UDC 621.39

Numerical Aspects of Statistical Pattern Recognition

Yurii N. Orlov [*][†], **Elvira R. Zaripova**[*], **Alexey V. Chukarin**[*]

[*] *Department of Applied Probability and Informatics,*
Peoples' Friendship University of Russia (RUDN University),
6 Miklukho-Maklaya st., Moscow, 117198, Russian Federation
[†] *Department of Kinetic Equations,*
Keldysh Institute of Applied Mathematics of Russian Academy of Sciences,
4 Miusskaya Sq., Moscow, 125047, Russian Federation

Email: yuno@kiam.ru,
zaripova_er@rudn.university, chukarin_av@rudn.university

The paper presents numerical restrictions of Bayesian method to a pattern recognition. The maximum probability of local state correspondence to one of the basis patterns is defined through the expansion of examined vector over the patterns. The practical recognition example without errors is presented in the frame of nearest neighbor method for the case, when the probability interpretation of the expansion coefficients is not valid. The spectral portraits of solving matrices are constructed for the literature texts author identification problem. Also, the statistical properties of letters frequencies in European literature texts are investigated. The determination of logarithmic dependence of letters sequence for one-language and two language texts are examined. The Voynich Manuscript structure was considered for numerical analysis. After our numerical analysis, we suppose, that the Voynich Manuscript was written in two languages having the same alphabet without vowel letters: one of the Germanic languages (Danish or German) and one of the Romance languages (Latin or Spanish).

Key words and phrases: non-stationary time series, probabilities of letters combination, basis patterns, pattern recognition, European language statistics, Voynich Manuscript.

In: K. E. Samouylov, L. A. Sevastianov, D. S. Kulyabov (eds.): Selected Papers of the 1st Workshop (Summer Session) in the framework of the Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems", Tampere, Finland, 20–23 August, 2018, published at http://ceur-ws.org

1. Introduction

Theory of pattern recognition concerns to various aspects of statistical analysis of big data. The methods and principles of the theory are presented in many special books, see [1–3]. The basic principles of characters recognition were investigated in [4]. Some algorithm aspects were discussed in [5–7].

Let us consider statistical recognition problem of belonging of the sample distribution function (SDF) to a certain class or a cluster. We suppose, that each class is characterized by a certain pattern or etalon distribution function. In practice very often there are situations, when patterns are given as frequencies of empirical estimated parameters. The corresponding functions $\{\varphi_1, ..., \varphi_m\}$, $\varphi_s \in R^n$, where m is a number of states (or basis patterns) and n is a dimension of parameter space are treated to be etalon probability distributions. If, as a result of observation, it was found, that the system is characterized by empirical discrete probability distribution (DPD) $f \in R^n$, $f \equiv (f_1, ..., f_n)$, the state, corresponding to this DPD, is recognized according to Bayesian method [3]. In mathematical sense the recognition problem is solved by the expansion of vector f over the basis $\{\varphi_1, ..., \varphi_m\}$. The problem is that the projection operator on the non-orthogonal (in general case) basis vector system may cause an unacceptably large error.

In this paper the concrete example of very exact recognition will be considered for the non-traditional case, when the scalar products between basis vectors φ_s are close to unit, but nevertheless the recognition is appeared to be possible as a consequence from Bayesian method.

Let us consider a vector $f(k)$, where k is a number of letter in alphabetical ordering, corresponding to empirical frequency of the letter k in a given text. For texts in Russian language $k = 1, 2, ..., 33$. For texts in English language $k = 1, 2, ..., 26$ etc. It appears, that literature texts by the same author have similar functions of letter frequencies distribution [8] and these functions are stable for each author. The error of recognition by the method of nearest neighbors for 100 authors and 1000 texts is equal to 0,15. If we construct the distribution of bigrams of the texts by the same author, the unknown text can be recognized in the library of author etalons with the accuracy, equal to 0.04. And if we consider a three letters composition distribution, the empirical recognition is absolutely accurate. The dimension of basis vectors in the last case is equal to $33^3 = 35937$ (for Russian language). If we have a library of ten (or even 1000) authors, then the technical problem is to project a vector with large dimension (number of letters compositions) in a space with small dimension (number of authors). According to Bayesian method this operation is numerically incorrect, but the method of nearest neighbors remains valid.

