Spectral Theory for Multidimensional Digital Data Matrices' Processing

Deykin I. I. [0000-0002-8151-6337], Syuzev V. V. [0000-0002-8689-8282], Smirnova E. V. [0000-0003-2275-3276]

Bauman Moscow State Technical University, Moscow, Russia
deykinii@student.bmstu.ru

Abstract. The article presents experimental results pertaining to an analysis of multidimensional digital data that describe applications to Russian universities within the constraints of spectral theory. Two-dimensional matrices extracted from the data contain average unified exam scores among first year students who enrolled under government-paid tuition to 9 engineering programs at 47 universities in 2020 and 2021. The matrices are processed as two-dimensional signals from which a spectral characteristic is formed which is then used to simulate new signals in Fourier and Hartley harmonic bases. The simulated signals are consequently presented and explored as predictive modeling data. The article presents preliminary experimental results confirming prospects of using the proposed technique and suggests a trajectory for future inquiry. The work is supported by the Russian Federation Ministry of Science and Higher Education (projects #FSFN-2023-0006 and "Priority 2030").

Keywords: Spectral theory, signal simulation, Fourier basis, Hartley basis, energy spectral density.

1 Introduction

The harmonic simulation within the framework of spectral theory allows simulating random signals for a given energy spectral density function. The resulting sets of random signals correspond to a given energy spectral density function. Harmonic simulation compares favorably to simulating with random fields or shaping filters with regards to theoretical and executional simplicity and computational complexity [1]. The possibility of using different harmonic bases allows customizing for specific conditions pertaining to a particular task or a particular area of expertise [2, 3].

Power and energy spectral densities can be estimated from various data arrays using the periodogram method or the Welch method [4]. All data are signals on a physical level. Simulating based on those densities reproduces statistical dependencies found in real systems and events and may be utilized for equipment preparations, personnel training, or forecasting [5]. The common scheme of the proposed modeling technique is shown in Figure 1.

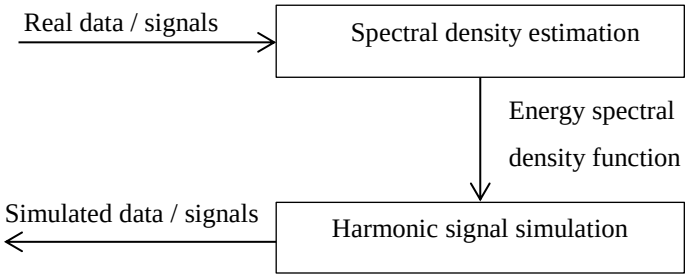


Fig. 1. Scheme of the process of simulating signals according to the spectral characteristic obtained on the basis of real-world data

This scheme is often used in radio electronics, but it is promising to study its applicability towards two-dimensional data arrays. Data arrays are more often studied with the help of artificial intelligence which often lacks in transparency compared to signal processing methods. This article considers 2D arrays of data on admission of applicants to Russian universities, the data is used to obtain a 2D energy spectral density function, from which new 2D signals or data sets are then reproduced. The inquiry concerns whether the signal simulation method can be used to simulate arbitrary data and therefore compete with led transparent artificial intelligence methods.

2 Data preprocessing

The raw data was collected by the Russian university Higher School of Economics (HSE) within their "Monitoring The Quality of Enrolment in Russian Universities" initiative, it contains average results of the unified state exam among applicants admitted to various programs at various universities for a government-paid education from 2011 to 2022 [6]. Applicants are admitted to universities based on their unified exam scores and internal exam scores. The programs within universities have a limited number of applicants they may accept for a government-paid tuition – those are given to the highest ranking applicants. Student-paid education is more flexible in terms of required academic scores and does not contain quite the same information about the quality of a university. Average scores among the applicants are helpful in determining which programs at which universities are the most popular among the highest scoring applicants. Assuming that the unified exam scores are adequate at indicating academic capabilities of applicants the described data might be used to develop tools that support the managerial decisions in the education sphere [7]. It was decided to limit the study to a set of 9 engineering areas, for each area in each year it was decided to consider only the top 10 universities with the highest average scores for the unified state exam among applicants for 2020 and 2021. As a result 47 universities were selected. The full list of the universities and programs can be found at the HSE web source [6].

