th percentile)	42
75th percentile	174
Maximum (max)	68529

The standard deviation (approximately 129.61) is quite large, indicating a significant diversity in message length. The maximum value (68529) is much larger than the 75th percentile (174), indicating that there are outliers with a very large number of tokens.

In order to reduce the impact of outliers, it was decided to keep in the set only those records with the number of tokens in the range from 5 to 600 (inclusive). This filter reduced the number of records in the set to 2761945 or by 12.07%. Figure 2 presents a one-dimensional box plot showing the distribution of the number of tokens in commit messages in the collected dataset.

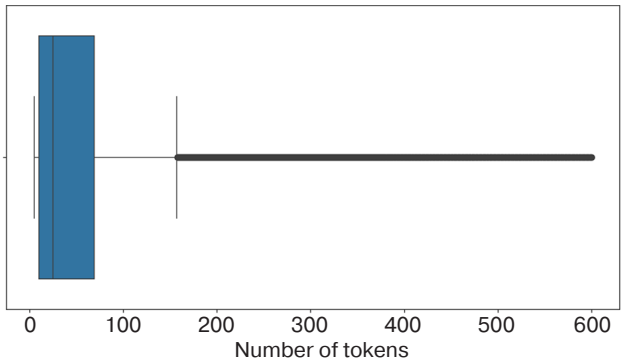


Fig. 2. Distribution of the number of tokens in the commit messages

While the bulk of the data is between 5 and 100 tokens, significant outliers are also visible, indicating a large number of messages with higher token counts.

Regular expressions were applied to filter trivial and generated messages. Keywords from CI-CD utilities were used as templates, e.g., *bump version to*. In addition, non-ASCII characters [9] and records with [bot] token in the *author* sign were removed. As a result of applying this filter, 185540 messages generated by automatic utilities were identified. After applying the filter, the number of records in the set amounted to 2576405.

3.3. Evaluating the semantic proximity of the first sentence and the whole commit message

After filtering, the hypothesis [13] that the first sentence in a commit message is a generalization of the whole message was tested.

In order to conduct experiments to test this hypothesis, a new feature (a column in the dataset) was created, then the first sentences and whole messages were mapped into numerical vectors using the HashingVectorizer utility from the scikit-learn library with a dimensionality of 1048576. Thus, the first sentences and whole messages were converted into vectors consisting of 1048576 numbers each. The program code for implementing the experiment is available in the online supplement to the article.[5]

The mean value of cosine similarity across all commit messages was 0.0969. This value reflects the degree of semantic proximity between the first sentence and the full text of commit messages. The relatively

[5] Online appendix to the article. https://gist.github.com/Malomalsky/a243e43c00adb56fd11c19242a239275. Accessed February 06, 2025.

low mean cosine similarity value may indicate that the first sentence often contains unique information that is not fully repeated in the rest of the message. For the purposes of this study, it was decided not to reduce the commit messages in the dataset to the first sentence.

Hypothesis B2, that "the first sentences in commit messages are a summary of the whole message," was not confirmed.

It is important to note that the cosine similarity procedure only measures the angle between vectors, not their length. This means that cosine similarity does not take into account the amount of information contained in commit messages. More sophisticated methods may be required to analyze the structure of commit messages in more depth.

3.4. Characteristics of the final dataset and its comparison with analogues

The cleaned dataset is available in the online appendix to the paper. The number of records in the dataset as a result of filtering was 2576405. The dataset contains change pairs for 20 programming languages thus extending the scope of a potential commit message generation model to be trained on the dataset. Automatically generated messages, along with trivial and empty messages, were removed from the dataset during the filtering phase.

In order to validate the quality of the dataset, a comparative analysis was performed on the collected dataset and the datasets mentioned earlier. A classification methodology using a pre-trained *commit-message-quality-codebert*[6] neural network was used to perform a comparative analysis of the quality of the commit messages in the proposed dataset. This model, based on the CodeBERT architecture [22], was pre-trained [23] for the task of classifying commit-message quality based on the previously mentioned message taxonomy [1]. According to this taxonomy, a message is "bad" that does not answer the questions 'Why were the changes made?' and 'What changed?', i.e., does not convey the context of the changes made.

The methodological procedure for the analysis included data preprocessing, which involved normalizing the text (lower case) and removing uninformative characters. Commit messages from the collected dataset and from other datasets available for comparison were then passed through a neural network for automatic classification based on the assigned "good" and "bad" labels. The classification results for each of the datasets are presented in Table 4.

From an analysis of the comparison results, the proportion of "bad" messages in the collected dataset was 16.82%, which is significantly lower than the other datasets.

CONCLUSIONS

The study collected an extensive multi-lingual dataset containing changes in program code, their natural language descriptions, and additional meta-information important in the context of filtering and cleaning the dataset. The dataset cleaned of generated and trivialized messages is hosted on the HuggingFace data hosting service.[7]

As part of the study, two hypotheses were proposed based on earlier work on the research topic. Hypothesis B1 was that approximately 14% of the commit messages in the studied dataset are likely to empty. However, when analyzing the collected data, it was found that only 0.0007% of the commit messages in the study sample are empty, which is much less than the expected value. Thus, hypothesis B1 was not confirmed during the study.

Hypothesis B2 was that the first sentences in the commit messages are a generalization of the whole message. To test this hypothesis, we vectorized the text

Table 4. Results of comparative dataset analysis

Dataset name	Total records	Classified as "good"	Classified as "bad"	Share of "bad" records, %
Collected dataset	2576405	2143000	433405	16.82
NNGen	27144	12415	14729	54.26
CoDiSum	90661	56305	34356	37.90
PtrGen	32663	16826	15837	48.49
MultiLang	126928	82819	44109	34.75

[6] Neural network classifier of messages to the commits. https://huggingface.co/saridormi/commit-message-quality-codebert. Accessed February 06, 2025.

[7] Collected dataset. https://huggingface.co/datasets/Malolmalsky/commit_dataset. Accessed February 06, 2025. https://doi.org/10.57967/hf/2216

and calculated the cosine similarity between the first sentence and the full text of the commit messages. In the study sample, the cosine similarity measure was 0.0969, indicating low semantic similarity between the first sentence and the full message text. Hence, hypothesis B2 was also not confirmed during the study.

The obtained results, which refute both hypotheses, indicate that empty commit messages are extremely rare, and that the first sentence of a commit message is an insufficient summary of the whole message. These findings can serve as a basis for further study of the structure and content of commit messages, as well as for the development of systems for automatic generation of commit messages.

The idea of conducting further research on vectorizing diff files (a representation of changes generated by a version control system) seems reasonable. If diff is a set of additions and deletions, then addition and subtraction operations should be defined for vectorized diff, where addition can be interpreted as adding program code and subtraction as removing fragments from it. Such a potential algorithm would solve the problem of reducing programming languages to a common formal notation under the assumption that identical program code changes made to files written in different programming languages would have close cosine distance. This question will be considered in further study.

It is also worth noting the feasibility of identifying generated messages using neural networks. While this approach will be much more computationally complex, it should provide greater accuracy in classifying commit messages.

Authors' contributions

I.A. Kosyanenko—research conceptualization, methodology development, data collection and analysis, computational experiments, preparation of the original draft of the manuscript.

R.G. Bolbakov—scientific supervision, manuscript review and editing, critical analysis of the obtained results.

REFERENCES

1. Tian Y., Zhang Y., Stol K., Jiang L., Liu H. What makes a good commit message? *Proceedings of the 44th International Conference on Software Engineerin*g. 2022;44:2389–2401. https://doi.org/10.1145/3510003.3510205
2. Kosyanenko I.A., Bolbakov R.G. About automatic generation of commit messages in version control systems. *International Journal of Open Information Technologies* (*INJOIT*). 2022;10(4):55–60 (in Russ.).
3. Liu Z., Xia X., Hassan A., Lo D., Xing Z. Neural-machine-translation-based commit message generation: how far are we? In: *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. 2018;33:373–384. https://doi.org/10.1145/3238147.3238190
4. Sun Z., Li L., Liu Y., Du X., Li L. On the importance of building high-quality training datasets for neural code search. In: *Proceedings of the 44th International Conference on Software Engineering*. 2022;44:1609–1620. https://doi.org/10.1145/3510003.3510160
5. Hawkins D.M. The problem of overfitting. *J. Chem. Inf. Comput. Sci*. 2004;44(1):1–12. https://doi.org/10.1021/ci0342472
6. Banko M., Brill E. Scaling to Very Very Large Corpora for Natural Language Disambiguation. In: *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. 2001;26–33. https://doi.org/10.3115/1073012.1073017
7. Halevy A., Norvig P., Pereira F. The unreasonable effectiveness of data. *IEEE Intell. Syst*. 2009;24(2):8–12. https://doi.org/10.1109/MIS.2009.36
8. Tao W., Wang Y., Shi E., Du L., Han S., Zhang H., Zhang D., Zhang W. On the evaluation of commit message generation models: An experimental study. In: *2021 IEEE International Conference on Software Maintenance and Evolution* (*ICSME*). *IEEE*. 2021;126–136. https://doi.org/10.48550/arXiv.2107.05373
9. Jiang S., McMillan C. Towards automatic generation of short summaries of commits. In: *2017 IEEE/ACM 25th International Conference on Program Comprehension* (*ICPC*). *IEEE*. 2017;320–323. https://doi.org/10.48550/arXiv.1703.09603
10. Myagkova E.Yu. To the problem of "formal" and "inner" grammar. *Vestnik Tverskogo gosudarstvennogo universiteta. Seriya: Filologiya = Herald of Tver State University. Series: Philology.* 2012;24(4):96–102 (in Russ.).
11. Xu S., Yao Y., Xu F., Gu T., Tong H., Lu J. Commit message generation for source code changes. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (*IJCAI-19*). 2019;3975–3981. https://doi.org/10.24963/ijcai.2019/552
12. Liu Q., Liu Z., Zhi H., Fan H., Du B., Qian Y. Generating commit messages from diffs using pointer-generator network. In: *2019 IEEE/ACM 16th International Conference on Mining Software Repositories* (*MSR*). *IEEE*. 2019;299–309. http://doi.org/10.1109/MSR.2019.00056
13. Loyola P., Marrese-Taylor E., Matsuo Y. A neural architecture for generating natural language descriptions from source code changes. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 2017;287–292. https://doi.org/10.18653/v1/P17-2045
14. Liu S., Gao C., Chen S., Yiu L., Liy Y. ATOM: Commit message generation based on abstract syntax tree and hybrid ranking. *IEEE Trans. Software Eng*. 2020;48(5):1800–1817. https://doi.org/10.48550/arXiv.1912.02972

15. Eliseeva A., Sokolov Y., Bogomolov E., Golubev Y., Dig D., Bryskin T. From Commit Message Generation to History-Aware Commit Message Completion. In: *2023 38th IEEE/ACM International Conference on Automated Software Engineering* (*ASE*). *IEEE.* 2023;723–735. https://doi.org/10.48550/arXiv.2308.07655
16. Dey T., Mousavi S., Ponce E. Detecting and characterizing bots that commit code. In: *Proceedings of the 17th international conference on mining software repositories.* 2020;209–219. https://doi.org/10.1145/3379597.3387478
17. Kuchnik M., Smith V., Amvrosiadis G. Validating large language models with ReLM. *Proceedings of Machine Learning and Systems*. 2023;5:457–476. https://doi.org/10.48550/arXiv.2211.15458
18. Haque S., Zachary E. Semantic similarity metrics for evaluating source code summarization. In: *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension.* 2022;36–47. https://doi.org/10.1145/3524610.3527909
19. Rahutomo F., Kitasuka T., Aritsugi M. Semantic cosine similarity. In: *The 7th International Student Conference on Advanced Science and Technology* (*ICAST*). 2012;4(1):1–2.
20. Roshan R., Bhacho I.A., Zai S. Comparative Analysis of TF–IDF and Hashing Vectorizer for Fake News Detection in Sindhi: A Machine Learning and Deep Learning Approach. *Eng. Proc.* 2023;46(1):5. https://doi.org/10.3390/engproc2023046005
21. Aggarwal C.C., Yu P.S. Outlier Detection in High Dimensional Data. In: *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*. 2001;30(2):37–46. http://dx.doi.org/10.1145/376284.375668
22. Feng Z., Guo D., Tang F., et al. CodeBERT: A pre-trained model for programming and natural languages. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. P. 1536–1547. Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.139
23. Qasim R., Bangyal W.H. A fine-tuned BERT-based transfer learning approach for text classification. *J. Healthc. Eng.* 2022;2022:3498123. https://doi.org/10.1155/2022/3498123

About the authors

Ivan A. Kosyanenko, Postgraduate Student, Department of Instrumental and Applied Software, Institute of Information Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: kosyanenko.edu@gmail.com. RSCI SPIN-code 2592-5015, https://orcid.org/0009-0009-1804-9412

Roman G. Bolbakov, Cand. Sci. (Eng.), Associate Professor, Head of the Department of Instrumental and Applied Software, Institute of Information Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: bolbakov@mirea.ru. Scopus Author ID 57202836952, RSCI SPIN-code 4210-2560, http://orcid.org/0000-0002-4922-7260

Об авторах

Косьяненко Иван Александрович, аспирант, кафедра инструментального и прикладного программного обеспечения, Институт информационных технологий, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: kosyanenko.edu@gmail.com. SPIN-код РИНЦ 2592-5015, https://orcid.org/0009-0009-1804-9412

Болбаков Роман Геннадьевич, к.т.н., доцент, заведующий кафедрой инструментального и прикладного программного обеспечения, Институт информационных технологий, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: bolbakov@mirea.ru. Scopus Author ID 57202836952, SPIN-код РИНЦ 4210-2560, http://orcid.org/0000-0002-4922-7260

Translated from Russian into English by Lyudmila O. Bychkova
Edited for English language and spelling by Thomas A. Beavitt

Information systems. Computer sciences. Issues of information security

Информационные системы. Информатика. Проблемы информационной безопасности

UDC 519.113.115+681.3
https://doi.org/10.32362/2500-316X-2025-13-2-18-26
EDN PJVWFG

RESEARCH ARTICLE

Informational ontological modeling

Viktor Ya. Tsvetkov [@], Nikita S. Kurdyukov

MIREA – Russian Technological University, Moscow, 119454 Russia
[@] Corresponding author, e-mail: cvj7@mail.ru

Abstract

Objectives. Despite the wide application of the term “ontology” in philosophy and social sciences, ontological modeling in the fields of computer science and information theory remains poorly studied. The purpose of the work is to develop a methodology for the ontological modeling of information and to clarify the theory of information retrieval technology both in a broad sense and as part of ontological modeling. Relevant problems in ontological modeling include the necessity of demonstrating the difference between regularity and functional dependence.

Methods. To achieve the stated goal, a logically structural approach is used, including the construction of conceptual schemes and their description in terms of logical formalism. The logically structural approach includes the construction of conceptual schemes that serve to apply logical formalism. The basis of logical modeling involves the selection of related models. The extended information retrieval technology proposed for this purpose searches not for individual objects, but for groups of objects. Since ontological research is based on a transition from qualitative to quantitative description, the methods used include quantitative-qualitative transitions.

Results. A new concept of ontological modeling of information is introduced. The conditions of ontological modeling are substantiated. Relationships between the concepts of regularity and functionality are investigated. On this basis, an interpretation of regularity and functional dependence is given. Structural and formal differences between information modeling, information retrieval technologies, and ontological modeling are demonstrated. Three information retrieval tasks are described, of which the second and third tasks involving the search for a group of related objects and the search for relationships or connections within a group of related objects, respectively, are solved using ontological modeling. Formal schemes of ontological modeling are provided. The transition from relations to connections in the case of ontological modeling is demonstrated.

Conclusions. Ontological modeling is shown to be applicable only to related models or to models between which there is a commonality. A technology of ontological modeling is proposed, in which version information retrieval is the initial part, while the second option involves the use of cluster analysis technology. Since ontological modeling uses qualitatively quantitative transitions, the proposed variant can be used to extract implicit knowledge.

Keywords: modeling, ontological modeling, information retrieval, information field, regularity, generalization, logical structural description, related models

• **Submitted:** 23.06.2024 • **Revised:** 02.08.2024 • **Accepted:** 22.01.2025

For citation: Tsvetkov V.Ya., Kurdyukov N.S. Informational ontological modeling. *Russian Technological Journal.* 2025;13(2):18–26. https://doi.org/10.32362/2500-316X-2025-13-2-18-26, https://elibrary.ru/PJVWFG

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Информационное онтологическое моделирование

В.Я. Цветков [@], **Н.С. Курдюков**

МИРЭА – Российский технологический университет, Москва, 119454 Россия
[@] *Автор для переписки, e-mail: cvj7@mail.ru*

Резюме

Цели. Несмотря на широкое применение термина «онтология» в философии и социальных науках, в области информатики онтология и, тем более, онтологическое моделирование остаются мало изученными. Также мало исследована онтология в области информационного поля. Цель работы – разработка методики информационного онтологического моделирования и исследование технологии информационного поиска в широком смысле и как части онтологического моделирования. На основе онтологического моделирования необходимо показать различие между закономерностью и функциональной зависимостью.

Методы. Для достижения цели применен логически структурный подход, включающий построение концептуальных схем и логический формализм их описания. Логически структурный подход включает построение концептуальных схем, которые служат для применения логического формализма. Основой логического моделирования является выделение родственных моделей. Для этой цели предлагается применить расширенную технологию информационного поиска, которая ищет не отдельные объекты, а группы объектов. Онтологическое исследование строится на применении перехода от качественного описания к количественному. К числу применяемых методов относится метод количественно-качественных переходов.

Результаты. Вводится новое понятие – информационное онтологическое моделирование. Обоснованы условия онтологического моделирования. Исследованы отношения между понятиями закономерности и функциональности. На этой основе дается трактовка закономерности и функциональной зависимости. Показано структурное и формальное различие между информационным моделированием, технологиями информационного поиска и онтологическим моделированием. Раскрыты три задачи информационного поиска. При онтологическом моделировании решают вторую и третью задачи информационного поиска, соответственно, поиск группы связанных между собой объектов и поиск отношений или связей внутри группы связанных между собой объектов. Даны формальные схемы онтологического моделирования. Показан переход от отношений к связям в случае онтологического моделирования.

Выводы. Доказано, что онтологическое моделирование можно применять только к родственным моделям или к моделям, между которыми существует общность. Предложена технология онтологического моделирования, в варианте которой информационный поиск является начальной частью онтологического моделирования. Вторым вариантом является применение технологии кластерного анализа. Онтологическое моделирование использует качественно-количественные переходы и в предлагаемом варианте может служить для извлечения неявного знания.

Ключевые слова: моделирование, онтологическое моделирование, информационный поиск, информационное поле, закономерность, обобщение, логически структурное описание, родственные модели

• **Поступила:** 23.06.2024 • **Доработана:** 02.08.2024 • **Принята к опубликованию:** 22.01.2025

Для цитирования: Цветков В.Я., Курдюков Н.С. Информационное онтологическое моделирование. *Russian Technological Journal*. 2025;13(2):18–26. https://doi.org/10.32362/2500-316X-2025-13-2-18-26, https://elibrary.ru/PJVWFG

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

In philosophy, ontology is used to conceptualize reality [1]. In computer science, by contrast, ontology is differentiated by subject areas and refers to an information artifact [2]. As such, an ontology includes a vocabulary used to describe a topic defined reality, conventions used for complementarity, concept matching and contextual relations, as well as construction and analysis schemes. Ontologies were originally proposed for the verification and construction of conceptual models. Since verification involves the domain of logic, logical constructs are widely used in ontological modeling. Nowadays, ontologies are applied for knowledge extraction and experience building. At the same time, information retrieval [3] should be mentioned as a one of the primary areas in which ontological modeling was subsequently used. Since ontology in computer science involves information field (IF) theory [4], the concept of information ontology acquires salience. An information ontology is one that acquires relevance as part of a finite element (FE) model. Various types of modeling are widely used in intellectual property (IP) systems, of which the main one is informational modeling. This feature gives reason to talk about informational ontological modeling [5].

Information ontological modeling is fully consistent with the theoretical provisions of ontology, which are related to the definition of types, properties, and relationships of entities [1]. Simply stated, ontology is a theory of entities of objects and entities of their relations [6]. Here, a distinction is made between formal, descriptive and formalized ontologies. Formal ontology was introduced by Edmund Husserl in his Logical Investigations. According to Husserl, the object of ontology is the study of essence and important categories. In information sciences, this formal approach links ontology with taxonomy. Nevertheless, it is necessary to distinguish between ontologies of information entities and ontologies of information systems [7]. The application of the ontological approach is driven by the needs of modern information societies, in which information support and knowledge sharing is a key development factor. In the context of global resource exchange, knowledge acquisition and its methods deserve special attention. While there is no single methodology for systematic information modeling, a proposed ontology-based approach provides a semantic representation of information [8].

Information modeling is used for different purposes, one of which is to extract meaning and knowledge. Moreover, information modeling is can be considered as conceptual modeling or semantic data modeling. Thus, information modeling and ontological modeling may have useful points of commonality. A variant of information modeling is aimed at building an information model that represents conceptual aspects of objective and subjective reality. Since the conceptual framework of this methodology relies on ontologies and concepts that arise in ontological constructs, this may be said to constitute the essence of ontological information modeling. Due to the diversity of models and information technologies generating redundant requirements and data exchange rules, the work presented in [9] utilizes ontological principles. In [10], research results are presented on an ontology-based approach for building information modeling to facilitate information exchange between different applications of a subject area. The described approach is based on a generic information entity ontology that models the types of IS elements and the relationships between them. The information systems to be integrated must be modeled using the common ontology, according to which each knowledge domain adds its own element properties to the common ontology.

In the artificial intelligence domain, ontologies are used to generalize and reduce the complexity [11] of information. The use of topological models in ontologies greatly simplifies their analysis. Ontological models that are extended to the field of information retrieval can be ontological models of information retrieval. Thus, the importance of ontologies in IS and the need for ontological modeling becomes clear. Ontological modeling [12] aims at generalizing the properties of a number of related models, finding patterns and knowledge in this generalization. However, information retrieval [3], which in a broad sense refers to scientific research aimed at obtaining knowledge, precedes ontological modeling. Therefore, the combination ontological modeling and information retrieval becomes a novel and relevant area of study.

1. METHODOLOGY

The logically structural approach of analysis was used. The second and third tasks of information search were applied, the essence of which is disclosed below. The method of qualitative-quantitative transitions, methods of comparative and qualitative analysis were used.

2. RESULTS

2.1. Conceptual diagrams

The logical structural approach implies the construction of graphical schemes, which further serve as a basis for the construction of logical constructs. In the presented ontological modeling system, it is reasonable to consider information modeling and information retrieval as related to modeling processes for the purposes of comparison. The conceptual scheme of information modeling is shown in Fig. 1.

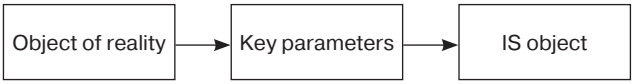


Fig. 1. Information modeling diagram

The basis of the model is formed by an object of reality, which is transformed into an IS object via key parameters, modeling conditions, and the set task. The IS object can be considered as synonymous with an information model. The structure of the information search model is shown in Fig. 2.

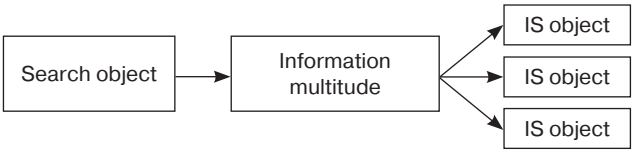


Fig. 2. Structure of information search

Information search can be performed for different purposes or tasks. The most common way to perform an information search is to find a single, user-required item. Here the first task will be to determine an object from the set of found objects in terms of the relevance of its features. The second task will be to find a group of related objects. In the third task of information search, it becomes necessary to find relations or connections within a group of related objects. Ontological modeling provides a means to solve the second and third search tasks. In all cases, the basis of the search is a search model, which may also be called a pattern. Linked objects in solving the second and third information search problem should be called *related objects*.

In the second and third search tasks, a discrete set of objects with commonalities is generated. The pattern generates a set of IP objects (Fig. 2). IP objects are identical to information models. Therefore, the information search in the second and third tasks generates a set of information models having commonalities.

The similarity between information retrieval and information modeling is that both technologies form information models. Single information modeling forms a single information model. Information retrieval forms a set of information models. In this totality, explicit and implicit patterns and relationships can be searched on the basis of ontological modeling. Figure 3 depicts a generalized model of ontological modeling.

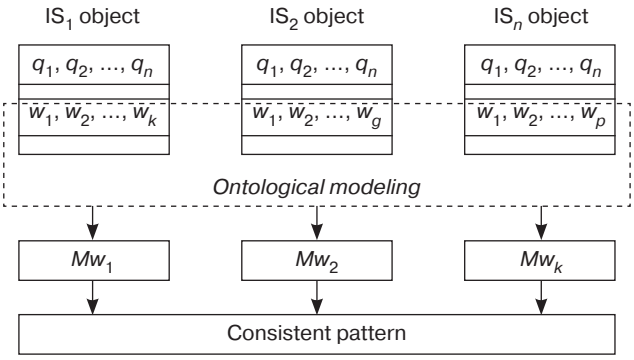


Fig. 3. Generalized model of ontological modeling

Three types of parameters are shown here: set or explicit $(q_1, q_2, \ldots, q_n)$; found or additional $(w_1, w_2, \ldots, w_k)$; generalized $(Mw_1, Mw_2, \ldots, Mw_k)$.

The number of these parameters is usually different. The number of given parameters may be greater or less than the number of found parameters. The number of generalized parameters is usually less than the number of found and set parameters.

2.2. Regularities of grouping of related objects

Objects that have commonalities and explicit or implicit connections may be described as related. Here it is necessary to distinguish between concepts of regularity, connection, and functional dependence. Regularity, as a rule, is a soft qualitative statement, having a logical or verbal form of representation. For example, an increase in the cost of a vehicle increases the cost of cargo transportation. Connection, on the other hand, is a hard dependence of one value on another or objects between each other. For example, the connection between cars in a train is realized by means of different types of couplings. Although there can be different kinds of couplings, all of them have a rigid connection in common. Finally, functional dependence is a relationship defined explicitly in the form of an analytic form. Any known law, such as Coulomb's law or the law of universal gravitation, is an example of functional dependence.

One way of finding related entities is information search in the aspect of the second and third search

tasks noted above. Information retrieval starts with the formation of a search entity using cognitive methods. A user can form a morphological pattern and supplement it with alternative parameters: type, size, date of file creation and others. Here it is fundamental that the search pattern be formed morphologically, not semantically.

A pattern (*Pat*) contains the query parameters $\{q\}$. The designation {} is used to describe a discrete set of values. In general, a search pattern can be represented as a regularity:

$$Pat\{q_i\} \rightarrow IR(IS) \rightarrow \{Ex_j\}, j = 1, \ldots, m; i = 1, \ldots, n. \quad (1)$$

Expression (1) is interpreted as follows. The query $Pat\{q_i\}$ is sent to the information set (*IS*) via the information retrieval (*IR*) technology. As a result, a discrete set $\{Ex_j\}$ is formed. The value n specifies the number of query parameters, while m specifies the number of instances selected in the information set based on the query. There is a regularity:

$$\uparrow n \rightarrow \uparrow m \rightarrow \uparrow t. \quad (2)$$

According to (2), an increase in the number of query parameters entails an increase in the number of instances and the search time t. This regularity leads to the need to minimize the number of search parameters.

Search $\{Ex_j\}$ as complete objects is an object search. In this case, the result of the query is an information model or an IS object, for example, a file. The result $\{Ex_j\}$ is a set of related objects for further analysis.

2.3. Modeling patterns

In contrast to information modeling, ontological modeling is a multi-stage process (Fig. 1). Ontological modeling begins with the selection of a group of objects that have commonalities. One variant of such selection is related to cognitive modeling. Another variant is based on the use of information search technology within the framework of tasks 2 and 3. Such information search can be represented as a process of clustering a heterogeneous multitude.

A search or clustering object can be an information model, a process model, a state model, relationships, or tacit knowledge. These objects have different degrees of abstraction. For task 1, there is an individual search, while for tasks 2 and 3 there is a group search. When forming a query for group search, an expert's experience or the search subject's cognitive abilities are used. The simplest search pattern is given in expression (1).

In expression (1) there are known, given parameters $(q_1, q_2, \ldots, q_n)$. Let us conditionally consider five instances in the group. We denote the newly found parameters by $(w_1, w_2, \ldots, w_k)$, where k is the total number of found parameters. As a result of the query, we have a total of $(n + k)$ parameters. The first sample of the group has the following form:

$$Ex_1(q_1, q_2, w_1, w_2, w_3, w_4). \quad (3)$$

From expression (3), we can see that the first instance contains two given and four found parameters. All six parameters describe the first sample.

The second sample of the group has the following form:

$$Ex_2(q_1, q_3, w_6, w_5, w_2, w_4) \quad (4)$$

and also contains two given and four found parameters. However, these parameters are different: q_2 appeared instead of q_3, while w_6, w_5 appeared instead of w_2, w_3. All six parameters describe the second sample.

The third sample of the group has the form:

$$Ex_3(q_1, q_3, q_n, w_1, w_3, w_6, w_8, w_4) \quad (5)$$

and contains three given and five found parameters. The parameters differ from the first instance. The parameter q_n appeared additionally, w_6, w_8 appeared instead of w_2. All eight parameters describe the third sample.

The fourth sample of the group has the form:

$$Ex_4(q_1, q_2, q_3, w_7, w_8, w_1, w_4) \quad (6)$$

and contains three given and four found parameters. The parameters differ from the first instance. Additionally, parameter q_3 appeared, while w_7, w_8 appeared instead of w_2, w_3. All seven parameters describe the fourth sample.

The fifth sample of the group has the form:

$$Ex_5(q_1, q_3, w_1, w_5, w_9, w_4) \quad (7)$$

and contains two given and four found parameters. The parameters differ from the first sample. Instead of the parameter q_2 there appeared q_3, instead of w_2, w_3 there appeared w_5, w_9. All six parameters describe the fifth sample.

The main disadvantage of group instance descriptions is that they exclude the description and influence of the situation in which the objects are located. Although it is acceptable to have different kinds of relations between parameters, different possible typical relationships between parameters should be emphasized:

$$Re_1(q_1, q_2, q_3, \ldots, q_n), \quad (8)$$

$$Re_2(q_1, q_2, w_1, \ldots, w_i), \quad (9)$$

$$Re_3(w_1, w_2, \ldots, w_k). \quad (10)$$

According to expression (8), there is a relationship between the query parameters. Expression (9) states that there is a relationship between a part of the query parameters and a part of the new found parameters, while expression (10) states that there is a relationship between the found parameters.

Relationships serve as a basis for establishing possible connections (*Con*) and functional dependencies (*F*). By analogy with (8)–(10) we can distinguish 3 possible groups of relations:

$$Con_1(q_1, q_2, q_3, \ldots, q_n), \quad (11)$$

$$Con_2(q_1, q_2, w_1, \ldots, w_i), \quad (12)$$

$$Con_3(w_1, w_2, \ldots, w_k). \quad (13)$$

Expression (11) states the possible existence of relations between query parameters, while expression (12) asserts the possible existence of relations between a part of query parameters and a part of new found parameters. Expression (13) states the possible existence of relations between found parameters.

The presence of relations can lead to functional dependence, for example, for expression (12) and (13) can appear a functional dependence of the following type:

$$Con_2(q_1, q_2, w_1, \ldots, w_i) \rightarrow \rightarrow Y = F_2(q_1, q_2, w_1, \ldots, w_i), \quad (14)$$

$$Con_2(w_1, w_2, \ldots, w_k) \rightarrow Y = F_3(w_1, w_2, \ldots, w_k). \quad (15)$$

Expressions (14) and (15) are of the qualitative-quantitative transition type. On the left side is a constant or a logical expression that serves as the basis for forming a functional dependence, while the functional dependence is indicated on the right side. Expression (14) hypothesizes that relationships between different parameters can lead to the formation of functional dependencies between different parameters. Expression (15) hypothesizes that links between new parameters can lead to the formation of functional dependence between new parameters. Expressions (14), (15), which can be considered as new knowledge, appeared after the identification of new parameters.

Ontological modeling is performed on the basis of additional analysis. For example, the analysis of instances in expressions (3)–(7) demonstrates the stability of occurrence of parameters q_1, w_1, w_4. This suggests that these parameters comprise common characteristics for different instances. This commonality is identified on a group of models related by a common theme. Common themes are organized either by the principle "from private to common" (information search) or by the principle "from common to private" (cluster analysis).

The result of further ontological modeling is three-level. On the first level, metaparameters are defined and emphasized. For expressions (3)–(7) these are q_1, w_1, w_4 and new metaparameters as functions are possible:

$$Mw_1 = \varphi_1(\{q\}, \{w\}), \quad (16)$$

$$Mw_2 = \varphi_2(\{q\}, \{w\}), \quad (17)$$

$$Mw_3 = \varphi_3(\{q\}, \{w\}). \quad (18)$$

Since the number and composition of arguments in functions φ_1, φ_2, φ_3 are different, we can generalize:

$$(\{q\}, \{w\}) \rightarrow Mw_k. \quad (19)$$

In expression (19), *Mw* are metaparameters whose number is equal to *k*.

Once the metaparameters are obtained, the relationships between them are found. This is the second stage of ontological modeling:

$$(Mw_1, Mw_2, \ldots, Mw_k) \rightarrow ReW. \quad (20)$$

In expression (20), *ReW* are the implicit relations between metaparameters, which are not initially identified by parameters *q*, *w* and determined only by metaparameters. New relations *ReW* give a reason to search for and establish new relations:

$$(Mw_1, Mw_2, \ldots, Mw_k) \rightarrow ConMw \rightarrow \rightarrow \psi(\varphi_1, \varphi_2, \ldots, \varphi_k). \quad (21)$$

In expression (21) *ConMw* are previously unknown relations, while φ_1, φ_2, …, φ_k are metaparameter functions and ψ is the ontological function.

Expression (21) describes a new dependence. This dependence is implicit before ontological modeling and is revealed only at its third stage.

3. DISCUSSION

Since ontological modeling is performed on specific objects or models, it requires related or related models. So far, such a concept has not been applied in the theory of ontological modeling. At the same time, it is a prerequisite for ontological modeling. Ontological analysis of models that are not related in any way will not give a reliable result. However, ontological analysis and ontological modeling of models related by internal properties can leads to the identification of new patterns and acquisition of new

knowledge. Ontological information modeling on related models is one of the methods for extracting tacit knowledge [13].

An important feature of ontological information modeling is the influence of cognitive factors on the modeling result. Cognitive modeling is required at the stage of forming a query to search for related models. This fact is also poorly taken into account in ontology theory. The disadvantage of the cognitive approach is that cognitive factors create ambiguity of search query formation, leading to ambiguity of related model formation.

Ontological information modeling, which uses the IS model [14, 15], is itself an IS technology. The information field creates an integral model of reality with all internal connections and relations, permitting their identification using ontological information modeling. The main advantage of IS is that it contains all internal connections and relations, which increases the adequacy of ontological modeling.

Nowadays, modeling—and especially ontological modeling—is affected by the problem of big data. In ontological modeling, it is necessary to carry out clustering using big data [16]. In addition, the task of data mining arises considering the volume of data [17]. Special methods are needed for this purpose. Therefore, modern ontological modeling methods must additionally include big data processing algorithms.

CONCLUSIONS

Ontological modeling can only be performed on models that have internal commonality and internal relationships. However, since ontological modeling and information retrieval are related, they can be considered as a single composite technology. In such a composite technology, information retrieval is a necessary preliminary stage, serving to select related models that form the basis for subsequent ontological analysis. Ontological modeling carried out as part of this composite technology identifies regularities and functional dependencies. As such, it represents a method for obtaining new knowledge. While researching the present work, the concepts of regularity and functional dependence were clarified to establish the presence of qualitative-quantitative transitions between them. Regularity is expressed with the help of logical descriptions. The relation of regularity is succession, while the main relation of functional dependence is equivalence. Regularity gives qualitative descriptions and qualitative evaluations, while functional dependence enables quantitative assessment of internal relations.

In this work, to obtain related models, we have proposed the technology of information inventory in the extended sense of group search. Cluster analysis, which can be used as an alternative approach, will be the subject of another paper.

Authors' contribution. All authors equally contributed to the research work.

REFERENCES

1. Gigi M., Tzfadia E. Frontieriphery: An anti-positivist ontological approach to intersectional investigation. *Ethnopolitics*. 2023;23(4):1–17. http://doi.org/10.1080/17449057.2023.2176586
2. Bader S., Pullmann J., Mader C., et al. The international data spaces information model–an ontology for sovereign exchange of digital content. In: Pan J.Z., et al. *The Semantic Web – ISWC 2020. ISWC 2020. Series: Lecture Notes in Computer Science.* Springer; 2020. V. 12507. P. 176–192. https://doi.org/10.1007/978-3-030-62466-8_12
3. Lin J., Ma X., Lin S.C., et al. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021. P. 2356–2362. https://doi.org/10.1145/3404835.3463238
4. Kudzh S.A. *Informacionnoe pole* (*Information Field*). Moscow: MAKS Press; 2017. 97 p. (in Russ.). ISBN 978-5-317-05530-1
5. Bolbakov R.G., Sinitsyn A.V., Tsvetkov V.Ya. Onomasiological modeling in the information field. *J. Phys.: Conf. Ser. The Third International Conference on Metrological Support of Innovative Technologies* (*ICMSIT-III-2022*). 2022;2373(2):2201. http://doi.org/10.1088/1742-6596/2373/2/022010
6. Sánchez-Zas C., Villagra V., Vega-Barbas M., et al. Ontology-based approach to real-time risk management and cyber-situational awareness. *Future Gener. Comput. Syst*. 2023;141(2):462–472. https://doi.org/10.1016/j.future.2022.12.006
7. Milton S., Kazmierczak E., Thomas L. Ontological foundations of data modeling in information systems. In: *AMCIS 2000 Proceedings*. 2000. P. 292. Available from URL: https://aisel.aisnet.org/amcis2000/292
8. Lu W., Xiong N., Park D.S. An ontological approach to support legal information modeling. *J. Supercomput*. 2012;62:53–67. https://doi.org/10.1007/s11227-011-0647-8
9. Lee Y.C., Eastman C.M., Solihin W. An ontology-based approach for developing data exchange requirements and model views of building information modeling. *Adv. Eng. Informatics.* 2016;30(3):354–367. https://doi.org/10.1016/j.aei.2016.04.008
10. Karshenas S., Niknam M. Ontology-based building information modeling. *Comput. Civil Eng.* 2013;2013:476–483. https://doi.org/10.1061/9780784413029.060

11. Sigov A.S., Tsvetkov V.Ya., Rogov I.E. Method for assessing testing difficulty in educational sphere. *Russian Technological Journal.* 2021;9(6):64−72. https://doi.org/10.32362/2500-316X-2021-9-6-64-72
12. Kogalovsky M.R., Kalinichenko L.A. Conceptual and ontological modeling in information systems. *Program. Comput. Soft*. 2009;35:241–256. https://doi.org/10.1134/S0361768809050016
13. Sigov A.S., Tsvetkov V.Ya. Tacit knowledge: Oppositional logical analysis and typologization. *Her. Russ. Acad. Sci.* 2015;85(5):429–433. https://doi.org/10.1134/S1019331615040073
[Original Russian Text: Sigov A.S., Tsvetkov V.Ya. Tacit knowledge: Oppositional logical analysis and typologization. *Vestnik Rossiiskoi Akademii Nauk*. 2015;85(9):800–804 (in Russ.). https://doi.org/10.7868/S0869587315080319]
14. Ostrom T.M., Pryor J.B., Simpson D.D. The organization of social information. In: *Social Cognition*. Routledge; 2022. P. 3–38.
15. Tsvetkov V.Ya., Romanchenko A., Tkachenko D., et al. The Information Field as an Integral Model. In: Silhavy R., Silhavy P. (Eds.). *Software Engineering Research in System Science. CSOC 2023. Series: Lecture Notes in Networks and Systems.* Springer. 2023;722:174–183. https://doi.org/10.1007/978-3-031-35311-6_19
16. Ikotun A.M., Ezugwu A.E., Abualigah L., et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci*. 2023;622(11):178–210. https://doi.org/10.1016/j.ins.2022.11.139
17. Thayyib P.V., Mamilla R., Khan M., et al. State-of-the-art of artificial intelligence and big data analytics reviews in five different domains: a bibliometric summary. *Sustainability*. 2023;15(5):4026. https://doi.org/10.3390/su15054026

СПИСОК ЛИТЕРАТУРЫ

1. Gigi M., Tzfadia E. Frontieriphery: An anti-positivist ontological approach to intersectional investigation. *Ethnopolitics*. 2023;23(4):1–17. http://doi.org/10.1080/17449057.2023.2176586
2. Bader S., Pullmann J., Mader C., et al. The international data spaces information model–an ontology for sovereign exchange of digital content. In: Pan J.Z., et al. *The Semantic Web – ISWC 2020. ISWC 2020. Series: Lecture Notes in Computer Science.* Springer; 2020. V. 12507. P. 176–192. https://doi.org/10.1007/978-3-030-62466-8_12
3. Lin J., Ma X., Lin S.C., et al. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021. P. 2356–2362. https://doi.org/10.1145/3404835.3463238
4. Кудж С.А. *Информационное поле*. М.: МАКС Пресс; 2017. 97 с. ISBN 978-5-317-05530-1
5. Bolbakov R.G., Sinitsyn A.V., Tsvetkov V.Ya. Onomasiological modeling in the information field. *J. Phys.: Conf. Ser. The Third International Conference on Metrological Support of Innovative Technologies* (*ICMSIT-III-2022*). 2022;2373(2):2201. http://doi.org/10.1088/1742-6596/2373/2/022010
6. Sánchez-Zas C., Villagra V., Vega-Barbas M., et al. Ontology-based approach to real-time risk management and cyber-situational awareness. *Future Gener. Comput. Syst*. 2023;141(2):462–472. https://doi.org/10.1016/j.future.2022.12.006
7. Milton S., Kazmierczak E., Thomas L. Ontological foundations of data modeling in information systems. In: *AMCIS 2000 Proceedings*. 2000. P. 292. URL: https://aisel.aisnet.org/amcis2000/292
8. Lu W., Xiong N., Park D.S. An ontological approach to support legal information modeling. *J. Supercomput*. 2012;62:53–67. https://doi.org/10.1007/s11227-011-0647-8
9. Lee Y.C., Eastman C.M., Solihin W. An ontology-based approach for developing data exchange requirements and model views of building information modeling. *Adv. Eng. Informatics.* 2016;30(3):354–367. https://doi.org/10.1016/j.aei.2016.04.008
10. Karshenas S., Niknam M. Ontology-based building information modeling. *Comput. Civil Eng.* 2013;2013:476–483. https://doi.org/10.1061/9780784413029.060
11. Сигов А.С., Цветков В.Я., Рогов И.Е. Методы оценки сложности тестирования в сфере образования. *Russian Technological Journal*. 2021;9(6):64–72. https://doi.org/10.32362/2500-316X-2021-9-6-64-72
12. Kogalovsky M.R., Kalinichenko L.A. Conceptual and ontological modeling in information systems. *Program. Comput. Soft*. 2009;35:241–256. https://doi.org/10.1134/S0361768809050016
13. Сигов А.С., Цветков В.Я. Неявное знание: оппозиционный логический анализ и типологизация. *Вестник Российской Академии Наук*. 2015;85(9):800–804. https://doi.org/10.7868/S0869587315080319
14. Ostrom T.M., Pryor J.B., Simpson D.D. The organization of social information. In: *Social Cognition*. Routledge; 2022. P. 3–38.
15. Tsvetkov V.Ya., Romanchenko A., Tkachenko D., et al. The Information Field as an Integral Model. In: Silhavy R., Silhavy P. (Eds.). *Software Engineering Research in System Science. CSOC 2023. Series: Lecture Notes in Networks and Systems.* Springer. 2023;722:174–183. https://doi.org/10.1007/978-3-031-35311-6_19
16. Ikotun A.M., Ezugwu A.E., Abualigah L., et al. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci*. 2023;622(11):178–210. https://doi.org/10.1016/j.ins.2022.11.139
17. Thayyib P.V., Mamilla R., Khan M., et al. State-of-the-art of artificial intelligence and big data analytics reviews in five different domains: a bibliometric summary. *Sustainability*. 2023;15(5):4026. https://doi.org/10.3390/su15054026

About the authors

Viktor Ya. Tsvetkov, Dr. Sci. (Eng.), Dr. Sci. (Econ.), Professor, Department of Instrumental and Applied Software, Institute of Information Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). Laureate of the Prize of the President of the Russian Federation, Laureate of the Prize of the Government of the Russian Federation, Academician at the Russian Academy of Education Informatization, Academician at the K.E. Tsiolkovsky Russian Academy of Cosmonautics. E-mail: cvj2@mail.ru. Scopus Author ID 56412459400, ResearcherID J-5446-2013. RSCI SPIN-code 3430-2415, http://orcid.org/0000-0003-1359-9799

Nikita S. Kurdyukov, Postgraduate Student, Department of Instrumental and Applied Software, Institute of Information Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: nskurdyukov@gmail.com. RSCI SPIN-code 8535-1612, https://orcid.org/0000-0001-6784-3369

Об авторах

Цветков Виктор Яковлевич, д.т.н., д.э.н., профессор, профессор кафедры инструментального и прикладного программного обеспечения, Институт информационных технологий, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). Лауреат Премии Президента РФ, Лауреат Премии Правительства РФ, академик Российской академии информатизации образования, академик Российской академии космонавтики им. К.Э. Циолковского. E-mail: cvj2@mail.ru. Scopus Author ID 56412459400, ResearcherID J-5446-2013, SPIN-код РИНЦ 3430-2415, http://orcid.org/0000-0003-1359-9799

Курдюков Никита Сергеевич, аспирант, кафедра инструментального и прикладного программного обеспечения, Институт информационных технологий, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: nskurdyukov@gmail.com. SPIN-код РИНЦ 8535-1612, https://orcid.org/0000-0001-6784-3369

Translated from Russian into English by L. Bychkova
Edited for English language and spelling by Thomas A. Beavitt

Information systems. Computer sciences. Issues of information security
Информационные системы. Информатика. Проблемы информационной безопасности

UDC 004.6
https://doi.org/10.32362/2500-316X-2025-13-2-27-35
EDN SRKXBR

RESEARCH ARTICLE

Logical integration of information systems based on expert systems

Evgeniy S. Shevtsov,
Roman V. Shamin @

MIREA – Russian Technological University, Moscow, 119454 Russia
@ *Corresponding author, e-mail: roman@shamin.ru*

Abstract

Objectives. The study set out to develop fundamental methodological principles for the logical integration of information systems (IS) in organizations and to quantitatively assess the topological significance of the IS integration process.

Methods. Methods based on expert systems were used for the logical integration of information in conjunction with data-mining approaches based on various IS. In order to quantitatively assess the topological significance of the IS integration procedure, graph theory methods were used. Discrete topology methods were also employed for calculating the topological invariants of the IS interconnection topology.

Results. Issues and challenges involved in the integration of IS in large organizations are considered in terms of integration methods based on physical and logical principles. While IS integration approaches based on logical principles offer distinct advantages over physical integration approaches, new problems arising in the context of logical integration approaches require innovative solutions. The proposed scheme for the logical integration of IS includes an algebraic method for quantitatively assessing the topological significance of integration, comprising an important numerical indicator in the logical integration of IS. Methods based on learning expert systems, which represent a fundamental solution for organizing the logical integration of IS for intelligent data analysis, are reviewed.

Conclusions. When integrating IS in organizations, it is advisable to use a logical integration approach that preserves the logic of existing information systems. The application of logical integration enables intelligent data analysis using various IS. The use of expert systems in logical integration enables the creation of a new logical layer for providing decision support within the organization.

Keywords: information systems, systems integration, expert systems, data mining, information systems topology

• **Submitted:** 11.03.2024 • **Revised:** 05.07.2024 • **Accepted:** 31.01.2025

For citation: Shevtsov E.S., Shamin R.V. Logical integration of information systems based on expert systems. *Russian Technological Journal.* 2025;13(2):27–35. https://doi.org/10.32362/2500-316X-2025-13-2-27-35, https://elibrary.ru/SRKXBR

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Логическая интеграция информационных систем на основе экспертных систем

Е.С. Шевцов,
Р.В. Шамин @

МИРЭА – Российский технологический университет, Москва, 119454 Россия
@ *Автор для переписки, e-mail: roman@shamin.ru*

Резюме

Цели. Целью статьи является разработка принципиальных основ для методов логической интеграции информационных систем (ИС) в организациях, а также получение количественной оценки топологической значимости процесса интеграции ИС.

Методы. Использованы методы экспертных систем для логической интеграции информации, а также методы интеллектуального анализа данных из различных ИС. Для количественной оценки топологической значимости процедуры интеграции ИС используются методы теории графов, а для вычисления топологических инвариантов топологии взаимной связи ИС – методы дискретной топологии.

Результаты. Рассмотрены вопросы и проблемы интеграции ИС в крупных организациях, а также методы интеграции ИС, основанные на физическом и логическом принципах. Показаны сложности, которые возникают при физической интеграции ИС, и преимущества их интеграции на основе логических принципов. Установлено, что логическая интеграция обладает рядом важных достоинств, но при этом возникают новые проблемы, которые необходимо решать. Предложены схема логической интеграции ИС и алгебраический метод количественной оценки топологической значимости интеграции – важного числового показателя при логической интеграции ИС. Рассмотрены методы обучающихся экспертных систем для интеллектуального анализа данных. Использование экспертных систем является принципиальным решением для организации логической интеграции ИС.

Выводы. При интеграции ИС в организациях целесообразно использовать логическую интеграцию, сохраняющую логику отдельных ИС. Применение логической интеграции позволяет проводить интеллектуальный анализ данных, используя различные ИС. Использование экспертных систем при логической интеграции дает возможность создать новый логический слой для осуществления поддержки принятия решений в организации.

Ключевые слова: информационные системы, интеграция систем, экспертные системы, интеллектуальный анализ данных, топология информационных систем

• **Поступила:** 11.03.2024 • **Доработана:** 05.07.2024 • **Принята к опубликованию:** 31.01.2025

Для цитирования: Шевцов Е.С., Шамин Р.В. Логическая интеграция информационных систем на основе экспертных систем. *Russian Technological Journal*. 2025;13(2):27–35. https://doi.org/10.32362/2500-316X-2025-13-2-27-35, https://elibrary.ru/SRKXBR

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

The ongoing digitalization of the economy and business processes involves the increasing use of various information systems (IS) for decision-making support in which basic information about the organization's activities is stored and processed [1–3]. While different IS have distinct purposes and may be based on different information technologies, the important task of integrating existing IS arises due to their extensive interrelation. However, challenges involved in the full-scale integration of large IS involve complex technical, financial, and organizational issues [4–8]. In addition, some IS cannot be integrated due to the potential violation of information security rules.

The present work proposes a logical approach to IS integration based on the creation of an intelligent information environment. Logical integration of IS, which offers a number of advantages over physical integration approaches [9], involves the use of various data mining methods to reveal additional (hidden) information. The use of learning expert systems to create a unified information field in the organization represents a new logical level of IS integration.

1. IS INTEGRATION METHOD

Many organizations simultaneously maintain several IS, which are connected by unified information flows. A characteristic feature of these IS consists in their mutual intersection by various objects, whose corresponding information must be stored and processed. Such objects may include employees, material objects, customers, etc. Since, for objective reasons, such IS are created at different times and using different technologies, a number of problems arise related to the integrity of information, its reliability, as well as issues related to the potential violation of information security rules [10]. The task of integrating different IS into a single information platform methods can approached based on either of the following principles:

1. Physical integration.
2. Integration based on business logic.

The physical integration of more than one IS implies the creation of an IS that fulfills all the functions of the merged systems. In this case, it is necessary to refactor the structure of databases and the corresponding logic of all software. While this process can provide a full-fledged integration of IS, it is generally very labor-intensive, in some cases comparable to the creation of a new IS [11].

IS integration carried out on the basis of business logic, which can thus be referred to as logical integration of IS, is understood not only in terms of the integration of databases, but also the creation of a single logic of combining information across different IS (Fig. 1) [12, 13].

IS integration sets out provide a unified logic of the merged IS without significant changes in the IS architecture [14, 15]. The advantages of logical integration of IS over physical integration approaches can include the following factors:

1. Lower cost.
2. Preserving diversity.
3. Technological heterogeneity.
4. Ability to process data at a higher level.

The use of a common logic in IS integration is based on the use of special protocols for mutual communication between existing IS. The development of such protocols should be based on a specially developed formal language [16].

2. SYSTEM INTERACTION DURING LOGICAL INTEGRATION

During the logical integration process, it becomes necessary to provide mechanisms of interaction between the merged IS. Such interaction can be described using a formal finite-automata language. The task of the interaction mechanism is to provide mappings of objects in one IS into the objects of another IS. However, the main difficulty that arises in this connection is that the mapped objects may have information overlap. For example, there may be information related to employees of an organization in different IS, but this information may be represented in different ways in different systems. An addition problem may arise due to the different scales used when describing the same objects in different information bases.

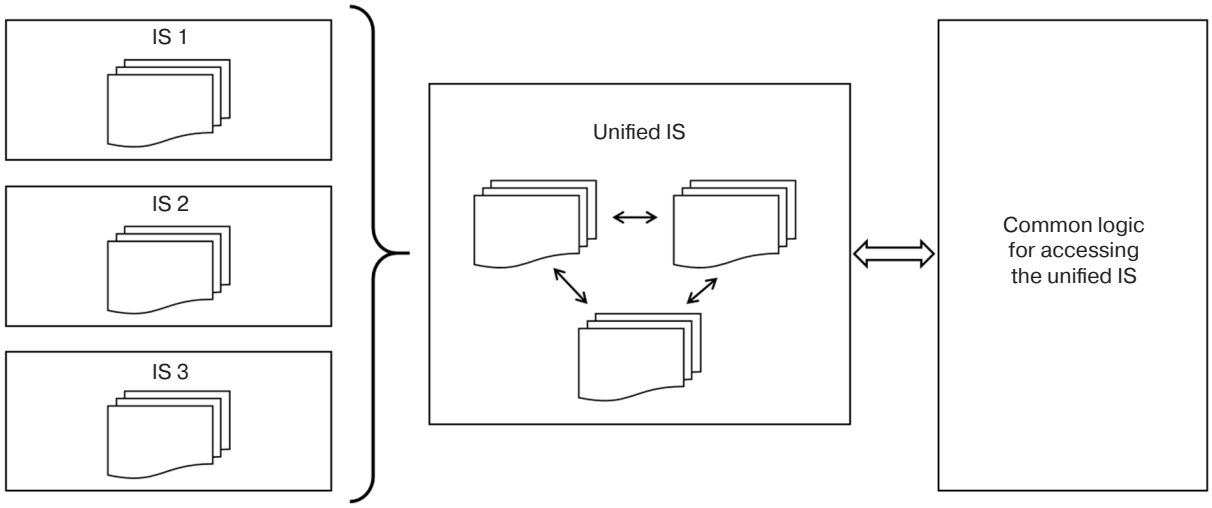


Fig. 1. Logical integration of IS

Thus, when designing and implementing an IS integration procedure, one of the main tasks is to create a mechanism for logical mapping of the different IS. Figure 2 shows an object-mapping scheme for application during logical integration.

For the logical display of the various objects in the different IS, it is advisable to use an intelligent environment for integration of information objects. This environment, which uses semantic information-processing methods, can be used to realize the mechanism of displaying objects in the IS.

3. CHANGING AN ORGANIZATION'S INTERCONNECTIVITY TOPOLOGY

When integrating different IS in an organization, an important issue arises in terms of changing the topology of mutual connection of these systems. The main point here is that, in order to obtain qualitative changes during IS integration, it becomes necessary to provide changes in the topology of their interconnection.

The topology of IS interconnection is described by an undirected graph, whose vertices are individual IS, while the edges represent information links between systems [17]. Since it is only the presence of a connection that is important for the topology of the IS network, i.e., irrespective of the direction of flow, we will consider undirected graphs. Here, the salient point consists in the fact that that even unidirectional flows include not only the data flow, but also the corresponding request for this data.

Figure 3 shows an example of the IS communication topology.

In this example, IS numbered 1–8 are linked by information links, while IS number 9 is not linked to the other IS. Two IS will be referred to as incident if there is an information link between them. We will define two IS S and P as connected if a chain of sequentially connected

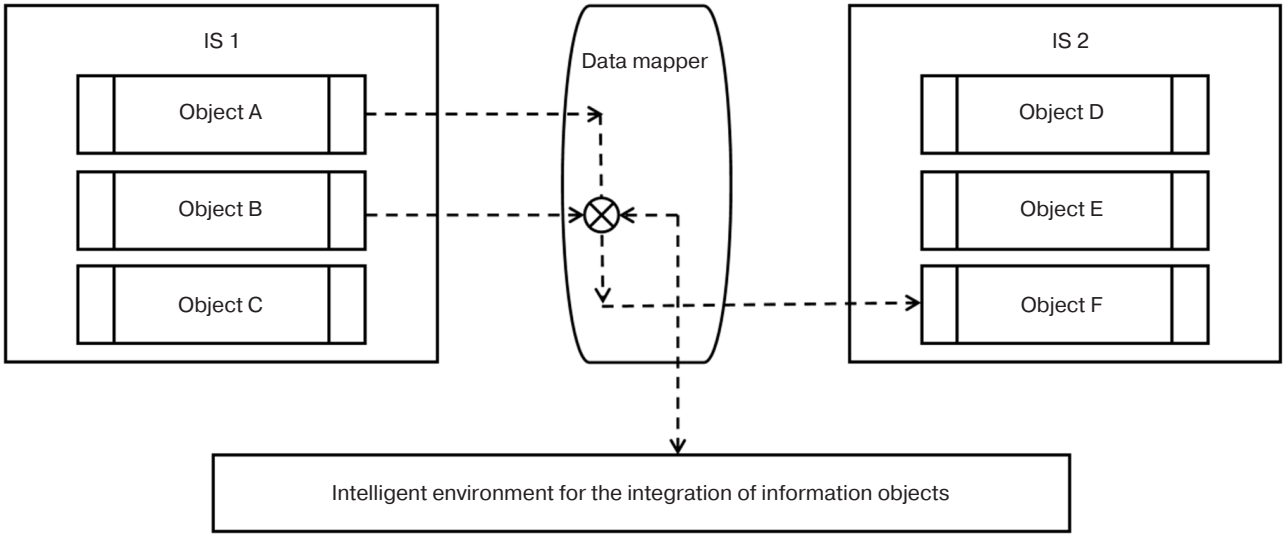


Fig. 2. Diagram of the logical object display mechanism

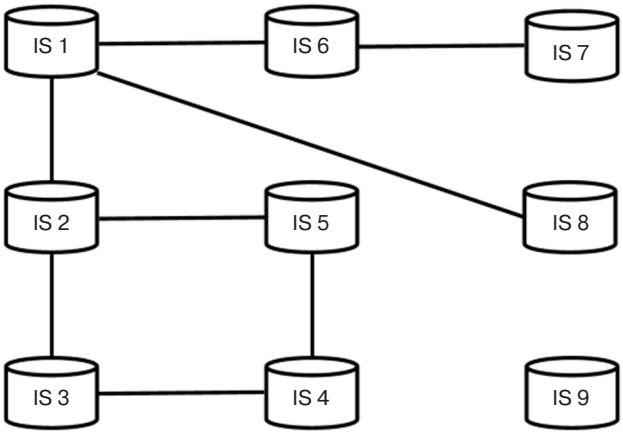


Fig. 3. Example of the IS communication topology

IS can be constructed from S to P. The entire IS network can be represented by connectivity components, where each connectivity component represents a set of individual IS that are connected in pairs.

We will consider the process of IS integration as a sequential operation of merging neighboring nodes (incident IS). In this case, the unified IS inherits all the information links of the united vertices.

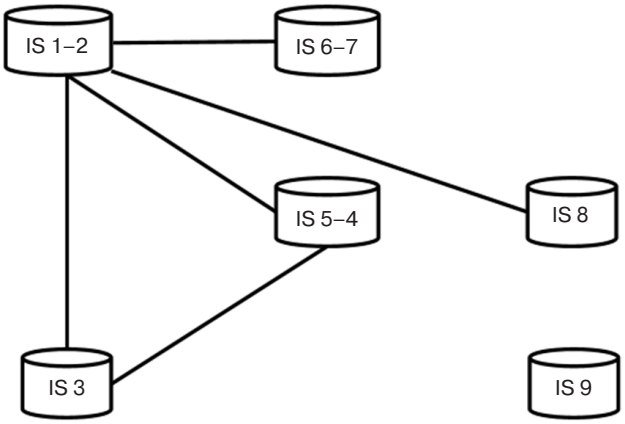


Fig. 4. Topology following IS unification

Figure 4 shows the IS topology following IS unification:

$$[1, 2] \rightarrow [1–2],$$
$$[4, 5] \rightarrow [5–4],$$
$$[6, 7] \rightarrow [6–7].$$

In order to distinguish significant changes in the IS communication topology, we will use topological invariants. As an example of such a variant, let us consider the fundamental group for the graph describing the IS communication topology. We will define the fundamental group as the set of equivalence classes of homotopy loops in the graph [18]. For the connected component of the IS network, the fundamental group defines the number of loops. If a connected graph representing the IS communication topology has N cycles, then the fundamental group is isomorphic to the $\mathbb{Z}^N$ group [19].

The presence of cycles in the organization's IS network indicates the need to integrate IS, since the presence of cycles in the network of information links implies risks of ambiguity in the presentation of information to the organization during information requests due to the possibility of ambiguous ways of transferring information between different IS.

While no new cycles can arise when integrating IS, existing cycles can be opened. In the language of the fundamental groups of the IS communication graph, this means that the following change in the representation of the fundamental groups occurs during the process of IS integration:

$$\mathbb{Z}^N \rightarrow \mathbb{Z}^{N-k}.$$

In this interpretation, we can define the topological significance of the IS integration procedure as the number k by which the degree of the fundamental group decreases.

4. USE OF EXPERT SYSTEMS FOR IS LOGICAL INTEGRATION

While the logical integration of IS in an organization may have various objectives, the main aim is to create a unified information field. This problem cannot be solved using "mechanical" methods since samples from different IS must be brought to a "common denominator" in order to obtain uniform information. Another challenge that arises in this connection consists in the need to obtain additional information about the organization's activities based on the heterogeneous information in the different IS.

In order to solve these problems inherent in the logical integration of IS, it is proposed to use trainable expert systems and knowledge bases. The objects described in the IS are used as the subject area of the knowledge base.

Figure 5 presents a scheme for the application of an expert system in logical integration of IS. The key element of the proposed scheme consists in the use of the IS query router and the intellectual environment of the integrated IS. The logical integration of IS implies an intellectual environment since it provides a means to obtain information from different sources on the incoming request for further logical integration of heterogeneous information. In order to determine which IS should form the basis for information queries, a query router is used to identify the most appropriate data sources based on the display of information in different IS.

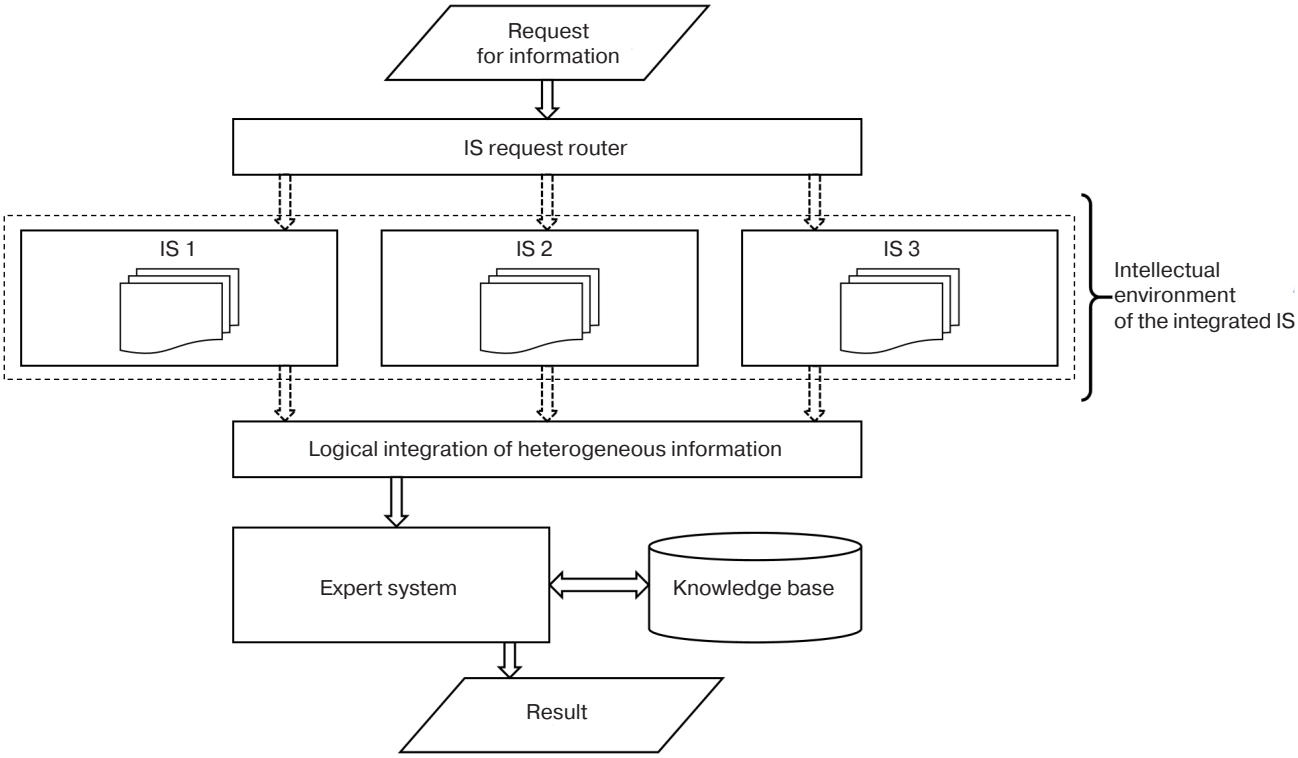


Fig. 5. Diagram of the expert system application

Following the realization of logical integration of information from different IS, a meaningful query is formed to the expert system, which forms the result according to the received query using the knowledge base [20, 21].

The architecture of an expert system depends on the nature of objects described by the IS, as well as on the completeness of information for each object [22]. The general scheme of the architecture of the expert system and knowledge base is presented in Fig. 6.

The architecture of the expert system includes a mechanism for performance-based learning as a means of improving the data-mining procedure on integrated data.

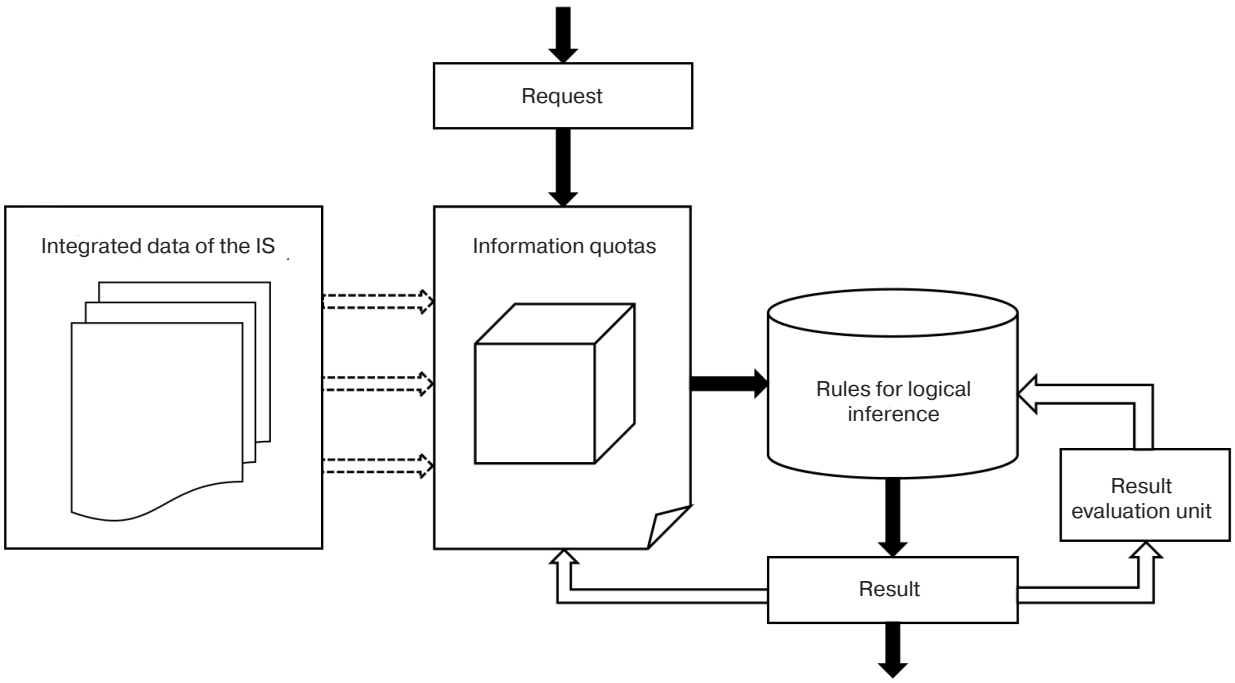


Fig. 6. Architecture of a learning expert system

In Fig. 6, solid arrows depict the sequential process of calculating and obtaining the result of the expert system operation, while data transferred to implement the training procedure of the expert system are represented by contour arrows. Dashed arrows show the data flow for the expert system from the integrated IS.

Within this scheme, a fundamental principle concerns the use of information quanta to represent integrated information from IS. The specific realization of the procedure of data representation in the form of information quanta depends on the nature of objects in IS and the structure of their interaction.

In order to implement the training process of the expert system, a result-evaluation block is used. This block can either be implemented using feedback from the user of the expert system or based on the evaluation by the artificial intelligence system. In cases where an artificial intelligence method is used (neural network, Bayesian networks, decision trees, etc.), training can be realized on the basis of machine learning methods with suitable reinforcement [23].

CONCLUSIONS

The present work considers fundamental issues of IS integration in organizations on the basis of intellectual methods. The introduced concept of logical IS integration is suitable for integrating heterogeneous IS. Issues connected with the topological significance of the logical IS integration procedure have also been considered.

The logical integration of IS and creation of a unified information environment involves the use of an expert system, which responds to queries associated with the operation of integrated information flows. Such an expert system allows a basis for intellectual analysis of the data to be provided.

The proposed architecture of expert systems includes mechanisms for training (self-learning) of the knowledge base of the expert system, by means of which the results of IS integration in the organization can be further improved.

Authors' contributions

E.S. Shevtsov—conceptual model of integration of information systems based on the creation of an intelligent information environment.

R.V. Shamin—mathematical model of integration of information systems based on the creation of an intelligent information environment.

REFERENCES

1. Spiridonov E.S., Klykov M.S., Rukin M.D., Grigoriev N.P., Balalaeva T.I., Smurov A.V. *Informatsionnaya ekonomika* (*Information Economy*). Moscow: URSS; 2021. 286 p. (in Russ.).
2. Kalyanov G.N., Lukinova O.V., Levochkina G.A., Vasilev R.B. *Strategicheskoe upravlenie informatsionnymi sistemami* (*Strategic Management of Information Systems*). Moscow: Prosveshchenie/Binom; 2019. 510 p. (in Russ.).
3. Prokhorov A., Konik L. *Tsifrovaya transformatsiya. Analiz, trendy, mirovoi opyt* (*Digital Transformation. Analysis, Trends, World Experience*). Moscow: KomN'yus Grup; 2019. 368 p. (in Russ.).
4. Loiko V.I., Lutsenko E.V., Orlov A.I. *Sovremennaya tsifrovaya ekonomika* (*Modern Digital Economy*). Krasnodar: KubSAU; 2018. 508 p. (in Russ.).
5. Chursin A.A., Yudin A.V. *Kiberekonomika v praktike: sozdanie radikal'no novoi produktsii v tsifrovuyu epokhu* (*Cybereconomics in Practice: Creating Radically New Products in the Digital Era*). Moscow: Ekonomika; 2021. 301 p. (in Russ.).
6. Belalova G.A. Analysis of information systems integration methods. *Tsifrovye modeli i resheniya = Digital Models and Solutions*. 2023;2(3):61–68 (in Russ.).
7. Karev A.N., Fedosin S.A. Ontological approach to information systems integration. *Perspektivy nauki = Science Prospects*. 2023;168(9):26–29 (in Russ.).
8. Belyaev A.K., Kritskaya S.N. Integration of information systems in action. *Informatsionnye tekhnologii v UIS = Information Technologies in the Penal System.* 2022;1:34–40 (in Russ.).
9. Weber R.H., Burri M. *Classification of Services in the Digital Economy*. Berlin, Heidelberg: Springer; 2012. 144 p.
10. Babash A.V., Baranova E.K. *Aktual'nye voprosy zashchity informatsii* (*Current Issues of Information Security*). Moscow: INFRA-M, RIOR Nauka; 2023. 111 p. (in Russ.). ISBN 978-5-36901-680-0
11. Martin R. *Chistaya arkhitektura. Iskusstvo razrabotki programmnogo obespecheniya* (*Clean Architecture. The Art of Software Development*). St. Petersburg: Piter; 2022. 352 p. (in Russ.).
12. Chernyak L. Data integration: syntax and semantics. *Otkrytye sistemy. SUBD = Open Systems. DBMS.* 2009;10 (in Russ.). Available from URL: https://www.osp.ru/os/2009/10/11170978
13. Antamoshin A.N., Bliznova O.V., Bobov A.V., Bolshakov A.A., Lobanov V.V., Kuznetsova I.N. *Intellektual'nye sistemy upravleniya organizatsionno-tekhnicheskimi sistemami* (*Intelligent Control Systems for Organizational and Technical Systems*). Moscow: Goryachaya liniya – Telekom; 2006. 160 p. (in Russ.).
14. Norenkov I.P. *Avtomatizirovannye informatsionnye sistemy* (*Automated Information Systems*). Moscow: Bauman Press; 2011. 344 p. (in Russ.).

15. Mazepa R.B., Mikhailov V.Yu. *Osnovy informatsionnykh tekhnologii. Vvedenie v protsessy informatsionnogo vzaimodeistviya* (*Fundamentals of Information Technology. Introduction to Information Interaction Processes*). Moscow: Vuzovskaya kniga; 2012. 60 p. (in Russ.).
16. Magazov S.S. *Teoriya formal'nykh yazykov. Regulyarnye yazyki* (*Theory of Formal Languages. Regular Languages*): eBook. Moscow: Bauman Press; 2023. 52 p. (in Russ.).
17. Harary F., Palmer E. *Perechislenie grafov* (*Graphical Enumeration*): transl. from Engl. Moscow: Mir; 1977. 328 p. (in Russ.). [Harary F., Palmer E. *Graphical Enumeration*. New York, London: Academic Press; 1973. 271 p.]
18. Fomenko A.T., Fuks D.B. *Kurs gomotopicheskoi topologii* (*Course on Homotopic Topology*). Moscow: URSS; 2024. 512 p. (in Russ.).
19. Hatcher A. *Algebraicheskaya topologiya* (*Algebraic Topology*): transl. from Engl. Moscow: MTsNMO; 2011. 688 p. (in Russ.).
[Hatcher A. *Algebraic Topology*. Cambridge University Press; 2005. 556 p.]
20. Giarratano J., Riley G. *Ekspertnye sistemy: printsipy razrabotki i programmirovanie* (*Expert Systems: Principles of Development and Programming*): transl. from Engl. Moscow: Vil'yams; 2007. 1152 p. (in Russ.).
[Giarratano J., Riley G. *Expert Systems: Principles and Programming*. Thomson Course Technology; 2005. 842 p.]
21. Bodrov O.A., Medvedev R.E. *Predmetno-orientirovannye ekonomicheskie informatsionnye sistemy* (*Subject-Oriented Economic Information Systems*). Moscow: Goryachaya liniya – Telekom; 2013. 244 p. (in Russ.).
22. Ruchkin V.N., Fulin V.A. *Universal'nyi iskusstvennyi intellekt i ekspertnye sistemy* (*Universal Artificial Intelligence and Expert Systems*). St. Petersburg: BKhV-Peterburg; 2009. 224 p. (in Russ.).
23. Sutton R.S., Barto A.G. *Reinforcement Learning: An Introduction*. Cambridge, Massachusetts: The MIT Press; 2018. 526 p.

