Computing, Telecommunication and Control, 2024, Vol. 17, No. 3, Pp. 61–70.
Информатика, телекоммуникации и управление. 2024. Том 17, № 3. С. 61–70.

Research article
DOI: https://doi.org/10.18721/JCSTCS.17306
UDC 004.8

RECONSTRUCTION OF ATTRACTORS OF SUPERCOMPUTER USER'S ACTIVITY AND IDENTIFICATION OF CRITICAL DEVIATIONS IN THEIR BEHAVIOR

I.O. Mezheneva, A.A. Lukashin ✉ *, S.K. Chatoyan*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

✉ lukash.spb.ru@gmail.com

Abstract. The modern job scheduling system in supercomputer platforms is based on the estimates of the request for computing resources provided by users (often based on subjective considerations). However, it has been found that such estimates can be significantly inaccurate. In this regard, a practically important task arises: building a behavior model of user tasks executed in a supercomputer, identifying and evaluating critical deviations from the predicted behavior profile (based on an assessment of user confidence). Methods of nonlinear dynamics and topological data analysis are used to solve this problem. The article presents the results of experimental studies for various data sets obtained at the “Polytechnic Supercomputer Center” of Peter the Great St. Petersburg Polytechnic University. The Betti curves of the supercomputer user profile are calculated. The results of the evaluation of the comparison of several user profiles with the reference profile are presented. A desirability scale and numerical intervals for the proposed classes are proposed.

Keywords: high performance systems, hybrid computing systems, topological data analysis, scalar time series, job scheduling

Citation: Mezheneva I.O., Lukashin A.A., Chatoyan S.K. Reconstruction of attractors of supercomputer user's activity and identification of critical deviations in their behavior. Computing, Telecommunications and Control, 2024, Vol. 17, No. 3, Pp. 61–70. DOI: 10.18721/JCSTCS.17306

Научная статья
DOI: https://doi.org/10.18721/JCSTCS.17306
УДК 004.8

РЕКОНСТРУКЦИЯ АТТРАКТОРОВ АКТИВНОСТЕЙ ПОЛЬЗОВАТЕЛЕЙ ВЫЧИСЛИТЕЛЬНЫХ РЕСУРСОВ СУПЕРКОМПЬЮТЕРНЫХ ПЛАТФОРМ И ВЫЯВЛЕНИЕ КРИТИЧЕСКИХ ОТКЛОНЕНИЙ В ИХ ПОВЕДЕНИИ

И.О. Меженева, А.А. Лукашин ✉ , С.К. Чатоян

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

✉ lukash.spb.ru@gmail.com

Аннотация. Современная система диспетчеризации задач в суперкомпьютерных платформах основана на оценках потребности в вычислительных ресурсах, предоставленных пользователями (зачастую на основе субъективных соображений). Однако было установлено, что такие оценки могут быть существенно неточными. В связи с этим возникает важная в практическом отношении задача – построение модели поведения пользовательских заданий при их выполнении в суперкомпьютере, выявление и оценка критических отклонений от прогнозируемого профиля поведения (на основе оценки доверия к пользователю). Для решения этой задачи используются методы нелинейной динамики и топологического анализа данных. Приводятся результаты экспериментальных исследований для различных наборов данных, полученных в «Суперкомпьютерном центре “Политехнический”» Санкт-Петербургского политехнического университета Петра Великого. Посчитаны кривые Бетти профиля пользователя суперкомпьютера. Представлены результаты оценки сравнения нескольких профилей пользователей с эталонным профилем. Предложена шкала желательности и числовые интервалы для предложенных классов.

