Inductive technology of the target clusterization of enterprise's economic indicators of Ukraine

Irina Lurie [1[0000-0001-8915-728X]], Andrii Podlevskyi[2[0000-0002-3166-7487]], Natalia Savina[2[0000-0001-8339-1219]], Maria Voronenko[1[0000-0002-5392-5125]], Anna Pashnina[2[000-0002-1425-1615]], Volodymyr Lytvynenko[1[0000-0002-1536-5542]]

[1]Kherson National Technical University, Department of Informatics & Computing Technology, Kherson, Ukraine
E-mail: lurieira@gmail.com immun56@gmail.com, mary_voronenko@i.ua
[2]National University of Water and Environmental Engineering, 11, Soborna Street, Rivne, Ukraine, 33000
E-mail: a.a.podlevskyi@nuwm.edu.ua,n.b.savina@nuwm.edu.ua

Abstract. The article presents an inductive model of objects clustering economic indicators based on the method of arguments group accounting. The basic principles of creating objective clustering inductive model are formed, the ways and prospects for the possible model implementation are shown, the advantages of an objective clustering model compared to traditional data clustering methods are defined.

Keywords. inductive modeling, economic objects clustering, method of arguments group accounting, k-means algorithm, external balance criterion.

1 Introduction

Relevance of the work. The economic objects clustering is currently receiving much attention. This is primarily due to the increased requirements for the accuracy of the recognition and identification systems under various conditions for obtaining information. Currently, there are a large number of diverse clustering algorithms for economic entities, each of which has its own advantages and disadvantages and is focused on a specific data type. One of the existing clustering algorithm drawbacks is their subjectivity, i.e. getting good results of clustering objects on one set does not guarantee to get similar results on another similar set. One of the ways to improve the clustering objectivity is the development of hybrid models based on the method of complex systems inductive modeling, which is a logical continuation of the method of group accounting of arguments (MGUA) [1-3]. In this regard, the development of hybrid models and clustering objects methods based on the methods of complex systems inductive modeling is an important problem from both theoretical and practical points of view.

2 Literature review

The use of a cluster approach is of particular importance for the Ukrainian enterprise's economic indicators clustering. Cluster policy is aimed at combining the capabilities, knowledge, and capacities of structures number with the aim of solving joint and private tasks. The immediate results from solving such problems will form the basis for the economic, social and technological development of the region.

Cluster policy is aimed at combining the capabilities of a structures number in order to solve joint and private tasks. The immediate results from solving such problems will form the basis for the economic, social and technological development of the region.

The theory and methodology of creating economic clusters in the regions are reflected in the writings of many authors. In scientific studies, it is noted that the successful development of the national economy depends on the development of the local concentration of specialized industries (industrial districts), which are the basis of the cluster approach [4]. This provision was first described in [5], where the synergistic effect of a merger of enterprises was first identified and analyzed.

In [6], such areas of knowledge as the new economic geography, business research of firms, regional studies and innovations that influence the development of cluster theory in the economy were identified. In [7], the authors proposed three clustering models: classical agglomeration models, industrial complex models, and network interaction models. In work [4], such five key concepts as externalities, innovative environment, interfirm struggle, competition of cooperation, dependency path, which constitute cluster theory, were considered. The authors of this work also focus on the geographical concentration and specialization of enterprises, the diversity of cluster participants, the critical mass and life cycle of the cluster, innovation and competition.

In [8], the authors succeeded in systematizing the extensive theoretical and empirical accumulated earlier, where the advantages of using national competitive relations in the economy were shown.

The authors of [9] believe that the cluster approach in the economy represents the synthesis of several areas, including local industrial specialization, spatial economic agglomeration, and regional development, as well as the provisions of strategic and venture management.

The diversity of the cluster theory indicates the absence of the only correct approach to its practical operationalization and makes it relevant to use the cluster approach for clustering the economic indicators of Ukrainian enterprises.

This study proposes inductive models of clustering objects of the economy to justify their creation at the regional level. The basic concepts of creating an inductive model of clustering objects based on the method of group accounting of arguments are described in [1-3, 10] and further developed in [11-16]. The authors formulated the basic principles of creating an inductive model of objective clustering, showed ways and prospects for a possible implementation of the model, determined the advantages of the model of objective clustering compared with traditional methods of data clustering. However, it should be noted that, despite the achievements achieved in this subject area, the inductive model of objective clustering currently has no practical implementation.

