Advances in Economics, Business and Management Research, volume 164
Proceedings of International Scientific and Practical Conference "Russia 2020 - a new reality: economy and society" (ISPCR 2020)

Development of Predictive Models of Socio-Economic Systems Based on Decision Trees with Multivariate Response

Kislyakov A.N.
Vladimir branch of RANEPA,
Vladimir, Russia,
ankislyakov@mail.ru

Filimonova N.M.
Vladimir branch of RANEPA,
Vladimir, Russia,
natal_f@mail.ru

Omarova N.Yu.
Yaroslav-the-Wise Novgorod State University,
Veliky Novgorod, Russia,
n-omarova@mail.ru

***Abstract*—The work is devoted to the actual problem of constructing decision trees with a multidimensional response of the optimal structure, which are used to create predictive models for the evolution of complex systems. The aim of the work is to generalize the experience of constructing decision trees with a multidimensional response and to study the homogeneity and violation of symmetry of classes of models of socio-economic systems based on decision trees, which most clearly show the process of changing the states of the system and filling the space of possibilities, as well as signs of self-organization, which is cause of evolutionary processes and a consequence of symmetry breaking. An example of building a tree with a multidimensional response for a credit scoring problem is shown. The approaches described in the work show the connection between the phenomenon of symmetry breaking and the phenomenon of heteroscedasticity of regression models. The possibility of overcoming the problem of instability of finite predictions of models based on decision trees by developing approaches to the study of the heteroscedasticity of predictive models of socio-economic systems and the homogeneity of groups of objects is considered.**

***Keywords*—economy, information, clustering, regression, decision trees, heteroscedasticity, asymmetry**

I. Introduction

In the modern world, the requirements for professional skills and abilities of a person in the field of analysis, interpretation, and application in practice of digital assets and information thesauri are extremely rapidly becoming complicated and growing. The beginning of the third millennium is marked by two major events for Russia – an attempt to carry out a new stage of economic reforms and accelerated integration into the world economy [1].

The economic reforms of the 1990s led to the transformation of a centrally planned economy, where the state acted as the main owner and manager of production resources, into a market economy. The former system of funded distribution of material resources, guaranteed sales of products, and planned pricing has been liquidated. Many industrial and scientific-production associations ceased to exist, as well as the middle management level – industrial ministries and central administrations. Sectoral and regional automated systems for collecting and processing production and economic information turned out to be unnecessary.

In recent years, fundamental work on the new information economy has appeared. However, it should not be opposed to the "old" industrial economy serving the material needs of society. The infrastructure of the modern information society, to which Russia is also striving, today is no longer conceivable without the World Wide Web. Internet expansion leaves no chance for latecomers, the slightest delay can push them very, very far [2]. The post-industrial society of the beginning of the XXI century has the following main features:

- changes in the economic structure of the national economy, an increase in the share of the secondary and tertiary sectors, primarily the service sector, due to material production;
- growth of science intensity and constant updating and introduction of new technologies;
- informatization of society, development of telecommunications;
- the primary role of management, improvement of management of all aspects of the life of society;
- human priorities in education, training, business and social activity [3].

The lockdown taking place on our planet throughout 2020 has led to the fact that governments, the world's leading scientists, prominent public figures are looking for working ways and models to strengthen the immunity of economies in the post-coronavirus world.

Among the main tools for supporting the immune system of the world economy are the following:

- support for domestic consumer demand and support for government demand;
- support for the corporate sector, including tax measures and increasing the availability of financial resources for enterprises;
- activating the resources of the financial and budget systems to support economic growth.

Thus, the priorities of Germany are as follows: increasing demand, maintaining jobs, and ensuring economic stabilization, stimulating private investment and investment by local authorities, stimulating investment in the future.

UK Priorities: Investing in infrastructure development, developing electric transport, attracting investment in R&D dedicated to green technologies, supporting green industrial clusters, simplifying requirements for developers, simplifying the procedure for reassigning land and premises, strengthening the commonwealth, and stimulating economic growth in Scotland, Northern Ireland and Wales, development of transport infrastructure in these areas and the creation of new jobs in them.