Ключевые слова: высокопроизводительные вычисления, гибридные вычислительные системы, топологический анализ данных, скалярные временные ряды, планирование задач

Для цитирования: Mezheneva I.O., Lukashin A.A., Chatoyan S.K. Reconstruction of attractors of supercomputer user's activity and identification of critical deviations in their behavior // Computing, Telecommunications and Control. 2024. Т. 17, № 3. С. 61–70. DOI: 10.18721/JCSTCS.17306

Introduction

Job scheduling is one of the key systems for supercomputer platforms, which significantly affects their performance [1]. The basis of this process is an assessment of the resource requirements of jobs, such as processor time and memory capacity. Based on this information, the dispatcher generates a schedule for completing tasks. However, existing dispatch systems rely on estimates provided by users, which often turn out to be inaccurate, resulting in inefficient use of valuable computing resources [2].

As the experience of operating the “Polytechnic Supercomputer Center” (SCC Polytechnic) shows, the inaccuracy of such estimates is usually due to the following reasons [3]:

- lack of experience of users of supercomputer platform resources in assessing the necessary needs for computing resources to solve a particular task;
- insufficient consideration of the specifics of the task being solved;
- complexity of predicting the behavior of complex algorithms, especially when using third-party libraries.

Moreover, since the dispatcher does not allocate resources beyond the requested amount, users tend to overestimate their estimates to ensure successful job completion.

The study is performed on the data, which was collected during the operation of the SCC Polytechnic. The dataset contains information about around 1.5 million of submitted jobs. SCC Polytechnic provides access to four different supercomputer clusters with the following parameters:

1. Cluster "Tornado" – consists of 612 nodes with 28-cores computers;
2. Cluster "Cascade" – consists of 81 nodes with 48-cores computers;
3. Cluster "Tornado-k40" – consists of 56 nodes with 28-cores computers with 2 GPUs each;
4. Cluster "NV" – consists of nodes with 48-cores computers with 8 GPUs each.

Each task in the dataset is submitted into one of these clusters and has information about its real execution and final task status.

In this regard, there is a need to develop methods for analyzing the behavior of users of supercomputer platforms – a method for reconstructing the dynamics of resource consumption, in particular, identifying deviations and evaluating them as critical. The proposed approach to computing behavior patterns is widely used for detecting anomalies in cybersecurity, retail, and other domains [4, 5]. Understanding the behavior patterns of supercomputer users allows to develop algorithms for improving the efficiency of using supercomputer resources.

The research methods include the theory of embedding time series in a reconstructed phase space, the theory of persistent homology, the theory of step functions, and decision theory methods based on the Harrington function.

The methodology of building profiles

The development of a methodology for building profiles based on complex and voluminous data, aimed at identifying deep patterns and abnormal behavior, as well as building descriptors reflecting various behavioral models, will be considered from the point of view of approaches based on topological data analysis (TDA) [6]. TDA, as a branch of data science, combines the principles of algebraic topology, differential geometry, functional analysis, mathematical statistics, and computer science.

In this approach, user behavioral profiles are built based on "data point clouds", which are disordered datasets that do not depend on a specific metric time (or similar) structure [7]. Topological spaces are mapped to these clouds of data points, to which TDA methods are then applied.

In particular, in task planning systems for supercomputer platforms, user behavior is often presented in the form of time series that cover multidimensional information about requests for computing resources: the type of resource, the amount of resources required (number of cluster nodes, processor requirements, memory, etc.) and the duration of the task (including actual operational data). Therefore, the first step in building a behavioral profile is to transform data from time series into point clouds and match them to the corresponding topological spaces. Thus, it is a process that ensures the integrity of information and the preservation of the existing "geometry" in the data, i.e. the choice of the appropriate topological space is carried out in such a way as to "cover" all the elements of the time series [8].

The main idea of TDA is to map a data set to the corresponding topological spaces, approximate them with simplicial complexes, and then apply the persistent homology technique to study the properties of these structures [6]. In the context of simplicial complexes, the theory of persistent homology relies on the mechanism of simplicial filtering, which systematically generates several nested, weakly dependent complexes, thereby revealing their evolution and stability at different levels of analysis. In this process, the key metrics are topological invariants, such as persistent homology groups and their numerical measure, the Betti numbers, which provide a deep understanding of the structural features of the data, and their geometry. The significance of each property is assessed through its "persistence" in filtration time – a concept that, although conditionally related to time, more accurately reflects the changing depth of analysis or the scale of consideration. The significance of this approach lies in the fact that the duration of the existence of such invariants directly correlates with the geometric structure of the studied simplicial complexes, which are approximating models of topological spaces corresponding to data clouds. Thus, persistent homology acts as a means for quantifying the stability and multilevel analysis of the topological (geometric) data characteristics.