Values in the tables extracted from the data are the average unified exam grades among the students accepted into the university represented in a column to study in the program represented by a row. Below is the Table 1 for year 2021. The table was abridged for the article, the full table containing 9 rows and 47 columns describing 9 programs among 47 universities may be accessed at GitHub [8]. The tables for different years were combined, the names of the departments were marked with the year, then the data was again separated by years - this was necessary in order to guarantee the same arrangement of universities in the columns in the tables for different years. The universities displayed are: Baltic State Technical University (BSTU) [9], Ufa State Petroleum Technological University (USPTU) [10], Ural Federal University (UrFU) [11], Voronezh State University (VSU) [12].

Table 1. The average results of the unified state exam among applicants admitted to 9 programs (rows) at 47 universities (columns) for a government-paid tuition

Programs	BSTU	USPTU	…	UrFU	VSU
Automation and control	0	0	…	0	0
Aviation and space instrumentation	73.2	0	…	0	0
Business Informatics	0	0	…	89.9	0
…	…	…	…	…	…
Information Security	0	0	…	0	0
Nuclear physics and technology	0	0	…	0	79.1
Technological machines	0	76.1	…	0	0

The Table 1 is the result of data pre-processing, the table represent a two-dimensional matrix or a two-dimensional signal suitable for processing within the framework of spectral theory. The full table for 2020 may be found at GitHub [8]. Figure 2 shows graphical representations for the 2020 and 2021 data matrices. The matrices are depicted as signals $x(i_1, i_2)$, i_1 represents universities, i_2 represents programs.

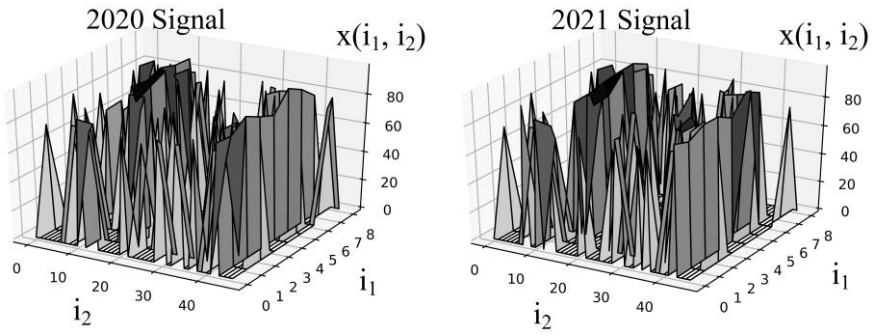


Fig. 2. Graphs of average unified exam scores among the applicants accepted to 9 general engineering programs at 47 Russian universities for 2020 and 2021

3 Signal simulation

3.1 Spectral density estimation

An important signal processing tool is the Fourier transform, which allows you to go from the time or space domain to the frequency domain. Figure 3 shows the Fast Fourier transforms (FFT) $X(\omega_1, \omega_2)$ of two-dimensional data matrices for 2020 and 2021, where ω_1 and ω_2 are frequencies along axis i_1, i_2 correspondingly.

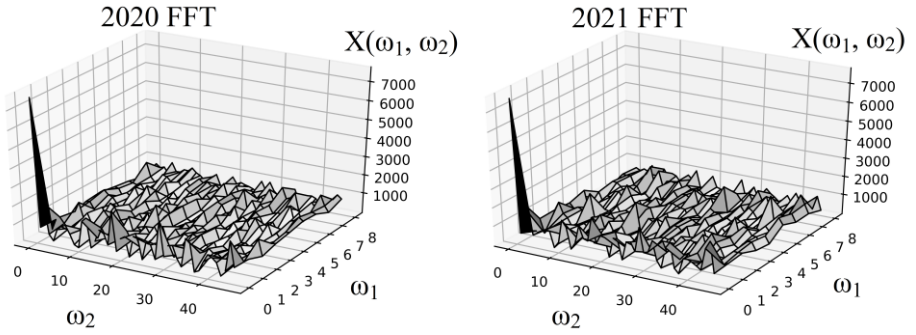


Fig. 3. Fourier transforms of admission data in 9 engineering areas at 47 Russian universities for 2020 and 2021