СПИСОК ЛИТЕРАТУРЫ

1. Спиридонов Э.С., Клыков М.С., Рукин М.Д., Григорьев Н.П., Балалаева Т.И., Смуров А.В. *Информационная экономика*. М.: URSS; 2021. 286 с.
2. Калянов Г.Н., Лукинова О.В., Левочкина Г.А., Васильев Р.Б. *Стратегическое управление информационными системами*. М.: Просвещение/Бином; 2019. 510 с.
3. Прохоров А., Коник Л. *Цифровая трансформация. Анализ, тренды, мировой опыт*. М.: ООО «КомНьюс Груп»; 2019. 368 с.
4. Лойко В.И., Луценко Е.В., Орлов А.И. *Современная цифровая экономика*. Краснодар: КубГАУ; 2018. 508 с.
5. Чурсин А.А., Юдин А.В. *Киберэкономика в практике: создание радикально новой продукции в цифровую эпоху*. М.: Экономика; 2021. 301 с.
6. Белалова Г.А. Анализ методов интеграции информационных систем. *Цифровые модели и решения*. 2023;2(3):61–68.
7. Карев А.Н., Федосин С.А. Онтологический подход к интеграции информационных систем. *Перспективы науки*. 2023;168(9):26–29.
8. Беляев А.К., Критская С.Н. Интеграция информационных систем в действии. *Информационные технологии в УИС*. 2022;1:34–40.
9. Weber R.H., Burri M. *Classification of Services in the Digital Economy*. Berlin, Heidelberg: Springer; 2012. 144 p.
10. Бабаш А.В., Баранова Е.К. *Актуальные вопросы защиты информации*. М.: ИНФРА-М, РИОР Наука; 2023. 111 с. ISBN 978-5-36901-680-0
11. Мартин Р. *Чистая архитектура. Искусство разработки программного обеспечения*. СПб.: Питер; 2022. 352 с.
12. Черняк Л. Интеграция данных: синтаксис и семантика. *Открытые системы. СУБД.* 2009;10. URL: https://www.osp.ru/os/2009/10/11170978
13. Антамошин А.Н., Близнова О.В., Бобов А.В., Большаков А.А., Лобанов В.В., Кузнецова И.Н. *Интеллектуальные системы управления организационно-техническими системами*. М.: Горячая линия – Телеком; 2006. 160 с.
14. Норенков И.П. *Автоматизированные информационные системы*. М.: Изд-во МГТУ им. Н.Э. Баумана; 2011. 344 с.
15. Мазепа Р.Б., Михайлов В.Ю. *Основы информационных технологий. Введение в процессы информационного взаимодействия*. М.: Вузовская книга; 2012. 60 с.
16. Магазов С.С. *Теория формальных языков. Регулярные языки*. М.: Изд-во МГТУ им. Н.Э. Баумана; 2023. 52 с.
17. Харари Ф., Палмер Э. *Перечисление графов*: пер. с англ. М.: Мир; 1977. 328 с.
18. Фоменко А.Т., Фукс Д.Б. *Курс гомотопической топологии*. М.: URSS; 2024. 512 с.
19. Хатчер А. *Алгебраическая топология*: пер. с англ. М.: МЦНМО; 2011. 688 с.
20. Джарратано Дж., Райли Г. *Экспертные системы: принципы разработки и программирование*: пер. с англ. М.: ИД Вильямс; 2007. 1152 с.
21. Бодров О.А., Медведев Р.Е. *Предметно-ориентированные экономические информационные системы*. М.: Горячая линия – Телеком; 2013. 244 с.
22. Ручкин В.Н., Фулин В.А. *Универсальный искусственный интеллект и экспертные системы*. СПб.: БХВ-Петербург; 2009. 224 с.
23. Sutton R.S., Barto A.G. *Reinforcement Learning: An Introduction*. Cambridge, Massachusetts: The MIT Press; 2018. 526 p.

About the authors

Evgeniy S. Shevtsov, Postgraduate Student, Department of Artificial Intelligence Technologies, Institute of Artificial Intelligence, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: shevcov@mirea.ru. http://orcid.org/0009-0007-8881-9406

Roman V. Shamin, Dr. Sci. (Phys.-Math.), Professor, Department of Industrial Programming, Institute for Advanced Technologies and Industrial Programming, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: roman@shamin.ru. Scopus Author ID 6506250832, RSCI SPIN-code 8966-0169, https://orcid.org/0000-0002-3198-7501

Об авторах

Шевцов Евгений Сергеевич, аспирант, кафедра технологий искусственного интеллекта, Институт искусственного интеллекта, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: shevcov@mirea.ru. http://orcid.org/0009-0007-8881-9406

Шамин Роман Вячеславович, д.ф.-м.н., профессор, кафедра индустриального программирования, Институт перспективных технологий и индустриального программирования, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: roman@shamin.ru. Scopus Author ID 6506250832, SPIN-код РИНЦ 8966-0169, https://orcid.org/0000-0002-3198-7501

Translated from Russian into English by L. Bychkova
Edited for English language and spelling by Thomas A. Beavitt

Modern radio engineering and telecommunication systems

Современные радиотехнические и телекоммуникационные системы

UDC 621.314.1+681.586.7
https://doi.org/10.32362/2500-316X-2025-13-2-36-45
EDN TIPRXB

RESEARCH ARTICLE

Zeta topology DC/DC converter design based on TPS40200 driver

Vladimir K. Bityukov [1],
Aleksey I. Lavrenov [1, @],
Daniil A. Malitskiy [2]

[1] *MIREA – Russian Technological University, Moscow, 119454 Russia*
[2] *Sputniks, Moscow, 121205 Russia*
[@] *Corresponding author, e-mail: lavrenov@mirea.ru*

Abstract

Objectives. The study set out to investigate typical characteristics of a Zeta converter developed by the authors based on the TPS40200 driver under various input voltages and loads and compare the experimental characteristics of the Zeta converter with those obtained through SPICE[1] simulation in the *Multisim* computer-aided design (CAD) system, as well as with the results derived from a continuous-time mathematical model.

Methods. A continuous-time mathematical model of the Zeta converter and the *Multisim* CAD system were used. The schematic diagram of the converter was developed according to the TPS40200 driver circuit design methodology presented in its datasheet. The printed circuit board layout was created using the *Altium Designer* CAD system.

Results. An experimental test bench of the Zeta topology DC/DC converter was designed and built using coupled chokes based on the TPS40200 driver. The results of the study showed a high correlation of both its load characteristics and its DC and AC components of currents flowing through the choke windings and capacitor voltages from the input voltage at two load resistances of 50 and 100 Ohm obtained by experimental, computational, and modeling methods.

Conclusions. The continuous-time mathematical model of the converter, along with the calculation method based on it, forms a foundation for the design of DC/DC converters using the Zeta topology. The experiment confirms the validity of both the mathematical model and the calculation method. The proposed design methods takes the magnetic coupling and the active resistance of inductors into account. The magnetic coupling permits a two-fold reduction of inductor values while maintaining the same ripple or a reduction in the ripple by up to half with unchanged inductor values.

Keywords: DC/DC converter, Zeta topology, converter, mathematical model, design method, TPS40200, Altium Designer, Multisim, printed circuit board

[1] SPICE (Simulation Program with Integrated Circuit Emphasis) is an open source simulator of general-purpose electronic circuits.

• **Submitted:** 05.06.2024 • **Revised:** 16.08.2024 • **Accepted:** 06.02.2025

For citation: Bityukov V.K., Lavrenov A.I., Malitskiy D.A. Zeta topology DC/DC converter design based on TPS40200 driver. *Russian Technological Journal.* 2025;13(2):36–45. https://doi.org/10.32362/2500-316X-2025-13-2-36-45, https://elibrary.ru/TIPRXB

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Проектирование DC/DC-преобразователя, построенного по Zeta-топологии на базе драйвера TPS40200

В.К. Битюков [1],
А.И. Лавренов [1, @],
Д.А. Малицкий [2]

[1] *МИРЭА – Российский технологический университет, Москва, 119454 Россия*
[2] *ООО «СПУТНИКС», Москва, 121205 Россия*
[@] *Автор для переписки, e-mail: lavrenov@mirea.ru*

Резюме

Цели. Целью работы является исследование типовых характеристик разработанного Zeta-преобразователя на основе драйвера TPS40200 (Texas Instruments, США) при различных входных напряжениях и нагрузках и сравнение экспериментальных характеристик Zeta-преобразователя с аналогичными, полученными при помощи SPICE[1]-моделирования в системе автоматизированного проектирования (САПР) *Multisim*, а также с помощью предельной непрерывной математической модели.

Методы. Использована предельная непрерывная математическая модель Zeta-преобразователя и САПР *Multisim*. Принципиальная электрическая схема преобразователя разработана по методике расчета обвязки драйвера TPS40200, представленной в его технической документации. С использованием САПР *Altium Designer* произведена разводка печатной платы.

Результаты. Спроектирован и создан экспериментальный стенд DC/DC-преобразователя, построенного по Zeta-топологии со связанными дросселями на базе драйвера TPS40200. Результаты исследования показали высокую корреляцию как его нагрузочных характеристик, так и его постоянных и переменных составляющих токов, протекающих через обмотки дросселей, и напряжений на конденсаторах от входного напряжения при двух сопротивлениях нагрузки 50 и 100 Ом, полученных различными методами: экспериментальным, расчетным и моделированием.

Выводы. Предельная непрерывная математическая модель преобразователя и метод расчета, основанный на ней, являются базой для проектирования DC/DC-преобразователей, построенных по топологии Zeta. Экспериментально доказана достоверность математической модели, а также метода проектирования. Предложенный метод проектирования позволяет учесть магнитную связь и активное сопротивление обмоток дросселей. Учет магнитной связи позволяет уменьшить номиналы дросселей до двух раз при неизменных пульсациях либо уменьшить пульсации до двух раз при неизменных номиналах дросселей.

Ключевые слова: DC/DC-преобразователь, топология Zeta, преобразователь, математическая модель, метод проектирования, TPS40200, Altium Designer, Multisim, печатная плата

[1] SPICE (англ. Simulation Program with Integrated Circuit Emphasis) – программа-симулятор электронных схем общего назначения с открытым исходным кодом. [SPICE (Simulation Program with Integrated Circuit Emphasis) is an open source simulator of general-purpose electronic circuits.]

• **Поступила:** 05.06.2024 • **Доработана:** 16.08.2024 • **Принята к опубликованию:** 06.02.2025

Для цитирования: Битюков В.К., Лавренов А.И., Малицкий Д.А. Проектирование DC/DC-преобразователя, построенного по Zeta-топологии на базе драйвера TPS40200. *Russian Technological Journal*. 2025;13(2):36–45. https://doi.org/10.32362/2500-316X-2025-13-2-36-45, https://elibrary.ru/TIPRXB

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

A relevant trend in the development of modern autonomous radio devices involves a reduction in mass-dimensional parameters and accompanying improvement in the tactical and technical characteristics of the power converters used in them [1–3]. Traditionally, choke DC/DC converters of various topologies are used to power such devices [4–6]. In topologies where two chokes are used for energy storage and transmission, it has long been common practice to use coupled chokes to reduce their mass-dimensional parameters and improve basic stabilizer characteristics [7, 8]. Examples of modern devices based on coupled chokes are given in [9, 10]. The design of such converters is typically based on their mathematical models [11–15]. The converter based on the calculation method proposed in [16] was validated by comparing the calculated characteristics with the modeling results rather than via an empirical study. To remedy this deficiency, we set out to experimentally study the DC/DC converter based on the Zeta topology with coupled chokes.

1. SCHEMATIC DIAGRAM OF THE CONVERTER BASED ON THE TPS40200 DRIVER

A TPS40200 microcircuit (Texas Instruments, USA) was chosen as the driver for the Zeta converter for a number of reasons. Firstly, this driver can deliver up to 95% efficiency at various load currents and over a wide range of input and output voltages[1]. Secondly, the chip has a fairly simple design with all the necessary functionality configured via external circuitry. Although this functionality is not declared by the manufacturer, this allows the driver to be used to control Zeta converters. Thirdly, an important factor in choosing this particular driver is its price and availability.

The circuit diagram of the DC/DC converter based on the TPS40200 driver is made up of two functional parts that are calculated separately (Fig. 1). The first part, responsible for the device logic, is the TPS40200 driver and all adjacent elements. The second is the power part of the converter based on Zeta topology, responsible for DC conversion. The nominal values of the elements of the Zeta topology are calculated using the design method [16], which is based on the continuous-time mathematical model of the converter.

A sawtooth signal of the required frequency is formed on the first pin of the RC driver TPS40200 using the frequency setting circuit R1C1. The switching frequency of the VT1 power switch is selected to be 500 kHz. However, the actual switching frequency may vary due to variations in basic parameters of electronic components within technological tolerances.

The part of the circuit responsible for the Zeta converter input current threshold consists of the current sensing resistor R7 and the smoothing filter R6C6, which is necessary to reduce the influence of the high-frequency component occurring when the power switch VT1 is switched. C2 is a start-up capacitor that determines the start-up time of 9 ms. C3C4R2 is a frequency compensation chain whose cut-off frequency is approximately 6–10 times lower than the switching frequency.

In the circuit diagram, the PLS-40 pins labelled “TP” and “P” are intended for the monitoring of currents and voltages, respectively.

2. ARRANGING THE ZETA CONVERTER PCB LAYOUT

The printed circuit board (PCB) layout is arranged in the *Altium Designer* computer-aided design (CAD) system[2], taking into account requirements for easy soldering and convenient multimeter or oscilloscope measurement (Figs. 2 and 3).

[1] TPS40200 Datasheet. TPS40200 Wide Input Range Non-Synchronous Voltage Mode Controller datasheet (Rev. G). Texas Instruments. SLUS659G–FEBRUARY 2006–REVISED NOVEMBER 2014.

[2] https://www.altium.com/altium-designer. Accessed May 24, 2024.

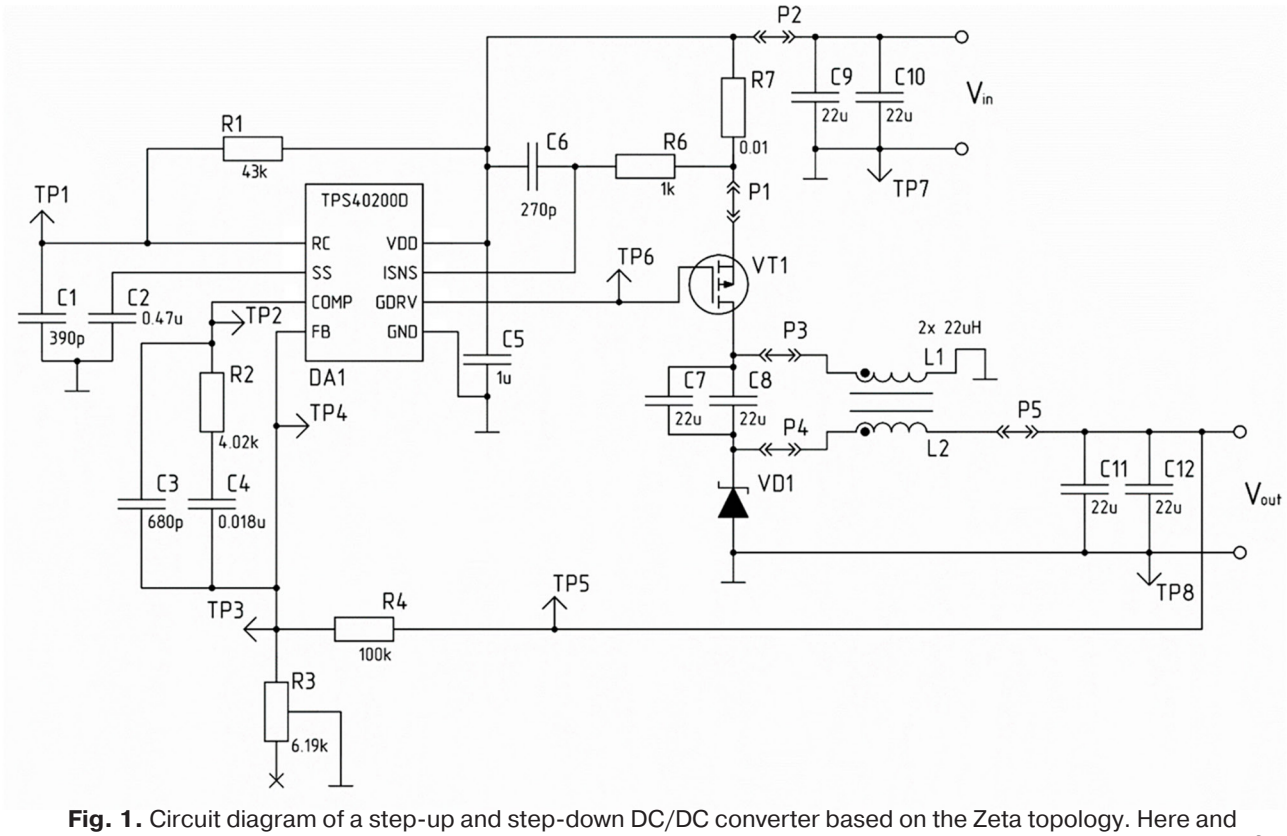


Fig. 1. Circuit diagram of a step-up and step-down DC/DC converter based on the Zeta topology. Here and in the following figures, the designations of the circuit elements correspond to those adopted in GOST 2.710-81[3]

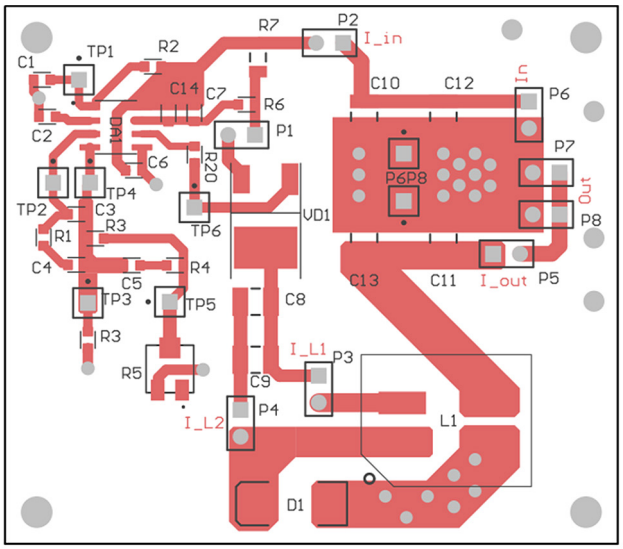


Fig. 2. Upper PCB layer

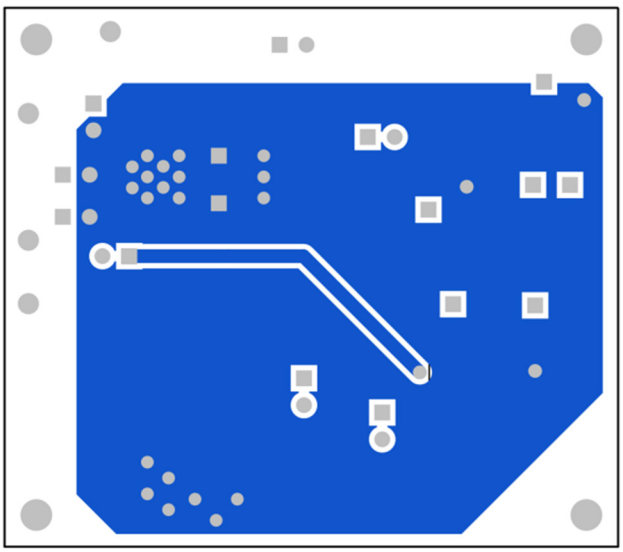


Fig. 3. Lower PCB layer

The VT1 transistor is a WMO25P06T1 p-channel MOSFET (Wayon Electronics Co., China) with a maximum power dissipation of up to 2.5 W and dynamic characteristics that allow operation with a switching frequency of up to 1 MHz.[4] The TPS40200 chip, having a high input voltage of up to 52 V, variable switching frequency of up to 500 kHz, and a gate current of up to 300 mA, is used as the transistor driver. This allows the gate voltage rise time (edge duration) of the transistor to be 0.025–0.040 μs.

[3] GOST 2.710-81. Interstate Standard. *Unified system for design documentation. Alpha-numerical designations in electrical diagrams*. Moscow: Standartinform; 2008 (in Russ.).

[4] WMO25P06T1 Datasheet. 60V P-Channel Enhancement Mode Power MOSFET. Rev. 3.0, 2020. P. 6.

The footprint inductance over the average width (2 mm) of the path from the transistor drain to the positive output pole is 0.3 nH/mm, while the path length is ~60 mm, corresponding to an inductance of 20 nH.

The interlayer capacitance is approximately 60 pF/cm^2. Taking into account the discrete element ratings, the influence of parasitic components on the device parameters is minimal. The considerable thickness of the dielectric has a major influence due to the weakness of the coupling of the signal lines in the upper layer to the lower ground layer, which can lead to significant cross-interference, especially in the feedback circuit.

The operating frequency of the converter is approximately 500 kHz, which is equivalent to a wavelength of 600 m. This is much larger than the PCB size and the length of the conductors.

A general view of the developed PCB with components is shown in Fig. 4.

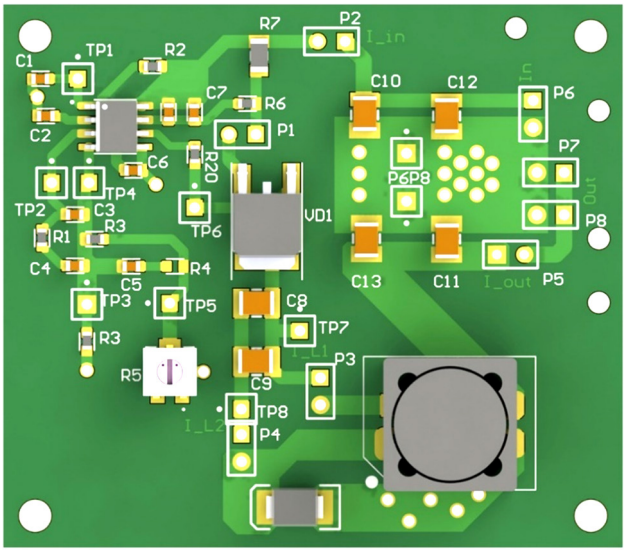


Fig. 4. General view of the developed PCB with components

The substrate material used in the PCB (Rezonit, Russia) is FR4 Tg135 with a relative dielectric constant of 4.3 and a thickness of 1.93 mm. The metallization thickness is 0.035 mm. PCB type: double-sided. PCB dimensions: 58 × 51 mm.

3. EXPERIMENTAL STUDY OF ZETA CONVERTER BASED ON TPS40200 DRIVER

The study was carried out at the Department of Radio Wave Processes and Technologies of the Institute of Radioelectronics and Informatics at RTU MIREA. The test bench shown in Fig. 5 consists of a Zeta converter, a PC, a laboratory power supply, an oscilloscope, a multimeter, a current sensor, and a set of samples/wires for connecting the PCB. The instruments and hardware used for the study, namely, the NGE100 power supply, the RTB2002 oscilloscope, and the HMC8012 universal multimeter, are from Rohde & Schwarz (Germany).[5]

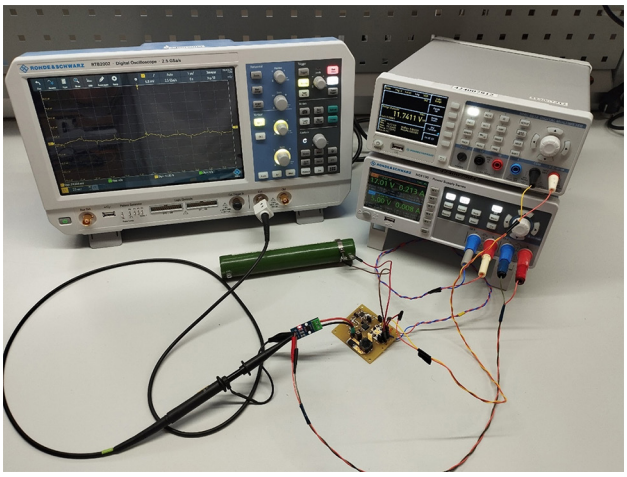


Fig. 5. Test bench for experimental study

The test bench is designed for experimental study of the typical characteristics of a DC/DC converter, with which similar characteristics can be compared as obtained by the design method based on the continuous-time mathematical model of the converter and by the simulation method using *Multisim* CAD.[6]

Typical characteristics of converters traditionally include the load characteristic (LC), which is the dependence of its output voltage U_{out} on the load current I_L at constant input voltage U_{in}, i.e., $U_{out} = f(I_L)$ with U_{in} = const, as well as the dependence of the constant and variable components of currents i_{L1}, i_{L2} and voltages $u_{C7, C8}$, $u_{C11, C12}$ on the input voltage U_{in} at different load resistances.

The converter LCs obtained in the experimental study at input voltages of 6.5, 12.0, and 17.5V are shown in Figs. 6–8.

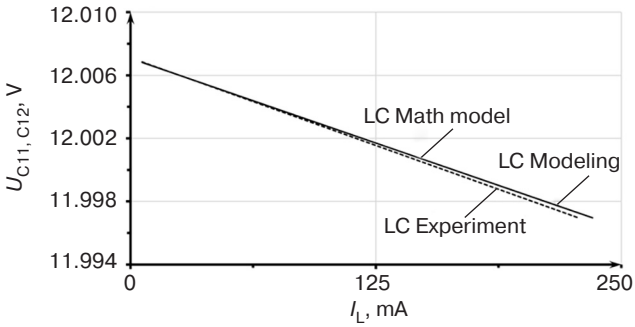


Fig. 6. Load characteristic at an input voltage of 6.5 V

[5] https://www.rohde-schwarz.com/. Accessed July 11, 2024.
[6] https://www.ni.com/ru-ru/shop/product/multisim.html. Accessed February 19, 2024.

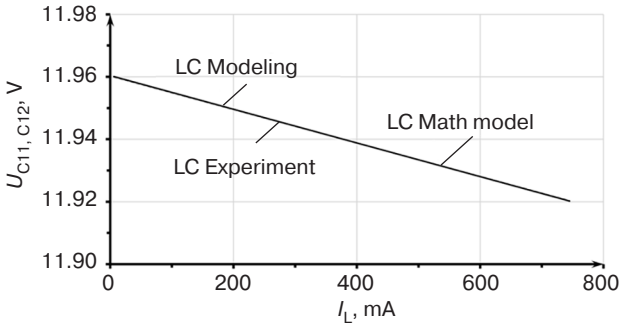


Fig. 7. Load characteristic at an input voltage of 12.0 V

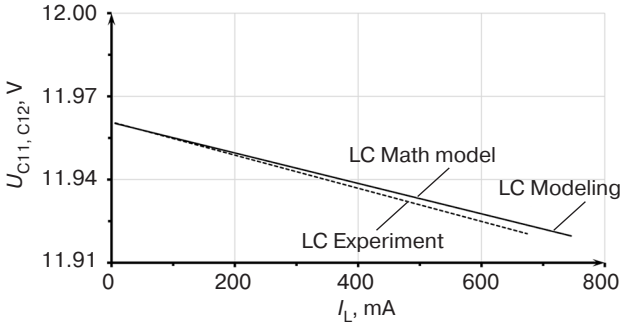


Fig. 8. Load characteristic at an input voltage of 17.5 V

Since the constant charging current of the output capacitor is negligible compared to the load current, the current of the second choke can be assumed to be equal to the load current $I_{L2} = I_L$. The deviation of the experimental load current I_L (Figs. 6–8) from the calculated value does not exceed 10% for an input voltage of 17.5 V. It should be noted that the maximum deviation is 3% for an input voltage of 6.5 V and 4% for 12.0 V.

The results of the study of the dependence of the constant and variable components of the currents i_{L1}, i_{L2} and the voltages $u_{C7, C8}$, $u_{C11, C12}$ on the input voltage U_{in} for two load resistances of 50 and 100 Ohm are shown in Figs. 9–16.

A good agreement between the experimental and calculated values is shown by the graphs of the dependence of the constant currents flowing through the windings of the chokes L1 and L2 on the input voltage U_{in} (Fig. 9) and of the voltages across the capacitors C7, C8 and C11, C12 on the input voltage U_{in} (Fig. 10) at a load resistance of 50 Ohm. The deviation in the constant components for the I_{L1} current amounts to 13% on average, while the deviation in the I_{L2} current is 15%, and the deviation in the $U_{C11, C12}$ voltage comprises 0.27%. The difference values obtained for the ripple spreads of corresponding currents and voltages obtained in the same way are as follows: Δi_{L1} is 5%, Δi_{L2} is 21%, while $\Delta u_{C11, C12}$ is 15%.

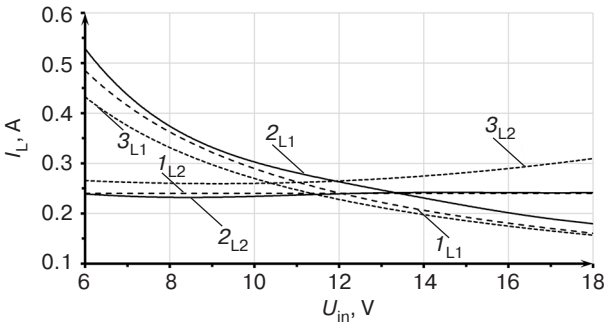


Fig. 9. Dependence of the constant currents flowing through the windings of chokes L1 and L2 on the input voltage U_{in} at a load resistance of 50 Ohm:
1 is calculation, *2* is modeling, and *3* is experimental

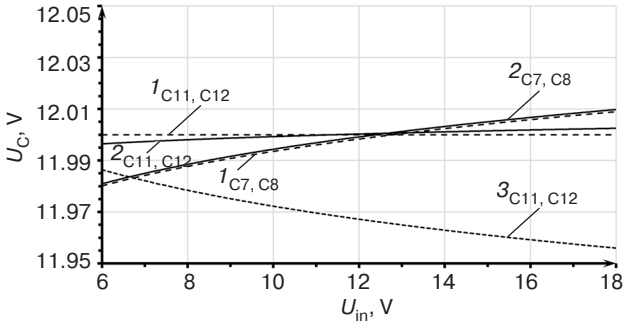


Fig. 10. Dependence of the voltages across the capacitors C7, C8 and C11, C12 on the input voltage U_{in} with a load resistance of 50 Ohm:
1 is calculation,
2 is modeling,
and *3* is experimental

The graphs of the current ripple spreads Δi_{L1} and Δi_{L2} flowing through the windings of the inductors L1 and L2 as a function of the input voltage U_{in} (Fig. 11), as well as those of the voltage ripple spreads $\Delta u_{C7, C8}$ and $\Delta u_{C11, C12}$ (Fig. 12) across the capacitors C7, C8 and C11, C12 as a function of the input voltage U_{in} at a load resistance of 50 Ohm, show good agreement between the experimental and calculated values.

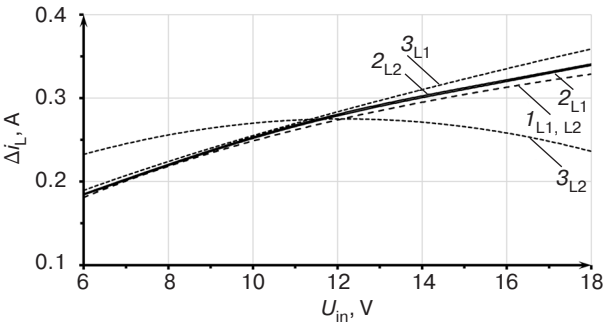


Fig. 11. Current ripple spreads through the windings of chokes L1 and L2 as a function of the input voltage U_{in} at a load resistance of 50 Ohm:
1 is calculated,
2 is modeling,
and *3* is experimental

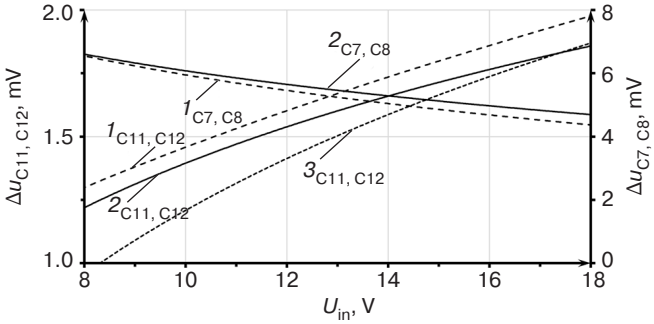


Fig. 12. Voltage ripple spreads across the capacitors C7, C8 and C11, C12 as a function of the input voltage U_{in} at a load resistance of 50 Ohm:
1 is calculation, *2* is modeling, and *3* is experimental

Similar dependencies of the constant and alternating components of the currents i_{L1}, i_{L2} (Fig. 13) and the voltages $u_{C7,\,C8}$, $u_{C11,\,C12}$ (Fig. 14) on the input voltage U_{in} are obtained at a load resistance R_L = 100 Ohm. The deviation of the constant current components of I_{L1} is on average 16%, while the deviation of I_{L2} is 9.0% and the deviation of $u_{C11,\,C12}$ is 0.10%. For the ripple spreads of the corresponding currents and voltages, the following deviations are obtained: Δi_{L1} is 13%, Δi_{L2} is 30%, while $\Delta u_{C11,\,C12}$ is 38%.

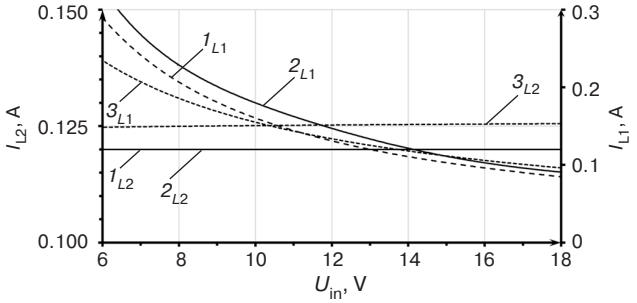


Fig. 13. Currents flowing through the windings of inductors L1 and L2 as a function of the input voltage U_{in} at a load resistance of 100 Ohm:
1 is calculation, *2* is modeling, and *3* is experimental

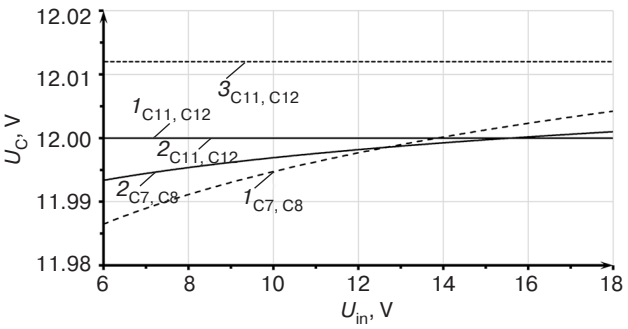


Fig. 14. Voltages across the capacitors C7, C8 and C11, C12 as a function of the input voltage U_{in} at a load resistance of 100 Ohm:
1 is calculation, *2* is modeling, and *3* is experimental

The graphs of the current ripple spreads Δi_{L1} and Δi_{L2} flowing through the windings of the inductors L1 and L2 as a function of the input voltage U_{in} (Fig. 15) and of the voltage ripple spreads $\Delta u_{C7,\,C8}$ and $\Delta u_{C11,\,C12}$ (Fig. 16) across the capacitors C7, C8 and C11, C12 as a function of the input voltage U_{in} at a load resistance of 100 Ohm show good agreement between experimental and calculated values.

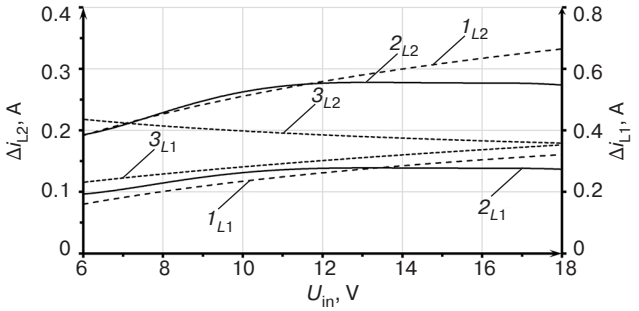


Fig. 15. Ripple currents flowing through the windings of chokes L1 and L2 as a function of the input voltage U_{in} at a load resistance of 100 Ohm:
1 is calculation, *2* is modeling, and *3* is experimental

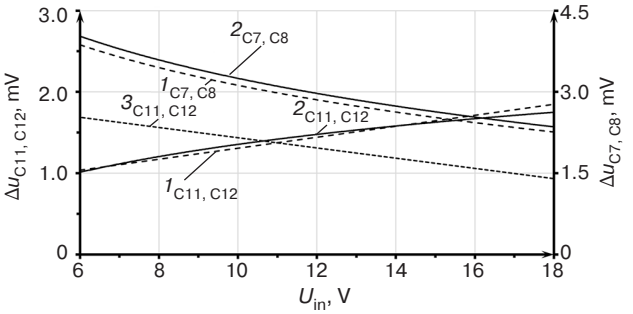


Fig. 16. Voltage ripple spreads across capacitors C7, C8 and C11, C12 as a function of the input voltage U_{in} at a load resistance of 100 Ohm:
1 is calculation, *2* is modeling, and *3* is experimental

The results of the study of the Zeta converter with inductively coupled chokes demonstrate the high correlation of both its LCs, as well as the dependence of the constant and alternating components of the currents i_{L1}, i_{L2} flowing through the windings of the chokes L1 and L2, and of the voltages $u_{C7,\,C8}$ and $u_{C11,\,C12}$ across the capacitors C7, C8 and C11, C12 on the input voltage U_{in} for two load resistances of 50 and 100 Ohm, which were obtained by different methods: experimental, calculated, and modeling. There is almost complete agreement between the calculated values and those obtained by SPICE[7] modeling. The differences between the experimental characteristics and those obtained by calculation and modeling can be considered negligible.

[7] SPICE (Simulation Program with Integrated Circuit Emphasis) is an open source simulator of general-purpose electronic circuits

CONCLUSIONS

We have described the design and construction a test bench of a DC/DC converter based on the Zeta topology with coupled chokes on the basis of a TPS40200 driver. An experimental study of the typical dependencies of the converter at different values of input voltage and load resistances was carried out. The experimental dependencies are compared with similar characteristics obtained by modeling using *Multisim* CAD and a calculation method based on the continuous-time mathematical model of the converter. A comparison of data obtained using these three methods demonstrates their high correlation.

Authors' contribution
All authors equally contributed to the research work

REFERENCES

1. Korotkov S.M., Lukin A.V. Power sources for LED lighting. *Prakticheskaya silovaya elektronika = Practical Power Electronics*. 2012;2(46):3–9 (in Russ.). https://www.elibrary.ru/papuhr
2. Obraztsov A., Obraztsov S. Circuit design of DC/DC converters. *Sovremennaya elektronika = Modern Electronics.* 2005;3:36–43 (in Russ.).
3. Bodin O.N., Bezborodova O.E., Mitroshin A.N., Chuvykin B.V., Martynov D.V., Edemskii M.V. Intelligent telemedicine information system. *Biomeditsinskaya radioelektronika = Biomedical Radioelectronics*. 2024;27(2):103–110 (in Russ.). https://doi.org/10.18127/j15604136-202402-14
4. Bityukov V.K., Simachkov D.S., Babenko V.P. *Skhemotekhnika elektropreobrazovatel'nykh ustroistv* (*Circuitry of Electrical Converter Devices*). Vologda: Infra-Inzheneriya; 2023. 384 p. (in Russ.). ISBN 978-5-9729-1439-5. https://www.elibrary.ru/pqyagy
5. Manannikova N.G., Shevtsov D.A. New Topology for the two-transistor power stage for a single-ended power converter. *Prakticheskaya silovaya elektronika = Practical Power Electronics*. 2023;1(89):17–20 (in Russ.). https://www.elibrary.ru/cuolqz
6. Anisimova T.V., Danilina A.N., Kryuchkov V.V. DC Boost Converter with Flying Capacitor. *Prakticheskaya silovaya elektronika = Practical Power Electronics*. 2021;1(81):28–33 (in Russ.). https://www.elibrary.ru/ijaakd
7. Minibaev L.M. Using zero ripple techniques in power supply designing. In: *Problems and Trends of Scientific Transformations in the Conditions of Society Transformation: Proceedings of the All-Russian Scientific and Practical Conference.* Ufa: Aeterna; 2020. P. 23–26 (in Russ.). https://www.elibrary.ru/pazzxq
8. Zhu F., Li Q. Coupled Inductors with an Adaptive Coupling Coefficient for Multiphase Voltage Regulators. *IEEE Trans. Power Electron*. 2023;38(1):739–749. https://doi.org/10.1109/TPEL.2022.3203855, https://www.elibrary.ru/hizbts
9. Zhang Ch., Yuan X., Wang J., et al. Si/WBG Hybrid Half-Bridge Converter Using Coupled Inductors for Power Quality Improvement and Control Simplification. *IEEE Trans. Power Electron*. 2024;39(3):3339–3352. https://doi.org/10.1109/TPEL.2023.3342133, https://www.elibrary.ru/kbwvtg
10. Tseng K.Ch., Huang G.Yu., Hsiung H.Yu. An isolated high step-down DC–DC converter with dual coupled inductors for ultracapacitor charger applications. *Int. J. Circuit Theor. Appl.* 2024;52(7):3341–3356. https://doi.org/10.1002/cta.3905, https://www.elibrary.ru/bfcarn
11. Bityukov V.K., Lavrenov A.I., Petrov D.R. Mathematical model of a ZETA-converter with inductively coupled chokes (Part 2). *Voprosy elektromekhaniki. Trudy VNIIEM = Elektromechanical Matters. VNIIEM Studies.* 2023;195(4):48–52 (in Russ.). https://elibrary.ru/mnusik
12. Bityukov V.K., Lavrenov A.I., Petrov D.R. Current and voltage pulsations of Zeta converter with inductively coupled inductors (Part 2). *Proektirovanie i tekhnologiya elektronnykh sredstv = Design and Technology of Electronic Means*. 2023;4:27–31 (in Russ.). https://www.elibrary.ru/dspqrt
13. Korshunov A.I. Limiting continuous model of a system with periodic high-frequency structure variation. *Silovaya elektronika = Power Electronics*. 2021;5(92):48−51 (in Russ.). https://www.elibrary.ru/sxwxqb
14. Belov G.A. Structural dynamic models of pulsed DC-DC switched mode converters in discontinuous current mode. *Prakticheskaya silovaya elektronika = Practical Power Electronics.* 2019;1(73):2–8 (in Russ.). https://www.elibrary.ru/jvniqr
15. Amelina M.A., Amelin S.A. Continuous Models of Composite DC-DC Converters. In: *Power Engineering, Computer Sciences, and Innovations – 2021: Proceedings of the 11th International Scientific and Technical Conference, Smolensk.* Smolensk: Universum; 2021. V. 1. P. 323–325 (in Russ.). https://www.elibrary.ru/klxdcg
16. Bityukov V.K., Lavrenov A.I. Method for designing DC/DC converters based on Zeta topology. *Russian Technological Journal.* 2025;13(1):59–67 (in Russ.). https://doi.org/10.32362/2500316X-2025-13-1-59-67

СПИСОК ЛИТЕРАТУРЫ

1. Коротков С.М., Лукин А.В. Источники питания для светодиодного освещения. *Практическая силовая электроника*. 2012;2(46):3–9. https://www.elibrary.ru/papuhr
2. Образцов А., Образцов С. Схемотехника DC/DC-преобразователей. *Современная электроника*. 2005;3:36–43.
3. Бодин О.Н., Безбородова О.Е., Митрошин А.Н., Чувыкин Б.В., Мартынов Д.В., Едемский М.В. Интеллектуальная телемедицинская информационная система. *Биомедицинская радиоэлектроника*. 2024;27(2):103–110. https://doi.org/10.18127/j15604136-202402-14
4. Битюков В.К., Симачков Д.С., Бабенко В.П. *Схемотехника электропреобразовательных устройств*. Вологда: Инфра-Инженерия; 2023. 384 с. ISBN 978-5-9729-1439-5. https://www.elibrary.ru/pqyagy
5. Мананникова Н.Г., Шевцов Д.А. Новая структура двухтранзисторного силового каскада для однотактного прямо-обратноходового преобразователя электроэнергии. *Практическая силовая электроника.* 2023;1(89):17–20. https://www.elibrary.ru/cuolqz
6. Анисимова Т.В., Данилина А.Н., Крючков В.В. Повышающий преобразователь постоянного напряжения с плавающим конденсатором. *Практическая силовая электроника*. 2021;1(81):28–33. https://www.elibrary.ru/ijaakd
7. Минибаев Л.М. Использовании техники нулевых пульсаций при проектировании источников питания. В сб.: *Проблемы и тенденции научных преобразований в условиях трансформации общества: сборник статей Всероссийской научно-практической конференции*. Уфа: Аэтерна; 2020. С. 23–26. https://www.elibrary.ru/pazzxq
8. Zhu F., Li Q. Coupled Inductors with an Adaptive Coupling Coefficient for Multiphase Voltage Regulators. *IEEE Trans. Power Electron*. 2023;38(1):739–749. https://doi.org/10.1109/TPEL.2022.3203855, https://www.elibrary.ru/hizbts
9. Zhang Ch., Yuan X., Wang J., et al. Si/WBG Hybrid Half-Bridge Converter Using Coupled Inductors for Power Quality Improvement and Control Simplification. *IEEE Trans. Power Electron*. 2024;39(3):3339–3352. https://doi.org/10.1109/TPEL.2023.3342133, https://www.elibrary.ru/kbwvtg
10. Tseng K.Ch., Huang G.Yu., Hsiung H.Yu. An isolated high step-down DC–DC converter with dual coupled inductors for ultracapacitor charger applications. *Int. J. Circuit Theor. Appl.* 2024;52(7):3341–3356. https://doi.org/10.1002/cta.3905, https://www.elibrary.ru/bfcarn
11. Битюков В.К., Лавренов А.И., Петров Д.Р. Математическая модель Zeta-преобразователя с индуктивно связанными дросселями (часть 2). *Вопросы электромеханики. Труды ВНИИЭМ*. 2023;195(4):48–52. https://elibrary.ru/mnusik
12. Битюков В.К., Лавренов А.И., Петров Д.Р. Пульсации токов и напряжений Zeta преобразователя с индуктивно связанными дросселями (часть 2). *Проектирование и технология электронных средств*. 2023;4:27–31. https://www.elibrary.ru/dspqrt
13. Коршунов А.И. Предельная непрерывная модель системы с периодическим высокочастотным изменением структуры. *Силовая электроника*. 2021;5(92):48−51. https://www.elibrary.ru/sxwxqb
14. Белов Г.А. Структурные динамические модели импульсных преобразователей постоянного напряжения в РПТ. *Практическая силовая электроника*. 2019;1(73):2–8. https://www.elibrary.ru/jvniqr
15. Амелина М.А., Амелин С.А. Непрерывные модели составных преобразователей постоянного напряжения. В сб.: *Энергетика, информатика, инновации – 2021: Сборник трудов XI Международной научно-технической конференции*. Т. 1. Смоленск: Универсум; 2021. С. 323–325. https://www.elibrary.ru/klxdcg
16. Битюков В.К., Лавренов А.И. Метод проектирования DC/DC-преобразователей, построенных по Zeta-топологии. *Russian Technological Journal*. 2025;13(1):59–67. https://doi.org/10.32362/2500316X-2025-13-1-59-67

About the authors

Vladimir K. Bityukov, Dr. Sci. (Eng.), Professor, Department of Radio Wave Processes and Technology, Institute of Radio Electronics and Informatics, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: bitukov@mirea.ru. ResearcherID Y-8325-2018, Scopus Author ID 6603797260, RSCI SPIN-code 3834-5360, https://orcid.org/0000-0001-6448-8509

Aleksey I. Lavrenov, Postgraduate Student, Assistant, Department of Radio Wave Processes and Technology, Institute of Radio Electronics and Informatics, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: lavrenov@mirea.ru. RSCI SPIN-code 6048-5027, https://orcid.org/0000-0001-5722-541X

Daniil A. Malitskiy, Circuit Engineer, SPUTNIX (Office 358, 359, 42/1, Bol'shoi bul'var, Skolkovo Technopark, Moscow, 121205 Russia). E-mail: malickij@mirea.ru. RSCI SPIN-code 4912-3018, https://orcid.org/0000-0003-4558-9085

Об авторах

Битюков Владимир Ксенофонтович, д.т.н., профессор, кафедра радиоволновых процессов и технологий, Институт радиоэлектроники и информатики, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: bitukov@mirea.ru. Scopus Author ID 6603797260, ResearcherID Y-8325-2018, SPIN-код РИНЦ 3834-5360, https://orcid.org/0000-0001-6448-8509

Лавренов Алексей Игоревич, аспирант, ассистент, кафедра радиоволновых процессов и технологий, Институт радиоэлектроники и информатики, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: lavrenov@mirea.ru. SPIN-код РИНЦ 6048-5027, https://orcid.org/0000-0001-5722-541X

Малицкий Даниил Александрович, инженер-схемотехник, ООО «СПУТНИКС» (121205, Россия, Москва, Технопарк «Сколково», Большой бульвар, д. 42, стр. 1, оф. 358, 359). E-mail: malickij@mirea.ru. SPIN-код РИНЦ 4912-3018, https://orcid.org/0000-0003-4558-9085

Translated from Russian into English by K. Nazarov
Edited for English language and spelling by Thomas A. Beavitt

Micro- and nanoelectronics. Condensed matter physics

Микро- и наноэлектроника. Физика конденсированного состояния

UDC 536.2
https://doi.org/10.32362/2500-316X-2025-13-2-46-56
EDN TNQTWK

RESEARCH ARTICLE

Distribution of temperature field strength on the surface of graphene inclusions in a matrix composite

Igor V. Lavrov [@],
Vladimir V. Bardushkin,
Victor B. Yakovlev

Institute of Nanotechnology of Microelectronics, Russian Academy of Sciences, Moscow, 119334 Russia
[@] Corresponding author, e-mail: iglavr@mail.ru

Abstract

Objectives. The study sets out to obtain an analytical expression for the distribution of the temperature field strength on the surfaces of anisotropic graphene inclusions taking the form of thin disks in the matrix composite and to use the obtained expressions to predict the strength of the temperature field on the surface of inclusions from the matrix side.

Methods. An inclusion taking the form of a thin circular disk represents a special limit case of an ellipsoidal inclusion. To obtain the corresponding analytical expressions, the authors use their previously derived more general expression for the operator of the concentration of the electric field strength on the surface of ellipsoidal inclusion. The approach is justified by the mathematical equivalence of problems of finding the electrostatic and temperature field in the stationary case. The operator relates the field strength on the inclusion surface from the matrix side to the average field strength in the composite sample; the corresponding expression is obtained in a generalized singular approximation.

Results. Analytical expressions were obtained for the operator of the concentration of the temperature field strength on the surface of the inclusion taking the form of a thin disk of multilayer graphene in a matrix composite. The expressions take into account inclusion anisotropy, the position of the point on the inclusion surface, the volume fraction of inclusions in the material, and the inclusion orientation. Two types of inclusion orientation distributions were considered: equally oriented inclusions and uniform distribution of inclusion orientations. Model calculations of the value for the temperature field strength at the points of the inclusion disk edge as a function of the angle between the radius vector of this point and the direction of the applied field strength were carried out.

Conclusions. In the case of graphene multilayer inclusions, it is shown that the field strength at points on their edges can exceed the applied field strength by several orders of magnitude.

Keywords: composite, matrix, graphene, inclusion, operators of temperature field strength concentration, generalized singular approximation

• **Submitted:** 17.07.2024 • **Revised:** 06.08.2024 • **Accepted:** 28.01.2025

For citation: Lavrov I.V., Bardushkin V.V., Yakovlev V.B. Distribution of temperature field strength on the surface of graphene inclusions in a matrix composite. *Russian Technological Journal.* 2025;13(2):46–56. https://doi.org/10.32362/2500-316X-2025-13-2-46-56, https://elibrary.ru/TNQTWK

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Распределение напряженности температурного поля на поверхности включений графена в матричном композите

И.В. Лавров [@], **В.В. Бардушкин, В.Б. Яковлев**

Институт нанотехнологий микроэлектроники, Российская академия наук, Москва, 119334 Россия
[@] *Автор для переписки, e-mail: iglavr@mail.ru*

Резюме

Цели. Цель работы – получить аналитическое выражение для распределения напряженности температурного поля на поверхностях анизотропных включений в форме тонких дисков в матричном композите и применить полученные выражения для прогнозирования величины напряженности температурного поля на поверхности графеновых включений со стороны матрицы.

Методы. Включение в форме тонкого кругового диска является частным предельным случаем эллипсоидального включения. Для получения требуемых аналитических выражений используется ранее полученное авторами более общее выражение для оператора концентрации напряженности электрического поля на поверхности эллипсоидального включения, поскольку задачи нахождения электростатического и температурного поля в стационарном случае математически эквивалентны. Данный оператор связывает напряженность поля на поверхности включения со стороны матрицы со средней напряженностью поля в образце композита, выражение для него получено в обобщенном сингулярном приближении.

Результаты. Получены аналитические выражения для оператора концентрации напряженности температурного поля на поверхности включения в форме тонкого диска из многослойного графена в матричном композите с учетом анизотропии включения в зависимости от положения точки на поверхности включения, от объемной доли включений в материале, от ориентации включения. Рассмотрены два вида распределения ориентаций включений: одинаково ориентированные включения и равномерное распределение ориентаций включений. Проведены модельные расчеты величины напряженности температурного поля в точках ребра включения-диска в зависимости от угла между радиус-вектором данной точки и направлением напряженности приложенного поля.

Выводы. Показано, что в случае графеновых многослойных включений в точках на их ребрах величина напряженности поля может на несколько порядков превышать напряженность приложенного поля.

Ключевые слова: композит, матрица, графен, включение, операторы концентрации напряженности температурного поля, обобщенное сингулярное приближение

• **Поступила:** 17.07.2024 • **Доработана:** 06.08.2024 • **Принята к опубликованию:** 28.01.2025

Для цитирования: Лавров И.В., Бардушкин В.В., Яковлев В.Б. Распределение напряженности температурного поля на поверхности включений графена в матричном композите. *Russian Technological Journal*. 2025;13(2):46–56. https://doi.org/10.32362/2500-316X-2025-13-2-46-56, https://elibrary.ru/TNQTWK

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

Graphene is a very promising material for various applications due to its exceptional electrical, thermal, and mechanical properties [1–4]. For example, the thermal conductivity coefficient of a single layer of graphene is up to 5000 W/(m·K) [4, 5]. In multilayer graphene, a lower thermal conductivity coefficient value is observed, which can be explained by an increase in phonon scattering due to the interactions between the layers [6]. However, even in multilayer graphene, the thermal conductivity in the plane of the layer remains high enough to be used in the development of composite materials to improve their thermal conductivity properties, which, together with their mechanical properties, are of great importance when undergoing intense external influences of different physical natures. For example, tribocomposite materials undergo uneven heating of the surface and bulk layers during operation, which affects diffusion and segregation processes in the material. As a result, the physical and mechanical properties of tribocomposites can change significantly [7, 8]. The use of materials with enhanced thermal conductivity represents one of the options to reduce the magnitude of the temperature field gradient during operation. Therefore, graphene, due to its very high thermal conductivity along the layers, is considered a very promising material to use as a small additive in composites to increase their thermal conductivity without sacrificing high mechanical and strength properties [9].

In inhomogeneous materials, a significant temperature field gradient value can occur at the micro-level close to the interfaces of homogeneous components that differ significantly in their thermal conductivity properties. This can lead to a change in the properties of the component particles of the inhomogeneous material, a weakening of the bond between inclusions and matrix in the composite, and ultimately a deterioration in the material's performance characteristics. In this regard, the ability to predict the local temperature fields at the interface between inclusions and binder (matrix) in the matrix composite is of great relevance.

There are a number of recent theoretical and experimental studies of the effective thermal conductivity properties of composites [9–12]. Some works also focus on predicting local temperature field distribution in composites, e.g. [13]. On the other hand, there are practically no studies on the distribution of the temperature field at the inclusion-matrix interface.

In [14], fundamental equations are derived for estimating the electric field strength distribution at the inclusion interface in a matrix composite. These results can be used to solve the problem of finding the temperature field strength distribution at the inclusion interface in a matrix composite due to the mathematical equivalence of the problem statements in the stationary case for the distribution of the electrostatic potential and the temperature field [15]. In this paper, a matrix composite with an ED-20 type polymer matrix and graphene multilayer inclusions in the form of thin flakes is considered. The shape of the flakes is approximated by thin circular disks. Analytical expressions are obtained for the concentration operator of the temperature field strength and the vector of the temperature field strength on the surface of the graphene inclusions from the matrix side as a function of the point location on the inclusion surface. Two cases of inclusion orientation distribution in the composite are considered: (1) equally oriented inclusions; (2) uniform spatial distribution of inclusion orientations.

PROBLEM STATEMENT. FIELD STRENGTH CONCENTRATION OPERATOR ON THE INCLUSION SURFACE IN A MATRIX COMPOSITE

We consider a sample of volume V of a statistically homogeneous matrix composite having ellipsoidal inclusions of a similar type. The matrix is isotropic with thermal conductivity k^{m}, while the inclusions (particles) are anisotropic with thermal conductivity tensor $\mathbf{k}^{\mathrm{p}}$ and the volume fraction of inclusions is equal to f. It is assumed that the shape of all inclusions is similar and that the principal axes of the thermal conductivity tensors coincide with the axes of the corresponding ellipsoids. All inclusions are assumed to be randomly distributed throughout the sample volume, while their orientations are distributed according to a probability law. It is further assumed that there are no internal heat sources in the material.

The temperature field in the sample is denoted by $T(\mathbf{r}, t)$, where $\mathbf{r}$ is the radius vector of a point in

space and t is time as per classical studies of heat conduction theory, e.g., as presented in [16, 17]. The concept of temperature field strength, which denotes a vector quantity opposite to the temperature field gradient, is neglected in a number of relevant works, i.e., the temperature field gradient is used directly in mathematical formulations [16–18]. However, in many studies dealing with the thermophysical properties of inhomogeneous media, a special notation for the intensity vector of the temperature field is introduced for convenience: $\mathbf{H}(\mathbf{r}, t) = -\nabla T(\mathbf{r}, t)$ (e.g., in [9, 19, 20]).

Let a uniform temperature field $T_0(\mathbf{r})$ with intensity $\mathbf{H}_0 = \text{const}$ (a uniform temperature field is a field with constant intensity, analogous to a uniform electrostatic field) be applied to the interface S of a given sample. A stationary temperature field $T(\mathbf{r})$ with intensity $\mathbf{H}(\mathbf{r})$ is then established. The task is to find the temperature field distribution at the S_p interface of any matrix-side inclusion in a given composite sample.

In [14], a similar problem of finding the electric field distribution at the inclusion interface in a composite is considered. Using the full mathematical analogy of the problems of electrostatic and temperature field determination in the stationary case, the expression for the temperature field strength at point $\mathbf{r}$ of the surface S_p of an ellipsoidal inclusion on the matrix side can be written as follows:

$$\mathbf{H}^m(\mathbf{r}) = \mathbf{K}^\text{H}(\mathbf{r})\langle\mathbf{H}\rangle, \quad \mathbf{r} \in S_\text{p}, \tag{1}$$

where $\langle\mathbf{H}\rangle$ is the average strength of the temperature field in the sample, which is equal to the applied field strength under the given boundary conditions of the problem [21]: $\langle\mathbf{H}\rangle = \mathbf{H}_0$; $\mathbf{K}^\text{H}(\mathbf{r})$ is the full concentration operator of the temperature field strength on the inclusion surface on the matrix side.

In turn, $\mathbf{K}^\text{H}(\mathbf{r})$ can be expressed in the following form [14]:

$$\mathbf{K}^\text{H}(\mathbf{r}) = \mathbf{K}^\text{sH}(\mathbf{r})\mathbf{K}^\text{vH}, \quad \mathbf{r} \in S_\text{p}, \tag{2}$$

where $\mathbf{K}^\text{sH}(\mathbf{r})$ is the surface field concentration operator relating the field strength at a given point of the inclusion surface on the matrix side to the average field strength in the matrix; $\mathbf{K}^\text{vH}$ is the volume field concentration operator relating the average field strength in the matrix to the average field strength in the sample.

In the generalized Maxwell–Garnett approximation, these operators have the following form [14]:

$$\mathbf{K}^\text{sH}(\mathbf{r}) = \left(\mathbf{I} + \mathbf{A}(\mathbf{r})(\mathbf{k}^\text{p} - k^\text{m}\mathbf{I})\right) \times \times \left[\mathbf{I} - \mathbf{g}(\mathbf{k}^\text{p} - k^\text{m}\mathbf{I})\right]^{-1}, \quad \mathbf{r} \in S_\text{p}, \tag{3}$$

$$\mathbf{K}^\text{vH} = \left[(1-f)\mathbf{I} + f\left\langle(\mathbf{I} - \mathbf{g}(\mathbf{k}^\text{p} - k^\text{m}\mathbf{I}))^{-1}\right\rangle\right]^{-1}, \tag{4}$$

where $\mathbf{I}$ is a unit tensor of rank 2; $\mathbf{A}(\mathbf{r})$ is a rank 2 tensor defined by the expression

$$\mathbf{A}(\mathbf{r}) = \frac{\mathbf{n}(\mathbf{r}) \otimes \mathbf{n}(\mathbf{r})}{\mathbf{n}(\mathbf{r}) \cdot (\mathbf{k}^\text{m}\mathbf{n}(\mathbf{r}))}, \ \mathbf{r} \in S_\text{p},$$

where $\mathbf{n}(\mathbf{r})$ is the external unit normal to the surface S_p at point $\mathbf{r}$; $\mathbf{k}^\text{m}$ is the heat conduction tensor of the matrix.

Since the matrix is isotropic, i.e., $\mathbf{k}^\text{m} = k^\text{m}\mathbf{I}$, the last expression can be rewritten in a simpler form, as follows:

$$\mathbf{A}(\mathbf{r}) = \frac{1}{k^\text{m}}(\mathbf{n}(\mathbf{r}) \otimes \mathbf{n}(\mathbf{r})). \tag{5}$$

The averaging in (4) is carried out over all the inclusions that are immersed in the matrix. The rank 2 tensor $\mathbf{g}$ related to the given inclusion and used in the generalized singular approximation [22] is also used in Eqs. (3) and (4). The components of tensor $\mathbf{g}$ in the coordinate system related to the ellipsoidal inclusion axes are calculated by the following equation [23]:

$$g_{ij} = -\frac{1}{4\pi}\int_0^\pi\int_0^{2\pi} \frac{n_i n_j}{n_\alpha k^\text{m}_{\alpha\beta} n_\beta}\sin\vartheta d\vartheta d\varphi, \ \ i,j = 1,2,3, \tag{6}$$

where the components of the normal n_i ($i = 1, 2, 3$) to the inclusion surface are expressed by the spherical angles ϑ, φ; α, β are the component numbers of the vector and tensor quantities.

Since in the case of an isotropic matrix $k^\text{m}_{\alpha\beta} = k^\text{m}\delta_{\alpha\beta}$ ($\delta_{\alpha\beta}$ is the Kronecker symbol), expression (6) can be rewritten as follows:

$$g_{ij} = -\frac{1}{4\pi k^\text{m}}\int_0^\pi\int_0^{2\pi} n_i n_j \sin\vartheta d\vartheta d\varphi, \ \ i,j = 1,2,3. \tag{7}$$

SPECIAL CASE OF ANISOTROPIC INCLUSIONS TAKING THE FORM OF THIN CIRCULAR DISKS

Let the inclusions in the matrix composite be anisotropic in the form of circular disks of radius a. We consider a particular inclusion occupying the region V_p with surface S_p. Let the plane of this disk form an angle α with the direction of the applied field intensity vector $\mathbf{H}_0$. We introduce the coordinate system $\xi\eta\zeta$ associated with this inclusion as follows. Proceeding from the origin O at the center of the disk, if $\alpha > 0$, we will orient the ξ axis along the projection of vector $\mathbf{H}_0$ on the plane of the disk, the ζ axis along the projection of $\mathbf{H}_0$ on the axis of rotation of the disk, and the η axis perpendicular to the ξ and ζ axes, so that the coordinate system $\xi\eta\zeta$ is right-handed.

If α = 0, i.e., vector $\mathbf{H}_0$ lies in the plane of the disk, we will orient the ξ axis along $\mathbf{H}_0$, the η axis perpendicular to the ξ axis in the plane of the disk, and the ζ axis perpendicular to the plane of the disk, so that the system ξηζ is right-handed. We will consider two points on the surface S_p of a given disk: a point M on the side surface of the disk and a point Q on the upper bound of the disk. Let the radius vector of point M make an angle θ with the ξ axis, then for the external unit normal to the surface S_p at point M we have: $\mathbf{n}(M) = (\cos\theta\ \sin\theta\ 0)^\text{T}$. For the normal at point Q: $\mathbf{n}(Q) = (0\ 0\ 1)^\text{T}$. Then, for the tensor $\mathbf{A}(\mathbf{r})$ at these points in the system ξηζ, we derive the following by Eq. (5):

$$\mathbf{A}(M) = \frac{1}{k^\text{m}}\begin{pmatrix} \cos^2\theta & \cos\theta\sin\theta & 0 \\ \cos\theta\sin\theta & \sin^2\theta & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad (8)$$

$$\mathbf{A}(Q) = \frac{1}{k^\text{m}}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \qquad (9)$$

For the tensor $\mathbf{g}$ of the disk-shaped inclusion, we derive from Eq. (7) the following:

$$\mathbf{g} = -\frac{1}{k^\text{m}}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \qquad (10)$$

For the multilayer graphene inclusion, the heat conduction tensor in the ξηζ system has the following form:

$$\mathbf{k}^\text{p} = \begin{pmatrix} k_\perp & 0 & 0 \\ 0 & k_\perp & 0 \\ 0 & 0 & k_\parallel \end{pmatrix}, \qquad (11)$$

where $k_\perp$ and $k_\parallel$ are the main components of the thermal conductivity along and across the graphene layers, respectively.

For convenience, we introduce a rank 2 tensor $\boldsymbol{\lambda}$ related to a particular inclusion, according to the following equation:

$$\boldsymbol{\lambda} = \left[\mathbf{I} - \mathbf{g}(\mathbf{k}^\text{p} - k^\text{m}\mathbf{I})\right]^{-1}. \qquad (12)$$

Given (10) and (11), we obtain the following form for the system ξηζ:

$$\boldsymbol{\lambda}' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & k^\text{m}/k_\parallel \end{pmatrix}. \qquad (13)$$

Taking into account Eqs. (2)–(4) and (12), the expression for the full concentration operator of the temperature field strength on the inclusion surface then takes the following form:

$$\mathbf{K}^\text{H}(\mathbf{r}) = \left(\mathbf{I} + \mathbf{A}(\mathbf{r})(\mathbf{k}^\text{p} - k^\text{m}\mathbf{I})\right)\boldsymbol{\lambda} \times \\ \times \left[(1-f)\mathbf{I} + f\langle\boldsymbol{\lambda}\rangle\right]^{-1}, \quad \mathbf{r} \in S_\text{p}, \qquad (14)$$

where the form of the tensor $\mathbf{A}(\mathbf{r})$ depends on the point on the inclusion surface; for the point M on the edge of the disk, it has the form (8), while for the point Q on the upper bound of the disk, it has the form (9).

The averaging in (14) is performed over all inclusions in the matrix. Since all inclusions are assumed to be identical, this averaging is performed over all inclusion orientations in the *xyz* coordinate system related to the texture of the composite sample.

For the distribution of inclusion orientations in a composite, we consider two cases: 1) inclusions with equal orientation; 2) uniform distribution of inclusion orientations. In the first case, the composite obtained is anisotropic, the orientations of all the systems ξηζ related to the inclusions are identical. Therefore, it is convenient to take a system ξηζ as the *xyz* coordinate system. Then $\langle\boldsymbol{\lambda}\rangle = \boldsymbol{\lambda}'$, and we obtain the following for $\mathbf{K}^\text{H}(\mathbf{r})$:

$$\mathbf{K}^\text{H}(\mathbf{r}) = \left(\mathbf{I} + \mathbf{A}(\mathbf{r})(\mathbf{k}^\text{p} - k^\text{m}\mathbf{I})\right)\boldsymbol{\lambda}' \times \\ \times \left[(1-f)\mathbf{I} + f\boldsymbol{\lambda}'\right]^{-1}, \quad \mathbf{r} \in S_\text{p}.$$

Given the form $\boldsymbol{\lambda}'$ (13), we find:

$$\mathbf{K}^\text{H}(\mathbf{r}) = \left(\mathbf{I} + \mathbf{A}(\mathbf{r})(\mathbf{k}^\text{p} - k^\text{m}\mathbf{I})\right) \times \\ \times \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \dfrac{k^\text{m}}{(1-f)k_\parallel + fk^\text{m}} \end{pmatrix}. \qquad (15)$$

In the case of a uniform distribution of inclusion orientations [24], we have:

$$\langle\boldsymbol{\lambda}\rangle = \frac{1}{3}(\lambda'_{11} + \lambda'_{22} + \lambda'_{33})\mathbf{I},$$

where $\lambda'_{11}, \lambda'_{22}, \lambda'_{33}$ are the main components of the tensor $\boldsymbol{\lambda}$, i.e., in this case, taking into account (13), we obtain:

$$\langle\lambda\rangle = \frac{1}{3}\left(2 + \frac{k^\text{m}}{k_\parallel}\right)\mathbf{I},$$

$$\left[(1-f)\mathbf{I}+f\langle\boldsymbol{\lambda}\rangle\right]^{-1}=\frac{3k_{\|}}{3k_{\|}-f(k_{\|}-k^{\mathrm{m}})}\mathbf{I},$$

while for the operator $\mathbf{K}^{\mathrm{H}}(\mathbf{r})$, we get:

$$\mathbf{K}^{\mathrm{H}}(\mathbf{r})=\left(\mathbf{I}+\mathbf{A}(\mathbf{r})(\mathbf{k}^{\mathrm{p}}-k^{\mathrm{m}}\mathbf{I})\right)\boldsymbol{\lambda}'\times \times\frac{3k_{\|}}{3k_{\|}-f(k_{\|}-k^{\mathrm{m}})},\quad \mathbf{r}\in S_{\mathrm{p}}. \tag{16}$$

In both cases of the distribution of the inclusion orientations, $\mathbf{K}^{\mathrm{H}}(\mathbf{r})$ is diagonal.

We consider the special case when the point M on an edge of the disk lies on the ξ axis, i.e., when θ = 0. In this case we have:

$$\mathbf{A}(M)=\frac{1}{k^{\mathrm{m}}}\begin{pmatrix}1&0&0\\0&0&0\\0&0&0\end{pmatrix}. \tag{17}$$

Then for the diagonal components of the operator $\mathbf{K}^{\mathrm{H}}(\mathbf{r})$ in the case of the same orientation of the inclusions from (15), taking into account (11), we have:

$$K_{11}^{\mathrm{H}}(M(0))=\frac{k_{\perp}}{k^{\mathrm{m}}},\quad K_{22}^{\mathrm{H}}(M(0))=1,\quad K_{33}^{\mathrm{H}}(M(0))=\frac{k^{\mathrm{m}}}{(1-f)k_{\|}+fk^{\mathrm{m}}}. \tag{18}$$

In the case of uniformly distributed inclusion orientations, from (16), (17), and (11) we obtain:

$$K_{11}^{\mathrm{H}}(M(0))=\frac{3k_{\perp}k_{\|}}{k^{\mathrm{m}}(3k_{\|}-f(k_{\|}-k^{\mathrm{m}}))},\quad K_{22}^{\mathrm{H}}(M(0))=\frac{3k_{\|}}{3k_{\|}-f(k_{\|}-k^{\mathrm{m}})},\quad K_{33}^{\mathrm{H}}(M(0))=\frac{3k^{\mathrm{m}}}{3k_{\|}-f(k_{\|}-k^{\mathrm{m}})}. \tag{19}$$

If the thermal conductivity of graphene multilayer inclusions is considered approximately equal to that of high quality graphite, in this case we have the following values of thermal conductivity component (W/(m·K)): $k_{\perp}=2000$, $k_{\|}=5.7$ [25], for an ED-20 type epoxy matrix k^{m} = 0.2 [26]. Then formula (18) gives $K_{11}^{\mathrm{H}}(M(0))=10^4$ for equally oriented inclusions, i.e., the temperature field strength component H_1 at the point M of the matrix-side inclusion interface is 10^4 times higher than the corresponding component of the applied field.

We now obtain the expressions for $\mathbf{K}^{\mathrm{H}}(\mathbf{r})$ at the point Q on the inclusion edge. Substituting (9) into (15), we obtain the following for the diagonal components of the operator $\mathbf{K}^{\mathrm{H}}(Q)$ for equally oriented inclusions:

$$K_{11}^{\mathrm{H}}(Q)=1,\quad K_{22}^{\mathrm{H}}(Q)=1,\quad K_{33}^{\mathrm{H}}(Q)=\frac{k_{\|}}{(1-f)k_{\|}+fk^{\mathrm{m}}}. \tag{20}$$

For uniformly oriented inclusions, we have:

$$K_{11}^{\mathrm{H}}(Q)=K_{22}^{\mathrm{H}}(Q)=K_{33}^{\mathrm{H}}(Q)=\frac{3k_{\|}}{3k_{\|}-f(k_{\|}-k^{\mathrm{m}})}. \tag{21}$$

It can be seen from Eq. (20) and (21) that the field strength at point Q on the inclusion upper bound is of the same order of magnitude as the applied field strength.

NUMERICAL MODELING RESULTS AND DISCUSSION

Based on the derived expressions for the full concentration operator of the temperature field strength, model calculations are carried out for a composite with an ED-20 type matrix and multilayer graphene inclusions taking the form of circular disks. The ratios of the components and the modulus of the temperature field strength at the point M on the edge of the inclusion disk to the modulus of the applied field strength are calculated as a function of the angle θ between the radius vector of this point and the ξ axis for different inclusion volume fractions, for different values of the angle α between the applied field strength and the inclusion plane. Some results are shown in Figs. 1–3. In all cases, the distribution of inclusion orientations is assumed to be uniform.