Step 1. Converting a time series to a point cloud

The development of an approach to the embedding space construction is based on the fundamental Takens theorem application, which is aimed at the attractor reconstructing of a dynamical system from a scalar observable [9]. The embedding theory establishes that to obtain a representation of the phase space of such a system, it is possible to replace true, often inaccessible, system variables with sequences of d-dimensional vectors with a delay collected from samples of the time series $x(t)$ at successive time points:

$$\vec{x}(t) = \left(x(t),\ x(t-\tau),\ \ldots,\ x\left(t-(d-1)\tau\right)\right)^T,$$

where τ is the time delay, d is the dimension of the embedding.

The main guarantee provided by the Takens theorem is that such an embedding structure preserves the key characteristics of the original time series up to continuous maps [8]. This means that when constructing a topological embedding, we can freely choose any continuous function, among which the shift introduced through delay is the simplest option among possible transformations.

We will determine the optimal parameters of the embedding dimension and the time delay using an algorithm developed based on the L-statistical methodology. This technique follows from the concept of the noise measure, first proposed by Casdagli [10, 11], and aimed at quantifying the embedding quality, based on the analysis of the disintegration of close trajectories in the reconstructed space. If the attachment is unreliable, even minor changes can significantly distort the true state of the system, increasing the influence of noise and reducing the accuracy of reconstruction.

Unlike the classical Casdagli approach, the improved L-statistics modifies, freeing itself from the need to determine a specific prediction horizon for measuring noise amplification [12]. The algorithm implementing this principle is based on the analysis of the proximity of neighbors and strives to preserve both the geometric and topological characteristics of the original and restored attractors. The goal is to maintain a correspondence between the structural features of the original time series and their projections in the embedding space, ensuring maximum informativeness without loss of significant properties. Combining the principles of redundancy and irrelevance into a single metric, L-statistics is formalized as an objective function, the optimization of which seeks to reduce both aspects simultaneously [12].

As a result, the time series is transformed into a discrete cloud of points inside a topological (usually Euclidean space) space, i.e. $\mathbb{R}^d$. Next, TDA procedures are applied sequentially to the point cloud.

Step 2. Topological data analyses

Let us start with the assumption that the point cloud is inscribed in a metric topological space (which is guaranteed for the embedding procedure) by introducing the Euclidean metric. The next step will be to triangulate this structure using the Vietoris–Rips complex. Taking into account the sensitivity of the triangulation process to the level of proximity of points (and, accordingly, the proximity parameter introduced during the construction), we use a strategy for calculating persistent homology [6], in which the Euclidean metric gradually increases and the evolution of topological features is recorded in the form of a filtered Vietoris–Rips complex [13]. Homology groups are calculated for each K^i complex:

$$H_k^i\left(K^i\right) = Z_k^i\left(K^i\right) \big/ B_k^i\left(K^i\right),\ k = \overline{1,\ \ldots,\ n}.$$

where $Z_k^i(\cdot)$ is the cycle group, $B_k^i(\cdot)$ is a group of complex boundaries.

Persistent homology groups track the changes that occur, when the Euclidean metric (proximity parameter) changes – the appearance and disappearance of topological features – and associate the corresponding persistence with them.