The calculation of the Fourier transforms is necessary when evaluating the signal's spectral density. Energy spectral density of a single signal equals its squared absolute Fourier transform values:

$$S_{Ei}(\omega_1, \omega_2) = |X_i(\omega_1, \omega_2)|^2 = |\mathcal{F}(x_i(i_1, i_2))|^2,$$

where the S_{Ei} is the energy spectral density function of an individual signal $x_i(i_1, i_2)$. $\mathcal{F}$ is a Fourier transform. $x_i(\tau_1, \tau_2)$ is a spatial representation of a signal, and $X_i(\omega_1, \omega_2)$ is a frequency representation that can be found as a Fourier transform of $x_i(i_1, i_2)$. Individual signals are considered as discrete finitely windowed samples of the same signal [2, 13]. Energy spectral densities or Fourier transforms of individual sections of the data are averaged to obtain collective general energy spectral density function.

$$S_E(\omega_1, \omega_2) = \frac{1}{N}\sum_{i=1}^{N}|X_i(\omega_1, \omega_2)|^2 = \frac{1}{N}\sum_{i=1}^{N}|\mathcal{F}(x_i(i_1, i_2))|^2.$$

Figure 4 shows a two-dimensional spectral density function obtained by this method from the input data for 2020 and 2021. In this paper energy spectral density is justified since the discrete time Fourier transforms exist for the chosen signals, however, the future inquiry might benefit from using power spectral density to cover the case of a so called power signal for which such a transform is meaningless [9, 14].

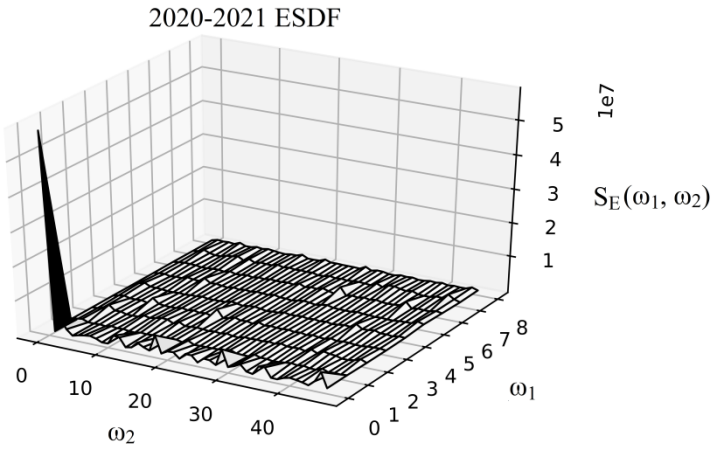


Fig. 4. Energy spectral density function for 2020 and 2021

3.2 Fourier signal simulation

The resulting spectral density function can be used to simulate signals in various harmonic bases within the framework of spectral theory. One of such bases is the complex-exponential Fourier basis, where the real part of the functions is taken equal to the cosine function

$$Re\left\{exp\left[2\pi\left(\frac{k_1 t_1}{T_1}+\frac{k_2 t_2}{T_2}\right)\right]\right\} = cos\left[2\pi\left(\frac{k_1 t_1}{T_1}+\frac{k_2 t_2}{T_2}\right)\right],$$

and the imaginary part is equal to the sine function [2, 15]:

$$Im\left\{exp\left[2\pi\left(\frac{k_1 t_1}{T_1}+\frac{k_2 t_2}{T_2}\right)\right]\right\} = sin\left[2\pi\left(\frac{k_1 t_1}{T_1}+\frac{k_2 t_2}{T_2}\right)\right].$$