Support measures in Canada are focused on the following: air travel, food inspection services, sports and heritage culture organizations, broadcasting industries, national museums in Canada, national arts centers, oil and gas sector, emission reduction grants, cleanup of ex-oil and gas wells, timber processing, agriculture farming, farming support, fishing and fish farming, research, development, support of the academic community.

China's main priorities are to create new jobs, launch major factories, export-oriented businesses, retail and service industries, support the most affected families financially, support small and medium-sized businesses that retain at least 80% of their employees, and countering volatility in prices for raw materials and agricultural products, ensuring the stability of production chains, localizing production in China, ensuring a stable social life, returning from isolation to an active social life.

Experts note that the Russian economy is 3 times underfunded, and therefore, like other countries, it is necessary to develop and implement its own strategy to strengthen the immune system of the economy, and in all its sectors [4, 5]. And in this regard, the creation of predictive models for the evolution of complex systems using the tool for constructing decision trees with a multidimensional response of the optimal structure is an urgent task.

Decision trees are a popular algorithm for data mining, description and modeling of socio-economic processes, and are actively used in practice both for classification problems and for forecasting problems. Algorithms based on decision trees make it possible to identify potentially possible patterns and relationships between individual components of a socio-economic system and predict new facts by assessing the value of the target feature y (response) for any object according to its description $X = (x_1, x_2, ...x_n)$ – a set of independent variables called predictors [6].

When creating predictive models, the main task is to predict the value of the target feature *y* based on the observed variation in the values of variables $x_1, x_2, ...x_n$, without examining the structure of internal relationships between the variables and/or a comparative assessment of the strength of their influence on the response.

However, in real conditions, the simulated processes have a sufficiently large or indefinite number of parameters, therefore, it is necessary to use a systematic approach to build predictive models, which is based on the principles: 1) the transition from the simplest options for describing the system to the most complex, when each feature describing the state of the system serves to obtain the best results – factor analysis; 2) modeling the evolution process, when a separate iteration of the refinement or fittin of the model is evaluated from the standpoint of utility and achievement of the result.

These principles allow describing the properties of self-organization of complex systems [7, 8], generalize approaches to their study, increase the accuracy and adequacy of predictive models. However, the main problem is that in the pursuit of accuracy, most models lose the most important characteristic – the interpretability of the results. Therefore, the development of highly interpretable models is an urgent task that can be solved using data analysis algorithms based on decision trees, the main advantage of which is the flexibility and interpretability of the analysis results. Another important advantage of decision trees is a low computational load when working with large amounts of data and features, high robustness to outliers, and the possibility of using them in dimensionality reduction problems. In addition, one of the most useful properties of decision trees is the ability to visually display the evolutionary process of fitting a model, which allows finding a relationship with the evolutionary processes of the system itself.

To date, there are several works [7, 9] linking the evolution of complex systems with the phenomenon of symmetry breaking in the context of random variability of the structure of interactions between elements (subsystems) in physical, biological, and socio-economic systems. Breaking the symmetry of complex systems is of particular importance in the study of early warning of financial crises and accounting for economic risks.

The aim of the work is to generalize the experience of constructing decision trees with a multidimensional response and to study the homogeneity and symmetry breaking of classes of models of socio-economic systems based on decision trees.

II. Current State of the Problem under Study

Research shows [10, 11] that physical, biological, and socio-economic systems are also characterized by the assessment of mutual dependencies between complexes of multivariate variables. In this case, the main task of building a model is to explain the variability of the multidimensional response $Y = (y_1, y_2, ...y_m)$. Such a response can be represented in the form of some relatively closed system of elements *N*, related to set $S = \{y\}$ of different types of these elements.

Using the classifier *S*, objects are divided into groups (classes), i.e. each group $y \in S$ corresponds to a subset *N(y)*, which determines the frequency of all occurrences of objects of this type in *N* [6].