As a topological descriptor — the basis of analysis, reflecting topological features in a form convenient for analysis — in this work, Betti curves were chosen, expressed through changes in Betti numbers during the process of filtration. They are step functions describing the life cycles of topological features and their resistance to changes in proximity scale. Betti numbers $\beta_k^i\left(K^i\right)$ are calculated using the formula (in the context of a vector space):

$$\beta_k^i\left(K^i\right)=\dim H_k\left(K^i\right)=\dim Z_k^i\left(K^i\right)-\dim B_k^i\left(K^i\right).$$

Betti curves were chosen because, being step functions, they allow us to calculate the average curve and to provide a simple method for estimating distances [14].

Step 3. Identification and evaluation of deviations from the basic profile

Our hypothesis is based on the idea that any deviation in the user's behavioral model entails a modification of the point cloud structure (a change in geometry in the data), which, in turn, manifests itself through noticeable shifts in topological properties. In this context, Betti curves act as a tool for visualizing the dynamics of these changes, providing a mapping of the metamorphoses of homology groups — key topological invariants.

The comparison of the obtained Betti curves with the reference profile generated according to a single methodology is performed using the Wasserstein and L_1. metrics. The evaluation process following the topological analysis requires a multidimensional approach to decision making, which involves the use of complex evaluation criteria — the construction of a generalized indicator. Within the framework of this task, we have chosen a methodology for constructing a generalized desirability indicator developed by E.K. Harrington [15].

Table 1

Desirability scale

Gradation names (linguistic meanings)	Numerical intervals
Vary bad	0–0.2
Bad	0.2–0.37
Acceptable	0.37–0.63
Good	0.63–0.8
Very good	0.8–1

The generalized Harrington desirability function provides a mechanism for converting complex topological characteristics into homogeneous numerical parameters, which greatly simplifies further interpretation and analysis. The application of this approach allows not only to more accurately assess the scale of deviations, but also to optimize the comparison process, making it more transparent and accessible to perception.

To determine the estimate, a "desirability curve" (one of Harrington's logistic functions) is used, given as follows [13]:

$$d\left(x\right)=\exp\left[-\exp\left(-y\left(x\right)\right)\right],$$

where $y\left(x\right)$ represents the encoded values of individual characteristics (scalar value), and x is a variable indicating the level or value of each characteristic. The x-axis is interpreted as a scale of individual indicators, and the d-axis is interpreted as a desirability scale, divided into five discrete ranges that determine the degrees of deviation.

The generalized desirability index is calculated as a geometric mean according to the formula [14]:

$$D = \sqrt[m]{\prod_{i=1}^{m} d_i},$$

where m is the number of individual quality criteria; d_i is the individual scores for each criterion.

The choice of the geometric mean in constructing a generalized desirability indicator is because the geometric mean plays the role of a "smoothing" mechanism that reduces the effect of random fluctuations in estimates and provides a more stable and adequate overall picture of quality [15].

Thus, the mechanism of forming a generalized desirability index works as a highly sensitive filter that allows to identify significant deviations and evaluate them by a comprehensive desirability scale.

Experimental results

In the framework of the study, the data obtained from the Supercomputer center Polytechnic and presented as a source of information on the effectiveness of performing computational tasks were selected as the analyzed data.

The studied dataset contains 1545793 records of launched tasks and their execution results. A task may be completed successfully or it may not be completed due to a user error or lack of the requested execution time. Information about each task contains the number of requested resources (processors and supercomputer nodes), as well as the results of the task, including how many and what resources were issued, when and how the task was completed. In addition, 10 problem areas were identified: astrophysics, bioinformatics, biophysics, energetics, geophysics, IT, mechanical engineering, mechanics, physics, and radiophysics. Each task belongs to one of them.

Data analysis was made based on the following job parameters from the dataset:

1. ReqNodes – Requested minimum amount of nodes for the job/step.
2. ReqCPUS – Number of requested CPUs.
3. CPUTimeRAW – Time used (Elapsed time * CPU count) by a job or step in CPU-seconds.
4. ElapsedRaw – Job's elapsed time in seconds.
5. AllocNodes – Number of nodes allocated to the job/step. 0 if the job is pending.
6. AllocCPUS – Count of allocated CPUs.
7. TimelimitRaw – What the time limit was/is for the job. Format is in number of minutes.
8. Priority – Slurm priority.
9. Partition – Partition on which the job ran (the name of the cluster, e.g., Tornado).