The dependencies of the H_i/H_0 ratio of the temperature field strength components at points M on the edge of the graphene inclusion to the applied field strength on the angle between the radius vector to point M and the applied field strength $\mathbf{H}_0$ for the case when $\mathbf{H}_0$ lies in the plane of the inclusion are shown in Fig. 1. An analysis of these dependencies shows that, for a fixed value of the applied field strength, the values of the components H_1 and H_2 at points on the edge of disks on the matrix side depend significantly on the angle θ between the radius vector of this point and the vector of the applied field strength. However, in the vast majority of such points the value of the corresponding strength component is rather high compared to the applied field. At the same time, the H_3 component has a negligible value close to zero.

Similar dependencies of the ratio $H(M)/H_0$ of the absolute magnitude of the temperature field strength to the applied field strength are shown in Fig. 2. It can be seen that the modulus of the field strength at the points on the disk edge in the ranges $\theta \in [0°; 84°]$ and $\theta \in [96°; 180°]$ is more than 10^3 times higher than the applied field strength. As the volume fraction of inclusions in the composite increases with a uniform distribution of their orientations, the absolute values of the components and the modulus of the field strength at the points on the disk edges also increase slightly. In the case of the same inclusion orientations in the composite, the change in the inclusion volume fraction has no effect on the values of the components H_1 and H_2. This follows directly from Eq. (18).

In the general case, with respect to the direction of the vector $\mathbf{H}_0$ of the applied field strength, the inclusion disk planes are oriented differently. The dependencies of the ratio $H(M)/H_0$ on the angle θ between the radius vector of the point M and the projection of the vector $\mathbf{H}_0$ onto the disk plane for different values of the angle α between the vector $\mathbf{H}_0$ and the inclusion plane are shown in Fig. 3. These dependencies show that increasing the angle between the disk plane and $\mathbf{H}_0$ leads to a decrease in the value of the field strength at points at the disk edge. At the same time, this value is still much higher than H_0. For example, for the angle $\alpha = 75°$, the ratio of the surface field strength to the applied field exceeds 10^3 for points in the ranges $\theta \in [0°; 66°]$ and $\theta \in [114°; 180°]$.

From the results, it can be concluded that the physical properties of the binder can be significantly modified by intensifying the diffusion and segregation processes taking place in these regions. This is due to the significant values of the temperature field strength in the regions near the edges of the graphene disks. For small volume fractions of graphene inclusions, these changes have no significant effect on the macroscopic properties of the composite. However, as the inclusion volume fraction increases, the proportion of the binder material regions in which these changes occur also increases. This can lead to a significant degradation of the performance characteristics of the material, which is consistent with the results obtained in [27].

CONCLUSIONS

The main result of the paper is Eqs. (15) and (16) for the concentration operators of the temperature field strength on the surface of anisotropic disk-shaped inclusions in a matrix composite. These expressions allow the prediction of these values at any point on the surface of the inclusions as a function of the external applied field, the volume fractions and

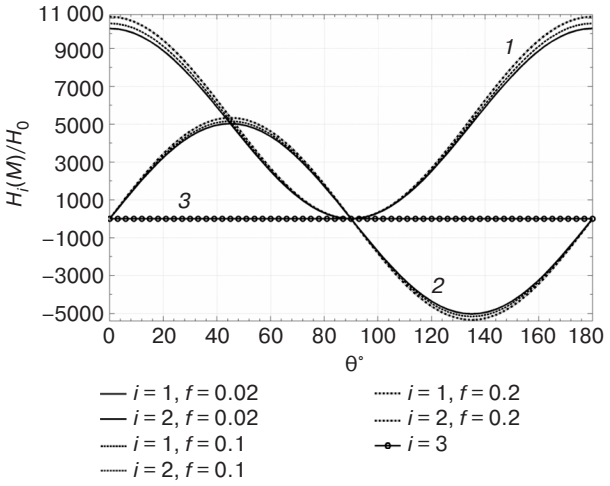


Fig. 1. Dependencies of the H_i/H_0 ratio on the angle between the radius vector to the point M and the applied field strength $\mathbf{H}_0$ for different inclusion volume fractions. The component numbers are given near the corresponding curves

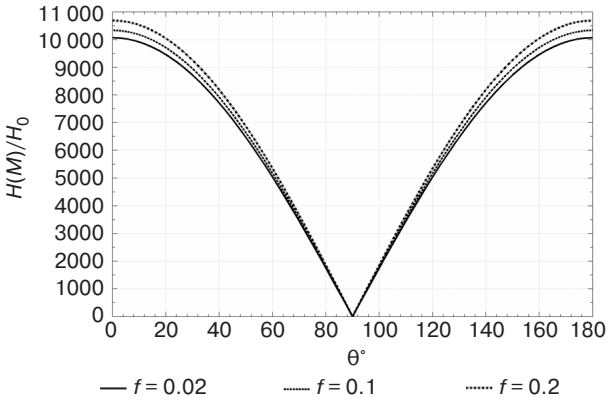


Fig. 2. Dependencies of the $H(M)/H_0$ ratio on the angle between the radius vector to the point M and the vector of the applied field strength $\mathbf{H}_0$ at different inclusion volume fractions

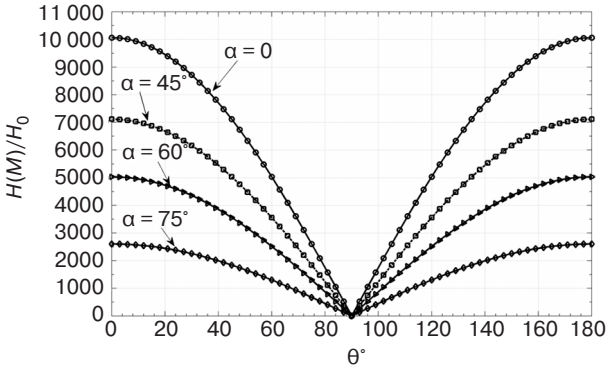


Fig. 3. Dependencies of the $H(M)/H_0$ ratio on the angle between the radius vector to the point M and the projection of the vector $\mathbf{H}_0$ onto the disk plane for different values of the angle α between the vector $\mathbf{H}_0$ and the inclusion plane. Inclusion volume fraction $f = 0.02$

material properties of the composite components, and the orientation of the inclusion with respect to the direction of the applied field strength. Modeling calculations have been carried out for the inclusions of graphene multilayers. It is shown that the temperature field strength can exceed the applied field strength by several orders of magnitude at the surface points on the edges of graphene inclusions.

Authors' contributions

I.V. Lavrov—literature review, deducing the calculation formulas, writing the computer programs, model calculations, plotting, discussion of the results.

V.V. Bardushkin—checking the mathematical correctness of deducing the calculation formulas, proofreading the text of the article, discussion of the results.

V.B. Yakovlev—problem statement, the idea of deducing the calculation formulas, discussion of the results.

REFERENCES

1. Novoselov K.S., Geim A.K., Morozov S.V., Jiang D., Zhang Y., Dubonos S.V., Grigorieva I.V., Firsov A.A. Electric field effect in atomically thin carbon films. *Science*. 2004;306(5696):666–669. https://doi.org/10.1126/science.1102896
2. Novoselov K.S. Graphene: Materials in the Flatland. *Uspekhi Fizicheskikh Nauk*. 2011;181(12):1299–1311 (in Russ.). https://doi.org/10.3367/UFNr.0181.201112f.1299
3. Bunch J.S., Van der Zande A.M., Verbridge S.S., Frank I.W., Tanenbaum D.M., Parpia J.M., Craighead H.G., McEuen P.L. Electromechanical resonators from graphene sheets. *Science*. 2007;315(5811):490–493. https://doi.org/10.1126/science.1136836
4. Yan Zh., Nika D.L., Balandin A.A. Thermal properties of graphene and few-layer graphene: applications in electronics. *IET Circuits, Devices & Systems*. 2015;9(1):4–12. https://doi.org/10.1049/iet-cds.2014.0093
5. Tkachev S.V., Buslaeva E.Y., Gubin S.P. Graphene: a novel carbon nanomaterial. *Neorg. Mater.* 2011;47(1):1–10. https://doi.org/10.1134/S0020168511010134
   [Original Russian Text: Tkachev S.V., Buslaeva E.Y., Gubin S.P. Graphene: a novel carbon nanomaterial. *Neorganicheskie materialy.* 2011;47(1):5–14 (in Russ.).]
6. Eletskii A.V., Iskandarova I.M., Knizhnik A.A., Krasikov D.N. Graphene: fabrication methods and thermophysical properties. *Phys.-Usp*. 2011;54(3):227–258. https://doi.org/10.3367/UFNe.0181.201103a.0233
   [Original Russian Text: Eletskii A.V., Iskandarova I.M., Knizhnik A.A., Krasikov D.N. Graphene: fabrication methods and thermophysical properties. *Uspekhi Fizicheskikh Nauk*. 2011;181(3):233–268 (in Russ.).]
7. Kolesnikov V.I. *Teplofizicheskie protsessy v metallopolimernykh tribosistemakh* (*Thermophysical Processes in Metal-Polymeric Tribosystems*). Moscow: Nauka, 2003. 279 p. (in Russ.). ISBN 5-02-002843-6
8. Kolesnikov V.I., Kozakov A.T., Sidashov A.V., Kravchenko V.N., Sychev A.P. Diffusion and segregation processes in metal-polymer tribosystem. *Trenie i iznos = Friction and Wear.* 2006;27(4):361–365 (in Russ.).
9. Lavrov I.V., Bardushkin V.V., Yakovlev V.B. Prediction of the effective thermal conductivity of composites with graphene inclusions. *Teplovye protsessy v tekhnike = Thermal Processes in Engineering*. 2023;15(7):299–308 (in Russ.).
10. Zarubin V.S., Zimin V.N., Kuvyrkin G.N., Savelyeva I.Y., Novozhilova O.V. Two-sided estimate of effective thermal conductivity coefficients of a textured composite with anisotropic ellipsoidal inclusions. *Z. Angew. Math. Phys.* (*ZAMP*). 2023;74(4):139. https://doi.org/10.1007/s00033-023-02039-0
11. Bonfoh N., Dinzart F., Sabar H. New exact multi-coated ellipsoidal inclusion model for anisotropic thermal conductivity of composite materials. *Appl. Math. Modell.* 2020;87(12):584–605. https://doi.org/10.1016/j.apm.2020.06.005
12. Shalygina T.A., Melezhik A.V., Tkachev A.G., et al. The Synergistic Effect of a Hybrid Filler Based on Graphene Nanoplates and Multiwalled Nanotubes for Increasing the Thermal Conductivity of an Epoxy Composite. *Tech. Phys. Lett.* 2021;47(7):364–367. https://doi.org/10.1134/S1063785021040143
    [Original Russian Text: Shalygina T.A., Melezhik A.V., Tkachev A.G., Voronina S.Yu., Voronchikhin V.D., Vlasov A.Yu. The Synergistic Effect of a Hybrid Filler Based on Graphene Nanoplates and Multiwalled Nanotubes for Increasing the Thermal Conductivity of an Epoxy Composite. *Pis'ma v Zhurnal tekhnicheskoi fiziki*. 2021;47(7):3–5 (in Russ.). https://doi.org/10.21883/PJTF.2021.07.50789.18609]
13. Kolesnikov V.I., Lavrov I.V., Bardushkin V.V., Sychev A.P., Yakovlev V.B. A method of the estimation of the local thermal fields' distribution in multicomponent composites. *Nauka Yuga Rossii = Science in the South Russia*. 2017;13(2):13–20 (in Russ.). https://doi.org/10.23885/2500-0640-2017-13-2-13-20
14. Kolesnikov V.I., Yakovlev V.B., Lavrov I.V., et al. Distribution of Electric Fields on the Surface of Inclusions in a Matrix Composite. *Dokl. Phys*. 2023;68(11):370–375. https://doi.org/10.1134/S1028335823110058
    [Original Russian Text: Kolesnikov V.I., Yakovlev V.B., Lavrov I.V., Sychev A.P., Bardushkin A.V. Distribution of Electric Fields on the Surface of Inclusions in a Matrix Composite. *Doklady Rossiiskoi akademii nauk. Fizika, tekhnicheskie nauki*. 2023;513(1):34–40 (in Russ.). https://doi.org/10.31857/S2686740023060093]
15. Milton G. *The Theory of Composites*. Cambridge: Cambridge University Press; 2004. 719 p.
16. Lykov A.V. *Teoriya teploprovodnosti* (*Theory of Thermal Conductivity*). Moscow: Vysshaya shkola; 1967. 600 p. (in Russ.).
17. Kartashov E.M., Kudinov V.A. *Analiticheskie metody teorii teploprovodnosti i ee prilozhenii* (*Analytical Methods of the Theory of Thermal Conductance and its Applications*). Moscow: Lenand; 2018. 1072 p. (in Russ.). ISBN 978-5-9710-4994-4

18. Kartashov E.M. New energy effect in non-cylindrical domains with a thermally insulated moving boundary. *Russian Technological Journal*. 2023;11(5):106–117 (in Russ.). https://doi.org/10.32362/2500-316X-2023-11-5-106-117
19. Benveniste Y., Miloh T. The effective conductivity of composites with imperfect thermal contact at constituent interfaces. *Int. J. Eng. Sci*. 1986;24(9):1537–1552. https://doi.org/10.1016/0020-7225(86)90162-X
20. Benveniste Y. On the effective thermal conductivity of multiphase composites. *Z. Angew. Math. Phys.* (*ZAMP*). 1986;37: 696–713. https://doi.org/10.1007/BF00947917
21. Stroud D. Generalized effective-medium approach to the conductivity of an inhomogeneous material. *Phys. Rev. B.* 1975;12(8):3368–3373. https://doi.org/10.1103/PhysRevB.12.3368
22. Shermergor T.D. *Teoriya uprugosti mikroneodnorodnykh sred* (*Micromechanics of Inhomogeneous Medium*). Moscow: Nauka; 1977. 399 p. (in Russ.).
23. Kolesnikov V.I., Yakovlev V.B., Bardushkin V.V., Lavrov I.V., Sychev A.P., Yakovleva E.N. A Method of Analysis of Distributions of Local Electric Fields in Composites. *Dokl. Phys.* 2016;61(3):124–128. https://doi.org/10.1134/S1028335816030101 [Original Russian Text: Kolesnikov V.I., Yakovlev V.B., Bardushkin V.V., Lavrov I.V., Sychev A.P., Yakovleva E.N. A Method of Analysis of Distributions of Local Electric Fields in Composites. *Doklady akademii nauk.* 2016;467(3): 275–279 (in Russ.). https://doi.org/10.7868/S0869565216090097]
24. Lavrov I.V. Permittivity of composite material with texture: ellipsoidal anisotropic inclusions. *Ekologicheskii vestnik nauchnykh tsentrov Chernomorskogo ekonomicheskogo sotrudnichestva = Ecological Bulletin of Research Centers of the Black Sea Economic Cooperation.* 2009;1:52–58 (in Russ.).
25. Grigor'ev I.S., Meilikhov E.Z. *Fizicheskie velichiny: spravochnik* (*Physical Quantities: A Handbook*). Moscow: Energoatomizdat; 1991. 1232 p. (in Russ.).
26. Lee H., Neville K. *Spravochnoe rukovodstvo po epoksidnym smolam* (*Handbook of Epoxy Resins*): transl. from Engl. Moscow: Energiya; 1973. 415 p. (in Russ.).
[Lee H., Neville K. *Handbook of Epoxy Resins.* N.-Y.: McGraw-Hill; 1967. 922 p.]
27. Sheinerman A.G., Krasnitskii S.A. Modeling of the Influence of Graphene Agglomeration on the Mechanical Properties of Ceramic Composites with Graphene. *Tech. Phys. Lett*. 2021;47(12):873–876. https://doi.org/10.1134/S106378502109011X [Original Russian Text: Sheinerman A.G., Krasnitskii S.A. Modeling of the Influence of Graphene Agglomeration on the Mechanical Properties of Ceramic Composites with Graphene. *Pis'ma v Zhurnal tekhnicheskoi fiziki*. 2021;47(17):37–40 (in Russ.). https://doi.org/10.21883/PJTF.2021.17.51385.18844]

СПИСОК ЛИТЕРАТУРЫ

1. Novoselov K.S., Geim A.K., Morozov S.V., Jiang D., Zhang Y., Dubonos S.V., Grigorieva I.V., Firsov A.A. Electric field effect in atomically thin carbon films. *Science*. 2004;306(5696):666–669. https://doi.org/10.1126/science.1102896
2. Новоселов К.С. Графен: материалы Флатландии. *Успехи физических наук* (*УФН*). 2011;81(12):1299–1311. https://doi.org/10.3367/UFNr.0181.201112f.1299
3. Bunch J.S., Van der Zande A.M., Verbridge S.S., Frank I.W., Tanenbaum D.M., Parpia J.M., Craighead H.G., McEuen P.L. Electromechanical resonators from graphene sheets. *Science*. 2007;315(5811):490–493. https://doi.org/10.1126/science.1136836
4. Yan Zh., Nika D.L., Balandin A.A. Thermal properties of graphene and few-layer graphene: applications in electronics. *IET Circuits, Devices & Systems*. 2015;9(1):4–12. https://doi.org/10.1049/iet-cds.2014.0093
5. Ткачев С.В., Буслаева Е.Ю., Губин С.П. Графен – новый углеродный наноматериал. *Неорганические материалы.* 2011;47(1):5–14.
6. Елецкий А.В., Искандарова И.М., Книжник А.А., Красиков Д.Н. Графен: методы получения и теплофизические свойства. *Успехи физических наук* (*УФН*). 2011;181(3):233–268.
7. Колесников В.И. *Теплофизические процессы в металлополимерных трибосистемах.* М.: Наука; 2003. 279 с. ISBN 5-02-002843-6
8. Колесников В.И., Козаков А.Т., Сидашов А.В., Кравченко В.Н., Сычев А.П. Диффузионные и сегрегационные процессы в металлополимерной трибосистеме. *Трение и износ*. 2006;27(4):361–365.
9. Лавров И.В., Бардушкин В.В., Яковлев В.Б. Прогнозирование эффективной теплопроводности композитов с графеновыми включениями. *Тепловые процессы в технике.* 2023;15(7):299–308.
10. Zarubin V.S., Zimin V.N., Kuvyrkin G.N., Savelyeva I.Y., Novozhilova O.V. Two-sided estimate of effective thermal conductivity coefficients of a textured composite with anisotropic ellipsoidal inclusions. *Z. Angew. Math. Phys.* (*ZAMP*). 2023;74(4):139. https://doi.org/10.1007/s00033-023-02039-0
11. Bonfoh N., Dinzart F., Sabar H. New exact multi-coated ellipsoidal inclusion model for anisotropic thermal conductivity of composite materials. *Appl. Math. Modell.* 2020;87(12):584–605. https://doi.org/10.1016/j.apm.2020.06.005
12. Шалыгина Т.А., Мележик А.В., Ткачев А.Г., Воронина С.Ю., Ворончихин В.Д., Власов А.Ю. Синергический эффект гибридного наполнителя на основе графеновых нанопластин и многостенных нанотрубок для повышения теплопроводности эпоксидного композита. *Письма в ЖТФ.* 2021;47(7):3–6. https://doi.org/10.21883/PJTF.2021.07.50789.18609
13. Колесников В.И., Лавров И.В., Бардушкин В.В., Сычев А.П., Яковлев В.Б. Метод оценки распределений локальных температурных полей в многокомпонентных композитах. *Наука Юга России.* 2017;13(2):13–20. https://doi.org/10.23885/2500-0640-2017-13-2-13-20

14. Колесников В.И., Яковлев В.Б., Лавров И.В., Сычев А.П., Бардушкин А.В. Распределение электрических полей на поверхности включений в матричном композите. *Доклады Российской академии наук. Физика, технические науки.* 2023;513(1):34–40. https://doi.org/10.31857/S2686740023060093, https://elibrary.ru/htskme
15. Milton G. *The Theory of Composites*. Cambridge: Cambridge University Press; 2004. 719 p.
16. Лыков А.В. *Теория теплопроводности*. М.: Высшая школа; 1967. 600 с.
17. Карташов Э.М., Кудинов В.А. *Аналитические методы теории теплопроводности и ее приложений*. М.: ЛЕНАНД; 2018. 1072 с. ISBN 978-5-9710-4994-4
18. Карташов Э.М. Новый энергетический эффект в областях нецилиндрического типа с термоизолированной движущейся границей. *Russian Technological Journal*. 2023;11(5):106–117. https://doi.org/10.32362/2500-316X-2023-11-5-106-117
19. Benveniste Y., Miloh T. The effective conductivity of composites with imperfect thermal contact at constituent interfaces. *Int. J. Eng. Sci*. 1986;24(9):1537–1552. https://doi.org/10.1016/0020-7225(86)90162-X
20. Benveniste Y. On the effective thermal conductivity of multiphase composites. *Z. Angew. Math. Phys.* (*ZAMP*). 1986;37: 696–713. https://doi.org/10.1007/BF00947917
21. Stroud D. Generalized effective-medium approach to the conductivity of an inhomogeneous material. *Phys. Rev. B.* 1975;12(8):3368–3373. https://doi.org/10.1103/PhysRevB.12.3368
22. Шермергор Т.Д. *Теория упругости микронеоднородных сред*. М.: Наука; 1977. 399 с.
23. Колесников В.И., Яковлев В.Б., Бардушкин В.В., Лавров И.В., Сычев А.П., Яковлева Е.Н. О методе анализа распределений локальных электрических полей в композиционном материале. *Доклады академии наук* (*ДАН*). 2016;467(3):275–279. https://doi.org/10.7868/S0869565216090097
24. Лавров И.В. Диэлектрическая проницаемость композиционных материалов с текстурой: эллипсоидальные анизотропные кристаллиты. *Экологический вестник научных центров Черноморского экономического сотрудничества*. 2009;1:52–58.
25. Григорьев И.С., Мейлихов Е.З. *Физические величины: справочник*. М.: Энергоатомиздат; 1991. 1232 с.
26. Ли Х., Невилл К. *Справочное руководство по эпоксидным смолам*: пер. с англ. М.: Энергия; 1973. 415 с.
27. Шейнерман А.Г., Красницкий С.А. Моделирование влияния агломерации графена на механические свойства керамических композитов с графеном. *Письма в ЖТФ.* 2021;47(17):37–40. https://doi.org/10.21883/PJTF.2021.17.51385.18844

About the authors

Igor V. Lavrov, Cand. Sci. (Phys.-Math.), Assistant Professor, Senior Researcher, Institute of Nanotechnology of Microelectronics, Russian Academy of Sciences (32A, Leninskii pr., Moscow, 119334 Russia). E-mail: iglavr@mail.ru. Scopus Author ID 35318030100, ResearcherID D-1011-2017, RSCI SPIN-code 2322-7217, https://orcid.org/0000-0002-1467-5100

Vladimir V. Bardushkin, Dr. Sci. (Phys.-Math.), Assistant Professor, Chief Researcher, Institute of Nanotechnology of Microelectronics, Russian Academy of Sciences (32A, Leninskii pr., Moscow, 119334 Russia). E-mail: bardushkin@mail.ru. Scopus Author ID 55620242900, ResearcherID D-1010-2017, RSCI SPIN-code 4294-9040, https://orcid.org/0000-0002-8805-5764

Victor B. Yakovlev, Dr. Sci. (Phys.-Math.), Professor, Chief Researcher, Scientific Secretary, Institute of Nanotechnology of Microelectronics, Russian Academy of Sciences (32A, Leninskii pr., Moscow, 119334 Russia). E-mail: yakvb@mail.ru. Scopus Author ID 7201907574, ResearcherID E-7995-2017, RSCI SPIN code 4318-0749, https://orcid.org/0000-0001-8515-3951

Об авторах

Лавров Игорь Викторович, к.ф.-м.н., доцент, старший научный сотрудник, ФГБУН «Институт нанотехнологий микроэлектроники Российской академии наук» (119334, Россия, Москва, Ленинский пр-т, д. 32А). E-mail: iglavr@mail.ru. Scopus Author ID 35318030100, ResearcherID D-1011-2017, SPIN-код РИНЦ 2322-7217, https://orcid.org/0000-0002-1467-5100

Бардушкин Владимир Валентинович, д.ф.-м.н., доцент, главный научный сотрудник, ФГБУН «Институт нанотехнологий микроэлектроники Российской академии наук» (119334, Россия, Москва, Ленинский пр-т, д. 32А). E-mail: bardushkin@mail.ru. Scopus Author ID 55620242900, ResearcherID D-1010-2017, SPIN-код РИНЦ 4294-9040, https://orcid.org/0000-0002-8805-5764

Яковлев Виктор Борисович, д.ф.-м.н., профессор, главный научный сотрудник и ученый секретарь, ФГБУН «Институт нанотехнологий микроэлектроники Российской академии наук» (119334, Россия, Москва, Ленинский пр-т, д. 32А). E-mail: yakvb@mail.ru. Scopus Author ID 7201907574, ResearcherID E-7995-2017, SPIN-код РИНЦ 4318-0749, https://orcid.org/0000-0001-8515-3951

Translated from Russian into English by K. Nazarov
Edited for English language and spelling by Thomas A. Beavitt

Micro- and nanoelectronics. Condensed matter physics

Микро- и наноэлектроника. Физика конденсированного состояния

UDC 621.382.3
https://doi.org/10.32362/2500-316X-2025-13-2-57-73
EDN TTUFNR

REVIEW ARTICLE

Thermal and mechanical degradation mechanisms in heterostructural field-effect transistors based on gallium nitride

Vadim M. Minnebaev @

Microwave Systems, Moscow, 105122 Russia
@ *Corresponding author, e-mail: vm@mwsystems.ru*

Abstract

Objectives. Gallium nitride heterostructural field-effect transistors (GaN HFET) are among the most promising semiconductor devices for power and microwave electronics. Over the past 10–15 years, GaN HFETs have firmly established their position in radio-electronic equipment for transmitting, receiving, and processing information, as well as in power electronics products, due to their significant advantages in terms of energy and thermal parameters. At the same time, issues associated with ensuring their reliability are no less acute than for devices based on other semiconductor materials. The aim of the study is to review the thermal and mechanical mechanisms of degradation in GaN HFETs due to the physicochemical characteristics of the materials used, as well as their corresponding growth and post-growth processes. Methods for preventing or reducing these mechanisms during development, production, and operation are evaluated.

Methods. The main research method consists in an analytical review of the results of publications by a wide range of specialists in the field of semiconductor physics, production technology of heteroepitaxial structures and active devices based on them, as well as the modeling and design of modules and equipment in terms of their reliable operation.

Results. As well as describing the problems of GaN HFET quality degradation caused by thermal overheating, mechanical degradation, problems with hot electrons and phonons in gallium nitride, the article provides an overview of research into these phenomena and methods for reducing their impact on transistor technical parameters and quality indicators.

Conclusions. The results of the study show that strong electric fields and high specific thermal loading of high-power GaN HFETs can cause physical, polarization, piezoelectric and thermal phenomena that lead to redistribution of mechanical stresses in the active region, degradation of electrical characteristics, and a decrease in the reliability of the transistor as a whole. It is shown that the presence of a field-plate and a passivating SiN layer leads to a decrease in the values of mechanical stress in the gate area by 1.3–1.5 times. The effects of thermal degradation in class AB amplifiers are more pronounced than the effects of strong fields in class E amplifiers; moreover, the mean time to failure sharply decreases at GaN HFET active zone temperatures over 320–350°C.

Keywords: GaN HFET, heterostructure, dual-channel HFET, coupled-channel HFET, current, self-heating, thermal conductivity, degradation, doping

• **Submitted:** 14.05.2024 • **Revised:** 12.07.2024 • **Accepted:** 11.02.2025

For citation: Minnebaev V.M. Thermal and mechanical degradation mechanisms in heterostructural field-effect transistors based on gallium nitride. *Russian Technological Journal.* 2025;13(2):57–73. https://doi.org/10.32362/2500-316X-2025-13-2-57-73, https://elibrary.ru/TTUFNR

Financial disclosure: The author has no financial or proprietary interest in any material or method mentioned.

The author declares no conflicts of interest.

ОБЗОР

Тепловые и механические механизмы деградаций в гетероструктурных полевых транзисторах на нитриде галлия

В.М. Миннебаев @

АО «Микроволновые системы», Москва, 105122 Россия
@ *Автор для переписки, e-mail: vm@mwsystems.ru*

Резюме

Цели. Гетероструктурные полевые транзисторы на нитриде галлия (GaN HFET, heterostructural field-effect transistor) являются наиболее перспективными полупроводниковыми устройствами для силовой и сверхвысокочастотной электроники. За последние 10–15 лет GaN HFET прочно заняли место в аппаратуре радиоэлектронных средств передачи, приема и обработки информации, а также в изделиях силовой электроники за счет существенных преимуществ в энергетических и тепловых параметрах. При этом вопросы обеспечения их долговременной надежности стоят не менее остро, чем для приборов на других полупроводниковых материалах. Целью исследования является обзор тепловых и механических механизмов деградаций в GaN HFET, обусловленных физико-химическими особенностями применяемых материалов, ростовыми и пост-ростовыми процессами, и способов купирования этих механизмов при разработке, производстве и эксплуатации.

Методы. Основным методом исследования является аналитический обзор результатов публикаций широкого круга специалистов в области физики полупроводников, технологии производства гетероэпитаксиальных структур и активных приборов на их основе, моделирования и проектирования модулей и аппаратуры, надежности и эксплуатации.

Результаты. Описаны причины снижения показателей качества GaN HFET, вызываемые тепловыми перегревами, механическими деградациями, проблемами с горячими электронами и фононами в нитриде галлия, а также представлен обзор исследований, посвященных этим явлениям и методам снижения их воздействия на технические параметры транзисторов и показатели качества.

Выводы. По итогам исследования отмечено, что сильные электрические поля и высокая удельная тепловая нагруженность мощных GaN HFET вызывают физические, поляризационные, пьезоэлектрические и тепловые явления, способные приводить к перераспределению механических напряжений в активной области, деградации электрических характеристик и снижению надежности транзистора в целом. Установлено, что наличие полевой платы и пассивирующего слоя из нитрида кремния SiN приводят к снижению значений механических напряжений в области затвора в 1.3–1.5 раз, эффекты тепловой деградации в усилителях класса AB выражены сильнее, чем эффекты воздействия сильных полей в усилителях класса E, при температуре активной зоны GaN HFET более 320–350 °C резко снижается время средней наработки до отказа.

Ключевые слова: GaN HFET, гетероструктура, двухканальный HFET, HFET со связанными каналами, ток, саморазогрев, теплопроводность, деградация, легирование

• **Поступила:** 14.05.2024 • **Доработана:** 12.07.2024 • **Принята к опубликованию:** 11.02.2025

Для цитирования: Миннебаев В.М. Тепловые и механические механизмы деградаций в гетероструктурных полевых транзисторах на нитриде галлия. *Russian Technological Journal*. 2025;13(2):57–73. https://doi.org/10.32362/2500-316X-2025-13-2-57-73, https://elibrary.ru/TTUFNR

Прозрачность финансовой деятельности: Автор не имеет финансовой заинтересованности в представленных материалах или методах.

Автор заявляет об отсутствии конфликта интересов.

INTRODUCTION

The achievements of recent years in the development of power and high-power microwave devices are mainly associated with III-nitride materials and the various devices based on them [1–3]. Heterostructured field-effect transistors (HFET) on gallium nitride (GaN) are the most promising for high-power microwave and power applications due to the sufficiently high mobility of electrons, high density of charge carriers, and high breakdown voltages, which is especially evident when operating in pulsed modes [4].

The idea of charge accumulation at the interface of a heterojunction first proposed in the late 1960s led to the possibility of creating an amplifying device on this basis. However, it was only after the development of high-quality high-precision epitaxial growth methods in the 1970s that the task of transforming a field-effect transistor (FET) into a heterostructural field-effect transistor (HFET)—also known as a high electron mobility transistor (HEMT)—could be solved [5].

In 1978, the achievement of high electron mobility obtained by modulating doping was demonstrated for the first time; in 1980, the University of Illinois demonstrated a device called a modulated-doped field-effect transistor (MODFET). During the following decades, the efforts of engineers and scientists from different countries led to more complex devices such as double heterojunction field-effect transistors [6]. The simplest GaN/AlGaN HFET structure incorporating aluminum–gallium nitride AlGaN is shown in Fig. 1 [7]. AlN/AlGaN/GaN/AlGaN/GaN/AlGaN multilayer heterostructures have become the basis for a new component base of solid-state microwave electronics.

The above-described structure is not the only possible one for GaN HFETs. For example, in InAlN/AlN/GaN devices incorporating indium-aluminum nitride (InAlN), a different structure is used for this purpose. Although both structures exhibit polarization, the AlGaN variety has stronger piezoelectric polarization, whose effect can be enhanced due to structural characteristics. For example, the piezoelectric effect is stronger in GaN/AlGaN than in InAlN/AlN due to the grid mismatch between the layers [8]. Here, the spontaneous polarization for the InAlN/AlN/GaN structure plays

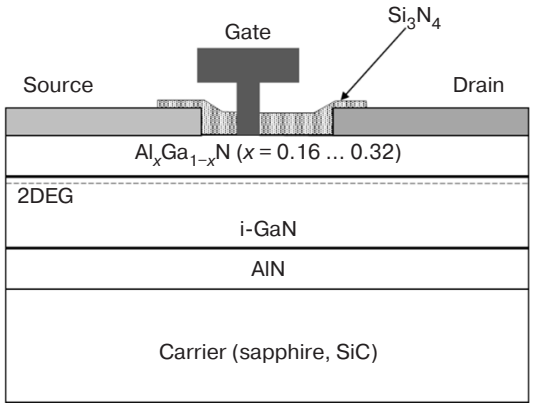


Fig. 1. The simplest structure of GaN/AlGaN HFET. 2DEG is two-dimensional electron gas; AlN is a buffer layer of aluminum nitride minimizing differences in step parameters of the chip lattice of heterostructure and carrier; SiC is a silicon carbide carrier

the only role except when the aluminum nitride (AlN) interface layer is used [9].

HFETs can be categorized into three main types depending on the GaN/AlGaN structure (Fig. 2). The typical structure, which consists of a single two-dimensional electron gas (2DEG) channel, is classical only because other HFET structures are often compared to it when evaluating the improvement of properties.

A typical HFET structure starts with a layer of AlN grown on a carrier (Al_2O_3, SiC, Si) followed by a thicker GaN buffer layer. The reason for using GaN on AlN (grid mismatch) is due to the first buffer layers serving as semi-isolating layers. After GaN growth, the growth of the AlN separating layer continues. Since the high frequency performance of AlGaN/GaN HFET is degraded due to the transfer of high-energy electrons from GaN to the AlGaN barrier, the AlN layer needs to be grown in between them to prevent high-energy electrons from moving to the AlGaN layer, thus keeping the electrons in GaN and creating a high-density 2DEG. However, this layer should not be thick (typically about 2 nm) due to the grid mismatch between GaN and AlN, which causes strain relaxation and cracking [10]. This layer is used to better confine the 2DEG channel due to the wider bandgap due to the penetration of electrons into the barrier material actually changing the effective

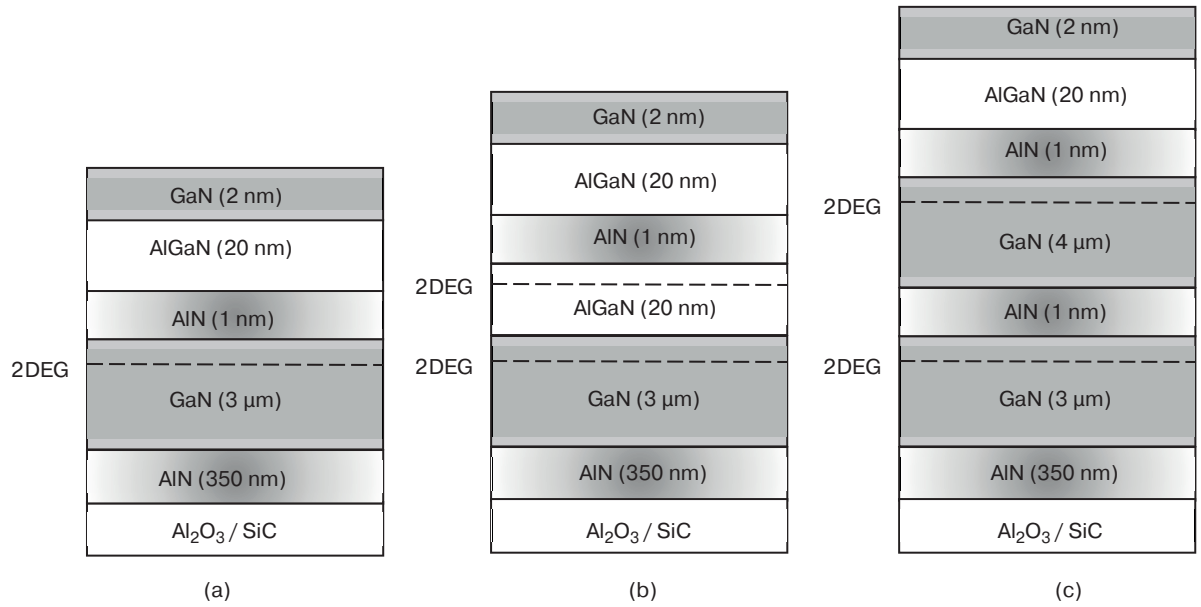


Fig. 2. Conventional schematic representation of GaN HFET heterostructure: (a) single-channel (classical), (b) dual-channel, (c) coupled-channel

mass and electron scattering rate. AlGaN is then grown in such a way as to force electrons to form a channel in GaN that relies on the polarization difference between it and the GaN layer below. A final 2-nm thick GaN layer is used to protect the AlGaN from oxidation and provide a better metallic contact to GaN than to AlGaN. The corresponding structure is depicted in Fig. 2a.

Another type of structure used for HFETs is the dual-channel HFET (Fig. 2b), which has a higher electron density. The third type of HFET structure, shown in Fig. 2c, is the coupled-channel HFET. In this structure, unlike the two-channel structure, the two channels are at the same energy level, so they can be coupled to form a channel called a three-dimensional electron gas or 3DEG.

1. SOURCES AND MECHANISMS OF GaN HFET FAILURES

Due to the strong electric fields present in GaN HFETs, as well as the interaction of thermal, physical, polarization fluxes, etc., there are multiple degradation paths in HFETs caused by overheating, resulting in a decrease in specific current, as well as an increase in gate leakage current and reliability limitations. Power dissipation associated with hot electron-hot phonon interactions can be caused by several mechanisms. A number of papers have investigated these degradation mechanisms and described approaches for reducing their deleterious effects. Depending on the operating time of a semiconductor device, there are three main failure periods (Fig. 3) [11]. As can be seen, the presence of device failure does not depend on the time of its operation; thus, designers strive to exclude failures at stages *I* and *II*, as well as to maximize their reduction at the aging period.

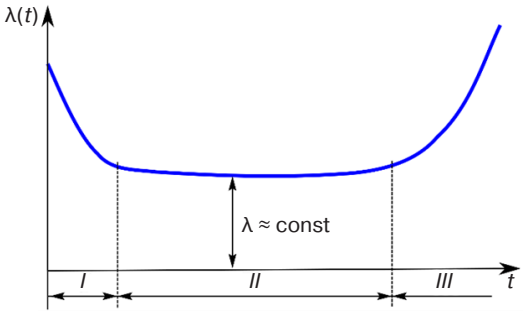


Fig. 3. Typical dependence of failure intensity λ on operation time *t*: *I* is a period of running-in and failures of low-quality products; *II* is a period of normal operation (failure intensity is approximately constant); *III* is a period of aging (failures are caused by wear and/or aging of materials)

In order to describe the above-mentioned degradations, let us define their mechanisms. Although it is rather difficult to classify all degradation mechanisms, for better understanding we will divide them into three main groups: electrical, thermal and mechanical. These mechanisms and their interrelation are shown in the so-called "magic triangle" of interactions that demonstrates the connections between electrical, mechanical, and thermal interactions (Fig. 4) [7].

We will discuss the location of the main sources of transistor failures typically manifested during operation (Fig. 5) on the example of the simplest GaN/AlGaN HFET structure.

From Figs. 4 and 5, it can be seen that the failure mechanisms in GaN HFETs are closely related. Here, failure sources *1* and *4* are peculiar to GaN devices due to the presence of spontaneous and piezoelectric polarization fields in AlGaN/GaN heterostructures,

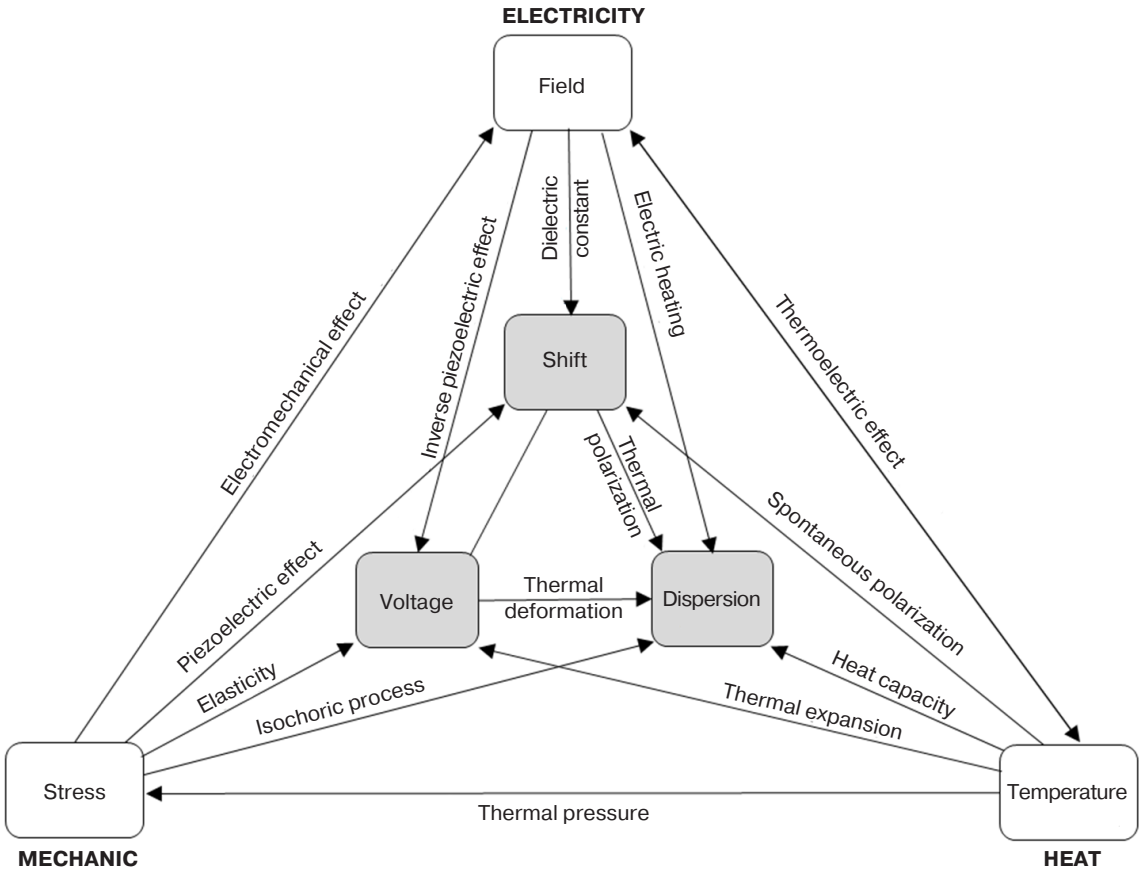


Fig. 4. "Magic triangle" of interactions

while sources *2* and *3* are caused by the presence of hot electrons occurring at high supply voltage levels, and sources *5–8* are activated by temperature increase [13].

The issues related to reliability improvement and approaches to solving them require understanding of GaN HFET degradation mechanisms, which can pose a serious problem due to the peculiarities of GaN device physics and imperfections of initial materials, as well as growth and post-growth processes of device fabrication. It was noted in [14–18] that some degradation effects occur even when the devices are in the off state (without bias voltages applied) or during double-state Schottky gate biases [14–18]. In this case, the most significant manifestation of degradation is a catastrophic increase in gate current leakage. The existence of a critical voltage above which degradation of GaN HFET parameters occurs led to the proposal of a degradation mechanism based on the formation of defects due to the inverse piezoelectric effect [13]. Failure mechanisms *5–8* refer to heat-activated degradation mechanisms, which were previously also observed in devices built on other semiconductor materials (Si, GaAs, InP, SiC, etc.). This suggests that these failure mechanisms are more related to metallization technologies and materials rather than to gallium nitride itself [13], but can be more pronounced in the latter due to the peculiarities of growth and post-growth technologies.

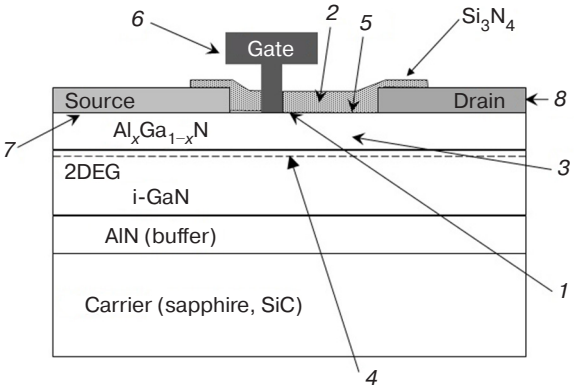


Fig. 5. Main identified failure mechanisms [12]:
(*1*) electric field causing degradation of the gate edge and pre-existing temperature defects;
(*2*) trapping of electrons in the passivation layer;
(*3*) generation of traps by hot electrons;
(*4*) generation of traps due to electrothermomechanical damage (temperature; electric field; mechanical deformation);
(*5*) thermally induced delamination of passivation;
(*6*) degradation of metal connection with the gate due to electrothermomechanical damage caused by traps;
(*7*) degradation of interlayer connections;
(*8*) degradation of ohmic contacts

Electrical mechanisms of GaN HFET degradation are discussed in detail in [19]. Now let us proceed to the analysis of thermal and mechanical failure mechanisms, as well as their interrelation with electrical mechanisms.

2. THERMAL MECHANISMS OF GaN HFET DEGRADATION

A. The problem of self-heating

When the chip grid of a semiconductor material cannot fully dissipate the heat generated by hot electrons through the emission of hot LO phonons[1], this excess heat accumulates in the structure, interacts with hotter electrons and causes even more heat, while there is no effective dissipation mechanism. Consequently, the temperature of the device rises, resulting in degraded device performance. This mechanism, called self-heating, is an important problem for GaN HFETs operating at high currents and powers.

In [20], the self-heating phenomenon of AlGaN/GaN HFETs was investigated using sapphire or SiC as a substrate. It is observed that in AlGaN/GaN HFETs grown on 6H-SiC substrates, the allowable maximum power dissipation is at least 3 times higher than that of those grown on sapphire under the same conditions. This is a result of the higher thermal conductivity of 6H-SiC compared to sapphire. However, the problem with using SiC as a carrier is its high cost. In order to understand better the effect of self-heating, we present the simulation results of temperature change in AlGaN/GaN HFETs with different geometry, heterostructure design, doping density, and substrate type obtained in [21].

In order to calculate the temperature rise, the authors started with the nonlinear flow equation and continued the simulation in doped and undoped AlGaN/GaN HFET channels on SiC. The structure used for the simulation and the simulation result are shown in Figs. 6 and 7, respectively [20]. It can be seen from the figures that the sample with undoped GaN 2DEG channel layer dissipates heat to a significantly lower extent compared to the sample with doped GaN channel layer.

It is known from solid state physics that there are two mechanisms that contribute to heat conduction. Thermal conduction can result from vibration of grid nodes as well as electronic conduction. The grid contribution to the thermal conductivity of pure chip s is defined by the expression:

$$k_{\text{grid}}(T) = \frac{1}{3} V_s C_{\text{grid}}(T) L(T), \qquad (1)$$

where T is the temperature, V_s is the speed of sound in the semiconductor, $C_{\text{grid}}(T)$ is the grid heat capacity, and $L(T)$ is the average phonon free path length.

[1] LO Phonon is a longitudinal optical phonon in semiconductor chips.

The contribution of electronic conduction to thermal conductivity is negligible at doping concentrations less than 10^{19} cm^{-3}. On the other hand, since penetrating dislocations decrease V_s and increase phonon scattering, the thermal conductivity of material decreases due to an increase in the dislocation density (GaN as a pure material has a much higher thermal conductivity compared to epitaxial GaN layers). At the same time, the increase in phonon scattering due to the increase in doping concentration dominates over the increase in the contribution of electronic conduction.

Consequently, the thermal conductivity decreases by increasing the doping concentration (k_{grid} decreases by a factor of about 2 for each decade of increasing n concentration), which is consistent with the findings of [20, 22].

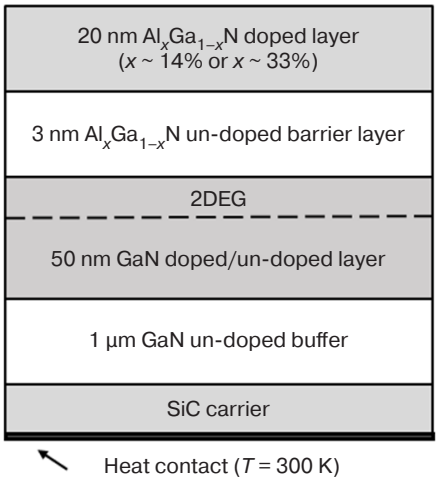


Fig. 6. Structure of the doped/undoped HFET used in [20]

In [21], it is shown that with a greater increase in temperature, the electron mobility begins to decrease. In addition, when the size of the transistor decreases, the negative effect of doping becomes even more pronounced due to an increase in the density of dislocations and an increase in the relative number of defects due to size reduction.

B. Degradations associated with heat exposure

In order to identify a specific device degradation effect, it is necessary to define the test conditions in such a way that other degradation mechanisms do not affect the results obtained. When the gate is reverse biased (the transistor is "locked"), the drain current is very small and therefore, as the temperature increases, it can be assumed that there are no other effects in the channel than the applied thermal energy. In this way, the degradation effects caused by thermal effects can be monitored. By applying this test condition to GaN HFETs, it is possible

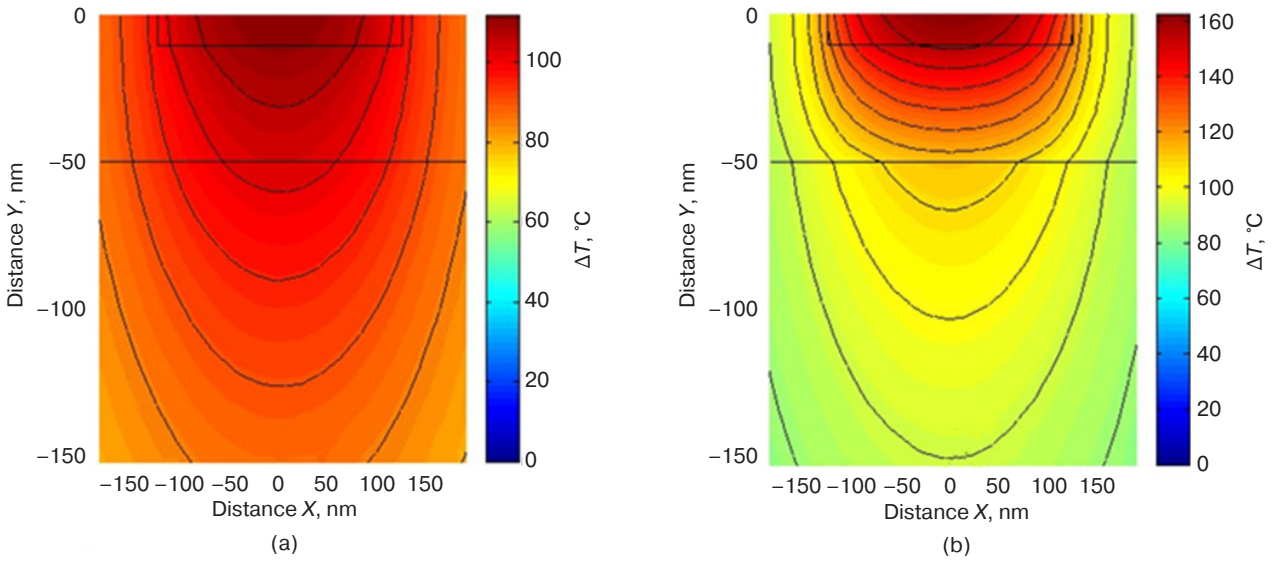


Fig. 7. Temperature rise profile in undoped (a) and doped (b) AlGaN/GaN HFET channels grown on SiC carrier [20]

to understand how thermal energy is dissipated in the channel.

Thermal scattering is related to the thermal conductivity of the material. It is known that the more acoustic phonon modes are occupied in the structure of a semiconductor, the greater its thermal conductivity. It was shown in [23] that the group velocity of acoustic phonons is much higher than that of optical phonons. Consequently, at low temperatures, optical phonon modes are not occupied, but only acoustic phonon modes are occupied. However, this is only true for small electric fields, which is usually not the case in high-power HFETs unless measurements are made at very small drain and gate biases. It can also be said that according to (1), at low temperatures the free path length L is relatively large and is dominated by the finite chip size (size effect), the number of defects (negligible in the case of a pure chip) and the thermal conductivity of the grid $C_{\text{grid}}(T) \sim \left(\frac{T}{\theta_D}\right)$, where θ_D is the Debye temperature. With increasing temperature, the thermal conductivity of the grid $C_{\text{grid}}(T)$ first begins to saturate, and at very high temperatures the thermal conductivity drops due to phonon–phonon and phonon–electron scattering processes [20] (Fig. 8).

In the study [24], various research methods such as Raman micro-thermography, micro-photoluminescence spectroscopy and thermal modeling have been applied to better understand the thermal properties of AlGaN/GaN HFETs. It is confirmed that the thermal conductivity is higher in bulk GaN than epitaxial layers due to the fact that the bulk material has lower dislocation density. At the same time, if the quality of bulk GaN is good enough, the epitaxial layers will also be of higher quality, other things being equal. It is proved that the thermal resistance in GaN-on-SiC is very close to that of GaN-on-GaN, although GaN has a slightly lower thermal conductivity of C_{GaN} ~ 260 W/(m·K) for bulk GaN compared to C_{SiC} ~ 480 W/(m·K) for SiC. The reason may be due to the lack of thermal boundary resistance between the device layers and the substrate [24]. Figure 9 shows the surface temperature and depth profile at the center of a 30 × 80 μm AlGaN/GaN HFET calculated in a 3D simulator and measured by photoluminescence and Raman micro-spectroscopy. Using Fig. 9 and the assumption of zero thermal boundary resistance (TBR_{eff}) at the GaN–GaN interface and uniform GaN thermal conductivity, as well as approximating the measured curve using the 3D-T thermal model with a standard thermal conductivity temperature dependence $T^{-1.22}$,

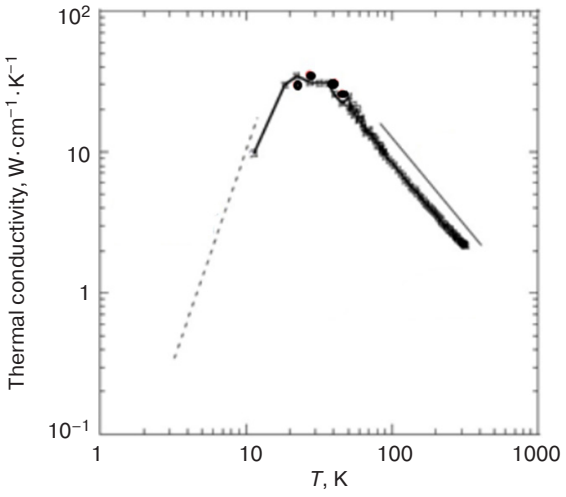


Fig. 8. Temperature dependence of thermal conductivity of GaN HFET with gate width W_g = 200 μm: dashed line—scattering limit at phonon free path length of 500 μm due to size effect, solid line—dependence $T^{-1.22}$ [20]

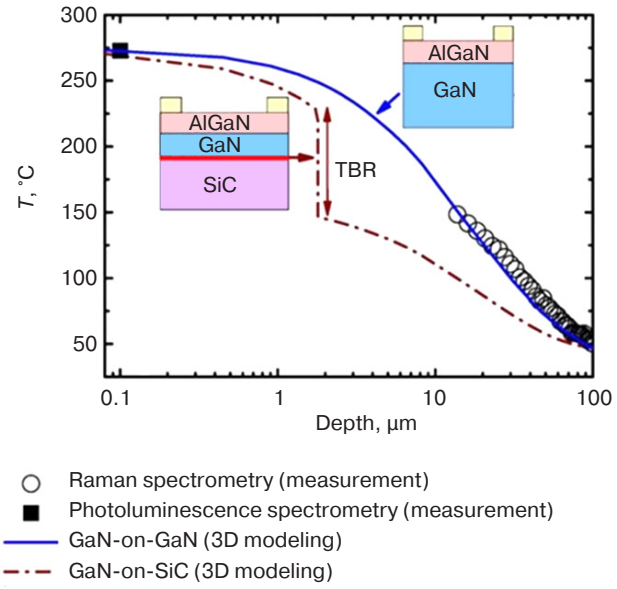


Fig. 9. Dependence of temperature T on the surface on the substrate thickness [24]

we obtain a C_{GaN} thermal conductivity C_{GaN} ~ 260 W/(m·K), in contrast to the thin epitaxial layer having a value C_{epiGaN} ~ 150 W/(m·K).

In [25], a developed ML-TCAD[2] coupled electronic model is presented, which enables prediction of the gain reduction and efficiency of GaN HEMTs induced by hot electron effects and does not require knowledge of reliability physics and results of long-term experimental tests. The resulting model predicts with high reliability the results of drain current variation in GaN HEMTs under the effect of hot electron voltage. In addition, the model allows us to determine the location, distribution, concentrations, and energy levels of traps in GaN HEMT from the current–voltage degradation curves, which is certainly useful for further studying the physical degradation mechanism of GaN HEMT under hot electron exposure.

C. Degradation of ohmic contacts and passivation coatings

AlGaN/GaN HFETs use ohmic contacts with standard gold metallization, which seems to guarantee sufficient stability under elevated temperature tests up to a certain limit. In [24], GaN HFETs with Ti/Al/Pt/Au ohmic contacts were subjected to step voltage for 48 h. It was found that the ohmic contacts begin to degrade at transition temperatures above 300°C—self-heating of the transistor leads to degradation of performances of devices and blocks due to degradation of the ohmic contacts [13, 24].

In [26], the degradation of AlGaN/GaN/SiC HFETs with 25 μm gate length and different passivation coatings related to the temperature regime of transistor operation was investigated. It was shown that the threshold voltage degradation starts in the temperature range of 310–330°C. Changes in the structure of the transistor were analyzed by measuring electroluminescence and using transmission electron microscopy (TEM). The formation of voids and gold diffusion in AlGaN/GaN were detected. These processes are responsible for device degradation with conventional passivation techniques.

The temperature increase of the active region and the transistor chip itself depends, among other things, on the choice of the operating point and the level of input microwave power. Figure 10 shows the thermal distribution over the chip surface in the input-to-output cross section of an AlGaN/GaN/SiC HFET with overall dimensions of 480 × 800 × 100 μm, gate length L_G = 25 μm, and gate width W_G = 6 × 200 μm. The tests were performed at a case base temperature of 120°C.

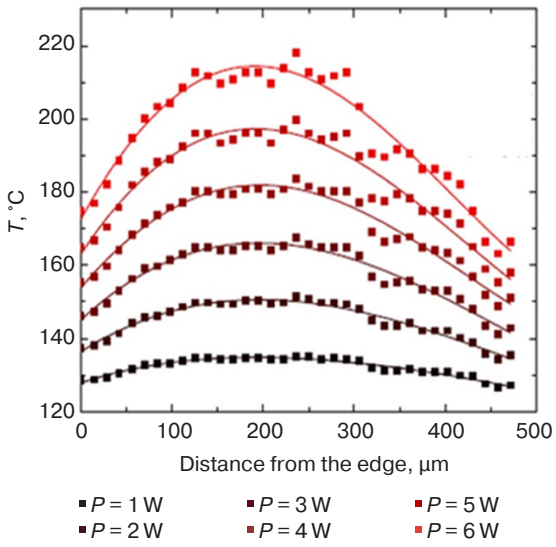


Fig. 10. Temperature distribution T on the chip surface at different power levels of the input microwave signal [26]

It can be seen that when the power load P is increased, the temperature distribution over the transistor area changes dramatically. The resulting temperature nonuniformity will certainly be the cause of subsequent transistor failures. Additional studies confirm that the temperature increase of the chip surface depends both on the case temperature and power dissipated (Fig. 11).

It is found that when exposed to elevated temperature, the gate current increases and the gate threshold voltage shifts to the negative side, and this is due to the passivation layers.

[2] ML-TCAD is an electronic model of a transistor created on machine learning (ML) basis to significantly speed up calculations in the TCAD (Technology Computer-Aided Design) environment by minimizing physical calculations.

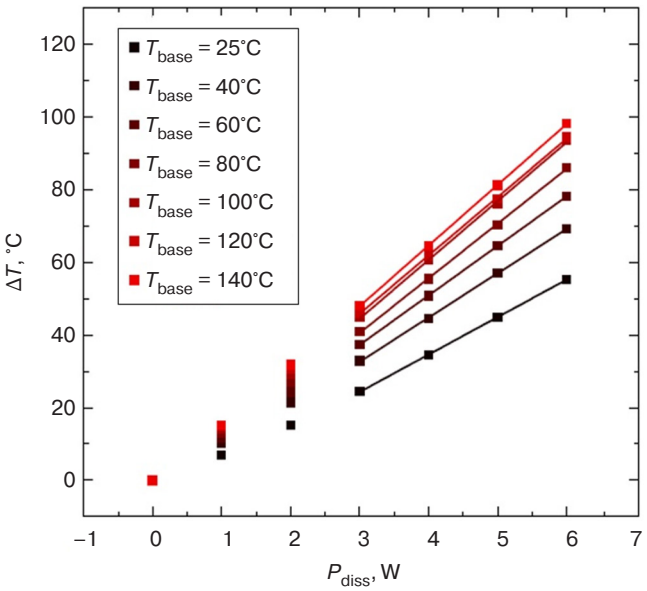


Fig. 11. Dependence of transistor surface temperature increment ΔT on power dissipation P_{diss} at different base temperatures [26]

In [27], the results of investigation of microscopic origin of the vulnerability of GaN HFET materials and devices based on them to high temperatures by monitoring the onset of structural degradation at different temperature conditions in real time are presented. The studies have been carried out by means of SEM. Electron-transparent samples were fabricated from bulk material and heated up to 800°C. High-resolution transmission electron microscopy, scanning transmission electron microscopy, energy-dispersive X-ray spectroscopy, and geometric phase analysis (GPA) were performed to assess the quality of chips, to study the diffusion of materials and the processes of strain propagation in the sample before and after heating. It was observed that the decrease of the gate contact area is noticeable starting from the temperature 470°C, and it is accompanied by Ni/Au mixing near the gate/AlGaN interface. Elevated temperatures cause significant out-of-plane lattice expansion at the SiNx/GaN/AlGaN interface, as shown by GPA strain maps with geometric phase, while in-plane strain remains relatively constant. In this study, it is shown that exposure to temperatures exceeding 500°C leads to 2 orders of magnitude increase in leakage current in GaN HFETs. The results of this study provide real-time visual information to determine the initial location of degradation and highlight the effect of temperature on GaN HFET structure, its electrical properties, and material degradation.

D. Failure and service life testing

Thermal effects are one of the major problems that reduce the performance and reliability of a semiconductor device. It is the most common mechanism because AlGaN/GaN HFETs mainly operate at relatively high temperatures.

Device reliability is a very important issue. Nowadays, every company has separate departments that focus on the quality and reliability of their devices. For any manufactured devices, it is required to determine the area of safe operation, the average time until failure, and the shelf life of devices and appliances. Therefore, the industry pays great attention to the reliability of its products—samples are subjected to numerous tests, including short-term and long-term failure and survivability. As you can understand from the name, these tests are conducted to evaluate the service life of the device. Since you cannot wait 10 or 25 years to see what happens to a semiconductor device, special conditions are applied to conduct relatively short time tests to evaluate the mean time to failure (MTTF).

These accelerated life tests are basically performed at three different temperatures and for each one the MTTF is measured. By extrapolating these values to the temperature at which the device is operating (with the device junction temperature being higher than the case temperature), the MTTF for the device can be obtained. Figure 12 shows how the MTTF is evaluated in a failure test: the MTTF is measured at three junction temperatures of 260, 285, and 310°C, and by extrapolation it is determined that at an operating junction temperature of 150°C the MTTF is more than 10^7 h, activation energy E_a = 2.0. Usually, the minimum possible number of devices is tested, but in such a way that the measurements do not lose accuracy [7].

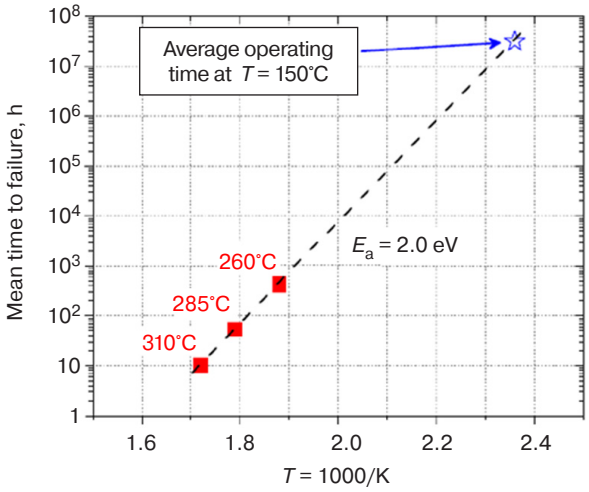


Fig. 12. Determination of MTTF by measurement at three temperatures

At the permissible operating temperature of the *p-n*-junction *Tj* of the active zone of the GaN HFET chip, equal to 200°C, the average MTTF is 10^5 h (11.57 years). The device resistance to load mismatch up to a voltage standing wave factor of 10 in the large-signal mode is also demonstrated.

Rapid (within several hours) destruction occurs in modern GaN heterotransistors at junction temperatures of 320–350°C [13].

In [28], the degradation effects observed in GaN HFETs with a gate length of 0.15 μm were experimentally investigated under real power amplifier conditions, i.e., when a high-level microwave power is applied to the input. The measurements were performed for a series of devices in the loud-pull measurement mode. Consequently, this mode of experimentation provides information relevant to microwave signal operation and enables preferentially detecting changes in electrical quantities that cannot be directly detected at current–voltage curve or high frequencies. Values such as gate resistance now play a fundamental role in reliability analysis of technologies. Experiments were performed on GaN HFETs with the same gate width while operating in class AB mode, a saturation mode that emphasizes degradation effects caused by high temperatures due to increased power dissipation, and in class E mode, where degradation is enhanced by strong electric fields. Experiments were conducted at $T_1 = 23$°C and $T_2 = 100$°C. As a result, it is found that:

- GaN HFET degradation level depends on the actual radio frequency mode;
- thermal degradation effects, which are enhanced in the class AB mode, are more pronounced than the effects of strong fields in the class E mode;
- characterization of GaN HFETs under real microwave loads should be used to accurately and deeply investigate degradation and failure mechanisms in order to determine MTTF in practical microwave applications.

3. MECHANICAL MECHANISMS OF GaN HFET DEGRADATION

A. Inverse piezoelectric effect

Mechanical stresses are also important as part of the triangle of interactions (Fig. 4). Mechanical stresses have an important influence on the parameters and reliability of microwave GaN HFETs. Gallium nitride is a polar material, i.e., valence electrons are not distributed between two neighbors uniformly—this causes local polarization of the semiconductor chip. However, globally the polarization vector of GaN is zero. Consequently, when mechanical pressure is applied, the chip structure can bend or expand (depending on the direction of the force) and cause a resultant polarization that can act as an electric field. This phenomenon is called the piezoelectric effect.

On the other hand, if we apply an electric field to this chip, the chip can again bend or expand (depending on the direction of the electric field). In this case, the electric field induces a mechanical force—this effect is called the inverse piezoelectric field. When the gate–source voltage $U_{g.s.}$ is applied, the inverse bias of the gate channel becomes more and more depleted. If too large an electric field is applied in the direction opposite to the relaxation direction of the chip, it will induce a large mechanical force and may cause mechanical damage to the chip. Such an effect is what is called the inverse piezoelectric effect. This is what happens to GaN when too large a voltage $U_{g.s.}$ is applied.

Fabrication processes and operating conditions also affect mechanical stresses. In [29], in particular, the relationship between mechanical stresses and reliability of gallium nitride heterotransistors is discussed. In this work, $Al_{0.2}Ga_{0.8}N$(20 nm)/GaN(2 μm) structures on a 75-μm thick carrier were investigated. The calculations show that a 100-nm thick SiN passivation layer creates a mechanical stress of up to 300 MPa at the gate junction [13]. Tensile stresses are critical from the point of view of transistor reliability due to the fact that they contribute to the formation of "pits" on the surface of the heterostructure. Figure 13 shows the calculated value of mechanical biaxial stresses in the gate region as a function of specific power dissipation $P_{diss.sp.}$ for two GaN HFET designs:

- traditional—without SiN passivation coating and field-plate;
- improved—in the presence of SiN and field-plate, usually used to increase breakdown voltages in HFETs.

The given data show that the presence of the field board and SiN layer leads to a 1.3–1.5-fold reduction of mechanical stresses in the gate region depending on the specific power dissipation. Of course, mechanical stresses depend on both the substrate temperature and the temperature of the *p-n*-junction.

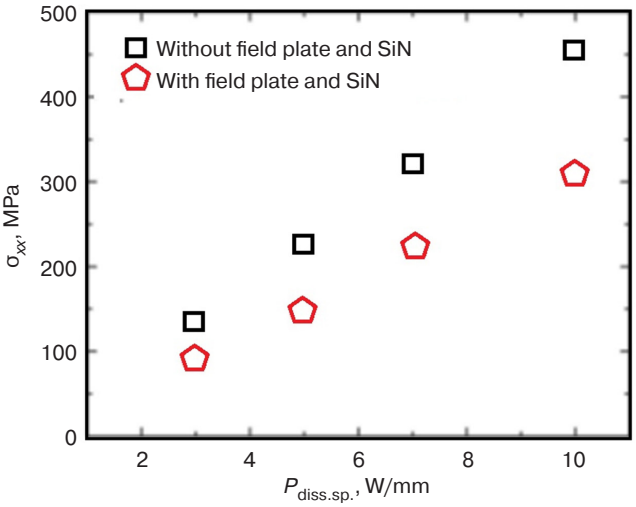


Fig. 13. Dependence of the magnitude of mechanical biaxial stresses σ_{xx} on the specific power dissipation $P_{diss.sp.}$ in GaN HFET [13]

In [29], calculations of the values of mechanical biaxial stresses σ_{xx} in GaN HFETs arising due to the inverse piezoelectric effect (Fig. 14), the consequence of which is the appearance of the electric field. For comparison, Fig. 14 outlines the area corresponding to the values of the electric field strength arising in the transistor at voltages ($V_{d.s.}$) between the drain and source $U_{d.s.}$ = 50–70 V.

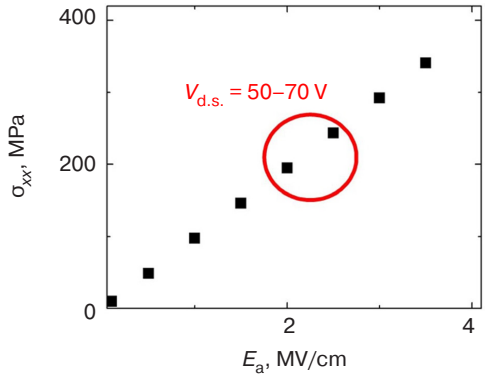


Fig. 14. Dependence of the magnitude of mechanical biaxial stresses σ_{xx} on the electric field strength in GaN HFET [29]

It follows from the data given in [29] that the values of the above mechanical stresses caused by the power dissipated in the transistor are comparable in magnitude to the stresses due to the inverse piezoeffect.

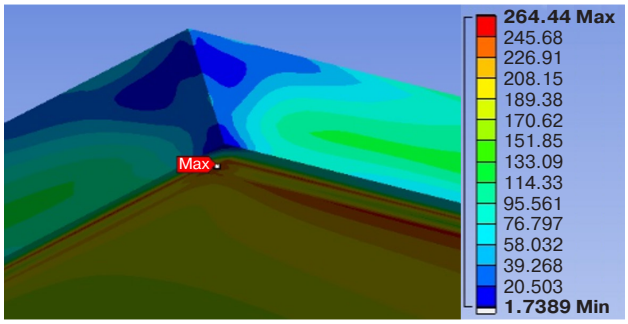
Ohmic contacts also create mechanical stresses. Thus, mechanical stresses in GaN heterotransistors are one of the reasons of their reliability decrease.

B. Interrelationship of thermal and mechanical degradation

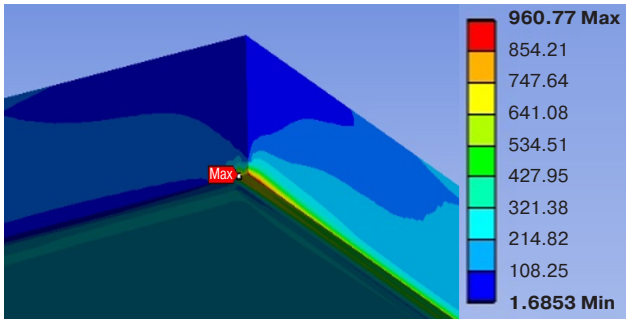
It was shown in [30] that the high specific thermal loading of high-power AlGaN/GaN/SiC transistors requires a particularly careful approach to heat dissipation in operating modes. Preventive measures to eliminate any assembly defects in high-power AlGaN/GaN/GiC transistors are essential. For example, the defect of vertical tilting of the chip after soldering it to the base leads to unequal thickness of the solder layer at the periphery between the chip and the base. The evaluation of stresses arising in the chip during heating depending on different variants of the chip arrangement in space, as well as their influence on the potential reliability of the chip structure are shown in [30], where the values and nature of the distribution of mechanical stresses in the chip were determined by calculating the stress–strain state of the chip model with a defect of displacement along the face or corner by the finite element method. The edge displacement in this case is a uniform increase/decrease in solder thickness between two parallel faces of the transistor chip, and the corner displacement is a change in solder thickness along the diagonal of the soldering plane.

Figure 15 shows the three-dimensional distribution of the main thermal stresses in the SiC layer. Based on the obtained values of equivalent stresses for the variants of chip position, the safety factor was calculated. The safety factor determines how much the actually designed structure can withstand the induced thermal stresses.

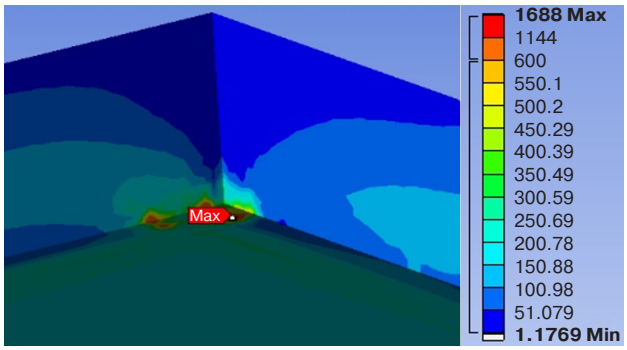
In [30], it is shown that the chip reliability of AlGaN/GaN/SiC transistors deteriorates by a factor of 6.5 at the maximum angle tilt and by a factor of 3.6 at the maximum edge tilt. Thus, it is shown that the displacement of the chip plane significantly increases mechanical stresses in the chip body in areas with thinning solder layer, which, in turn, means the potential development of mechanical failures of the structure, especially under cyclic thermal loads.



