ECONOMETRICAL MODELING OF THE STRUCTURE OF MULTIDIMENSIONAL STATISTICAL INTERRELATIONS

Alexander K. Rozentsvaig[1]
Alexey G. Isavnin[2]
Anton N. Karamyshev[3]

Abstract: In economics, the general theory is largely descriptive, and mathematical models are not only statistical but also partial. Therefore, an economic phenomenon usually requires using partial methods and getting only private solutions limited by particular conditions - the type of activity, its place and time of implementation. The real idea of the nature of the economic phenomenon that interests us is given only by statistical data. Correlation analysis is a time-consuming and completely non-formalizable task when it is necessary to justify the relationship structure of a large number of factors. In addition, the quality and interpretation of the results of statistical analysis are predetermined by the nature of the statistical models used to obtain sample estimates of their parameters. Due to the complexity of multidimensional statistical models, general theoretical concepts are usually limited by the assumption that the sampled data does not contradict the normal multidimensional distribution law. This greatly simplifies multivariate statistical analysis and therefore it always leads to linear regression relationships, which corresponds to a trivial system of correlation relationships and is rarely observed in reality. The structure of each economic object is unique, therefore, it is proposed to refine it using a system of correlation matrices of various orders. It is shown that the generalization of large volumes of multidimensional sample data in the form of "portraits" of correlation matrices clearly represents the specific features of the object of study. Moreover, the empirical system of statistically significant relationships is transformed into the corresponding

[1] Kazan Federal University. e-mail: antonkar2005@yandex.ru. **Tel.: +7-960-067-65-50**
[2] Kazan Federal University. e-mail: antonkar2005@yandex.ru. **Tel.: +7-960-067-65-50**
[3] Kazan Federal University. e-mail: antonkar2005@yandex.ru. **Tel.: +7-960-067-65-50**

model of economic relationships. Prerequisites are being created for the practical use of universal systems analysis methods based on modern theoretical and software tools of information technologies.

Keywords: econometrics, correlation analysis, multivariate sampling, statistical model, relationship structure

1 Introduction

The interconnection of disciplines that study complex socio-economic phenomena and processes significantly increases the necessary amount of basic knowledge in connection with new ideas of science and the needs of the practice. The basic concepts and methods of economic statistics have evolved over the centuries as the ever-growing demands of practical activity have accumulated [1, 2]. However, at the same time, not only the statistical analysis tools were created and improved. Along with them, methods countering with modern conditions were established.

The same problems arise in the development of hardware and software for computer technology. Therefore, the role of systematic methods of organization and management is growing not only in information technology but in technical and socio-economic systems [3]. A less formal apparatus of discrete mathematics turned out to be in demand as a source of adequate mathematical representations and economic models, as well as a tool for formalizing problems.

Undoubtedly, the role of classical mathematical disciplines is also increasing. However, the generalization of mathematical and statistical methods for solving economic problems under accelerated development, which is initiated by global informatization, is largely associated with methods of system analysis and discrete mathematics [4, 5]. Thus, the effective organization and management of modern production are associated with typical facilities and industries rather than with unique enterprises. They are designed based on system representations according to design standards such as BRP or MRP [6]. Models of business processes formalize production activities, which should ensure the quality of products in accordance with ISO standards. The generally accepted, typical structures of interconnections that implement these standards predetermine the final choice of a certain type (class) of mathematical models corresponding to them [7].

However, the specifics of the place, time, and nature of economic activity always remain a unique characteristic of each object. This area is the subject of statistics, which, using statistical models, estimates such non-random deviations from the standards that take shape under non-deterministic conditions [8-10]. The structure of the statistical relationships of the system of economic indicators $X_1, X_2, \ldots, X_P$ is fixed by real statistics in the form of a multidimensional sample. It is predetermined by the law of the joint distribution of their probabilities $f(x_1, x_2, \ldots, x_p)$ if the indicators are considered as a system p of random variables *CpCB* $(X_1, X_2, \ldots, X_P)$. A classical regression analysis proceeds from theoretical ideas about the normal nature of multidimensional distribution. Moreover, each random variable also has a normal distribution law. Therefore, all its results are a direct consequence or generalization of the mathematical properties of the normal multidimensional law. In particular, a linear, additive model of multiple regression is adequate only when the distribution of the sample is close enough to the normal distribution [11-13].