A decision tree is a hierarchical structure in which each internal node denotes an attribute test using the *S* classifier, each branch represents the test result, and each leaf (terminal node) contains a class label. This construction allows flexible and evolutionary fitting of the tree-based model. However, such models can be significantly unstable. Small changes in the fitting data set can lead to significant changes in the tree structure, and in the end, to the final predictions. Decision tree fitting can create super-complex trees that do not generalize well from the fitting data (the effect of "overfitting").

The creation of optimal classification rules, and therefore sufficiently accurate and adequate predictive models, are based on two key ideas: the idea of recursive partitioning of the space of variables, when the *n*-dimensional space of variables is recursively divided into many non-intersecting regions – rectangles that refine the classification results on smaller groups of points and the idea of truncation (pruning), when the tree is reduced depending on the result of work on the test set. In the first case, the model uses the so-called "greedy" tree construction algorithm and is not resistant to overfitting; in the second case, it often converges on a local solution. In this regard, more complex ensemble methods have been developed, such as, for example, a random forest with subsequent data sampling for fitting the model (bagging) or stochastic gradient boosting to increase the adequacy of the developed models.

It should not be forgotten that any statistical averaging or simplification of the model negatively affects the interpretability of the work results, therefore, a comprehensive study of the feature space of the system objects is necessary.

Decision trees with multivariate response or Multivariate Regression Trees (MRT) are a model [12] that predicts response values, which is specified as a two-dimensional table containing several columns of observed features. When constructing a model, the main task is to determine the degree of influence of predictors on the total variability of quantitative relationships between individual components of the response. This allows us to conclude which factors are the most significant and determine the stability of the process.

MRTs are formed because of a recursive procedure for dividing rows of a data table into subsets, which is implemented using a set of external quantitative and/or categorical independent variables *X*. The "leaves" of the resulting tree are clusters of objects arranged in such a way as to minimize differences between points in a multidimensional space in within each population [6].

One can use, for example, the sum of squared deviations as a metric of the distance between classes (1):

$$SS_D = \sum_{ij} \left(y_{ij} - \overline{y_j} \right)^2 \qquad (1)$$

where y_{ij} – is response rate value *j* for observation *i*; *j*, $\overline{y_j}$ – the average values of this indicator for the cluster being formed, which includes the *i*-th observation.

The multidimensional classification procedure consists of a sequence of steps, at each of which the following actions are performed synchronously: first, a binary partitioning of objects into groups is performed due to the value of one of the independent variables, and then cross-validation and grouping of the response by each variable is performed. Do not forget about pre-scaling your data. Fig. 1 shows an example of visualizing a hierarchical tree structure representing a decision tree with a multidimensional response.

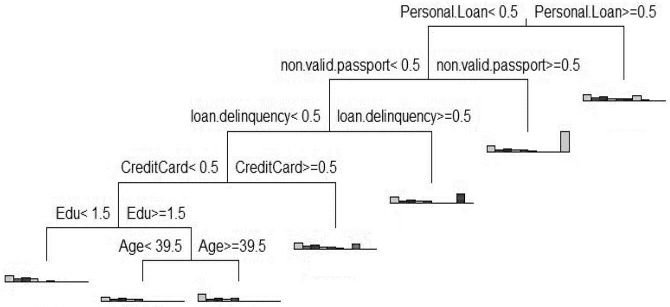


Fig. 1. An example of a decision tree with multidimensional response for a credit scoring problem

Fig. 1 shows an example of building a tree with a multidimensional response for the credit scoring problem (forecasting the fact of loan repayment/non-repayment based on the characteristics of a credit institution's customer). The nodes of the tree indicate the criteria that determine the value of the variables (age, income, education, etc.) for the binary classification, in the terminal nodes there are column diagrams showing the proportion of groups of points for each feature in each cluster. Fig. 1 shows that each cluster of points is composed with the dominance of certain characteristic features. In addition, it is possible to assess visually which features dominate in the formation of the entire set of clusters. Fitting of such a model is accompanied by refinement of criteria and creation of new rules by branching. In order not to "overfit" the model based on the decision tree, it is necessary to assess the consistency of the available data. It is necessary to estimate the magnitude of the prediction error on the number of tree nodes to do this (Fig. 2).