Theoretical aspects and practical examples of the method of nearest neighbors are discussed below.

2. Bayesian Recognition

Let vectors $\{\varphi_1, ..., \varphi_m\}$, $\varphi_s \in R^n$ are histograms, which correspond to definite author letter distribution etalons. Then

$$\forall s \in \{1, ..., m\} \quad \sum_{k=1}^{n} \varphi_s(k) = 1. \tag{1}$$

Let also vector $f \in R^n$ belongs to the convex hull of vectors $\{\varphi_1, ..., \varphi_m\}$, so that

$$f(k) = \sum_{s=1}^{m} y_s \varphi_s(k), \quad 0 \le y_s \le 1. \tag{2}$$

From (1) and (2) it follows, that $\sum_{s=1}^{m} y_s = 1$.

If the expansion (2) is obtained, then we can define a number s^*, so that

$$s^* = \arg\max y_s. \tag{3}$$

Then the most probable solution is that $f = \varphi_{s^*}$. In this case the following consequence is valid:

$$s^* = \arg\min \|f - \varphi_s\|. \tag{4}$$

In general the projection problem is solved by QR-expansion method [9, 10], where optimal in a sense of 2-norm expansion is given by formula

$$\begin{gathered} f - \Phi y = If - QRy = \left(I - QQ^T + QQ^T\right) f - QRy = \\ = Q\left(Q^T f - Ry\right) + \left(I - QQ^T\right) f \end{gathered} \tag{5}$$

Here $\Phi_{n\times m}$ is a matrix of basis vectors $\{\varphi_1, ..., \varphi_m\}$, $Q_{n\times m}$ is a matrix, for which $Q^T Q = I_{m\times m}$, and $R_{m\times m}$ is an upper triangular matrix. Since the last two terms in the expression (5) are orthogonal, the optimal expansion is given by a formula

$$y_{opt} = R^{-1} Q^T f. \tag{6}$$

The relative error of expansion (2), (6) is defined as (7):

$$\varepsilon = \frac{\delta}{\|f\|} = \frac{\left\|\left(I - QQ^T\right) f\right\|}{\|f\|}. \tag{7}$$

If we have a statistical inaccuracy in numerical estimations of basis patterns and vector *f*, the error, defined by (7), can be bounded as (8) from [10]:

$$\frac{\|\Delta y\|}{\|y\|} \le \xi \cdot \left(\frac{2\kappa(\Phi)}{\cos\theta} + \kappa^2(\Phi) tg\theta\right) + O\left(\xi^2\right). \tag{8}$$

In (8) was used $\xi = \max\left(\frac{\|\Delta\Phi\|}{\|\Phi\|}, \frac{\|\Delta f\|}{\|f\|}\right)$, and $\kappa(\Phi)$ - is a condition number of the matrix Φ and $\sin\theta = \varepsilon$.

So, in practice the condition of non-negativity of coefficients y_s and a probability interpretation of the expansion according to the condition $\sum_{s=1}^{m} y_s = 1$ may be violated. In the paper [5] it was exhibited, that for $m = 4$ authors the violation of probability interpretation of Bayesian recognition in the form (3) is equal to 0.55 for $n = 33$ letters, 0.35 for $n = 33^2$ letters combinations and 0.30 for $n = 33^3$ letters combinations.

The spectral portrait [9] of matrix Φ for 4 author etalons is presented in Fig. 1. The color regions correspond to the same accuracies of matrix element in terms of decimal power in the legend. Contours correspond to the eigenvalues regions.

The condition number of this matrix Φ is equal to 3500.

It should be emphasized, that the recognition according the method [4] for the same statistical experiment lead to the accuracy of recognition 0.15, 0.04 and 0.00 respectively. So in practice we should use the most effective method. The practical example is given below.