The energy spectral density function S_E is used to obtain the Fourier coefficients X_F:

$$X_{\mathrm{F}}(k_1,k_2) = \sqrt{\frac{S_E\left(\frac{2\pi}{T_1}k_1,\frac{2\pi}{T_1}k_2\right)}{T_1^2T_2^2\left(1+\lambda_{k_1,k_2}^2\right)}} = \frac{1}{T_1T_2}\sqrt{\frac{S_E\left(\frac{2\pi}{T_1}k_1,\frac{2\pi}{T_1}k_2\right)}{1+\lambda_{k_1,k_2}^2}},$$

$$k_1,k_2 = 0,1,\ldots,$$

where T_1, T_2 are periods of the two-dimensional signal; λ_{k_1,k_2} is a tangent function of a 2D phase density which is needed since the energy spectral density only describes amplitudes but not phases and therefore is calculated as an average phase spectrum of the original signals.

$$\lambda_{k_1,k_2} = \frac{1}{N}\sum_{i=1}^{N}\left(Im\left[\mathcal{F}\left(x_i(i_1,i_2)\right)\right]/Re\left[\mathcal{F}\left(x_i(i_1,i_2)\right)\right]\right).$$

Fourier coefficients are used to obtain the simulated signal:

$$x(i_1, i_2) = \sum_{k_1=0}^{\frac{N_1-1}{2}} \sum_{k_2=0}^{\frac{N_2}{2}} X_F(k_1, k_2)\, exp\left[j2\pi\left(\frac{k_1 i_1}{T_1} + \frac{k_2 i_2}{T_2}\right)\right], i_1 \in [0, N_1), i_2 \in [0, N_2).$$

The deterministic signal received by running such a calculation without adding random coefficients is presented on Figure 5. The resulting signal has a complex nature, the real components were depicted both for deterministic and random signal simulation in Fourier basis.

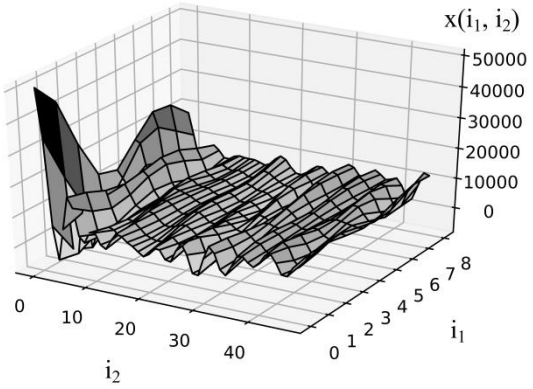


Fig. 5. Deterministic signal obtained by simulating in the Fourier basis for the energy spectral density function based on the real-world data

Random coefficients might be added. Random coefficients may be random signs thus taking on values of "-1" or "+1" randomly or random values in the interval [-1, +1] [1]. A set of random signals generated by adding random coefficients to the formula is presented on Figure 6.

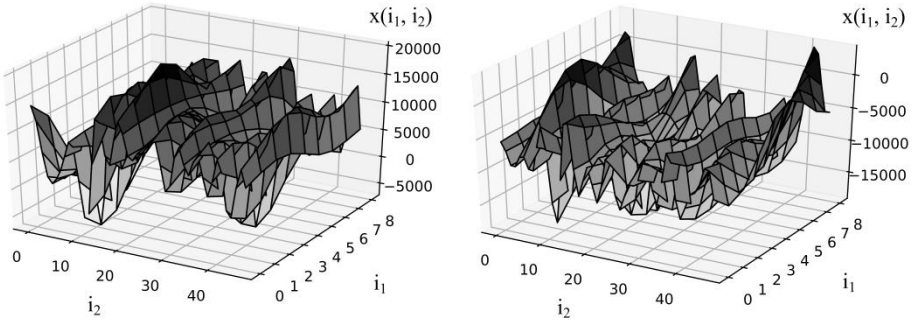


Fig. 6. Random signals obtained by simulating in the Fourier basis for the energy spectral density function based on the real-world data

The leftmost signal on the figure 6 was picked for post processing and comparison. It's range was normalized and then fitted to the average of the two original signals. The values were moved up so that the minimum is at 0 and then divided by the new maximum, multiplied by the average maximum 98.5 found from the original data. The values below the average original minimum 71.35 were dropped. Simulated data table after the post-processing is shown on the Table 2.