2 Estimation Of The Structure Of Statistic Relationships Based On Multidimensional Sampling Data (Main Part)

An econometric analysis distinguishes explained (dependent) and explanatory (independent) variables. There can be any number of them but several explained variables lead to model representations associated with regression systems in the form of recursive or simultaneous equations. This is a separate, more general section of econometrics [8], therefore, in the future, we will to considering an explained variable only.

The following designations are accepted: Y – explained variable, and – $X_1, X_2, \ldots, X_P$ – explanatory (factorial) variables. Their total number is $p+1$. Any variable can be explained, depending on the nature of the subject area and the objectives of the study. It is denoted as *Y*, unlike the rest of the explanatory variables - X_i, where $1 \le i \le p$. All subsequent analysis is performed within this limitation. The analysis is repeated, if it is necessary to study the behavior of another variable in the population, due to the influence of other variables.

Statistical estimates (empirical values) of the theoretical characteristics of the joint probability distribution of the system *CB* ($Y, X_1, X_2, \ldots, X_P$) are usually obtained from *n* random measurements *(p+1)th CB*. It is believed that the general population is subordinate to the normal multidimensional distribution, and $n >> (p+1)$.

Table 1. Presentation of multidimensional sampling data

$$\begin{array}{ccccccc}
Y & X_1 & X_2 & \ldots & X_i & \ldots & X_p \\
y_1 & x_{11} & x_{21} & \ldots & x_{i1} & \ldots & x_{p1} \\
y_2 & x_{12} & x_{22} & \ldots & x_{i2} & \ldots & x_{p2} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
y_j & x_{1j} & x_{2j} & \ldots & x_{ij} & \ldots & x_{pj} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
y_n & x_{1n} & x_{2n} & \ldots & x_{in} & \ldots & x_{pn}
\end{array}. \tag{1}$$

To refine the composition and structure of correlation relationships in the form of regression relationships, based on this initial information, which represents the real state of the studied economic phenomenon, the coefficients of partial, multiple-partial, and multiple correlations are calculated.

Estimates of the linear coefficients of pair correlation ρ_{ij} of any variables X_i and X_j are calculated according to a sample of *n* real measurements of economic factors by the formulas:

$$r_{x_i x_j} = \frac{S_{x_i x_j}}{\sqrt{S^2_{x_i}} \cdot \sqrt{S^2_{x_j}}}, \tag{2}$$

where $S_{X_iX_j} = \frac{1}{n}\sum_{k=1}^{n}\left(x_{ik} - \overline{X_i}\right)\left(x_{jk} - \overline{X_j}\right)$, $S^2_{X_i} = \frac{1}{n}\sum_{k=1}^{n}\left(x_{ik} - \overline{X_i}\right)^2$, $S^2_{X_j} = \frac{1}{n}\sum_{k=1}^{n}\left(x_{jk} - \overline{X_j}\right)^2$.

For a multivariate sample for *p* variables, they are represented by matrices of sample or empirical values of the corresponding theoretical indicators:

$$S = \begin{pmatrix} S_1^2 & S_{12} & \dots & S_{1p} \\ S_{21} & S_2^2 & \dots & S_{2p} \\ \dots & \dots & \dots & \dots \\ S_{p1} & S_{p2} & \dots & S_p^2 \end{pmatrix} \qquad R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

All other indicators of statistical communication are calculated on the basis of these basic characteristics.

To calculate the partial coefficients of the pair correlation of the first order, linear coefficients $r_{X_iX_j}$ of zero-order are used:

$$r_{X_iX_j/X_k} = \frac{r_{X_iX_j} - r_{X_iX_k} r_{X_jX_k}}{\sqrt{1 - r^2_{X_iX_k}}\sqrt{1 - r^2_{X_jX_k}}} \qquad (3)$$

At the same time, the linear relationship of the variables X_i and X_j is "cleared" of the influence of X_k – one of the remaining *p* variables of the sample population.

The partial coefficients of pair correlation of the second, third, and higher orders are calculated using the recurrence formula and the previously calculated coefficients of the previous order:

$$r_{X_iX_j/X_kX_m} = \frac{r_{X_iX_j/X_k} - r_{X_iX_k/X_m} r_{X_jX_k/X_m}}{\sqrt{1 - r^2_{X_iX_k/X_m}}\sqrt{1 - r^2_{X_jX_k/X_m}}} \qquad (4)$$

If it is necessary to “clear” the relationship between the variables X_i and X_j from the influence of all other variables in the population, then it is convenient to replace the recurrence relation with a sufficiently large number of variables by another using the original correlation matrix R:

$$r_{X_iX_j/\overline{X_iX_j}} = \frac{-R_{ij}}{\sqrt{R_{ii} \cdot R_{jj}}} \qquad (5)$$

where R_{ij} is the algebraic complement of the element r_{ij} of the correlation matrix, $\overline{X_iX_j}$ are the variables complementing the variables X_i and X_j to the full composition of the set of interconnected CBs.