This data combines user requests for computing resources with detailed performance metrics in the form of time series, including key parameters such as the requested and actual task execution time, the amount of processor time used, the degree of launch success, and other critical performance indicators.

Fig. 1 provides a visual representation of the multidimensional nature of these time series for one particular user. On the graph, each point represents a snapshot of the system state for one of the tasks – thus, a sequence of 15 such points reflects the results of the analysis for 15 sequentially completed tasks.

For each user, data is extracted and examined individually, taking into account all characteristics as part of a single multidimensional time series, using an approach where the transformation is carried out using a Takens embedding based on an algorithm for selecting parameters based on L-statistics. Due to the complexity of time series that require embedding in a space with a dimension of at least $2n + 1$, a direct visual representation of this point cloud becomes unrealizable at dimensions $d > 3$.

However, using the theory of persistent homology, we transform these point clouds into analytically controlled information. We create filtered Vietoris–Rips complexes that allow to extract persistent homology and construct average Betti curves for each user. This process forms a unique "topological portrait" of the user, which reflects his characteristics and behavior within the framework of computational tasks [16].

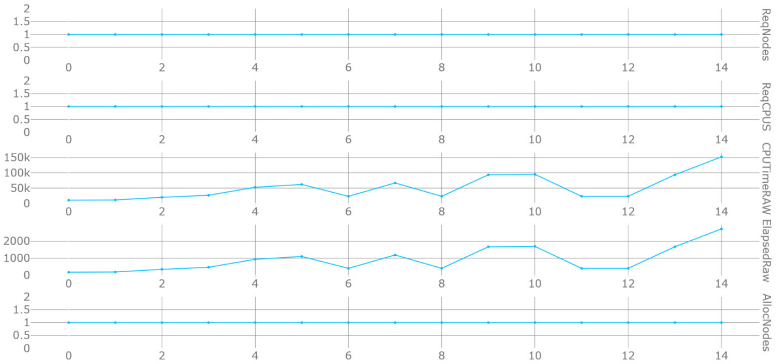

Fig. 1. Time series

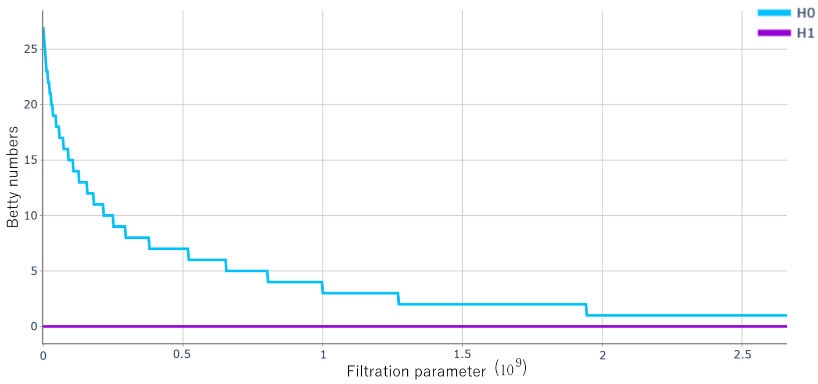

Fig. 2. Betti curves of the user profile

Fig. 2 illustrates an example of such a user profile in the form of Betti curves, where each step on the curve reflects changes in the topological structure of user data, taking into account different levels of detail and time scales.

To create a reference profile of the "ideal" user, we used a time series of the most successful users selected according to two criteria: more than 95% successful completion of tasks and a minimum deviation in the use of resources from the stated needs. The results of this analysis are presented in Fig. 3, showing Betti curves reflecting optimal topological behavior characteristics.

Table 2 shows the results of evaluating the comparison of several user profiles with a reference profile.