Table 2. Data obtained by simulating in the Fourier basis for the energy spectral density function based on the real-world data

Programs	BSTU	USPTU	…	UrFU	VSU
Automation and control	78.397	0	…	0	0
Aviation and space instrumentation	0	0	…	0	0
Business Informatics	0	0	…	78.262	0
…	…	…	…	…	…
Information Security	0	0	…	0	0
Nuclear physics and technology	0	0	…	0	0
Technological machines	0	0	…	72.767	0

The Table 2 does not allow viewing the entire array of data, so figure 7 shows the original data for 2021 (7a) and the simulated data after post-processing (7b).

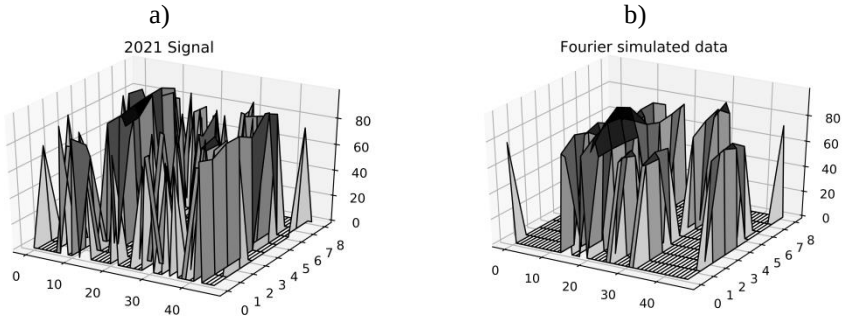


Fig. 7. Graph of the original data matrix for 2021 (a) and a graph of the Fourier simulated and adjusted data matrix (b)

The data being spread across i_1 axis is consistent with the original data. However, the simulation "created" new programs and "closed" the old ones. Such variations are contingent on the random coefficients generated for each experiment. The average difference of values in the simulated data and averaged 2020 and 2021 data is 6.7276. The average difference across 500 experiments with different random coefficients and therefore different simulated data with the same post processing is 7.9279.

3.3 Hartley signal simulation

Another harmonic basis is a Hartley basis that uses a special cas function [15]:

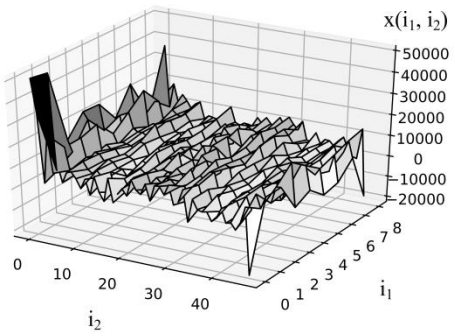
$$\mathrm{cas}(\omega_1 t_1 + \omega_2 t_2) = \cos\left[2\pi\left(\frac{k_1 * t_1}{T_1} + \frac{k_2 * t_2}{T_2}\right)\right] + \sin\left[2\pi\left(\frac{k_1 * t_1}{T_1} + \frac{k_2 * t_2}{T_2}\right)\right].$$

Fourier coefficients are used to obtain a simulated signal:

$$x(i_1, i_2) = \sum_{k_1=0}^{\frac{N_1-1}{2}} \sum_{k_2=0}^{\frac{N_2}{2}} X_F(k_1, k_2)\, cas\left[2\pi\left(\frac{k_1 i_1}{T_1} + \frac{k_2 i_2}{T_2}\right)\right], i_1 \in [0, N_1), i_2 \in [0, N_2).$$

Simulation according to the given rules led to the generation of various signals. Figure 8a shows a deterministic signal, Figure 8b shows a series of random signals.

a)



b)