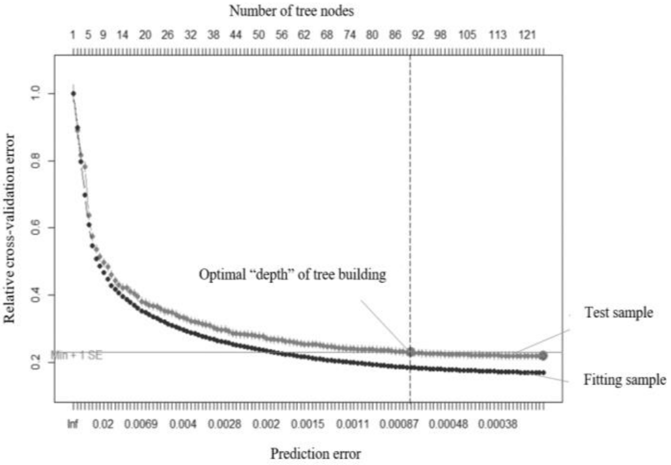


Fig. 2. Dependence of the classification error value on the tree learning depth

However, it is necessary to project data from multidimensional space onto a plane with axes of the first two principal components to show clearly how the factors influence the development of the process and to assess how large the heterogeneity between the selected clusters is. Fig. 3 shows an example of such a projection for the optimal fitting depth of the tree. The tree learning depth is determined by the number of nodes participating in the formation of refinement rules and classification criteria, as well as the minimum number of points in the class.

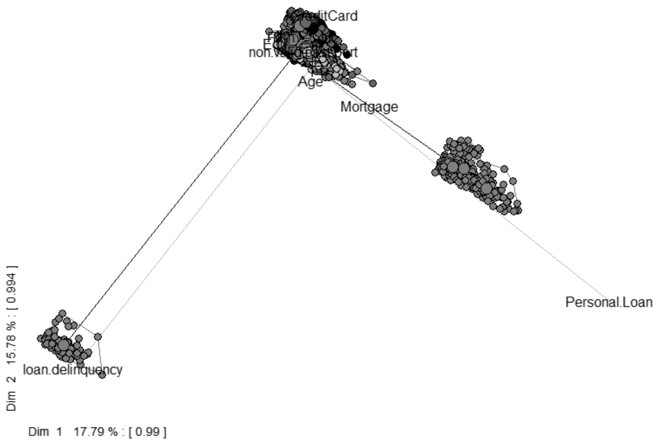


Fig. 3. Diagram for 30 groups of clients in the space of two main components

In Fig. 3, specific observations assigned to different clusters are highlighted in shades of gray and indicated by a contour drawn through the extreme points. In the centers of gravity of the areas of each of the data blocks, a larger circle is placed, denoting their centroid. The distances between the centers of the regions determine the degree of similarity of the points combined into clusters. Additional ordination axes are drawn from the center of the chart coordinates, the cosines of the angles between which correspond to the correlation coefficients between each pair of client groups. The projections of the points on each ordination axis determine the character of the indicator distribution over clusters with different external factors. Such a complex structure, shown in Fig. 3, also determines the nature of self-organization of the system and allows drawing a conclusion about the features of evolutionary processes, but for this it is necessary to turn to group theory and the principle of symmetry breaking.

In Fig. 3, specific observations assigned to different clusters are highlighted in shades of gray and indicated by a contour drawn through the extreme points. In the centers of gravity of the areas of each of the data blocks, a larger circle is placed, denoting their centroid. The distances between the centers of the regions determine the degree of similarity of the points combined into clusters. Additional ordination axes are drawn from the center of the chart coordinates, the cosines of the angles between which correspond to the correlation coefficients between each pair of client groups. The projections of the points on each ordination axis determine the character of the indicator distribution over clusters with different external factors. Such a complex structure, shown in Fig. 3, also determines the nature of self-organization of the system and allows drawing a conclusion about the features of evolutionary processes, but for this it is necessary to turn to group theory and the principle of symmetry breaking.