Partial pair correlation coefficients of the highest order are most convenient for practical use. They are the only characteristics for each pair of variables X_i and X_j, represent the close relationship of each of all pairs of variables in a “pure” form.

The aggregate, multiple correlation coefficient characterizing the close relationship of the explained variable Y with all other factorial variables of the system

$$R_{YX_1X_2\ldots X_p} = \sqrt{1 - \frac{|R'|}{|R|}} \qquad (6)$$

where □*R*□ is the determinant of the correlation matrix R for the system of the *(p+1)*-th economic indicator *Y*, X_1, X_2, …, X_P, and $|R'|$ is the determinant of this matrix of internal factor correlation with the excluded first row and column corresponding to the indicator *Y*.

Table 2: The matrix of linear coefficients of pairwise and partial correlations of the CB system (**Y**, X_1, X_2, …X_p).

ij	***Y***	***X1***	***X2***	***X3***		***p***
	1	r_{YX_1}	r_{YX_2}	r_{YX_3}	..	
1	$r_{YX_1/X_2X_3...Xp}$	1	$r_{X_1X_2}$	$r_{X_1X_3}$	..	
2	$r_{YX_2/X_1X_3...Xp}$	$r_{X_1X_2/YX_3...Xp}$	1	$r_{X_2X_3}$	..	
3	$r_{YX_3/X_1X_2...Xp}$	$r_{X_1X_3/YX_2...Xp}$	$r_{X_2X_3/YX_1...Xp}$	1	..	
	…	…	…	…	..	..
p	$r_{YXp/X_1X_2...X_{P-1}}$	$r_{X_1Xp/YX_2...X_{P-1}}$	$r_{X_2Xp/YX_1...X_{P-1}}$	$r_{YX_P/X_1X_2...X_{P-1}}$	..	

The upper right part of Table 2 shows the linear coefficients of pair correlation. They are elements of the correlation matrix R = $(r_{ij})_{(p+1)□(p+1)}$, supplemented by another explained variable *Y*. A comparison of the partial pair correlation coefficients of various orders is used to justify the model representation of the multiple statistical relationships.

The structural features of the composition of the explanatory variables are multicollinearity, which is important not only for creating the conditions for the justified use of classical OLS. The presence of groups of interrelated indicators serves as the basis for identifying internal processes that shape the behavior of the studied economic object. Their interaction with other uncorrelated indicators reveals the mechanisms of complex, non-linear formation of the behavior of the explained variable.

The independence condition in the aggregate of the system *p* of random variables requires the independence of not only each pair but also all possible combinations of three, four, up to the (*p* -1)-th component of the system. Based on this, the analysis of the characteristics of pair correlations must be verified using triple, quadruple, etc. characteristics of multiple relationships. Unlike pair correlations, they are called multiple correlations, and their analogs of partial correlation are called multiple-partial correlations. All of them are also calculated based on the correlation matrix *R* using multiple correlation coefficients:

$$r_{YX_1X_2\ldots X_k\,/\,X_{k+1}\ldots X_p} = \sqrt{\frac{R^2_{YX_1X_2\ldots X_p} - R^2_{YX_{k+1}\ldots X_p}}{1 - R^2_{YX_{k+1}\ldots X_p}}} \qquad (20)7$$

Under fixation of the explained variable *Y*, a matrix of linear and partial coefficients of the triple set-private correlation *(p-2)* of the second-order (with any pairs of explanatory variables) is also formed. As a result of this, as in the case of pair correlation, the multifactor correlation matrix remains two-dimensional but of a smaller order $(r_{YxiXj})_{p\square p}$. If all variables are selected for the regression, then the intrafactual correlation will be statistically insignificant. The inter-factual correlation is the multiple correlation coefficient (6), which is also a zero-order multiple-partial coefficient.