Thus, the proposed methodology allows to compare current user behavior with historical data as well as to provide a quantitative assessment of their behavioral effectiveness, the basis for determining the level of trust in each user.

However, it is worth noting the limitation of this approach: for new users, it is required to collect a sufficient amount of initial data to accurately build their profile and reliable assessment. A lack of initial data can make it difficult to accurately model behavior and leads to inaccuracies in the assessment.

Conclusion

Within the framework of this study, an algorithm based on topological data analysis is proposed, which builds user profiles of a supercomputer center with acceptable computational complexity using simple and effective procedures for identifying behavioral patterns. This approach demonstrates a wide

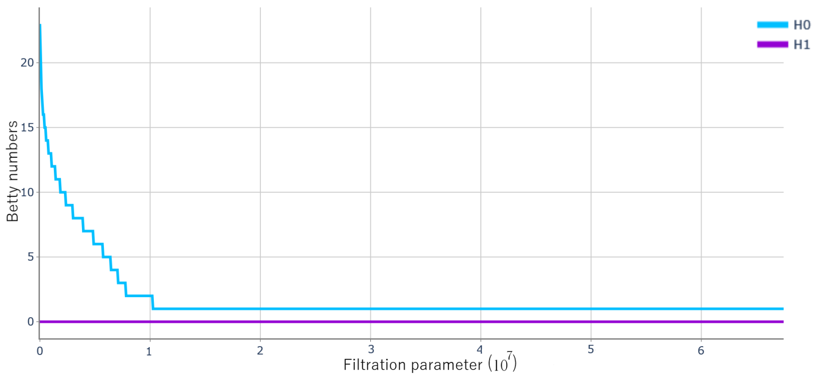

Fig. 3. Betti curves of the reference profile

Table 2

User profile ratings

User	Desirability assessment		User performance	
	Numerical evaluation	**Linguistic value**	**Percentage of successfully completed tasks**	**Estimation error (average absolute logarithmic error)**
0	0.115	Very bad	0.369	3.875
1	0.824	Very good	0.698	2.838
2	0.339	Bad	0.33	2.556
3	0.085	Very bad	0.001	4.261
4	0.787	Good	0.611	3.06

potential in the field of detection and evaluation of deviations. Experimental verification of the method based on real data from the SCC Polytechnic confirmed the high applicability of the methodology in the context of the development of intelligent task allocation management systems.

This approach helps to increase the efficiency of using the resources of supercomputer platforms, ensuring optimal allocation of tasks and reducing unnecessary costs, ensuring fairness and transparency of access to resources for all users, based on their behavioral characteristics. It also potentially provides tools for predicting system load and strengthens the security of platforms, allowing timely detection of abnormal activity and prevention of possible threats.

The development of adaptive dispatch systems that would take into account the dynamics of user behavior and quickly respond to changes in their requests and needs, and the integration of the developed technology with existing supercomputer infrastructure management systems, which will enhance their functionality and increase the efficiency of resource management, are promising for the development of this method.

REFERENCES

1. **Hou Z., Shen H., Zhou X. et al.** Prediction of job characteristics for intelligent resource allocation in HPC systems: a survey and future directions. Frontiers of Computer Science, 2022, Vol. 16, No. 5, Article no. 165107. DOI: 10.1007/s11704-022-0625-8

2. **Uchroński M., Bożejko W., Krajewski Z., Tykierko M., Wodecki M.** User estimates inaccuracy study in HPC scheduler. Contemporary Complex Systems and Their Dependability: Proceedings of the 13th

International Conference on Dependability and Complex Systems (DepCoS-RELCOMEX), 2019, Pp. 504–514. DOI: 10.1007/978-3-319-91446-6_47

3. **Zaborovsky V.S., Utkin L.V., Muliukha V.A., Lukashin A.A.** Improving Efficiency of Hybrid HPC Systems Using a Multi-agent Scheduler and Machine Learning Methods. Supercomputing Frontiers and Innovations, 2023, Vol. 10, No. 2, Pp. 104–126. DOI: 10.14529/jsfi230207

