

The model of informative ordering in the analysis of socio-psychological processes

Yury Pichugin¹, Valentina Ivashova²

***, Olga Chudnova², Yulia Nadtochiy³, and Irina Makhova²

¹Saint-Petersburg State University of Aerospace Instrumentation, 67, Bolshaya Morskaya, 190000, St, Saint Petersburg, Russia

²Stavropol State Agrarian University, 12, Zootechnicheskiy lane, 355017, Stavropol, Russia

³Financial University under the Government of the Russian Federation, 49 Leningradsky Prospekt, 125993, Moscow, Russia

Abstract. The article presents the possibilities of informative ordering in the analysis of socio-psychological processes. We carried out a brief review of publications covering the use of mathematical apparatus in sociological and psychological research. The main groups of research strategies that have found application in the study of modern socio-psychological processes are identified. We substantiated the necessity and advantages of using the method of information ordering. In our research we described the stages of application of the proposed method, the mathematical apparatus based on the main components, estimates of structural similarity and the amount of information. The developed model of informative ordering can be used to determine the significance of information obtained in the course of sociological, psychological, social, economic and other types of research, where information about the state of the object and subject of research is organized in the form of variational series.

1 Introduction

Mathematical methods of data analysis are needed to provide evidence-based conclusions of socio-psychological, sociological, pedagogical, and other humanitarian studies. The complication of cause-and-effect relationships, factors of influence in the subject field of the humanities requires the development of mathematical methods of data analysis. In addition, the use of a complex mathematical apparatus requires a high level of qualification of researchers already at the stage of hypothesis development, the choice of research methods, tools for testing hypotheses using mathematical methods.

Thus, the research issue is actualized – the development of mathematical models of structural ordering of information in databases of social and socio-psychological research.

During the analysis of modern scientific publications (in the field of the use of mathematical methods that provide evidence for conclusions in the humanitarian subject field) we have identified various approaches to the design of research strategies. The authors

* Corresponding author: vivashov@mail.ru

of the article K.S. Berenhaut, K.E. Moore, R.L. Melvin develop approaches to the structural analysis of the community [1]. Assessing the level of cohesion of the social community, the authors of the study propose an approach for collecting significant structural information obtained as a result of induced local comparisons. Through the inclusion of mathematical results together with applications in linguistics, psychology, and comparative clustering data, the authors determine the significant structure of the community.

The significant growth of the network community is a modern reality. The network community as a discussion platform and an area for the promotion of meaningful information by stakeholders is of great interest to researchers. It is necessary to obtain new approaches of using mathematical methods for providing analytical procedures related to the structuring of large amounts of information. The authors of the article R. Banez, H. Gao, L. Li, Z. Han, H.V. Poor propose to study the behaviour of users of social networks using a multiple-population mean field games (MPMFG) [2]. The MPMFG model developed by the authors can be used to evaluate and predict the behaviour of a group in a social network, as well as their influence on the beliefs and opinions of other groups. Thus, we see how the increase in the volume and complexity of information has an impact on the development of mathematical methods of analysis and the search for research strategies that can become a data analysis tool adequate to their characteristics. In general, this confirms the relevance of our research.

Content analysis, as a method of studying large volumes of textual information, is actively used at the present time. An interesting approach to the analysis of economic discourse in studies published by the authors in different time periods in *Journal of Economic Issues* (N = 763) is shown by A. Oleinik [3]. The author emphasizes the higher readiness of economists to use mathematical methods than researchers in the social sphere, carriers of humanitarian knowledge. This conclusion is consistent with our conclusions about the need to analyze large volumes of high-quality information (where mathematical methods are needed); the insufficient willingness of the research community in the humanitarian sphere to resort to analysis and mathematically evidence-based research conclusions. There is an obvious increase in the demand for the development of mathematical methods and applied programs for their implementation. This position corresponds to the conclusions of our study.