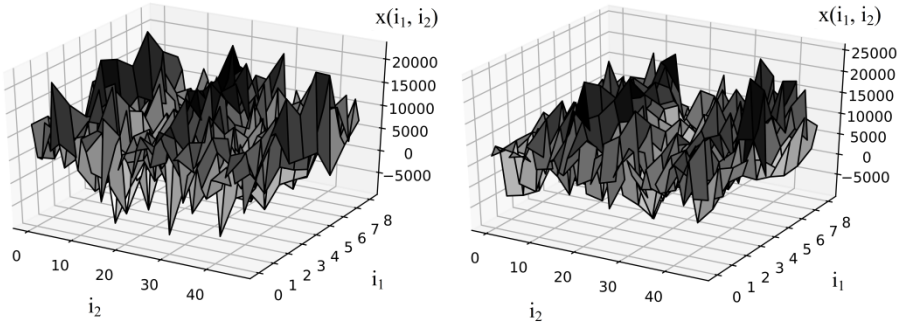


Fig. 8. Deterministic signal (a) and random signals (b) obtained by simulating in the Fourier basis for the above spectral density

For Hartley simulated data post-processing is the same. As a result of such post-processing a simulation table was obtained that is shown on Table 3.

Table 3. Data obtained by simulating in the Hartley basis for the energy spectral density function based on the real-world data

Programs	BSTU	USPTU	…	UrFU	VSU
Automation and control	82.785	0	…	79.721	0
Aviation and space instrumentation	0	0	…	0	0
Business Informatics	0	72.259	…	0	0
…	…	…	…	…	…
Information Security	77.039	0	…	73.975	0
Nuclear physics and technology	0	0	…	0	0
Technological machines	0	72.989	…	0	0

Let's compare graphical representations of the data. Figure 8 shows the original data for 2021 (8a) and the simulated data after post-processing (8b).

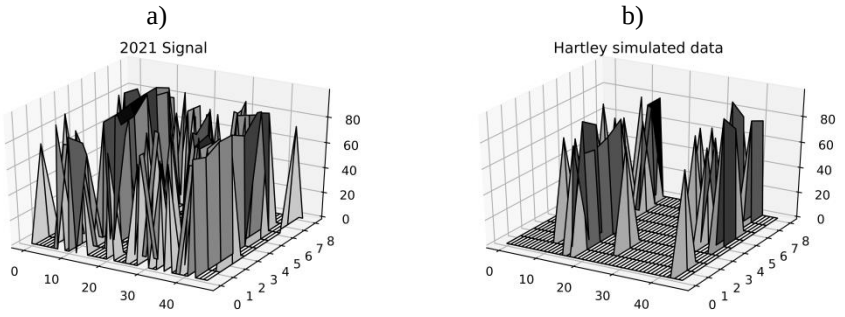


Fig. 8. Graph of the original data matrix for 2021 (a) and a graph of the Hartley simulated and adjusted data matrix (b)

The discrepancy mentioned in the case of the Fourier simulation remains for the Hartley simulation too. The amount of zeroes increased noticeably. Directionality of the data is preserved. The average difference after post processing is 11.6539. The difference across 500 experiments with different random coefficients is 9.4748.

4 Conclusion

The proposed technique includes extracting energy spectral density function and phase density from data or signals. Simulation yields signals or data where some statistical characteristics of the original are repeated, namely, most non-zero values are spanned across same axis both in the original and in the simulation. That and arguable arbitrariness to the post processing used should be discussed with experts in the field. The average differences between the simulated and original data are quite significant and are around 8.834% of the maximal value in the original data. Very small piece of

the original data was used and further progress demands expanding it. Statistical characteristics of generated signals may fit different fields or tasks better therefore other two-dimensional data must be inspected including images and graphs which can too be presented as two-dimensional matrices [16, 17].

5 Acknowledgments

The work has been financially supported by the Russian Federation Ministry of Science and Higher Education. Applying simulation methods towards different 2D data was a part the State assignment on the topic " Exploratory research in the field of creating algorithmic, software and hardware for high-performance hybrid intelligent systems for multimodal merging and analytical processing of heterogeneous data on geographically distributed industrial infrastructure facilities" (project #FSFN-2023-0006). The particular data, tasks concerning it, and an adaptation of a simulation method are all parts of the project "Priority 2030".