In this case, the main task is not only to determine the evolution vector, but also the limits of applicability of the developed model. Decision trees that form the spatial structures of connections of individual elements of the system, as one of the components of self-organization, have a relatively stable nature, they are recognizable both in the process of formation and in the form of an emerging organization.

The violation of the symmetry of structures is the reason for the heteroscedasticity [14, 15] of the model, when there is a non-constant variance of the prediction error on the initial data set, but with an increase in the number of variables. Heteroscedasticity in cross section data and in panel data (a set of characteristics of different objects collected at the same time) arises due to the fact that objects have different characteristics, and therefore the variance of errors for them will be different, and the more features are taken into account in the model, then the model is more complex and it is necessary to define areas of competence (model applicability). If the constructed model looks like (2):

$$Y = \alpha + X\beta + \varepsilon \, , \tag{2}$$

and the model consists of a fixed ($\alpha + X\beta$) and random (ε) part, then the model has heteroscedasticity if, contrary to the assumptions of the Gauss-Markov theorem, the variance of the random error is different for different observations, that is, $\exists\, i, j\ D(\varepsilon_i) \neq D(\varepsilon_j)$, where $D(\varepsilon_i) = \sigma_i^2$, then the covariance matrix (3):

$$\operatorname{cov}(\varepsilon) = \begin{vmatrix} \sigma_1^2 & & 0 \\ & \dots & \\ 0 & & \sigma_n^2 \end{vmatrix} \neq \sigma^2 \cdot \left(X^T X \right)^{-1}, \tag{3}$$

However, if the data are averaged over groups and the number of objects in the groups may differ, then the error variance for the *i*-th group will be equal to σ^2 / n_i, where n_i is the number of objects in the -th group *i = 1, 2, ..., N*, where *N* is the number of groups. The error covariance matrix for such data will have the form (4):

$$\operatorname{cov}(\varepsilon) = \begin{vmatrix} \sigma^2 / n_1 & & 0 \\ & \dots & \\ 0 & & \sigma_n^2 / n_N \end{vmatrix} . \tag{4}$$

Thus, heteroscedasticity is directly related both to the number of groups and points in a group, and to the characteristics of the homogeneity of the identified clusters. Large amounts of data make it possible to evaluate complex models of heteroscedasticity, which take into account not the symmetry of the influence of positive and negative influences (factors), but the fact that the response of the system to factors is not proportional. There is an approach to identify heteroscedasticity. The first approach considers the dependence of the error variance on a large number of factors, the second is based on checking the dependence of the error variance for each variable. The results of the second type of tests are easier to interpret, and the heteroscedasticity revealed with their help is easier to eliminate [15, 16]. This is the type of test for breaking the symmetry of the structure of the response diagram by groups of points of the decision tree.

Observation of symmetry breaking in decision trees with a multidimensional response allows estimating the symmetry of the system's response to disturbances with respect to variables.

This is expressed in the structure of the diagram for groups of objects in the space of the main components, and is estimated according to the following criteria:

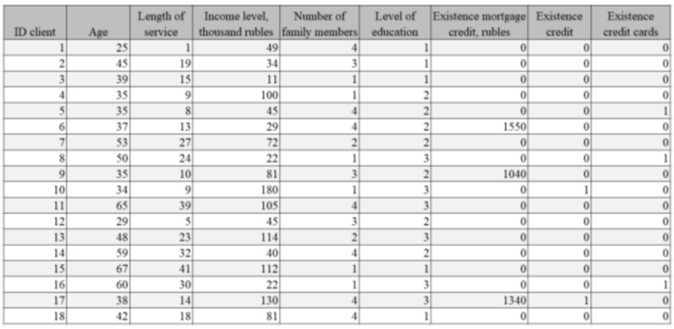
1) If the sets of objects on the diagram are compact and homogeneous, then the model is homoscedastic with respect to the set of factors, which indicates the homogeneity of the variance of the random error of the model.

2) If, on the contrary, heterogeneous groups of objects are observed, which, moreover, are separated into separate groups on a separate basis, this indicates the heteroscedasticity of the constructed model.