In the paper [16, 17] an inductive model of objective clustering of objects based on the k-medium clustering algorithm was developed, an estimation of the stability of the model to the noise component was made, ways of further improvement of the proposed model with purpose of increasing the objectivity of the clustering of the studied data. Approbation of the work of the proposed model was performed using the data “Compound” and “Aggregation” of the database of the computer school of the East Finnish University. It is presented studies to assess the stability of the model to the noise component using data "Seeds".

In this paper, a more improved version of the k-means inductive clustering algorithm is used. In contrast to [16], the centers of masses of the clusters are not calculated, and the Silhouette, Entropy, Dunn’s index and Calinski-Harabasz index are used as an internal criterion for the quality of clustering (in [16] uses only the Calinski-Harabasz index).

3 Formal problem statement

The unresolved parts of a common problem include:

- efficient algorithms lack for extracting equal-power subsets from the initial data set;
- lack of research on the influence of external and internal criteria on the clustering quality;
- insufficient implementation of the objective clustering inductive model in various areas of society, especially economic.

The aim of the article is to develop and study the influence of internal and external criteria on the quality of the objective clustering inductive model of objects based on the k-means clustering algorithm in the study of the enterprise's different types economic data in Ukraine.

4 Method Description

Let $A = \{x_{ij}\}, i = 1,...,n; j = 1,...,m$ – the matrix of the objects features under study, where n is the number of rows or observed objects, m is the number of features characterizing the object. The clustering task is reduced to splitting the set of objects into non-empty subsets of disjoint clusters, and the plane separating the clusters can take any form:

$$\begin{aligned} &K = \{K_s\}, s = 1,...,k; K_1 \cup K_2 \cup ... \cup K_k = A; \\ &K_i \cap K_j = \varnothing, i \neq j; i, j = 1,...,k \end{aligned} \quad (1)$$

where k – the number of clusters.

The methodology of complex systems inductive modeling is based on three fundamental principles borrowed from various scientific fields:

1. The principle of heuristic self-organization, the i.e. search of applicants various models with a choice of the best from the point of view of the relativity external criterion, the value of which is determined on two equally powerful data sets;

2. the external addition principle, the idea of which is the need for objective verification of the model using additional "fresh" information;
3. The inconclusive decisions principle, i.e. generation of certain solutions set with the optimal variant subsequent choice.

The implementation of these principles in the framework of an inductive objective clustering model implies the following steps:

• normalization of the studied objects signs, i.e. bringing them to the same range with the same median of the attribute space attributes;

• splitting the original objects set into two equally powerful submultiples;

• determination of an external criterion or relevance criteria group for choosing the optimal clustering on two equally powerful subsets;

• selection or development of a basic clustering algorithm used as a component in an objective clustering inductive model of objects.

Data normalization was performed according to the signs in accordance with the formula:

$$x'_{ij} = \frac{x_{ij} - med_j}{\max\left(\left|x_{ij} - med_j\right|\right)} \tag{2}$$

where x_{ij} – the value of attribute i in the column j, x'_{ij} – normalized value of this feature, med_j – column median j. The choice of this normalization method was determined by the fact that as a result, the data features set in all columns had the same median with a maximum variation attributes range from -1 to 1, while the data volume for each column falling into the inter-quantile distance (50%) is the largest compared to other normalization methods.

Algorithm for the separation of the original objects set Ω to 2 equipotent disjoint subsets Ω^A and Ω^B consists of the following steps [13]:

1. calculation $n\cdot(n-1)$ pairwise distances between objects in the original data sample;

2. selection of objects pair X_s, X_p, the distance between which is minimal:

$$d\left(X_s, X_p\right) = \min_{i,j} d\left(X_i, X_j\right) \tag{3}$$

3. object distribution X_s into a subset Ω^A, and object X_p into a subset Ω^B;

4. repetition steps 2–3 for the remaining objects. If the objects number is odd, the last object is distributed into both subsets.