References

1. Liu, Yang & Li, Jingfa & Sun, Shuyu & Yu, Bo. (2019). Advances in Gaussian random field generation: A review. Computational Geosciences. 10.1007/s10596-019-09867-y.
2. Smirnova E., Syuzev V., Samarev R., Deykin I., Proletarsky A. High-Dimensional Simulation Processes in New Energy Theory: Experimental Research (Extended Abstract). Data Analytics and Management in Data Intensive Domains: XXII International Conference DAMDID/RCDL' 2020 (October 13–16, 2020, Voronezh, Russia): Extended Abstracts of the Conference. Edited by Bernhard Thalheim, Sergey Makhortov, Alexander Sychev. – Voronezh: Voronezh State University, 2020. – 246 p.
3. Shakhnov, V.A., Kurnosenko, A.E., Demin, A.A., Vlasov, A.I. (2020). Industry 4.0 Visual Tools for Digital Twin System Design. In: Silhavy, R., Silhavy, P., Prokopova, Z. (eds) Software Engineering Perspectives in Intelligent Systems. CoMeSySo 2020. Advances in Intelligent Systems and Computing, vol 1295. Springer, Cham. https://doi.org/10.1007/978-3-030-63319-6_80
4. Shumway R. H., Stoffer D. S. Time series analysis and its applications: with R examples. Springer International Publishing. 2017. – 567 p.
5. Andreev, A., Berezkin, D., Kozlov, I. (2018). Approach to Forecasting the Development of Situations Based on Event Detection in Heterogeneous Data Streams. In: Kalinichenko, L., Manolopoulos, Y., Malkov, O., Skvortsov, N., Stupnikov, S., Sukhomlin, V. (eds) Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2017. Communications in Computer and Information Science, vol 822. Springer, Cham. https://doi.org/10.1007/978-3-319-96553-6_16
6. Higher School of Economics (HSE): Monitoring The Quality of Enrolment in Russian Universities. URL: https://ege.hse.ru/ (last accessed on 28.05.2020)
7. Ismagilov, K., Vlasov, A., Karpunin, A., Kurnosenko, A., Strukova, A. (2023). Open Engineering Education Tools in the Context of Education 4.0. In: Kumar, V., Kyriakopoulos, G.L., Akberdina, V., Kuzmin, E. (eds) Digital Transformation in Industry . DTI 2022. Lecture Notes in Information Systems and Organisation, vol 61. Springer, Cham. https://doi.org/10.1007/978-3-031-30351-7_33

8. Full tables for the paper on GitHub. URL: https://github.com/vandeyk/UniApplicationsSimulation (last accessed on 30.05.2023)
9. Baltic State Technical University's official web site. URL: https://www.voenmeh.ru/ (last accessed on 31.05.2023)
10. Ufa State Petroleum Technological University's official web site. URL: https://study.rusoil.net/ (last accessed on 31.05.2023)
11. Ural Federal University's official web site. URL: https://urfu.ru/en/ (last accessed on 31.05.2023)
12. Voronezh State University's official web site. URL: https://www.vsu.ru/english/ (last accessed on 31.05.2023)
13. Alessio S.M. Digital signal processing and spectral analysis for scientists. Concepts and applications – Cham: Springer International Publishing Switzerland, 2016. – 909 p.
14. Marwan, Norbert & Braun, Tobias. (2023). Power spectral estimate for discrete data. Chaos (Woodbury, N.Y.). 33. 10.1063/5.0143224.
15. Tuan, Trinh. (2022). Operational Properties of the Hartley Convolution and Its Applications. Mediterranean Journal of Mathematics. 19. 10.1007/s00009-022-02173-5.
16. Rajinikanth, V., Priya, E., Lin, H., & Lin, F. (2021). Hybrid Image Processing Methods for Medical Image Examination (1st ed.). CRC Press. https://doi.org/10.1201/9781003082224
17. Dong, Xiaowen & Thanou, Dorina & Rabbat, Michael & Frossard, Pascal. (2019). Learning Graphs From Data: A Signal Representation Perspective. IEEE Signal Processing Magazine. 36. 44-63. 10.1109/MSP.2018.2887284.