Education as a social institution of society plays an important role in the development of the socio-economic potential of the country. Multidimensional socio-psychological, social and socio-economic processes that occur in educational organizations (with the participation of students and all stakeholders) are an object for in-depth analysis, including the use of mathematical methods. So, the authors of the article L.F. Panchenko, H.O. Korzhov, A.O. Khomiak, et al. emphasize the need to strengthen the multidimensional social dimension of higher education in order to develop models of retention and involvement of students [4]. The authors propose to update the training courses “Mathematical and Statistical Methods of Social Information Analysis”, “Social Statistics and Demography”, “Multidimensional Data Analysis”, “Structural Equation Modelling” based on modern research. Thus, it is formulated both the need for the development of mathematical methods for analyzing the social sphere of society and training courses in mathematics for future researchers.

The authors of the article S. Wojcik, A.S. Bijral, R. Johnston, A. Vespignani, D. Lazer note that it is necessary to develop a subject area that combines data on digital and real behaviour of people. In the future, such studies will make it possible to predict the spread of infectious diseases [5]. Research using digital data streams, in turn, requires the development of mathematical analysis tools. Thus, we see the relevance of our research – the development of an informative ordering model in the analysis of complex socio-psychological processes.

A study by the authors R. Vala, M. Valova, P. Drazdilova, et al. presents a new set of methods for processing and interpreting data on physical activity and lifestyle of adolescents [6]. The database of research on the health behaviour of school-age children has been analyzed using modern machine learning methods. The developed behavioural models are

presented graphically, which, according to the authors of the article, facilitates the work of specialists who do not have expert knowledge in the field of sociology, statistics, mathematical modelling. The authors' comments on the low level of knowledge in the application of mathematical methods among representatives of the humanities are in the logic of our research hypothesis.

M. Bastos in his work introduces concepts, theories and methods that are at the intersection of spatial analysis and social network analysis [7]. The author shows such processes as the complication of social practices, the intersection of real and virtual practices, and the increase in the intensity of social activity online and offline. It is legitimate to raise the question of new methods of analyzing modern social reality. The author suggests approaches to building a graph of a social network on a map, including a specially created package R Spatial social media. Graphical representation of social information is another important track in the analysis of humanitarian information.

Thus, based on a brief review of modern publications, the relevance of developing a mathematical apparatus for analyzing complex processes in the socio-humanitarian sphere of modern society is confirmed. The review examines research strategies for using mathematical methods in the field of structural analysis of a community, assessment and forecasting of group behaviour in a social network, content analysis of economic discourse in studies of different time periods, multidimensional social dimension of higher education and others [8, 9, and 10]. The authors of the publications noted the need for the development of mathematical methods in response to the complexity of social practices, the demand for graphical representation of the results of mathematical analysis and increasing the level of competence of humanitarian specialists in the field of mathematical education [11].

2 Materials and methods

To develop a mathematical apparatus that facilitates work with databases of social and socio-psychological research, we have developed a new method of informative ordering. The general concept of the method is presented in the article.

The concept of information or information entropy was first defined by Claude Elwood Shannon. According to Shannon, the average entropy $H(\xi)$, or the information that is transmitted in some message using n characters or values $\{u_1, u_2, \dots, u_n\}$, that appear in the message with probabilities $\{P_{\xi}(1), P_{\xi}(2), \dots, P_{\xi}(n)\}$, is calculated by the formula [12].

$$H(\xi) = - \sum_{i=1}^n P_{\xi}(i) \cdot \log_2 P_{\xi}(i) = \sum_{i=1}^n P_{\xi}(i) H_i.$$

Here the quantity $H_i(\xi) = -\log_2 P_{\xi}(i)$ is the so-called partial entropy. The base of the logarithm, in principle, may not be a 2, but any other number $a > 1$, as this number specifies the scale of units of information. Suppose there is another random variable η , which takes m values $\{v_1, v_2, \dots, v_m\}$, and the values themselves appear with probabilities $\{P_{\eta}(1), P_{\eta}(2), \dots, P_{\eta}(m)\}$, then, by Shannon's definition, the quantity of information $I(\xi, \eta)$ that the message ξ contains relative to the message η or the message η to the message ξ , is expressed as follows

$$I(\xi, \eta) = - \sum_{i=1}^n \sum_{j=1}^m P_{\xi\eta}(i, j) \cdot \log_a \frac{P_{\xi\eta}(i, j)}{P_{\xi}(i)P_{\eta}(j)}.$$

Here the quantity $P_{\xi\eta}(i, j)$ is the probability of ξ taking the value of u_i , and η taking the value of v_j , respectively. In the case of a quantity $P_{\xi\eta}(i, j) = 0$, the corresponding term of the given sum is assumed to be zero.