As internal criteria (IC) for the quality of clustering used:

1. Silhouette [18]

$$SWC = \frac{1}{K}\sum_{j=1}^{K} S_{x_j} \tag{4}$$

where K - number of clusters, S_{x_j} - the "best" element belonging x_j to cluster p.

The best partition is characterized by the maximum SWC, which is achieved when the distance inside the cluster is small, and the distance between the elements of the neighboring clusters is large.

2. Dunn's index [19]

Compares intercluster distance with cluster diameter. The higher the index value, the better the clustering.

$$DI\left(k\right) = \min_{i \in k} \tag{5}$$

3. Calinski – Harabasz index [20]

$$QC_{CH} = \frac{QCB \cdot \left(N - K\right)}{QCW \cdot \left(K - 1\right)} \to \max \tag{6}$$

where N - number of objects, K - number of clusters. The maximum index value corresponds to the optimal cluster structure.

4. Entropy [21]

$$\left[PE = \frac{\sum_{q=1}^{Q} \sum_{k=1}^{K} \ln\left(u_{qk}\right)}{Q}, PE \in \left[0, \ln Kk\right] \right] \tag{7}$$

Entropy is known as the numerical expression of the system orderliness. The entropy of the partition reaches a minimum with the highest orderliness in the system (in the case of a clear partition, the entropy is zero). That is, the greater the belonging degree of an element to one cluster (and the smaller the belonging degree to all other clusters), the smaller the entropy value and the more qualitatively clustering is performed. The main disadvantage of these methods is that their computation becomes more and more complex, both with an increase in the clusters number k and with an increase in the objects number included in the data. To calculate the external criterion of balance, the approach taken from [16] was taken as a basis. In this paper, the external criterion of balance (ECB) of controlled clustering is defined as the normalized optimal value of the sum of squared deviations between the values of the internal criteria of clustering quality (4) – (7):

$$ECB = \sqrt{\frac{\left(IC_A - IC_B\right)^2}{\left(IC_A + IC_B\right)^2}} \to opt \tag{8}$$

To create equal clustering conditions on subsets and using the k-average clustering algorithm, an initial number of clusters is determined at the initialization stage, and the initial value of criterion (8) is zero. The experiment showed that at subsequent iterations the criterion value at the first step increases, then monotonously changes until it reaches saturation, which corresponds to stable clustering on two equally powerful subsets. The block diagram of the inductive clustering model based on the k-means algorithm is shown in Fig. 1.