4. **Martín A.G., Beltrán M., Fernández-Isabel A., Martín de Diego I.** An approach to detect user behavior anomalies within identity federations. Computers & Security, 2021, Vol. 108, Article no. 102356. DOI: 10.1016/j.cose.2021.102356

5. **Martín A.G., Fernández-Isabel A., Martín de Diego I. et al.** A survey for user behavior analysis based on machine learning techniques: current models and applications. Applied Intelligence, 2021, Vol. 51, No. 8, Pp. 6029–6055. DOI: 10.1007/s10489-020-02160-x

6. **Chazal F., Michel B.** An introduction to topological data analysis: Fundamental and practical aspects for data scientists. Frontiers in Artificial Intelligence, 2021, Vol. 4, Article no. 667963. DOI: 10.3389/frai.2021.667963

7. **Yan H., Yang N., Peng Y., Ren Y.** Data mining in the construction industry: Present status, opportunities, and future trends. Automation in Construction, 2020, Vol. 119, Article no. 103331. DOI: 10.1016/j.autcon.2020.103331

8. **Seversky L.M., Davis S., Berger M.** On time-series topological data analysis: New data and opportunities. 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2016, Pp. 1014–1022. DOI: 10.1109/CVPRW.2016.131

9. **Choi K., Yi J., Park C., Yoon S.** Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. IEEE Access, 2021, Vol. 9, Pp. 120043–120065. DOI: 10.1109/ACCESS.2021.3107975

10. **Taieb S.B., Bontempi G., Atiya A.F., Sorjamaa A.** A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. Expert Systems with Applications, 2012, Vol. 39, No. 8, Pp. 7067–7083. DOI: 10.1016/j.eswa.2012.01.039

11. **Tan E., Algar S., Corrêa D., Small M., Stemler T., Walker D.** Selecting embedding delays: An overview of embedding techniques and a new method using persistent homology Chaos, 2023, Vol. 33, No. 3, Article no. 032101. DOI: 10.1063/5.0137223

12. **Uzal L.C., Grinblat G.L., Verdes P.F.** Optimal reconstruction of dynamical systems: A noise amplification approach. Physical Review E, 2011, Vol. 84, No. 1, Article no. 016223. DOI: 10.1103/PhysRevE.84.016223

13. **Leykam D., Angelakis D.G.** Topological data analysis and machine learning. Advances in Physics: X, 2023, Vol. 8, No. 1, Article no. 2202331. DOI: 10.1080/23746149.2023.2202331

14. **Mezheneva I.O., Lukashin A.A.** Topological descriptors for data analysis in behavioral analytics systems. Matematicheskie metody v tekhnike i tekhnologiiakh [Mathematical methods in technology and engineering], 2023, Vol. 10, Pp. 104–108. DOI: 10.52348/2712-8873_MMTT_2023_10_104

15. **Thorisingam Y., Mustafa M.S.** Augmented approach to desirability function based on principal component analysis. International Journal of Academic Research in Progressive Education and Development, 2022, Vol. 11, No. 2. DOI: 10.6007/IJARPED/v11-i2/13270

16. **Nashivochnikov N.V., Pustarnakov V.F.** Topological methods of analysis in behavioral analytics systems. Voprosy kiberbezopasnosti, 2021, Vol. 2, No. 42, Pp. 26–36. DOI: 10.21681/2311-3456-2021-2-26-36

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Mezheneva Irina O.
Меженева Ирина Олеговна
E-mail: Mezheneva-I@gaz-is.ru

Lukashin Alexey A.
Лукашин Алексей Андреевич
E-mail: lukash.spb.ru@gmail.com

Chatoyan Sergey K.
Чатоян Сергей Камалович
E-mail: Chatoyan-S@gaz-is.ru

Submitted: 05.07.2024; Approved: 17.09.2024; Accepted: 04.10.2024.
Поступила: 05.07.2024; Одобрена: 17.09.2024; Принята: 04.10.2024.