It should also be taken into account that the change in the states of the system occurs constantly and dynamically, therefore the variance will depend on its values in previous periods of time. Such a test allows determining visually the variables that have a non-linear response.

III. Results and Discussion

As an illustrative example, let us consider the above mentioned credit scoring problem. The response is specified in the form of a multi-level matrix of features (Fig. 4). The decision tree is formed as a result of dividing the rows of the data table into subsets, taking into account the minimization of differences in the feature vectors of each object. In this task, it is necessary to assess the solvency of individuals and answer the question: what are the key signs in determining the solvency of individuals and predict the solvency of individuals in relation to loan repayment/default on time.

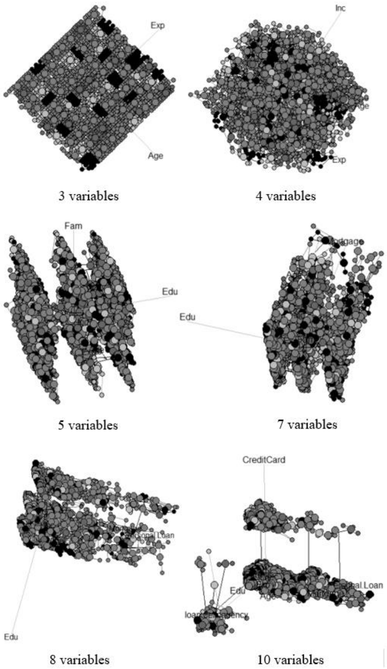


ID client	Age	Length of service	Income level, thousand rubles	Number of family members	Level of education	Existence mortgage credit, rubles	Existence credit	Existence credit cards
1	25	1	49	4	1	0	0	0
2	45	19	34	3	1	0	0	0
3	39	15	11	1	1	0	0	0
4	35	9	100	1	2	0	0	0
5	35	8	45	4	2	0	0	1
6	37	13	29	4	2	1550	0	0
7	53	27	72	2	2	0	0	0
8	50	24	22	1	3	0	0	1
9	35	10	81	3	2	1040	0	0
10	34	9	180	1	3	0	1	0
11	65	39	105	4	3	0	0	0
12	29	5	45	3	2	0	0	0
13	48	23	114	2	3	0	0	0
14	59	32	40	4	2	0	0	0
15	67	41	112	1	1	0	0	0
16	60	30	22	1	3	0	0	1
17	38	14	130	4	3	1340	1	0
18	42	18	81	4	1	0	0	0

Fig. 4. Signs of statistical sampling (part of the data is hidden)

Fig. 5 shows diagrams for different numbers of customer groups in the space of two main components. In this case, it is not the number of partition groups that is important, but the structure of the relationships of the resulting groups. This structure can be used to judge the nature of the system's response to disturbing influences with respect to variables. The diagrams show how great the heterogeneity between the identified clusters and the asymmetry of their structure are. Obviously, with an increase in the number of variables, the symmetry of the structure of the diagram is violated and the direction of evolution changes.

The overlap between points of different clusters in all cases is explained by the prediction error set value at the 10^{-4} level, as well as by the minimum cluster size equal to one data point. When the number of variables is small, symmetric structures with approximately the same cluster size are observed at the optimal fitting depth of the tree. With an increase in the number of variables, the optimal structure of the tree changes and some isolated groups of objects are distinguished, which are located symmetrically about one of the axes. With a further increase in the number of variables to 7-8, a violation of symmetry along this axis is observed, and then the isolation of individual asymmetric structures of the diagram.



3 variables

4 variables

5 variables

7 variables

8 variables

10 variables

Fig. 5. Diagrams for a different number of customer attributes, deployed in the space of two main components

IV. Conclusion

Decision trees are a flexible tool for building models for describing socio-economic processes, in turn, decision trees with a multidimensional response expand the possibilities of studying the space of attributes of system objects, allowing not only to study the behavioral activity of individual objects of the system and their groups (clusters), but also to develop recommendations for identifying predictors with a nonlinear response, allowing to build a more adequate predictive model.