As can be seen from the formula given here, the amount of information $I(\xi, \eta)$ is a symmetric value, i.e., the equality $I(\xi, \eta) = I(\eta, \xi)$ is satisfied.

The notion of information was substantially developed in the work of I.M. Gelfand and A.M. Yaglom. In this work it is shown that in the case when there are two random Gaussian vectors ξ and η , the amount of information $I(\xi, \eta)$, that is contained in the vector ξ with respect to the vector η , and vice versa, is equal

$$I(\xi, \eta) = -\frac{1}{2} \log_a \det(\mathbf{I} - \mathbf{V}_{\xi\eta} \mathbf{V}_{\eta}^{-1} \mathbf{V}_{\eta\xi} \mathbf{V}_{\xi}^{-1}). \quad (2)$$

In this formula \mathbf{V}_{ξ} and \mathbf{V}_{η} there are matrices of mutual covariances of vector components ξ and η , respectively, \mathbf{I} is a identity matrix of appropriate dimension; $\mathbf{V}_{\xi\eta}$ is a matrix of mutual covariances of vector components ξ and vector components η . The following equality $\mathbf{V}_{\eta\xi} = \mathbf{V}_{\xi\eta}^T$ is fulfilled.

The notion of information is further developed in [4], where a regression model is considered

$$\mathbf{y} = \mathbf{F}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \quad (3)$$

Which differs significantly from the standard model of regression analysis in that the regression parameters (vector components $\boldsymbol{\theta}$) are not constant values, but have a stochastic nature. Recall that in the model (3) \mathbf{F} is a constant matrix, and there $\boldsymbol{\varepsilon}$ is a vector of random errors in the regression. It is assumed that all components of the vector $\boldsymbol{\varepsilon}$ are mutually independent, have zero mean and the same variance σ^2 . It follows that the matrix of mutual covariances of this vector has a diagonal structure $\mathbf{V}_{\boldsymbol{\varepsilon}} = \sigma^2 \mathbf{I}$. In this case, the amount of information contained in the vector \mathbf{y} with respect to the vector of parameters $\boldsymbol{\theta}$ is equal to

$$I(\mathbf{y}, \boldsymbol{\theta}) = \frac{1}{2} \log_a \det(\mathbf{I} + \sigma^{-2} \mathbf{F}^T \mathbf{F} \mathbf{V}_{\boldsymbol{\theta}}), \quad (4)$$

Where $\mathbf{V}_{\boldsymbol{\theta}}$ is a matrix of mutual covariances of the parameters.

Suppose that we choose some subset \mathbf{g} , i.e., $\mathbf{g} \subset \{1, 2, \dots, m\}$ where m is the dimension of vector \mathbf{y} . Let us denote by $\mathbf{y}_{\mathbf{g}}$ the vector which contains components of the original vector \mathbf{y} only with numbers from the subset \mathbf{g} . Then the amount of Shannon information, which is contained in the vector $\mathbf{y}_{\mathbf{g}}$ with respect to the vector of parameters, according to formula (4), is

$$I(\mathbf{y}_{\mathbf{g}}, \boldsymbol{\theta}) = \frac{1}{2} \log_a \det(\mathbf{I} + \sigma^{-2} \mathbf{F}_{\mathbf{g}}^T \mathbf{F}_{\mathbf{g}} \mathbf{V}_{\boldsymbol{\theta}}), \quad (5)$$

Where $\mathbf{F}_{\mathbf{g}}$ is a matrix that contains the rows of the original matrix \mathbf{F} , but only with numbers from a subset \mathbf{g} .

Suppose that we successively increase the number of elements of a subset \mathbf{g} from zero to m , adding one element at a time, which we choose each time from the maximum condition $I(\mathbf{y}_{\mathbf{g}}, \boldsymbol{\theta})$.