(a)



(b)



(c)

Fig. 15. Distribution of main stresses in the chip under continuous power dissipation for the case of:
(a) uniform constant solder thickness,
(b) nonconstant thickness along the face,
(c) nonconstant thickness along the angle [30]

C. Relationship between electrical and mechanical damage

Degradation may be reversible or irreversible. If the product returns to its normal state after the end of stress impact, we can talk about reversible degradation. However, sometimes after the end of degradation the device is irreversibly changed, which is interpreted as damage or irreversible degradation.

Electrical degradation is characterized by the value of critical applied voltage below which degradation is reversible. Application of voltages higher than the critical voltage causes irreversible degradation. In [31], significant crystallographic damage induced by strong electromagnetic fields is shown, as well as the correlation between electrical degradation (sudden drop in current consumption, current collapse, increase in gate leakage current, avalanche injection, etc.) and material degradation as a mechanical effect. To find the correlation between electrical and physical damage, the depth and width of pits for different samples are measured from images obtained by transmission electron microscopy. Then, a degradation plot of the percentage of saturation current $I_{d.s.sat}$ between drain and source and collapse current $I_{d.s.coll}$ values as a function of the depth of the defect region is plotted to obtain a quantitative comparison between electrical and mechanical damage. This is shown in Fig. 16 [32] and implies that these mechanisms are interrelated and crystallographic damage is responsible for electrical degradation.

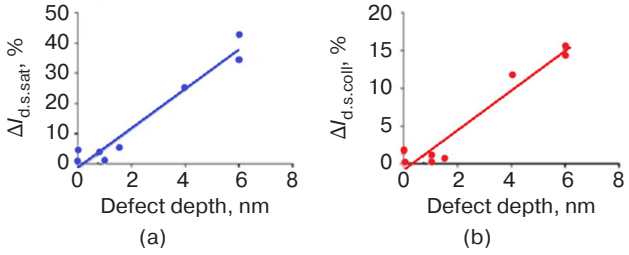


Fig. 16. Correlation between the values of $I_{d.s.sat}$, $I_{d.s.coll}$ and the depth of defects [32]

Figure 17 schematically shows the mechanical stress distribution in the gate region when a negative voltage is applied to the gate with respect to the drain and source [33].

When the transistor operates in pulsed mode, the mechanical stresses shown in Fig. 17 increase and decrease. While amplifiers based on microwave GaN HFETs should operate for 15–20 years, billions or even trillions of such cycles occur in the transistor. Eventually, a crack occurs in the gate region on the drain side. This is due to the fact that the tensile mechanical stresses on the drain side relative to the gate in operating mode when voltage is applied to the drain are greater than on the source side. In the initial stage of the process,

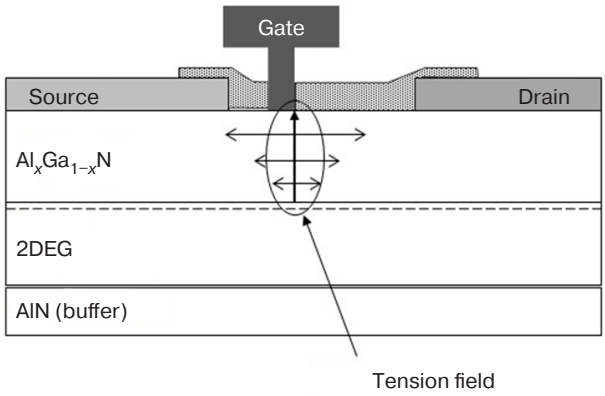
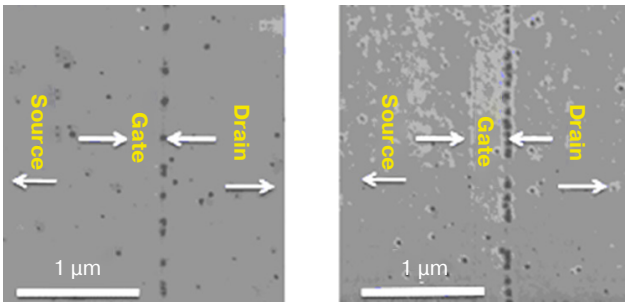


Fig. 17. Stretching regions of the AlGaN layer resulting from the inverse piezoelectric effect

defects in the form of pits are formed—less deep and rare on the source side and deeper and more frequent on the drain side. Figure 18 shows the time evolution of the process [34].



(a) (b)

Рис. 18. Evolution of crack formation from pitted defects in time [34]:
(a) after 10 min, (b) after 1000 min

Jimenez investigated the degradation mechanisms of a power amplifier in the form of a 3-stage microwave monolithic integral circuit (MMIC) W-band based on gallium nitride under the influence of an input microwave signal of high-power level[3]. The same experiments were performed on a discrete transistor with a gate periphery equal to the gate periphery of the MIC output stage. The studies did not reveal any shift in the threshold voltage (U_{thres}) after exposure to a high-power microwave signal; however, the modeled dynamic load lines showed that the output voltage fluctuations exceeded the breakdown voltage ($U_{d.s.break}$) when exposed to an input microwave signal with a high power level. Thus, it can be concluded from the experiment that the inverse piezoelectric effect will be the main factor of performance degradation, leading to defects and dislocations in the chip on the drain side.

[3] Jimenez J. *Advanced Reliability Aspects of GaN FETs*. Presented at European Microwave Week (EuMW), 2010.

Tsao et al. [35] confirms that the rather rapid development of gallium nitride-based power amplifiers, focused on high output power and efficiency, has created a critical problem for temperature control of devices in general. As a result of the thermal design and analysis of transistors and MMIC based on thermal maps measured by an infrared camera in continuous and pulsed modes of operation, it was found that the thermal resistance R_t "junction-to-case GaN-chip" at DC operation is 1.63°/W, and in pulsed mode R_t = 1.05°/W. Thus, it has been experimentally proved that to ensure the required reliability performance in the design of GaN MIC, only a thermal model reliably confirmed by functional testing should be used.

The presence of dislocation pits on the surface of the initial substrate leads to the subsequent formation of cracks [33]. Fine chemical treatment of its surface leads to the elimination or, at least, to the reduction of the formation of pitting. Foreign impurities (contaminants) also stimulate the degradation process. The double top layer of GaN ("cap") over the AlGaN barrier reduces the probability of subsequent degradation. These phenomena, which reduce the reliability of GaN HFETs, were taken into account in the developments of Cree[4] (USA) and UMS[5] (Germany) when developing microwave GaN HFETs for space applications [36].

The European Social Innovation Competition report (European high-quality GaN wafers on SiC substrates for space applications) reflects the following[6]:

- in order to improve the reliability and reduce the stresses shown in Fig. 17, the aluminum concentration in the barrier layer was reduced; the 22-nm thick barrier layer consists of an Al(16%)Ga(84%)N barrier and a 3-nm thick top protective GaN layer;
- layer carrier concentration in the 2DEG channel was slightly below $6 \cdot 10^{12}$ cm^{-2} (Cree) and below $3 \cdot 10^{12}$ cm^{-2} (SiCrystal). These values are typical values for HEMT structures with 18% Al (Cree) and 16% Al (SiCrystal) at 22-nm AlGaN barrier layer, respectively.
- average curvature value is 10.4 μm, average bend value is 4.96 μm.

It was shown in [37] that ultraviolet (UV) illumination very strongly reduces the breakdown voltage in GaN-on-Si, but this effect is negligible for GaN-on-SiC. Considering that the quality of the grown material is quite good (which may not be the case, especially for GaN-on-SiC), it can be assumed that UV illumination causes electrons to split and fall into traps due to the action of a field caused by the inverse piezoelectric effect. Therefore, this effect is more pronounced in GaN-on-Si due to the larger number of traps. In addition, this is another of the effects that confirm the correlation between the magnitude of the breakdown voltage of GaN HFETs and the number of traps in GaN, in other words, between electrical degradation and physical changes.

CONCLUSIONS

The strong electric fields present in GaN HFETs result in thermal, physical and polarization phenomena that degrade active element performance in the form of reduced specific drain current and increase gate leakage current and threshold voltage offset, as well as reducing breakdown voltages and specific output power and leading to lower reliability in general.

The redistribution of mechanical stresses in the chip due to the high specific thermal loading of high-power AlGaN/GaN transistors and consequent inverse piezoelectric effect results in mechanical damage of the structure and reduced reliability. However, the presence of the field board and SiN layer can reduce mechanical stress values in the gate region by 1.3–1.5 times depending on the specific power dissipated.

The degradation of GaN HFETs depends on the actual power supply modes, input power level, and class of operation of the amplifier. In this case, the thermal degradation effects that are amplified in class AB mode are more pronounced than the effects of strong fields in class E mode. Therefore, in order to determine the MTTF in practical microwave applications, GaN HFETs should be characterized under the conditions of actual supply voltages, microwave power, and the class of operation of the power amplifier.

A sharp decrease in MTTF and rapid (within a few hours) destruction occurs in contemporary GaN-heterotransistors at junction temperatures greater than 320–350°C.

While GaN HFETs have outperformed other semiconductors currently used in industry (such as GaAs, Si, InP, etc.) in terms of their technical and performance characteristics, it is necessary to take into account physical limitations in their design and fabrication to eliminate the possibility of degradation in operation due to the presence of the described degradation mechanisms, as well as to control the thermal and electrical modes of GaN HFETs in operation.

[4] www.wolfspeed.com. Accessed January 12, 2024.

[5] www.ums-rf.com. Accessed January 12, 2024.

[6] Final Report Summary – EUSIC (High Quality European GaN-Wafer on SiC Substrates for Space Applications). https://cordis.europa.eu/project/id/242360/reporting/pl.Accessed January 12, 2024.

REFERENCES

1. Akinin V.E., Borisov O.V., Ivanov K.A., Kolkovskiy Yu.V., Minnebaev V.M., Redka Al.V. 350-Watt solid-state amplifier of X-band frequencies with air cooling. *Nanoindustriya = Nanoindustry.* 2020;13;S4(99):465–467 (in Russ.). https://doi.org/10.22184/1993-8578.2020.13.4s.465.467
2. Belolipeckiy A.V., Borisov O.V., Kolkovsky Yu.V., Lega1 G.V., Minnebaev V.M., Redka Al.V., Redka An.V. Electronic antenna unit for X-band space application AESA. *Elektronnaya tekhnika. Seriya 2: Poluprovodnikovye pribory = Electronic Engineering. Series 2. Semiconductor Devices.* 2017;3(246):15–25 (in Russ.).
3. Borisov O.V., Zubkov A.V., Ivanov K.A., Minnebaev V.M., Redka A.V. X-band 70-W GaN broadband power amplifier. *Elektronnaya tekhnika. Seriya 2: Poluprovodnikovye pribory = Electronic Engineering. Series 2. Semiconductor Devices.* 2014;2(233):4–9 (in Russ.).
4. Abolduyev I.M., Garber G.Z., Zubkov A.V., Ivanov K.A., Kolkovsky Yu.V., Minnebaev V.M., Redka A.V., Ushakov A.V. The pulse mode operation of the microwave power AlGaN/GaN HFE. *Elektronnaya tekhnika. Seriya 2: Poluprovodnikovye pribory = Electronic Engineering. Series 2. Semiconductor Devices.* 2012;1(228):48–53 (in Russ.).
5. Ghovanloo M. *Dual-Heterojunction High Electron Mobility Transistors on GaAs Substrate.* University of Michigan. Ann Arbor MI 48109-2122. 2008. 18 p.
6. Hamaguchi C., Miyatsuji K., Hihara H. Proposal of single quantum well transistor (SQWT) self-consistent calculations of 2D electrons in a quantum well with external voltage. *Jpn. J. Appl. Phys. Part 2.* 1984;23(3):132–134. https://doi.org/10.1143/JJAP.23.L132
7. Morkoc H. *Handbook of Nitride Semiconductors and Devices.* V. 3. *GaN-based Optical and Electronic Devices*. Wiley-VCH Verlag GmbH & Co. 2008. 902 p. http://doi.org/10.1002/9783527628445
8. Butte R., Carlin J.-F., Feltin E., Gonschorek M., Nicolay S., Christmann G., Simeonov D., Castiglia A., Dorsaz J., Buehlmann H.J., Christopoulos S., von Hogersthal B.H., Grundy G.A.J.D., Mosca M., Pinquier C., Py M.A., Demangeot F., Frandon J., Lagoudakis P.G., Baumberg J.J., Grandjean N. Current status of AlInN layers lattice-matched to GaN for photonics and electronics. *J. Phys. D: Appl. Phys.* 2007;40(20):6328–6344. https://doi.org/10.1088/0022-3727/40/20/S16
9. Ramonas M., Matulionis A., Liberis J., Eastman L.F., Chen X., Sun Y.-J. Hot-phonon effect on power dissipation in a biased AlGaN/AlN/GaN channel. *Phys. Rev. B.* 2005;71(7):075324. https://doi.org/10.1103/PhysRevB.71.075324
10. Kasahara K., Miyamoto N., Ando Y., Okamoto Y., Nakayama T., Kuzuhara M. Ka-band 2.3W power AlGaN–GaN heterojunction FET. *IEDM Tech. Dig*. 2002:667–680. http://doi.org/10.1109/IEDM.2002.1175929
11. Polovko A.M. *Osnovy teorii nadezhnosti* (*Fundamentals of Reliability Theory*). Moscow: Nauka; 1964. 446 p. (in Russ.).
12. Meneghesso G., Meneghini M., Tazzoli A., et al. Reliability issues of Gallium Nitride High Electron Mobility Transistors. *Int. J. Microw. Wirel. Technol*. 2010;2(1):39–50. https://doi.org/10.1017/S1759078710000097
13. Kolkovskiy Yu.V, Kontcevoi Yu.A. Problems of reliability of GaN microwave heterotransistors. (Review). *Elektronnaya tekhnika. Seriya 2: Poluprovodnikovye pribory = Electronic Engineering. Series 2. Semiconductor Devices.* 2022;4(267):27–41 (in Russ.). https://elibrary.ru/kacktk
14. Joh J., del Alamo J.A. Critical voltage for electrical degradation of GaN high electron mobility transistors. *IEEE Elect. Device Lett*. 2008;29(4):287–289. https://doi.org/10.1109/LED.2008.917815
15. Joh J., del Alamo J.A. Mechanisms for electrical degradation of GaN high-electron mobility transistors. In: *Proc. of the IEEE Int. Elect. Device Meeting* (*IEDM*), *Tech. Dig.* 2006. P. 415–418. https://doi.org/10.1109/IEDM.2006.346799
16. Joh J., Xia L., del Alamo J.A. Gate current degradation mechanisms of GaN high electron mobility transistors. In: *Proc. of the IEEE Int. Elect. Device Meeting* (*IEDM*). 2007. P. 385–388. http://doi.org/10.1109/IEDM.2007.4418953
17. Meneghesso G., Verzellesi G., Danesin F., et al. Reliability of GaN high-electron-mobility transistors: state of the art and perspectives. *IEEE Trans. Device Mater. Reliabil.* 2008;8(2):332–343. https://doi.org/10.1109/TDMR.2008.923743
18. Zanoni E., Meneghesso G., Verzellesi G., et al. A review of failure modes and mechanisms of GaN-based HEMTs. In: *Proc. of the IEEE Int. Elect. Device Meeting* (*IEDM*). 2007. P. 381–384. https://doi.org/10.1109/IEDM.2007.4418952
19. Minnebaev V.M. Electrical Degradation of GaN heterostructure field-effect transistors. *Elektronnaya tekhnika. Seriya 2: Poluprovodnikovye pribory = Electronic Engineering. Series 2. Semiconductor Devices.* 2021;3(262):4–24 (in Russ.). https://elibrary.ru/catpkn
20. Charles Kittel. *Introduction to Solid State Physics.* 8th ed. N.Y.: John Wiley & Sons Inc.; 2005. 703 p.
21. Filippov K.A., Balandin A.A. Self-Heating Effects in GaN/AlGaN Heterostructure Field-Effect Transistors and Device Structure Optimization. In: TechConnet. Briefs. V. 3. *Proceed. of the ACRS Nanotech. Conf.* 2003;3:333–336.
22. Manoi A., Pomeroy J.W., Killat N., Kuball M. Benchmarking of thermal boundary resistance in AlGaN/GaN HEMTs on SiC substrates: Implications of the nucleation layer microstructure. *IEEE Elect. Device Lett.* 2010;31(12):1395–1397. https://doi.org/10.1109/LED.2010.2077730
23. Killat N., Montes M., Pomeroy J.W., et al. Thermal Properties of AlGaN/GaN HFETs on Bulk GaN Substrates. *IEEE Elect. Device Lett.* 2012;33(3):366–368. https://doi.org/10.1109/LED.2011.2179972
24. Rampazzo F., Pierobon R., Pacetta D., et al. Hot carrier aging degradation phenomena in GaN based MESFETs. *Microelectron. Reliability*. 2004;44(9-11):1375–1380. https://doi.org/10.1016/j.microrel.2004.07.017
25. Wang K., Jiang H., Liao Y., Xu Y., Yan F., Ji X. Degradation Prediction of GaN HEMTs under Hot-Electron Stress Based on ML-TCAD Approach. *Electronics.* 2022;11(21):3582. https://doi.org/10.3390/electronics11213582

26. Stopel A., Khramtsov A., Katz O., et al. Direct monitoring of hot-carrier accumulated charge in GaN HEMT and PHEMT devices. *Proc. Of the Int. Conf. on GaAs Manufact. Technol.* New Orleans. 2005;14–19. Available from URL: https://cris.tau.ac.il/en/publications/direct-monitoring-of-hot-carrier-accumulated-charge-in-gan-hemt-a
27. Rasel M.A.J., Zhang D., Chen A., Thomas M., House S.D., Kuo W., Watt J., Islam A., Glavin N., Smyth M., Haque A., Wolfe D.E., Pearton S.J. Temperature-Induced Degradation of GaN HEMT: An *In situ* Heating Study. *J. Vacuum Sci. Technol. B*. May 2024;42(3):032209. https://doi.org/10.1116/6.0003490
28. Bosi G., Raffo A., Vadalà V., Giofrè R., Crupi G., Vannini G. A Thorough Evaluation of GaN HEMT Degradation under Realistic Power Amplifier Operation. *Electronics.* 2023;12(13):2939. https://doi.org/10.3390/electronics12132939
29. Dammann M., Baeumler M., Brückner P., et al. Degradation of 0.25 µm GaN HEMTs under high temperature stress test. *Microelectron. Reliability.* 2015;55(9–10):1667–1671. https://doi.org/10.1016/j.microrel.2015.06.042
30. Joglekar A., Lian C., Baskaran R., et al. Finite Element Analysis of Fabrication- and Operation-Induced Mechanical Stress in AlGaN/GaN Transistors. *IEEE Trans. Semiconduct. Manufact*. 2016;29(4):349–354. https://doi.org/10.1109/TSM.2016.2600593
31. Klimov A.O. Thermomechanical response study of the FET crystal changing its vertical orientation in Solder. *Elektronnaya tekhnika. Seriya 2: Poluprovodnikovye pribory = Electronic Engineering. Series 2. Semiconductor Devices.* 2019;2(253):64–71 (in Russ.).
32. Joh J., del Alamo J.A., Langworthy K., Xie S., Zheleva T. Role of stress voltage on structural degradation of GaN high-electron-mobility transistors. *Microelectron. Reliability*. 2011;51(2):201–206. https://doi.org/10.1016/j.microrel.2010.08.021
33. Morkoc H. *Handbook of Nitride Semiconductors and Devices.* V. 3. *Materials Properties, Physics and Growth*. Weinheim: Wiley-VCH Verlag GmbH & Co; 2008. 850 p. ISBN 978-3-527-40838-2
34. Ancona M.G., Binari S.C., Meyer D.J. Fully coupled thermoelectromechanical analysis of GaN high electron mobility transistor degradation. *J. Appl. Phys.* 2012;111(7):074504. https://doi.org/10.1063/1.3698492
35. Tsao Y.-F., Wang Y., Chiu P.-H., Hsu H.-T. Reliability Assessment of 60-GHz GaN Power Amplifier Under High-Level Input RF Stress. *IEEE Trans. Elect. Devices.* 2024;71(7):4087–4092. https://doi.org/10.1109/TED.2024.3397634
36. Han Y., Tang G., Lau B.L. Thermal Characterization and Management of GaN-on-SiC High Power Amplifier MMIC. In: *IEEE 73rd Electronic Components and Technology Conference* (*ECTC*). IEEE; 2023. P. 1989–1993. http://doi.org/10.1109/ectc51909.2023.00342
37. Demirtas S., del Alamo J.A. Effect of Trapping on the Critical Voltage for Degradation in GaN High Electron Mobility Transistors. In: *IEEE International Reliability Physics Symposium* (*IRPS*). IEEE; 2010. P. 134–138. https://doi.org/10.1109/IRPS.2010.5488838

СПИСОК ЛИТЕРАТУРЫ

1. Акинин В.Е., Борисов О.В., Иванов К.А., Колковский Ю.В., Миннебаев В.М., Редька Ал.В. 350-Ваттный твердотельный усилитель мощности X-диапазона частот с воздушным охлаждением. *Наноиндустрия.* 2020;13;S4(99):465–467. https://doi.org/10.22184/1993-8578.2020.13.4s.465.467
2. Белолипецкий А.В., Борисов О.В., Колковский Ю.В., Легай Г.В., Миннебаев В.М., Редька А.В., Редька А.В. Антенный электронный блок для спутниковой АФАР X-диапазона. *Электронная техника. Серия 2: Полупроводниковые приборы.* 2017;3(246):15–25.
3. Борисов О.В., Зубков А.М., Иванов К.А., Миннебаев В.М., Редька А.В. Широкополосный 70-ваттный GaN усилитель мощности X-диапазона. *Электронная техника. Серия 2: Полупроводниковые приборы.* 2014;2(233):4–9.
4. Аболдуев И.М., Гарбер Г.З., Зубков А.М., Иванов К.А., Колковский Ю.В., Миннебаев В.М., Редька А.В., Ушаков А.В. Импульсный режим работы мощных СВЧ гетерополевых AlGaN/GaN транзисторов. *Электронная техника. Серия 2: Полупроводниковые приборы.* 2012;1(228):48–53.
5. Ghovanloo M. *Dual-Heterojunction High Electron Mobility Transistors on GaAs Substrate.* University of Michigan. Ann Arbor MI 48109-2122. 2008. 18 p.
6. Hamaguchi C., Miyatsuji K., Hihara H. Proposal of single quantum well transistor (SQWT) self-consistent calculations of 2D electrons in a quantum well with external voltage. *Jpn. J. Appl. Phys.* 1984;23(3):132–134. https://doi.org/10.1143/JJAP.23.L132
7. Morkoc H. *Handbook of Nitride Semiconductors and Devices.* V. 3. *GaN-based Optical and Electronic Devices*. Wiley-VCH Verlag GmbH & Co. 2008. 902 p. http://doi.org/10.1002/9783527628445
8. Butte R., Carlin J.-F., Feltin E., Gonschorek M., Nicolay S., Christmann G., Simeonov D., Castiglia A., Dorsaz J., Buehlmann H.J., Christopoulos S., von Hogersthal B.H., Grundy G.A.J.D., Mosca M., Pinquier C., Py M.A., Demangeot F., Frandon J., Lagoudakis P.G., Baumberg J.J., Grandjean N. Current status of AlInN layers lattice-matched to GaN for photonics and electronics. *J. Phys. D: Appl. Phys.* 2007;40(20):6328–6344. https://doi.org/10.1088/0022-3727/40/20/S16
9. Ramonas M., Matulionis A., Liberis J., Eastman L.F., Chen X., Sun Y.-J. Hot-phonon effect on power dissipation in a biased AlGaN/AlN/GaN channel. *Phys. Rev. B.* 2005;71(7):075324. https://doi.org/10.1103/PhysRevB.71.075324
10. Kasahara K., Miyamoto N., Ando Y., Okamoto Y., Nakayama T., Kuzuhara M. Ka-band 2.3W power AlGaN–GaN heterojunction FET. *IEDM Tech. Dig*. 2002:667–680. http://doi.org/10.1109/IEDM.2002.1175929

11. Половко А.М. *Основы теории надежности*. М.: Наука; 1964. 446 с.
12. Meneghesso G., Meneghini M., Tazzoli A., et al. Reliability issues of Gallium Nitride High Electron Mobility Transistors. *Int. J. Microw. Wirel. Technol*. 2010;2(1):39–50. https://doi.org/10.1017/S1759078710000097
13. Колковский Ю.В., Концевой Ю.А. Проблемы надежности GaN СВЧ гетеротранзисторов. Обзор. *Электронная техника. Серия 2. Полупроводниковые приборы.* 2022;4(267);27–41. https://elibrary.ru/kacktk
14. Joh J., del Alamo J.A. Critical voltage for electrical degradation of GaN high electron mobility transistors. *IEEE Elect. Device Lett*. 2008;29(4):287–289. https://doi.org/10.1109/LED.2008.917815
15. Joh J., del Alamo J.A. Mechanisms for electrical degradation of GaN high-electron mobility transistors. In: *Proc. of the IEEE Int. Elect. Device Meeting* (*IEDM*), *Tech. Dig.* 2006. P. 415–418. https://doi.org/10.1109/IEDM.2006.346799
16. Joh J., Xia L., del Alamo J.A. Gate current degradation mechanisms of GaN high electron mobility transistors. In: *Proc. of the IEEE Int. Elect. Device Meeting* (*IEDM*). 2007. P. 385–388. http://doi.org/10.1109/IEDM.2007.4418953
17. Meneghesso G., Verzellesi G., Danesin F., et al. Reliability of GaN high-electron-mobility transistors: state of the art and perspectives. *IEEE Trans. Device Mater. Reliabil.* 2008;8(2):332–343. https://doi.org/10.1109/TDMR.2008.923743
18. Zanoni E., Meneghesso G., Verzellesi G., et al. A review of failure modes and mechanisms of GaN-based HEMTs. In: *Proc. of the IEEE Int. Elect. Device Meeting* (*IEDM*). 2007. P. 381–384. https://doi.org/10.1109/IEDM.2007.4418952
19. Миннебаев В.М. Электрические механизмы деградации полевых гетероструктурных транзисторов на основе нитрида галлия. *Электронная техника. Серия 2. Полупроводниковые приборы*. 2021;3(262):4–24. https://elibrary.ru/catpkn
20. Kittel Ch. *Introduction to Solid State Physics.* 8th ed. N.Y.: John Wiley & Sons Inc.; 2005. 703 p.
21. Filippov K.A., Balandin A.A. Self-Heating Effects in GaN/AlGaN Heterostructure Field-Effect Transistors and Device Structure Optimization. In: TechConnet. Briefs. V. 3. *Proceed. of the ACRS Nanotech. Conf.* 2003;3:333–336.
22. Manoi A., Pomeroy J.W., Killat N., Kuball M. Benchmarking of thermal boundary resistance in AlGaN/GaN HEMTs on SiC substrates: Implications of the nucleation layer microstructure. *IEEE Elect. Device Lett.* 2010;31(12):1395–1397. https://doi.org/10.1109/LED.2010.2077730
23. Killat N., Montes M., Pomeroy J.W., et al. Thermal Properties of AlGaN/GaN HFETs on Bulk GaN Substrates. *IEEE Elect. Device Lett.* 2012;33(3):366–368. https://doi.org/10.1109/LED.2011.2179972
24. Rampazzo F., Pierobon R., Pacetta D., et al. Hot carrier aging degradation phenomena in GaN based MESFETs. *Microelectron. Reliability*. 2004;44(9–11):1375–1380. https://doi.org/10.1016/j.microrel.2004.07.017
25. Wang K., Jiang H., Liao Y., Xu Y., Yan F., Ji X. Degradation Prediction of GaN HEMTs under Hot-Electron Stress Based on ML-TCAD Approach. *Electronics.* 2022;11(21):3582. https://doi.org/10.3390/electronics11213582
26. Stopel A., Khramtsov A., Katz O., et al. Direct monitoring of hot-carrier accumulated charge in GaN HEMT and PHEMT devices. *Proc. Of the Int. Conf. on GaAs Manufact. Technol.* New Orleans. 2005;14–19. URL: https://cris.tau.ac.il/en/publications/direct-monitoring-of-hot-carrier-accumulated-charge-in-gan-hemt-a
27. Rasel M.A.J., Zhang D., Chen A., Thomas M., House S.D., Kuo W., Watt J., Islam A., Glavin N., Smyth M., Haque A., Wolfe D.E., Pearton S.J. Temperature-Induced Degradation of GaN HEMT: An *In situ* Heating Study. *J. Vacuum Sci. Technol. B*. May 2024;42(3):032209. https://doi.org/10.1116/6.0003490
28. Bosi G., Raffo A., Vadalà V., Giofrè R., Crupi G., Vannini G. A Thorough Evaluation of GaN HEMT Degradation under Realistic Power Amplifier Operation. *Electronics.* 2023;12(13):2939. https://doi.org/10.3390/electronics12132939
29. Dammann M., Baeumler M., Brückner P., et al. Degradation of 0.25 μm GaN HEMTs under high temperature stress test. *Microelectron. Reliability.* 2015;55(9–10):1667–1671. https://doi.org/10.1016/j.microrel.2015.06.042
30. Joglekar A., Lian C., Baskaran R., et al. Finite Element Analysis of Fabrication- and Operation-Induced Mechanical Stress in AlGaN/GaN Transistors. *IEEE Trans. Semiconduct. Manufact*. 2016;29(4):349–354. https://doi.org/10.1109/TSM.2016.2600593
31. Климов А.О. Исследование термомеханического отклика кристалла ПТБШ при изменении его вертикальной ориентации в припое. *Электронная техника. Серия 2. Полупроводниковые приборы*. 2019;2(253):64–71.
32. Joh J., del Alamo J.A., Langworthy K., Xie S., Zheleva T. Role of stress voltage on structural degradation of GaN high-electron-mobility transistors. *Microelectron. Reliability*. 2011;51(2):201–206. https://doi.org/10.1016/j.microrel.2010.08.021
33. Morkoc H. *Handbook of Nitride Semiconductors and Devices.* V. 3. *Materials Properties, Physics and Growth*. Weinheim: Wiley-VCH Verlag GmbH & Co; 2008. 850 p. ISBN 978-3-527-40838-2
34. Ancona M.G., Binari S.C., Meyer D.J. Fully coupled thermoelectromechanical analysis of GaN high electron mobility transistor degradation. *J. Appl. Phys.* 2012;111(7):074504. https://doi.org/10.1063/1.3698492
35. Tsao Y.-F., Wang Y., Chiu P.-H., Hsu H.-T. Reliability Assessment of 60-GHz GaN Power Amplifier Under High-Level Input RF Stress. *IEEE Trans. Elect. Devices.* 2024;71(7):4087–4092. https://doi.org/10.1109/TED.2024.3397634
36. Han Y., Tang G., Lau B.L Thermal Characterization and Management of GaN-on-SiC High Power Amplifier MMIC. In: *IEEE 73rd Electronic Components and Technology Conference* (*ECTC*). IEEE; 2023. P. 1989–1993. http://doi.org/10.1109/ectc51909.2023.00342
37. Demirtas S., del Alamo J.A. Effect of Trapping on the Critical Voltage for Degradation in GaN High Electron Mobility Transistors. In: *IEEE International Reliability Physics Symposium* (*IRPS*). IEEE; 2010. P. 134–138. https://doi.org/10.1109/IRPS.2010.5488838

About the author

Vadim M. Minnebaev, Cand. Sci. (Eng.), Assistant Professor, Deputy General Director on the Development of Electronic Components, Microwave Systems JSC (5-1, Shchelkovskoye sh., Moscow, 105122 Russia). E-mail: vm@mwsystems.ru. Scopus Author ID 6602931676, RSCI SPIN-code 8336-0490, https://orcid.org/0000-0002-3992-5196

Об авторе

Миннебаев Вадим Минхатович, к.т.н., доцент, заместитель генерального директора по развитию ЭКБ, АО «Микроволновые системы» (105122, Россия, Москва, Щёлковское шоссе, д. 5, стр. 1). E-mail: vm@mwsystems.ru. Scopus Author ID 6602931676, SPIN-код РИНЦ 8336-0490, https://orcid.org/0000-0002-3992-5196

Translated from Russian into English by L. Bychkova
Edited for English language and spelling by Thomas A. Beavitt

Mathematical modeling

Математическое моделирование

UDC 539.3, 621.762
https://doi.org/10.32362/2500-316X-2025-13-2-74-92
EDN KPQMIJ

RESEARCH ARTICLE

Mathematical modeling of hot isotatic pressing of tubes from powder materials

Vasiliy A. Goloveshkin [1, 2, @], Artem A. Nickolaenko [1], Victor N. Samarov [3], Gerard Raisson [4], Daria M. Fisunova [1]

[1] *MIREA – Russian Technological University, Moscow, 119454 Russia*
[2] *Institute of Applied Mechanics, Russian Academy of Sciences, Moscow, 125040 Russia*
[3] *Laboratory of new technologies, Moscow, 121352 Russia*
[4] *Clermond Ferrand, France*
[@] *Corresponding author, e-mail: vag-1953@yandex.ru*

Abstract

Objectives. The work set out to create a mathematical model to investigate the process of hot isostatic pressing (HIP) process of long tubes from powder materials in metal capsules. By analyzing the stress-strain state in the areas far from the top and bottom borders in the cylindrical system of coordinates, the axial strain rate at every moment of the process can be considered to be constant through the entire volume.

Methods. Mathematical modeling methods were used to describe mechanical properties in the process of HIP deformation by Green's model of porous compressible media. The HIP capsule material, which is considered to be non-compressible, is described by the ideal plasticity model. The temperature field is assumed to be uniform over the volume and constant during the time of deformation.

Results. The hypothesis of the uniform density over the cross section at each moment of the process was considered during analysis to the extent that the wall thickness of the tube is substantially less than its diameter. This hypothesis allowed us to reduce the task of determining the strain rates at every step of the process to a solution comprising two equations having two variables. When the strain rates are determined, the deformation field is built to obtain the final dimensions of the tube when the powder material is fully consolidated at the end of the HIP process.

Conclusions. The proposed model for describing the process hot isostatic pressing of long tubes from powder materials takes all the features of this process into account depending on the system parameters. The possibility of using tubular samples to determine the functions included in the Green's condition is demonstrated.

Keywords: mathematical modeling, plastically compressible media, Hot Isostatic Pressing, powder material, plastically irreversible compressible media, Green's plasticity criterion, ideal plasticity

• **Submitted:** 25.06.2024 • **Revised:** 29.08.2024 • **Accepted:** 06.02.2025

For citation: Goloveshkin V.A., Nickolaenko A.A., Samarov V.N., Raisson G., Fisunova D.M. Mathematical modeling of hot isotatic pressing of tubes from powder materials. *Russian Technological Journal.* 2025;13(2):74–92. https://doi.org/10.32362/2500-316X-2025-13-2-74-92, https://elibrary.ru/KPQMIJ

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Математическое моделирование процесса горячего изостатического прессования труб из порошковых материалов

В.А. Головешкин [1, 2, @], А.А. Николаенко [1], В.Н. Самаров [3], Ж. Рейссон [4], Д.М. Фисунова [1]

[1] *МИРЭА – Российский технологический университет, Москва, 119454 Россия*
[2] *Институт прикладной механики, Российская академия наук, Москва, 125040 Россия*
[3] *ООО «Лаборатория новых технологий», Москва, 121352 Россия*
[4] *Клермон Ферран, Франция*
[@] *Автор для переписки, e-mail: vag-1953@yandex.ru*

Резюме

Цели. Цель работы – создание модели, которая позволяет с помощью математического моделирования исследовать процесс горячего изостатического прессования (ГИП) длинных труб из порошковых материалов. Напряженно-деформируемое состояние исследуется вдали от верхней и нижней границ капсулы в цилиндрической системе координат, поэтому осевая скорость деформации в каждый момент процесса предполагается постоянной по объему.

Методы. Используются методы математического моделирования. Порошковый материал моделируется как пластически сжимаемая сплошная среда. Для описания его механических свойств в процессе деформации используется модель Грина. Для анализа механического поведения материала капсулы применяется модель идеальной пластичности при условии несжимаемости. Температурное поле предполагается постоянным по объему и по времени в течение всего процесса.

Результаты. Поскольку, как правило, толщина стенок труб существенно меньше их радиуса, то в процессе исследования принималась гипотеза о постоянстве относительной плотности порошкового материала по объему в каждый момент процесса. Принятая гипотеза позволила свести задачу определения скоростей деформаций на каждом шаге процесса к решению некоторой системы двух уравнений с двумя неизвестными. По известным скоростям деформации определяются скорости перемещений, что позволяет получить конечные размеры трубы (при относительной плотности порошкового материала равной единице). Анализируются усадки всех размеров трубы (вертикального, внутреннего радиуса, наружного радиуса), как функции относительной плотности.

Выводы. Предложенная модель описания процесса ГИП длинных труб из порошковых материалов позволяет учитывать все особенности данного процесса в зависимости от параметров системы. Показана возможность использования трубчатых образцов для определения функций, входящих в условие Грина.

Ключевые слова: математическое моделирование, пластически сжимаемая среда, горячее изостатическое прессование, порошковый материал, условие Грина, идеальная пластичность

• **Поступила:** 25.06.2024 • **Доработана:** 29.08.2024 • **Принята к опубликованию:** 06.02.2025

Для цитирования: Головешкин В.А., Николаенко А.А., Самаров В.Н., Рейссон Ж., Фисунова Д.М. Математическое моделирование процесса горячего изостатического прессования труб из порошковых материалов. *Russian Technological Journal*. 2025;13(2):74–92. https://doi.org/10.32362/2500-316X-2025-13-2-74-92, https://elibrary.ru/KPQMIJ

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

Hot isostatic pressing (HIP) is a process of high-temperature compaction (~1000°C) of powder materials under the action of external pressure (~1000 atm). Products manufactured using the HIP powder metallurgy process have high performance characteristics. However, it is precisely because of their high strength properties that their subsequent processing presents certain difficulties, requiring maximum precision at the manufacturing stage. As part of the HIP process, the powder material is placed in a metal container (capsule). This capsule is deformed together with the powder material until the latter is completely compacted, then removed chemically or mechanically. The function of the capsule is to induce the powder monolith to take the desired shape following the completion of the process. The mathematical modeling of the process sets out to design the capsule in such a way that the powder monolith takes the desired geometric shape after the end of the process.

Two main problems of mathematical modeling of the HIP process can be distinguished. Firstly, the HIP process is characterized by large deformations (initial powder density is about 65% of the monolith density). In mathematical terms, this means that the defining relations will be nonlinear with the boundary conditions set on a time-variable boundary. The second more fundamental problem consists in the difficulty of constructing the constitutive relations (under constitutive relations, we understand the relations defining the relationship between the stress tensor in the medium and the parameters characterizing the state of the medium). This problem is characteristic of all problems of mechanics of deformable solids that investigate their behavior beyond the elasticity limit. Since any defining relations will be approximate, any calculation will be approximate even if mathematical problems are excluded. Therefore, the actual powder manufacturing process must be an iterative process, a schema of which is outlined in [1]. The essence is as follows: a mathematical model is built, a capsule is designed on the basis of this model, and the product is manufactured. The product parameters are then compared with the required to provide a basis for refining the mathematical model. This method is some analog of the complex loading computer method proposed by Ilyushin in [2, 3]. A model satisfying the following requirements is considered to be an acceptable mathematical model of the HIP process:

1) it provides a close first approximation;
2) it correctly considers the influence of parameters;
3) it enables changes to be made to the model parameters based on the experimental results and, if necessary, additional parameters to be introduced.

Typically, 2–3 experimental iterations are required to put a product into production. At the present stage, the task of mathematical modeling is to consistently obtain the desired geometry at the second experimental iteration. For this purpose, experience shows that it is necessary to have an error of about 10% at the first iteration.

There are various approaches to describing the behavior of powder medium, some of which, for example [4], consider the medium as discrete. Within such approaches, when considering the interaction of individual particles, it is necessary to take into account the effects arising on the surface of their interaction [5, 6]. The use of such an approach requires the application of statistical methods [7]. More often, powder material is considered as a single continuum; thus, since we are interested in the kinematic aspects of behavior in the process of HIP as shown in [8–10], the kinematic aspects of the behavior of powder materials do not differ significantly from the behavior of continuous media.

The defining relations for powder materials have one essential difference from those used in classical theories of plasticity [11–14] due to the fact that these works proceed, as a rule, from small volume deformations or their equality to zero. For powder materials, the volume strain (or equivalent parameters: relative density, porosity) is an important parameter characterizing the state of the medium. It should be noted that it is the shear part of the strain tensor that is of real interest in describing the HIP process. Since the purpose of ISU is to obtain a monolithic product and the initial density can be determined with a high degree of accuracy, the volume component of the strain tensor can be considered as known. The time of the compaction process can be determined quite accurately (at known temperature and pressure) by compaction diagrams (Ashby diagram) [15–18]. Various models describing the behavior of plastically compressible media are presented in the works of Druyanov [19], Green [20], Shtern [21], Skorokhod [22].

The presence of the capsule causes its walls to “shield” the external pressure differently in different directions. This is particularly evident in the fabrication of pipes made of powder materials. In this process, there are actually three external boundaries: an outer and inner radial boundary, and another boundary at the ends. As shown in [23], under certain conditions this can lead to a looping motion of the inner wall. The peculiarities of the behavior of powder tubes at the initial stage of the process are investigated in [24]. The purpose of this work is to develop a model that for qualitatively assessing the change of the main parameters to determine the pipe geometry at the entire stage of the HIP process, as well as to establish the possibility of using pipe pressing experiments to determine the parameters characterizing the mechanical behavior of powder material.

1. MATHEMATICAL TASK STATEMENT

Let there be the following system in the cylindrical coordinate system (r, z): at $R_1 < r < R_2$ and $0 < z < H$, there is a capsule with perfectly plastic material with yield strength T_1; at $R_3 < r < R_4$ and $0 < z < H$, there is a capsule with perfectly plastic material with yield strength T_2; at $R_2 < r < R_3$, and $0 < z < H$, there is a powder material, whose behavior is described by the Green's model, with a monolithic yield strength Y. Here R_1, R_2, R_3, R_4, and H are the current geometrical dimensions of the pipe.

The elliptic Green's yield condition is used to describe the mechanical properties of powder material [20, 22]:

$$\frac{\sigma^2}{f_2^2} + \frac{s^2}{f_1^2} = Y^2, \tag{1.1}$$

where $\sigma = \sigma_{ij}/3$ is the first invariant of the stress tensor (σ_{ij} are the components of the stress tensor); s is the intensity of the deviator of the stress tensor, $s^2 = (3/2)s_{ij}s_{ij}$, $s_{ij} = \sigma_{ij} - \sigma\delta_{ij}$; indices i, j take integer values 1, 2, 3, with the r axis corresponding to index 1, φ axis—to index 2, z axis—to index 3; δ_{ij} is the Kronecker symbol ($\delta_{ij} = 0$ at $i \neq j$, $\delta_{ij} = 1$ at $i = j$); f_1 and f_2 are the relative density ρ functions known from experiment.

According to the flow law

$$\varepsilon_{ij} = \omega \frac{\partial \Phi}{\partial \sigma_{ij}}, \tag{1.2}$$

where ε_{ij} are the components of the strain rate tensor; $\Phi(\sigma_{ij}) = 0$ is the yield surface equation (1.1); ω is the proportionality factor determined at each point of space during the solution process.

The process is investigated away from the pipe ends. This allows us to assume that the strain rate ε_z is constant throughout the entire volume of the system. Then, by virtue of the assumption made and axial symmetry, $\varepsilon_{rz} = \varepsilon_{r\varphi} = \varepsilon_{z\varphi} = 0$, $\sigma_{rz} = \sigma_{r\varphi} = \sigma_{z\varphi} = 0$. Using (1.1), (1.2), taking into account the last remarks, we obtain:

$$\begin{aligned}
\varepsilon_r &= \omega/9\left[\left(2/f_2^2 + 18/f_1^2\right)\sigma_r + \left(2/f_2^2 - 9/f_1^2\right)\sigma_\varphi + \left(2/f_2^2 - 9/f_1^2\right)\sigma_z\right],\\
\varepsilon_\varphi &= \omega/9\left[\left(2/f_2^2 - 9/f_1^2\right)\sigma_r + \left(2/f_2^2 + 18/f_1^2\right)\sigma_\varphi + \left(2/f_2^2 - 9/f_1^2\right)\sigma_z\right],\\
\varepsilon_z &= \omega/9\left[\left(2/f_2^2 - 9/f_1^2\right)\sigma_r + \left(2/f_2^2 - 9/f_1^2\right)\sigma_\varphi + \left(2/f_2^2 + 18/f_1^2\right)\sigma_z\right].
\end{aligned} \tag{1.3}$$

Considering (1.3) as a system of equations with respect to stresses, we obtain

$$\begin{aligned}
\sigma_r &= 1/(18\omega)\left[\left(9f_2^2 + 4f_1^2\right)\varepsilon_r + \left(9f_2^2 - 2f_1^2\right)\varepsilon_\varphi + \left(9f_2^2 - 2f_1^2\right)\varepsilon_z\right],\\
\sigma_\varphi &= 1/(18\omega)\left[\left(9f_2^2 - 2f_1^2\right)\varepsilon_r + \left(9f_2^2 + 4f_1^2\right)\varepsilon_\varphi + \left(9f_2^2 - 2f_1^2\right)\varepsilon_z\right],\\
\sigma_z &= 1/(18\omega)\left[\left(9f_2^2 - 2f_1^2\right)\varepsilon_r + \left(9f_2^2 - 2f_1^2\right)\varepsilon_\varphi + \left(9f_2^2 + 4f_1^2\right)\varepsilon_z\right].
\end{aligned} \tag{1.4}$$

Substituting (1.4) into (1.1), we obtain:

$$\frac{1}{\omega} = \frac{6Y}{\sqrt{\left(9f_2^2 - 2f_1^2\right)\left(\varepsilon_r + \varepsilon_\varphi + \varepsilon_z\right)^2 + 6f_1^2\left(\varepsilon_r^2 + \varepsilon_\varphi^2 + \varepsilon_z^2\right)}}. \tag{1.5}$$

The power of internal forces in a unit volume w is determined by the relation

$$w = \sigma_{ij}\,\varepsilon_{ij}. \tag{1.6}$$

According to (1.4)–(1.6)

$$w = \frac{Y}{3}\sqrt{\left(9f_2^2 - 2f_1^2\right)\left(\varepsilon_r + \varepsilon_\varphi + \varepsilon_z\right)^2 + 6f_1^2\left(\varepsilon_r^2 + \varepsilon_\varphi^2 + \varepsilon_z^2\right)}. \tag{1.7}$$

The capsule material is assumed to be incompressible. Its behavior is described by the law of perfect plasticity.

$$s^2 = T^2, \tag{1.8}$$

where T is the yield strength.

According to the flow law (1.2) and the yield surface Eq. (1.8),

$$\varepsilon_r = \omega\left[2\sigma_r - \sigma_\varphi - \sigma_z\right],\ \varepsilon_\varphi = \omega\left[2\sigma_\varphi - \sigma_r - \sigma_z\right],\ \varepsilon_z = \omega\left[2\sigma_z - \sigma_\varphi - \sigma_r\right]. \tag{1.9}$$

We suppose

$$\sigma_r + \sigma_\varphi + \sigma_z = -3p. \tag{1.10}$$

Since due to the incompressibility condition

$$\varepsilon_r + \varepsilon_\varphi + \varepsilon_z = 0, \tag{1.11}$$

then from (1.9)–(1.11) we have:

$$\sigma_r = -p + \varepsilon_r / 3\omega,\ \sigma_\varphi = -p + \varepsilon_\varphi / 3\omega,\ \sigma_z = -p + \varepsilon_z / 3\omega. \tag{1.12}$$

According to (1.8), (1.12),

$$\frac{1}{\omega} = \frac{T\sqrt{6}}{\sqrt{\left(\varepsilon_r^2 + \varepsilon_\varphi^2 + \varepsilon_z^2\right)}}. \tag{1.13}$$

Power of internal forces per unit volume:

$$w = T\sqrt{\frac{2}{3}}\sqrt{\left(\varepsilon_r^2 + \varepsilon_\varphi^2 + \varepsilon_z^2\right)}. \tag{1.14}$$

Taking into account the axial symmetry, the radial equation of equilibrium in the quasi-static approximation has the form:

$$\frac{\partial\sigma_r}{\partial r} + \frac{\left(\sigma_r - \sigma_\varphi\right)}{r} = 0. \tag{1.15}$$

The equilibrium equation along the z axis is satisfied integrally, i.e., the total stress forces σ_z in each section z = const are balanced by the external pressure on the capsule face:

$$2\pi\int_{R_1}^{R_4}\sigma_z r dr = -\pi P\left(R_4^2 - R_1^2\right), \tag{1.16}$$

where P is the external pressure.

At the boundaries $r = R_2$, $r = R_3$ we assume the condition of equality of radial velocities U_r. At the boundaries $r = R_1$, $r = R_4$ we assume

$$\sigma_r = -P. \tag{1.17}$$

By assumption of the constancy of the strain rate ε_z over the volume, since the problem statement allows the system to move as a rigid body along the z axis, we can put the following expression for the axial velocity U_z:

$$U_z = \left(V / H\right)z, \tag{1.18}$$

where V is the value of velocity U_z at $z = H$, determined during the solution process.

The radial velocity at each moment is a function of radius only $U_r = U_r(r)$. For strain rates we have the following relations:

$$\varepsilon_r = \frac{\partial U_r}{\partial r},\ \varepsilon_\varphi = \frac{U_r}{r},\ \varepsilon_z = \frac{\partial U_z}{\partial z} = V / H. \tag{1.19}$$

Relations (1.1)–(1.13), (1.15), (1.19) with additional conditions (1.16)–(1.18) mathematically define the problem of finding the velocity field at each moment of time at known distribution of relative density of powder material.

The law of change of powder material density is determined by the continuity equation:

$$\frac{\partial \rho}{\partial t} + \frac{1}{r}\frac{\partial \left(\rho U_r r\right)}{\partial r} = 0, \tag{1.20}$$

where t is time.

2. FIELD OF VELOCITIES

As shown in [24], a nonuniform density distribution along the radius occurs when the tube is pressed. However, as a rule, the wall thickness of the tube is significantly less than its radius. Under these conditions, it can be approximated that the relative density of the powder material is constant along the radius at each moment of the process. Consequently, for the radial velocity of displacements U_r in the powder material we have an approximate representation:

$$U_r = Ar + B/r, \tag{2.1}$$

where A, B are coefficients determined in the solution process.

The value of the velocity U_z is determined by Eq. (1.18).

From the incompressibility condition (1.11) in the capsule material we have: $\frac{dU_r}{dr} + \frac{U_r}{r} + \varepsilon_z = 0$. Consequently:

$$U_r = -1/2\varepsilon_z r + C/r, \tag{2.2}$$

where C is the coefficient determined in the solution process.

We suppose that

$$U_r = U_1 \text{ at } r = R_2;\ U_r = U_2 \text{ at } r = R_3. \tag{2.3}$$

Finally, taking into account (2.3) and the condition of continuity of velocities at $r = R_2$, $r = R_3$, the velocities in the total area are equal to:

at $R_1 < r < R_2$, $0 < z < H$ (capsule)

$$\begin{gathered} U_z = V / H, \\ U_r = -\left(1/2\right)\varepsilon_z r + C_1 / r, \\ C_1 = \left(1/2\right)\varepsilon_z R_2^2 + U_1 R_2; \end{gathered} \tag{2.4}$$

at $R_3 < r < R_4$, $0 < z < H$ (capsule)

$$\begin{gathered} U_z = V / H, \\ U_r = -\left(1/2\right)\varepsilon_z r + C_3 / r, \\ C_3 = \left(1/2\right)\varepsilon_z R_3^2 + U_2 R_3; \end{gathered} \tag{2.5}$$

at $R_2 < r < R_3$, $0 < z < H$ (powder)

$$\begin{gathered} U_z = \left(V / H\right)z, \\ U_r = Ar + B_1 R_3^2 / r, \\ A = \left(U_2 R_3 - U_1 R_2\right) / \left(R_3^2 - R_2^2\right),\ B_1 = \left(U_1 R_2 - U_2 R_2^2 / R_3\right) / \left(R_3{}^2 - R_2^2\right). \end{gathered} \tag{2.6}$$

According to (2.4)–(2.6), (1.19) strain rates in the entire domain:
at $R_1 < r < R_2$, $0 < z < H$ (capsule)

$$\varepsilon_z = V / H,\ \ \varepsilon_r = -(1/2)\varepsilon_z - C_1 / r^2,\ \ \varepsilon_\varphi = -(1/2)\varepsilon_z + C_1 / r^2; \tag{2.7}$$

at $R_3 < r < R_4$, $0 < z < H$ (capsule)

$$\varepsilon_z = V / H,\ \varepsilon_r = -(1/2)\varepsilon_z - C_3 / r^2,\ \varepsilon_\varphi = -(1/2)\varepsilon_z + C_3 / r^2; \tag{2.8}$$

at $R_2 < r < R_3$, $0 < z < H$ (powder)

$$\varepsilon_z = V / H,\ \varepsilon_r = A - B_1 R_3^2 / r^2,\ \varepsilon_\varphi = A + B_1 R_3^2 / r^2. \tag{2.9}$$

The total internal force power W consists of three components: $W = W_1 + W_2 + W_3$, where W_1 is the internal force power in the inner capsule at $R_1 < r < R_2$, $0 < z < H$; W_3 is the internal force power in the outer capsule at $R_3 < r < R_4$, $0 < z < H$; W_2 is the internal force power in the powder at $R_2 < r < R_3$, $0 < z < H$.

At $R_1 < r < R_2$, $0 < z < H$ the power of internal forces per unit volume, according to (1.14), (2.7), is equal to $w_1 = T_1\sqrt{2/3}\sqrt{(3/2)\varepsilon_z^2 + 2C_1^2 / r^4}$, where T_1 is the yield strength. Then the total power in the internal capsule:

$$W_1 = 2\pi H T_1 \sqrt{2/3}\int_{R_1}^{R_2} \sqrt{(3/2)\varepsilon_z^2 + 2C_1^2 / r^4}\, r dr.$$

After the corresponding calculations:

$$W_1 = \pi H T_1 \sqrt{\frac{1}{3}}\left\{\left[\sqrt{3\varepsilon_z^2 R_2^4 + 4C_1^2} - \sqrt{3\varepsilon_z^2 R_1^4 + 4C_1^2}\right] + 2C_1 \ln\left(\frac{R_2^2}{R_1^2}\cdot\frac{2C_1 + \sqrt{3\varepsilon_z^2 R_1^4 + 4C_1^2}}{2C_1 + \sqrt{3\varepsilon_z^2 R_1^4 + 4C_1^2}}\right)\right\}. \tag{2.10}$$

At $R_3 < r < R_4$, $0 < z < H$ the power of internal forces per unit volume, according to (1.14), (2.8) is equal to $w_3 = T_2\sqrt{2/3}\sqrt{(3/2)\varepsilon_z^2 + 2C_3^2 / r^4}$, where T_2 is the yield strength. Total power:

$$W_3 = 2\pi H T_2 \sqrt{2/3}\int_{R_3}^{R_4} \sqrt{(3/2)\varepsilon_z^2 + 2C_3^2 / r^4}\, r dr.$$

By calculating, we get:

$$W_3 = \pi H T_2 \sqrt{\frac{1}{3}}\left\{\left[\sqrt{3\varepsilon_z^2 R_4^4 + 4C_3^2} - \sqrt{3\varepsilon_z^2 R_3^4 + 4C_3^2}\right] + 2C_3 \ln\left(\frac{R_4^2}{R_3^2}\cdot\frac{2C_3 + \sqrt{3\varepsilon_z^2 R_3^4 + 4C_3^2}}{2C_3 + \sqrt{3\varepsilon_z^2 R_4^4 + 4C_3^2}}\right)\right\}. \tag{2.11}$$

At $R_2 < r < R_3$, $0 < z < H$ ((in powder), the power of internal forces in a unit volume, according to (1.7), (2.9), is equal to $w_2 = \frac{Y}{3}\sqrt{\left[\left(9f_2^2 - 2f_1^2\right)\left(2A + \varepsilon_z\right)^2 + 6f_1^2\left(\varepsilon_z{}^2 + 2A^2\right)\right] + \frac{12 f_1^2 R_3^4 B_1^2}{r^4}}$. Total power:

$$W_2 = (2/3)\pi H Y \int_{R_2}^{R_3} \sqrt{\left[\left(9f_2^2 - 2f_1^2\right)\left(2A + \varepsilon_z\right)^2 + 6f_1^2\left(\varepsilon_z^2 + 2A^2\right)\right] + 12 f_1^2 R_3^4 B_1^2 / r^4}\, r dr.$$

Integrating, we come to the expression:

$$W_2 = \pi HYR_3^2 / 3\left\{\sqrt{\left[\left(9f_2^2 - 2f_1^2\right)\left(2A+\varepsilon_z\right)^2 + 6f_1^2\left(\varepsilon_z^2 + 2A^2\right)\right] + 12f_1^2B_1^2} - \right.$$
$$- \sqrt{\left[\left(9f_2^2 - 2f_1^2\right)\left(2A+\varepsilon_z\right)^2 + 6f_1^2\left(\varepsilon_z^2 + 2A^2\right)\right]R_2^4 / R_3^4 + 12f_1^2B_1^2} + \qquad (2.12)$$
$$\left. + 2\sqrt{3}f_1B_1 \ln\left(\frac{R_3^2}{R_2^2}\frac{\sqrt{12}f_1B_1 + \sqrt{\left[\left(9f_2^2 - 2f_1^2\right)\left(2A+\varepsilon_z\right)^2 + 6f_1^2\left(\varepsilon_z^2 + 2A^2\right)\right]R_2^4 / R_3^4 + 12f_1^2B_1^2}}{\sqrt{12}f_1B_1 + \sqrt{\left[\left(9f_2^2 - 2f_1^2\right)\left(2A+\varepsilon_z\right)^2 + 6f_1^2\left(\varepsilon_z^2 + 2A^2\right)\right] + 12f_1^2B_1^2}}\right)\right.\!.$$

Let P be the external pressure. Let us determine the power of external forces. The total power of external forces contains three components: N_1 is the power of external forces at the boundary $z = H$; N_2 is the power of external forces at the boundary $r = R_1$; N_3 is the power of external forces at the boundary $r = R_4$. After the corresponding calculations, we obtain:

$$N_1 = -\pi PV\left(R_4^2 - R_1^2\right),$$
$$N_2 = 2\pi PR_1H\left(-1/2\varepsilon_zR_1 + 1/2\varepsilon_zR_2^2 / R_1 + U_1R_2 / R_1\right),$$
$$N_3 = -2\pi PR_4H\left(-1/2\varepsilon_zR_4 + 1/2\varepsilon_zR_3^2 / R_4 + U_2R_3 / R_4\right).$$

Total power $N = N_1 + N_2 + N_3$.
Considering that $\varepsilon_z = V/H$:

$$N = -PH\left(R_3^2 - R_2^2\right)\pi\left[V / H + 2\left(U_2R_3 - U_1R_2\right) / \left(R_3^2 - R_2^2\right)\right]. \qquad (2.13)$$

From the condition $N = W_1 + W_2 + W_3$ we obtain:

$$P = -\frac{W_1 + W_2 + W_3}{H\left(R_3^2 - R_2^2\right)\pi\left[\frac{V}{H} + 2\left(U_2\frac{R_3}{R_3^2 - R_2^2} - U_1\frac{R_2}{R_3^2 - R_2^2}\right)\right]}. \qquad (2.14)$$

Since the velocities are determined with accuracy up to a constant multiplier and there is no characteristic time in the problem formulation, we can put:

$$\left[\frac{V}{H} + 2\left(U_2\frac{R_3}{R_3^2 - R_2^2} - U_1\frac{R_2}{R_3^2 - R_2^2}\right)\right] = -1. \qquad (2.15)$$

Let us represent the external pressure P in the form:

$$P = T_1M_1 + YM_2 + T_2M_3. \qquad (2.16)$$

Let us denote:

$$\beta_1 = T_1 / Y,\ \beta_2 = T_2 / Y,$$
$$\alpha_1 = R_2 / R_1,\ \alpha_2 = R_3 / R_2,\ \alpha_3 = R_4 / R_3, \qquad (2.17)$$
$$\mu_1 = U_1 / R_2,\ \mu_2 = U_2 / R_3.$$

Relations (2.15), (2.16) will take the form:

$$P/Y = \beta_1M_1 + M_2 + \beta_2M_3, \qquad (2.18)$$

$$\varepsilon_z + 2\left(\alpha_2^2\mu_2 - \mu_1\right)/\left(\alpha_2^2 - 1\right) = -1. \tag{2.19}$$

Then, according to (2.10)–(2.14), (2.16)–(2.19):

$$M_1 = M_1\left(\mu_1, \varepsilon_z\right) = \frac{1}{\sqrt{3}\left(\alpha_2^2 - 1\right)}\left\{\left[\sqrt{3\varepsilon_z^2 + \left(2\mu_1 + \varepsilon_z\right)^2} - \sqrt{\frac{3\varepsilon_z^2}{\alpha_1^4} + \left(2\mu_1 + \varepsilon_z\right)^2}\right] + \left(2\mu_1 + \varepsilon_z\right)\ln\left(\frac{\left(2\mu_1 + \varepsilon_z\right)\alpha_1^2 + \sqrt{3\varepsilon_z^2 + \alpha_1^4\left(2\mu_1 + \varepsilon_z\right)^2}}{\left(2\mu_1 + \varepsilon_z\right) + \sqrt{3\varepsilon_z^2 + \left(2\mu_1 + \varepsilon_z\right)^2}}\right)\right\}, \tag{2.20}$$

$$M_3 = M_3\left(\mu_2, \varepsilon_z\right) = \frac{\alpha_2^2}{\sqrt{3}\left(\alpha_2^2 - 1\right)}\left\{\left[\sqrt{3\varepsilon_z^2\alpha_3{}^4 + \left(2\mu_2 + \varepsilon_z\right)^2} - \sqrt{3\varepsilon_z^2 + \left(2\mu_2 + \varepsilon_z\right)^2}\right] + \left(2\mu_2 + \varepsilon_z\right)\ln\left(\alpha_3^2\frac{\left(2\mu_2 + \varepsilon_z\right) + \sqrt{3\varepsilon_z^2 + \left(2\mu_2 + \varepsilon_z\right)^2}}{\left(2\mu_2 + \varepsilon_z\right) + \sqrt{3\varepsilon_z^2\alpha_3^4 + \left(2\mu_2 + \varepsilon_z\right)^2}}\right)\right\}, \tag{2.21}$$

$$M_2 = M_2\left(A, B_1\right) = \frac{\alpha_2^2}{3\left(\alpha_2^2 - 1\right)}\left\{\sqrt{\left[9f_2^2 + 4f_1^2\left(1 + 3A\right)^2\right] + 12f_1^2B_1^2} - \sqrt{\frac{1}{\alpha_2^4}\left[9f_2^2 + 4f_1^2\left(1 + 3A\right)^2\right] + 12f_1^2B_1^2} + 2\sqrt{3}f_1B_1\ln\frac{\sqrt{12}\alpha_2^2f_1B_1 + \sqrt{\left[9f_2^2 + 4f_1^2\left(1 + 3A\right)^2\right] + 12\alpha_2^4f_1^2B_1^2}}{\sqrt{12}f_1B_1 + \sqrt{\left[9f_2^2 + 4f_1^2\left(1 + 3A\right)^2\right] + 12f_1^2B_1^2}}\right\}, \tag{2.22}$$

where

$$A = \left(\mu_2\alpha_2^2 - \mu_1\right)/\left(\alpha_2^2 - 1\right),\ B_1 = \left(\mu_1 - \mu_2\right)/\left(\alpha_2^2 - 1\right). \tag{2.23}$$

3. SYSTEM OF EQUATIONS FOR DETERMINING VARIABLES UNKNOWN

At each step of the HIP process it is necessary to determine the unknown values μ_1, μ_2, ε_z, and P. At known values of μ_1 and μ_2, the values ε_z and P are determined by Eqs. (2.18) and (2.19). The values μ_1, μ_2 are determined from the condition of minimum P in the area bounded by the lines $\mu_1 \geq \mu_2$; $\mu_2 \leq 0$, $\varepsilon_z \leq 0$ on the plane of parameters μ_1, μ_2 (see Figure).

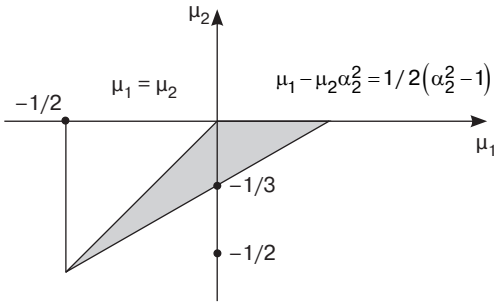


Figure. Area of finding the minimum

According to (2.18) the system of equations for determining μ_1, μ_2 is obtained from the conditions:

$$\frac{\partial}{\partial\mu_1}\left(\beta_1 M_1 + M_2 + \beta_2 M_3\right) = 0, \tag{3.1}$$

$$\frac{\partial}{\partial\mu_2}\left(\beta_1 M_1 + M_2 + \beta_2 M_3\right) = 0. \tag{3.2}$$

So we have:

$$\frac{\partial\varepsilon_z}{\partial\mu_1} = 2/\left(\alpha_2^2-1\right), \quad \frac{\partial\varepsilon_z}{\partial\mu_2} = -2\alpha_2^2/\left(\alpha_2^2-1\right);$$
$$\frac{\partial A}{\partial\mu_1} = -1/\left(\alpha_2^2-1\right), \quad \frac{\partial A}{\partial\mu_2} = \alpha_2^2/\left(\alpha_2^2-1\right);$$
$$\frac{\partial B_1}{\partial\mu_1} = 1/\left(\alpha_2^2-1\right), \quad \frac{\partial B_1}{\partial\mu_2} = -1/\left(\alpha_2^2-1\right).$$

Consequently, Eqs. (3.1), (3.2) will take the form:

$$\beta_1\frac{\partial M_1}{\partial\mu_1} + \frac{\partial M_1}{\partial\varepsilon_z}\frac{2\beta_1}{\left(\alpha_2^2-1\right)} - \frac{\partial M_2}{\partial A}\frac{1}{\left(\alpha_2^2-1\right)} + \frac{\partial M_2}{\partial B_1}\frac{1}{\left(\alpha_2^2-1\right)} + \frac{\partial M_3}{\partial\varepsilon_z}\frac{2\beta_2}{\left(\alpha_2^2-1\right)} = 0, \tag{3.3}$$

$$-\frac{\partial M_1}{\partial\varepsilon_z}\frac{2\beta_1\alpha_2^2}{\left(\alpha_2^2-1\right)} + \frac{\partial M_2}{\partial A}\frac{\alpha_2^2}{\left(\alpha_2^2-1\right)} - \frac{\partial M_2}{\partial B_1}\frac{1}{\left(\alpha_2^2-1\right)} + \beta_2\frac{\partial M_3}{\partial\mu_2} - \frac{\partial M_3}{\partial\varepsilon_z}\frac{2\beta_2\alpha_2^2}{\left(\alpha_2^2-1\right)} = 0. \tag{3.4}$$

For the above derivatives the corresponding relations are as follows:

$$\frac{\partial M_1}{\partial\mu_1} = \frac{1}{\left(\alpha_2^2-1\right)}\frac{2}{\sqrt{3}}\ln\left(\frac{\left(2\mu_1+\varepsilon_z\right)\alpha_1^2 + \sqrt{3\varepsilon_z^2+\alpha_1^4\left(2\mu_1+\varepsilon_z\right)^2}}{\left(2\mu_1+\varepsilon_z\right)+\sqrt{3\varepsilon_z^2+\left(2\mu_1+\varepsilon_z\right)^2}}\right), \tag{3.5}$$

$$\frac{\partial M_1}{\partial\varepsilon_z} =$$
$$= \frac{1}{\left(\alpha_2^2-1\right)}\sqrt{\frac{1}{3}}\left\{\ln\left(\frac{\left(2\mu_1+\varepsilon_z\right)\alpha_1^2 + \sqrt{3\varepsilon_z^2+\alpha_1^4\left(2\mu_1+\varepsilon_z\right)^2}}{\left(2\mu_1+\varepsilon_z\right)+\sqrt{3\varepsilon_z^2+\left(2\mu_1+\varepsilon_z\right)^2}}\right) + \frac{1}{\alpha_1^4}\frac{3\varepsilon_z\left(\alpha_1^4-1\right)}{\sqrt{3\varepsilon_z^2+\left(2\mu_1+\varepsilon_z\right)^2}+\sqrt{3\varepsilon_z^2/\alpha_1^4+\left(2\mu_1+\varepsilon_z\right)^2}}\right\}, \tag{3.6}$$

$$\frac{\partial M_3}{\partial\mu_2} = \frac{\alpha_2^2}{\left(\alpha_2^2-1\right)}\frac{2}{\sqrt{3}}\ln\left(\alpha_3^2\frac{\left(2\mu_2+\varepsilon_z\right)+\sqrt{3\varepsilon_z^2+\left(2\mu_2+\varepsilon_z\right)^2}}{\left(2\mu_2+\varepsilon_z\right)+\sqrt{3\varepsilon_z^2\alpha_3^4+\left(2\mu_2+\varepsilon_z\right)^2}}\right), \tag{3.7}$$

$$\frac{\partial M_3}{\partial\varepsilon_z} =$$
$$= \frac{\alpha_2^2}{\left(\alpha_2^2-1\right)}\sqrt{\frac{1}{3}}\left\{\ln\left(\alpha_3^2\frac{\left(2\mu_2+\varepsilon_z\right)+\sqrt{3\varepsilon_z^2+\left(2\mu_2+\varepsilon_z\right)^2}}{\left(2\mu_2+\varepsilon_z\right)+\sqrt{3\varepsilon_z^2\alpha_3^4+\left(2\mu_2+\varepsilon_z\right)^2}}\right) + \frac{3\varepsilon_z\left(\alpha_3^4-1\right)}{\sqrt{3\varepsilon_z^2\alpha_3^4+\left(2\mu_2+\varepsilon_z\right)^2}+\sqrt{3\varepsilon_z^2+\left(2\mu_2+\varepsilon_z\right)^2}}\right\}, \tag{3.8}$$

$$\frac{\partial M_2}{\partial A} =$$
$$= \frac{4f_1^2\left(1+3A\right)\left(\alpha_2^2+1\right)}{\alpha_2^2}\left\{\frac{1}{\sqrt{\left[9f_2^2+4f_1^2\left(1+3A\right)^2\right]+12f_1^2B_1^2}+\sqrt{\left[9f_2^2+4f_1^2\left(1+3A\right)^2\right]\frac{1}{\alpha_2^4}+12f_1^2B_1^2}}\right\}, \tag{3.9}$$

$$\frac{\partial M_2}{\partial B_1}=\frac{\alpha_2^2}{\left(\alpha_2^2-1\right)}\frac{2}{\sqrt{3}}f_1\left[\ln\left(\alpha_2^2\frac{\sqrt{12}f_1B_1+\sqrt{\left[9f_2^2+4f_1^2\left(1+3A\right)^2\right]\frac{1}{\alpha_2^4}+12f_1^2B_1^2}}{\sqrt{12}f_1B_1+\sqrt{\left[9f_2^2+4f_1^2\left(1+3A\right)^2\right]+12f_1^2B_1^2}}\right)\right]. \quad (3.10)$$

Equation (3.3) implicitly determines μ_1 at a given value of μ_2. Under certain conditions, the equation has no solutions for the parameters under study, and then the value of μ_1 is at the boundary of the area. That is, it can be argued, taking into account the above remark, that Eq. (3.3) implicitly defines μ_1 as a function of μ_2, i.e., $-\mu_1 = \mu_1(\mu_2)$. Then, using Eq. (3.4), given that $\mu_1 = \mu_1(\mu_2)$, we can determine μ_2. Again, the equation has no solutions, and the point of minimum is on the boundary of the domain. Knowing the parameters μ_1 and μ_2 allows us to determine all other parameters of the process. In order to determine the nature of change of the pipe parameters, it remains to find out what was meant by the concept of time when the condition (2.20) was accepted. From the law of conservation of mass of powder material it follows:

$$\rho\pi H\left(R_3^2-R_2^2\right)=\rho_0\pi H_0\left(R_{30}^2-R_{20}^2\right),$$

where ρ_0 is the initial relative density; R_{20} and R_{30} are the initial pipe dimensions.

Differentiating this relation, after some simplifications we obtain:

$$d\rho H\left(R_3^2-R_2^2\right)+\rho V\left(R_3^2-R_2^2\right)dt+2\rho H\left(R_3U_2-R_2U_1\right)dt=0. \quad (3.11)$$

Let us convert it to the form

$$d\rho/\rho+\left[\varepsilon_z+2\left(\alpha_2^2\mu_2-\mu_1\right)/\left(\alpha_2^2-1\right)\right]dt=0. \quad (3.12)$$

Then, considering (2.19), we have:

$$dt=d\rho/\rho. \quad (3.13)$$

Consequently, the relative density of the powder material ρ can be taken as a process parameter instead of time t. Then the laws of change of values of H, R_2 and R_3 are determined by the relations:

$$dH=\varepsilon_z Hdt\Rightarrow dH=\varepsilon_z Hd\rho/\rho\Rightarrow dH/d\rho=\varepsilon_z H/\rho, \quad (3.14)$$

$$dR_2=U_1dt\Rightarrow dR_2=\mu_1R_2d\rho/\rho\Rightarrow dR_2/d\rho=\mu_1R_2/\rho, \quad (3.15)$$

$$dR_3=U_2dt\Rightarrow dR_3=\mu_2R_3d\rho/\rho\Rightarrow dR_3/d\rho=\mu_2R_3/\rho. \quad (3.16)$$

When H, R_2, R_3 are known, the values of R_1, R_4 are determined from the condition of incompressibility of the capsule material.

$$\pi\left(R_2^2-R_1^2\right)H=\pi\left(R_{20}^2-R_{10}^2\right)H_0, \quad (3.17)$$

$$\pi\left(R_4^2-R_3^2\right)H=\pi\left(R_{40}^2-R_{30}^2\right)H_0, \quad (3.18)$$

where R_{10} and R_{40} are the initial dimensions of the pipe.

Table 1 shows the results of calculation of the process of HIP of the pipe with initial parameters—initial dimensions in millimeters: $R_{10} = 18$, $R_{20} = 20$, $R_{30} = 30$, $R_{40} = 32$, $H_0 = 100$, initial relative density $\rho_0 = 0.6$.