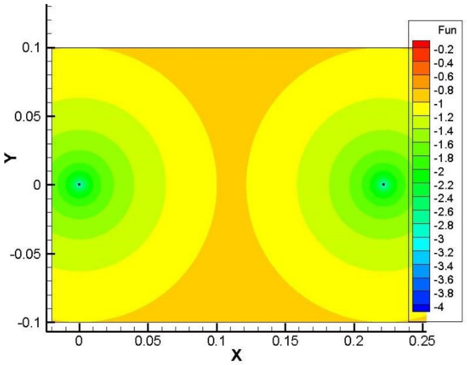

Figure 1. Spectral portrait of matrix of author patterns

3. Statistical Properties of European Languages and Voynich Manuscript

It is possible to construct the etalons not only for authors, but also for various languages. In the most languages the dependence of ordered frequencies is logarithmic [11], its determination is more than 0.98. Parameters of the logarithmic dependence are determined by the number n of characters in an alphabet and allow the interpretation of the redundancy or the failure. Namely, the frequency of the letter with number k is given by formula

$$f(k) = \frac{1}{n}\left(1 + \frac{1}{n}\ln\frac{n!}{k^n}\right). \tag{9}$$

Would it be correct using statistics of the letters frequency to make sufficiently reliable supposition about texts language? This question has arisen from paper authors' Voynich Manuscript discussion. The Voynich Manuscript (MV) [12] – is a hand-written codex, dating from the XVI c. It consists of over 170,000 characters referred to as letters, which are united by transcriptioners in 22 distinct characters. These characters are not elements of any known alphabet. At the present time the manuscript is kept in the Beinecke Library and has the status of a cryptographic puzzle.

Numerous studies in order to decrypt the text carried out more than a hundred years and are still unsuccessfully. Versions of the authorship, content and language of the manuscript [13–16] are not supported enough by the full-fledged statistical studies. Here we try to get the answer to the following question: Is the MV an encrypted meaningful text (and in which language it is written), or it is a hoax, i. e. meaningless set of characters?

We consider transcription of MV into Latin alphabet according to [16]. Our goal is to study statistic properties of the Manuscript. Researchers have proposed numerous hypotheses about the structure of the Manuscript. There are some known theories:

- it was written with permutation of letters;
- two letters of the well-known alphabet correspond to one character of the manuscript;
- there is a key without which you can not read the text, because the same characters in different parts of the manuscript correspond to different letters;
- the manuscript is an encoded two-language text;
- vowels have been removed from the originally meaningful text;
- the text contains false spaces between words.

At the same time in various concepts (unproven in the statistical sense) for the role of the original language are proposed: Hebrew, Spanish, Russian, Manchurian, Vietnamese and much more (even Arabic or "something Indian"). At the same time we can consider the existence of false spaces as a real component of Manuscript structure. In addition, if the text contains no vowels, the vowel recovery is not uniquely.

For finding linguistic invariants of European languages the following statistics are used:

- the distance between distributions of empirical frequencies of letter combinations in norm L1;
- determination level of logarithmic approximations of one-letter distributions for texts without vocalisation;
- Hurst index distribution for a series of the number of letters concluded between the two most frequently encountered same letters;
- spectral matrix portrait of two-letter combinations.

These indicators allow to make the formal clusterization of languages from Indo-European family. As result our clusters have coincided with groups formed on the basis of studies in Historical Linguistics.

For modern languages of Indo-European family and the same group (for example German) logarithmic dependence of letter frequency on its rang is typical with accuracy 0.93-0.98 (see Fig. 2).

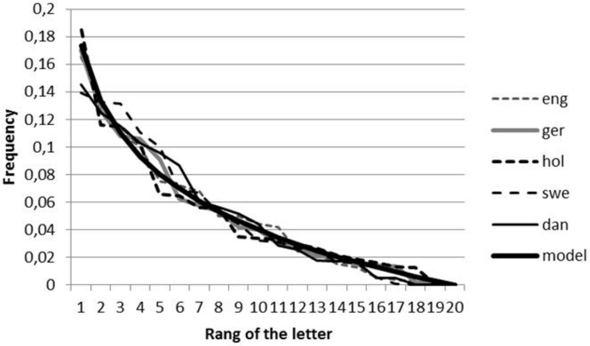

Figure 2. Distributions of letter frequencies for German language group

Actual distribution of odered frequencies for texts written in the same language differs from logarithmic approximation in L1 norm within 0.08–0.13.