The result of this procedure will be a sequence of vector \mathbf{y} components, i.e., a sequence of component numbers \mathbf{y} , denote $i = \mu(j)$, where $j=1, 2, \dots, m$. This sequence will correspond to an increasing sequence of information quantity values I_j ($j=1, 2, \dots, m$) in ascending order. It is convenient to convert the values of I_j into percentages of the total information calculated by formula (4) $I_j = 100\% \times \left[\frac{I_j}{I(\mathbf{y}, \theta)} \right]$. This conversion is also convenient because it eliminates the dependence of I_j values on the choice of the base of the logarithm a .

A principal component regression model can be used as model (3)

$$\mathbf{y} = \mathbf{P}\mathbf{z} + \boldsymbol{\varepsilon} \quad (6)$$

Where \mathbf{P} is a matrix with mutually orthogonal columns (k first columns of matrix \mathbf{Q} , see above), which form the basis of the principal components of vector \mathbf{y} , and \mathbf{z} is the vector of principal components.

A factor analysis model can also be used

$$\mathbf{y} = \mathbf{L}\mathbf{f} + \boldsymbol{\varepsilon} \quad (7)$$

Where \mathbf{L} is the matrix of factor loadings and \mathbf{f} is the vector of factors determining the observed vector \mathbf{y} .

If the vector of factors \mathbf{f} and the vector of principal components \mathbf{z} have the same dimensions ($\dim \mathbf{f} = \dim \mathbf{z} = k$), the vector $\boldsymbol{\varepsilon}$ in models (6) and (7) have the same values, and the condition $I(\mathbf{y}_g, \mathbf{f}) = I(\mathbf{y}_g, \mathbf{z})$ or, in expanded form, according to formula (5) is fulfilled

$$\frac{1}{2} \log_a \det(\mathbf{I} + \sigma^{-2} \mathbf{P}_g^T \mathbf{P}_g \mathbf{V}_z) = \frac{1}{2} \log_a \det(\mathbf{I} + \sigma^{-2} \mathbf{L}_g^T \mathbf{L}_g \mathbf{V}_f), \quad (8)$$

Where matrices \mathbf{V}_f and \mathbf{V}_z are matrices of mutual covariances of factor vectors \mathbf{f} and \mathbf{z} , respectively, \mathbf{L}_g and \mathbf{P}_g are defined similarly to the matrix \mathbf{F}_g . Condition (formula) (8) follows from the fact that Shannon information quantity, is invariant to linear transformations of vectors, and vectors \mathbf{f} and \mathbf{z} are linked exactly by linear transformation. Taking into account that the factors have unit variance and remain independent under rotation, i.e. the matrix $\mathbf{V}_f = \mathbf{I}$ and can be omitted, finally we have

$$I(\mathbf{y}_g, \mathbf{z}) = I(\mathbf{y}_g, \mathbf{f}) = \frac{1}{2} \log_a \det(\mathbf{I} + \sigma^{-2} \mathbf{L}_g^T \mathbf{L}_g).$$

However, there is some technical advantage of model (6). When calculating the principal component basis, we need to diagonalize the mutual covariance matrix of the initial vector \mathbf{y} or the correlation matrix when considering the normalized initial vector. The significant part of the spectrum $\{\lambda_1, \lambda_2, \dots, \lambda_k, \lambda_{k+1}, \dots, \lambda_m\}$ (diagonal of this matrix) gives us the matrix \mathbf{V}_z which, due to the mutual independence of the principal components, has a diagonal structure $\mathbf{V}_z = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$, while the rest of the spectrum goes to calculate the biased estimate σ^2 .

$$\sigma^2 = \frac{1}{m-k} \sum_{i=k+1}^m \lambda_i.$$

As noted in [13], where the information was considered in the vector case (see above), the computation of information is in principle based precisely on biased estimation.

Note. Since the principal component method is based on the correlation matrix (see above), the vector \mathbf{y} in model (6) is not the original column of the sampling matrix \mathbf{Y} , but the column with centred and normalised elements, i.e. $y_{ij} = \frac{(y_{ij} - \bar{y}_i)}{\sigma_i}$ ($i = 1, 2, \dots, m; j = 1, 2, \dots, n$), where \bar{y}_i and σ_i , respectively, the mean value and the standard deviation of the i -th row of the sampling matrix. However, this transformation is performed automatically in SPSS when calculating the correlation matrix \mathbf{R} .