To assess the structure of the regression model, it is necessary to additionally use empirical sample information in the form of a system of coefficients of linear, partial, and multiple-partial correlation. Although this compresses a large amount of the initial statistical data, the total number of communication characteristics remains quite large. The next step is to highlight essential, statistically significant relationships. The statistically insignificant matrix elements are replaced by zeros. The number of non-zero elements for structured economic phenomena is much less than the number of zeros. They are scattered throughout the matrix, forming what is called a portrait of the matrix or its pattern of zeros - not zeros [14]. The portrait of the matrix becomes a model of the relationship structure, formalizing the choice of a statistical model for generalizing real sample data.

The correct choice of the structure of the statistical model improves the quality and reliability of the results of processing sample data [15]. The sparseness of the correlation matrix is a clear reflection of the fact that a fairly simple idea of the most characteristic relationships of the studied object is obtained. Further, to illustrate the foregoing, several characteristic situations are considered.

Example 1. A detailed “portrait” of the correlation matrix *R* (8) is obtained for the system of (*p*+*1*)-th economic indicator. All explanatory variables X_i are in a statistical relationship with the explained variable *Y*. But at the same time, they remain uncorrelated among themselves. In other words, intrafactual correlation is completely absent.

$$R = \begin{pmatrix} 1 & r_{YX_1} & r_{YX_2} & r_{YX_3} & \dots & r_{YX_{P-1}} & r_{YX_P} \\ r_{YX_1} & 1 & 0 & 0 & \dots & 0 & 0 \\ r_{YX_2} & 0 & 1 & 0 & \dots & 0 & 0 \\ r_{YX_3} & 0 & 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ r_{YX_{P-1}} & 0 & 0 & 0 & \dots & 1 & 0 \\ r_{YX_P} & 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix} \qquad (8)$$

The mathematical model of such a system for the relationship of indicator *Y* with explanatory variables is the following form of the distribution law:

$$f(y,x_1,x_2,...,x_P) = f_{x_1x_2\cdots x_P}(y) f_{x_1}(x_1) f_{x_2}(x_2) \cdots f_{x_p}(x_p) \qquad (9)$$

It is graphically represented by the following structural diagram:

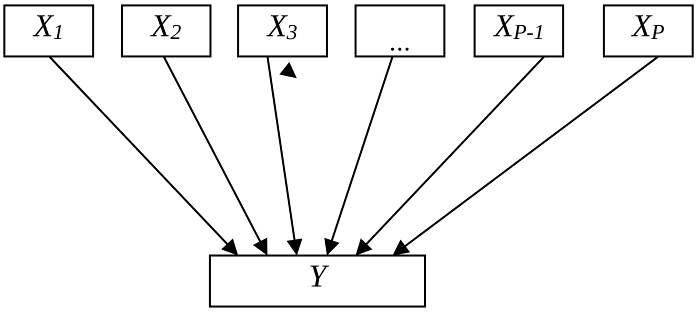

Fig. 1. The correlation graph of uncorrelated factors

This is the most important particular case when the application of the linear model of statistical correlation is most justified and explains the behavior of *Y* by the influence of *p* independent variables in the aggregate X_i. It gives the highest quality estimates when using the ordinary least squares method (OLS) [8]. Since the basic theoretical prerequisites of the regression analysis are fulfilled, the structure of the interconnections of such a system of indicators turns out to be the simplest, which allows using pair regression methods with a slight generalization.

Example 2. A characteristic "portrait" of the correlation matrix *R* for a system of seven economic indicators (10) represents a more complex structure of statistical relationships:

$$R=\begin{pmatrix} 1 & r_{YX_1} & r_{YX_2} & r_{YX_3} & 0 & 0 & r_{YX_6} \\ r_{YX_1} & 1 & 0 & 0 & 0 & 0 & 0 \\ r_{YX_2} & 0 & 1 & 0 & 0 & 0 & 0 \\ r_{YX_3} & 0 & 0 & 1 & r_{X_3X_4} & r_{X_3X_5} & 0 \\ 0 & 0 & 0 & r_{X_3X_4/\overline{X_3X_4}} & 1 & r_{X_4X_5} & 0 \\ 0 & 0 & 0 & r_{X_3X_5/\overline{X_3X_5}} & r_{X_4X_5/\overline{X_4X_5}} & 1 & 0 \\ r_{YX_6} & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad (10)$$