The following function values were taken in the calculation:

$$f_1(\rho)=\sqrt{(\rho-\rho_0)/(1-\rho_0)},\quad f_2(\rho)=\sqrt{(\rho-\rho_0)/(1-\rho)}.$$

Table 1. Results of parameter calculation as a function of relative density

ρ	0.659	0.719	0.778	0.838	0.897	0.937	0.977
P/Y	0.447	0.682	0.932	1.249	1.742	2.357	4.083
H/H_0	0.977	0.954	0.933	0.915	0.898	0.889	0.882
R_1/R_{10}	0.976	0.949	0.923	0.901	0.881	0.870	0.863
R_2/R_{20}	0.983	0.963	0.946	0.930	0.916	0.909	0.905
R_3/R_{30}	0.973	0.948	0.925	0.905	0.887	0.876	0.867
R_4/R_{40}	0.978	0.957	0.939	0.923	0.909	0.900	0.893

Here is another example of calculation of initial and final dimension ratios under other conditions.
$\beta_1 = \beta_2 = 2/9$, $R_{10} = 15$, $R_{20} = 20$, $R_{30} = 30$, $R_{40} = 45$, $H_0 = 100$, $\rho_0 = 0.6$, $\rho = 0.978$, $P/Y = 4.309$, $H/H_0 = 1$, $R_1/R_{10} = 1$, $R_2/R_{20} = 1$, $R_3/R_{30} = 0.886$, $R_4/R_{40} = 0.918$.

In the latter case, the shrinkage for the powder was directed towards the radius reduction. Vertical shrinkage was not observed. The inner capsule remained nondeformable. The minimum was reached at the boundary of the area.

4. POSSIBILITY OF EXPERIMENTAL DETERMINATION OF FUNCTIONS

Traditionally, the functions $f_1(\rho)$, $f_2(\rho)$ are determined on the basis of two experiments [25, 26]. The first experiment involves the process of HIP of a cylindrical sample in a thin-walled capsule up to a certain relative density ρ, while the second investigates the free deposition of the obtained sample after removing the capsule. It should be assumed that the first experiment (considered to be the main one for determining the function $f_2(\rho)$) does not provide uniform all-round compression due to the influence of the capsule. The second precipitation experiment (considered basic for determining the function $f_1(\rho)$) is not easy to perform, especially for small values of relative density ρ. The second drawback of the free settlement experiment for a cylindrical specimen is as follows. In the real HIP process, all deformations are overwhelmingly compressive in nature, while in the settlement experiment, two out of three main deformations are tensile in nature. As shown in [27], the real vector of principal strain rates in the HIP process makes a significantly smaller angle with the vector of uniform compression as compared to that obtained in free settlement. While the ideal experiment in this respect involves one-dimensional pressing of a powder layer, it is difficult to realize due to technical issues. In [27–29] it is shown that both functions can in principle be determined in one experiment if the ratio of strain rates is known at each moment of the experiment. In [28], the possibility of determining the desired functions based on experiments with the same cylindrical specimens interrupted at different values of relative density is demonstrated. The main disadvantage of these experiments is that their results lie rather close to the hydrostatic axis (uniform all-round compression). However, the use of tubular specimens allows us to eliminate this disadvantage to a certain extent. The purpose of this section is to determine the feasibility in principle of determining the functions $f_1(\rho)$, $f_2(\rho)$.

Adding (3.3) to (3.4), we obtain:

$$\beta_1 \partial M_1 / \partial \mu_1 - 2\beta_1 \partial M_1 / \partial \varepsilon_z + \partial M_2 / \partial A + \beta_2 \partial M_3 / \partial \mu_2 - 2\beta_2 \partial M_3 / \partial \varepsilon_z = 0. \quad (4.1)$$

Using (3.9), we rewrite Eq. (4.1) in the form:

$$\frac{4 f_1^2 (1+3A)\left(\alpha_2^2+1\right)}{\alpha_2^2\left\{\sqrt{\left[9 f_2^2 + 4 f_1^2 (1+3A)^2\right] + 12 f_1^2 B_1^2} + \sqrt{\left[9 f_2^2 + 4 f_1^2 (1+3A)^2\right]\frac{1}{\alpha_2^4} + 12 f_1^2 B_1^2}\right\}} = \quad (4.2)$$
$$= -\beta_1 \partial M_1 / \partial \mu_1 + 2\beta_1 \partial M_1 / \partial \varepsilon_z - \beta_2 \partial M_3 / \partial \mu_2 + \beta_2 \partial M_3 / \partial \varepsilon_z.$$

Let us denote:

$$\Psi_1 = \varepsilon_z / \mu_2, \ \Psi_2 = \mu_1 / \mu_2. \quad (4.3)$$

Then

$$(3A+1)/B_1=\Delta_1=\left(\alpha_2^2-\Psi_2\right)/\left(\Psi_2-1\right)-\Psi_1\left(\alpha_2^2-1\right)/\left(\Psi_2-1\right), \tag{4.4}$$

$$1/B_1=\Delta_2=-\Psi_1\left(\alpha_2^2-1\right)/\left(\Psi_2-1\right)-2\left(\alpha_2^2-\Psi_2\right)/\left(\Psi_2-1\right). \tag{4.5}$$

According to (3.5)–(3.10) and (4.3)–(4.5), we obtain

$$\frac{\partial M_1}{\partial\mu_1}=\frac{1}{\left(\alpha_2^2-1\right)}\frac{2}{\sqrt{3}}\ln\left(\frac{-\left(2\Psi_2+\Psi_1\right)\alpha_1^2+\sqrt{3\Psi_1^2+\alpha_1^4\left(2\Psi_2+\Psi_1\right)^2}}{-\left(2\Psi_2+\Psi_1\right)+\sqrt{3\Psi_1^2+\left(2\Psi_2+\Psi_1\right)^2}}\right),$$

$$\frac{\partial M_3}{\partial\mu_2}=\frac{\alpha_2^2}{\left(\alpha_2^2-1\right)}\frac{2}{\sqrt{3}}\ln\left(\alpha_3^2\frac{-\left(2+\Psi_1\right)+\sqrt{3\Psi_1^2+\left(2+\Psi_1\right)^2}}{-\left(2+\Psi_1\right)+\sqrt{3\Psi_1^2\alpha_3^4+\left(2+\Psi_1\right)^2}}\right),$$

$$\frac{\partial M_1}{\partial\varepsilon_z}=$$

$$=\frac{1}{\left(\alpha_2^2-1\right)}\sqrt{\frac{1}{3}}\left\{\ln\left(\alpha_1^2\frac{-\left(2\Psi_2+\Psi_1\right)+\sqrt{3\Psi_1^2\frac{1}{\alpha_1^4}+\left(2\Psi_2+\Psi_1\right)^2}}{-\left(2\Psi_2+\Psi_1\right)+\sqrt{3\Psi_1^2+\left(2\Psi_2+\Psi_1\right)^2}}\right)-\frac{3\Psi_1\left(\alpha_1^4-1\right)}{\alpha_1^4\left[\sqrt{3\Psi_1^2+\left(2\Psi_2+\Psi_1\right)^2}+\sqrt{3\Psi_1^2\frac{1}{\alpha_1^4}+\left(2\Psi_2+\Psi_1\right)^2}\right]},\right.$$

$$\frac{\partial M_3}{\partial\varepsilon_z}=\frac{\alpha_2^2}{\left(\alpha_2^2-1\right)}\sqrt{\frac{1}{3}}\left\{\ln\left(\alpha_3^2\frac{-\left(2+\Psi_1\right)+\sqrt{3\Psi_1^2+\left(2+\Psi_1\right)^2}}{-\left(2+\Psi_1\right)+\sqrt{3\Psi_1^2\alpha_3^4+\left(2+\Psi_1\right)^2}}\right)-\frac{3\Psi_1\left(\alpha_3^4-1\right)}{\left[\sqrt{3\Psi_1^2\alpha_3^4+\left(2+\Psi_1\right)^2}+\sqrt{3\Psi_1^2+\left(2+\Psi_1\right)^2}\right]}.\right.$$

Let $x^2=\left[(3/4)\Delta_2^2f_2^2+(1/3)\Delta_1^2f_1^2\right]$. Then, according to (4.2):

$$\frac{f_1^2}{\sqrt{x^2+f_1^2}+\sqrt{x^2\gamma^2+f_1^2}}=\Omega_1\left(\Psi_1,\Psi_2\right), \tag{4.6}$$

where

$$\gamma=\frac{1}{\alpha_2^2}<1;\ \ \Omega_1\left(\Psi_1,\Psi_2\right)=\frac{\alpha_2^2\sqrt{3}}{2\Delta_1\left(\alpha_2^2+1\right)}\left(-\beta_1\frac{\partial M_1}{\partial\mu_1}+2\beta_1\frac{\partial M_1}{\partial\varepsilon_z}-\beta_2\frac{\partial M_3}{\partial\mu_2}+2\beta_2\frac{\partial M_3}{\partial\varepsilon_z}\right). \tag{4.7}$$

Let us denote

$$z=x/f_1. \tag{4.8}$$

Consequently, we have:

$$f_1\left(\sqrt{z^2+1}+\sqrt{z^2\gamma^2+1}\right)=\Omega_1. \tag{4.9}$$

We represent Eq. (2.19) in the form:

$$M_2=P/Y-\beta_1M_1-\beta_2M_3. \tag{4.10}$$

According to (4.7) and (2.23), we have:

$$\frac{\alpha_2^2}{\left(\alpha_2^2-1\right)}\frac{1}{3}\left\{\sqrt{\left[9f_2^2+4f_1^2\left(1+3A\right)^2\right]+12f_1^2B_1^2}-\sqrt{\left[9f_2^2+4f_1^2\left(1+3A\right)^2\right]\frac{1}{\alpha_2^4}+12f_1^2B_1^2}+\right.$$
$$\left.+2\sqrt{3}f_1B_1\ln\left(\alpha_2^2\frac{\sqrt{12}f_1B_1+\sqrt{\left[9f_2^2+4f_1^2\left(1+3A\right)^2\right]\frac{1}{\alpha_2^4}+12f_1^2B_1^2}}{\sqrt{12}f_1B_1+\sqrt{\left[9f_2^2+4f_1^2\left(1+3A\right)^2\right]+12f_1^2B_1^2}}\right)\right\}=\frac{P}{Y}-\beta_1M_1-\beta_2M_3.$$

Taking into account (4.3)–(4.5) and (4.8), we transform this equation to the form:

$$\sqrt{x^2+f_1^2}-\sqrt{x^2\gamma^2+f_1^2}+f_1\ln\left(\frac{1}{\gamma}\frac{f_1+\sqrt{x^2\gamma^2+f_1^2}}{f_1+\sqrt{x^2+f_1^2}}\right)=\frac{\sqrt{3}\left(\alpha_2^2-1\right)\Delta_2}{2\alpha_2^2}\left(\frac{P}{Y}-\beta_1M_1-\beta_2M_3\right).\qquad(4.11)$$

We note that according to (4.3), (2.21), and (2.22):

$$M_1=\frac{1}{\sqrt{3}}\frac{1}{\Psi_1\left(\alpha_2^2-1\right)+2\left(\alpha_2^2-\Psi_2\right)}\left\{\left[\sqrt{3\Psi_1^2+\left(2\Psi_2+\Psi_1\right)^2}-\sqrt{3\Psi_1^2\frac{1}{\alpha_1^4}+\left(2\Psi_2+\Psi_1\right)^2}\right]-\right.$$
$$\left.-\left(2\Psi_2+\Psi_1\right)\ln\left(\frac{-\left(2\Psi_2+\Psi_1\right)\alpha_1^2+\sqrt{3\Psi_1^2+\alpha_1^4\left(2\Psi_2+\Psi_1\right)^2}}{-\left(2\Psi_2+\Psi_1\right)+\sqrt{3\Psi_1^2+\left(2\Psi_2+\Psi_1\right)^2}}\right)\right\},$$

$$M_3=\sqrt{\frac{1}{3}}\frac{\alpha_2^2}{\Psi_1\left(\alpha_2^2-1\right)+2\left(\alpha_2^2-\Psi_2\right)}\left\{\left[\sqrt{3\Psi_1^2\alpha_3^4+\left(2+\Psi_1\right)^2}-\sqrt{3\Psi_1^2+\left(2+\Psi_1\right)^2}\right]-\right.$$
$$\left.-\left(2+\Psi_1\right)\ln\left(\alpha_3^2\frac{-\left(2+\Psi_1\right)+\sqrt{3\Psi_1^2+\left(2+\Psi_1\right)^2}}{-\left(2+\Psi_1\right)+\sqrt{3\Psi_1^2\alpha_3^4+\left(2+\Psi_1\right)^2}}\right)\right\}.$$

Consequently, Eq. (4.11) can be represented in the form:

$$f_1\left\{\sqrt{z^2+1}-\sqrt{z^2\gamma^2+1}+\ln\frac{1}{\gamma}\frac{1+\sqrt{z^2\gamma^2+1}}{1+\sqrt{z^2+1}}\right\}=\Omega_2\left(\Psi_1,\Psi_2\right),\qquad(4.12)$$

where $\Omega_2=\sqrt{3}\left(\alpha_2^2-1\right)\Delta_2/\left(2\alpha_2^2\right)\left(P/Y-\beta_1M_1-\beta_2M_3\right)$.

From Eqs. (4.9) and (4.12) it follows

$$\left(\sqrt{z^2+1}+\sqrt{z^2\gamma^2+1}\right)\left\{\sqrt{z^2+1}-\sqrt{z^2\gamma^2+1}+\ln\left(\frac{1}{\gamma}\frac{1+\sqrt{z^2\gamma^2+1}}{1+\sqrt{z^2+1}}\right)\right\}=\Omega\left(\Psi_1,\Psi_2\right),\qquad(4.13)$$

where $\Omega(\Psi_1, \Psi_2) = \Omega_2/\Omega_1$.

Let us consider the function

$$f(z)=\left(\sqrt{z^2+1}+\sqrt{z^2\gamma^2+1}\right)\left\{\sqrt{z^2+1}-\sqrt{z^2\gamma^2+1}+\ln\left(\frac{1}{\gamma}\frac{1+\sqrt{z^2\gamma^2+1}}{1+\sqrt{z^2+1}}\right)\right\}.$$

This function is monotonically increasing at $z > 0$. Consequently, Eq. (4.13) has a single solution $z > 0$ provided that $f(0) < \Omega$.

Let $H^-, H^+; R_2^-, R_2^+; R_3^-, R_3^+$ are the previous and subsequent values of the corresponding parameters. Then ε_z, μ_1, μ_2 can be assumed as:

$$\varepsilon_z \approx 2\left(H^+ - H^-\right)/\left(H^+ + H^-\right),$$
$$\mu_1 \approx 2\left(R_2^+ - R_2^-\right)/\left(R_2^+ + R_2^-\right),$$
$$\mu_2 \approx 2\left(R_3^+ - R_3^-\right)/\left(R_3^+ + R_3\right).$$

Consequently, we have $\Psi_1 \approx \varepsilon_z / \mu_2; \Psi_2 \approx \mu_1 / \mu_2$.

If the value of z is found, than $f_1 = H_1\left(\sqrt{z^2+1} + \sqrt{z^2\gamma^2+1}\right)$. Since $x^2 = z^2 f_1^2$, from the equation $(3/4)\Delta_2^2 f_2^2 + (1/3)\Delta_1^2 f_1^2 = z^2 f_1^2$ we have:

$$f_2 = \frac{2}{\sqrt{3}}\sqrt{\frac{\left(z^2 - \Delta_1^2/3\right)}{\Delta_2^2}} f_1.$$

Below are the results of approximate determination of the functions based on the calculation under the conditions used in Table 1.

Table 2. Initial data for calculation of the functions $f_1(\rho)$, $f_2(\rho)$

Parameter	1 (initial state)	2	3	4
ρ	0.6198	0.6396	0.6594	0.6792
P/*Y*	0.2544	0.3594	0.4468	0.5272
H	99.2634	98.4711	97.6770	96.8966
R_2	19.9027	19.7805	19.6517	19.5218
R_3	29.7277	29.4550	29.1885	28.9299

The value presented in column 1 of Table 2 was taken as the initial state.

1. The final state is column 2. The relative density increment is 2%. Theoretical values: $f_2 = 0.3324, f_1 = 0.3154$. Calculated values: $f_2 = 0.3329, f_1 = 0.3180$.
2. The final state is column 3. The relative density increment is 4%. Theoretical values: $f_2 = 0.4184, f_1 = 0.3860$. Calculated values: $f_2 = 0.42281, f_1 = 0.3646$.
3. The final state is column 4. The relative density increment is 6%. Theoretical values: $f_2 = 0.4977, f_1 = 0.4455$. Calculated values: $f_2 = 0.5026, f_1 = 0.4079$.

These results show that even at a relative density step of 6%, we obtain quite acceptable agreement between theoretical and calculated data.

Nevertheless, the above method of determining the functions is not applicable if plane deformation is realized (see the second calculation example). In this case, as a rule, one of the capsule walls remains nondeformable. That is, it is in a rigid state. And in this case, the equations for determining the functions are already incorrect, because the minimum is reached at the boundary of the area, and Eqs. (4.1) and (4.10) are obtained under the assumption that the entire system is deformed.

CONCLUSIONS

A model of the HIP process for a powder tube has been developed. To describe the mechanical properties of the powder material, Green's model is adopted; for the capsule material, this involves a model of an ideal plastic body with the condition of incompressibility. However, the application of this model requires knowledge of the following mechanical characteristics of materials: yield strength of capsule

materials, yield strength of powder material monolith, two experimentally determined functions of relative density $f_1(\rho)$ and $f_2(\rho)$. The model can be used to analyze different possible variants of the process—complete deformation of the system and the variant of plane deformation with one fixed boundary. The application of this model permits the use of a relatively simple mathematical apparatus.

The possibility in principle of using tubular specimens for experimental determination of functions included in Green's yield condition is confirmed.

Authors' contributions

V.A. Goloveshkin, V.N. Samarov, G. Raisson—problem statement, mathematical model creation, model research, analysis of results.

A.A. Nickolaenko, D.M. Fisunova—model research, calculations, analysis of results.

REFERENCES

1. Anokhina A.V., Goloveshkin V.A., Samarov V.N., Seliverstov D.G., Raisson G. A Mathematical model for calculating the process of hot isostatic pressing of parts of complex-shaped parts in the presence of a periodic structure of embedded elements. *Mekhanika kompozitsionnykh materialov i konstruktsii = Mechanics of Composite Materials and Structures.* 2002;8(2):245–254 (in Russ.). https://elibrary.ru/jwpwnd
2. Ilyushin A.A. *Plastichnost'. Osnovy obshchei matematicheskoi teorii* (*Plasticity. Fundamentals of General Mathematical Theory*). St. Petersburg: Lenand; 2020. 272 p. (in Russ.). ISBN 978-5-9710-7092-4
3. Ilyushin A.A. *Mekhanika sploshnoi sredy* (*Continuum Mechanics*). Moscow: MSU; 1990. 310 p. (in Russ.). ISBN 5-211-00940-1
4. Cundall P.A., Strack O.D.L. A discrete numerical model for granular assemblies. *Geotechnique*. 1979;29(1):47–65. https://doi.org/10.1680/geot.1979.29.1.47
5. Gordon V.A., Shorkin V.S. The nonlocal theory of the near-surface layer of a solid. In: *Results of the Development of Mechanics in Tula. International Conference: Abstracts of the Reports*. Tula; 1998. P. 24 (in Russ.).
6. Gordon V.A., Shorkin V.S. The nonlocal theory of the near-surface layer of a solid. *Izvestiya Tul'skogo gosudarstvennogo universiteta. Seriya: Matematika. Mekhanika. Informatika = Bulletin of Tula State University. Series: Mathematics. Mechanics. Computer Science*. 1998;4(2):55–57 (in Russ.).
7. Lomakin V.A. *Statisticheskie zadachi mekhaniki tverdykh deformiruemykh tel* (*Statistical Problems of Mechanics of Solid Deformable Bodies*). Moscow: Nauka; 1970. 138 p. (in Russ.).
8. Balshin M.Yu., Kiparisov S.S. *Osnovy poroshkovoi metallurgii* (*Fundamentals of Powder Metallurgy*). Moscow: Metallurgiya; 1978. 184 p. (in Russ.).
9. Balshin M.Yu. *Nauchnye osnovy poroshkovoi metallurgii i metallurgii volokna* (*Scientific Foundations of Powder Metallurgy and Fiber Metallurgy*). Moscow: Metallurgiya; 1972. 336 p. (in Russ.).
10. Fedorchenko I.M., Andrievskii R.A. *Osnovy poroshkovoi metallurgii* (*Fundamentals of Powder Metallurgy*). Kiev: Publishing House of the Academy of Sciences of the Ukrainian SSR; 1963. 420 p. (in Russ.).
11. Ilyushin A.A. *Plastichnost'. Ch. 1. Uprugo-plasticheskie deformatsii* (*Plasticity. Part 1. Elastic-plastic deformations*). Moscow; Leningrad: Gostekhizdat; 1948. 376 p. (in Russ.).
12. Kachanov L.M. *Osnovy teorii plastichnosti* (*Fundamentals of the Theory of Plasticity*). Moscow: Nauka; 1969. 420 p. (in Russ.).
13. Sokolovskii V.V. *Teoriya plastichnosti* (*Theory of Plasticity*). Moscow: Vysshaya Shkola; 1969. 608 p. (in Russ.).
14. Hill R. *Matematicheskaya teoriya plastichnosti* (*Mathematical Theory of Plasticity*): transl. from Engl. Moscow: Gostekhizdat; 1956. 408 p. (in Russ.).
[Hill R. *The Mathematical Theory of Plasticity*. Oxford: Clarendon Press; 1950. 356 p.]
15. Frost H.J., Ashby M.F. *Karty mekhanizmov deformatsiy* (*Deformation Mechanism Maps*): transl. from Engl. Chelyabinsk: Metallurgiya; 1989. 327 p. (in Russ.).
[Frost H.J., Ashby M.F. *Deformation Mechanism Maps*. Oxford: Pergamon Press; 1982.]
16. Arzt E., Ashby M.F., Easterling K.E. Practical application of Hot-Isostatic Pressing diagrams: four case studies. *Metall. Trans*. 1983;14A(1):211–221. https://doi.org/10.1007/BF02651618
17. Ashby M.F. A first report of sintering diagrams. *Acta Metall*. 1974;22(3):275–289. https://doi.org/10.1016/0001-6160(74)90167-9
18. Helle A.S., Easterling K.E., Ashby M.F. Hot Isostatic Pressing diagrams: New development. *Acta Metall*. 1985;33(12): 2163–2174. https://doi.org/10.1016/0001-6160(85)90177-4
19. Druyanov B.A. *Prikladnaya teoriya plastichnosti poristykh tel* (*Applied Theory of Plasticity of Porous Bodies*). Moscow: Mashinostroenie; 1989. 168 p. (in Russ.).
20. Green R.J. A plasticity theory for porous solids. *Int. J. Mech. Sci*. 1972;14(4):215–224. https://doi.org/10.1016/0020-7403(72)90063-X
21. Shtern M.B., Serdyuk G.G., Maksimenko L.A., et al. *Fenomenologicheskie teorii pressovaniya poroshkov* (*Phenomenological Theories of Powder Pressing*). Kiev: Naukova dumka; 1982. 140 p. (in Russ.).

22. Skorokhod V.V. *Reologicheskie osnovy teorii spekaniya* (*Rheological Foundations of Sintering Theory*). Kiev: Naukova dumka; 1972. 152 p. (in Russ.).
23. Goloveshkin V.A., Kazberovich A.M., Samarov V.N., Seliverstov D.G. New Regularities of the Shape-Changing of Hollow Parts During HIP. In: Koizumi M. (Ed.). *Hot Isostatic Pressing Theory and Applications.* Springer; 1992. P. 281–287. https://doi.org/10.1007/978-94-011-2900-8_43
24. Anokhina A.V., Goloveshkin V.A., Pirumov A.R., Flaks M.Ya. Investigation of the initial process of pressing pipes made of powder materials, taking into account vertical shrinkage. *Mekhanika kompozitsionnykh materialov i konstruktsii = Mechanics of Composite Materials and Structures*. 2003;9(2):123–132 (in Russ.).
25. Dutton R.E., Shamasundar S., Semiatin S.L. Modeling the Hot Consolidation of Ceramic and Metal Powders. *Metall. Mater. Trans*. A. 1995;26A:2041–2051. https://doi.org/10.1007/BF02670676
26. Vlasov A.V., Seliverstov D.G. Determination of the plasticity functions of powder materials used in HIP. In: *Research in the Field of Theory, Technology and Equipment for Stamping Production*: *Collection of Scientific Papers*. Tula; 1998. P. 46–49 (in Russ.).
27. Raisson G., Goloveshkin V., Samarov V. Identification of Porous Materials Rheological Coefficient Using Experimental Determination of the Radial and Longitudinal Strain Rate Ratio. In: *Hot Isostatic Pressing – HIP'22. Materials Research Proceedings*. 2023. V. 38. P. 150–159. https://doi.org/10.21741/9781644902837-21
28. Raisson G., Goloveshkin V., Khomyakov E., Samarov V. Effect of Experimental Determination Process on Shear Stress Coefficient of Green Equation Describing HIP. In: *Hot Isostatic Pressing – HIP'22. Materials Research Proceedings*. 2023. V. 38. P. 172–176. https://doi.org/10.21741/9781644902837-24
29. Bochkov A., Kozyrev Yu., Ponomarev A., Raisson G. Theoretical Evaluation of Capsule Material Strain Hardening on the Deformation of Long Cylindrical Blanks During HIP Process. In: *Hot Isostatic Pressing – HIP'22. Materials Research Proceedings*. 2023. V. 38. P. 177–183. https://doi.org/10.21741/9781644902837-25

СПИСОК ЛИТЕРАТУРЫ

1. Анохина А.В., Головешкин В.А., Самаров В.Н., Селиверстов Д.Г., Raisson G. Математическая модель расчета процесса горячего изостатического прессования деталей сложной формы при наличии периодической структуры закладных элементов. *Механика композиционных материалов и конструкций*. 2002;8(2):245–254. https://elibrary.ru/jwpwnd
2. Ильюшин А.А. *Пластичность. Основы общей математической теории*. СПб.: Ленанд; 2020. 272 с. ISBN 978-5-9710-7092-4
3. Ильюшин А.А. *Механика сплошной среды*. М.: Изд-во МГУ; 1990. 312 с. ISBN 5-211-00940-1
4. Cundall P.A., Strack O.D.L. A discrete numerical model for granular assemblies. *Geotechnique*. 1979;29(1):47–65. https://doi.org/10.1680/geot.1979.29.1.47
5. Гордон В.А., Шоркин В.С. Нелокальная теория приповерхностного слоя твердого тела. В сб.: *Итоги развития механики в Туле. Международная конференция: Тезисы докладов*. Тула: ТулГУ; 1998. С. 24.
6. Гордон В.А., Шоркин В.С. Нелокальная теория приповерхностного слоя твердого тела. *Известия ТулГУ. Серия: Математика. Механика. Информатика.* 1998;4(2):55–57.
7. Ломакин В.А. *Статистические задачи механики твердых деформируемых тел*. М.: Наука; 1970. 138 с.
8. Бальшин М.Ю., Кипарисов С.С. *Основы порошковой металлургии*. М.: Металлургия; 1978. 184 с.
9. Бальшин М.Ю. *Научные основы порошковой металлургии и металлургии волокна*. М.: Металлургия; 1972. 336 с.
10. Федорченко И.М., Андриевский Р.А. *Основы порошковой металлургии*. Киев: Изд-во АН УССР; 1963. 420 с.
11. Ильюшин А.А. *Пластичность. Ч. 1. Упруго-пластические деформации*. М.; Л.: Гостехиздат; 1948. 376 с.
12. Качанов Л.М. *Основы теории пластичности*. М.: Наука; 1969. 420 с.
13. Соколовский В.В. *Теория пластичности*. М.: Высшая Школа; 1969. 608 с.
14. Хилл Р. *Математическая теория пластичности:* пер. с англ. М.: Гостехиздат; 1956. 408 с.
15. Фрост Г., Эшби М.Ф. *Карты механизмов деформаций:* пер. с англ. Челябинск: Металлургия; 1989. 327 с.
16. Arzt E., Ashby M.F., Easterling K.E. Practical application of Hot-Isostatic Pressing diagrams: four case studies. *Metall. Trans*. 1983;14A(1):211–221. https://doi.org/10.1007/BF02651618
17. Ashby M.F. A first report of sintering diagrams. *Acta Metall*. 1974;22(3):275–289. https://doi.org/10.1016/0001-6160(74)90167-9
18. Helle A.S., Easterling K.E., Ashby M.F. Hot Isostatic Pressing diagrams: New development. *Acta Metall*. 1985;33(12):2163–2174. https://doi.org/10.1016/0001-6160(85)90177-4
19. Друянов Б.А. *Прикладная теория пластичности пористых тел*. М.: Машиностроение; 1989. 168 с.
20. Green R.J. A plasticity theory for porous solids. *Int. J. Mech. Sci*. 1972;14(4):215–224. https://doi.org/10.1016/0020-7403(72)90063-X
21. Штерн М.Б., Сердюк Г.Г., Максименко Л.А., Трухан Ю.В., Шуляков Ю.М. *Феноменологические теории прессования порошков*. Киев: Наукова думка; 1982. 140 с.
22. Скороход В.В. *Реологические основы теории спекания*. Киев: Наукова думка; 1972. 152 с.
23. Goloveshkin V.A., Kazberovich A.M., Samarov V.N., Seliverstov D.G. New Regularities of the Shape-Changing of Hollow Parts During HIP. In: Koizumi M. (Ed.). *Hot Isostatic Pressing Theory and Applications.* Springer; 1992. P. 281–287. https://doi.org/10.1007/978-94-011-2900-8_43

24. Анохина А.В., Головешкин В.А., Пирумов А.Р., Флакс М.Я. Исследование начального процесса прессования труб из порошковых материалов с учетом вертикальной усадки. *Механика композиционных материалов и конструкций*. 2003;9(2):123–132.
25. Dutton R.E., Shamasundar S., Semiatin S.L. Modeling the Hot Consolidation of Ceramic and Metal Powders. *Metall. Mater. Trans*. A. 1995;26A:2041–2051. https://doi.org/10.1007/BF02670676
26. Власов А.В., Селиверстов Д.Г. Определение функций пластичности порошковых материалов, применяемых при ГИП. В сб.: *Исследование в области теории, технологии и оборудования штамповочного производства: Сб. научных трудов.* Тула; 1998. С. 46–49.
27. Raisson G., Goloveshkin V., Samarov V. Identification of Porous Materials Rheological Coefficient Using Experimental Determination of the Radial and Longitudinal Strain Rate Ratio. In: *Hot Isostatic Pressing – HIP'22. Materials Research Proceedings*. 2023. V. 38. P. 150–159. https://doi.org/10.21741/9781644902837-21
28. Raisson G., Goloveshkin V., Khomyakov E., Samarov V. Effect of Experimental Determination Process on Shear Stress Coefficient of Green Equation Describing HIP. In: *Hot Isostatic Pressing – HIP'22. Materials Research Proceedings*. 2023. V. 38. P. 172–176. https://doi.org/10.21741/9781644902837-24
29. Bochkov A., Kozyrev Yu., Ponomarev A., Raisson G. Theoretical Evaluation of Capsule Material Strain Hardening on the Deformation of Long Cylindrical Blanks During HIP Process. In: *Hot Isostatic Pressing – HIP'22. Materials Research Proceedings*. 2023. V. 38. P. 177–183. https://doi.org/10.21741/9781644902837-25

About the authors

Vasiliy A. Goloveshkin, Dr. Sci. (Eng.), Professor, Higher Mathematics Department, Institute of Cybersecurity and Digital Technologies, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia); Leading Researcher, Institute of Applied Mechanics of Russian Academy of Sciences (7-1, Leningradskii pr., Moscow, 125040 Russia). E-mail: vag-1953@yandex.ru. Scopus Author ID 6602872377, https://orcid.org/0000-0002-5413-8625

Artem A. Nickolaenko, Student, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: temanickolaenko2004@yandex.ru. https://orcid.org/0009-0003-2483-4392

Victor N. Samarov, Dr. Sci. (Eng.), Technical Director, Laboratory of New Technologies (15, Simferopol'skii bul., Moscow, 117556 Russia). E-mail: Samarov13@Aol.com. Scopus Author ID 6603606878

Gerard Raisson, Retired Consultant, Clermond Ferrand, France, E-mail: gerard.raisson@gmail.com. Scopus Author ID 6603152593

Daria M. Fisunova, Student, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: daria.fisunova@gmail.com. https://orcid.org/0009-0008-2857-3499

Об авторах

Головешкин Василий Адамович, д.т.н., профессор, кафедра высшей математики, Институт кибербезопасности и цифровых технологий, ФГБОУ ВО «МИРЭА – Российский технологический университет (119454, Россия, Москва, пр-т Вернадского, д. 78); ведущий научный сотрудник, ФГБУН «Институт прикладной механики Российской академии наук» (125040, Россия, Москва, Ленинградский пр-т, д. 7, стр. 1). E-mail: vag-1953@yandex.ru. Scopus Author ID 6602872377, https://orcid.org/0000-0002-5413-8625

Николаенко Артем Андреевич, студент, ФГБОУ ВО «МИРЭА – Российский технологический университет (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: temanickolaenko2004@yandex.ru. https://orcid.org/0009-0003-2483-4392

Самаров Виктор Наумович, д.т.н., технический директор, ТОО «Лаборатория Новых Технологий» (117556, Россия, Москва, Симферопольский б-р, д. 15). E-mail: Samarov13@Aol.com. Scopus Author ID 6603606878

Рейссон Жерар, консультант, Клермон Ферран, Франция. E-mail: gerard.raisson@gmail.com. Scopus Author ID 6603152593

Фисунова Дарья Михайловна, студент, ФГБОУ ВО «МИРЭА – Российский технологический университет (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: daria.fisunova@gmail.com. https://orcid.org/0009-0008-2857-3499

Translated from Russian into English by L. Bychkova
Edited for English language and spelling by Thomas A. Beavitt

Mathematical modeling

Математическое моделирование

UDC 519.224.22, 519.246.8, 330.322.54
https://doi.org/10.32362/2500-316X-2025-13-2-93-110
EDN JCRKUO

RESEARCH ARTICLE

Dynamic model of BSF portfolio management

Artur A. Mitsel,
Elena V. Viktorenko @

Tomsk State University of Control Systems and Radioelectronics, Tomsk, 634050 Russia
@ *Corresponding author, e-mail: viktorenko.e@gmail.com*

Abstract

Objectives. The work compares studies on BSF portfolios consisting of a risk-free Bond (B) asset, a Stock (S), and a cash Flow (F) that represents risky asset prices in the form of a tree structure. On the basis of existing models for managing dynamic investment portfolios, the work develops a dynamic model for managing a BSF portfolio that combines risk-free and risky assets with a deposit. Random changes in the prices of a risky asset are reflected in the developed model according to a tree structure. Two approaches to portfolio formation are proposed for the study: (1) initial capital is invested in a risk-free asset, while management is conducted at the expense of a risky asset; (2) the initial capital is invested in a risky asset, but management is carried out at the expense of a risk-free asset.

Methods. A binomial model was used to predict the prices of risky assets. Changes in risky asset prices in the model are dynamically managed via a branching tree structure. A comparative analysis of modeling results reveals the optimal control method.

Results. A dynamic model for unrestricted management of a BSF portfolio has been developed. By presenting risky asset prices according to a tree structure, the model can be used to increase the accuracy of evaluating investments by from 2.4 to 2.7 times for the first approach and from 1.7 to 2.7 times for the second. The increased accuracy of evaluating investments as compared with previously proposed models is achieved by averaging prices at various vertices of the tree.

Conclusions. The results of the research suggest that the use of a dynamic management model based on a tree-like price structure can significantly increase the accuracy of evaluating investments in an investment portfolio.

Keywords: optimal control, dynamic system with random parameters, dynamic programming, investment portfolio, tracking a reference portfolio, binomial price structure of a risky asset

• **Submitted:** 26.02.2024 • **Revised:** 24.06.2024 • **Accepted:** 17.02.2025

For citation: Mitsel A.A., Viktorenko E.V. Dynamic model of BSF portfolio management. *Russian Technological Journal.* 2025;13(2):93–110. https://doi.org/10.32362/2500-316X-2025-13-2-93-110, https://elibrary.ru/JCRKUO

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Динамическая модель управления BSF-портфелем без ограничений

А.А. Мицель,
Е.В. Викторенко [@]

Томский государственный университет систем управления и радиоэлектроники, Томск, 634050 Россия
[@] *Автор для переписки, e-mail: viktorenko.e@gmail.com*

Резюме

Цели. Рассматриваются модели управления инвестиционными портфелями, носящими динамический характер, проводится сравнение исследований, посвященных BSF-портфелям (состоящим из безрискового актива (bond), акции (stock) и потока платежей (cash flow)) с древовидной структурой цен рискового актива. Целью работы является разработка динамической модели управления BSF-портфелем, включающим безрисковый, рисковый активы и депозит. В отличие от проведенных ранее исследований, в разрабатываемой модели цены рискового актива изменяются случайным образом, следуя древовидной структуре. К исследованию предлагается два подхода формирования портфеля: 1) начальный капитал вкладывается в безрисковый актив, управление происходит за счет рискового актива; 2) начальный капитал вкладывается в рисковый актив, управление происходит за счет безрискового актива.

Методы. Использована биномиальная модель для моделирования цен рискового актива. Динамическая модель управления на основе древовидной структуры цен рискового актива позволяет учитывать изменения в ценах активов. Сравнительный анализ результатов моделирования выявляет оптимальный способ управления.

Результаты. Разработана динамическая модель управления BSF-портфелем без ограничений. Показано, что динамическая модель управления на основе древовидной структуры цен рискового актива позволяет повысить точность оценки объема вложений от 2.4 до 2.7 раз для первого подхода и от 1.7 до 2.7 раз – для второго. Повышение точности оценки объемов вложений по сравнению с ранее предложенными моделями достигается путем усреднения цен по различным вершинам дерева.

Выводы. Проведенное исследование позволяет говорить о том, что применение динамической модели управления, основанной на древовидной структуре цен, позволяет значительно повысить точность оценки объема вложений в инвестиционный портфель.

Ключевые слова: оптимальное управление, динамическая система со случайными параметрами, динамическое программирование, инвестиционный портфель, слежение за эталонным портфелем, биномиальная структура цен рискового актива

• **Поступила:** 26.02.2024 • **Доработана:** 24.06.2024 • **Принята к опубликованию:** 17.02.2025

Для цитирования: Мицель А.А., Викторенко Е.В. Динамическая модель управления BSF-портфелем без ограничений. *Russian Technological Journal*. 2025;13(2):93–110. https://doi.org/10.32362/2500-316X-2025-13-2-93-110, https://elibrary.ru/JCRKUO

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

The management of investment portfolios (IP) can be analyzed in terms of multi-period dynamic decision-making problems pertaining to transactions that occur at discrete points in time. An evaluation carried out by the investor concerning possible future changes in interest rates, prices, or cash flows from securities forms the basis for further decisions to buy or sell, open deposits or lend, i.e., decisions to reshape the IP. The present work sets out to develop and verify a dynamic model for the management of a portfolio that combines a risky asset (RA), a riskless asset (RLA), and a deposit.

Dynamic models of the IP management have been studied in detail in a number of works [1–12]. The study carried out by V.V. Dombrovsky discusses problems involved in controlling discrete stochastic systems and applying the quadratic criterion in this area. The author considers systems characterized by their functional dependency on states and control actions whose various random parameters include additive and multiplicative sources of noise. As well as deriving equations for optimal linear static and dynamic output regulators, the study applies the obtained conclusions to solve the dynamic IP optimization problem for a portfolio whose financial assets having variable price volatility are analyzed in discrete time. The practical significance of the work lies in the possibility of developing effective IP management strategies [1]. In the study [3], the problems of synthesizing control strategies in discrete systems using a predictive model are considered. These systems also include random parameters comprising additive and multiplicative noise phenomena that depend on the states and controls. The work develops control strategies using prediction for closed-loop and open-loop systems taking into account random factors and restrictions. The results are applied to solve the dynamic optimization problem of IP taking into account restrictions on trading operations. Another study [8] considers the problem of managing an IP consisting of RA and RLA taking into account dynamic tracking of the benchmark portfolio. Price changes on RA are described by stochastic equations with Gaussian and impulsive Poisson perturbations. The method for determining an optimal control strategy using feedback based on the application of a quadratic criterion can be used to evaluate the quality of control and select the best strategy for minimizing uncertainty and achieving the best results. The main scientific contribution of the study to the field of IP control consists in its innovative use of stochastic analysis and feedback techniques. The study by D.V. Dombrovsky and E.A. Lyashenko [9] analyzes the dynamics of the IP control model taking into account restrictions on the trading operations. The model includes stochastic difference equations with random volatility to describe the dynamics of prices of risky financial assets within the given IP. An important problem of IP management arises when trying to ensure effective investment management under conditions of restrictions on trading operations. In order to minimize risks and achieve the best results under financial market conditions where asset prices are subject to random fluctuations, volatility is considered as a random variable. The dissertation by T.Yu. Pashinskaya[1] synthesizes the results of research devoted to the control of nonlinear discrete systems with random parameters under constraints. The author develops a methodology for tracking a hypothetical benchmark portfolio with a predetermined growth trajectory in the field of IP management. The results of the study are used to derive equations for determining optimal strategies of IP management with feedback in the presence of constraints. In [13], a dynamic model of IP management using a linear quality criterion is developed.

Studies based on BSF-portfolios comprising RLA Bond (B), Stock (S), and cash Flow (F) with a tree-like RA price structure have also been conducted [14–21]. These studies analyze market structures including such assets as stocks, RLA bonds, and cash flow. The essence of the model is revealed under certain conditions for completeness and absence of arbitrage in the market. A numerical approach to the development of a self-financing strategy provides a payment function superior to the one established in the terminal vertices of the price tree given an initial portfolio of minimum value. The works [17–21] analyze the properties of the (B, S)-market when market completeness and arbitrage-free conditions are violated. Particular attention is paid to the problems related to the inadequacy of the model representation of the RA price evolution in the process of exchange trading using the binomial pricing mechanism in an incomplete (B, S)-market. The described methods take the impact of market trends on the process of RA price evolution into account.

The present study proposes a new dynamic model of BSF portfolio management including RLA, RA, and deposits. Unlike those described in works [1–12], the presented model considers random RA price changes according to a tree structure. The novelty of the model consists in the increased accuracy of investment evaluation as compared to the model described in the work of T.Yu. Pashinskaya. This effect is achieved by averaging prices across different vertices of the tree.

[1] Pashinskaya T.Yu. *Control with prediction of nonlinear discrete systems with random parameters under constraints*: Cand. Sci. Thesis (Phys.-Math.). Tomsk: Tomsk State University; 2021. http://vital.lib.tsu.ru/vital/access/manager/Repository/koha:000702951 (in Russ.). Accessed February 26, 2024.

MODEL CONSTRUCTION

Let us consider a portfolio consisting of RLA, RA, and deposits at discrete moments of time 0, 1, 2, …, n. We will denote the RLA return rate as $r_1(t)$. The price of RLA is known at each moment of time. The randomly changing price of RLA can take one of two possible values at any one time, i.e., possible prices of the stock have the structure of a binary tree (Fig. 1) with terminal vertices (Fig. 2). Let us denote the probability of the RA price increasing by a random value η as p, while the probability that the asset price will decrease by a random value η will be denoted as $q = 1 - p$.

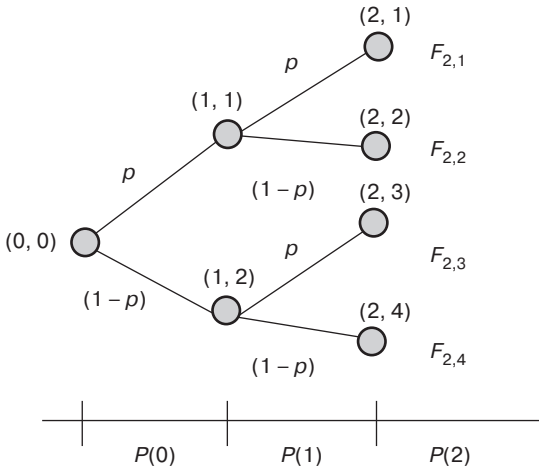


Fig. 1. Two-period RA tree. $F_{t,\,i}$ is the payment function for the point with number (t, i); $P(t)$ is the payment for time step t

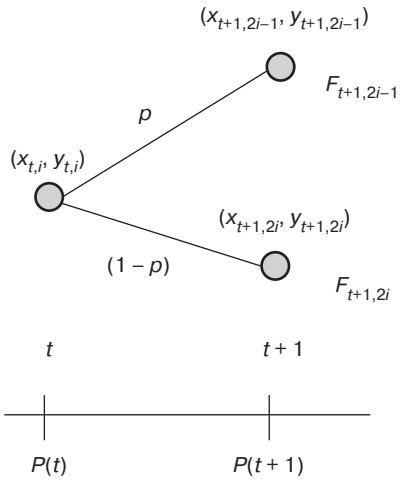


Fig. 2. Terminal vertices of the RA tree

Let us denote the share of RLA as x, and the share of RA as y. The successors of the (t, i)th vertex are the vertices with the numbers $(t + 1, 2i - 1)$ and $(t + 1, 2i)$. The price of RLA in the vertex with number (t, i) corresponding to the moment of time t is equal to $C(t)$, while the price of RLA in the vertices with numbers $(t + 1, 2i - 1)$ and $(t + 1, 2i)$ corresponding to the moment of time $t + 1$ is equal to $C'(t + 1)$. The price of RA in the vertex with number (t, i) corresponding to the moment of time t is equal to $C''_{t,i}$, while the prices of RA in the vertices with numbers $(t + 1, 2i - 1)$ and $(t + 1, 2i)$ corresponding to the moment of time $t + 1$ are equal to $C''_{t+1,2i-1} = C''_{t,i}(1+\eta_i)$ and $C''_{t+1,2i} = C''_{t,i}(1-\eta_i)$ respectively. For each time step t, the payments $P(t)$ are set.

We will suppose that at the initial stage all available funds were invested in RLA and no borrowed funds were used. It is important to note that both RLA and RA can be acquired or alienated at any time, which implies their high availability and readiness for trading [15, 16]. One of the key features of payment flow is its limited liquidity according to which payments are constrained. Since each path from the initial vertex of the price tree structure to the terminal vertex represents a particular scenario, it can be randomized.

The method of asset portfolio management consists in determining at each point of the price tree the RLA x_i and RA y_i under the following conditions [16, 17]:

a) for each endpoint of the tree, a payment function $F_{t,\,i} \geq 0$, $i = 1, 2, \ldots, 2^t$, is defined, representing the amount that the investor expects to receive when asset prices reach the corresponding tree vertex, after selling assets and making payments or receipts of funds along the payment stream;

b) there is a fee for borrowing assets. For example, if x units of RLA are borrowed, at the next moment of time λx units (RLA) should be returned, and μ is the RA loan fee;

c) the market is self-financing, i.e., the investor can buy and sell assets, providing payments and receipts on the payment flow so that the portfolio value at each moment of time does not change, but at the same time the vertex-average value of the portfolio with time changes according to the given law in accordance with the law of change of the vertex-average payment function. Specific vertices can be mathematically written as follows:

$$C'(t+1)\lambda x_{t,i} + C''_{t+1,2i-1}\,\mu\, y_{t,i} + P(t+1) = \\ = C'(t+1)x_{t+1,2i-1} + C''_{t+1,2i-1}y_{t+1,2i-1}, \quad (1)$$

$$C'(t+1)\lambda x_{t,i} + C''_{t+1,2i}\,\mu\, y_{t,i} + P(t+1) = \\ = C'(t+1)x_{t+1,2i} + C''_{t+1,2i}y_{t+1,2i} \quad (2)$$

at $i = \overline{1, 2^t}$.

In the terminal vertices the following inequalities must be met:

$$C'(t+1)\lambda x_{t,i} + C''_{t+1,2i-1}\,\mu\, y_{t,i} + P(t+1) \geq F_{t+1,2i-1}, \quad (3)$$

$$C'(t+1)\lambda x_{t,i} + C''_{t+1,2i}\,\mu\, y_{t,i} + P(t+1) \geq F_{t+1,2i}. \quad (4)$$

Constructing a dynamic model of a BSF portfolio with one RA and one RLA

Let us analyze the IP, where the components are RA with variable returns and risk-free deposits with constant return. At the moment of time t the funds invested in RA are equal to $V''(t)$, and the funds invested in RLA are equal to $V'(t)$. Then the total amount of investments at the moment of time t will be equal, taking into account the deposit

$$V(t) = V'(t) + V''(t) - P(t). \quad (5)$$

Using formulas (1), (2) for the moment of time $t = 1$ (vertices (1, 1) and (1, 2)), we obtain

$$C'(1)x_{1,1} + C''_{1,1}y_{1,1} - P(1) = C'(1)\lambda x_{0,0} + \mu C''_{1,1}y_{0,0},$$
$$C'(1)x_{1,2} + C''_{1,2}y_{1,2} - P(1) = C'(1)\lambda x_{0,0} + \mu C''_{1,2}y_{0,0}.$$

Given the fact that RA prices in vertices (1, 1) and (1, 2) are random, accepting values $C''_{1,1}$ and $C''_{1,2}$ with probabilities p and $q = 1 - p$ respectively, the value of the portfolio at the moment of time $t = 1$ will be equal to

$$V(1) = \lambda C'(1)x_{0,0} + \mu C''(1)y_{0,0}. \quad (6)$$

Here $C'(1)x_{0,0}$ is the RLA cost at the moment of time $t = 1$; $C''(1)y_{0,0}$ is the RA cost at the moment of time $t = 1$;

$$C''(1) = pC''_{1,1} + qC''_{1,2}. \quad (7)$$

$C''(1)$ is the average value of RA price at the moment of time $t = 1$.

For the moment of time $t = 2$ (vertices (2, 1), (2, 2), (2, 3) and (2, 4)) we obtain:

$$V(2) = \lambda C'(2)x_1 + \mu C''(2)y_1. \quad (8)$$

Here

$$C''(2) = p\left(pC''_{2,1} + qC''_{2,2}\right) + q\left(pC''_{2,3} + qC''_{2,4}\right). \quad (9)$$

$C''(2)$ is the vertex-averaged value of RA price at the moment of time $t = 2$; x_1 is the average value of RLA share at the moment of time $t = 1$; y_1 is the average value of RA share at the moment of time $t = 1$.

For the moment of time $t = 3$ (vertices (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), (3, 7), (3, 8)) we obtain:

$$V(3) = \lambda C'(3)x_2 + \mu C''(3)y_2, \quad (10)$$

where x_2 is the average value of the RLA share at the moment of time $t = 2$; y_2 is the average value of the RA share at the moment of time $t = 2$;

$$C''(3) = p\left(p\left(pC''_{3,1} + qC''_{3,2}\right) + q\left(pC''_{3,3} + qC''_{3,4}\right)\right) + q\left(p\left(pC''_{3,5} + qC''_{3,6}\right) + q\left(pC''_{3,7} + qC''_{3,8}\right)\right). \quad (11)$$

$C''(3)$ is the average value of RA price at the moment of time $t = 3$.

Continuing this process, we obtain:

$$V(t) = \lambda C'(t)x_{t-1} + \mu C''(t)y_{t-1},\ t = 1, 2, 3, \ldots, \quad (12)$$

where $x_0 = x_{0,0}$, $y_0 = y_{0,0}$. Here x_{t-1} is the vertex average of the RLA share at the moment of time $t - 1$; y_{t-1} is the average value of the RA share at the moment of time $t - 1$; $C''(t)$ is the average value of the RA price at the moment of time t; y_t is the average value of the RA share at the moment of time t.

Let us introduce the values

$$m_{t,i} = pC''_{t,2i-1} + qC''_{t,2i},\ i = \overline{1, 2^{t-1}}. \quad (13)$$

Then the average price of RA at the moments of time 1, 2, 3, ... can be represented as

$$C''(1) = m_{1,1}, \quad (14)$$

$$C''(2) = pm_{1,1} + qm_{2,2}, \quad (15)$$

$$C''(3) = p(pm_{3,1} + qm_{3,2}) + q(pm_{3,3} + qm_{3,4}). \quad (16)$$

It is easy to show that for any moment of time t the sum of probabilities is equal to 1. Indeed,

$$\sum_{k=0}^{n} C_n^k p^k q^{n-k} = (p+q)^n = 1, \quad (17)$$

where $C_n^k = \dfrac{n!}{k!(n-k)!}$ are the binomial coefficients.

In the terminal vertices the following inequalities must be met:

for the moment of time $t = 1$

$$\lambda C'(1)x_{0,0} + \mu C''(1)y_{0,0} + P(1) \geq F(1); \quad (18)$$

for the moment of time $t = 2$

$$\lambda C'(2)x_1 + \mu C''(2)y_1 + P(2) \geq F(2); \quad (19)$$

for the moment of time t

$$\lambda C'(t)x_{t-1} + \mu C''(t)y_{t-1} + P(t) \geq F(t),\ t = \overline{1, n}. \quad (20)$$

Here $F(t)$ is the average value of the payment function at the moment of time t. In our case it is a deterministic a priori known value.

In accordance with the approach outlined in the dissertation of T.Yu. Pashinskaya, let us introduce the RA return rate for the period of time $[t, t+1]$:

$$v(t+1) = \frac{C''(t+1) - C''(t)}{C''(t)}. \quad (21)$$

Earlier we have introduced the value $r_1(t)$—the RLA return rate. Let us introduce the value $r_2(t)$ the rate on RLA loan (deposit rate).

The dynamics of RLA price and risk-free borrowing are defined by the expressions:

$$C'(t+1) = C'(t)(1 + r_1(t+1)), \quad (22)$$

$$P(t+1) = P(t)(1 + r_2(t+1)). \quad (23)$$

Then from the formulas (20), (22), and (23) for the moment of time $t+1$ it follows:

$$\lambda C'(t)(1 + r_1(t+1))x_t + \mu C''(t)(1 + v(t+1))y_t + \\ + P(t)(1 + r_2(t+1)) \geq F(t+1)$$

or

$$\lambda V'(t)(1 + r_1(t+1)) + \mu V''(t)(1 + v(t+1)) + \\ + P(t)(1 + r_2(t+1)) \geq F(t+1). \quad (24)$$

Model 1

Let us consider the change in IP capital in discrete time. Such a change can be written using the equation taking into account (24) and (5):

$$V(t+1) = \lambda(1 + r_1(t+1))V(t) + \\ + V''(t)[\mu(1 + v(t+1)) - \lambda(1 + r_1(t+1))] + \\ + [\lambda(1 + r_1(t+1)) - (1 + r_2(t+1))]P(t), \\ t = 0, 2, \ldots, n-1, \quad (25)$$

where n is the depth of the tree.

The capital placed in RLA is equal to

$$V'(t) = V(t) - V''(t) + P(t). \quad (26)$$

Note that expression (25) coincides with a similar formula for the dynamics of capital at $\lambda = 1$, $\mu = 1$, obtained in the works of V.V. Dombrovsky [10] and T.Yu. Pashinskaya for the RA random rate.

Let us define the equation of the benchmark portfolio by an expression for the payment function:

$$F(t+1) = [1 + \mu_0(t)]F(t), \quad (27)$$

where $\mu_0(t)$ is a given benchmark portfolio rate. This indicator characterizes the investor's risk aptitude: the larger it is, the higher the risk aptitude. $F(0) = V(0)$ (at the initial moment of time the capital of the reference portfolio coincides with the capital of real IP).

Let us introduce the notations $u1_1(t) = V''(t)$, $u1_2(t) = P(t)$,

$$\mathbf{A1}(t) = \begin{pmatrix} \lambda\left(1 + r_1(t+1)\right) & 0 \\ 0 & (1 + \mu_0(t+1)) \end{pmatrix}, \quad (28)$$

$$\mathbf{B1}(t) = \\ = \begin{pmatrix} \mu(1 + v(t+1)) - \lambda(1 + r_1(t+1)) & \lambda(1 + r_1(t+1)) - (1 + r_2(t+1)) \\ 0 & 0 \end{pmatrix}, \quad (29)$$

$$\mathbf{z}(t) = (V(t)\ F(t))^{\mathrm{T}}. \quad (30)$$

Taking into account (29), (30), and (31), the expression (27) will take the form:

$$\mathbf{z}(t+1) = \mathbf{A1}(t)\mathbf{z}(t) + \mathbf{B1}(t)\mathbf{ul}(t), \quad (31)$$

where

$$\mathbf{u1}(t) = (V''(t)\ P(t))^{\mathrm{T}}. \quad (32)$$

The control variables here are the values

$$u1_1(t) = V''(t),\ u1_2(t) = P(t).$$

The cost of the riskless part of the portfolio in this case is equal to

$$V'(t) = V(t) - V''(t) + P(t) = V(t) - u1_1(t) + u1_2(t). \quad (33)$$

Model 2

Let us describe the IP capital dynamics in discrete time by the equation taking into account (24) and (5):

$$V(t+1) = \mu(1 + v(t+1))V(t) + \\ + V'(t)[\lambda(1 + r_1(t+1)) - \mu(1 + v(t+1))] + \\ + [\mu(1 + v(t+1)) - (1 + r_2(t+1))]P(t), \\ t = 0, 2, \ldots, n-1, \quad (34)$$

where n is the depth of the tree.

Then the matrices for Model 2 will have the following form:

$$\mathbf{A2}(t) = \begin{pmatrix} \mu\left(1 + v(t)\right) & 0 \\ 0 & (1 + \mu_0(t)) \end{pmatrix}, \quad (35)$$

$$\mathbf{B2}(t) = \\ = \begin{pmatrix} \lambda(1 + r_1(t)) - \mu(1 + v(t)) & \mu(1 + v(t)) - (1 + r_2(t)) \\ 0 & 0 \end{pmatrix}. \quad (36)$$

The control vector will now be

$$\mathbf{u2}(t) = (V'(t)\ P(t))^{\mathrm{T}}. \quad (37)$$

The value of the risk part of the portfolio in this case is equal to

$$V''(t) = V(t) - V'(t) + P(t) = V(t) - u2_1(t) + u2_2(t). \quad (38)$$

Tracking task

As an optimality criterion we choose a quadratic functional

$$J = \sum_{t=0}^{n-1}\left[[V(t)-F(t)]^2 + (\mathbf{u}(t))^{\mathrm{T}}\mathbf{R}(t)\mathbf{u}(t) + [V(n)-F(n)]^2\right] \rightarrow$$

$$\rightarrow \min_{u} \frac{-b \pm \sqrt{b^2-4ac}}{2a}, \qquad (39)$$

$\mathbf{R}(t)$ is a diagonal matrix of weight coefficients of dimension 2 × 2. Here $\mathbf{u}(t)$ means either $\mathbf{u1}(t)$ or $\mathbf{u2}(t)$.

The second summand in the functional (39) imposes restrictions on the size of monetary amounts that are used to buy/sell securities.

Let us write the functional (39) as follows:

$$J = \sum_{t=0}^{n-1}\left[\mathbf{z}^{\mathrm{T}}(t)\mathbf{h}^{\mathrm{T}}hz(t) + (\mathbf{u}(t))^{\mathrm{T}}\mathbf{R}(t)\mathbf{u}(t) + \mathbf{z}^{\mathrm{T}}(n)\mathbf{h}^{\mathrm{T}}\mathbf{h}z(n)\right], \quad (40)$$

where $h = [1, -1]$.

In order to determine the optimal control strategy with quadratic criterion feedback, a linear control law of the following form is used

$$\mathbf{u}(t) = K_1(t)V(t) + K_2(t)F(t) = \mathbf{K}(t)\mathbf{z}(t), \qquad (41)$$

where $\mathbf{K}(t) = [K_1(t), K_2(t)]$—the matrix of feedback coefficients—is chosen from the condition of the minimum of the functional (40).

The functional (40) can be rewritten in the form

$$J = tr\left\{\sum_{t=0}^{n-1}\left[\mathbf{h}^{\mathrm{T}}\mathbf{h}\mathbf{S}(t) + \mathbf{K}^{\mathrm{T}}(t)\mathbf{R}(t)\mathbf{K}(t)\mathbf{S}(t)\right] + \mathbf{h}^{\mathrm{T}}\mathbf{h}\mathbf{S}(n)\right\}, \quad (42)$$

where $tr\{\cdot\}$ is the trace of the matrix, and the matrix is

$$\mathbf{S}(t) = \mathbf{z}(t)\mathbf{z}^{\mathrm{T}}(t) = \begin{pmatrix} (V(t))^2 & V(t)\cdot F(t) \\ V(t)\cdot F(t) & (F(t))^2 \end{pmatrix}.$$

Equation of state

Based on (31) and (41), the dynamics of the matrix $\mathbf{S}(t) = \mathbf{z}(t)\mathbf{z}^{\mathrm{T}}(t)$ is determined by the expression:

$$\mathbf{S}(t+1) = \left[\mathbf{A}(t) + \mathbf{B}(t)\mathbf{K}(t)\right]\mathbf{S}(t)\left[\mathbf{A}(t) + \mathbf{B}(t)\mathbf{K}(t)\right]^{\mathrm{T}}. \quad (43)$$

Here, either $\mathbf{A1}(t)$, $\mathbf{B1}(t)$, or $\mathbf{A2}(t)$, $\mathbf{B2}(t)$ is taken as $\mathbf{A}(t)$ and $\mathbf{B}(t)$.

The optimal control strategy is determined by solving the system optimization problem [22, 23]. In this problem, the equation of state dynamics (43) is considered, where the matrix $\mathbf{K}(t)$ represents the control action and the functional (44) serves as a quality criterion.

In the context of this task it is required to minimize the criterion (42) under dynamic constraints, which are described by the difference matrix equation (43). To solve this problem, the maximum principle in the matrix formulation, which was developed earlier in [3, 4], is applied.

Algorithm for finding a solution

1. We find $\mathbf{Q}(t)$, $t = n, n-1, \ldots, 1, 0$ from the equation

$$\mathbf{Q}(t) = \mathbf{A}(t)\mathbf{Q}(t+1)\mathbf{A}(t) + \mathbf{A}(t)\mathbf{Q}(t+1)\mathbf{B}(t)\times$$
$$\times\left(\mathbf{R}(t) - \mathbf{B}^{\mathrm{T}}(t)\mathbf{Q}(t+1)\mathbf{B}(t)\right)^{-1}\left(\mathbf{B}^{\mathrm{T}}(t)\mathbf{Q}(t+1)\mathbf{A}(t)\right) - \mathbf{h}^{\mathrm{T}}\mathbf{h}.$$

2. Then, we calculate $\mathbf{K}(t)$, $t = 0, 1, \ldots, n-1$ in accordance with the formula

$$\mathbf{K}(t) = \left(\mathbf{R}(t) - \mathbf{B}^{\mathrm{T}}(t)\mathbf{Q}(t+1)\mathbf{B}(t)\right)^{-1}\left(\mathbf{B}^{\mathrm{T}}(t)\mathbf{Q}(t+1)\mathbf{A}(t)\right).$$

3. By found $\mathbf{K}(t)$, we calculate $\mathbf{S}(t)$, $t = 1, 2, \ldots, n$, where $\mathbf{S}(t) = \begin{pmatrix} (V(t))^2 & V(t)F(t) \\ V(t)F(t) & (F(t))^2 \end{pmatrix}$.

The elements of the matrix $\mathbf{S}(t)$ and the matrix $\mathbf{K}(t)$ are the desired solution to the benchmark portfolio tracking problem.

Knowing the matrix $\mathbf{S}(t)$, we have:

$$F(t) = \sqrt{S_{22}(t)};\ V(t) = S_{12}(t)/F(t),$$

where $V(t)$ is the investments in the real portfolio.

The portfolio management is calculated by the formula

$$\mathbf{u}(t) = K_1(t)V(t) + K_2(t)F(t).$$

4. In order to calculate the amount of investment in the portfolio, it is necessary to solve the system of relations:

$$\begin{aligned}&\mathbf{u}(t) = K_1(t)V(t) + K_2(t)F(t) = \mathbf{K}(t)\mathbf{z}(t),\ t = \overline{0, n-1};\\ &\mathbf{z}(t+1) = \mathbf{A}(t)\mathbf{z}(t) + \mathbf{B}(t)\mathbf{u}(t).\end{aligned} \quad (44)$$

Here for Model 1: $\mathbf{A}(t) = \mathbf{A1}(t)$, $\mathbf{B}(t) = \mathbf{B1}(t)$, $\mathbf{z}(t) = \begin{pmatrix} V(t) \\ F(t) \end{pmatrix}$, $\mathbf{u}(t) = \begin{pmatrix} V''(t) \\ P(t) \end{pmatrix}$; for Model 2: $\mathbf{A}(t) = \mathbf{A2}(t)$, $\mathbf{B}(t) = \mathbf{B2}(t)$, $\mathbf{z}(t) = \begin{pmatrix} V(t) \\ F(t) \end{pmatrix}$, $\mathbf{u}(t) = \begin{pmatrix} V'(t) \\ P(t) \end{pmatrix}$.

5. The RLA investments is calculated by (33) (for Model 1)

$$V'(t) = V(t) - V''(t) + P(t)$$

or the RA investments in is calculated by (38) (for Model 2)

$$V''(t) = V(t) - V'(t) + P(t).$$

6. Let us calculate the RA and RLA shares in the portfolio

$$x_t = \begin{cases} [V'(t)/C'(t)], & \text{if } V'(t) \geq 0, \\ [|V'(t)|/(\lambda C'(t))], & \text{if } V'(t) < 0, \end{cases} \quad t = \overline{0,n}, \quad (45)$$

$$y_t = \begin{cases} [V''(t)/C''(t)], & \text{if } V''(t) \geq 0, \\ [|V''(t)|/(\mu C''(t))], & \text{if } V''(t) < 0, \end{cases} \quad t = \overline{0,n}. \quad (46)$$

Here [·] is the integer part of the number.

NUMERICAL MODELING RESULTS

RA prices were modeled on the basis of a mixture of two normal distributions with parameters:

$$m1_j = C01 + h1 \cdot j, \; j = \overline{0,500}, \; \sigma 1;$$

$$m2_j = C02 + h2 \cdot j, \; j = \overline{0,500}, \; \sigma 2,$$

where $m1_j$, $m2_j$ are the price distributions; C01, C02 is the initial price for the corresponding distribution; $h1$, $h2$ are possible price fluctuations at a given sample size; σ1 and σ2 are standard deviations.

The parameter values were as follows: C01 = 100, C02 = 90, $h1$ = 0.04, $h2$ = 0,02, σ1 = 10, σ2 = 15.

Figure 3 shows the calculated RA prices. Hereinafter monetary values are shown in conventional units.

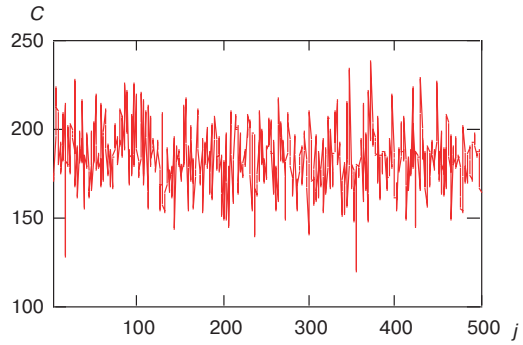


Fig. 3. RA prices. *C* is the RA price, units of money; *j* is the number of realizations

Figure 4 shows the obtained probability distribution of RA prices.

This distribution is treated as an analog of the empirical distribution. Then RA prices at the tree nodes are modeled based on this distribution. The probability of price growth was estimated based on the constructed distribution for the price difference $C''_i - C''_1$, $i = \overline{1,500}$. The probability of price growth (the probability that $C''_i - C''_1 > 0$) was $p = 0.495$.

The values of the other parameters were as follows: $C''(0) = 150$, $C'(0) = 10$, $\lambda = 1.02$, $\mu = 1.02$, $F(0) = 10000$, $V(0) = 10000$, $P(0) = 0$, $r_1(t) = 0.02$, $r_2(t) = 0.015$, $\mu_0(t) = 0.02$, tree depth $n = 5$.

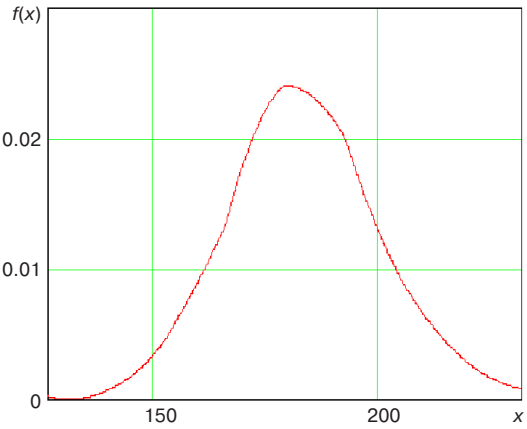


Fig. 4. Probability distribution of RA prices. *f*(*x*) is the probability distribution density; *x* is the RA price, unit of money

The values of investments were taken as follows. For Model 1, $V'(0) = 10000$, $V''(0) = 0$, i.e., at the initial moment of time all funds were invested in RLA. For Model 2, $V'(0) = 0$, $V''(0) = 10000$, i.e., all funds were invested in RA.

Modeling results are given in Tables 1–5.

Figure 5 shows graphs of tracking the desired portfolio value. Here and further in the graphs time t is given in relative units.

Figure 6 shows the necessary changes in the RA investments to achieve the portfolio value not less than the desired one.

Figure 7 shows the necessary changes in the RLA investments to achieve the portfolio value not less than the desired one.

Figure 8 shows the tracking for the payment function (desired portfolio value).

Figure 9 shows the necessary changes in the RA investments to achieve the portfolio value not less than the desired one.

Figure 10 shows the necessary changes in the RLA investments to achieve the portfolio value not less than the desired one.

Figure 11 shows the tracking for the payment function (desired portfolio value).

Figure 12 shows the necessary changes in RA investments to achieve the portfolio value not less than the desired one.

Figure 13 shows the necessary changes in the RLA investments in to achieve the portfolio value not less than the desired one.

Figure 14 shows the tracking for the payment function (desired portfolio value).

Figure 15 shows the necessary changes in RA investments to achieve the portfolio value not less than the desired one.

Figure 16 shows the necessary changes in the RLA investments to achieve the portfolio value not less than the desired one.