3 Results and Discussion

An example of the implementation of the method can be seen in the results of the study (Table 1).

Table 1. Results of information sequencing (first age group). Source: [14].

N ^o selection step j	N ^o question μ (j)	Amount of information I_j (%)	Increment of the amount of information ΔI_j (%)	Ratio of consecutive increments of the amount of information $\frac{\Delta I_{j+1}}{\Delta I_j}$
1	12. I enjoy doing my job	11.74	11.74	
2	11. The job I do cannot be done by someone with lower qualifications	22.03	10.28	0.88
3	9. There are often situations in life where you can't do all the work you're asked to do	32.22	10.19	0.99
4	13. I am not satisfied with the organisation of work in our team	41.86	9.64	0.95
5	4. Job satisfaction is more important than earning a high salary	51.09	9.23	0.96
6	3. I have good relationships with people in my team	59.47	8.38	0.91
7	16. Even if I was offered a higher salary. I would not change my job	65.05	5.58	0.67
8	17. My line manager often doesn't understand or doesn't want to understand me	69.65	4.59	0.82
9	8. People with whom I work respect me	73.96	4.31	0.94
10	2. In recent years. I have made progress in my profession	78.10	4.14	0.96
11	6. What I like about my job is that I get to learn new things	82.11	4.01	0.97

12	14. I often have disagreements with my colleagues at work	85.63	3.51	0.88
13	18. Good working conditions at my company	88.38	2.75	0.78
14	15. I am rarely rewarded for my work	91.06	2.69	0.98
15	10. My bosses have been very satisfied with my work recently	93.55	2.49	0.93
16	5. My job title does not match my abilities	95.37	1.82	0.73
17	1. What I do at work interests me	97.14	1.77	0.97
18	7. From year to year. I feel my knowledge and skills are improving	98.68	1.54	0.87
19	19. I set and achieve goals in my work	100.00	1.32	0.86

The possibilities of the proposed method of informative ordering of data are evident. The discussion, which revolved around the subject of the study (strategies for using mathematical methods in the humanitarian field), collects best practices of mathematical data analysis. The author of the article Y. Ozhigov considers the prospects of quantum computing in models of social behaviour [15]. The transition to a quantum language in computer science will allow the use of a single mathematical apparatus and computational methods in such different fields, including in the field of social behaviour research. The importance of the quantum approach in sociology is justified in the concept of the “social laser”. In the article Y. Ozhigov defines the “crowd effect” – the peak nature of the excitement of large social groups, based on the Tavis-Cummings-Hubbard model. Thus, we see a certain trend in the development of mathematical methods for the validity and objectivity of the conclusions of the humanitarian sphere, which was the purpose of our study.

In the book by the authors R. Startup, E.T. Whittaker we see a detailed consideration of the possibilities of using mathematical statistics in sociological, psychological, political, and demographic studies [16]. The examples given by the authors show the transformation and complication of social processes and the humanitarian sphere as a whole. It is important that the proposed material is designed for the perception of not specially trained students, but is possible for the assimilation of a representative of the humanities of university training. It also corresponds to the focus of our research on ensuring ease of use of the proposed mathematical method of informative ordering.

The authors’ approach is used in solving the problem of estimating the repeatability of processes whose mathematical model is stochastic processes [17]. The prospects for the application of this method are related to the fields of science and social practice, where it is necessary to assess the frequency of the process in conditions of information scarcity, when the entire process is observed for no more than a limited time. The development of this mathematical approach also contributes to the development of the mathematical apparatus for the study of social processes, as well as a number of other studies [18, 19, and 20].

Thus, the discussion on the development of mathematical methods of social sphere research is highly relevant and the scientific community. It demonstrates opinions on the need to develop means of objective proof of conclusions and building up the mathematical competencies of humanitarians [21, 22, 23].

4 Conclusion

The conducted theoretical review of modern publications devoted to the results of social, economic, socio-psychological research shows the relevance of the method developed by our research group. Its capabilities consist in identifying the most significant factors that influence the result – making decisions about actions, behaviour, motivation, etc.