Letter frequency for Danish, Serbian, Croatian and Romanian texts without vocalization have a lower approximation accuracy (0.93).

Distances between frequency distributions for texts in Cirillic for Slavic group show that Russian, Bulgarian and Serbian are related: the closest are Russian and Bulgarian (with a distance 0.06), Russian and Serbian as well as Bulgarian and Serbian have a distance 0.12.

For texts with the Latin alphabet distances between frequency distributions form clusters in accordance with language groups in sense of closeness between themselves in norm L1. It was found that Indo-European languages united in German, Romance and

Slavic groups subgroups have the same statistical properties. The distances in norm L1 between frequencies from one language group vary quite narrow (0.08–0.13). The distance between different groups belongs to interval 0.14–0.22.

Now let us compare the MV transcription symbol distribution with analogical distributions in European languages (Fig. 3).

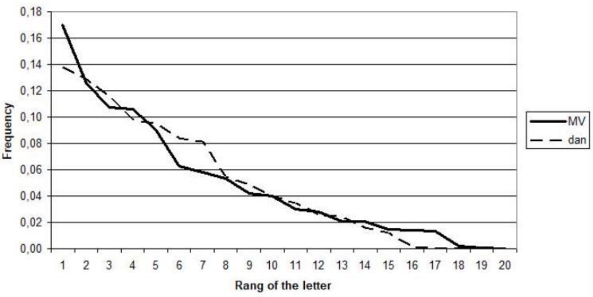

Figure 3. Distribution of letter frequencies for MV transcription

The distance between MV transcription distribution and logarithmic approximation (9) in the norm L1 is equal to 0.17. The distance MV transcription distribution and Danish language letters distribution is equal to 0,10. So MV language could be treated as Danish without vowels, but the logarithmic approximation of the last language differs from actual distribution on the value 0.12, that is sufficiently less, than 0.17 for MV. It should be emphasized, the distributions of text symbols for the same language have the same accuracy of approximation. Hence, Danish is not MV language.

The distributions of Hurst exponents for the Manuscript and ordinary texts are completely different [11]. For MV it is shifted to the right and have much less acute maximum compared to texts on one language. It means that statistics of the Manuscript does not agree with statistics of texts written in one particular language. The symbols in the Manuscript are placed “more randomly”. The main suggestion is that the Manuscript is written in several languages. So we accept the following working hypotheses regarding the MV:

1. The manuscript is a bilingual text with a common alphabet.

2. Vowels have been deleted from the text before the decoding.

3. Decoding was a bijective letter replacing by a symbol.

4. Spaces in the text are not considered as characters.

After that we need to find out which pairs of languages with a common alphabet and in which proportion could be considered as the Manuscript languages, whether they have the same or different linguistic groups and which groups exactly. To test the bilingual hypothesis we join two texts without vowels with about equal volumes, each one written in its own language, but with the same alphabets in both texts. It appears, that the mixture of pairs of texts from one linguistic group has the same statistical properties, as a group etalon. Languages with the same group not only have close ordered frequency distribution in the texts without vowels, but also a mixture of these languages has the same logarithmic approximation determination to its components.

It appears, that the MV transcription is even closer to the mixture of Latin and Danish languages in ratio 2:1, the distance between these distributions is 0.09. But as it was mentioned above, all languages from the same linguistic group have the same statistical properties. So we can consider the MV text in detail – by pages, each of them contains approximately 1 thousand symbols. The page distribution is compared with language etalons and its language is recognized by the method (4). The result is presented in Fig. 4 as a corresponding language "coloring".