Table 1. Tree depth up to $n = 1$

Tree depth	Model 1				Model 2			
	Investments in the portfolio	RLA share	RA share	Deposit	Investments in the portfolio	RLA share	RA share	Deposit
0	10000	1	0	0	10000	0	1	0
1	10340	1.028	−0.028	−0.368	11750	0.725	0.272	−35.890

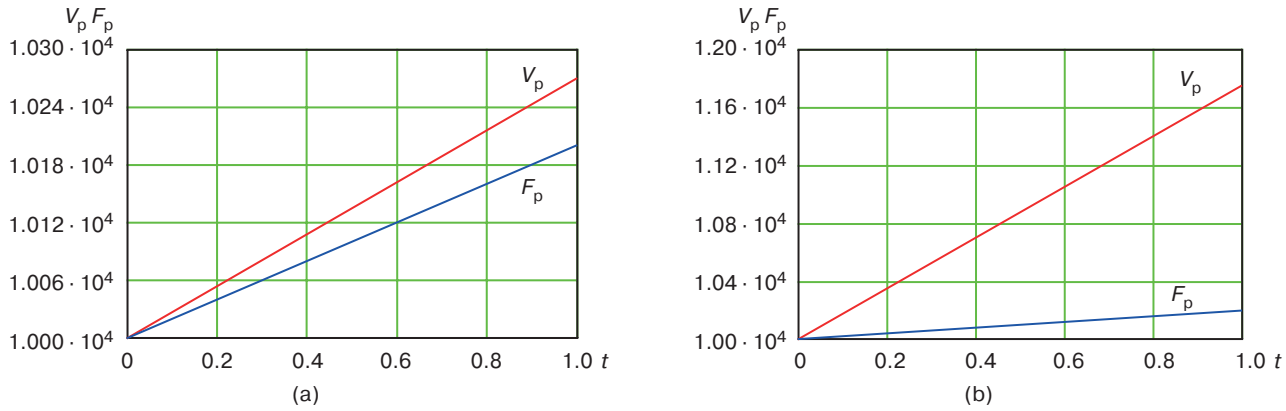


Fig. 5. Tracking the desired portfolio value: (a) Model 1, (b) Model 2.
V_p is the investments or portfolio capital, units of money; F_p is the payment function, units of money; t is time, arb. units

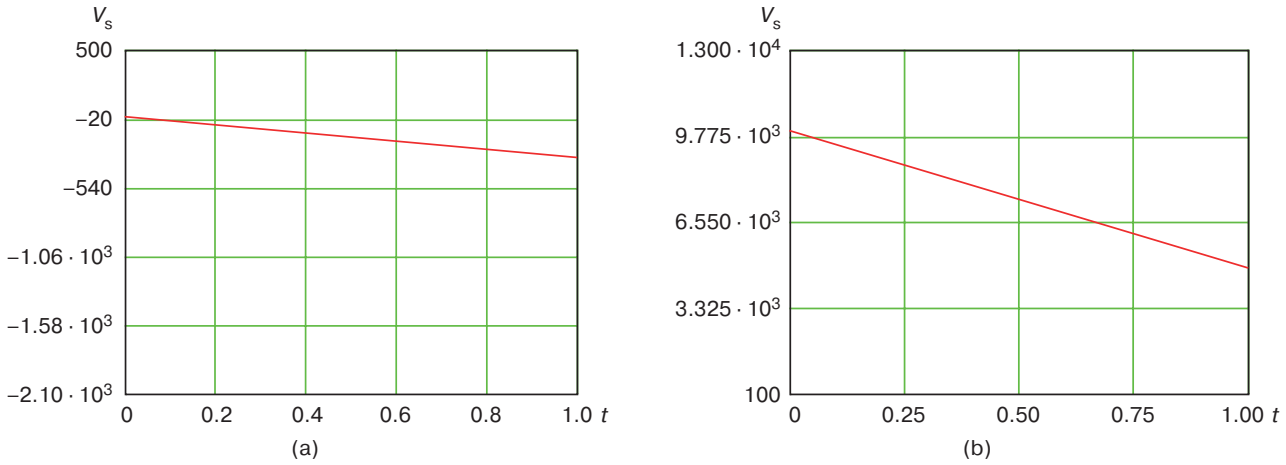


Fig. 6. Necessary changes in the RA investments: (a) Model 1, (b) Model 2.
V_s is the RA investments, units of money; t is time, arb. units

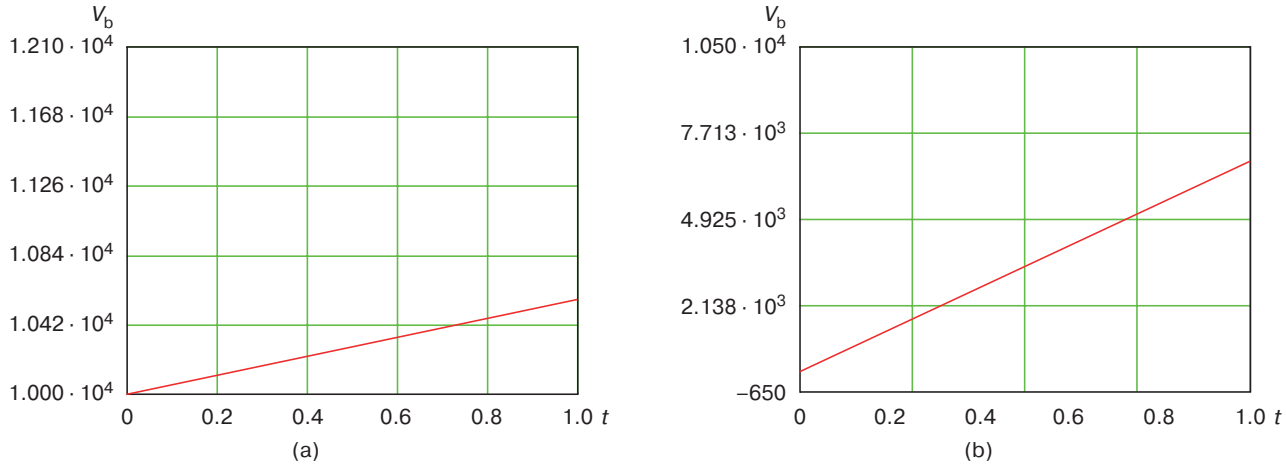


Fig. 7. Necessary changes in the RLA investments: (a) Model 1, (b) Model 2.
V_b is the investments in RLA, units of money; t is time, arb. units

Table 2. Tree depth up to $n = 2$

Tree depth	Model 1				Model 2			
	Investments in the portfolio	RLA share	RA share	Deposit	Investments in the portfolio	RLA share	RA share	Deposit
0	10000	1	0	0	10000	0	1	0
1	10260	1.067	−0.067	−0.874	11400	0.427	0.568	−54.988
2	10690	1.079	−0.079	−1.089	11820	0.698	0.294	−93.140

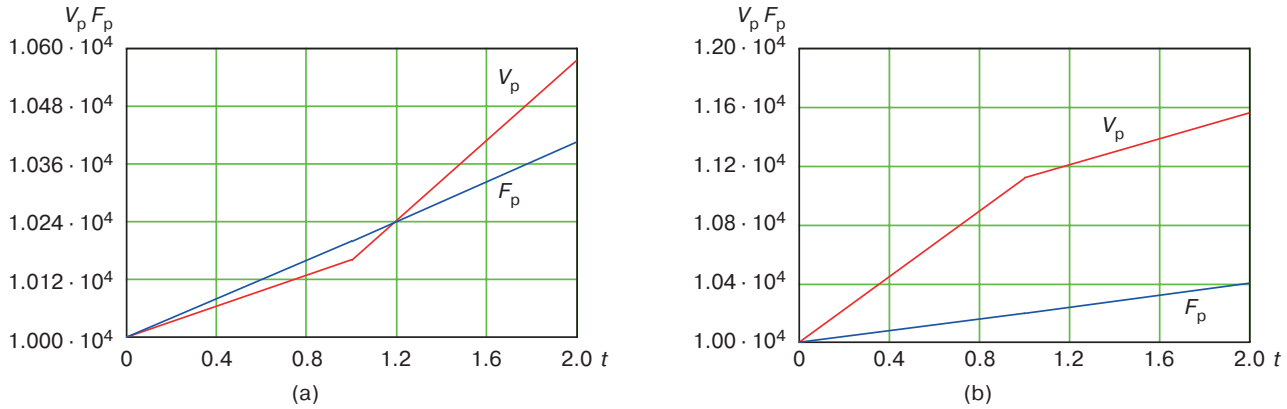


Fig. 8. Tracking for the payment function: (a) Model 1, (b) Model 2.
V_p is the investments or portfolio capital, units of money; F_p is the payment function, units of money; t is time, arb. units

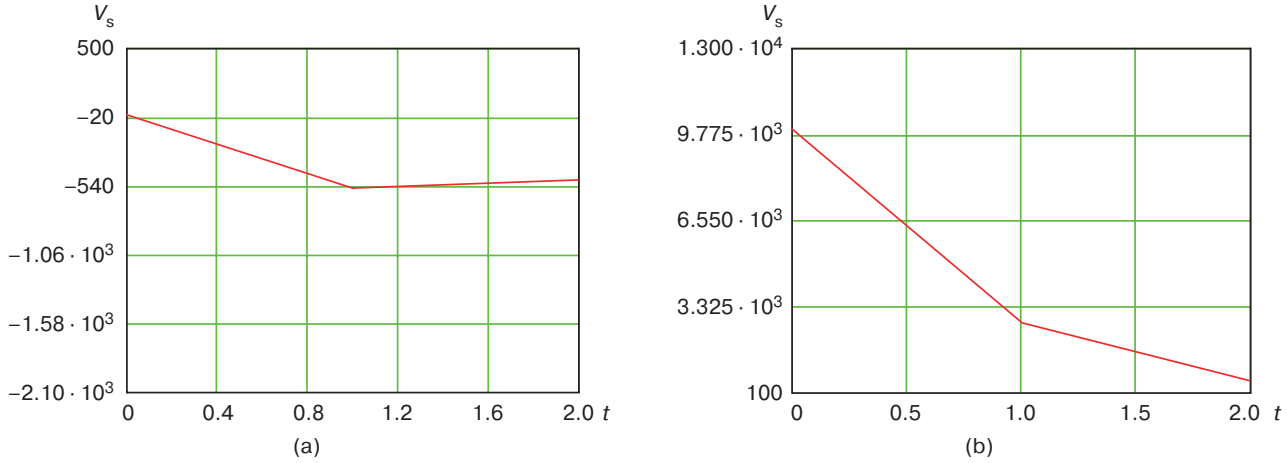


Fig. 9. Necessary changes in the RA investments: (a) Model 1, (b) Model 2.
V_s is the RA investments, units of money; t is time, arb. units

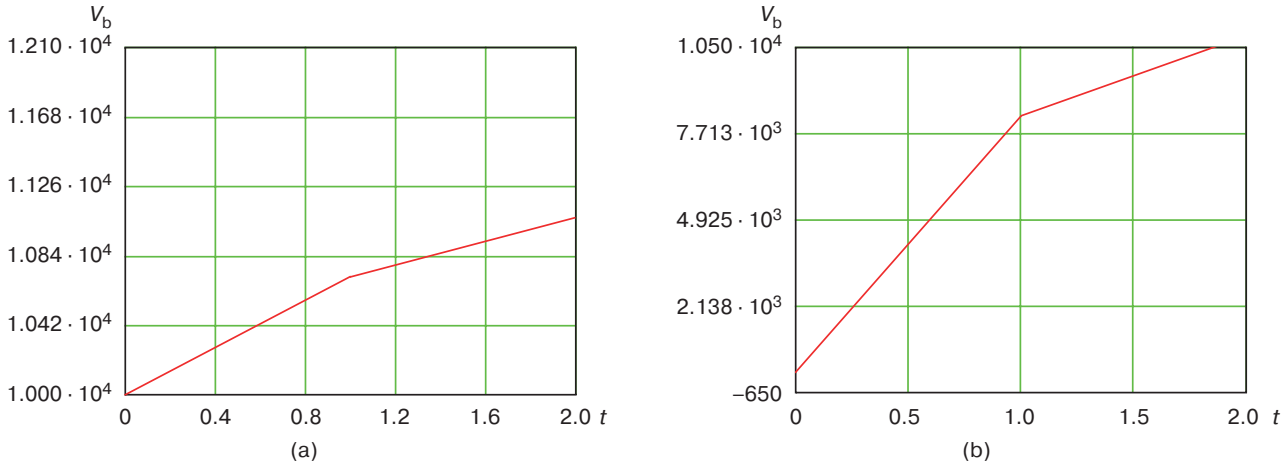


Fig. 10. Necessary changes in the RLA investments:
(a) Model 1, (b) Model 2. V_b is the RLA investments, units of money; t is time, arb. units

Table 3. Tree depth up to $n = 3$

Tree depth	Model 1				Model 2			
	Investments in the portfolio	RLA share	RA share	Deposit	Investments in the portfolio	RLA share	RA share	Deposit
0	10000	1	0	0	10000	0	1	0
1	10180	1.110	−0.111	−1.484	11190	0.538	0.456	−67.938
2	10610	1.100	−0.103	−1.450	11610	0.814	0.177	−106.597
3	11030	0.975	0.018	−2.705	11850	−0.024	1.024	−3.375

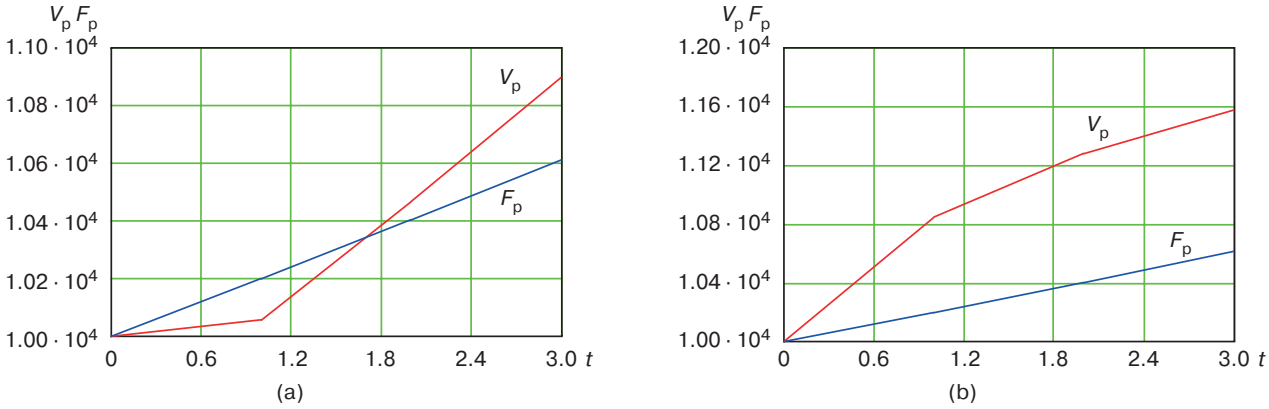


Fig. 11. Tracking for the payment function: (a) Model 1, (b) Model 2.
V_p is the investments or portfolio capital, units of money; F_p is the payment function, units of money; t is time, arb. units

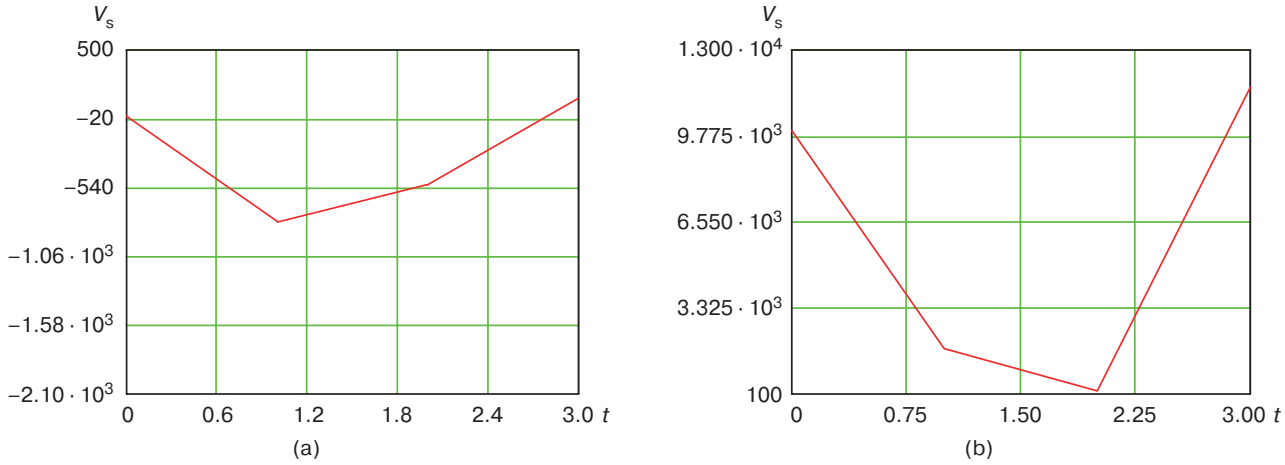


Fig. 12. Necessary changes in RA investments: (a) Model 1, (b) Model 2.
V_s is the RA investments, units of money; t is time, arb. units

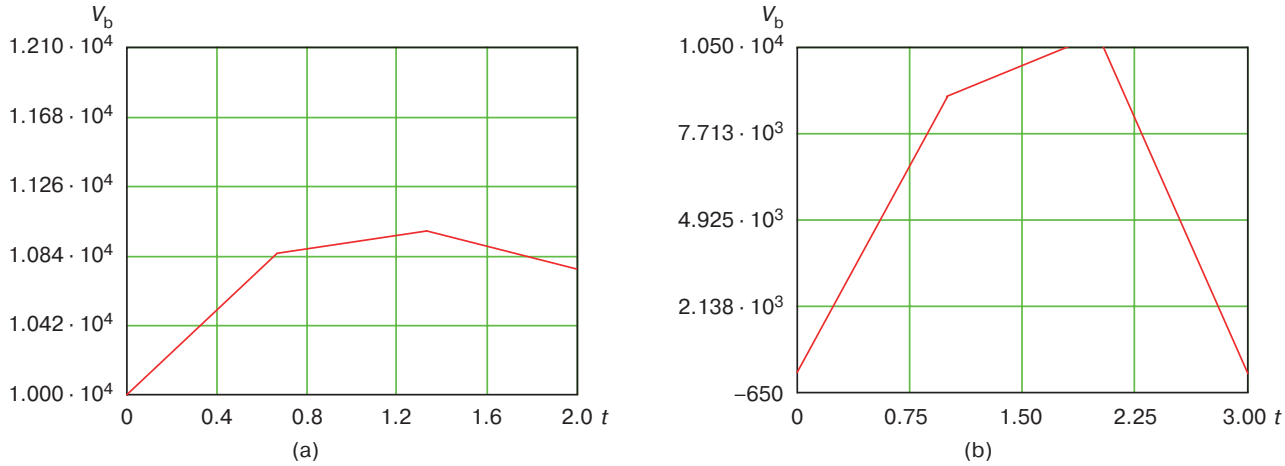


Fig. 13. Necessary changes in the RLA investments: (a) Model 1, (b) Model 2.
V_b is the RLA investments, units of money; t is time, arb. units

Table 4. Tree depth up to $n = 4$

Tree depth	Model 1				Model 2			
	Investments in the portfolio	RLA share	RA share	Deposit	Investments in the portfolio	RLA share	RA share	Deposit
0	10000	1	0	0	$1 \cdot 10^4$	0	1	0
1	10080	1.159	−0.159	−2.078	$1.057 \cdot 10^4$	0.913	0.078	−100.38
2	10510	1.119	−0.119	−1.653	$1.1 \cdot 10^4$	0.977	0.013	−111.746
3	10930	0.970	0.030	−4.141	$1.124 \cdot 10^4$	−0.027	1.026	−1.617
4	11370	0.977	0.023	−3.411	$1.146 \cdot 10^4$	$3.502 \cdot 10^{-3}$	0.996	−3.804

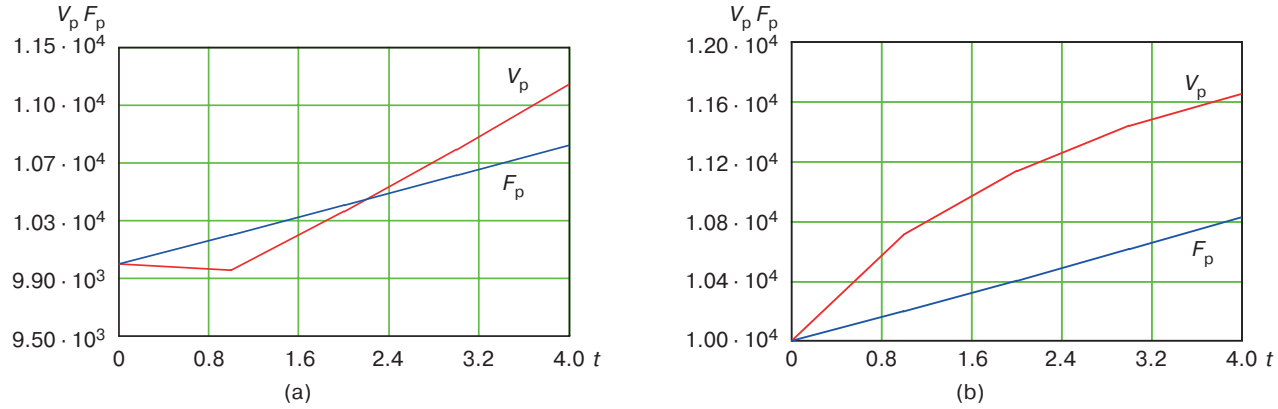


Fig. 14. Tracking for the payment function: (a) Model 1, (b) Model 2.
V_p is the investments or portfolio capital, units of money; F_p is the payment function, units of money; t is time, arb. units

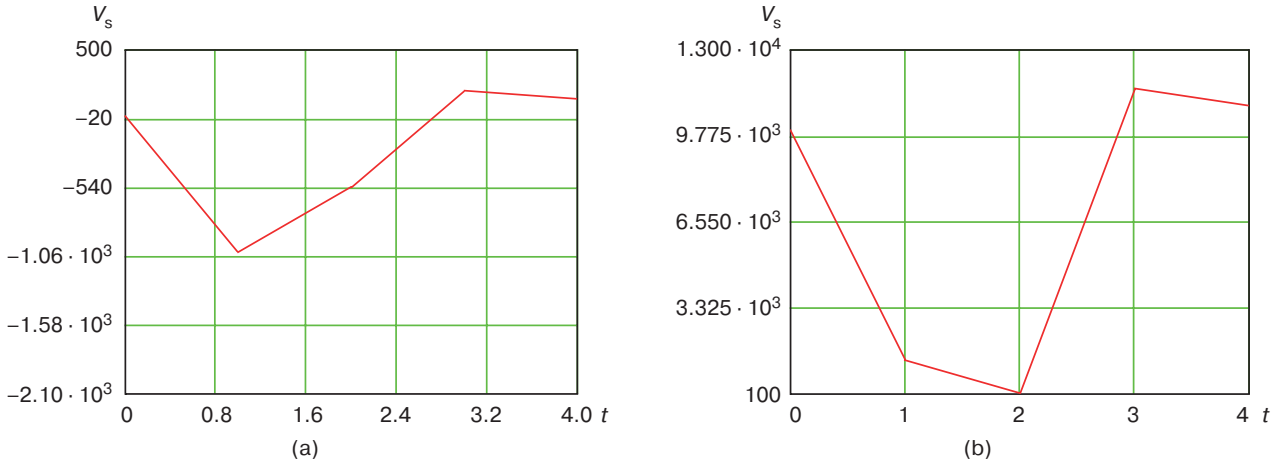


Fig. 15. Necessary changes in the RA investments: (a) Model 1, (b) Model 2.
V_s is the RA investments, units of money; t is time, arb. units

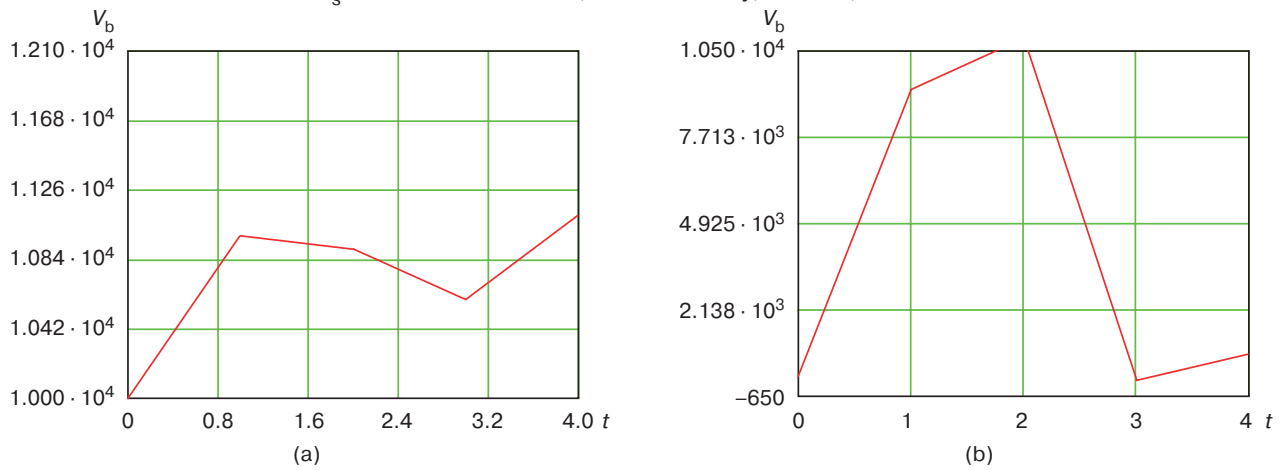


Fig. 16. Necessary changes in the RLA investments: (a) Model 1, (b) Model 2.
V_b is the RLA investments, units of money; t is time, arb. units

Table 5. Tree depth up to $n = 5$

Tree depth	Model 1				Model 2			
	Investments in the portfolio	RLA share	RA share	Deposit	Investments in the portfolio	RLA share	RA share	Deposit
0	10000	1	0	0	10000	0	1	0
1	9815	1.095	−0.095	−0.379	10480	0.934	0.056	−101.88
2	10220	1.034	−0.034	−0.140	10910	0.979	0.011	−111.062
3	10630	0.971	0.028	−3.587	11130	−0.025	1.025	−2.593
4	11060	0.976	0.023	−3.602	11350	0.003	0.996	−4.519
5	11500	0.980	0.019	−2.986	11580	−0.01	1.01	−1.555

Figure 17 shows the tracking for the payment function (desired portfolio value).

Figure 18 shows the necessary changes in RA investments to achieve the portfolio value not less than the desired one.

Figure 19 shows the necessary changes in the RLA investments to achieve the portfolio value not less than the desired one.

Negative deposit shares observed in Tables 1–5 can be interpreted as "short sales." Negative deposit fractions present in Tables 1–5 mean borrowing of funds. Such results within the framework of this dynamic model are explained by the fact that no restrictions were imposed on the investments and deposits.

It can be seen that portfolio reforming according to the dynamic model allows us to provide a given level of the payment function.

It is of interest to compare the error of the dynamic model based on the tree structure of RA price changes with the general model of RA price changes [13]. Tables 6 and 7 summarize the errors of investment estimation. Here σV is the error of portfolio value; σV_s is the error of RA investment; σV_b is the error of RLA investment; σx is the error of RA quantity; σy is the error of RLA quantity.

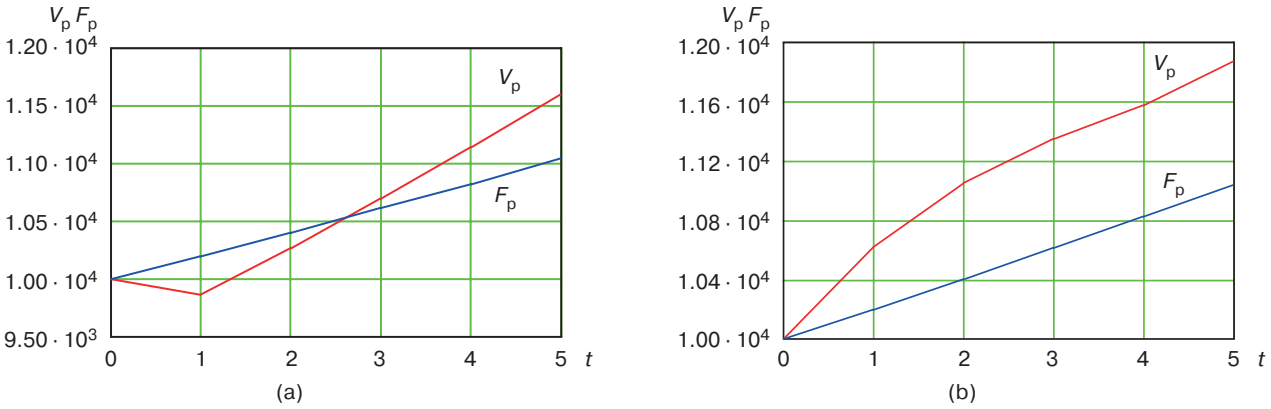


Fig. 17. Tracking for the payment function: (a) Model 1, (b) Model 2.
V_p is the investments or portfolio capital, units of money; F_p is the payment function, units of money; t is time, arb. units

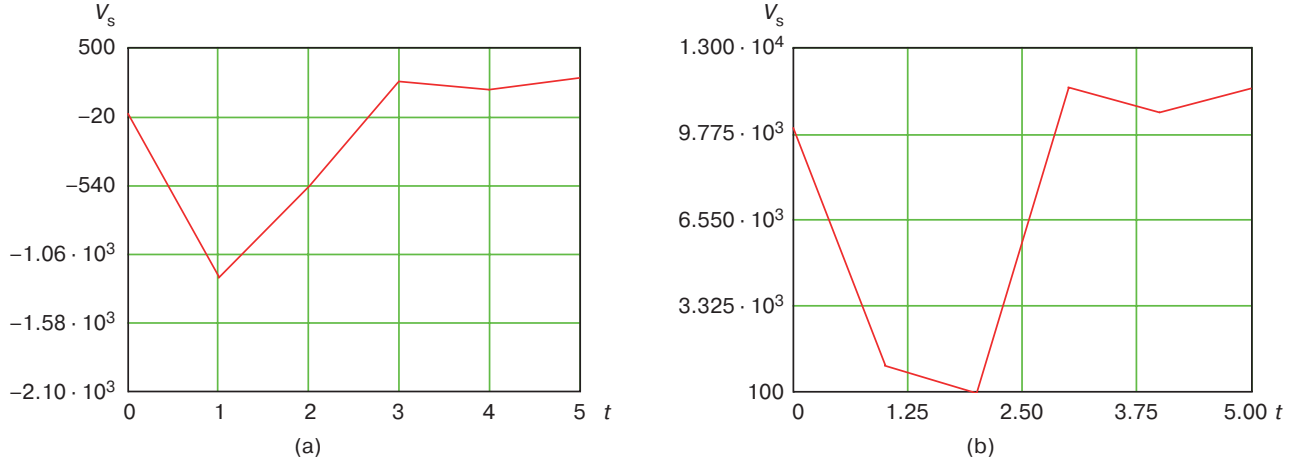


Fig. 18. Necessary changes in RA investments:
(a) Model 1, (b) Model 2. V_s is the RA investments, units of money; t is time, arb. units

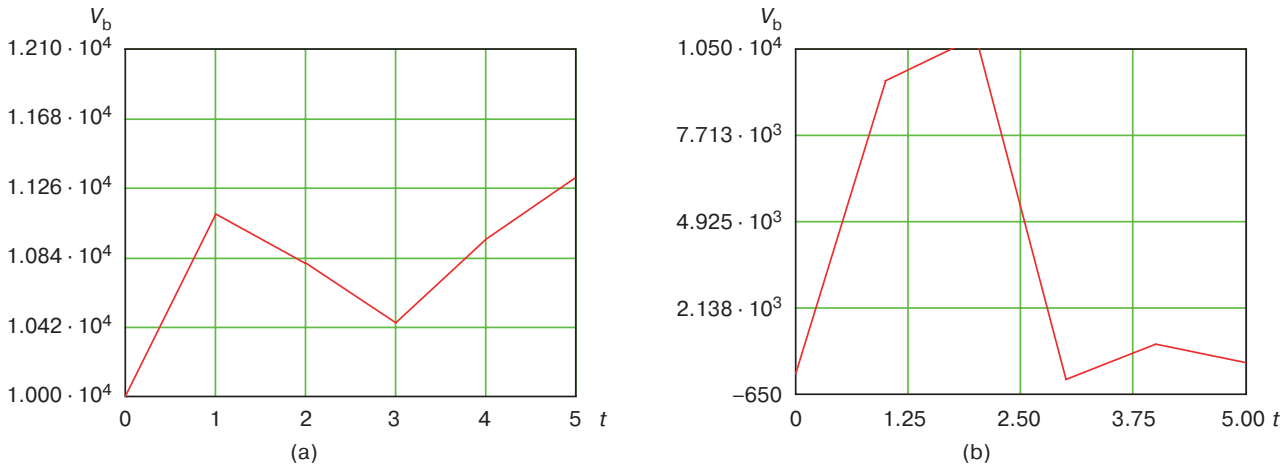


Fig. 19. Necessary changes in the RLA investments. (a) Model 1, (b) Model 2. V_b is the RLA investments, units of money; t is time, arb. units

Table 6. Estimation of the Model 1 error

Tree depth / investment horizon n	Tree structure of RA prices			Natural change in RA prices			Gain in model accuracy		
	σV	σV_s	σV_b	$\sigma V1$	σV_s1	σV_b1	$\frac{\sigma V1}{\sigma V}$	$\frac{\sigma V_s1}{\sigma V_s}$	$\frac{\sigma V_b1}{\sigma V_b}$
1	6.44	61.5	61.7	15.4	84.9	83.3	2.4	1.4	1.3
2	12.3	26.3	21.9	21.25	37.0	34.0	1.7	1.4	1.6
3	18.64	417	410	39.0	581	575	2.1	1.4	1.4
4	25.1	340	354	61.49	625	654	2.4	1.8	1.8
5	26.9	234	238	73.7	552	556	2.7	2.4	2.3

Table 7. Estimation of the Model 2 error

Tree depth / investment horizon n	Tree structure of RA prices			Natural change in RA prices			Gain in model accuracy		
	σV	σV_s	σV_b	$\sigma V2$	σV_s2	σV_b2	$\frac{\sigma V2}{\sigma V}$	$\frac{\sigma V_s2}{\sigma V_s}$	$\frac{\sigma V_b2}{\sigma V_b}$
1	194.4	5003	302	339.7	875	531	1.7	1.7	1.7
2	210.3	52.8	157	366.2	98.2	269	1.7	1.9	1.7
3	370.3	981	637	743.6	1726	1083	2.0	1.8	1.7
4	367.8	560	493	781.0	1900	1959	2.1	3.4	4.0
5	370.0	559	280	843.2	1685	1288	2.3	3.0	4.6

As follows from Table 6, the error of the dynamic model based on the tree structure of RA price changes is smaller than for the conventional model, and the gain in accuracy of the model increases with increasing investment horizon. Thus, for the portfolio value the gain in model accuracy varies from 2.4 (n = 1) to 2.7 (n = 5); for the RA investments, the gain in accuracy varies from 1.4 (n = 1) to 2.4 (n = 5); for the RLA investments, the gain in accuracy varies from 1.3 (n = 1) to 2.3 (n = 5).

Accuracy gains are also observed for Model 2. For example, the gain in accuracy of portfolio value estimation ranges from 1.7 (n = 1) to 2.3 (n = 5).

CONCLUSIONS

The study analyzes a securities portfolio comprising assets having different risk levels, as well as risk-free assets and deposits. A binomial model was used to model the RA price structure. The main objective of the study was to develop a management model for tracking the benchmark portfolio. For this purpose, a quality criterion in the form of a quadratic function served as the basis for the construction of the management model. The developed model belongs to the class of dynamic programming models for determining the optimal management strategy using feedback.

The study considers two approaches to portfolio formation. The first approach involves initial investment of capital in RLA and subsequent management carried out through RA. The second approach, conversely, includes initial capital investment in RA, with management is carried out through RLA. In order to optimize management for achieving the desired objective, the study applies a linear control law to determine the optimal values of the control parameters based on the current state of the system and the target value of the benchmark portfolio.

By using the described dynamic management model based on the tree structure of RA prices the accuracy of evaluating investments in the portfolio can be significantly increased. For the first approach (Model 1), there is an increase in evaluation accuracy from 2.4 to 2.7 times, while the second approach (Model 2) increases evaluation accuracy from 1.7 to 2.7 times.

Thus, the developed model can become a useful tool for financial analysts and investors by allowing them to make better informed decisions when forming and managing a securities portfolio. The model can be used to carry out a more accurate determination of the optimal amount of investments, leading to higher investment efficiency and better results for investors.

Authors' contributions

A.A. Mitsel—concept, structure, and scientific leadership of the study.

E.V. Viktorenko—analysis and interpretation of data, writing and editing the text of the manuscript.

REFERENCES

1. Dombrovskii V.V., Lyashenko E.A. A Linear Quadratic Control for Discrete Systems with Random Parameters and Multiplicative Noise and Its Application to Investment Portfolio Optimization. *Autom. Remote Control*. 2003;64(10): 1558–1570. https://doi.org/10.1023/A:1026057305653
   [Original Russian Text: Dombrovskii V.V., Lyashenko E.A. A Linear Quadratic Control for Discrete Systems with Random Parameters and Multiplicative Noise and Its Application to Investment Portfolio Optimization. *Avtomatika i telemekhanika*. 2003;10:50–65 (in Russ.).]
2. Dombrovskii V.V., Lyashenko E.A. Dynamic model of investment portfolio management in the financial market with stochastic volatility with regard transaction costs and restrictions. *Vestnik Tomskogo gosudarstvennogo universiteta = Tomsk State University J.* 2006;S16:217–225. (in Russ.).
3. Dombrovskii V.V., Dombrovskii D.V., Lyashenko E.A. Predictive control of random-parameter systems with multiplicative noise. Application to investment portfolio optimization. *Autom. Remote Control*. 2005;66(4):583–595. https://doi.org/10.1007/s10513-005-0102-5
   [Original Russian Text: Dombrovskii V.V., Dombrovskii D.V., Lyashenko E.A. Predictive control of random-parameter systems with multiplicative noise. Application to investment portfolio optimization. *Avtomatika i telemekhanika.* 2005;4: 84–97 (in Russ.).]
4. Gerasimov E.S., Dombrovskii V.V. Dynamic network model of investment management control for quadratic risk function. *Autom. Remote Control*. 2002;63(2):280–288. https://doi.org/10.1023/A:1014251725737
   [Original Russian Text: Gerasimov E.S., Dombrovskii V.V. Dynamic network model of investment management control for quadratic risk function. *Avtomatika i telemekhanika.* 2002;2:119–128 (in Russ.).]
5. Dombrovskii V.I., Galperin V.A. Dynamic model of investments portfolio selection by quadratic risk function. *Vestnik Tomskogo gosudarstvennogo universiteta = Tomsk State University J.* 2000;269:73–75 (in Russ.).
6. Galperin V.A., Dombrovskii V.I. Dynamic management of a self-financing investment portfolio with a quadratic risk function in discrete time. *Vestnik Tomskogo gosudarstvennogo universiteta = Tomsk State University J.* 2002;(S1-1):141–146 (in Russ.).
7. Dombrovskii V.I., Galperin V.A. Investment portfolio management in continuous time with a quadratic risk function. In: *Proceedings of the 10th Anniversary Symposium on Nonparametric and Robust Statistical Methods in Cybernetics.* Tomsk: TSU; 2004. P. 185–192 (in Russ.). https://elibrary.ru/xwjkax
8. Galperin V.A., Dombrovskii V.I. Dynamic management of an investment portfolio taking into account abrupt changes in prices of financial assets. *Vestnik Tomskogo gosudarstvennogo universiteta = Tomsk State University J.* 2003;280:112–117 (in Russ.).
9. Dombrovskii V.V., Dombrovskii D.V., Lyashenko E.A. Dynamic optimization of the investment portfolio under restrictions on the volume of investments in financial assets. *Vestnik Tomskogo gosudarstvennogo universiteta = Tomsk State University J.* 2008;1:13–17 (in Russ.).
10. Dombrovskii V.V., Pashinskaya T.Yu. Predictive control strategies for investment portfolio in the financial market with hidden regime switching. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie vychislitelnaja tehnika i informatika = Tomsk State University Journal of Control and Computer Science*. 2020;50:4–13 (in Russ.).
11. Grineva N.V. Dynamic optimization of the investment portfolio management trajectory. *Problemy ekonomiki i yuridicheskoi praktiki = Economic Problems and Legal Practice*. 2021;17(3):73–77. https://doi.org/10.33693/2541-8025-2021-17-3-73-77
12. Ivanyuk V. Proposed Model of a Dynamic Investment Portfolio with an Adaptive Strategy. *Mathematics.* 2022;10(23):4394. https://doi.org/10.3390/math10234394

13. Mitsel A.A., Krasnenko N.P. Dynamic model of investment portfolio management with linear criterion of quality. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki* (*Doklady TUSUR*) = *Proceedings of TUSUR University*. 2014;34:176–182 (in Russ.).
14. Kolyasnikova E.R. Hedging strategy in the (B, S, F)-market model. *Obozrenie prikladnoi i promyshlennoi matematiki* = *OP&PM Surveys of Applied and Industrial Mathematics*. 2009;16(3):467–468 (in Russ.).
15. Bronshtein E.M., Kolyasnikova E.R. The (B, S, F)-market Model and hedging strategies. *Upravlenie riskom* = *Management of Risk*. 2010;2:55–64 (in Russ.).
16. Bronshtein E.M., Kolyasnikova E.R. Approximate hedging strategy in the (B, S, F)-market model. *Matematicheskoe modelirovanie* = *Math. Model.* 2010;22(11):29–38 (in Russ.).
17. Davnis V.V., Bogdanova S.Yu., Suyunova G.B. Models of (B, S)-market and risk-neutral price of options. *Vestnik OrelGIET* = *OrelSIET Bulletin*. 2010;1:134–140 (in Russ.).
18. Davnis V.V., Fedoseev A.M. Adaptive model-bulding of (B, S)-market. *Sovremennaya ekonomika: problemy i resheniya* = *Modern Economics: Problems and Solutions*. 2011;6(18):202–213 (in Russ.).
19. Fedoseev A.M., Korotkikh V.V. Features valuation of options on complete and incomplete markets. *Sovremennaya ekonomika: problemy i resheniya* = *Modern Economics: Problems and Solutions*. 2011;4(16):137–144 (in Russ.).
20. Almeida C., Freire G. Pricing of index options in incomplete markets. *J. Fin. Economic*. 2022;144(1):174–205. https://doi.org/10.1016/j.jfineco.2021.05.041
21. Davnis V.V., Davnis V.V. Econometric options for the (B, S, I)-market models. *Sovremennaya ekonomika: problemy i resheniya* = *Modern Economics: Problems and Solutions*. 2013;10(46):154–165 (in Russ.). Available from URL: https://journals.vsu.ru/meps/article/view/7987
22. Krotov V.F., Lagosha B.A., Lobanov S.M., Danilov N.I., Sergeev S.I. *Osnovy teorii optimal'nogo upravleniya* (*Fundamentals of Optimal Control Theory*). Moscow: Vysshaya shkola; 1990. 430 p. (in Russ.).
23. Athans M. The Matrix Minimum Principle. *Information and Control*. 1967;11(5–6):592–606. https://doi.org/10.1016/S0019-9958(67)90803-0

СПИСОК ЛИТЕРАТУРЫ

1. Домбровский В.В., Ляшенко Е.А. Линейно-квадратичное управление дискретными системами со случайными параметрами и мультипликативными шумами с применением к оптимизации инвестиционного портфеля. *Автоматика и телемеханика*. 2003;10:50–65.
2. Домбровский В.В., Ляшенко Е.А. Модель управления инвестиционным портфелем на финансовом рынке со стохастической волатильностью с учетом трансакционных издержек и ограничений. *Вестник Томского государственного университета*. 2006;S16:217–225.
3. Домбровский В.В., Домбровский Д.В., Ляшенко Е.А. Управление с прогнозированием системами со случайными параметрами и мультипликативными шумами и применение к оптимизации инвестиционного портфеля. *Автоматика и телемеханика*. 2005;4:84–97.
4. Герасимов Е.С., Домбровский В.В. Динамическая сетевая модель управления инвестициями при квадратичной функции риска. *Автоматика и телемеханика*. 2002;2:119–128.
5. Домбровский В.И., Гальперин В.А. Динамическая модель управления инвестиционным портфелем при квадратической функции риска. *Вестник Томского ГУ.* 2000;269:73–75.
6. Гальперин В.А., Домбровский В.И. Динамическое управление самофинансируемым инвестиционным портфелем при квадратической функции риска в дискретном времени. *Вестник Томского ГУ.* 2002;(S1-1):141–146.
7. Домбровский В.И., Гальперин В.А. Управление инвестиционным портфелем в непрерывном времени при квадратической функции риска. В сб.: *Труды Десятого юбилейного симпозиума по непараметрическим и робастным статистическим методам в* кибернетике. Томск: ТГУ; 2004. С. 185–192. https://elibrary.ru/xwjkax
8. Гальперин В.А., Домбровский В.И. Динамическое управление инвестиционным портфелем с учетом скачкообразного изменения цен финансовых активов. *Вестник Томского ГУ.* 2003;280:112–117.
9. Домбровский В.В., Домбровский Д.В., Ляшенко Е.А. Динамическая оптимизация инвестиционного портфеля при ограничениях на объемы вложений в финансовые активы. *Вестник ТГУ*. 2008;1:13–17.
10. Домбровский В.В., Пашинская Т.Ю. Стратегии прогнозирующего управления инвестиционным портфелем на финансовом рынке со скрытым переключением режимов. *Вестник ТГУ. Управление, вычислительная техника и информатика*. 2020;50:4–13.
11. Гринева Н.В. Динамическая оптимизация траектории управления инвестиционным портфелем. *Проблемы экономики и юридической практики*. 2021;17(3):73–77.
12. Ivanyuk V. Proposed Model of a Dynamic Investment Portfolio with an Adaptive Strategy. *Mathematics.* 2022;10(23):4394. https://doi.org/10.3390/math10234394
13. Мицель А.А., Красненко Н.П. Динамическая модель управления инвестиционным портфелем с линейным критерием качества. *Доклады Томского государственного университета систем управления и радиоэлектроники* (*Доклады ТУСУР*). 2014;4(34):176–182.
14. Колясникова Е.Р. Хеджирующая стратегия в модели (B, S, F)-рынка. *Обозрение прикладной и промышленной математики*. 2009;16(3):467–468.

15. Бронштейн Е.М., Колясникова Е.Р. Модель (B, S, F)-рынка и хеджирующие стратегии. *Управление риском*. 2010;2:55–64.
16. Бронштейн Е.М., Колясникова Е.Р. Приближенные хеджирующие стратегии в модели (B, S, F)-рынка. *Математическое моделирование*. 2010;22(11):29–38.
17. Давнис В.В., Богданова С.Ю., Суюнова Г.Б. Модели (B, S)-рынка и риск-нейтральная цена опционов. *Вестник ОрёлГИЭТ*. 2010;1:134–140.
18. Давнис В.В., Федосеев А.М. Адаптивное моделирование (B, S)-рынка. *Современная экономика: проблемы и решения*. 2011;6(18):202–213.
19. Федосеев А.М., Коротких В.В. Особенности оценки стоимости опционов на полном и неполных рынках. *Современная экономика: проблемы и решения*. 2011;4(16):137–144.
20. Almeida C., Freire G. Pricing of index options in incomplete markets. *J. Fin. Economic*. 2022;144(1):174–205. https://doi.org/10.1016/j.jfineco.2021.05.041
21. Давнис В.В., Коротких В.В. Эконометрические варианты модели (B, S, I)-рынка. *Современная экономика: проблемы и решения*. 2013;10(46):154–165. URL: https://journals.vsu.ru/meps/article/view/7987
22. Кротов В.Ф., Лагоша Б.А., Лобанов С.М., Данилов Н.И., Сергеев С.И. *Основы теории оптимального управления.* М.: Высшая школа; 1990. 430 с.
23. Athans M. The Matrix Minimum Principle. *Information and Control*. 1967;11(5–6):592–606. https://doi.org/10.1016/S0019-9958(67)90803-0

About the authors

Artur A. Mitsel, Dr. Sci. (Eng.), Professor, Department of Automated Control Systems, Tomsk State University of Control Systems and Radioelectronics (40, Lenina pr., Tomsk, 634050 Russia). E-mail: artur.a.mitsel@tusur.ru. Scopus Author ID 6603150769, ResearcherID G-8307-2014, RSCI SPIN-code 9698-2160, https://orcid.org/0000-0002-2624-4383

Elena V. Viktorenko, Senior Lecturer, Postgraduate Student, Department of Economics, Tomsk State University of Control Systems and Radioelectronics (40, Lenina pr., Tomsk, 634050 Russia). E-mail: viktorenko.e@gmail.com. ResearcherID AEJ-4949-2022, RSCI SPIN-code 8664-3235, https://orcid.org/0000-0003-3871-8993

Об авторах

Мицель Артур Александрович, д.т.н., профессор, кафедра автоматизированных систем управления, ФГАОУ ВО «Томский государственный университет систем управления и радиоэлектроники» (634050, Россия, Томск, пр-т Ленина, д. 40). E-mail: artur.a.mitsel@tusur.ru. Scopus Author ID 6603150769, ResearcherID G-8307-2014, SPIN-код РИНЦ 9698-2160, https://orcid.org/0000-0002-2624-4383

Викторенко Елена Владимировна, старший преподаватель, аспирант кафедры экономики, ФГАОУ ВО «Томский государственный университет систем управления и радиоэлектроники» (634050, Россия, Томск, пр-т Ленина, д. 40). E-mail: viktorenko.e@gmail.com. ResearcherID AEJ-4949-2022, SPIN-код РИНЦ 8664-3235, https://orcid.org/0000-0003-3871-8993

Translated from Russian into English by L. Bychkova
Edited for English language and spelling by Thomas A. Beavitt

Mathematical modeling

Математическое моделирование

UDC 577.337; 538.931
https://doi.org/10.32362/2500-316X-2025-13-2-111-120
EDN ATOWXW

RESEARCH ARTICLE

Lateral proton transport induced by acoustic solitons propagating in lipid membranes

Vasiliy N. Kadantsev [@], Alexey N. Goltsov

MIREA – Russian Technological University, Moscow, 119454 Russia
[@] Corresponding author, e-mail: appl.synergy@yandex.ru

Abstract

Objectives. The study of proton transport in membrane structures represents a significant technological task in the development of hydrogen energy as well as a fundamental problem in bioenergetics. Investigation in this field aims at finding out the physical mechanisms of fast proton transport in the meso-porous structures in polymer electrolyte membranes, which serve as electrochemical components of hydrogen fuel cells. The objectives of the research in the field of bioenergetics are to elucidate the molecular mechanisms of effective proton transport in transmembrane channel proteins, as well as along the surface proton-conducting structures in biological membranes. To investigate the molecular mechanisms of the direct proton transport along the water-membrane interface, we developed a model of proton movement along quasi-one-dimensional lateral domain structures in multicomponent lipid membranes.

Methods. The developed approach is based on a model of collective excitations spreading along the membranes in the form of acoustic solitons, which represent the regions of local compression of polar groups and structural defects in hydrocarbon chains of lipid molecules.

Results. The results of modeling showed that the interaction between an excess proton on the membrane surface and a soliton of membrane compression leads to the proton being trapped by an acoustic soliton, followed by its transport by moving soliton. The developed model was applied to describe effective proton transport along the inner mitochondrial membrane and its role in the local coupling function of molecular complexes in cell bioenergetics.

Conclusions. The developed soliton model of proton transport demonstrated that collective excitations within lipid membranes can determine one of the factors affecting the efficiency of proton transport along interphase boundaries. Further development of the theoretical approaches, taking into account dynamic properties of polymer and biological proton-conducting membranes, can contribute to the study of a role of surface proton transport in cell bioenergetics, as well as to the investigation of transport characteristics of the proton-exchange polymer membranes developed for the hydrogen energy industry.

Keywords: proton transport, proton-conducting structures, lipid membranes, domain structures, collective dynamics, solitons

• **Submitted:** 15.07.2023 • **Revised:** 24.09.2024 • **Accepted:** 23.01.2025

For citation: Kadantsev V.N., Goltsov A.N. Lateral proton transport induced by acoustic solitons propagating in lipid membranes. *Russian Technological Journal.* 2025;13(2):111–120. https://doi.org/10.32362/2500-316X-2025-13-2-111-120, https://elibrary.ru/ATOWXW

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Латеральный протонный транспорт, индуцированный распространением акустических солитонов в липидных мембранах

В.Н. Каданцев [@], **А.Н. Гольцов**

МИРЭА – Российский технологический университет, Москва, 119454 Россияя
[@] *Автор для переписки, e-mail: appl.synergy@yandex.ru*

Резюме

Цели. Исследование протонного транспорта в мембранных структурах является важной технологической задачей в области водородной энергетики, а также представляет собой фундаментальную проблему биоэнергетики. Целью этих исследований является выяснение физических механизмов быстрого протонного транспорта в мезо-пористых структурах полимерных электролитных мембран, являющихся электрохимическими компонентами водородных топливных элементов. В области биоэнергетики эти исследования направлены на выяснения молекулярных механизмов эффективного протонного транспорта в трансмембранных белках-каналах и в поверхностных протонпроводящих структурах биологических мембран в системах биоэнергетики клетки. С целью исследования молекулярных механизмов направленного транспорта протонов в работе рассматривается модель движения протонов в квазиодномерных латеральных доменных структурах в многокомпонентных липидных мембранах.

Методы. В основе развиваемого подхода лежит модель коллективных возбуждений типа акустических солитонов, которые представляют собой перемещающиеся вдоль мембраны области локального сжатия полярных групп и структурных дефектов в подсистеме углеводородных цепей липидных молекул.

Результаты. Показано, что учет в модели взаимодействия избыточного протона на поверхности мембраны с солитоном сжатия мембраны приводит к захвату протона акустическим солитоном с его последующим транспортом. Разработанная модель применяется к описанию механизма эффективного протонного транспорта вдоль внутренней митохондриальной мембраны и его роли в сопряжении функционирования молекулярных комплексов в системе биоэнергетики клетки.

Выводы. Развитая солитонная модель протонного транспорта показала, что коллективные возбуждения в липидных мембранах могут определять факторы, влияющие на эффективность протонного транспорта вдоль межфазных границ. Дальнейшее развитие теоретических подходов, учитывающих динамические свойства полимерных и биологических протонпроводящих мембран, может внести вклад в исследование роли поверхностного транспорта протонов в биоэнергетику клетки, а также в исследование транспортных характеристик разрабатываемых протонно-обменных полимерных мембран водородной энергетики.

Ключевые слова: протонный транспорт, протонпроводящие структуры, липидные мембраны, доменные структуры, коллективная динамика, солитоны

• **Поступила:** 15.07.2023 • **Доработана:** 24.09.2024 • **Принята к опубликованию:** 23.01.2025

Для цитирования: Каданцев В.Н., Гольцов А.Н. Латеральный протонный транспорт, индуцированный распространением акустических солитонов в липидных мембранах. *Russian Technological Journal*. 2025;13(2):111–120. https://doi.org/10.32362/2500-316X-2025-13-2-111-120, https://elibrary.ru/ATOWXW

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

Experimental and theoretical studies of proton conductivity of materials and systems are currently carried out in two distinct scientific and technological fields: hydrogen energy and bioenergetics of living systems. The development of proton-conducting materials is of particular interest today for the production of components of electrochemical devices, especially fuel cell membranes for the manufacture of hydrogen power plants, batteries, electric vehicle power units, etc. [1]. With proton batteries already emerging as a competitive alternative to lithium-ion technologies, research in this field is aimed at creating efficient polymeric (e.g., perfluorinated sulfopolymers) and solid-state proton electrolytes. The replacement of lithium ions with protons as charge carriers in hydrogen fuel cells significantly increases the conductivity of the electrolyte due to the high mobility of protons in the electrolyte membrane.

In most polymeric proton-exchange membranes of hydrogen fuel cells, fast proton transport takes place due to hydrate layers of water in the membrane structure—or, more precisely, along the nanoscale structures formed by water molecules in mesoporous structures of polymer materials. The specific proton conductivity of proton exchange membranes can reach values in the range of 10^{-3} to 10^{-1} S/cm [2], while the proton conductivity of bulk water is in the range of 10^{-6} S/cm, i.e., five orders of magnitude lower than the conductivity of polymer membranes. However, the molecular mechanisms of the high proton conductivity of near-surface water in porous polymeric materials, whose properties are anomalously different from those of bulk water, are still not fully understood.

While the study of physicochemical mechanisms of fast proton transport in membrane structures represents an important technological task of hydrogen energetics, it is also a fundamental problem in the field of bioenergetics. Here, studies are aimed at elucidating the molecular mechanisms of proton transport in the system of oxidative phosphorylation in cell mitochondria, in transmembrane channels, and in surface proton-conducting structures of biomembranes.

Artificial polymeric proton-exchange membranes have much in common with biological lipid membranes in terms of their structure and molecular composition: both polymeric membranes and biological membranes consist of amphiphilic molecules with hydrophobic chains and acidic groups. In both systems, the proton-conducting hydrate layer of water molecules forms a hydrogen bonding structure at the interface, along which fast two-dimensional diffusion of protons is assumed to occur. The physical mechanisms of efficient proton transport in hydrophobic channels of protein molecules and the surface layer of mitochondrial membranes are currently the subject of intensive research in cell bioenergetics. However, although Mitchell's chemiosmotic theory gives a general idea of the functioning of the mitochondrial bioenergetic system, it needs further development based on new experimental data on the functioning of individual components of the oxidative phosphorylation system and on the spatial organization of the entire system of adenosine triphosphate (ATP) molecule synthesis in the mitochondrial membrane [3–5]. In particular, new experimental data on the structure of inner mitochondrial membranes have not only demonstrated their structural and organizational function, but also revealed the important coupling and integrating role that they fulfill in the functioning of the whole system of electron transport processes and ATP synthesis [6, 7]. In particular, recent experimental data on fast proton transport across mitochondrial and artificial membranes [8, 9] supported the hypothesis of a local coupling of respiration and phosphorylation due to the near-membrane transport of partially dehydrated protons [10–12].

Several possible molecular mechanisms for the fast lateral transport of protons along the membrane interface through bound water molecules over long distances are currently under consideration. The first of these is the Grotthuss mechanism of proton transport along the hydrogen bond chain (structural diffusion) [13]. A second potential mechanism is based on the diffusion of protons within the hydroxonium ion H_3O^+ (vesicular diffusion) [14]. A third approach describes the process of co-diffusion of a lipid molecule with a strongly bound proton. Effective proton transport along the membrane-bound water interface can also be envisaged in terms of a combination of structural and vesicular diffusion [15]. To date, no conclusive experimental evidence can be adduced in favor of one or another mechanism of proton transport. However, all proposed mechanisms are based on experimental data confirming proton retention at the membrane-water interface that ensure the efficient two-dimensional diffusion of protons with limited release into the bulk phase [7, 16, 17]. The retention of protons at the membrane surface has been studied in experiments on bilayer membranes [18, 19] and liposomes [20]. Moreover, part of the energy stored in the form of partially dehydrated proton has been shown to be incorporated into ATP synthesis [21]. The causes of proton affinity to interfaces and surface proton transport have also been studied theoretically [22–24]. The results of these studies showed that the mechanism of proton retention at the membrane surface is determined by electrostatic interaction and the entropic barrier. The polar groups (PG) of lipid molecules have also been found to significantly affect the rate of proton surface transport [25]. Here, the PG composition is assumed to influence the formation of one-dimensional

proton-conducting structures of hydrogen bonds of water molecules bound to the membrane [9].

In work [5], proton-conducting lateral structures, representing quasi-one-dimensional domain structures (DS) in the cristae of inner mitochondrial membranes enriched with cardiolipin molecules were considered. Based on the Grotthuss mechanism, the authors have developed a model of proton transport along hydrogen-bonded chains of water molecules interacting with cardiolipin PGs in proton-conducting membrane structures. The interaction of the proton subsystem and the lipid PG subsystem has been shown to lead to the formation of a two-component soliton, whose motion corresponds to the coordinated movement of the proton and soliton of compression along the lipid membrane [5].

A similar theoretical approach to modeling the soliton transport of protons has also been developed for polymer membranes [26]. In the proposed model, proton transport as part of the hydroxonium ion H_3O^+ occurs due to collective excitations of the soliton-like type spreading in ordered chains of hydrogen bonds formed by water molecules on the membrane with sulfide surface groups. The model establishes the relationship between the soliton mobility and the parameters of the spatial structure of the surface sulfide groups.

The present work considers a model of alternative proton transport realized by proton trapping by an acoustic soliton moving along proton-conducting structures in lipid membranes. This mechanism is closer to the vesicular mechanism, where an acoustic soliton rather than a hydroxonium ion acts as a proton carrier (vesicle). The work is based on the model of acoustic soliton formation and propagation in quasilinear DSs proposed in our previous study [27]. Nonlinear excitations of the soliton-like type represent regions of local compression of lipid PGs and structural defects in the subsystem of hydrocarbon chains (HC) moving along the membrane molecular structures.

The experimental observation of soliton-like excitations in lipid monolayers and bilayers has been carried out in a number of experiments using different excitation and registration methods. The excitation of acoustic soliton-like pulses and their dissipationless motion were observed in experiments with optical generation of solitary waves in lipid monolayers [28]. Excitation of elastic soliton-like pulses was also found in liposomes in the temperature range of the lipid phase transition. Acoustic waves accompanying the propagation of a nerve impulse in an axon were shown to have one of the characteristic properties of solitons, i.e. two colliding nerve impulses pass through each other without changing their shape [29].

The present work considers a model of proton trapping by an acoustic soliton and its subsequent transport. The possibility of such trapping and the resulting mechanism of charge transport by acoustic solitons through one-dimensional nonlinear molecular structures has been discussed in [30, 31]. It is assumed that lateral proton transport in lipid membranes can occur in a similar way as a result of proton trapping by a soliton, and that quasi-one-dimensional DSs in multi-component membranes may represent proton-conducting channels on the membrane surface.

1. MODEL OF THE LATERAL TRANSPORT OF A PROTON TRAPPED BY A SOLITON IN A QUASI-ONE-DIMENSIONAL DS OF A LIPID MEMBRANE

We consider the model of a one-dimensional chain of lipid molecules forming a quasi-one-dimensional DS in mixed lipid bilayers. In the model, two subsystems are distinguished: the membrane surface formed by PGs of lipid molecules and the inner hydrophobic region of the membrane formed by lipid HCs [32]. In [27], we developed a model of soliton-type collective excitations representing the regions of local displacements of lipid PGs $\rho(x, t)$ from equilibrium positions and structural defects of the kink type in the lipid HC subsystem, which is described by the HC deviation from the normal to the membrane surface $u(x, t)$:

$$\rho(x,t) = -\rho_0 \operatorname{sech}^2 \frac{x - Vt}{\Delta}, \quad (1)$$

$$u(x,t) = u_0 \tanh \frac{x - Vt}{\Delta}, \quad (2)$$

where V is the soliton velocity and the lipid PG equilibrium position ρ_0 is determined by the following equation:

$$\rho_0 = \frac{\chi u_0^2}{M\Omega_0^2}.$$

Here, M is the mass of the PG of lipid molecule; χ is the constant of interaction between lipid PGs and HCs, which accounts for the change in HC conformation upon PG displacement from the equilibrium position; Ω_0 is the characteristic frequency of a chain of the lipid PG; Δ is the kink width. $u_0 = \pm(|G|/B)^{1/2}$ is the HC equilibrium state in the Ginzburg–Landau double-well potential:

$$U_T(u) = -\frac{1}{2}|G(T)|u^2 + \frac{1}{4}Bu^4,$$

where $G(T) = E_0(T/T_c - 1)$ and B are potential parameters; E_0 is the potential barrier height; T_c is the temperature of the main phase transition of the membrane, at which lipid melting occurs [27, 32].

The soliton solution for the PG displacement $\rho(x, t) < 0$ (1) describes the compression strain in the lipid PG subsystem associated with the defect in the lipid HC system. For protons in the near-membrane layer, which are trapped in such a structure, the presence of a soliton appears as an additional interaction energy. The solution $u(x, t)$ (2) in the form of a kink in the region of coordinate $x = Vt$ describes the deviations of lipid molecule HCs in opposite directions, which is characteristic of structural defects like dislocations in liquid crystals.

In the proposed model, the motion of a proton trapped by soliton of compression in a lipid PG chain is described by the wave function $\psi(x, t)$ which satisfies the time-dependent Schrödinger equation:

$$i\hbar\frac{\partial\psi(x,t)}{\partial t}+\frac{\hbar^2}{2m}\cdot\frac{\partial^2\psi(x,t)}{\partial x^2}-U(x,t)\psi(x,t)=0, \quad (3)$$

where $U(x, t) = \sigma\rho(x, t)$ is the potential well generated by negatively charged PGs of lipid molecules in the soliton region; σ is the parameter of the electrostatic interaction between the proton and the PGs of lipid molecules in the soliton region; m is the mass of the proton; $\hbar$ is the Planck constant.

The solution of the time-dependent Schrödinger equation Eq. (3) was sought in the following form:

$$\psi(x,t)=\varphi(\xi)e^{-\frac{i}{\hbar}Et}, \quad (4)$$

where the spatial coordinate $\xi = x - Vt$ associated with the soliton motion at velocity V is introduced.

Substituting (4) into (3), we obtained the stationary Schrödinger equation for the real part of the amplitude $\varphi(\xi)$ of the proton in the potential well $U(\xi)$:

$$\frac{\hbar^2}{2m}\cdot\frac{\partial^2\varphi}{\partial\xi^2}+[E-U(\xi)]\varphi=0, \quad (5)$$

where $U(\xi)=-\sigma\rho_0\,\mathrm{sech}^2\left(\frac{\xi}{\Delta}\right)$.

Equation (5) can be transformed into the equation for generalized Lagrangian functions, as follows:

$$\frac{d}{dz}\left[1-z^2\frac{d\varphi}{dz}\right]+\left[s(s+1)-\frac{\varepsilon^2}{1-\varepsilon^2}\right]\varphi=0, \quad (6)$$

where the variable $z = \tanh(\Delta^{-1}\xi)$ and the following notations were introduced:

$$\varepsilon=\frac{\Delta\sqrt{-2mE}}{\hbar};\ s=\frac{1}{2}\left(-1+\sqrt{1+8m\sigma\rho_0\Delta^2\hbar^{-2}}\right). \quad (7)$$

In this case, Eq. (6) has a solution in the following form:

$$\varphi_n(\xi)=A_n\,\mathrm{sech}^{\varepsilon}\left(\frac{\xi}{\Delta}\right)\times$$
$$\times F\left(\varepsilon-s,\ \varepsilon+s+1,\ \varepsilon+1,\ \frac{1}{2}\left(1-\tanh(\Delta^{-1}\xi)\right)\right), \quad (8)$$

where A_n is the normalization factor of the wave function; F is a hypergeometric function representing a polynomial of degree n under the condition $\varepsilon - s = -n$ ($n = 0, 1, 2, \ldots$) [33].

This condition gives the following expression for the proton energy levels in the potential well $U(\xi)$:

$$E_n=-\frac{\hbar}{8m\Delta^2}\left(-(1+2n)+\sqrt{1+8m\sigma\rho_0\Delta^2\hbar^{-2}}\right)^2.$$

Thus, the potential well $U(\xi)$ contains a finite number of stationary energy levels for a proton trapped by a soliton in the PG chain of lipid molecules. For the ground level at $n = 0$, the wave function (8) has the following form:

$$\varphi_0(\xi)=A_0\,\mathrm{sech}^{\varepsilon}\left(\frac{\xi}{\Delta}\right). \quad (9)$$

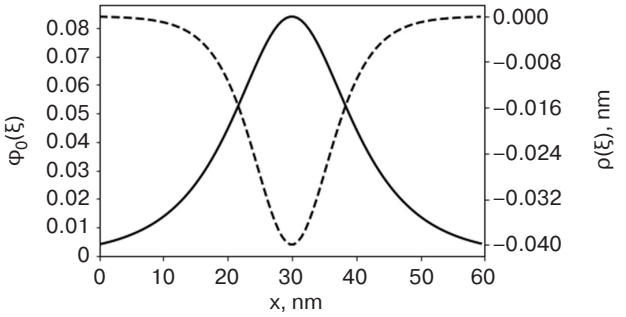


Figure. Wave function of the ground state of proton $\varphi_0(\xi)$ (solid line) in the potential well $U(\xi)$ generated by soliton of compression $\rho(\xi)$ (dashed line) of lipid molecule PGs in a quasi-one-dimensional lipid DS

The figure shows the proton wave function $\varphi_0(\xi)$ (9) calculated for $\varepsilon = 1$ with the normalization factor $A_0 = 0.42$. The calculation was performed for the soliton parameters obtained based on the following experimental data. The characteristic frequency $\Omega_0 = 10^{11}$ Hz was estimated according to the experimental data for the oscillation frequency of the lipid membrane [32]. The velocity of the trapped proton was determined by soliton propagation velocity V (Eqs. (1) and (2)), which was estimated by us to be in the range of 50–100 m/s [5]. The kink width $\Delta \approx 8$ nm was estimated on the basis

of experimental data on the size of the defect region in the HC subsystem formed in membrane structures at the temperature near the main phase transition temperature [34]. In this case, the soliton region includes about 10 lipid molecules located along the one-dimensional DS. The kink solution for the HC displacement $u(x, t)$ describes a dislocation-type defect. Negative and positive values of $u(x, t)$ correspond to the deviations of lipid molecule HCs in opposite directions. The soliton solution for the PG displacement $\rho(x, t)$ describes the compressive deformation in the PG subsystem caused by a defect in the HC subsystem.

2. DISCUSSION

In the present work, we developed a model of proton motion in quasi-one-dimensional lateral DSs, which are assumed to play the role of proton conducting structures in multi-component cell membranes [27, 35]. The proposed approach is based on the model of collective excitations in lipid membranes such as acoustic solitons which represent regions of local PG compression and structural defects of lipid molecule HCs coordinately moving along the membrane.

The model takes into account the electrostatic interaction of a membrane-bound proton with soliton of compression. This interaction leads to the proton trapping by a moving acoustic soliton and its subsequent transport. In contrast to the model of proton soliton transport previously developed by us on the basis of the structural proton diffusion mechanism (Grotthuss mechanism) [5], in this work, we considered an alternative model of proton transport. The proposed model overcomes two problems of the structural proton diffusion model. Firstly, in contrast to the Grotthuss mechanism, a proton trapped by the soliton does not undergo a large number of jumps along the hydrogen bond chain over a short distance of 2.5 Å, which significantly increases the velocity of its movement. Secondly, the described approach provides proton transport that is strongly coupled to the region of local compression of PGs moving along the membrane surface in the form of a soliton. The experimentally observed effect of proton retention at the membrane surface implies the strong local interaction of the proton with lipid PGs that is inconsistent with the data on proton delocalization and its movement along the membrane surface.

The developed model can be applied to describe proton transport along the surface of the inner mitochondrial membrane in the oxidative phosphorylation coupling system, which has been experimentally established to possess a unique spatial organization. The cryo-electron tomography method has revealed an ordered cluster oligomeric structure formed by parallel rows of respiratory complexes and ATP synthase dimers located on the folds of the cristae of mitochondrial inner membranes [7]. The formation, morphology, and dynamics of mitochondrial cristae are determined by structural rearrangements of lipid membranes, which are highly sensitive to the physiological state of mitochondria [36]. The small distance (~50 Å) between the rows of proton pumps and ATP synthase molecules provides conditions for direct and fast proton transport to ATP synthases along the cristae membrane. Currently, considerable interest is shown in investigating the molecular mechanisms of proton transport in mitochondrial membranes and determining the factors that influence its efficiency [16, 17, 37]. Based on the developed approach, we proposed that this research should not only consider the influence of the membrane surface structure on proton transport, but should also take into account the dynamic properties of biomembranes, in particular the formation of collective excitations in lipid bilayers. Through the consideration of elastic excitations of the membrane in the proposed model, proton transport is accompanied by the transfer of membrane deformation energy stored by the acoustic soliton. This approach links lateral proton motion to the non-equilibrium dynamics of mitochondrial cristae by coupling transport and dynamic processes at the biomembrane surface [38]. The energy of the local membrane elastic oscillations transferred together with the charge may additionally be hypothesized to contribute to functioning and synchronizing membrane proteins, receptors, and ion channels [39, 40] in particular, those involved in synchronizing the functioning of oligomeric protein complexes making up the mitochondrial oxidative phosphorylation system. The experimental detection of proton transport, accompanied by the propagation of elastic excitations along membrane surfaces, may confirm the contribution of collective excitations to the effective proton transport in the inner mitochondrial membranes, as well as to the coupling mechanism in the oxidative phosphorylation system.

CONCLUSIONS

The results of theoretical and experimental studies in the fields of bioenergetics of mitochondrial membranes and polymeric proton-exchange membrane technology of hydrogen fuel cells allowed the elucidation of many common features of surface proton transport in biological and artificial membranes. Two primary mechanisms of efficient proton transport—structural diffusion (Grotthuss mechanism) and vesicular transport—are considered in the study of both systems to confirm the role of a membrane bounded layer of structured water, in which proton transport is localized. In polymer membranes, this was demonstrated by the detection of fast proton transport at low membrane

hydration. In mitochondrial membranes, the same effect was confirmed in the experiments that showed the localization of protons in the surface layer of the inner mitochondrial membrane in the oxidative phosphorylation system. In both artificial and biological membranes, a significant influence of membrane composition and structure (surface acidic groups) on proton conductance was discovered. The soliton model of proton transport developed in this paper showed that the collective excitations of lipid membranes together with their structural properties can determine the factors that influence the proton transport efficiency. The further development of theoretical approaches that take into account both structural and dynamic properties of polymeric and biological proton-conducting membranes can contribute to the study of the role of surface proton transport in cell bioenergetics, as well as to the investigation of transport characteristics of polymeric proton-exchange membranes developed for hydrogen energetics.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (grant No. FGFZ-2023-0004).

Authors' contribution. All authors equally contributed to the research work.

REFERENCES

1. Dobrovolsky Y.A., Chikin A.I., Sanginov E.A., Chub A.V. Proton-exchange membranes based on heteropoly compounds for low temperature fuel cells. *Al'ternativnaya energetika i ekologiya = Alternative Energy and Ecology.* 2015;4(165):22–45 (in Russ.). https://doi.org/10.15518/isjaee.2015.04.02
2. Lebedeva O.V. Proton conducting membranes for hydrogen-air fuel elements. *Izvestiya vuzov. Prikladnaya khimiya i biotekhnologiya = Proceedings of Universities. Applied Chemistry and Biotechnology*. 2016;1(16):7–19 (in Russ.).
3. Eremeev S.A., Yaguzhinsky L.S. On local coupling of the electron transport and ATP-synthesis system in mitochondria. Theory and experiment. *Biochemistry* (*Moscow*). 2015;80(5):576–581. https://doi.org/10.1134/S0006297915050089 [Original Russian Text: Eremeev S.A., Yaguzhinsky L.S. On local coupling of the electron transport and ATP synthesis system in mitochondria. Theory and experiment. *Biokhimiya.* 2015;80(5):682–688 (in Russ.).]
4. Kell D.B. A protet-based model that can account for energy coupling in oxidative and photosynthetic phosphorylation. *Biochim. Biophys. Acta Bioenerg*. 2024;1865(4):149504. https://doi.org/10.1016/j.bbabio.2024.149504
5. Nesterov S.V., Yaguzhinsky L.S., Vasilov R.G., Kadantsev V.N., Goltsov A.N. Contribution of the Collective Excitations to the Coupled Proton and Energy Transport along Mitochondrial Cristae Membrane in Oxidative Phosphorylation System. *Entropy* (*Basel*). 2022;24(12):1813. https://doi.org/10.3390/e24121813
6. Davies K.M., Strauss M., Daum B., Kief J.H., Osiewacz H.D., Rycovska A., et al. Macromolecular organization of ATP synthase and complex I in whole mitochondria. *Proc. Natl. Acad. Sci. USA.* 2011;108(34):14121–14126. https://doi.org/10.1073/pnas.1103621108
7. Nesterov S., Chesnokov Y., Kamyshinsky R., Panteleeva A., Lyamzaev K., Vasilov R., et al. Ordered Clusters of the Complete Oxidative Phosphorylation System in Cardiac Mitochondria. *Int. J. Mol. Sci*. 2021;22(3):1462. https://doi.org/10.3390/ijms22031462
8. Mulkidjanian A.Y., Heberle J., Cherepanov D.A. Protons @ interfaces: Implications for biological energy conversion. *Biochimica et Biophysica Acta* (*BBA*) – *Bioenergetics*. 2006;1757(8):913–930. https://doi.org/10.1016/j.bbabio.2006.02.015
9. Weichselbaum E., Österbauer M., Knyazev D.G., Batishchev O.V., Akimov S.A., Nguyen T.H., et al. Origin of proton affinity to membrane/water interfaces. *Sci. Rep*. 2017;7(1):4553. https://doi.org/10.1038/s41598-017-04675-9
10. Yaguzhinsky L.S., Boguslavsky L.I., Volkov A.G., Rakhmaninova A.B. Synthesis of ATP coupled with action of membrane protonic pumps at the octane-water interface. *Nature*. 1976;259(5543):494–496. https://doi.org/10.1038/259494a0

11. Kell D.B. On the functional proton current pathway of electron transport phosphorylation. An electrodic view. *Biochim. Biophys. Acta*. 1979;549(1):55–99. https://doi.org/10.1016/0304-4173(79)90018-1
12. Morelli A.M., Ravera S., Calzia D., Panfoli I. An update of the chemiosmotic theory as suggested by possible proton currents inside the coupling membrane. *Open Biol*. 2019;9(4):180221. https://doi.org/10.1098/rsob.180221
13. Wraight C.A. Chance and design—Proton transfer in water, channels and bioenergetic proteins. *Biochimica et Biophysica Acta* (*BBA*) – *Bioenergetics*. 2006;1757(8):886–912. https://doi.org/10.1016/j.bbabio.2006.06.017
14. Kreuer K.D. Proton Conductivity: Materials and Applications. *Chem. Mater*. 1996;8(3):610–641. https://doi.org/10.1021/cm950192a
15. Ludueña G.A., Kühne T.D., Sebastiani D. Mixed Grotthuss and Vehicle Transport Mechanism in Proton Conducting Polymers from *Ab initio* Molecular Dynamics Simulations. *Chem. Mater*. 2011;23(6):1424–1429. https://doi.org/10.1021/cm102674u
16. Weichselbaum E., Galimzyanov T., Batishchev O.V., Akimov S.A., Pohl P. Proton Migration on Top of Charged Membranes. *Biomolecules*. 2023;13(2):352. https://doi.org/10.3390/biom13020352
17. Knyazev D.G., Silverstein T.P., Brescia S., Maznichenko A., Pohl P. A New Theory about Interfacial Proton Diffusion Revisited: The Commonly Accepted Laws of Electrostatics and Diffusion Prevail. *Biomolecules*. 2023;13(11):1641. https://doi.org/10.3390/biom13111641
18. Antonenko Y.N., Kovbasnjuk O.N., Yaguzhinsky L.S. Evidence in favor of the existence of a kinetic barrier for proton transfer from a surface of bilayer phospholipid membrane to bulk water. *Biochimica et Biophysica Acta* (*BBA*) – *Biomembranes*. 1993;1150(1):45–50. https://doi.org/10.1016/0005-2736(93)90119-k
19. Tashkin V.Yu., Vishnyakova V.E., Shcherbakov A.A., Finogenova O.A., Ermakov Yu.A., Sokolov V.S. Changes of the Capacitance and Boundary Potential of a Bilayer Lipid Membrane Associated with a Fast Release of Protons on Its Surface. *Biochem. Moscow Suppl. Ser. A*. 2019;13(2):155–160. https://doi.org/10.1134/S1990747819020077
20. Sjöholm J., Bergstrand J., Nilsson T., Šachl R, Ballmoos C., Widengren J., et al. The lateral distance between a proton pump and ATP synthase determines the ATP-synthesis rate. *Sci. Rep*. 2017;7(1):1–12. http://doi.org/10.1038/s41598-017-02836-4
21. Yaguzhinsky L.S., Boguslavsky L.I., Volkov A.G., Rakhmaninova A.B. Synthesis of ATP coupled with action of membrane protonic pumps at the octane–water interface. *Nature*. 1976;259(5543):494–496. https://doi.org/10.1038/259494a0
22. Lee J.W. Mitochondrial energetics with transmembrane electrostatically localized protons: do we have a thermotrophic feature? *Sci Rep*. 2021;11(1):14575. https://doi.org/10.1038/s41598-021-93853-x
23. Medvedev E., Stuchebrukhov A. Mechanism of long-range proton translocation along biological membranes. *FEBS Lett*. 2012;587(4):345–349. https://doi.org/10.1016/j.febslet.2012.12.010
24. Cherepanov D.A., Junge W., Mulkidjanian A.Y. Proton transfer dynamics at the membrane/water interface: dependence on the fixed and mobile pH buffers, on the size and form of membrane particles, and on the interfacial potential barrier. *Biophys J*. 2004;86(2):665–80. https://doi.org/10.1016/s0006-3495(04)74146-6
25. Amdursky N., Lin Y., Aho N., Groenhof G. Exploring fast proton transfer events associated with lateral proton diffusion on the surface of membranes. *Proc. Natl. Acad. Sci. USA.* 2019;116(7):2443–2451. https://doi.org/10.1073/pnas.1812351116
26. Golovnev A., Eikerling M. Theory of collective proton motion at interfaces with densely packed protogenic surface groups. *J. Phys.: Condens. Matter*. 2012;25(4):045010. https://doi.org/10.1088/0953-8984/25/4/045010
27. Kadantsev V.N., Goltsov A.N. Collective dynamics of domain structures in liquid crystalline lipid bilayers. *Russian Technological Journal* . 2022;10(4):44–54 https://doi.org/10.32362/2500-316X-2022-10-4-44-54
28. Shrivastava S., Schneider M.F. Evidence for two-dimensional solitary sound waves in a lipid controlled interface and its implications for biological signalling. *J. Royal Soc. Interface.* 2014;11(97):20140098. https://doi.org/10.1098/rsif.2014.0098
29. Gonzalez-Perez A., Budvytyte R., Mosgaard L.D., Nissen S., Heimburg T. Penetration of Action Potentials During Collision in the Median and Lateral Giant Axons of Invertebrates. *Phys. Rev. X.* 2014;4(3):031047. http://doi.org/10.1103/PhysRevX.4.031047
30. Lupichev L.N., Savin A.V., Kadantsev V.N. *Synergetics of Molecular Systems*. Series: Springer Series in Synergetics. Cham: Springer; 2015. 332 p. https://doi.org/10.1007/978-3-319-08195-3
31. Kadantsev V.N., Goltsov A.N., Kondakov M.A. Electrosoliton dynamics in a thermalized molecular chain. *Rossiiskii tekhnologicheskii zhurnal.* 2020;8(1):43–57 (in Russ.). https://doi.org/10.32362/2500-316X-2020-8-1-43-57
32. Bolterauer H., Tuszyński J.A., Satarić M.V. Fröhlich and Davydov regimes in the dynamics of dipolar oscillations of biological membranes. *Phys. Rev. A*. 1991;44(2):1366–1381. https://doi.org/10.1103/physreva.44.1366
33. Landau L.D., Lifshits E.M. *Teoreticheskaya fizika* (*Theoretical physics*): in 10 v. V. 3. *Kvantovaya Mekhanika* (*nerelyativistskaya teoriya*) (*Quantum Mechanics* (*Non-Relativistic Theory*)). Moscow: Fizmatlit; 2024. 800 p. (in Russ.). ISBN 5-9221-0057-2, 978-5-9221-0530-9

34. Wack D.C., Webb W.W. Synchrotron X-ray study of the modulated lamellar phase in the lecithin-water system. *Phys. Rev. A*. 1989;40(5):2712–2730. https://doi.org/10.1103/PhysRevA.40.2712
35. Goltsov A.N. Formation of quasilinear structure in lipid membranes. *Biofizika*. 1997;42(1):174–181.
36. Joubert F., Puff N. Mitochondrial Cristae Architecture and Functions: Lessons from Minimal Model Systems. *Membranes* (*Basel*). 2021;11(7):465. https://doi.org/10.3390/membranes11070465
37. Toth A., Meyrat A., Stoldt S., Santiago R., Wenzel D., Jakobs S., et al. Kinetic coupling of the respiratory chain with ATP synthase, but not proton gradients, drives ATP production in cristae membranes. *Proc. Natl. Acad. Sci. USA*. 2020;117(5): 2412–2421. https://doi.org/10.1073/pnas.1917968117
38. Patil N., Bonneau S., Joubert F., Bitbol A.F., Berthoumieux H. Mitochondrial cristae modeled as an out-of-equilibrium membrane driven by a proton field. *Phys. Rev. E*. 2020;102(2):022401. https://doi.org/10.1103/physreve.102.022401
39. Johnson A.S., Winlow W. The Soliton and the Action Potential – Primary Elements Underlying Sentience. *Front. Physiol*. 2018;9:779. https://doi.org/10.3389/fphys.2018.00779
40. Li S., Yan Z., Huang F., Zhang X., Yue T. How a lipid bilayer membrane responds to an oscillating nanoparticle: Promoted membrane undulation and directional wave propagation. *Colloids Surf. B. Biointerfaces*. 2020;187:110651. https://doi.org/10.1016/j.colsurfb.2019.110651

About the authors

Vasiliy N. Kadantsev, Dr. Sci. (Phys.-Math.), Professor, Department of Biocybernetic Systems and Technologies, Institute of Artificial Intelligence, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: appl.synergy@yandex.ru. Scopus Author ID 6602993607, https://orcid.org/0000-0001-9205-6527

Alexey N. Goltsov, Dr. Sci. (Phys.-Math.), Professor, Department of Biocybernetic Systems and Technologies, Institute of Artificial Intelligence, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: golcov@mirea.ru. Scopus Author ID 56234051200, RSCI SPIN-code 1288-9918, https://orcid.org/0000-0001-6725-189X

Об авторах

Каданцев Василий Николаевич, д.ф.-м.н., профессор, кафедра биокибернетических систем и технологий, Институт искусственного интеллекта, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: appl.synergy@yandex.ru. Scopus Author ID 6602993607, https://orcid.org/0000-0001-9205-6527

Гольцов Алексей Николаевич, д.ф.-м.н., профессор, кафедра биокибернетических систем и технологий, Институт искусственного интеллекта, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: golcov@mirea.ru. Scopus Author ID 56234051200, SPIN-код РИНЦ 8852-2616, https://orcid.org/0000-0001-6725-189X

Translated from Russian into English by K. Nazarov
Edited for English language and spelling by Thomas A. Beavitt

Mathematical modeling

Математическое моделирование

UDC 004.023, 519.677
https://doi.org/10.32362/2500-316X-2025-13-2-121-131
EDN EWCRYQ

RESEARCH ARTICLE

Method for estimating objective function landscape convexity during extremum search

Alexander V. Smirnov @

MIREA – Russian Technological University, Moscow, 119454 Russia
@ *Corresponding author, e-mail: av_smirnov@mirea.ru*

Abstract

Objectives. The work set out to develop a method for estimating the objective function (OF) landscape convexity in the extremum neighborhood. The proposed method, which requires no additional OF calculations or complicated mathematical processing, relies on the data accumulated during extremum search.