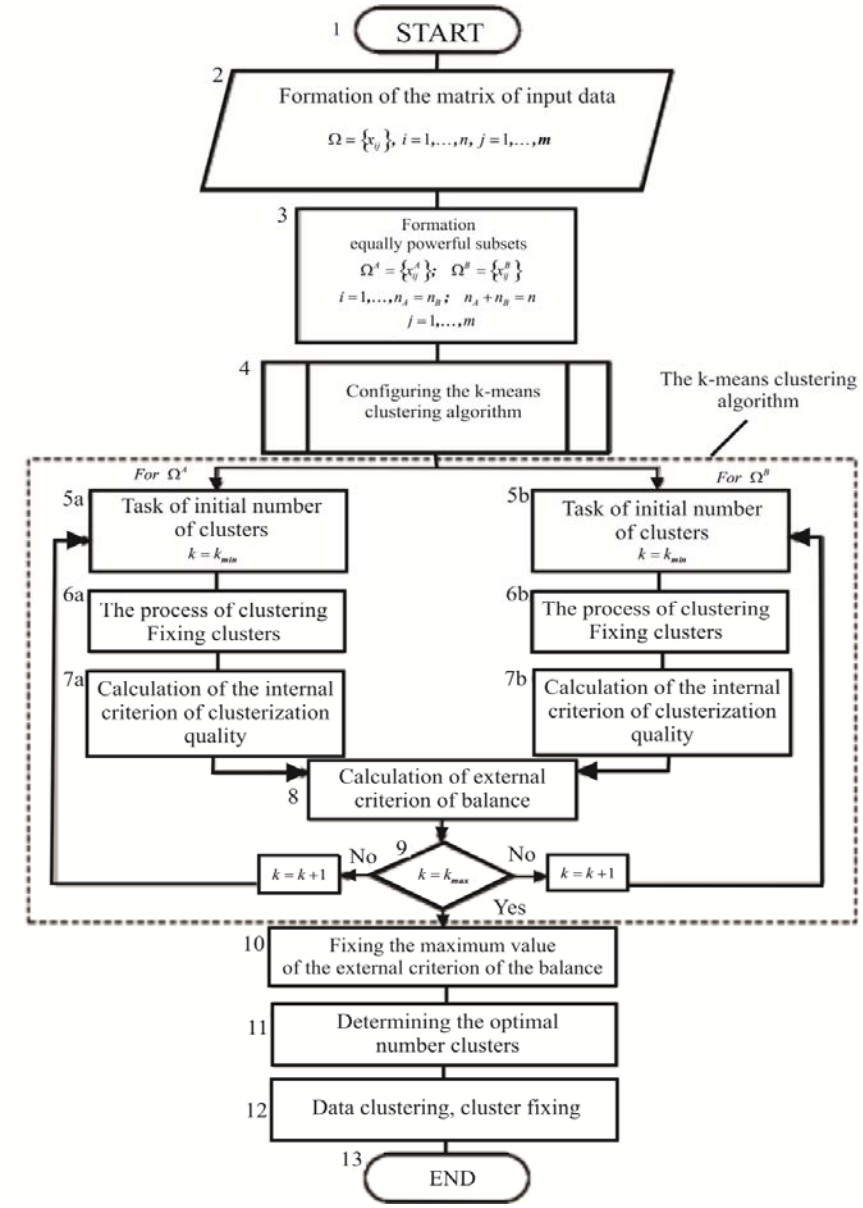

Fig. 1. Block diagram of an inductive model of objective clustering based on the k-means algorithm

The implementation of the algorithm requires the following steps:

Step 1. Start

Step 2. Formation of the initial set Ω of objects under study. Data preprocessing (filtering, normalization, analysis for missing values). Representation of data in the

form of a matrix $n \times m$, where n - the number of rows or the number of objects studied, m - the number of columns or the number of signs characterizing the objects.

Step 3. The division Ω into two equally powerful subsets in accordance with the above algorithm. The resulting subsets Ω^A and Ω^B formally can be represented as follows:

$$\Omega^A = \left\{x_{ij}^A\right\}, \Omega^B = \left\{x_{ij}^B\right\}, j = 1,...,m$$
$$i = 1,..., n_A = n_B, n_A + n_B = n \quad (9)$$

Step 4. Configure the k-means clustering algorithm.
For each equally powerful subset:
Step 5. Select the number of clusters.
Step 6. Sequential clustering and cluster fixing
Step 7. Calculation of the internal criterion of clustering quality.
Step 8. Calculation of the external balance criterion in accordance with formula (8).
Step 9. If the value of the balance criterion reaches the optimum, then:
Step 10 Fixation of the received clustering is performed.
otherwise, the number of clusters is increased by 1 and repeated Step 5-9
Step 11. Determining the optimal number of clusters.
Step 12. Clustering data (sets Ω of objects under study), fixing clusters.
Step 13. End

5 Characteristics of the data used

5.1 Analysis of the ratio of small, medium and large enterprises, and their importance for economic development in the regions of Ukraine

For Ukraine, unbalanced, resource-intensive, with significant territorial-branch disproportions is a model of a national economic complex that requires significant financial and organizational efforts to remedy the situation. Small, medium and large enterprises play an important role in the Ukrainian economy and have well-known advantages and disadvantages. To strengthen the benefits and reduce the impact of shortcomings for each group of companies in the context of accelerated socio-economic national development can be considered a system of production cooperation state regulation, which has its own characteristics in the sectoral and regional dimension.

Therefore, it is important to study the main indicators of large, medium and small businesses of Ukraine with the help of cluster analysis tools, which will reveal certain differences in the functioning of entrepreneurship in a regional dimension and offer more effective and flexible mechanisms of regions socio-economic development.

The matrix of the objects under study contained 26 rows (objects) and 18 columns (signs characterizing the objects). In accordance with the goals of the problem to be solved, clustering was carried out according to signs, i.e. after transformation, the data matrix under investigation had a dimension of 18 × 26. When dividing this set into

two closed subsets in accordance with the algorithm described above, we get two subsets with dimensions 9 × 13 (9).