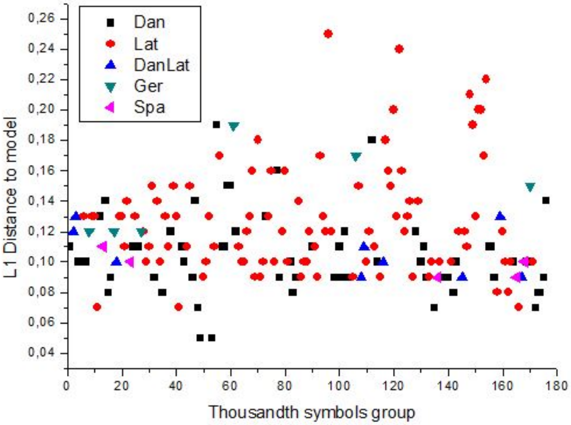

Figure 4. The recognized languages of MV pages

As it was mentioned above, the accuracy of method for one-symbol distribution is equal to 0.15. So approximately 15 percents of text pages are recognized incorrectly. From Fig. 4 it follows, that the level of critical distance is equal to 0.13, and this value coincides with upper level of distances between texts distributions for one language. So our results are consistent with properties of recognition model.

However, this identification method is effective only when we have a complete set of references. Otherwise, the identification will be incorrect. Proposed method of identification is sufficiently accurate if there is a correct reference among the reference distributions. If it is not the case, the most similar reference will be found, but it is no guarantee of the correct recognition of course. Once again, we would like to emphasize that we are talking about the European language that is the closest to the MV transcription instead of discussing which language the MV is actually written on. If the distance to the nearest reference distribution becomes too large, it's possible, that the intended distribution is missing in the library. It seems interesting that along with the expected Dutch and Latin you can see the pair of German and Spanish. As it was already mentioned, the distributions of the frequencies correspond to the language group, therefore it makes sense to identify Spanish and Latin as a single Roman language group, and German and Dutch as a single German group.

Nevertheless, an abundance of arguments support the fact that the text to be written in two or more European languages.

Finally let us consider the spectral portraits of bigram matrix for MV and any European language of Germanic and Romanic groups without vowels. Using the calculation procedure, described in [9], it is possible to construct domains of eigenvalues. The areas with the same color have eigenvalues of the matrices if the elements of these matrices are known with the precision noted in the legend. The typical comparison is presented in Fig. 5.

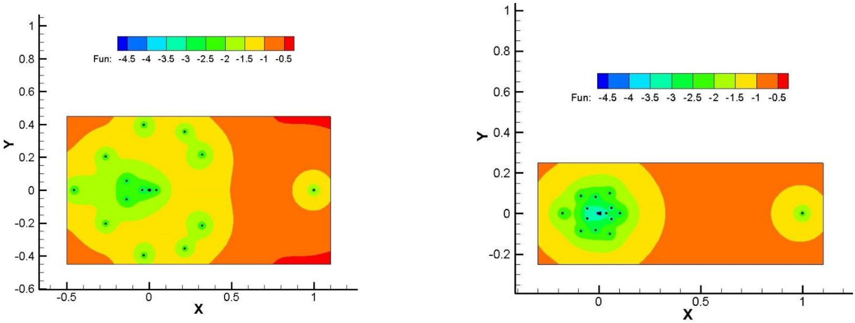

Figure 5. The MV spectral portrait (on the left) and spectral portrait of the text in English (on the right).

For all European languages the spectrum area is approximately limited by the circle with the radius 0.2. According to [8] the area of the spectrum for the texts with the full alphabet has the form not of a circle but of an ellipse with semimajor axis equal to approximately 0.5 and semiminor axis still equal to 0.2.

Comparing pictograms in Fig. 5, we see that the regions with equal accuracy are markedly different in finding eigenvalues of the matrices for MV and conventional texts. For MV transcription the circle (it is not an ellipse!) of the location of the eigenvalues has approximately two times larger radius than for natural languages. It has fundamental importance.

It is important to emphasize, all discussed arguments are fundamentally different, i.e. they express the peculiar properties of independent statistics, indicating that the interpretation of the MV as part of the composite manuscript is acceptable.