Methods. Landscape convexity is characterized by the index of power approximation of the OF in the vicinity of the extremum. The estimation of this index is carried out for pairs of test points taking into account their distances to the found extremum and OF values in them. Based on the analysis of estimation errors, the method includes the selection of test points by their distances from the found extremum and the selection of pairs of test points by the angle between the directions to them from the found extremum. Test functions having different convexities, including concave, were used to experimentally validate the method. The particle swarm optimization algorithm was used as an extremum search method. The experimental results were presented in the form of statistical characteristics and histograms of distributions of the estimation values of the degree of the OF approximation index.

Results. The conductive experiments confirm that the proposed method provides a reliable estimation of power index range bounds upon condition of appropriate definition of trial points and trial point pair selection parameters.

Conclusions. The proposed method may be a part of OF landscape analysis. It is necessary to complement it with the algorithms for automatic adjustment of trial points and pairs of trial points selection parameters. Additional information may be provided by analyzing the dependencies of power index estimations and trial point distances from extrema.

Keywords: objective function landscape, convex function, concave function, power approximation, power index, histogram

• **Submitted:** 28.05.2024 • **Revised:** 26.07.2024 • **Accepted:** 12.02.2025

For citation: Smirnov A.V. Method for estimating objective function landscape convexity during extremum search. *Russian Technological Journal.* 2025;13(2):121–131. https://doi.org/10.32362/2500-316X-2025-13-2-121-131, https://elibrary.ru/EWCRYQ

Financial disclosure: The author has no financial or proprietary interest in any material or method mentioned.

The author declares no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Метод оценки выпуклости рельефа целевых функций в процессе поиска экстремума

А.В. Смирнов @

МИРЭА – Российский технологический университет, Москва, 119454 Россия
@ *Автор для переписки, e-mail: av_smirnov@mirea.ru*

Резюме

Цели. Целью работы является разработка метода оценки выпуклости рельефа целевой функции (ЦФ) в окрестностях экстремума, не требующего выполнения дополнительных расчетов ЦФ и сложной математической обработки, а использующего только данные, собираемые в процессе поиска экстремума.

Методы. Выпуклость рельефа характеризуется показателем степени степенной аппроксимации ЦФ в окрестностях экстремума. Оценка этого показателя осуществляется по парам пробных точек с учетом их расстояний до найденного экстремума и значений ЦФ в них. На основе анализа погрешностей такой оценки в методе предусмотрены отбор пробных точек по их расстояниям от найденного экстремума и отбор пар пробных точек по углу между направлениями на них из найденного экстремума. Для экспериментальной проверки метода использовались тестовые функции с различной выпуклостью, как выпуклые, так и вогнутые. В качестве метода поиска экстремума применялся алгоритм роя частиц (particle swarm optimization, PSO). Результаты экспериментов представлялись в виде статистических характеристик и гистограмм распределений значений оценки показателя степени степенной аппроксимации ЦФ.

Результаты. Эксперименты показали, что при соответствующем выборе параметров отбора пробных точек и их пар метод дает достоверные значения границ диапазона, в который попадают оценки показателя степени степенной аппроксимации.

Выводы. Предложенный метод может стать частью методики анализа свойств рельефа ЦФ. Для этого необходимо дополнить его алгоритмами автоматической настройки параметров отбора пробных точек и их пар. Повышение информативности метода может быть достигнуто путем анализа распределения оценок показателя степени по расстояниям пробных точек от экстремума и направлениям на них.

Ключевые слова: рельеф целевой функции, выпуклая функция, вогнутая функция, степенная аппроксимация, показатель степени, гистограмма

• **Поступила:** 28.05.2024 • **Доработана:** 26.07.2024 • **Принята к опубликованию:** 12.02.2025

Для цитирования: Смирнов А.В. Метод оценки выпуклости рельефа целевых функций в процессе поиска экстремума. *Russian Technological Journal*. 2025;13(2):121–131. https://doi.org/10.32362/2500-316X-2025-13-2-121-131, https://elibrary.ru/EWCRYQ

Прозрачность финансовой деятельности: Автор не имеет финансовой заинтересованности в представленных материалах или методах.

Автор заявляет об отсутствии конфликта интересов.

INTRODUCTION

One of the most promising directions for the development and improvement of methods for searching for optimal solutions involves the study of the landscape properties of the optimized target objective functions (OFs) and a consideration of these properties when selecting a search algorithm or/and tuning its parameters [1]. This direction is usually referred to as exploratory landscape analysis (ELA). ELA methods are based on a definition and classification of the OF landscape properties and the development of algorithms for their quantitative evaluation by processing the results of OF calculations at trial points [2–5].

In this paper, we will be interested in the convexity characteristics of landscape properties, according to which the OF landscape areas can be divided into convex and concave ones.

Let us give the definitions [6, 7]. Function $f(\mathbf{x})$ is called convex on the set X if for $\forall(\mathbf{x}_1, \mathbf{x}_2) \in X$ and $\forall\lambda \in [0,1]$ the following condition is satisfied:

$$f(\mathbf{x}_\lambda) \le \lambda f(\mathbf{x}_1) + (1-\lambda) f(\mathbf{x}_2), \quad (1)$$

where $\mathbf{x}_\lambda = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$.

Function $f(\mathbf{x})$ is called strictly convex if the inequality in condition (1) is satisfied strictly. Function $f(\mathbf{x})$ is called concave if the function $-f(\mathbf{x})$ is convex. A strictly concave function is defined similarly. The characteristics of convexity are important for understanding the properties of OFs. In particular, if the function is concave in the neighborhood of the minimum point, such a minimum will be unstable in the sense that an insignificant shift from this point can lead to a significant increase in the value of the OF [6, 8].

The set of ELA properties includes convexity characteristics. The methodology of their estimation is as follows [2, 3]. In the search area, a set of trial points $\{\mathbf{x}_i\}$ is formed, where the values of OF $f(\mathbf{x}_i)$ are determined. From this set, pairs of points $\{\mathbf{x}_{j1}, \mathbf{x}_{j2}\}$ are randomly selected, for which the value of $f(\mathbf{x}_{j\lambda})$ at $\lambda = 0.5$ is determined, after which the difference Δ of the left and right parts of (1) is calculated. Next, the convexity probability of the OF is defined as the fraction of pairs of points for which $\Delta < \Delta_{conv}$, where $\Delta_{conv} < 0$ is a given threshold. Such a property characterizes the OF on average over the entire search area, rather than individual landscape regions, in particular, the neighborhoods of local extrema, which are of most interest. In addition, to obtain each value of $f(\mathbf{x}_{j\lambda})$ it is required to perform an additional calculation of OF, which in cases where such a calculation is performed by modeling the object, as in many optimization problems of the characteristics of radio engineering devices [9], may require significant time consumption.

In cases where the calculation of the OF gradient is performed, the convexity of the OF can be checked at each iteration by fulfilling the inequality [7]:

$$(\mathbf{x}_2 - \mathbf{x}_1)^T \cdot (\nabla f(\mathbf{x}_2) - \nabla f(\mathbf{x}_1)) > \varepsilon, \quad (2)$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ are the coordinate vectors of the initial and final iteration points; T is the transpose operation; $\nabla f(\mathbf{x})$ the OF gradient at $\mathbf{x}$; ε is a small positive number.

Calculation of the gradient requires analytical expressions for partial derivatives of the OF on coordinates or application of the finite difference method. In the latter case, the number of OF calculations that require to be performed increases significantly.

The convexity of the OF landscape is also characterized by the eigenvalues of the hessian $\nabla^2 f(\mathbf{x})$ —the matrix of second partial derivatives. The function is convex if all eigenvalues of the Hessian are nonnegative. The convexity of the landscape is characterized by absolute values of the eigenvalues along the corresponding directions. In [3], a set of properties determined by the statistics of the ratio of the maximum and minimum eigenvalues of the hessian is introduced. In [10], a measure of the degree of convexity in the form of the number of nonnegative eigenvalues is proposed. However, the computation of the hessian requires a significant number of additional calculations of the OF values.

In recent years, the use of so-called *surrogate* OF models for solving optimization problems has attracted much attention. While such a model should preserve the most important properties of the OF for the extremum search algorithm, the calculation of the values of the modeling function should require significantly less time than determining the value of the OF itself [11, 12]. A sufficiently accurate OF model will also correctly reproduce the convexity of the landscape. Although this approach has excellent prospects, the construction of corresponding models is associated with a large number of calculations.

The task of this work is to develop a method for estimating the convexity of the OF landscape during the search for extrema, which does not require the calculation of the OF derivatives and additional calculations of the OF values beyond those performed by the search algorithm itself, as well as does not require the construction of the surrogate OF models.

ANALYSIS OF THE METHOD FOR ESTIMATING THE CONVEXITY OF THE OF LANDSCAPE

Let us consider the problem of estimating the convexity characteristics of the OF landscape $f(\mathbf{x})$ in the vicinity Ω_X of the local minimum $\mathbf{x}^*$, where the following condition is satisfied:

$$f(\mathbf{x}) > f(\mathbf{x}^*), \forall\mathbf{x} \in \Omega_X. \quad (3)$$

We will search for a degree approximation of the OF changes in the vicinity of $\mathbf{x}^*$ in the form:

$$f(\mathbf{x}) - f(\mathbf{x}^*) \approx \hat{f}(\mathbf{x}) = k\left\|\mathbf{x} - \mathbf{x}^*\right\|^\alpha, \quad (4)$$

where $\|\mathbf{x}\|$ is the Euclidean norm of the vector $\mathbf{x}$. The index of degree α is an objective characteristic of the convexity of the OF landscape. At $\alpha > 1$, OF is convex, while at $\alpha < 1$, it is concave.

However, the index α does not depend on the value of OF $f(\mathbf{x}^*)$ at the point of extremum, because when this value changes by the same amount, the values of OF at other points will also shift. Therefore, in order to simplify the record, we will assume $f(\mathbf{x}^*) = 0$ without

loss of generality and consider (4) as an approximation of the OF itself.

Suppose that the point $\mathbf{x}^*$ is known, the OF is indeed a power function of the form (4), and the values of α and k are the same at all points of Ω_X. Let there be two trial points $\mathbf{x}_1$ and $\mathbf{x}_2$ and the values of the OF at them are $f(\mathbf{x}_1), f(\mathbf{x}_2)$ respectively. Then from the system of equations

$$\begin{cases} f(\mathbf{x}_1) = k\|\mathbf{x}_1 - \mathbf{x}^*\|^\alpha, \\ f(\mathbf{x}_2) = k\|\mathbf{x}_2 - \mathbf{x}^*\|^\alpha \end{cases} \quad (5)$$

we find:

$$\alpha = \frac{\ln(f(\mathbf{x}_1)) - \ln(f(\mathbf{x}_2))}{\ln(\|\mathbf{x}_1 - \mathbf{x}^*\|) - \ln(\|\mathbf{x}_2 - \mathbf{x}^*\|)}. \quad (6)$$

If the above assumptions are not fulfilled, this estimate will be approximate. Let us estimate the errors arising in this case.

Suppose that the local minimum position $\mathbf{x}'$ found in the search process differs from the true position $\mathbf{x}^*$ (Fig. 1):

$$\mathbf{x}' = \mathbf{x}^* + \Delta\mathbf{x}. \quad (7)$$

In this case we have the estimation:

$$\hat{\alpha} = \frac{\ln(f(\mathbf{x}_1)) - \ln(f(\mathbf{x}_2))}{\ln(\|\mathbf{x}_1 - \mathbf{x}'\|) - \ln(\|\mathbf{x}_2 - \mathbf{x}'\|)}. \quad (8)$$

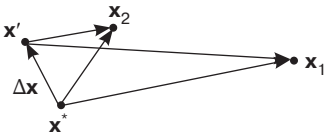


Fig. 1. Analysis of errors at inaccurate determination of the position of the OF minimum

By dividing (8) by (6) and expressing the distances from trial points $\mathbf{x}_1$, $\mathbf{x}_2$ to the true minimum $\mathbf{x}^*$ through the known distances to $\mathbf{x}'$ using the cosine theorem, we get:

$$K_\alpha = \frac{\hat{\alpha}}{\alpha} = \frac{\ln(\|\mathbf{x}_1 - \mathbf{x}^*\|) - \ln(\|\mathbf{x}_2 - \mathbf{x}^*\|)}{\ln(\|\mathbf{x}_1 - \mathbf{x}'\|) - \ln(\|\mathbf{x}_2 - \mathbf{x}'\|)} =$$

$$= \frac{0.5\ln\left(\|\mathbf{x}_1 - \mathbf{x}'\|^2 + \|\Delta\mathbf{x}\|^2 - 2\|\mathbf{x}_1 - \mathbf{x}'\| \cdot \|\Delta\mathbf{x}\| \cdot \cos\psi_1\right)}{\ln(\|\mathbf{x}_1 - \mathbf{x}'\|)} -$$

$$- \frac{0.5\ln\left(\|\mathbf{x}_2 - \mathbf{x}'\|^2 + \|\Delta\mathbf{x}\|^2 - 2\|\mathbf{x}_2 - \mathbf{x}'\| \cdot \|\Delta\mathbf{x}\| \cdot \cos\psi_2\right)}{\ln(\|\mathbf{x}_2 - \mathbf{x}'\|)}. \quad (9)$$

Here ψ_1 and ψ_2 are the angles between the vectors $(\mathbf{x}_1 - \mathbf{x}')$, $(\mathbf{x}_2 - \mathbf{x}')$ and the vector $\Delta\mathbf{x}$, respectively.

The value of K_α, which does not depend on the values of OFs in the trial points, is invariant to changes in the scale of distance measurements, making it a convenient characteristic of the estimation $\hat{\alpha}$ error. We will assume that $\|\mathbf{x}_1 - \mathbf{x}'\| > \|\mathbf{x}_2 - \mathbf{x}'\|$ and normalize all distances to $\|\mathbf{x}_2 - \mathbf{x}'\|$. Figure 2 shows the results of calculating by (9) the dependencies of the value of K_α on the distance $\|\Delta\mathbf{x}\|$ from the true to the found position of the minimum for several combinations of parameters given in Table 1. This assumption is based on the fact that, as will be seen from the following analysis, the angles between the directions to the sample points must be sufficiently small to obtain reliable estimates $\hat{\alpha}$.

Table 1. Parameters of examples of calculation of the K_α dependence on the distance $\|\Delta\mathbf{x}\|$

Examples	$\|\mathbf{x}_1 - \mathbf{x}'\|$	$\|\mathbf{x}_2 - \mathbf{x}'\|$	ψ_1	ψ_2
Example 1	10	1	90	90
Example 2	10	1	100	80
Example 3	10	1	80	100
Example 4	10	1	30	30
Example 5	10	1	150	150
Example 6	3	1	90	90
Example 7	30	1	90	90

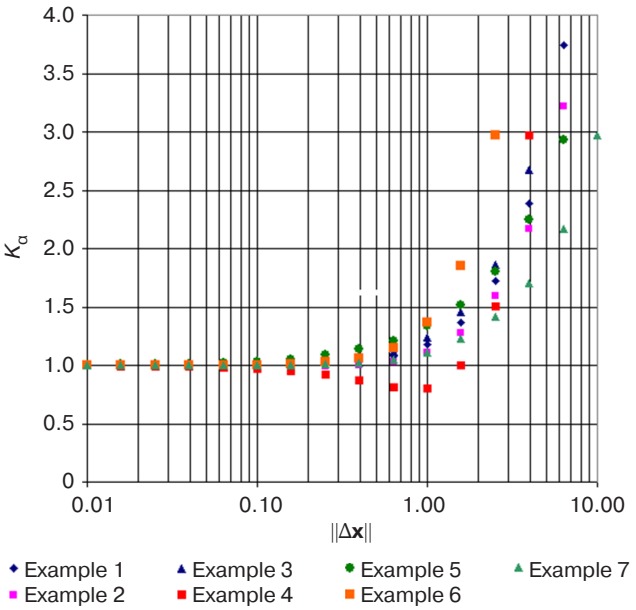


Fig. 2. Dependencies of the ratio K_α of the estimate $\hat{\alpha}$ to the true value of α on the distance $\|\Delta\mathbf{x}\|$ from the true to the found minimum position

The above results allow us to conclude that the error of the index estimation is small in cases when the distances to both trial points are significantly larger than the distance from the true position to the found position of the minimum. More specifically, when the inequality $\|\Delta\mathbf{x}\| \le 0.1\|\mathbf{x}_2 - \mathbf{x}'\|$ is satisfied, the deviation of K_α from unity does not exceed 0.1, which can be considered acceptable for approximate estimation of the convexity of the OF landscape.

Next, we consider the estimation $\hat{\alpha}$ error due to the differences in the values of α_1 and α_2, as well as k_1 and k_2 along the directions from the point of minimum $\mathbf{x}^*$ to the points $\mathbf{x}_1$ and $\mathbf{x}_2$. From (6) we obtain:

$$\hat{\alpha} = \frac{\ln\left(f(\mathbf{x}_1)\right) - \ln\left(f(\mathbf{x}_2)\right)}{\ln\left(\|\mathbf{x}_1 - \mathbf{x}^*\|\right) - \ln\left(\|\mathbf{x}_2 - \mathbf{x}^*\|\right)} = \frac{\ln\left(k_1\|\mathbf{x}_1 - \mathbf{x}^*\|^{\alpha_1}\right) - \ln\left(k_2\|\mathbf{x}_2 - \mathbf{x}^*\|^{\alpha_2}\right)}{\ln\left(\|\mathbf{x}_1 - \mathbf{x}^*\|\right) - \ln\left(\|\mathbf{x}_2 - \mathbf{x}^*\|\right)} = \bar{\alpha} + \frac{\ln\left(k_1/k_2\right)}{\ln\left(\|\mathbf{x}_1 - \mathbf{x}^*\|/\|\mathbf{x}_2 - \mathbf{x}^*\|\right)} - \frac{\Delta\alpha \ln\left(\|\mathbf{x}_1 - \mathbf{x}^*\| \cdot \|\mathbf{x}_2 - \mathbf{x}^*\|\right)}{\ln\left(\|\mathbf{x}_1 - \mathbf{x}^*\|/\|\mathbf{x}_2 - \mathbf{x}^*\|\right)}, \quad (10)$$

where $\bar{\alpha} = \frac{\alpha_1 + \alpha_2}{2}$, $\Delta\alpha = \frac{\alpha_2 - \alpha_1}{2}$.

Let us take the arithmetic mean of the indices for the two sample points $\bar{\alpha}$ as the correct estimate of the indicator α. From (10), we obtain the ratio for calculating the absolute error of this estimation.

$$E_\alpha = \hat{\alpha} - \bar{\alpha} = \frac{\ln\left(k_1/k_2\right)}{\ln\left(\|\mathbf{x}_1 - \mathbf{x}^*\|/\|\mathbf{x}_2 - \mathbf{x}^*\|\right)} - \frac{\Delta\alpha \ln\left(\|\mathbf{x}_1 - \mathbf{x}^*\| \cdot \|\mathbf{x}_2 - \mathbf{x}^*\|\right)}{\ln\left(\|\mathbf{x}_1 - \mathbf{x}^*\|/\|\mathbf{x}_2 - \mathbf{x}^*\|\right)}. \quad (11)$$

The first summand shows the contribution to the estimation $\hat{\alpha}$ error of the difference in the k coefficients at the two trial points, and the second summand shows the contribution of the difference in the α indices.

Figure 3 shows examples of dependencies of the error magnitude E_α on the distance of the second trial point from the minimum $\|\mathbf{x}_2 - \mathbf{x}^*\|$. The parameters are the distance of the first trial point from the minimum $\|\mathbf{x}_1 - \mathbf{x}^*\|$, as well as the ratio k_1/k_2 and the value Δα introduced above, which characterize the differences of the parameters of the degree approximation at the two points. The values of these parameters for each example are given in Table 2.

Table 2. Parameters of examples of calculation of the E_α dependence on $\|\mathbf{x}_2 - \mathbf{x}^*\|$ values

Examples	$\|\mathbf{x}_2 - \mathbf{x}^*\|$	k_1/k_2	Δα
Example 1	1	2	0
Example 2	1	1	0.2
Example 3	10	1	0.2
Example 4	100	1	0.2
Example 5	10	2	0.2
Example 6	10	0.5	0.2

Example 1 shows the case when the exponent α is constant in all directions, but the coefficient k varies. The error increases with distance $\|\mathbf{x}_2 - \mathbf{x}^*\|$ as the denominator of the first summand decreases. In the next three examples, only the exponent α changes. The dependencies are different for different values of $\|\mathbf{x}_1 - \mathbf{x}^*\|$ due to the fact that the second summand in (11) is not invariant to changes in the scale of distances. The absolute value of E_α with increasing $\|\mathbf{x}_2 - \mathbf{x}^*\|$ can both increase and decrease, or even turn to 0 if the equation $\|\mathbf{x}_1 - \mathbf{x}^*\| \cdot \|\mathbf{x}_2 - \mathbf{x}^*\| = 1$ is satisfied. In the examples presented in rows 5 and 6, both error components are present. The direction of change and the sign of the total error can be different depending on the ratio of parameters.

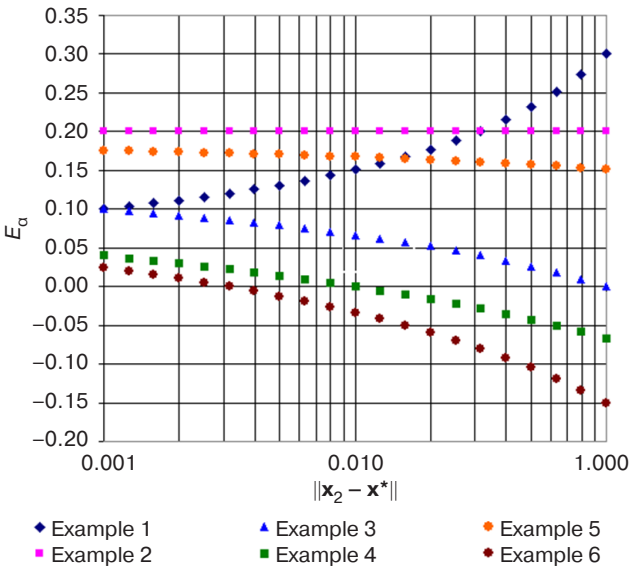


Fig. 3. Dependencies of the difference E_α of the degree index estimation $\hat{\alpha}$ and the accepted as true value $\bar{\alpha}$ on the distance of the nearest sample point to the point of minimum

Thus, the value of the error E_α is affected by the values of the differences between the parameters k and α at the two sample points, and these differences in most cases

will be less the smaller the angle between the directions to the sample points from the point of minimum.

The real OF is approximated by a step function of the form (4). In the general case, the approximation will have the form of a step series. Let us consider what information about the convexity of the landscape can be given by the estimation $\hat{\alpha}$ by two trial points. Let the OF be the sum of two degree functions:

$$f(\mathbf{x}) = k_1 \left\|\mathbf{x} - \mathbf{x}^*\right\|^{\alpha_1} + k_2 \left\|\mathbf{x} - \mathbf{x}^*\right\|^{\alpha_2}. \qquad (12)$$

The relation (8) takes the form:

$$\hat{\alpha} = \frac{\ln\left(k_1 \left\|\mathbf{x}_1 - \mathbf{x}^*\right\|^{\alpha_1} + k_2 \left\|\mathbf{x}_1 - \mathbf{x}^*\right\|^{\alpha_2}\right)}{\ln\left(\left\|\mathbf{x}_1 - \mathbf{x}^*\right\|\right)} - \frac{\ln\left(k_1 \left\|\mathbf{x}_2 - \mathbf{x}^*\right\|^{\alpha_1} + k_2 \left\|\mathbf{x}_2 - \mathbf{x}^*\right\|^{\alpha_2}\right)}{\ln\left(\left\|\mathbf{x}_2 - \mathbf{x}^*\right\|\right)}. \qquad (13)$$

Figure 4 shows examples of dependencies $\hat{\alpha}$ on the distance between the first trial point and the extremum $\|\mathbf{x}_1 - \mathbf{x}^*\|$ for the combinations of parameters given in Table 3.

Table 3. Parameters of examples of calculation of dependence $\hat{\alpha}$ on the values $\|\mathbf{x}_1 - \mathbf{x}^*\|$

Examples	α_1	α_2	k_1	k_2	$\|\mathbf{x}_1 - \mathbf{x}^*\|/\|\mathbf{x}_2 - \mathbf{x}^*\|$
Example 1	1	2	0.5	0.5	10
Example 2	1	2	0.2	0.8	10
Example 3	1	2	0.8	0.2	10
Example 4	1	2	0.5	0.5	3

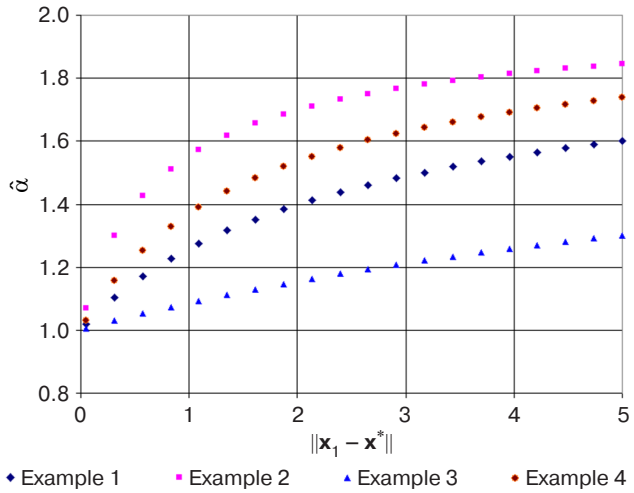


Fig. 4. Dependencies of the degree index $\hat{\alpha}$ estimation on the distance $\|\mathbf{x}_1 - \mathbf{x}^*\|$ at different combinations of parameters in relation (13)

In all the considered examples, the estimation of the degree index $\hat{\alpha}$ changes from a smaller value α_1 to a larger value α_2 as the distances of the reference points to the minimum point increase. The rate of this change depends on the ratios of the weight coefficients k_1, k_2 in (12) (examples 2 and 3), as well as on the ratio of the distances of the two reference points to the minimum point (example 4). Similar regularities will occur with a larger number of summands of the step series. These results should be taken into account when analyzing the convexity of the real OFs.

EXPERIMENTAL

The aim of the experiments was to test the possibility of obtaining reliable estimates $\hat{\alpha}$ using the described method. The methodology of experiments included obtaining sets of trial points in the process of searching for the minimum of the OF and subsequent processing of the collected data to obtain estimates $\hat{\alpha}$ at different parameters of selection of pairs of trial points. The experiments were performed using *MATLAB*[1] programs.

The well-known and widely used particle swarm optimization (PSO) algorithm [13], which, as the experience of its use shows, allows finding extrema of both convex and nonconvex OFs [14], was used as a minimum search method. With the help of this algorithm we searched for the minimum of test functions from the set [15] often used in such studies, as well as specially developed test functions. Information about the test functions will be given below together with the results of experiments. *MATLAB* function implementing the PSO algorithm was modified to return to the program calling it a data array containing the coordinates of all swarm particles in all iterations and the corresponding OF values. Subsequent processing of this data included the following steps:

1. Determination of the coordinates of the found minimum $\mathbf{x}'$ and the value of the OF at this point $f(\mathbf{x}')$.
2. Calculation of distances of all trial points $\mathbf{x}$ from the found minimum $\mathbf{x}'$ and selection by fulfillment of the inequalities $d_{\min} \le \left\|\mathbf{x} - \mathbf{x}'\right\| \le d_{\max}$, where $d_{\min}$, $d_{\max}$ are the set thresholds. The value $d_{\min}$ affects the estimation $\hat{\alpha}$ error determined by the relation (9). The value $d_{\max}$ determines the size of the vicinity $\mathbf{x}'$, within which the estimation α is calculated.
3. Calculation of the entropy of the distribution of trial points along the orthants of the coordinate system centered on the point of the found minimum $\mathbf{x}'$. The entropy value is determined by the formula:

[1] https://www.mathworks.com/products/matlab.html. Accessed February 14, 2025.

$$H = -\sum_{i=1}^{Nort} P_i \log_2 P_i, \qquad (14)$$

where P_i is the probability of the point getting into the ith orthant; *Nort* is the number of orthants equal to 2^{ND}; *ND* is the dimensionality of the search space. This value gives an estimate of uniformity of distribution of trial points in different directions from the found minimum.

4. Calculation of the angles φ_{ij} between the directions to the trial points $\mathbf{x}_i$, $\mathbf{x}_j$ included in all possible pairs from the previously selected trial points.
5. Selection of pairs of points $\mathbf{x}_i$, $\mathbf{x}_j$ for estimation of the parameters of the degree approximation. The selection conditions are formulated on the basis of the above analysis of errors of the method.

$$\varphi_{ij} \le \varphi_{\max}, \quad \ln \frac{\|\mathbf{x}_i - \mathbf{x}'\|}{\|\mathbf{x}_j - \mathbf{x}'\|} \ge C_1, \qquad (15)$$

where $\varphi_{\max}$ and C_1 are the given parameters, and it is assumed that the point $\mathbf{x}_i$ is farther from the found minimum than the point $\mathbf{x}_j$. The value of C_1 determines the minimum of the denominator in (11). The value of $\varphi_{\max}$ determines the maximum angle between the directions to the points of the pair.

6. Calculation of the entropy of the distribution of the selected pairs by orthants, similarly to item 3, which gives an estimate of the completeness of information about the indicator α in different directions.
7. Calculation of estimates of the degree approximation index $\hat{\alpha}$ for the selected pairs of points according to relation (8). Formation of the histogram of the values of these estimates. Calculation of statistical characteristics of their distribution.

Examples of the results of application of the described method are given below. In the cases of isotropic OFs, in which the parameters of the power function (4) are the same in all directions from the minimum, the proposed method finds the values of these parameters with high accuracy. Such examples are not considered here, and attention is paid to anisotropic OFs, for which it is expected that there are errors due to differences in the parameters of the power function in different directions. For all used OFs, the equation $f(\mathbf{x}^*) = 0$ is satisfied, which, as explained earlier, does not lead to a loss of generality of the results.

The data are divided into two tables. Table 4 shows the initial parameters of 12 experiments. The dimensionality of the search space in all experiments is 4. The column "N_{point}" gives the total number of trial points collected during the search for the minimum. The next column gives the distance between the found minimum $\mathbf{x}'$ and the true minimum position $\mathbf{x}^*$. This

Table 4. Initial parameters of the experiments

Exp.	Function	N_{point}	$\|\mathbf{x}' - \mathbf{x}^*\|$	$d_{\min}$	$d_{\max}$	$\varphi_{\max}$	C_1
1	ellips	1980	$7.11 \cdot 10^{-5}$	$1.00 \cdot 10^{-8}$	10	10	2
2	ellips	1980	$7.11 \cdot 10^{-5}$	0.001	10	10	2
3	ellips	1980	$7.11 \cdot 10^{-5}$	0.001	10	2	2
4	ellips	1980	$7.11 \cdot 10^{-5}$	0.001	10	10	6
5	ellips	1980	$7.11 \cdot 10^{-5}$	0.001	10	2	6
6	diffpowers	1120	$1.03 \cdot 10^{-2}$	$1.00 \cdot 10^{-8}$	10	10	2
7	diffpowers	1120	$1.03 \cdot 10^{-2}$	0.001	10	10	2
8	diffpowers	1120	$1.03 \cdot 10^{-2}$	0.1	10	10	2
9	diffpowers	1120	$1.03 \cdot 10^{-2}$	0.1	10	30	2
10	TestLE4	1420	$1.20 \cdot 10^{-3}$	$1.00 \cdot 10^{-8}$	10	10	2
11	TestLE4	1420	$1.20 \cdot 10^{-3}$	0.01	10	10	2
12	TestLE4	1420	$1.20 \cdot 10^{-3}$	0.1	10	10	2

value is given for reference and is not used by the algorithm since the true position of the minimum is assumed to be unknown. The following columns contain the values of the parameters by which the sample points and their pairs are selected.

Table 5 shows the results of these experiments. Here $N_{\text{sel. point}}$ and $H_{\text{sel. point}}$ are the number of points selected according to item 2 and the entropy of their distribution over orthants, N_{pair}, H_{pair} are the same parameters for pairs of points selected according to item 5. The following columns contain the parameters of the distribution of the estimations $\hat{\alpha}$ for the selected pairs: minimum (min), mean (mean), median (med), maximum (max), standard deviation (std), skewness (skew), and kurtosis (kurt). Histograms of the estimation $\hat{\alpha}$ values for the experiments 5, 9, and 12 are shown in Fig. 5.

Let us proceed to analyze the results of the experiments.

In experiments 1–5, we studied the function ellips(**x**) [15], formed according to the equation:

$$f(\mathbf{x}) = \sum_{n=1}^{ND} \left(x_n - x_n^*\right)^2 \cdot 10^{\left(6(n-1)/(ND-1)\right)}, \qquad (16)$$

where $\mathbf{x} = (x_1, ..., x_{ND})$ are the coordinates of the point, $\mathbf{x}^* = (x_1^*, ..., x_{ND}^*)$ are the coordinates of the minimum.

Table 5. Results of experiments

Exp.	$N_{\text{sel.point}}$	$H_{\text{sel.point}}$	N_{pair}	H_{pair}	min	mean	med	max	std	skew	kurt
1	1978	3.845	202413	3.659	0.0003	1.916	1.926	6.545	0.414	0.590	9.351
2	1698	3.775	148938	3.515	0.0003	1.943	1.951	6.545	0.432	0.702	9.199
3	1698	3.775	41786	3.499	0.052	1.951	1.972	4.517	0.301	−0.045	9.988
4	1698	3.775	33095	3.263	0.929	1.940	1.947	3.523	0.231	0.658	7.901
5	1698	3.775	8982	3.222	1.025	1.953	1.969	2.799	0.142	−0.038	7.830
6	1118	3.766	506	3.367	2.447	4.803	4.769	6.564	0.937	−0.381	2.319
7	1118	3.766	506	3.367	2.447	4.803	4.769	6.564	0.937	−0.381	2.319
8	744	3.668	123	3.305	2.755	4.924	4.941	6.249	0.877	−0.382	2.277
9	744	3.668	3373	3.329	2.015	4.562	4.578	6.287	0.954	−0.222	2.271
10	1419	3.706	2448	3.012	0.568	2.615	2.763	3.561	0.435	−2.237	7.655
11	1196	3.710	1078	3.155	0.568	2.534	2.823	3.033	0.578	−1.482	3.981
12	805	3.654	165	2.707	0.568	1.685	1.494	3.016	0.635	0.467	2.355

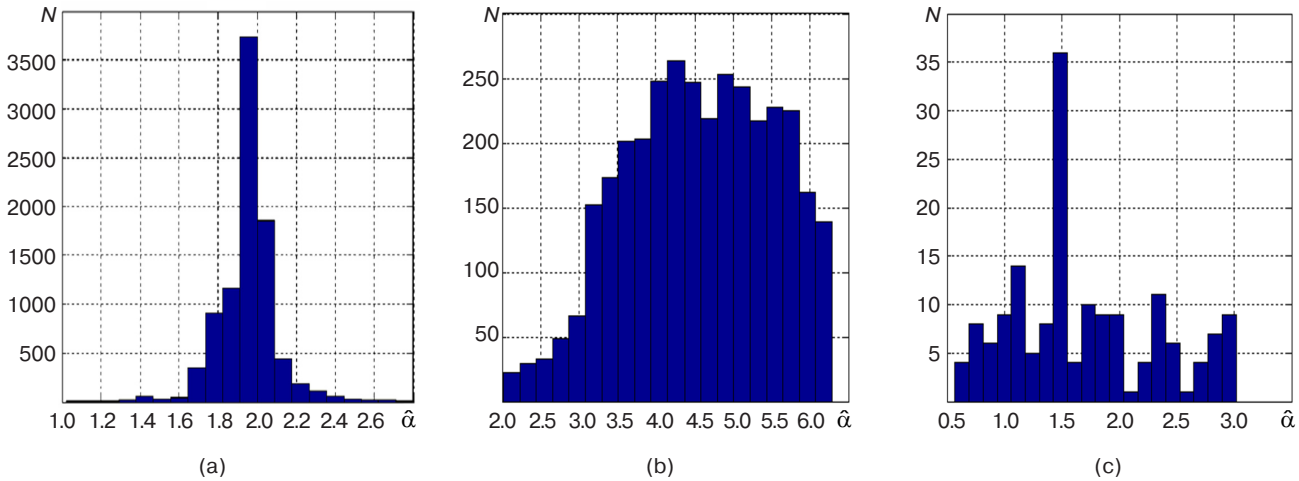


Fig. 5. Histograms of evaluation $\hat{\alpha}$: values: (a) experiment 5, (b) experiment 9, (c) experiment 12

For this OF, the degree exponent α = 2 in all directions, and the coefficient k varies in different directions in the range from 1 to 10^6.

In all experiments with this function, the mean and median values of the estimate $\hat{\alpha}$ are close to the correct value of 2. The range of estimates from minimum to maximum narrows as the constraints on pair selection become stronger, and the standard deviation decreases and reaches in experiment 5 a value of about 7% of the mean value, which can be recognized as quite satisfactory. At the same time, the shape of the distribution function of estimates turns out to be symmetric and with a sharp peak (Fig. 5a). The entropy of the distribution of selected points by orthants is close to the maximum value of 4. The entropy of the distribution of the selected pairs is smaller, but from the histogram of this distribution (not given here) we can see that in experiments 1–5 all orthants are represented, i.e., all directions are taken into account in the first approximation. This is also true for the other functions considered below.

In experiments 6–9, the function diffpowers(**x**) [15] defined by the relation:

$$f(\mathbf{x}) = \sum_{n=1}^{ND} \left(x_n - x_n^* \right)^{\left(2+4(n-1)/(ND-1)\right)}, \qquad (17)$$

where the notation is the same as in (16). This function is the sum of degree functions from different components of the point coordinate vector. Degree exponents vary in the range from 2 to 6.

In experiments 6–8, the parameter $d_{\min}$ increases successively, and the number of selected pairs of points decreases. In experiment 7, this leads to narrowing of the range of estimates $\hat{\alpha}$, but in experiment 8, the number of sampled pairs of points becomes too small, and the lower limit of the range is shifted downward. In experiment 9, the tolerance $\varphi_{\max}$ on the angle between the points of a pair is increased. As a result, the number of selected pairs has increased significantly, and the boundaries of the range of estimates $\hat{\alpha}$ (from 2 to 6) are defined with acceptable errors. At the same time, the histogram of $\hat{\alpha}$ values for this experiment is significantly different from zero in the whole range from 2 to 6 (Fig. 5b), which is an indication of the difference of the index in the degree approximation in different directions.

In the standard set of test functions [15] there is no function whose landscape in the region of minimum can be made both convex and concave. To obtain such properties, several additional test functions were developed. Below we present the results of experiments with one of them—TestLE4(**x**) calculated by the following relations:

$$f(\mathbf{x}) = k\|\mathbf{z}\|^{\alpha},$$
$$\mathbf{z} = \mathbf{x} - \mathbf{x}^*,$$
$$k = \frac{1}{\|\mathbf{z}\|^2} \sum_{n=1}^{ND} (K_{1n} z_n^2 h(z_n) + K_{2n} z_n^2 h(-z_n)), \qquad (18)$$
$$\alpha = \frac{1}{\|\mathbf{z}\|^2} \sum_{n=1}^{ND} (W_{1n} z_n^2 h(z_n) + W_{2n} z_n^2 h(-z_n)),$$
$$h(y) = \begin{cases} 1, & y > 0, \\ 0, & y \le 0. \end{cases}$$

The variables K_{ij} and W_{ij} are elements of matrices **K** and **W**, which have dimensions 2 × ND, and represent the values of coefficients and degree exponents, respectively, along the positive and negative directions of all coordinates of the search space. The resulting values of the degree exponent k and coefficient α along the direction to the trial point are obtained by interpolation between the values of these quantities along the coordinate axes. Thus, the possibility of arbitrary setting of the parameters of the degree function along different coordinates and smooth changes of these parameters along intermediate directions is provided.

In experiments 10–12, the following parameter matrices were specified:

$$\mathbf{W} = \begin{pmatrix} 3 & 1.5 & 0.5 & 1 \\ 1.5 & 2 & 1 & 0.7 \end{pmatrix}, \qquad \mathbf{K} = \begin{pmatrix} 1 & 2 & 3 & 5 \\ 3 & 1 & 0.5 & 1 \end{pmatrix}.$$

The function is convex in some directions and concave in others, and the rate of change of the function is also different in different directions. The range of values of the degree exponent is from 0.5 to 3.

In experiments 10–12, the point selection threshold $d_{\min}$ was consistently increased. As a result, the number of selected points and pairs decreased. At the same time, the maximum value of the estimate $\hat{\alpha}$ decreased insignificantly, the minimum value remained unchanged, and the value of the distribution excess decreased significantly, i.e., the distribution became more uniform. The accuracy of estimation of the range $\hat{\alpha}$ boundaries can be considered acceptable. The histogram of estimation values is different from zero in the whole range from the lower to the upper boundaries.

These examples represent a part of the experimental data obtained using different test functions. In addition, besides the PSO algorithm, the differential evolution algorithm [13] and covariance matrix adaptation evolution strategy [16] were used.

CONCLUSIONS

The experimental results confirm the feasibility of the described method to obtain objective information about the convexity of the OF in the neighborhood of the found minimum at appropriate setting of parameters of sampling points and their pairs.

The development of a more detailed method for setting the selection parameters will require further work. One of the possible options in this respect is to automate the process of sequential change of these parameters, rather than performing this operation manually as was done when obtaining the results described above. In this connection, the criteria for selecting parameters can be obtained from statistical characteristics and the shape of the histogram of the distribution of estimates $\hat{\alpha}$. To obtain more information about the convexity of the landscape, in addition to that presented in the above histogram, it is necessary to analyze the distribution of values $\hat{\alpha}$ by distances from the point of the found minimum, as well as the multivariate distribution by distances and directions.

The described method of convexity estimation can become an integral part of the technique of analyzing the OF landscape properties.

REFERENCES

1. Malan K.M. A Survey of Advances in Landscape Analysis for Optimisation. *Algorithms.* 2021;14(2):40. https://doi.org/10.3390/a14020040
2. Mersmann O., Bischl B., Trautmann H., Preuss M., Weihs C., Rudolf G. Exploratory Landscape Analysis. In: *GECCO'11: Proceedings of the 13th Annual Genetic and Evolutionary Computation.* 2011. P. 829–836. https://doi.org/10.1145/2001576.2001690
3. Kerschke P., Trautmann H. Comprehensive Feature-Based Landscape Analysis of Continuous and Constrained Optimization Problems Using the R-package flacco. In: Bauer N., Ickstadt K., Lübke K., Szepannek G., Trautmann H., Vichi M. (Eds.). *Applications in Statistical Computing. Book Series: Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin/Heidelberg, Germany: Springer; 2019. P. 93–123. https://doi.org/10.1007/978-3-030-25147-5_7
4. Trajanov R., Dimeski S., Popovski M., Korosec P., Eftimov T. *Explainable Landscape-Aware Optimization Performance Predicion*. Preprint. 2021. http://arxiv.org/pdf/2110.11633v1, https://doi.org/10.48550/arXiv.2110.11633
5. van Stein B., Long F.X., Frenzel M., Krause P., Gitterle M., Back T. *DoE2Vec: Deep-learning Based Features for Exploratory Landscape Analysis*. Preprint. 2023. https://arxiv.org/pdf/2304.01219v1
6. Polyak B.T. *Vvedenie v optimizatsiyu* (*Introduction into Optimization*). Moscow: Nauka; 1983. 384 p. (in Russ.).
7. Nocedal J., Wright S. *Numerical Optimization*: 2nd ed. Springer; 2006. 684 p.
8. Bertsimas D., ten Eikelder S.C.M., den Hertog D., Trichakis N. Pareto Adaptive Robust Optimality via a Fourier-Motzkin Elimination Lens. *Math. Program.* 2024;205(9):485–538. https://doi.org/10.1007/s10107-023-01983-z
9. Smirnov A.V. Comparison of algorithms for multi-objective optimization of radio technical device characteristics. *Russian Technological Journal.* 2022;10(6):42−51. https://doi.org/10.32362/2500-316X-2022-10-6-42-51
10. Doikov N., Stich S.U., Jaggi M. *Spectral Preconditioning for Gradient Methods on Graded Non-convex Functions*. Preprint. 2024. https://arxiv.org/pdf/2402.04843v1
11. Yaochu J. A Comprehensive Survey of Fitness Approximation in Evolutionary Computation. *Soft Computing*. 2005;9(1):3–12. https://doi.org/10.1007/s00500-003-0328-5
12. Hong L.J., Zhang X. *Surrogate-Based Simulation Optimization*. Preprint. 2021. https://arxiv.org/pdf/2105.03893v1
13. Karpenko A.P. *Sovremennye algoritmy poiskovoi optimizatsii. Algoritmy, vdokhnovlennye prirodoi* (*Modern Search Optimization Algorithms. Nature-Inspired Optimization Algorithms*): 3rd ed. Moscow: Baumann Press; 2021. 448 p. (in Russ.).
14. Smirnov A.V. Properties of objective functions and search algorithms in multi-objective optimization problems. *Russian Technological Journal*. 2022;10(4):75−85. https://doi.org/10.32362/2500-316X-2022-10-4-75-85
15. Hansen N., Finck S., Ros R., Auger A. *Real-Parameter Black-Box Optimization Benchmarking 2009*: *Noiseless Functions Definitions*. [Research Report] RR-6829. INRIA; 2009. Available from URL: https://hal.inria.fr/inria-00362633v2
16. Hansen N. *The CMA Evolution Strategy: A Tutorial*. Preprint. 2016. https://arxiv.org/abs/1604.00772v2

СПИСОК ЛИТЕРАТУРЫ

1. Malan K.M. A Survey of Advances in Landscape Analysis for Optimisation. *Algorithms.* 2021;14(2):40. https://doi.org/10.3390/a14020040
2. Mersmann O., Bischl B., Trautmann H., Preuss M., Weihs C., Rudolf G. Exploratory Landscape Analysis. In: *GECCO'11: Proceedings of the 13th Annual Genetic and Evolutionary Computation.* 2011. P. 829–836. https://doi.org/10.1145/2001576.2001690

3. Kerschke P., Trautmann H. Comprehensive Feature-Based Landscape Analysis of Continuous and Constrained Optimization Problems Using the R-package flacco. In: Bauer N., Ickstadt K., Lübke K., Szepannek G., Trautmann H., Vichi M. (Eds.). *Applications in Statistical Computing. Book Series: Studies in Classification, Data Analysis, and Knowledge Organization*. Berlin/Heidelberg, Germany: Springer; 2019. P. 93–123. https://doi.org/10.1007/978-3-030-25147-5_7
4. Trajanov R., Dimeski S., Popovski M., Korosec P., Eftimov T. *Explainable Landscape-Aware Optimization Performance Predicion.* Preprint. 2021. http://arxiv.org/pdf/2110.11633v1, https://doi.org/10.48550/arXiv.2110.11633
5. Van Stein B., Long F.X., Frenzel M., Krause P., Gitterle M., Back T. *DoE2Vec: Deep-learning Based Features for Exploratory Landscape Analysis*. Preprint. 2023. https://arxiv.org/pdf/2304.01219v1
6. Поляк Б.Т. *Введение в оптимизацию*. М.: Наука; 1983. 384 с.
7. Nocedal J., Wright S. *Numerical Optimization*: 2nd ed. Springer; 2006. 684 p.
8. Bertsimas D., ten Eikelder S.C.M., den Hertog D., Trichakis N. Pareto Adaptive Robust Optimality via a Fourier-Motzkin Elimination Lens. *Math. Program.* 2024;205(9):485–538. https://doi.org/10.1007/s10107-023-01983-z
9. Смирнов А.В. Сравнение алгоритмов многокритериальной оптимизации характеристик радиотехнических устройств. *Russian Technological Journal*. 2022;10(6):42−51. https://doi.org/10.32362/2500-316X-2022-10-6-42-51
10. Doikov N., Stich S.U., Jaggi M. *Spectral Preconditioning for Gradient Methods on Graded Non-convex Functions*. Preprint. 2024. https://arxiv.org/pdf/2402.04843v1
11. Yaochu J. A Comprehensive Survey of Fitness Approximation in Evolutionary Computation. *Soft Computing*. 2005;9(1): 3–12. https://doi.org/10.1007/s00500-003-0328-5
12. Hong L.J., Zhang X. *Surrogate-Based Simulation Optimization*. Preprint. 2021. https://arxiv.org/pdf/2105.03893v1
13. Карпенко А.П. *Современные алгоритмы поисковой оптимизации. Алгоритмы, вдохновленные природой*: 3-е изд. М.: Изд-во МГТУ им. Н.Э. Баумана; 2021. 448 с.
14. Смирнов А.В. Свойства целевых функций и алгоритмов поиска в задачах многокритериальной оптимизации. *Russian Technological Journal*. 2022;10(4):75−85. https://doi.org/10.32362/2500-316X-2022-10-4-75-85
15. Hansen N., Finck S., Ros R., Auger A. *Real-Parameter Black-Box Optimization Benchmarking 2009*: *Noiseless Functions Definitions*. [Research Report] RR-6829. INRIA; 2009. URL: https://hal.inria.fr/inria-00362633v2
16. Hansen N. *The CMA Evolution Strategy: A Tutorial*. Preprint. 2016. https://arxiv.org/abs/1604.00772v2

About the author

Alexander V. Smirnov, Cand. Sci. (Eng.), Professor, Department of Telecommunications, Institute of Radio Electronics and Informatics, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: av_smirnov@mirea.ru. Scopus Author ID 56380930700, https://orcid.org/0000-0002-2696-8592

Об авторе

Смирнов Александр Витальевич, к.т.н., доцент, профессор кафедры телекоммуникаций, Институт радиоэлектроники и информатики, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: av_smirnov@mirea.ru. Scopus Author ID 56380930700, https://orcid.org/0000-0002-2696-8592

Translated from Russian into English by L. Bychkova
Edited for English language and spelling by Thomas A. Beavitt

Mathematical modeling

Математическое моделирование

UDC 51-74:621.791.92
https://doi.org/10.32362/2500-316X-2025-13-2-132-142
EDN EATLRM

RESEARCH ARTICLE

Mathematical modeling of technological parameters of laser powder surfacing based on approximation of the deposition track profile

Mikhail E. Soloviev [1, @], Denis V. Malyshev [1], Sergey L. Baldaev [2], Lev Kh. Baldaev [2]

[1] *Yaroslavl State Technical University, Yaroslavl, 150023 Russia*
[2] *Technological Systems of Protective Coatings, Moscow, Shcherbinka, 108851 Russia*
[@] *Corresponding author, e-mail: me_s@mail.ru*

Abstract

Objectives. Laser powder surfacing is a promising mechanical engineering technology used to effectively restore worn surfaces of parts and create special coatings with valuable properties. In the research and development of laser cladding technology, mathematical modeling methods are of crucial importance. The process of applying powder coating involves moving the spray head relative to the surface of the part to form a roller or spray path, whose sequential application results in the formation of coatings. The study sets out to evaluate methods of profile approximation and optimization of technological parameters in laser powder cladding processes.

Methods. In order to describe the dependencies of the profile parameters of the deposition paths during laser surfacing on the technological parameters of the process, mathematical modeling methods were used. The contours of the profiles of the surfacing section were obtained by analyzing images of microphotographs of thin sections of the cross sections of parts with applied surfacing. To approximate the curves of the section contours, methods of linear and nonlinear regression analysis were used. The dependence of the parameters of the profile contours of the surfacing section on the technological parameters of the spraying was represented by a two-factor parabolic regression equation. The search for optimal values of spraying technological parameters was carried out using the method of conditional optimization with linear approximation of the confidence region.

Results. A nonlinear two-parameter function was selected from three options for approximating functions of the section profile of a surfacing track. Technological surfacing parameters were mapped onto a set of parameters of the approximating contour line. Optimal values of the technological parameters of surfacing were obtained using regression models of these mappings to provide the maximum value of the area of the surfacing contour under restrictions on the proportion of the sub-melting area to the total cross-sectional area. The approximating function of the cross-sectional profile of the surfacing track was used to calculate the optimal pitch of the tracks that provides the most even surface.

Conclusions. The results of the study represent a technique for optimizing the technological parameters of laser surfacing with powder metals to ensure specified characteristics of the deposition track profile and select the track deposition step at which the most even deposition surface is achieved.

Keywords: mathematical modeling, laser cladding, section contour, approximation, regression analysis, optimization

• **Submitted:** 16.05.2024 • **Revised:** 18.08.2024 • **Accepted:** 29.01.2025

For citation: Soloviev M.E., Malyshev D.V., Baldaev S.L., Baldaev L.Kh. Mathematical modeling of technological parameters of laser powder surfacing based on approximation of the deposition track profile. *Russian Technological Journal.* 2025;13(2):132–142. https://doi.org/10.32362/2500-316X-2025-13-2-132-142, https://elibrary.ru/EATLRM

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Математическое моделирование технологических параметров порошковой лазерной наплавки на основе аппроксимации профиля дорожки напыления

М.Е. Соловьев [1, @], Д.В. Малышев [1], С.Л. Балдаев [2], Л.Х. Балдаев [2]

[1] *Ярославский государственный технический университет, Ярославль, 150023 Россия*
[2] *ООО «Технологические системы защитных покрытий», Москва, Щербинка, 108851 Россия*
[@] *Автор для переписки, e-mail: me_s@mail.ru*

Резюме

Цели. Лазерная порошковая наплавка – перспективная технология в машиностроении, позволяющая эффективно восстанавливать изношенные поверхности деталей и создавать специальные покрытия с ценными свойствами. Методы математического моделирования имеют решающее значение в исследовании и развитии технологии лазерной наплавки. Процесс нанесения порошкового покрытия предполагает перемещение распылительной головки относительно поверхности детали, образуя валик – дорожку напыления. Покрытия формируются путем последовательного нанесения этих дорожек. Целью исследования является изучение различных методов аппроксимации профиля и оптимизация технологических параметров в процессах порошковой лазерной наплавки.

Методы. Использованы методы математического моделирования для описания зависимостей параметров профиля дорожек напыления при лазерной наплавке от технологических параметров процесса. Получение контуров профилей сечения наплавки осуществлялось методами анализа изображений микрофотографий шлифов поперечных сечений деталей с наплавкой. Для аппроксимации кривых контуров сечений использовались методы линейного и нелинейного регрессионного анализа. Зависимость параметров контуров профилей сечения наплавки от технологических параметров напыления аппроксимировалась двухфакторным уравнением параболической регрессии. Поиск оптимальных значений технологических параметров напыления осуществляли методом условной оптимизации с линейной аппроксимацией доверительной области.

Результаты. Рассмотрены три варианта аппроксимирующих функций профиля сечения дорожки наплавки, из которых была выбрана нелинейная двухпараметрическая функция. Получены отображения множества технологических параметров наплавки во множество параметров аппроксимирующей линии контура. С использованием регрессионных моделей данных отображений найдены оптимальные значения технологических параметров наплавки, обеспечивающие максимальную величину площади контура наплавки при ограничениях на долю области подплавления к общей площади сечения. Аппроксимирующая функция профиля сечения дорожки наплавки использована для расчета оптимального шага нанесения дорожек, обеспечивающего наиболее ровную поверхность наплавки.

Выводы. Результаты проведенного исследования могут рассматриваться в качестве методики оптимизации технологических параметров лазерной наплавки порошковых металлов, позволяющей обеспечивать заданные характеристики профиля дорожки напыления и выбирать шаг нанесения дорожек, при котором достигается наиболее ровная поверхность наплавки.

Ключевые слова: математическое моделирование, лазерная наплавка, контур сечения, аппроксимация, регрессионный анализ, оптимизация

• **Поступила:** 16.05.2024 • **Доработана:** 18.08.2024 • **Принята к опубликованию:** 29.01.2025

Для цитирования: Соловьев М.Е., Малышев Д.В., Балдаев С.Л., Балдаев Л.Х. Математическое моделирование технологических параметров порошковой лазерной наплавки на основе аппроксимации профиля дорожки напыления. *Russian Technological Journal*. 2025;13(2):132–142. https://doi.org/10.32362/2500-316X-2025-13-2-132-142, https://elibrary.ru/EATLRM

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

A promising technology in mechanical engineering, laser powder cladding is used to effectively restore the surfaces of worn parts, as well as to create special coatings for parts having valuable properties such as increased heat resistance, wear resistance, and chemical resistance [1–3]. Various methods of gas-thermal spraying of powder coatings [4–6] involve heating powdered materials to temperatures higher than their melting points and applying them to the surface of parts using high-speed gas flows. In the case of laser cladding, the source of particle heating is an infrared laser, whose beam is focused directly on or near the surface of the part. This allows more precise control of the temperature of the material to be clad and more accurate positioning of the cladding track. Due to these advantages, laser cladding forms the basis of state-of-the-art additive metal technologies [7].

Mathematical modeling methods are the most important research tool in the development of laser cladding technology [1, 8, 9]. Traditionally, the problems of heat flux distribution over the cross-section of the cladding material and the adjacent area of the workpiece are solved. Both analytical and numerical methods are used, among which finite element methods [10–12], including those using universal engineering analysis packages, have become very common in recent years. At the same time, in practical terms, a very important characteristic of the cladding process is the shape of the cross-sectional profile of the cladding track [13]. The laser cladding process, like the processes used in other powder spraying methods, consists in the movement of the atomizer head relative to the surface of the workpiece. As a result, a roller—a spraying (surfacing) track—is formed on the surface of the workpiece. By successive application of the tracks the coating is formed. The shape of the cross-sectional profile of the spraying track determines the thickness of the coating and the quality of its surface [14]. Due to the complexity of physical and chemical processes occurring during the formation of the sprayed track, modeling of the track cross-section profile on the basis of physical principles is difficult; therefore, in practice, track cross-section profiles are approximated by processing microphotographs of experimentally obtained cross-sectional slides of sprayed tracks [15]. As approximating functions of the cross-sectional profiles, rather simple mathematical functions such as parabola, circle arc or ellipse have been used in [16–18], although in practice the shape of the profile can be more complex [15, 19]. In this regard, the aim of the present work was to investigate by mathematical modeling methods various methods of approximation of cross-sectional profiles of spraying tracks with subsequent optimization of technological parameters of laser powder cladding.

APPROXIMATION AND OPTIMIZATION METHODS

The contours of the cross-sectional profiles of the sputtering tracks were obtained by processing micrographs of cross-sectional slides of the sputtering tracks with image analysis methods using the Python OpenCV library. For this purpose, auxiliary inscriptions (if any) were manually removed from the image and the color space was converted to grayscale. Next, an array of contours was extracted from the image using the algorithm [20], in which the contour with the maximum number of elements corresponded to the track contour. To perform the procedures of image file conversion and track contour extraction, we created a Python program module that enables batch processing of the scanned image array, returning a set of files in csv format containing an array of coordinates of the selected contour.

An approximation of the section profile contour and construction of mathematical models of dependencies of approximating function parameters on technological parameters of track spraying was carried out by methods of linear and nonlinear regression analysis [21, 22]. The mathematical formulation of these tasks was as follows.

Let there be a random function $y_j(x_j, \omega_j) \in Y$ values for the fixed $x_j \in X \subset \mathbb{R}$, $j = 1, \ldots, n$, where ω_j is a random event from Ω for a given sigma algebra A and probability measure P. The purpose of the approximation is to recover in X the function $Ey(x, \omega) = \eta(x)$, which is referred to as the regression function. In this work, we consider three variants of regression functions of the type $\eta_i(x, \theta_i)$, $i = 1, 2, 3$. Here η_i are known functions comprising regression models, whose specific type will be described in the main part of the article, while θ_i are parameters from the given parametric sets Θ_i as determined by the values of y_j.

Among the three regression models studied in this paper, two are parametrically linear, while one is parametrically nonlinear. The responses of the linear in parameters model y_j can be represented in the form of:

$$y_j = \eta_{\text{lin}}(x_j, \boldsymbol{\theta}) + \varepsilon_j = \boldsymbol{\theta}^{\text{T}} f(x_j) + \varepsilon_j, \quad (1)$$

where ε_j are random variables with distribution assumed to be normal with zero expectation $E\varepsilon_j = 0$ and diagonal covariance matrix $E\varepsilon_j\varepsilon_k = \sigma^2\delta_{jk}$; $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^{\text{T}}$ is a vector of unknown parameters from $\mathbb{R}^m$; $\mathbf{f}(x) = (f_1(x), \ldots, f_m(x))^{\text{T}}$ is a vector of given, linearly independent functions on the set X.

In matrix notation $\mathbf{Y} = (y_1, \ldots, y_n)^{\text{T}}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\text{T}}$, $\mathbf{F} = (f_1(x_j), \ldots, f_m(x_j))_{j=1}^n$ the system (1) is written in the form:

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad (2)$$

where $E\mathbf{Y} = \mathbf{F}\boldsymbol{\theta}$ and the covariance matrix $\mathbf{DY}$ is equal to $\sigma^2\mathbf{I}_n$, $\mathbf{I}_n$ is a unity matrix.

In the present work, the estimates $\hat{\boldsymbol{\theta}}$ of the unknown parameters $\boldsymbol{\theta}$ were computed using the least squares method (LSM):

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{j=1}^{n} \sigma_j^{-2} (y_j - \eta(x_j, \boldsymbol{\theta}))^2 \quad (3)$$

or in the matrix notations:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} (\mathbf{Y} - \mathbf{F}\Theta)^{\text{T}} (\mathbf{Y} - \mathbf{F}\Theta). \quad (4)$$

The solution of problem (4) is reduced to the well-known formula of regression analysis

$$\hat{\boldsymbol{\theta}} = (\mathbf{F}^{\text{T}}\mathbf{F})^{-1}\mathbf{F}^{\text{T}}\mathbf{Y}. \quad (5)$$

The model adequacy dispersion s^2, which is an unbiased estimate of the variance σ^2, in this case is calculated by the formula $s^2 = SS_{\text{reg}}/(n - m)$, where

$$SS_{\text{reg}} = (\mathbf{Y} - \mathbf{F}\hat{\boldsymbol{\theta}})^{\text{T}} (\mathbf{Y} - \mathbf{F}\hat{\boldsymbol{\theta}}). \quad (6)$$

Since it was not possible to estimate the error variance from parallel experiments in the present work, the adequacy of the model could not be validated by comparing the adequacy and error variance. Therefore, the adequacy of the models was assessed qualitatively by the closeness to unity of the value of the coefficient of determination

$$R^2 = 1 - SS_{\text{reg}}/SS_{\text{tot}}, \quad (7)$$

where $SS_{\text{tot}} = (\mathbf{Y} - \bar{\mathbf{Y}})^{\text{T}} (\mathbf{Y} - \bar{\mathbf{Y}})$, $\bar{\mathbf{Y}}$ is the average value of the responses.

For calculations according to formulas (5)–(7), a Python program module was created using the linear algebra package numpy.linalg, which is used to process csv files in batch mode with coordinates of track contours obtained as a result of image processing from microphotographs of cross-sectional profiles and plot points of the original contours and regression lines obtained as a result of calculating parameter estimates of regression equations.

For the regression model that is parametrically nonlinear, instead of representing the responses in the form (2), we used the representation of:

$$\mathbf{Y} = \mathbf{H}(\text{X}, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}, \quad (8)$$

where $\mathbf{H}(\text{X}, \boldsymbol{\theta}) = (\eta(x_1, \boldsymbol{\theta}), \ldots, \eta(x_n, \boldsymbol{\theta}))^{\text{T}}$ is the vector of values of the nonlinear function $\eta(x, \boldsymbol{\theta})$ at points x_j with parameters $\boldsymbol{\theta}$.

The formulation (4) of the LSM in this case takes the following form:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} (\mathbf{Y} - \mathbf{H}(\text{X}, \Theta))^{\text{T}} (\mathbf{Y} - \mathbf{H}(\text{X}, \Theta)). \quad (9)$$

Since the system of normal equations of LSM becomes nonlinear, it is not possible to use the simple formula (5) to calculate the parameter estimates. Therefore, we used a numerical optimization method to solve this problem [23]. Formula (6) for the nonlinear model is as follows:

$$SS_{\text{reg}} = (\mathbf{Y} - \mathbf{H}(\text{X}, \hat{\boldsymbol{\theta}}))^{\text{T}} (\mathbf{Y} - \mathbf{H}(\text{X}, \hat{\boldsymbol{\theta}})), \quad (10)$$

while the general form of formula (7) for calculating the coefficient of determination remains the same.

One of the three regression models, which was selected according to the results of the adequacy analysis,

was further used to build the dependencies of the shape of the cladding track cross-sectional profile on the technological parameters of the spraying process. For this purpose, the calculated parameter estimates of the selected model were used to construct mappings of the set of technological spraying parameters $u_k \in U \subset \mathbb{R}^p$, $k = 1, \ldots, p$, into the set of profile line parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$. Since the parameter estimates $\hat{\boldsymbol{\theta}}$ are random variables and the technological parameters u_k are set parameters, linear regression analysis was used to construct such mappings. The specific type of regression functions is described in the main part of the paper. These functions were further used to optimize the technological mode of sputtering using the algorithms described in [24].

CALCULATION RESULTS AND DISCUSSION

The cross-sectional profile of the clad track can be divided into two areas [13]: the part located above the workpiece surface comprising the cladding area, while the part located below the workpiece surface is referred to as the under-melting area, which is formed as a result of deepening of the molten metal bath into the volume of the workpiece. A comparison of methods of approximating the cross-sectional profiles of the cladding region by simple approximating functions—circle arc, elliptical arc, sinusoid, and parabola—is presented in [15]. According to the comparison of residual dispersions of the compared approximating functions, it was concluded that the parabola approximation is the best option among the studied ones. It should be noted, however, that the profiles of the roll sections studied in this paper were quite regular in contrast to the profiles presented in other works, including [13]. In addition, the above approximating functions do not describe the sub-melting region, whose profile turns out to be more complex. In this connection, in the present work, first of all, polynomials of higher order compared to the parabola were studied as approximating functions.

Figure 1 shows the profile contour points obtained by image processing in [13] and their approximations by two types of polynomials. The first type of regression model represented a seventh degree polynomial of the following form:

$$y = (1 - x^2)\sum_{i=0}^{5} \theta_i^1 x^i, \qquad (11)$$

where θ_i^1 are the regression parameters, $x \in [-1; 1] \subset \mathbb{R}$ are the normalized values of the argument.

Hereinafter, the upper index y of the parameter θ denotes the number of the approximating function, while the upper index y of the argument x denotes the degree.

The second type of polynomial approximating function for the profile of the roll section, proposed in [25] where it is called a biquadratic approximator, takes the form:

$$y = (1 - x^2)(\theta_0^2 + \theta_1^2 x^2 + \theta_2^2 x^4). \qquad (12)$$

In contrast to model (11), only even degrees of the argument are retained here, thus achieving symmetry of the profile with respect to the OY axis. Theoretically, this symmetry, which should be fulfilled in coaxial surfacing, was also taken into account in [15] when choosing simple symmetric approximating functions. In both models, the common term $(1 - x^2)$ is removed to zero the functions at points $x = \pm 1$. This may be explained in terms of the practical convenience of representing the argument of regression functions in a dimensionless normalized form:

$$x^{\text{cond}} = \frac{x^{\text{nat}} - x_0}{\Delta x}, \qquad (13)$$

where x^{nat} is the value of the argument in natural units;
x^{cond} is the value in dimensionless scale;
$x_0 = (\max(x^{\text{nat}}) + \min(x^{\text{nat}}))/2$ is the mean level;
$\Delta x = (\max(x^{\text{nat}}) - \min(x^{\text{nat}}))/2$ is the step of variation.

Since both models (11) and (12) are parametrically linear, their estimates are easily calculated using formula (5). As can be seen from Fig. 1, both functions adequately approximate the cladding region and the sub-melting region. At the same time, the coefficient of determination of function *1* was $R^2 = [0.995; 0.996]$ for approximating functions of the cladding and sub-melting regions, respectively, while that of function *2* was slightly lower: $R^2 = [0.985; 0.976]$. Based on this, it would be possible to adopt, for example, function (11) as the main tool for approximating the track cross-section profiles. However, when processing micrographs of less regular profiles, the disadvantage of polynomial approximation when processing complex or irregular profiles became clear due to the appearance of additional extrema on the approximating curve, which should not exist in the physical sense. To address this issue, we propose a nonlinear approximating function of the following form:

$$y = A\cos^B\left(\frac{\pi}{2}x\right), \qquad (14)$$

where the approximation parameters, which were calculated by the nonlinear estimation method, are denoted as A, B.

The convenience of the function (14) is its symmetry due to having only two parameters: similar

to functions (11), (12), it has zeros at $x = \pm 1$. At the same time, it always has only one extremum in the interval $x \in [-1; 1]$.