The mathematical model of such a system for the relationship of indicator *Y* with explanatory variables is the following form of the distribution law:

$$f(y,x_1,x_2,x_3,x_4,x_5,x_6)=f_{x_1x_2x_6}(y)f_{x_3x_4x_5}(x_1,x_2,x_6)f_{x_4x_5}(x_3)f(x_4,x_5) \qquad (11)$$

It is graphically represented by the following structural diagram in Fig. 2:

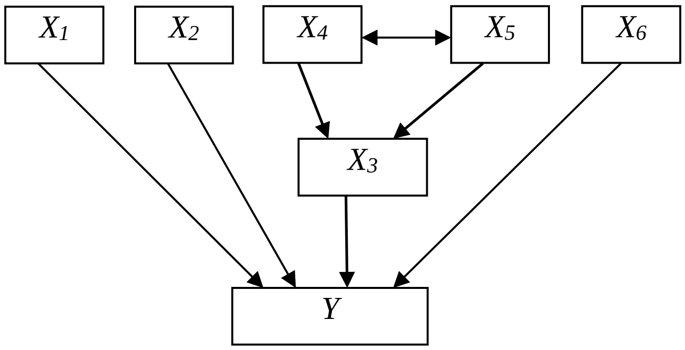

Fig. 2. The correlation graph of partially correlated factors

In this case, there is multicollinearity of the explanatory variables *X₃*, *X₄*, and *X₅*, which are in a statistical relationship. Moreover, only one of them, *X₃*, is associated with the explained variable *Y*. If the correlation strength $r_{X_3X_4}$, $r_{X_4X_5}$, and $r_{X_3X_5}$ turns out to be quite high, the estimates of the usual regression analysis using OLS are unsuitable for practical decisions.

In addition, this example indicates that low-correlation coefficients r_{YX_4} and r_{YX_5} cannot always serve as a basis for excluding them from consideration. Here, the usual OLS must be used to estimate the regression dependence of X_3 on X_4 and X_5 and then to explain the behavior of the variable Y. If now only X_1, X_2, $X_3=f(X_4,X_5\ )$ and X_6 are explanatory variables, then multicollinearity will be eliminated. Then all the prerequisites for the use of OLS are satisfied, and the result of statistical analysis is a recursive regression.

Example 3. Example 2 excludes the relationship between X_3 and the explained variable Y. However, the statistical insignificance of the coefficients of their pair correlations with the explained variable Y is not enough to exclude from the correlation analysis the subsystem of interrelated variables X_3, X_4, and X_5. The “portrait” of the correlation matrix R, characteristic of this rather simple situation, is as follows:

$$R=\begin{pmatrix} 1 & r_{YX_1} & r_{YX_2} & 0 & 0 & 0 & r_{YX_6} \\ r_{YX_1} & 1 & 0 & 0 & 0 & 0 & 0 \\ r_{YX_2} & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & r_{X_3X_4} & r_{X_3X_5} & 0 \\ 0 & 0 & 0 & r_{X_3X_4/\overline{X_3X_4}} & 1 & r_{x_4x_5} & 0 \\ 0 & 0 & 0 & r_{X_3X_5/\overline{X_3X_5}} & r_{X_4X_5/\overline{X_4X_5}} & 1 & 0 \\ r_{YX_6} & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \qquad (12)$$

The model representation of a multidimensional sample, in this case, is two separate regressions:

$$\begin{gathered} f(y,x_1,x_2,x_6)=f_{x_1x_2x_6}(y)f_{x_1}(x_1)f_{x_2}(x_2)f_{x_3}(x_3) \\ g(x_3,x_4,x_5)=g_{x_4x_5}(x_3)f_{x_4}(x_5)f_{x_5}(x_5) \end{gathered} \qquad (13)$$

It is graphically represented by the following structural diagram in Fig. 3:

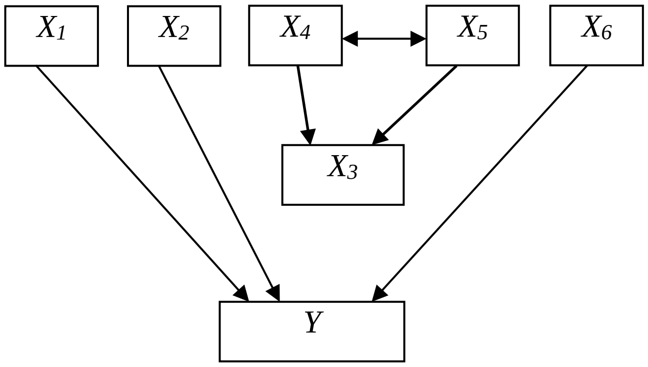

Fig. 3. The correlation graph of partially isolated factors

Since both regressions represent a single economic object, they cannot be evaluated individually using OLS. In such cases, regression analysis can lead to poor-quality (biased and untenable) estimates. In the general case, to solve such problems, model representations in the form of systems of simultaneous equations are used. For the simultaneous assessment of several regression equations, it is necessary to carry out a preliminary study and justify the conditions for the use of OLS. It is usually used in the form of an indirect, two-step or three-step OLS [8].

3 Methods

The study applied the following methods:

1. A selective analysis of specialized literature with a high citation index for the topics indicated in the title of the article. In particular, information has been collected on econometric modeling methods for complex relationships.

2. The generated array of information was systematized for the purpose of further analysis.

3. The authors interpreted the results of the study and made conclusions.

4 Results And Discussion

The article deals with the task of evaluating the system of characteristics of multifactor linear relationships and means of their systematization using multidimensional sample data. It resembles the mathematical problem of a piecewise linear approximation of unknown functions given by their numerical values. But even numerical values of the correlation coefficients that are close to zero do not always mean the complete absence of a statistical relationship. This may also mean a significant deviation of the regression dependence on the linear form. The presence of nonlinear relationships must be additionally tested using more general characteristics of the statistical relationship such as correlation relationships.

5 SUMMARY

Since the economy exists only in the form of a system organization, the methods of system analysis are largely applicable to the characteristics of the relationship of the system of its statistical indicators. Systematization of information in the form of matrices of linear coefficients of multiple, partial, and multiple-partial correlations of various orders generalizes large volumes of statistical information. The nature of the structure of these matrices contributes to a reasonable choice of the type of regression model and the involvement of specially developed econometric analysis methods.

6 Conclusions

The authors propose to use the “portraits” of correlation matrices to justify the statistical model of multidimensional sample data. At the same time, a circle of statistically significant relationships, which model the internal structure of economic phenomena, is substantiated. From a practical point of view, the analysis of the structure models of multidimensional statistical relationships and their visualization provides great opportunities by the use of universal methods of system analysis.

7 Acknowledgments

The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

References

H. Ahrens, J. Leiter, Multivariate analysis of variance. Trans. from German, M.: Finance and Statistics, 1985, 230 p.

A. Afifi, S. Eisen, Statistical analysis: a computer-based approach. Trans. from English - M.: Mir, 1982. - 488 p.

I.N. Drohobytskii. System analysis in economics, UNITI-DANA, 2012, 423 p.
R.L. Ackoff, The art of problem solving, John Wiley & Sons, 1978.

K. Nakamatsu, G.Phillips-Wren, C. Jain Lakhmi, R. Howlett, New Advances in Intelligent Decision Technologies. Results of the First KES International Symposium IDT. Springer, (2009).

Process management / Ed. Becker J., Vilkova L. et al.; [trans. from German]. - M.: Eksmo, 2007. - 384 p.

V.V. Repin, V.G. Eliferov. Process approach to management. Modeling of business processes, Moscow: “Standards and Quality”, 2005, 408 p.

Econometrics: a textbook / I.I. Eliseeva, S.V. Kurysheva, T.V. Kosteev, et al.; Ed. I.I. Eliseeva. - 2nd ed., revised. - M.: Finance and Statistics, 2007. - 576 p. ISBN 978-5-279-02786-6

Econometrics: a study guide / D.F. Fedorov, A.K. Rosenzweig, A.N. Karamyshev, I.F. Nazmiev. - Naberezhnye Chelny: Publishing House of the Naberezhnye Chelny Institute of KFU, 2016. - 104 p.

T. T. Soong, Fundamentals of probability and statistics for engineers, John Wiley & Sons, 2004.

Ch. Dougerty, Introduction to econometrics, Oxford University Press, 1992.

W.H. Greene, Econometric Analysis, 7th ed. Prentice Hall, 2011.

B. Bolch, K.J. Huan, Multidimensional statistical methods for economics. Trans. from English, Moscow: Statistika, 1979, 317 p.

S. Pissanetzky, Sparse matrix technology, London Academic Press, 1984.

George, J. Liu, Computer solution of large sparse positive defines systems, Prentice-Hall, 1981