It is necessary to indicate the accuracy of the results presented in this paper. We work on statistical pattern recognition by comparison with the standard. The critical point is the accuracy with known standards. In this case standard refers to the probability distribution of text characters. If the text is made up of N signs and written by alphabet of n signs, the distribution of these characters in the text is determined with a precision ε that find numerically from equation, obtained in [8]:

$$\frac{u_{1-\varepsilon/2}}{\varepsilon} = \frac{\sqrt{N}}{\Sigma_N(n)}\,, \quad \Sigma_N(n) = \sum_{j=1}^{n} \sqrt{f_N(j)\cdot(1-f_N(j))}. \tag{10}$$

Here u is a quantile of the normal distribution. The quantile of Student's distribution with large values of N is approximated by corresponding u, and $f_N(j)$ is the empirical frequency of symbol j in this text with length N. In particular, for the logarithmic ordering model (9) when $n = 20$ the value of the sum in (10) is equal to 3.93; in relation to the MV with the number of signs $N=$ 170 thousands its actual distribution leads to $\Sigma_N(n) = 3.65$. The right side of the equation (10) with respect to ε for a theoretical model is equal to 105, and for the MV is equal to 113. These values correspond to similar accuracies $\varepsilon = 0.02$, which differ in the third decimal place. It is similarly found out that the accuracy of the frequency distribution for a single page (1500 characters) is 0.1. Consequently, the differences between the distributions of the fragments at the level of

0.08–0.13, and sheets of fragments at the level of 0.20–0.40 are not caused by statistical noise of samples, they caused by objective reasons. Thus, the difference between the samples with the specified accuracy of statistical estimations is well defined.

4. Conclusions

The results of presented statistical investigations can be summarized as follows.

First of all, It should be noted, the Bayesian recognition may be incorrect in statistical sense, because the probability interpretation of projections over system of non-orthogonal patterns may be violated. Nevertheless the recognition with the use of consequence of Bayesian method, i.e. the nearest neighbours method, appeared to be very exact in some practical cases.

With the use of this recognition method the classification of the Indo-European languages into distinct groups can be performed very accurately according to a formal statistical procedure [17, 18], i.e., pairwise clusterization of symbol frequencies distributions in texts without vowel letters. Within these sub-groups languages can be mixed together without changes in the corresponding frequency distribution. Concerning the Voynich Manuscript, it seems most plausible that it was written in two languages having the same alphabet without vowel letters: 30% of the text is written in one of the Germanic languages (Danish or German) and the rest 70% – in one of the Romance languages (Latin or Spanish).

As our further research, it will be very interesting to apply methods for data analysis from application areas [19–22], where new quality indicators, such as as traditional indicators, i.e. SIR, are analyzed. Another methods of analysis, such as a combination of linear topology and automated control [23, 24], or information description cluster method and multidimensional approache [25, 26] are also very actual and could be applied to our current tasks.

Acknowledgments

The publication has been prepared with the support of the “RUDN University Program 5-100” and funded by RFBR according to the research project No 17-07-00845.

References

1. B. Cyganek, Object Detection and Recognition in Digital Images: Theory and Practice John Wiley & Sons, Ltd, (2013) pp. 1–548.
2. B. Javidi, Image Recognition and Classification Marcel Dekker, Inc. NY, Basel, (2002).
3. V. Vapnik, A. Chervonenkis, Pattern Recognition Theory, Statistical Learning Problems (in Russian), Nauka, Moskva, (1974) pp. 1– 416.
4. N. A. Jebril, H. R. Al-Zoubi, Q. Abu Al-Haija, Recognition of Handwritten Arabic Characters using Histograms of Oriented Gradient, Pattern Recognition and Image Analysis, April (2018) , 28 (2), pp 321–345.
5. J. K. Aggarwal and R. 0. Duda, Special issue on digital filtering and image processing, IEEE Trans. Circuits Syst., vol. CAS-2, (1975) pp. 161- 304.
6. M. Arafah, Q. Abu Moghli, Efficient Image Recognition Technique Using Invariant Moments and Principle Component Analysis, Journal of Data Analysis and Information Processing, 5 (2017) 1–10.
7. I. Biederman, Recognition-by-Components: A Theory of Human Image Understanding Psychological Review 94 (2) (1987) 115–147.
8. Yu. N. Orlov, K. P. Osminin, Methods of Statistical Analysis of Literature Texts, (in Russian), Moscow: Editorial URSS, (2012) pp. 1–326 p.
9. S. K. Godunov, Modern Aspects of Linear Algebra, (in Russian), Novosibirsk: Nauchnaya kniga, (1997). pp. 1–388.