The violation of symmetry with respect to the projections of the main components shows the directions of changes in the structure of the system, thus, the structure of the tree characterizes an asymmetric response to changes in the external environment. This approach makes it possible to assess the nature of the system's response to disturbing influences with respect to variables; therefore, even with frequent occurrence of the phenomenon of symmetry breaking and, as a consequence of the heteroscedasticity of one of the variables, one should not immediately try to eliminate it using known methods.

Acknowledgment

The study was supported by a grant from the Russian Foundation for Basic Research (RFBR) No 18.07.00170.

References

[1] N.M. Filimonova, N.V. Kapustina, V.V. Bezdenezhnykh, and N.A. Kobiashvili, “Trends in the sharing economy: bibliometric analysis”, Lecture Notes in Networks and Systems, 2020, vol. 87, pp. 145-154.

[2] M.M. Omarov, N.Y. Omarova, and D.L. Minin, “Territory branding

development as a regional economy activation factor", Lecture Notes in Networks and Systems, 2020, vol. 87, pp. 270-277.

[3] N.P. Kuznetsova, M.M. Omarov, and N.Y. Omarova, Information Economy [Informatsionnaya ekonomika], Veliky Novgorod: Yaroslav-the-Wise Novgorod State University, 2014, 135 p. (In Russ.).

[4] U.A. Dmitriev, and M.M. Omarov, "Increasing the efficiency of light and textile industry through the creation of regional clusters production", Proceedings of Higher Education Institutions. Textile Industry Technology, 2015, vol. 4(358), pp. 52-56. (In Russ.).

[5] T.N. Kashicina, E.S. Lovkova, and N.Yu. Omarova, "Import substitution of textile industry through innovation project management branch", Proceedings of Higher Education Institutions. Textile Industry Technology, 2015, vol. 4(358), pp. 203-207. (In Russ.).

[6] V.K. Shitikov, and S.E. Mastitsky, Classification, Regression and Other Data Mining Algorithms Using R [Klassifikatsiya, regressiya i drugiye algoritmy Data Mining s ispol'zovaniyem R], 2017, 351 p. (In Russ.) Retrieved from https://github.com/ranalytics/data-mining

[7] V.G. Rau, K.A. Gorshkov, S.V. Polyakov, T.F. Rau, A.N. Kislyakov, I.A. Togunov, and N.E. Tikhonyuk, Research of the theory of groups of broken symmetry in natural, biological and socio-economic systems [Issledovaniye teorii grupp narushennoy simmetrii v prirodnykh, biologicheskikh i sotsial'no-ekonomicheskikh sistemakh], Vladimir, Vladimir branch of RANEPA, 2020, 261 p. (In Russ.).

[8] M.A. Deryabina, "Theoretical and methodological foundations of self-organization of socio-economic systems", Voprosy Ekonomiki, 2019, vol. 7, pp. 73-94. (In Russ.).

[9] G. Nicolis, and I. Prigozhin, Knowledge of the Complex. Introduction [Poznaniye slozhnogo. Vvedeniye], Moscow: Book on Demand, 2012, 345 p. (In Russ.).

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer, 2013.

[11] O.P. Ivanova, V.A. Trifonov, and D.N. Nesteruk, "Directions and possibilities of predictive analytics in managing the development of single-industry towns", Espacios, 2019, vol. 40(3), p. 04.

[12] G. De'ath, "Multivariate Regression Trees: A New Technique for Modeling Species-Environment Relationships", Ecology, 2002, vol. 83(4), pp. 1105-1117. DOI: 10.2307/3071917

[13] A.N. Kislyakov, "Asymmetry of information in the analysis of socio-economic processes", Vestnik NSUEM, 2020, vol. 1, pp. 64-75. (In Russ.).

[14] W. Ruth, and T. Loughin, The Effect of Heteroscedasticity on Regression Trees, Cornel Univesity, 2013.

[15] J.M. Wooldridge, Introductory Econometrics: A Modern Approach, South-Western Pub., 2004. 910 p.

[16] S.J. Grossman, and J.E. Stiglitz, "On the Impossibility of Informationally Efficient Markets", American Economic Review, 1980, vol. 70(3), pp. 393-408.