The theoretical review analyzes the mathematical methods applied by the authors in the field of structural analysis of the community, assessment and forecasting of the behaviour of a group in a social network, content analysis of economic discourse in studies of different time periods, multidimensional social dimension of higher education and others. It is established the necessity of developing mathematical methods in response to the complexity of social practices, the need for graphical representation of the results of mathematical analysis and increasing the level of competence of humanitarian specialists in the field of mathematical education.

References

1. K.S. Berenhaut, K.E. Moore, R.L. Melvin, *Sciences of the United States of America* **119(4)**, e2003634119 (2022)
2. R. Banez, H. Gao, L. Li, Z. Han, H.V. Poor, *Modeling and Analysis of Opinion Dynamics in Social Networks Using Multiple-Population Mean Field Games*, *IEEE Transactions on Signal and Information Processing over Networks* (2022)
3. A. Oleinik, *Journal of Economic Issues* **56(1)**, 259-280 (2022)
4. L.F. Panchenko, H.O. Korzhov, A.O. Khomiak, V.Ye. Velychko, V.N. Soloviev, *CEUR Workshop Proceedings* **3085**, 124-138 (2022)
5. S. Wojcik, A.S. Bijral, R. Johnston, A. Vespignani, D. Lazer, *Nature Communications* **12(1)**, 194 (2021)
6. R. Vala, M. Valova, P. Drazdilova, P. Krömer, J. Platos, *Children and Youth Services Review* **128**, 106150 (2021)
7. M. Bastos, *Spatializing Social Media: Social Networks Online and Offline, Spatializing Social Media: Social Networks* (2021)
8. W. An, *Sociological Methods and Research* **50(3)**, 939-943 (2021)
9. N. Gabdrakhmanova, M. Pilgun, *Applied Sciences (Switzerland)* **11(14)**, 6579 (2021)
10. S.A. Borz, E. Iordache, M.V. Marcu, *Forests* **12(7)**, 926 (2021)
11. R. Azen, C.M. Walker, *Categorical Data Analysis for the Behavioral and Social Sciences, Categorical Data Analysis for the Behavioral and Social Sciences* (2021)
12. C.E. Shannon, W. Weaver, *Publisher Foreign Literature* **1953**, 288 (1949)
13. Yu.A. Pichugin, *Journal Physics and Mathematics* **11(3)**, 74-89 (2018)
14. Y.A. Pichugin, V.A. Ivashova, V.N. Goncharov, O.U. Kolosova, *European Journal of Contemporary Education* **11(1)**, 138–146 (2022)
15. Y. Ozhigov, *Communications in Computer and Information Science* **1510**, 365-375 (2021)
16. R. Startup, E.T. Whittaker, *Introducing Social Statistics (Book)*, *Introducing Social Statistics* (2021)
17. A.V. Toropova, M.V. Abramov, T.V. Tulupyeva, *Scientific and Technical Journal of Information Technologies, Mechanics and Optics* **21(5)**, 727-737 (2021)

18. Y.I. Brodsky, IFAC-Papers On Line **54(13)**, 46-51 (2021)
19. J. Xue, M. Zhang, M. Xu, *Modeling the Impact of Social Distancing on the COVID-19 Pandemic in a Low Transmission Setting*, *IEEE Transactions on Computational Social Systems*, in print (2021)
20. D. Nikitin, C. Canudas de Wit, P.A. Frasca, *Continuation Method for Large-Scale Modeling and Control: from ODEs to PDE, a Round Trip*, *IEEE Transactions on Automatic Control*, in print (2021)
21. W. Gao, Y. Li, *Solving a New Test Set of Nonlinear Equation Systems by Evolutionary Algorithm*, *IEEE Transactions on Cybernetics*, in print (2021)
22. T. Weron, K. Sznajd-Weron, Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics **12744**, 309-315 (2021)
23. J. Brittin, O.M. Araz, A. Ramirez-Nafarrate, T.T. Huang, *An Agent-Based Simulation Model for Testing Novel Obesity Interventions in School Environment Design*, *IEEE Transactions on Engineering Management* (2021)