Figure 2 shows the points of the profile of the cladding track cross-section according to the processing of microphotography [13] for cladding parameters: laser power is 310 W, while powder feed is 29 g/m and approximating curves are obtained by the three regression functions considered above. As can be seen, the polynomial functions adequately approximate the upper part of this profile, but are not workable for the lower part (the sub-melting region), whereas function (14) adequately describes the entire profile, including both the cladding region and the sub-melting region.

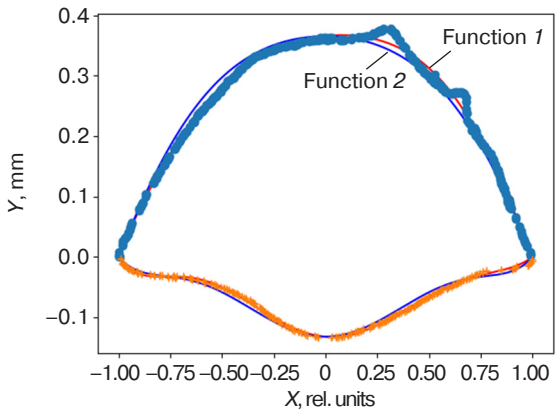


Fig. 1. Points of the cladding track cross-section profile according to microphotography processing [13] (positive *Y* values—cladding area, negative *Y* values—sub-melting area) and approximating curves: by model (11)—function *1*, by model (12)—function *2*

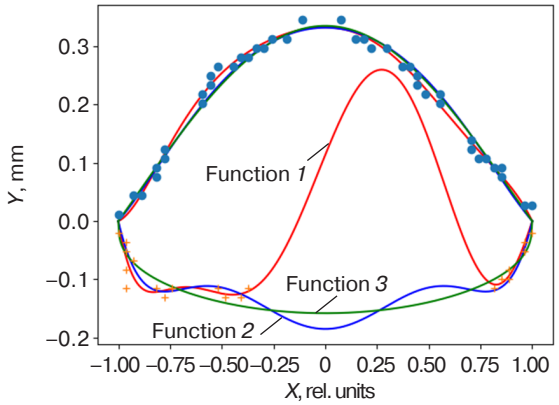


Fig. 2. Points of the cladding track cross-section profile by microphotography processing [13] and approximating curves: by model (11)—function *1*, by model (12)—function *2*, by model (14)—function *3*

Tables 1 and 2 show the parameters A, B of the approximating functions of the model (14) and the values of Δx for various technological parameters of sputtering NiCr16 nickel-chromium alloy on a steel billet obtained from the results of processing the images of profiles given in [13]. These data can be used for validation of mathematical models of cladding based on numerical solution of the equations of hydrodynamics and heat transfer using finite element methods [11, 12, 18]. By considering a wide enough range of variation of cladding parameters, it is possible to use the approximation results to construct mappings of a set of technological parameters of cladding into a set of profile line parameters and thus solve the problem of optimization of technological parameters. To construct such mappings, the method of regression analysis with approximation of profile parameter dependencies on technological parameters by two-factor parabolic regression equations of the following form was used in the present work:

$$\theta_i = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 + b_5 x_1 x_2, \quad (15)$$

where θ_i are the profile parameters; x_1 and x_2 are the laser power and powder feed in normalized units (13); b_j, $j = 0, \ldots, 5$ are the regression coefficients calculated with the help of LSM.

Table 1. Parameters of approximation of cladding path profiles at different spraying process parameters (upper part of the contour)

Laser power, W	Powder feed, g/m	Δx, mm	A_0, mm	B_0
310	12	0.333	0.145	0.661
310	29	0.429	0.335	1.022
310	45	0.500	0.606	0.689
570	12	0.476	0.228	1.035
570	29	0.524	0.495	0.558
570	45	0.587	0.521	0.664
570	85	0.802	1.084	0.541
720	12	0.508	0.233	0.973
720	29	0.516	0.481	0.597
720	45	0.603	0.598	0.678
720	63	0.746	0.815	0.769
1150	12	0.619	0.248	0.893
1150	29	0.643	0.572	0.904
1150	45	0.778	0.721	0.654
1150	63	0.873	0.959	0.463
1150	85	1.095	0.988	0.506
1150	100	1.159	0.919	0.513

Table 2. Parameters of approximation of sub-melting area profiles at different technological spraying parameters (lower part of the contour)

Laser power, W	Powder feed, g/m	Δx, mm	A_1, mm	B_1
310	12	0.333	−0.054	0.550
310	29	0.429	−0.158	0.367
310	45	0.500	−0.106	0.264
570	12	0.476	−0.245	2.518
570	29	0.524	−0.139	1.191
570	45	0.587	−0.323	0.365
570	85	0.802	−0.357	0.320
720	12	0.508	−0.213	1.088
720	29	0.516	−0.208	0.769
720	45	0.603	−0.317	0.431
720	63	0.746	−0.388	0.473
1150	12	0.619	−0.603	2.246
1150	29	0.643	−0.445	2.384
1150	45	0.778	−0.523	1.553
1150	63	0.873	−0.675	1.715
1150	85	1.095	−0.676	0.792
1150	100	1.159	−0.577	0.414

The values of the parameters Δx, A_0, A_1, B_0, B_1 given in Tables 1 and 2, as well as the areas under the profile curve S_0, S_1 calculated on their basis, were used as profile parameters. Indices 0 and 1 of the parameters correspond to the upper and lower parts of the profile, respectively. The areas were calculated by numerical integration of the profile functions (14) for the corresponding technological parameters.

Table 3 shows the calculated values of estimates of regression coefficients (15) for the studied parameters and the coefficients of determination of the models. In practice, the most interesting parameters are Δx, A_0, A_1, of which the first two characterize the half-width and height of the cladding roll, respectively, while the third characterizes the depth of the underfusion area into the part volume. Regression models for these parameters, as can be seen from Table 3, are characterized by R^2 values close to unity, which indicates their adequacy.

The areas under the profile curve are also important. They are used to calculate the relative share of the under-melting area (under-melting coefficient):

$$D = \frac{S_1}{S_0 + S_1}. \quad (16)$$

For high-quality cladding, the value of the D parameter should be optimal: small values of D provide an insufficiently strong bond between the cladding and the substrate, while excessively large values worsen the properties of the base material of the part.

The parameters of the mathematical models were derived on the basis of the calculated values of the regression equation coefficients: coordinates of critical points and eigenvalues of Hesse matrices. The graphs of regression surfaces were also plotted. According to the analysis of the obtained regression models, the dependencies of profile parameters Δx, A_0, A_1, S_0, S_1 on technological parameters x_1, x_2 are monotonic in the studied area of technological parameter changes, while the dependence $D(x_1, x_2)$ has the character of a hyperbolic paraboloid. As an example, Figs. 3–5 are plots of surfaces $A_0(x_1, x_2)$, $S_0(x_1, x_2)$, $D(x_1, x_2)$.

Table 3. Coefficients of regression equations of profile parameter dependencies on technological cladding parameters

Parameter	b_0	b_1	b_2	b_3	b_4	b_5	R^2
Δx	0.6932	0.1514	0.2539	−0.0083	0.0543	0.0356	0.9877
a_0	0.7772	0.0339	0.4114	−0.0055	−0.1445	−0.0717	0.9543
a_1	−0.3444	−0.1908	−0.1000	−0.0445	−0.0048	0.0406	0.9052
B_0	0.6222	−0.0459	−0.1436	0.0225	0.1185	−0.0918	0.5243
B_1	0.6396	0.6522	−0.7602	0.0881	0.2115	−0.2000	0.7039
S_0	0.9716	0.2314	0.8311	0.0162	0.0200	0.0824	0.9704
S_1	0.4409	0.3013	0.3442	0.0466	0.0876	0.1216	0.9594

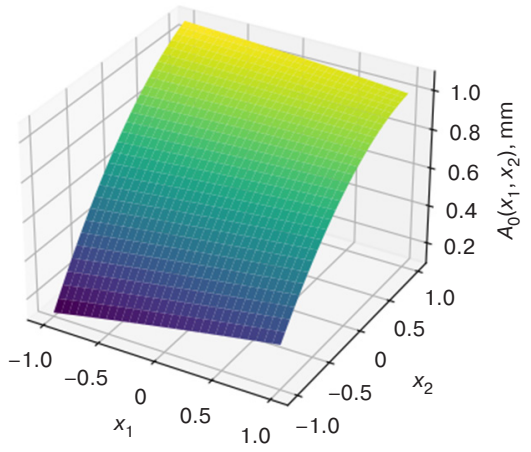


Fig. 3. Dependence of the cladding roll height on normalized values of laser power and powder feed rate

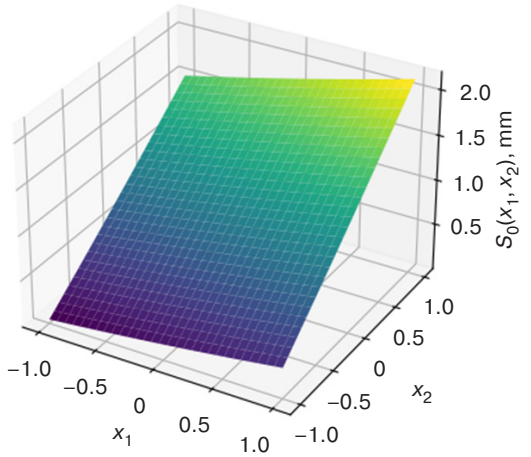


Fig. 4. Dependence of the cross-sectional area of the cladding roll on the normalized values of laser power and powder feed rate

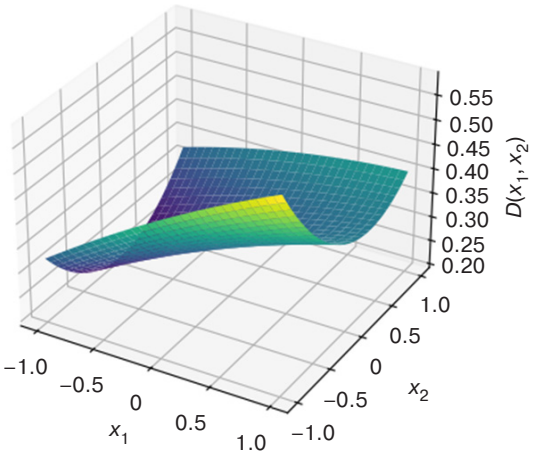


Fig. 5. Dependence of the relative share of the sub-melting area on normalized values of laser power and powder feed rate

The choice of values for the cladding profile parameters depends on the specific requirements of the product. If there is no optimization problem and it is only necessary to provide some specified profile characteristics, the construction of contour curves of regression functions can be used. As an example, Fig. 6 shows contour curves for functions $S_0(x_1, x_2)$ and $D(x_1, x_2)$. Having selected the required values of each function, the values of technological parameters x_1 and x_2 necessary for their achievement can be obtained as coordinates of the intersection points of the corresponding isolines. Thus, in particular, the values of coordinates $x_1 = -0.753$, $x_2 = -0.283$ meet the requirement $S_0 = 0.6$, $D = 0.25$. The coordinates are calculated from the solution of the nonlinear system of equations:

$$\begin{cases} S_0(x_1, x_2) - 0.6 = 0, \\ D(x_1, x_2) - 0.25 = 0. \end{cases}$$

An alternative option formulates the problem of finding the conditional extremum of one of the indicators in terms of constraints on the values of the others. For example, the coordinates of the conditional maximum of the function $S_0(x_1, x_2)$ under the constraint $D(x_1, x_2) = 0.35$, which is calculated by the conditional optimization method with a linear approximation of the confidence region [23], were the values $x_{opt} = (0.350, 0.748)$.

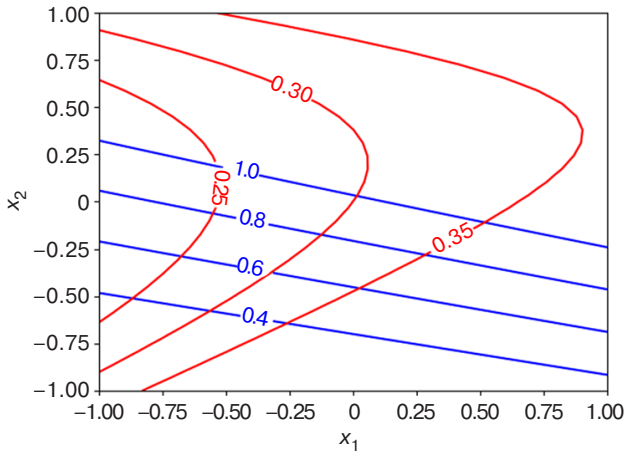


Fig. 6. Contour curves of regression functions $S_0(x_1, x_2)$ for levels {0.4, 0.6, 0.8, 1.0} and $D(x_1, x_2)$—for levels {0.25, 0.30, 0.35}

Once the optimum values of the technological parameters of surfacing have been selected, the optimum step of track application can be calculated. The technological process of surfacing consists in the sequential application of powdered material on the surface in the form of tracks with a certain step, which we will further denote w. If the step of application is

smaller than the width of the track profile, the profiles overlap, and the material of the second track fills the space between these profiles [15]. In this case, the problem arises of choosing the optimal value of w at which the excess material formed during the overlapping of profiles completely fills the free space between them. In [26], the optimal values of w for the tracks of some simple profiles are calculated. Let us calculate the optimal value of w at approximation of a profile by function (14), which parameters A_0, B_0 are calculated by regression equation (15).

Figure 7 shows the scheme for calculating the optimal value of w: curve *1* corresponds to the profile of the first track, while curve *2* corresponds to the profile of the second track overlapped with the first one. The excess material formed due to overlapping profiles theoretically forms curve *3*, which is obtained by summing up the profiles of tracks of curves *1* and *2* in the area of their overlapping. The optimum step of overlapping of tracks will be such that the area of the area ABC of intersection of profiles of tracks will be equal to the area of the area BDE located above the overlap zone. In this case, the molten metal from the area under curve *3* will evenly fill the free space between the tracks and the surface of the coating will be optimally smooth.

Taking into account the symmetry of the figures with respect to the vertical line FG, passing through the point of intersection of profiles B with dimensionless coordinate ζ, the optimal superposition of tracks assumes the equality of areas of areas ABG and BDF, which can be expressed as follows:

$$A_0\zeta - \int_0^{\zeta} A_0 \cos^{B_0}\left(\frac{\pi}{2}x\right)dx = \int_{\zeta}^{1} A_0 \cos^{B_0}\left(\frac{\pi}{2}x\right)dx. \quad (17)$$

Hence, it follows that

$$\zeta = \int_0^{1} \cos^{B_0}\left(\frac{\pi}{2}x\right)dx. \quad (18)$$

It is clear that the optimal value of the track spacing in dimensionless units can be expressed in terms of ζ by means of the formula:

$$w_{opt} = 2 - 2(1 - \zeta) = 2\zeta. \quad (19)$$

In particular, for $B_0(0.35,\ 0.75) = 0.544$, the optimal step value is $w_{opt} = 1.498$, and for $B_0(-0.753, -0.283) = 0.700$ is equal to $w_{opt} = 1.409$.

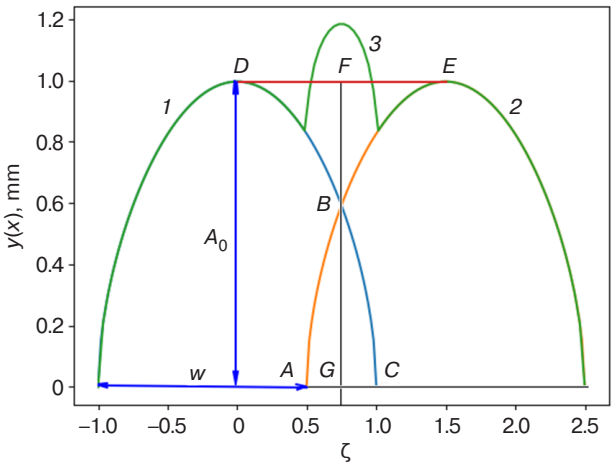


Fig. 7. Scheme for calculating the optimal value of the track overlap step

CONCLUSIONS

A method of approximating the profile of the spraying track cross-section during laser cladding of powder metals has been proposed for optimizing the cladding process parameters on its basis.

A variant of nonlinear dependence that includes two approximation parameters has been selected from three profile approximating function variants. The coefficients of the regression dependence equations of the approximating function parameters were calculated along with the contour cross-sectional area and the underfusion coefficient on the laser power and powder feed technological cladding parameters.

As a result of mathematical modeling of the dependence of the parameters of the spraying track cross-sectional profile function, the contour cross-sectional area is shown to increase monotonically with increasing laser power and powder feed, while the dependence of the sub-melting factor on the above parameters has the form of a surface with a saddle point. For these contour characteristics, the problem of conditional optimization of the contour cross-sectional area with a restriction on the value of the sub-melting coefficient is solved.

The approximating function of the cladding track cross-section profile proposed in the present work can be used to solve the problem arising in practice of calculating the optimal step of track application to ensure the achievement of an even cladding surface.

Authors' contribution. All authors equally contributed to the research work.

REFERENCES

1. Toyserkani E., Khajepour A. Corbin S. *Laser Cladding*. Boca Raton: CRC Press; 2005. 263 p.
2. Ghasempour-Mouziraji M., Lagarinhos J., Afonso D., de Sousa R.A. A review study on metal powder materials and processing parameters in Laser Metal Deposition. *Opt. Laser Technol*. 2024;170:110226. https://doi.org/10.1016/j.optlastec.2023.110226
3. Cheng J., Xing Y., Dong E., Zhao L., Liu H., Chang T., Chen M., Wang J., Lu J., Wan J. An Overview of Laser Metal Deposition for Cladding: Defect Formation Mechanisms, Defect Suppression Methods and Performance Improvements of Laser-Cladded Layers. *Materials.* 2022;15(16):5522. https://doi.org/10.3390/ma15165522
4. Davis J.R. *Handbook of Thermal Spray Technology*. ASM International; 2004. 338 p.
5. Baldaev L.H. (Ed.). *Gazotermicheskoe napylenie* (*Gas Thermal Spraying*); Moscow: Market DS; 2007. 344 p. (in Russ.).
6. Il'yushchenko A.F., Shevtsov A.I., Okovityi V.A., Gromyko V.F. *Protsessy formirovaniya gazotermicheskikh pokrytii i ikh modelirovanie* (*Processes of Formation of Gas-Thermal Coatings and Their Modeling*). Minsk: Belarus. nauka; 2011. 357 p. (in Russ.).
7. Bian L., Shamsaei N., Usher J. (Eds.). *Laser-Based Additive Manufacturing of Metal Parts: Modeling, Optimization, and Control of Mechanical Properties*. Boca Raton: CRC Press; 2018. 328 p.
8. Steen W.M., Mazumder J. *Laser Material Processing*. London: Springer; 2010. 558 p.
9. Dowden J.M. *The Mathematics of Thermal Modeling an Introduction to the Theory of Laser Material Processing*. Boca Raton: CRC Press; 2001. 292 p.
10. Pinkerton A.J. Advances in the modeling of laser direct metal deposition. *J. Laser Appl.* 2015;27:S15001. https://doi.org/10.2351/1.4815992
11. Kovalev O.B., Bedenko D.V., Zaitsev A.V. Development and application of laser cladding modeling technique: From coaxial powder feeding up to the surface deposition and bead formation. *Appl. Math. Modell*. 2018;57:339–359. https://doi.org/10.1016/j.apm.2017.09.043
12. Khamidullin B.A., Tsivilskiy I.V., Gorunov A.I., Gilmutdinov A.Kh. Modeling of the effect of powder parameters on laser cladding using coaxial nozzle. *Surf. Coat. Technol*. 2019;364:430–443. https://doi.org/10.1016/j.surfcoat.2018.12.002
13. De Oliveira U., Ocelík V., De Hosson J.Th.M. Analysis of coaxial laser cladding processing conditions. *Surf. Coat. Technol*. 2005;197(2–3):127–136. https://doi.org/10.1016/j.surfcoat.2004.06.029
14. Ocelík V., De Oliveira U., De Hosson J.Th.M. Thick tool steel coatings with laser cladding. *WIT Trans. Eng. Sci.* 2007;55. https://doi.org/10.2495/SECM070021
15. Ocelik V., Nenadl O., Palavra A., De Hosson J.Th.M. On the geometry of coating layers formed by overlap. *Surf. Coat. Technol*. 2014;242:54–61. https://doi.org/10.1016/j.surfcoat.2014.01.018
16. Jhavar S., Jain N.K., Paul C.P. Development of micro-plasma transferred arc (μ-PTA) wire deposition process for additive layer manufacturing applications. *J. Mater. Process. Technol.* 2014;214(5):1102–1110. https://doi.org/10.1016/j.jmatprotec.2013.12.016
17. Jhavar S., Jain N.K., Paul C.P. Enhancement of Deposition Quality in Micro-plasma Transferred Arc Deposition Process. *Mater. Manuf. Process.* 2014;29(8):1017–1023. https://doi.org/10.1080/10426914.2014.892984
18. Jain N.K., Sawant M.S., Nikam S.H., Jhavar S. Metal Deposition: Plasma-Based Processes. In: *Encyclopedia of Plasma Technology.* 1st ed. V. II. New York: Taylor and Francis; 2016. 19 p. http://doi.org/10.1081/E-EPLT-120053919
19. Yu T., Yang L., Zhao Yu., Sun J., Li B. Experimental research and multi-response multi-parameter optimization of laser cladding Fe313. *Opt. Laser Technol.* 2018;108:321–332. https://doi.org/10.1016/j.optlastec.2018.06.030
20. Suzuki S., KeiichiA be. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing.* 1985;30(1):32–46. https://doi.org/10.1016/0734-189X(85)90016-7
21. Draper N.R., Smith H. *Applied Regression Analysis:* 3rd ed. New York: Wiley-Interscience; 1998. 736 p.
22. Bates D.M., Watts D.G. *Nonlinear Regression Analysis and its Applications*. New York: Wiley & Sons; 1988. 365 p.
23. Vugrin K.W., Swiler L.P., Roberts R.M., Stucky-Mack N.J., Sullivan S.P. Confidence region estimation techniques for nonlinear regression in groundwater flow: Three case studies. *Water Resour. Res*. 2007;43(3):W03423. https://doi.org/10.1029/2005WR004804
24. Powell M.J.D. Direct search algorithms for optimization calculations. *Acta Numerica*. 1998;7:287–336. https://doi.org/10.1017/S0962492900002841
25. Soloviev M.E., Kokarev S.S., Baldaev S.L., Baldaev L.Kh., Mishchenko V.I., Fedorova M.O. Approximation of the profile of the section of the spray spot during gas thermal deposition of powder coating. *Informatsionno-tekhnologicheskii vestnik = Information Technology Bulletin.* 2022;3(33):138–163 (in Russ.).
26. Niz'ev V.G., Khomenko M.D., Mirzade F.Kh. Process planning and optimisation of laser cladding considering hydrodynamics and heat dissipation geometry of parts. *Quantum Electron*. 2018;48(8):743–748. https://doi.org/10.1070/QEL16708
27. Cao Y., Zhu S., Liang X., Wang W. Overlapping model of beads and curve fitting of bead section for rapid manufacturing by robotic MAG welding process. *Robotics and Computer-Integrated Manufacturing* (*RCIM*). 2011;27(3):641–645. https://doi.org/10.1016/j.rcim.2010.11.002
28. Baldaev S.L., Soloviev M.E., Raukhvarger A.B., Baldaev L.Kh., Mishchenko V.I. The Influence of Aluminum Oxide Powder Plasma Spraying Parameters on the Adhesive Strength of Ceramic Coatings Applied to the Gas Turbine Engine Thermally Stressed Components. *Vestnik MEI = Bulletin of MPEI*. 2024;1:93–102 (in Russ.). https://doi.org/10.24160/1993-6982-2024-1-93-102

About the authors

Mikhail E. Soloviev, Dr. Sci. (Phys.-Math.), Professor, Department of Information Systems and Technologies, Institute of Digital Systems, Yaroslavl State Technical University (88, Moskovskii pr., Yaroslavl, 150023 Russia). E-mail: me_s@mail.ru. Scopus Author ID 57190224257, ResearcherID A-4328-2014, RSCI SPIN-code 7444-3564, https://orcid.org/0000-0002-8840-248X

Denis V. Malyshev, Assistant, Department of Information Systems and Technologies, Institute of Digital Systems, Yaroslavl State Technical University (88, Moskovskii pr., Yaroslavl, 150023 Russia). E-mail: deniscs49@gmail.com. https://orcid.org/0009-0009-9861-1531

Sergey L. Baldaev, Cand. Sci. (Eng.), Deputy General Director, Technologies of Technological Systems of Protective Coatings (9A, Yuzhnaya ul., Shcherbinka, Moscow, 108851 Russia). E-mail: s.baldaev@tspc.ru. ResearcherID B-8056-2018, RSCI SPIN-code 6954-6407, https://orcid.org/0000-0002-1917-7979

Lev Kh. Baldaev, Dr. Sci. (Eng.), General Director, Technologies of Technological Systems of Protective Coatings (9A, Yuzhnaya ul., Shcherbinka, Moscow, 108851 Russia). E-mail: l.baldaev@tspc.ru. RSCI SPIN-code 8991-5015, https://orcid.org/0000-0002-9084-8771

Об авторах

Соловьев Михаил Евгеньевич, д.ф.-м.н. профессор, кафедра информационных систем и технологий, Институт цифровых систем, ФГБОУ «Ярославский государственный технический университет» (150023, Россия, Ярославль, Московский пр-т, д. 88). E-mail: me_s@mail.ru. Scopus Author ID 57190224257, ResearcherID A-4328-2014, SPIN-код РИНЦ 7444-3564, https://orcid.org/0000-0002-8840-248X

Малышев Денис Владимирович, ассистент, кафедра информационных систем и технологий, Институт цифровых систем, ФГБОУ «Ярославский государственный технический университет» (150023, Россия, Ярославль, Московский пр-т, д. 88). E-mail: deniscs49@gmail.com. https://orcid.org/0009-0009-9861-1531

Балдаев Сергей Львович, к.т.н., заместитель генерального директора по технологиям, ООО «Технологические системы защитных покрытий» (108851, Россия, Москва, г. Щербинка, ул. Южная, д. 9А). E-mail: s.baldaev@tspc.ru. ResearcherID B-8056-2018, SPIN-код РИНЦ 6954-6407, https://orcid.org/0000-0002-1917-7979

Балдаев Лев Христофорович, д.т.н., генеральный директор ООО «Технологические системы защитных покрытий» (108851, Москва, г. Щербинка, ул. Южная, д. 9А). E-mail: l.baldaev@tspc.ru. SPIN-код РИНЦ 8991-5015, https://orcid.org/0000-0002-9084-8771

Translated from Russian into English by L. Bychkova
Edited for English language and spelling by Thomas A. Beavitt

Mathematical modeling

Математическое моделирование

UDC 621.391:53.08
https://doi.org/10.32362/2500-316X-2025-13-2-143-154
EDN GXAGAW

RESEARCH ARTICLE

Image restoration using a discrete point spread function with consideration of finite pixel size

Victor B. Fedorov [@], Sergey G. Kharlamov, Alexey V. Fedorov

MIREA – Russian Technological University, Moscow, 119454 Russia
[@] Corresponding author, e-mail: feodorov@mirea.ru

Abstract

Objectives. The problem of restoring defocused and/or linearly blurred images using a Tikhonov-regularized inverse filter is considered. A common approach to this problem involves solving the Fredholm integral equation of the first convolution type by means of discretization based on quadrature formulas. The work sets out to obtain an expression of the point scattering function (PSF) taking into account pixel size finiteness and demonstrate its utility in application.

Methods. The research is based on signal theory and the method of digital image restoration using Tikhonov regularization.

Results. Taking into account the finiteness of the pixel size, discrete PSF formulas are obtained both for the case of a defocused image and for the case of a linearly blurred image at an arbitrary angle. It is shown that, while differences between the obtained formulas and those traditionally used are not significant under some conditions, under other conditions they can become significant.

Conclusions. In the case of restoring images at the resolution limit, i.e., when the pixel size cannot be considered negligibly small compared to the details of the image, the proposed approach can slightly improve the resolution. In addition, the derived formula for the discrete PSF corresponding to linear blur in an arbitrarily specified direction can be used to solve the problem without the need for prior image rotation and account for the blur value with sub-pixel accuracy. This offers an advantage in terms of improving the resolution of extremely fine details in the image, allowing the obtained formula to be used in solving the adaptive deconvolution problem, where precise adjustment of PSF parameters is required.

Keywords: blurred image, defocused image, resolution limit, finite pixel size, discrete PSF, image restoration, Tikhonov regularization, regularization parameter

• **Submitted:** 14.05.2024 • **Revised:** 01.07.2024 • **Accepted:** 30.01.2025

For citation: Fedorov V.B., Kharlamov S.G., Fedorov A.V. Image restoration using a discrete point spread function with consideration of finite pixel size. *Russian Technological Journal.* 2025;13(2):143–154. https://doi.org/10.32362/2500-316X-2025-13-2-143-154, https://elibrary.ru/GXAGAW

Financial disclosure: The authors have no financial or proprietary interest in any material or method mentioned.

The authors declare no conflicts of interest.

НАУЧНАЯ СТАТЬЯ

Восстановление изображений с использованием дискретной функции рассеяния точки, получаемой с учетом конечности размера пикселя

В.Б. Федоров [@], **С.Г. Харламов**, **А.В. Федоров**

МИРЭА – Российский технологический университет, Москва, 119454 Россия
[@] *Автор для переписки, e-mail: feodorov@mirea.ru*

Резюме

Цели. Рассматривается задача восстановления расфокусированного и/или линейно смазанного изображения с использованием регуляризированного по Тихонову инверсного фильтра. Распространенным подходом к решению этой задачи является решение интегрального уравнения Фредгольма 1-го рода типа свертки путем его дискретизации на основе квадратурных формул. Цель работы – получить выражение функции рассеяния точки (ФРТ) с учетом конечности размера пикселя и продемонстрировать его полезность.

Методы. Исследование основывается на теории сигналов и методе восстановления цифровых изображений с использованием тихоновской регуляризации.

Результаты. Получены формулы дискретной ФРТ как для случая расфокусированного, так и для случая линейно смазанного под произвольным углом изображения, с учетом конечности размера пикселя. Рассмотрены отличия полученных формул от традиционно используемых, показано при каких условиях эти отличия практически исчезают, а при каких – могут оказаться существенными.

Выводы. При восстановлении изображений на пределе разрешающей способности, т.е. когда размеры пикселя не могут считаться пренебрежимо малыми в сравнении с деталями изображения, предлагаемый подход может несколько улучшать разрешение. Кроме того, полученная формула дискретной ФРТ, соответствующей линейному смазу изображения в произвольно заданном направлении, позволяет не только решать задачу без необходимости предварительного поворота изображения, но и учитывать величину смаза с точностью до долей пикселя. Это дает преимущество в плане повышения разрешения предельно мелких деталей изображения и позволяет использовать данную формулу при решении задачи адаптивной деконволюции, когда требуется точная подстройка параметров ФРТ.

Ключевые слова: смазанное изображение, расфокусированное изображение, разрешающая способность, конечный размер пикселя, дискретная ФРТ, восстановление изображения, регуляризация по Тихонову, коэффициент регуляризации

• **Поступила:** 14.05.2024 • **Доработана:** 01.07.2024 • **Принята к опубликованию:** 30.01.2025

Для цитирования: Федоров В.Б., Харламов С.Г., Федоров А.В. Восстановление изображений с использованием дискретной функции рассеяния точки, получаемой с учетом конечности размера пикселя. *Russian Technological Journal*. 2025;13(2):143–154. https://doi.org/10.32362/2500-316X-2025-13-2-143-154, https://elibrary.ru/GXAGAW

Прозрачность финансовой деятельности: Авторы не имеют финансовой заинтересованности в представленных материалах или методах.

Авторы заявляют об отсутствии конфликта интересов.

INTRODUCTION

In the contemporary world, the quality of images of various objects is critical in many fields. These include medical imaging, astronomy, earth remote sensing from satellites, security monitoring, and video surveillance. In order to meet the growing demand for high quality images, researchers are challenged to improve image reconstruction and processing techniques. One of the main challenges involves the recovery of images that have been distorted by uniform linear motion of the object or camera, leading to linear blur and defocus.

The present work continues the authors' earlier study [1] to explore the issue of restoring a linearly blurred or defocused image for a case where the blur parameters are known. So far, this problem has been the subject of many investigations. For example, the theory of solving inverse non-correlated problems, which includes the problem of image restoration, is the subject of fundamental works [2–5]. The image restoration problem is also specifically addressed in the fundamental works [6–10] published in the period leading up to the early 1990s. The state of the art in this field is described in [11–15]. However, all the above studies are based on point spread function (PSF) expressions that assume that the pixel size is infinitesimally small. By contrast, the present work derives PSF expressions that take pixel size finiteness into account, which offers several advantages. Firstly, considering the finiteness of the pixel size allows for some improvement in recovery quality when recovering images captured at the resolution limit of the camera, where the pixel size cannot be considered as negligibly small compared to the image details. This is true for both linearly blurred and defocused image reconstruction. In addition, the obtained PSF expressions are continuously dependent on the blur parameters, which allows easy adjustment of these parameters to the required values within fractions of a pixel. In particular, the value and direction of linear blur values can be easily selected.

The study aims to demonstrate the advantages of the proposed discrete PSF model that accounts for the finite pixel dimensions. The paper includes a rigorous mathematical derivation of the specified PSF equations and their comparison with traditional approaches. The theoretical results are confirmed by numerical simulation of the distortions under consideration and their elimination by deconvolution using the A.N. Tikhonov regularization.

1. THE 2D DISCRETE PSF WITH A LINEAR BLUR OF THE IMAGE IN AN ARBITRARY DIRECTION

We consider a rectangular panel of light-sensitive elements, which is an $M \times N$ pixel matrix. The pixels are assumed to be square-shaped and to fill the entire panel without gaps; let w be the pixel size. Each pixel is assigned a pair of indices (m,n), $m \in \overline{0, M-1}$; $n \in \overline{0, N-1}$, the pixel in the upper left corner of the panel having indices (0, 0). We relate this panel to the Cartesian coordinate system Oxy, with the origin in the upper left corner of the panel, such that the center of the pixel with indices (m, n) lies at the point with coordinates $(mw+w/2, nw+w/2)$. The Ox axis is vertically down, while the Oy axis is vertically to the right.

Let the function $p(x, y)$ define the luminance field of the points of the panel generated by the light flux forming the image at some instant of time t. The function $p(x, y)$ is logically independent of t. Then the luminance energy accumulated by the pixel with indices (m, n) for the exposure time τ of the image moving relative to the panel (focused flux) is equal to

$$q[m,n] = \\ = \int_{mw}^{(m+1)w} dx \int_{nw}^{(n+1)w} dy \int_0^{\tau} p(x - v_x(x,y)t,\ y - v_y(x,y)t)dt,$$

where $(v_x(x, y), v_y(x, y))$ are Cartesian components of the velocity vector of the image point with coordinates (x, y). So far, we have been considering the general case where different pixels can have different velocities.

The 2D Kotelnikov interpolation series can be used to represent the luminance field, as follows:

$$p(x,y) = \sum_{m\in\mathbb{Z}} \sum_{n\in\mathbb{Z}} p[m,n] \operatorname{sinc}\left(\frac{x}{w} - m\right) \operatorname{sinc}\left(\frac{y}{w} - n\right), \quad (1)$$

where $p[m, n] = p(mw, nw)$.

Substituting this expression into the integral, we obtain the following:

$$q[k,l] = \sum_{m\in\mathbb{Z}} \sum_{n\in\mathbb{Z}} h_{k,l}[k-m, l-n] p[m,n],$$

where

$$h_{k,l}[m,n] = \\ = w^2 \int_0^{\tau} \operatorname{sinc}\left(m - \frac{v_x(kw,lw)t}{w}\right) \operatorname{sinc}\left(n - \frac{v_y(kw,lw)t}{w}\right) dt = \\ = w^2\tau \int_0^{1} \operatorname{sinc}\left(m - \frac{v_x(kw,lw)\tau}{w}t\right) \operatorname{sinc}\left(n - \frac{v_y(kw,lw)\tau}{w}t\right) dt.$$

Here, it is taken into account that the velocity field of the image motion within a pixel can be considered almost constant and equal to its value in the upper left corner of the pixel; in this case, the multiplier $w^2\tau$ is considered equal to one.

Under the assumption that the velocity field is constant over the entire pixel matrix, the 2D convolution is the following:

$$q[k,l] = \sum_{m\in\mathbb{Z}} \sum_{n\in\mathbb{Z}} h[m,n]p[k-m,l-n], \qquad (2)$$

where the kernel of this convolution is defined by the following equation:

$$h[m,n] = \int_0^1 \mathrm{sinc}(m-u_x t)\mathrm{sinc}(n-u_y t)dt, \qquad (3)$$

where $u_x = v_x\tau/w$, $u_y = v_y\tau/w$ are the displacement components in pixels for the exposure time.

With $u_x = u_y = 0$, we have $h[m, n] = \mathrm{sinc}(m)\mathrm{sinc}(n) = \delta[n]\delta[m]$, as it should be.

The examples of the graphs of the discrete kernel calculated by Eq. (3) are shown in Fig. 1.

In the general case, taking into account the finiteness of the pixel matrix size, we have a 2D finite convolution:

$$q[m,n] = \sum_{k=0}^{\min(m,\,K-1)} \sum_{l=0}^{\min(n,\,L-1)} h[k,l]p[m-k,n-l], \quad (4)$$

where $m \in \overline{0,M-1}$; $n \in \overline{0,N-1}$ and array $p[:, :]$ is assumed to be of size $M \times N$; array $h[:, :]$ is of size $K \times L$; and array $q[:, :]$ is of size $(M + K) \times (N + L)$.

In particular, when $u_x = 0$ (no vertical displacement),

$$h[m,n] = \delta[m]\int_0^1 \mathrm{sinc}(n-u_y t)dt,$$

where $\delta[m]$ is a discrete delta function, i.e., in the absence of the vertical velocity component, the 2D convolution actually reduces to the 1D convolution with the kernel, as follows:

$$h[n] = \int_0^1 \mathrm{sinc}(n-u_y t)dt.$$

In this case, taking into account the finiteness of the pixel matrix size, we get

$$q[k] = \sum_{m=0}^{\min(k,M-1)} h[m]p[k-m],$$

where $k \in \overline{0,M-1}$.

If we add the multiplier $1/w^2$ to the right-hand side of Eq. (3) and then proceed to the limit at $w \to 0$, taking into account that the kernel does not depend on the integer indices m, n but on the corresponding continuous variables $x = mw$, $y = nw$, we obtain the equation of the following form:

$$h(x,y) = \int_0^1 \delta(x-v_x\tau t)\delta(y-v_y\tau t)dt.$$

Although this equation is used in some literature on optics (e.g., [16]), it is not suitable for direct discretization in this form. It can only be discretized by replacing the delta function it contains by a suitable regular function; such a replacement by the scaled sinc function leads back to Eq. (3). However, a slightly different transformation procedure is also possible to obtain an expression suitable for discretization:

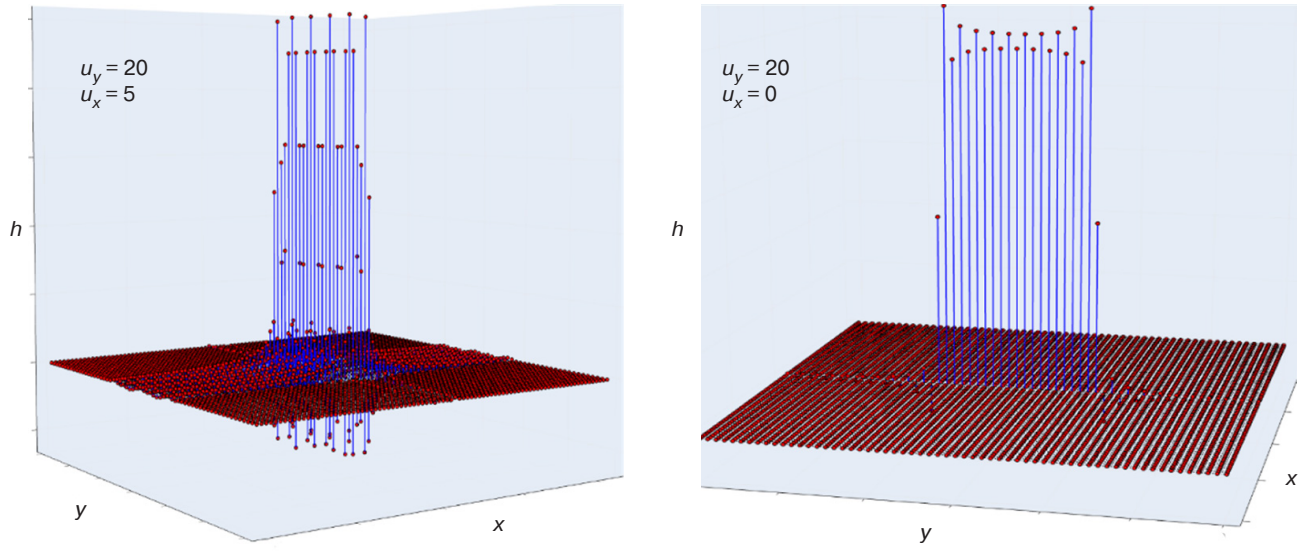


Fig. 1. Examples of graphs of the 2D discrete kernel of linear blur calculated by Eq. (3)

$$h(x,y)=\int_{-\infty}^{+\infty} I_{(0;1)}(t)\delta(x-v_x\tau t)\delta(y-v_y\tau t)dt=$$
$$=I_{(0;1)}\left(\frac{y}{v_y\tau}\right)\delta\left(x-\frac{v_x}{v_y}y\right),$$

where $I_{(0;1)}(y)$ is the indicator function of the interval (0; 1). Thus, given that the sinc function, when appropriately scaled, plays the role of the Dirac delta function in the space of functions with a finite frequency spectrum, we obtain

$$h(x,y)=I_{(0;v_y\tau)}(y)\operatorname{sinc}\left(\frac{1}{w}\left(x-\frac{v_x}{v_y}y\right)\right).$$

The scaling factor $1/w$ appearing in this substitution is discarded for convenience. Then, assuming again $x = mw$ and $y = nw$, we obtain the discrete analogue of the last equation, as follows:

$$h[m,n]=hu_y[n]\operatorname{sinc}\left(m-\frac{u_x}{u_y}n\right), \qquad (5)$$

where $u_y=\frac{v_y\tau}{w}\in\mathbb{N}$, $hu_y[n]$ is the function of the integer argument at the extreme values of the argument $n = 0$, u_y equal to 1/2; at $n = 1, 2, ..., u_y - 1$ equal to 1, and at all other n equal to 0, that correspond to the quadrature formula of trapezoids.

It should be noted that the value of the horizontal blur u_y in Eq. (5) is assumed to be a positive integer, whereas this equation imposes no such restriction on the value of the vertical component of the blur u_x; u_x can take any real value in this equation.

In particular, if we set $u_x = 0$ in (5), then considering the identity $\operatorname{sinc}(m) = \delta[m]$, we get $h[m, n] = h_{u_y}[n]\delta[m]$. Since $\delta[m]$ only differs from zero when $m = 0$, the 2D kernel $h[m, n]$ is replaced by a 1D kernel, which is most commonly used in the literature to describe horizontal linear blur (e.g., [8, 14, 15]).

Equation (5) can be considered as an alternative to Eq. (3). Like Eq. (3), it allows the recovery of a linearly blurred image with arbitrary blur direction. However, in contrast to Eq. (5), Eq. (3) removes the restrictions on the values of the horizontal blur u_y, which can be any real number in Eq. (3), as well as the value of the vertical component of the blur u_x. Thus, the use of Eq. (3) offers a number of advantages. First, considering the real value of blur within a fraction of a pixel can increase the resolution of details of the restored image when restoring an image at the resolution limit, as shown in [1]. Second, the discreteness of the parameter determining the value of the horizontal blur can be an obstacle when applying the discussed image restoration method as a basis for solving the problem of adaptive deconvolution when the direction and value of the blur are not precisely known.

2. THE 2D DISCRETE PSF AT IMAGE DEFOCUSING

For simplicity, we consider a model where the image is defocused according to the Gaussian law. Although the Gaussian model is not usually used for high quality optical systems such as telescopes and microscopes, it can be used to demonstrate the method for constructing a discrete PSF taking into account the finiteness of pixel sizes. In addition, Gaussian defocus is typically used for demonstration purposes only (e.g., [14]). If required, the Gaussian function can be replaced by any other function, e.g., the Airy function, which corresponds to the case where diffraction is the only cause of defocusing. Here, there are no fundamental restrictions.

Let the function $p(x, y)$ be the intensity of the light flux entering the aperture of the lens. Then, due to the assumed defocus of this flux, the luminance field of points on the light-sensitive panel is defined by the convolution integral

$$q(x,y)=\frac{1}{2\pi(\sigma w)^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(x-u,y-v)\mathrm{e}^{-\frac{u^2+v^2}{2\pi(\sigma w)^2}}dudv,$$

forming the image during the exposure time τ, where σ is the parameter determining the degree of defocus and w is the pixel size.

Substituting Eq. (1) into this integral, we get the following:

$$q(x,y)=\frac{1}{2\pi\sigma^2w^2}\sum_{m\in\mathbb{Z}}\sum_{n\in\mathbb{Z}}p[m,n]\times$$
$$\times\int_{-\infty}^{+\infty}\operatorname{sinc}\left(\frac{x-u}{w}-m\right)\mathrm{e}^{-\frac{u^2}{2\sigma^2w^2}}\times$$
$$\times du\int_{-\infty}^{+\infty}\operatorname{sinc}\left(\frac{y-v}{w}-n\right)\mathrm{e}^{-\frac{v^2}{2\sigma^2w^2}}dv$$

Thus, assuming that $x = wl$, $y = wk$, $k,l\in\mathbb{Z}$, we obtain the following:

$$q(k,l)=\frac{1}{2\pi\sigma^2w^2}\sum_{m\in\mathbb{Z}}\sum_{n\in\mathbb{Z}}p[m,n]\times$$
$$\times\int_{-\infty}^{+\infty}\operatorname{sinc}\left(k-m-\frac{u}{w}\right)\mathrm{e}^{-\frac{u^2}{2\sigma^2w^2}}du\times$$
$$\times\int_{-\infty}^{+\infty}\operatorname{sinc}\left(1-n-\frac{v}{w}\right)\mathrm{e}^{-\frac{v^2}{2\sigma^2w^2}}dv.$$

Going to the limit at $w \to 0$ in the obtained formula, given that $\mathrm{sinc}(x/w)/w \to \delta(x)$, we arrive at the discrete convolution of the image $p[m, n]$ with the traditional kernel representing a Gaussian grid function. In fact, the discrete convolution kernel traditionally used in this problem is obtained in the limit, as follows:

$$h_{(w\to 0)}[m,n] = \frac{1}{2\pi\sigma^2} \mathrm{e}^{-\frac{m^2+n^2}{2\sigma^2}}.$$

However, without going to the limit, replacing u/w and v/w in the last two integrals by u and v, respectively, gives the following:

$$q(k,l) = \frac{1}{2\pi\sigma^2} \sum_{m\in\mathbb{Z}} \sum_{n\in\mathbb{Z}} p[m,n] \times$$
$$\times \int_{-\infty}^{+\infty} \mathrm{sinc}\left(k-m-u\right) \mathrm{e}^{-\frac{u^2}{2\sigma^2}} du \times$$
$$\times \int_{-\infty}^{+\infty} \mathrm{sinc}\left(1-n-v\right) \mathrm{e}^{-\frac{v^2}{2\sigma^2}} dv.$$

Thus, similar to linear blur, we have a 2D discrete convolution of the form (2), whereas in the case of defocus, the corresponding kernel is separable, as follows:

$$h[m, n] = h_1[m]h_1[n], \quad (6)$$

where

$$h_1[k] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{+\infty} \mathrm{sinc}(k-u) \mathrm{e}^{-\frac{u^2}{2\sigma^2}} du = (\mathrm{sinc} * g)(k),$$

$g(z) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{z^2}{2\sigma^2}}$; here, the asterisk stands for the 1D analogue convolution operation.

We consider the convolution $f(z) = (\mathrm{sinc} * g)(z)$. According to the convolution theorem, the Fourier transform of the function $f(z)$ is the following:

$$F(v) = I_{(-0.5;0.5)}(v) \mathrm{e}^{-2\pi^2(v\sigma)^2},$$

where it is taken into account that the Fourier image of the function $\mathrm{sinc}(z)$ is an indicator function of the interval (−0.5; 0.5), while the Fourier image of the Gaussian function $g(z)$ is the function $\frac{1}{\pi}\mathrm{e}^{-2(v\sigma)^2}$. To verify the latter, recall that the Fourier image of the function $\mathrm{e}^{-\pi(z/\lambda)^2}/\lambda$ is the function $\mathrm{e}^{-\pi(v/\lambda)^2}$ (in this case, $\lambda = \sqrt{2\pi}\sigma$).

Since $h_1[k]$ is the Fourier original of the function $F(v)$ at the point $z = k$, we have the following:

$$h_1[k] = \int_{-0.5}^{0.5} \mathrm{e}^{-2(\pi v\sigma)^2} \mathrm{e}^{i2\pi vk} dv =$$
$$= \int_{-0.5}^{0.5} \mathrm{e}^{-2(\pi v\sigma)^2} \cos(2\pi vk) dv. \quad (7)$$

The last equation is due to the fact that the imaginary part of this integral should be zero. This can be verified directly, since there will be an odd function in the imaginary part under the integral. In the limit at $\sigma \to 0$, we have $h_1[k] = \delta[k]$. The graphs of the 1D kernel (7) for different values of the parameter σ are shown in Fig. 2. It can be seen that already at $\sigma = 1.0$ the values of the kernel (7) are practically indistinguishable from the limit values at $w \to 0$.

3. DECONVOLUTION

We consider Eq. (4), which is a finite 2D linear (in each dimension) discrete convolution. The equation is solved using Discrete Fourier Transform (DFT; i.e., 2D DFT). For this, the linear convolution under consideration should first be represented as a cyclic convolution as follows:

$$q[m,n] = \sum_{k=0}^{M+K-1} \sum_{l=0}^{N+L-1} h[k,l] p[(m-k)_{M+K}, (n-l)_{N+L}], \quad (8)$$

where $(m-k)_{M+K}, (n-l)_{N+L}$ are modulo $M + K$ and modulo $N + L$ residuals, respectively, $m \in \overline{0,(M+K-1)}$, $n \in \overline{0,(N+L-1)}$, and all arrays are assumed to be of equal size $(M + K) \times (N + L)$. This requires adding M null rows and N null columns to the array $h[:, :]$, and K null rows and L null columns to the array $p[:, :]$ (null rows and columns can be added, for example, to the number of the last rows and columns).

Then, according to the discrete cyclic convolution theorem, we get the following:

$$Q[m, n] = H[m, n]P[m, n], \quad (9)$$

where $m \in \overline{0,(M+K-1)}$, $n \in \overline{0,(N+L-1)}$; $H[:, :] = \mathrm{fft}(h[:, :])$, $Q[:, :] = \mathrm{fft}(q[:, :])$, $P[:, :] = \mathrm{fft}(p[:, :])$ are the 2D DPFs of the corresponding arrays.

The problem of reversing the convolution consists in solving Eq. (8) with respect to the array $p[:, :]$, given the array $q[:, :]$. This task is known to be ill-conditioned, i.e., very sensitive to errors in the original data, as well as to noise. Therefore, it is not possible to use the Eq. (9) directly for its solution; rather, it is necessary to use special regularization methods [2–8]. We use the A.N. Tikhonov regularization method, which considers

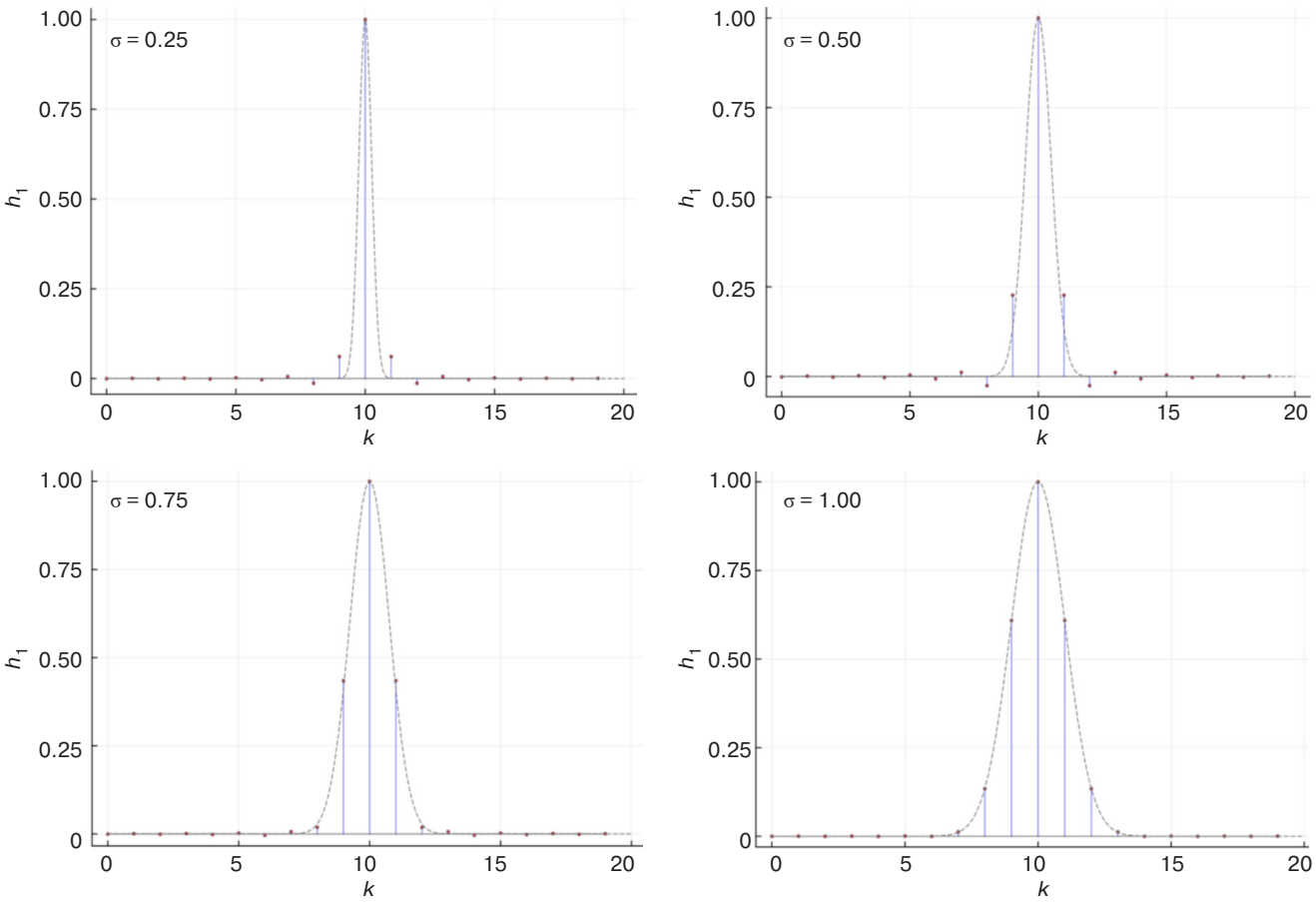


Fig. 2. Examples of 1D kernel graphs (normalized to the maximum) corresponding to the Gaussian defocus; the dashed line shows the plots of the Gaussian curves giving the kernel limits at $w \to 0$

the reciprocal formula with a regularization term instead of directly reversing Eq. (9), as follows:

$$P[m,n] = \frac{\overline{H[m,n]}}{|H[m,n]|^2 + \alpha(R[m,n])^s} Q[m,n], \quad (10)$$

where $\alpha \geq 0$ is the regularization parameter to choose for maximum restored image quality ($\alpha = 0$ means no regularization), $R[:, :]$ is an array corresponding to a chosen regularization function, and $s \geq 0$ is the regularization order. In each case, the regularization function and order are chosen individually.

For example, the regularizing array $R[:, :]$ can be calculated as follows:

$$R[m, n] = R_1[m] + R_2[n], \quad (11)$$

where

$$R_1[m] = \begin{cases} \pi\left(\frac{m}{M+K}\right)^2, & \overline{m \in 0,(M+K)/2-1}, \\ R_{M+K}\left[m - \frac{M+K}{2}\right], & \overline{m \in (M+K)/2-1,(M+K)-1}, \end{cases}$$

$$R_2[n] = \begin{cases} \pi\left(\frac{n}{N+L}\right)^2, & \overline{n \in 0,(N+L)/2-1}, \\ R_{N+L}\left[n - \frac{N+L}{2}\right], & \overline{n \in (N+L)/2-1,(N+L)-1} \end{cases}$$

(if one of the numbers here, $M + K$ or $N + L$, is odd, then dividing that number by 2 means the integer part of such a division), or as follows:

$$R[:, :] = \text{fft}(\Delta[:, :]), \quad (12)$$

where $\Delta[:, :]$ is some difference approximation of the 2D Laplace differential operator (expanded to a matrix of the desired size with zero rows and columns). The regularization order s is usually chosen low: $s = 0, 1, 2$.

It should be noted that, since we are always restoring an image of finite size, the so-called "edge effect" is inevitable. This is due to the fact that the real image to be restored does not have edges with smoothly decreasing brightness, which are always obtained when modeling a blurred or defocused image (when blurring an image of finite size). Therefore, when modeling such an image,

the smoothly decreasing edges should first be cut off in order to make the image resemble reality. Additionally, prior to reconstruction, its edges should be restored or smoothed in some way. Otherwise, the reconstructed image may contain strong artefacts in the form of the so-called Gibbs effect.

4. MODELING RESULTS

The original image used to model the defocused image, the resulting defocused image and the result of restoring it using the kernel that considers finite pixel sizes are shown in Fig. 3. The image with the larger pixel size is shown in Fig. 4. The results of deconvolution using two different PSFs are shown in Fig. 5, where the first does not consider pixel size finiteness (Fig. 5a), while the second does (Fig. 5b). The Gaussian defocus parameter is chosen such that there are noticeable differences between the graphs of the two PSFs. Comparing the results shown in Fig. 5, it is clear that the PSF taking into account the finiteness of the pixel size produces a significantly sharper image.

A similar result is shown in Fig. 6, which shows the reconstruction of the image linearly blurred in a given direction (6 pixels horizontally and 2 pixels vertically) using a kernel that accounts for the finiteness of the pixel size. A series of reconstructed images of different images with different errors in the parameters of the reconstruction kernel that determine the estimated blur vector is shown in Fig. 7. Here, the error values are 25%, 12.5%, 6%, and 0% of the true blur components. It can be seen that, firstly, there may exist situations where error values, even when expressed in fractions of pixels, can significantly worsen the result of the image restoration. Secondly, a successive monotonic reduction of the error values provides a monotonic improvement in image quality. This suggests the possibility of optimizing the parameters of the kernel used to solve the adaptive deconvolution problem.

(a)

(b)

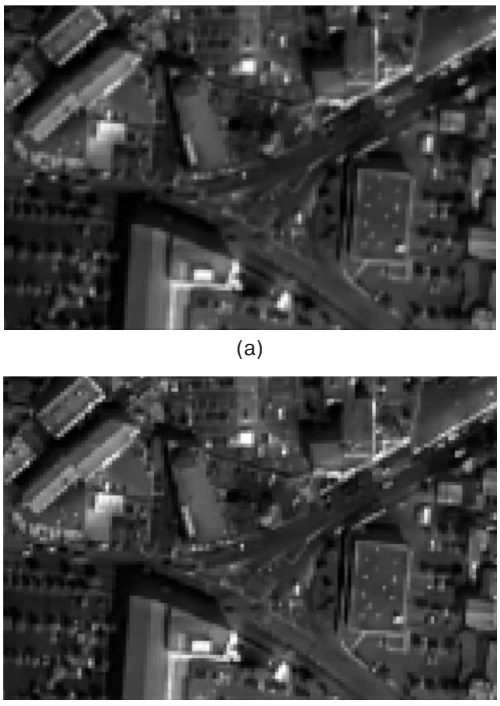
(c)

Fig. 3. Reference image and its Gaussian defocus at σ = 2 along with the result of the convolution with the regularization parameter $\alpha = 10^{-5}$ and the regularization order $s = 1$

(a)

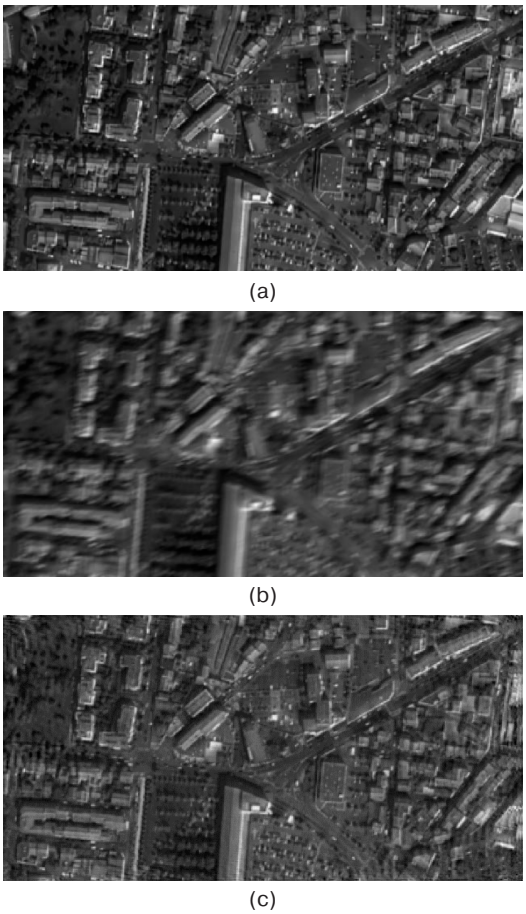
(b)

Fig. 4. Reference image (double grain size compared to Fig. 3) and its Gaussian defocus at σ = 0.4



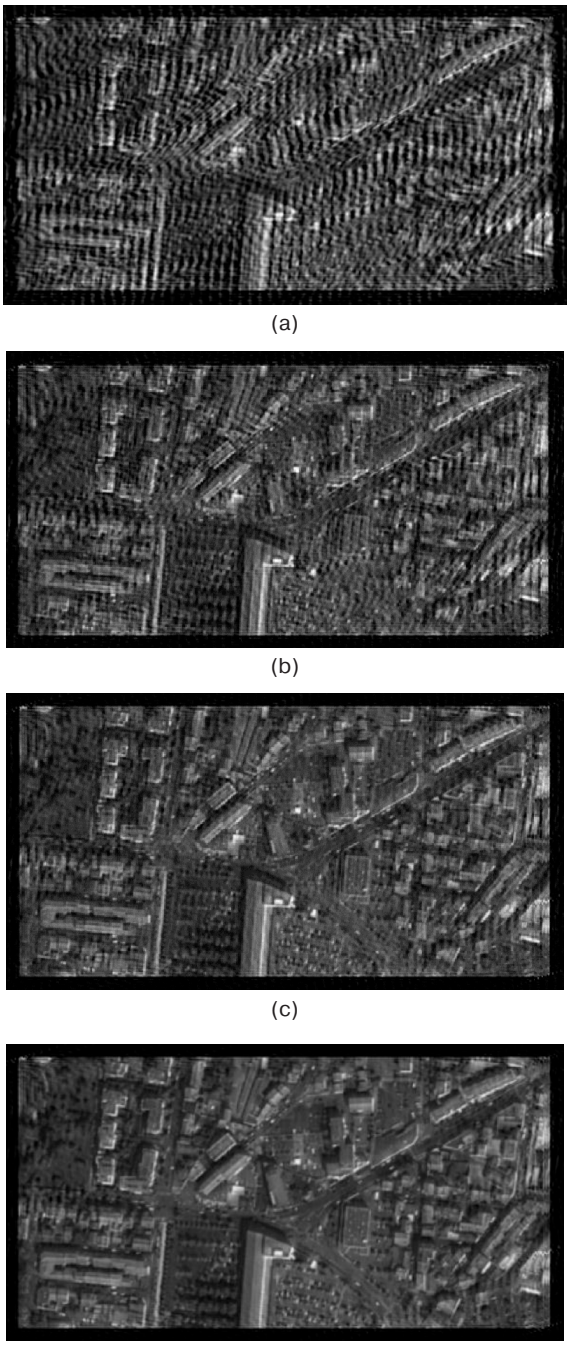
(a) (b)

Fig. 5. Convolution results of the defocused image from Fig. 4 with regularization parameter $\alpha = 10^{-5}$ and regularization order $s = 1$



(a) (b) (c)

Fig. 6. Reference image, its linear blur, and the result of its restoration with regularization parameter $\alpha = 10^{-3}$ and regularization order $s = 1$



(a) (b) (c) (d)

Fig. 7. Reconstruction results of the linearly blurred image from Fig. 6 with successively decreasing error of blur parameters, regularization parameter $\alpha = 10^{-3}$, and regularization order $s = 1$

CONCLUSIONS

The numerical modeling demonstrates the good performance of the proposed method offering the following advantages. Firstly, by taking into account the finiteness of the pixel size—or more precisely, taking into account the blur parameters within a fraction of a pixel—the resolution of the image details can be improved when the pixel size limit is reached. Importantly, this is achieved without image interpolation. Secondly, the

resulting convolution kernel equation for the linear blur makes it possible to recover the image blurred at any angle, not only horizontally. This does not require any prior image rotation to reduce the problem to restoring the horizontally blurred image. In this case, the blur values can be quite high, for example several tens of pixels. Thirdly, since the equation permits the use of blur parameters that are not necessarily expressed in terms of the integer number of pixels, a convenient opportunity arises to use it in solving the adaptive deconvolution problem, where its continuous dependence on both blur parameters may be required.

Authors' contributions

V.B. Fedorov—idea and theoretical part of the study.

S.G. Kharlamov—development of algorithms and conducting computer calculations.

A.V. Fedorov—processing results and assistance in computer calculations.

REFERENCES

1. Fedorov V.B., Kharlamov S.G., Starikovskiy A.I. Restoration of a blurred photographic image of a moving object obtained at the resolution limit. *Russian Technological Journal*. 2023;11(4):94–104. https://doi.org/10.32362/2500-316X-2023-11-4-94-104
2. Bakushinskii A.B., Goncharskii A.V. *Nekorrektnye zadachi. Chislennye metody i prilozheniya* (*Incorrect Tasks. Numerical Methods and Applications*). Moscow: MSU; 1989. 199 p. (in Russ.).
3. Goncharskii A.V., Leonov A.S., Yagola A.G. Methods for solving Fredholm integral equations of the 1st kind of convolution type. In: *Some Problems of Automated Processing and Interpretation of Physical Experiments*. V. 1. Moscow: MSU; 1973. P. 170– 191 (in Russ.).
4. Tikhonov A.N., Goncharskii A.V., Stepanov V.V., Yagola A.G. *Regulyariziruyushchie algoritmy i apriornaya informatsiya* (*Regularizing Algorithms and A Priori Information*). Moscow: Nauka; 1983. 198 p. (in Russ.).
5. Tikhonov A.N., Arsenin V.Ya. *Metody resheniya nekorrektnykh zadach* (*Methods for Solving Ill-Posed Problems*). Moscow: URSS; 2022. 288 p. (in Russ.). ISBN 978-5-9710-9341-1
6. Tikhonov A.N., Goncharskii A.V., Stepanov V.V. Inverse problems of photo image processing. In: Tikhonov A.N., Goncharskii A.V. (Eds.). *Incorrect Problems of Natural Sciences*. Moscow: MSU; 1987. P. 185–195 (in Russ.).
7. Vasilenko G.I. *Teopiya vosstanovleniya signalov*: *o reduktsii k ideal'nomu priboru v fizike i tekhnike* (*Theory of Signal Recovery*: *On the Reduction to an Ideal Device in Physics and Technology*). Moscow: Sovetskoe Radio; 1979. 272 p. (in Russ.).
8. Vasilenko G.I., Taratorin A.M. *Vosstanovlenie izobrazhenii* (*Image Restoration*). Moscow: Radio i svyaz'; 1986. 302 p. (in Russ.).
9. Bates R., McDonnell M. *Vosstanovlenie i rekonstruktsiya izobrazhenii* (*Image Restoration and Reconstruction*): transl. from Engl. Moscow: Mir; 1989. 336 p. (in Russ.).
   [Bates R., McDonnell M. *Image Restoration and Reconstruction*. New York: Oxford University Press; 1986. 312 p.]
10. Medoff B.P. Image Reconstruction from Limited Data: Theory and Application in Computed Tomography. In: Stark G. (Ed.). *Image Reconstruction*. Moscow: Mir; 1992. P. 384–436 (in Russ.).
11. Gonzales R., Woods R. *Tsifrovaya obrabotka izobrazhenii* (*Digital Image Processing*): transl. from Engl. Moscow: Tekhnosfera; 2012. 1104 p. (in Russ.).
    [Gonzales R., Woods R. *Digital Image Processing.* Pearson/Prentice Hall; 2008. 954 p.]
12. Russ J.C. *The Image Processing Handbook*. Boca Raton: CRC Press; 2007. 852 p.
13. Gruzman I.S., Kirichuk V.S., Kosykh V.P., Peretyagin G.I., Spektor A.A. *Tsifrovaya obrabotka izobrazhenii v informatsionnykh sistemakh* (*Digital Image Processing in Information Systems*). Novosibirsk: NSTU Publ.; 2002. 352 p. (in Russ.).
14. Sizikov V.S., Dovgan' A.N., Lavrov A.V. *Ustoichivye metody matematiko-komp'yuternoi obrabotki izobrazhenii i spektrov* (*Stable Methods of Mathematical and Computer Processing of Images and Spectra*). St. Petersburg: ITMO University; 2022. 70 p. (in Russ.).
15. Sizikov V.S., Ruschenko N.G. New sustainable methods for distorted image recovering. *Izvestiya vysshikh uchebnykh zavedenii. Priborostroenie = J. Instrument Eng*. 2023;66(7):559–567 (in Russ.). https://doi.org/10.17586/0021-3454-2023-66-7-559-567
16. Domnenko V.M., Bursov M.V., Ivanova T.V. *Modelirovanie formirovaniya opticheskogo izobrazheniya* (*Modeling of Optical Image Formation*). St. Petersburg: ITMO Research Institute; 2011. 141 p. (in Russ.).

СПИСОК ЛИТЕРАТУРЫ

1. Федоров В.Б., Харламов С.Г., Стариковский А.И. Восстановление смазанного фотографического изображения движущегося объекта, получаемого на пределе разрешающей способности. *Russian Technological Journal*. 2023;11(4): 94–104. https://doi.org/10.32362/2500-316X-2023-11-4-94-104
2. Бакушинский А.Б., Гончарский А.В. *Некорректные задачи. Численные методы и приложения*. М.: Изд-во МГУ; 1989. 199 с.

3. Гончарский А.В., Леонов А.С., Ягола А.Г. Методы решения интегральных уравнений Фредгольма 1-го рода типа свертки. В: *Некоторые вопросы автоматизированной обработки и интерпретации физических экспериментов*. Вып. 1. М.: Изд-во МГУ; 1973. С. 170–191.
4. Тихонов А.Н., Гончарский А.В., Степанов В.В., Ягола А.Г. *Регуляризирующие алгоритмы и априорная информация*. М.: Наука; 1983. 198 с.
5. Тихонов А.Н., Арсенин В.Я. *Методы решения некорректных задач*. М.: URSS; 2022. 288 с. ISBN 978-5- 9710-9341-1
6. Тихонов А.Н., Гончарский А.В., Степанов В.В. Обратные задачи обработки фотоизображений. В кн.: *Некорректные задачи естествознания*; под ред. А.Н. Тихонова, А.В. Гончарского. М.: Изд-во МГУ; 1987. С. 185–195.
7. Василенко Г.И. *Теория восстановления сигналов*: *о редукции к идеальному прибору в физике и технике*. М.: Сов. радио; 1979. 272 с.
8. Василенко Г.И., Тараторин А.М. *Восстановление изображений*. М.: Радио и связь; 1986. 302 с.
9. Бейтс Р., Мак-Доннелл М. *Восстановление и реконструкция изображений*: пер. с англ. М.: Мир; 1989. 336 с.
10. Медофф Б.П. Реконструкция изображений по ограниченным данным: Теория и применение в компьютерной томографии. В кн.: *Реконструкция изображений*; под ред. Г. Старка. М.: Мир; 1992. С. 384–436.
11. Гонсалес Р., Вудс Р. *Цифровая обработка изображений*: пер. с англ. М.: Техносфера; 2012. 1104 с.
12. Russ J.C. *The Image Processing Handbook*. Boca Raton: CRC Press; 2007. 852 p.
13. Грузман И.С., Киричук В.С., Косых В.П., Перетягин Г.И., Спектор А.А. *Цифровая обработка изображений в информационных системах*. Новосибирск: Изд-во НГТУ; 2002. 352 с.
14. Сизиков В.С., Довгань А.Н., Лавров А.В. *Устойчивые методы математико-компьютерной обработки изображений и спектров*. СПб.: Ун-т ИТМО; 2022. 70 с.
15. Сизиков В.С., Рущенко Н.Г. Новые устойчивые методы восстановления искаженных изображений. *Известия высших учебных заведений. Приборостроение*. 2023;66(7):559–567. https://doi.org/10.17586/0021-3454-2023-66-7-559-567
16. Домненко В.М., Бурсов М.В., Иванова Т.В. *Моделирование формирования оптического изображения*. СПб.: НИУ ИТМО; 2011. 141 с.

About the authors

Victor B. Fedorov, Cand. Sci. (Eng.), Associate Professor, Higher Mathematics Department, Institute of Artificial Intelligence, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: feodorov@mirea.ru. Scopus Author ID 57208924592, RSCI SPIN-code 2622-7666, https://orcid.org/0000-0003-1011-5453

Sergey G. Kharlamov, Postgraduate Student, Higher Mathematics Department, Institute of Artificial Intelligence, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: serhar2000@mail.ru. https://orcid.org/0000-0003-4470-6323

Alexey V. Fedorov, Master Student, Higher Mathematics Department, Institute of Artificial Intelligence, MIREA – Russian Technological University (78, Vernadskogo pr., Moscow, 119454 Russia). E-mail: alexis.sasis7@gmail.com. https://orcid.org/0009-0003-2314-7400

Об авторах

Федоров Виктор Борисович, к.т.н., доцент, кафедра высшей математики, Институт искусственного интеллекта, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: feodorov@mirea.ru. Scopus Author ID 57208924592, SPIN-код РИНЦ 2622-7666, https://orcid.org/0000-0003-1011-5453

Харламов Сергей Григорьевич, аспирант, кафедра высшей математики, Институт искусственного интеллекта, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: serhar2000@mail.ru. https://orcid.org/0000-0003-4470-6323

Федоров Алексей Викторович, магистрант, кафедра высшей математики, Институт искусственного интеллекта, ФГБОУ ВО «МИРЭА – Российский технологический университет» (119454, Россия, Москва, пр-т Вернадского, д. 78). E-mail: alexis.sasis7@gmail.com. https://orcid.org/0009-0003-2314-7400

Translated from Russian into English by K. Nazarov
Edited for English language and spelling by Thomas A. Beavitt

MIREA – Russian Technological University.
78, Vernadskogo pr., Moscow, 119454 Russian Federation.
Publication date March 28, 2025.
Not for sale.

МИРЭА – Российский технологический университет.
119454, РФ, г. Москва, пр-т Вернадского, д. 78.
Дата опубликования 28.03.2025 г.
Не для продажи.