10. J. W. Demmel, Applied Numerical Linear Algebra, SIAM, Philadelphia, (1997) pp. 1–432.
11. A. A. Arutyunov, L. A. Borisov, D. A. Zeniuk, A. Yu. Ivchenko, E. P. Kirina-Lilinskaya, Yu. N. Orlov, K. P. Osminin, S. L. Fedorov, S. A. Shilin, Statistical Regularity of European Languages and Voynich Manuscript Analysis, (in Russian), Preprint KIAM of RAS, 52 (2016) pp. 1–32.
12. B. A. Shailor, Voynich Catalog Record. Yale University Beinecke Rare Book & Manuscript Library.
13. N. J. Pelling, The Curse of the Voynich: the Secret History of the World's Most Mysterious Manuscript, Surbiton, Surrey: Compelling Press (2006) pp. 1–230.
14. J. G. Barabe, Materials analysis of the Voynich Manuscript, Yale University Beinecke Rare, Book & Manuscript Library.
15. L. Levitov, Solution of the Voynich Manuscript: A liturgical Manual for the Endura Rite of the Cathari Heresy, the Cult of Isis, Walnut Creek, California: Aegean Park Press (1987) pp. 1–182.
16. G. Landini, R. Zandbergen, A Well-Kept Secret of Mediaeval Science: The Voynich Manuscript. Aesculapius 18 (1988) 77–82.
17. K. Zdeněk, V. Jan, Imperfection Sensitivity Analysis of Steel Columns at Ultimate Limit State. Archives of Civil and Mechanical Engineering, 18 (4) September (2018) 1207–1218, doi:10.1016/j.acme.2018.01.009.
18. S. Jan, K. Zdeněk, M. Lumír, N. Arnoldas, Global Sensitivity Analysis for Transformation of Hoek-Brown Failure Criterion for Rock Mass. Journal of Civil Engineering and Management. (2018) 24 5 390–398, doi:10.3846/jcem.2018.5194.
19. Y. Gaidamaka, A. Pechinkin, R. Razumchik, K. Samouylov, E. Sopin, Analysis of an $M|G|1|R$ queue with batch arrivals and two hysteretic overload control policies, International Journal of Applied Mathematics and Computer Science, 24 (3) (2014) 519–534.
20. V. Begishev, R. Kovalchukoz, A. Samuylov, A. Ometov, D. Moltchanov, Y. Gaidamaka, S. Andreev, An analytical approach to SINR estimation in adjacent rectangular cells, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9247 (2015) 446–458.
21. A. Samuylov, D. Moltchanov, Y. Gaidamaka, S. Andreev, Y. Koucheryavy, Random Triangle: A Baseline Model for Interference Analysis in Heterogeneous Networks, IEEE Transactions on Vehicular Technology, 65 (8) art. no. 7275184 (2016) 6778–6782.
22. K. Samouylov, P. Abaev, Y. Gaidamaka, A. Pechinkin, R. Razumchik, Analytical modelling and simulation for performance evaluation of sip server with hysteretic overload control, Proceedings - 28th European Conference on Modelling and Simulation, ECMS 2014, (2014) pp. 603-609.
23. V. Vishnevsky, A. Krishnamoorthy, D. Kozyrev, A. Larionov. Review of Methodology and Design of Broadband Wireless Networks with Linear Topology / Indian Journal of Pure and Applied Mathematics, June (2016), 47 (2) 329-–342, doi:10.1007/s13226-016-0190-7.
24. D. A. Aminev, D. V. Kozyrev, A. P. Zhurkov, A. Y. Romanov and I. I. Romanova, Method of Automated Control of Distributed Radio Direction Finding System, Dynamics of Systems, Mechanisms and Machines (Dynamics), Omsk, Russia (2017) pp. 1–9, doi:0.1109/Dynamics.2017.8239426.
25. M. Fomin, Cluster Method of Description of Information System Data Model based on Multidimensional Approach. Communications in Computer and Information Science 678 (2016) 657–668, doi:10.1007/978-3-319-51917-3_56.
26. M. B. Fomin, I. V. Smirnov, Methods for Identifying Clusters of Cells in Sparse Data Cubes of Multidimensional Information Systems. CEUR Workshop Proceedings, 2064 (2017) 187–194.