6 Clustering Results

Using the NbClust package from the programming language and the environment for developing, analyzing data and statistical calculations R, we construct a diagram that shows the dependence of the clusters number on the NbClust criteria (Fig. 2).

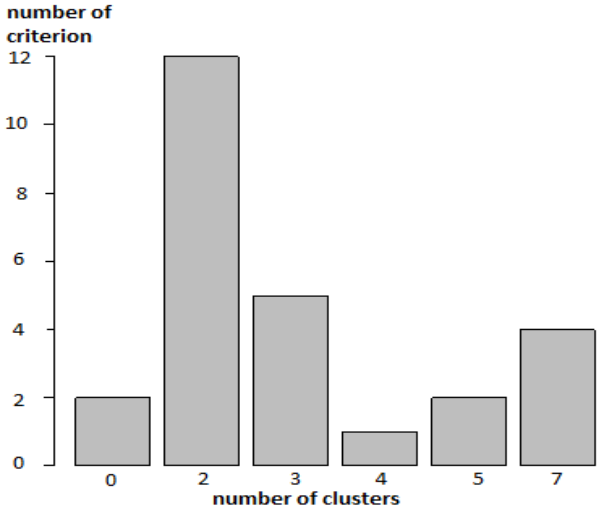

Fig. 2. Cluster number diagram for the k-means clustering algorithm

Analysis of the chart allows us to conclude that, from the point of view of the criteria used, it is optimal to divide the test signs into two, three or seven groups. However, since, in accordance with the goals set, the division of the features set into a large clusters number is inexpedient, we consider the most optimal clustering when dividing objects into three groups. The results of the work of objective clustering inductive models based on the above algorithm are presented in Table 1. Analyzing the values of the criteria, it is clear that splitting into three clusters shows the best results for the values of Silhouette, Dunn index and Calinski-Harabasz index. The minimum Entropy value is achieved when split into two clusters.

Table 1. The results of the work of objective clustering inductive models

Internal criterion of clustering quality	Number of clusters		
	2	3	7
	Exterior balance creature	Exterior balance creature	Exterior balance creature
Silhouette	0,0243161	0,1375	0,0766
Dunn;s index	0,15297984	0,3497	0,0369
Calinski-Harabasz index	0,01035773	0,0204	0,0175
Entropy	0	0,2115	0,0353

Table 2 shows the results of clustering into two, three and seven clusters. As can be seen from the table, the divisions into three and seven clusters distinguish the economic indicators of Kiev city into a separate cluster. Given that Kyiv is the capital of Ukraine not only in the administrative-territorial field, but also in the international and economic, in most cases it is clustered with clustering as it is in a cluster, because in most economic indicators of the enterprises it is significantly ahead of the nearest cities-"competitors" (Kharkiv, Dnipro, Odessa, Lviv, etc.) and the whole area. The remaining groups are formed, taking into account the economic development and territorial regions location.

7 Economic interpretation of the results

2-cluster model. The results of such clustering can clearly identify 2 clusters with clear socio-economic characteristics. The first cluster contains Ukraine regions, which traditionally are centers of economic Ukraine macro-regions (Dnipropetrovsk, Donetsk, Lviv, Odessa, Kharkiv, Kyiv and Kyiv). This cluster can also be positioned as industrial since in these regions, the main industries are concentrated and the historically concentrated largest share of large enterprises, which also interact with a significant number of small and medium enterprises. Accordingly, this cluster is more productive in socio-economic terms.

The second cluster focuses on the remaining Ukraine regions with greater economic diversification of enterprises various types (mostly medium and small) and relatively lower economic indicators of regional development.

3-cluster model. The main results are duplicated by the above-mentioned model, the main difference is the isolation of the Ukraine capital -- the city of Kiev - as a separate cluster, given its significant socio-economic potential.

7-cluster model. In our opinion, this model is more relevant in interpreting the impact of enterprises different types on the Ukrainian regions economic development. To a large extent, such a division is correlated with studies of Ukraine regions by the integrated indicator of specialization [19,20]. In the first cluster, only Kyiv is represented, because according to the main indicators of social and economic development, and as the capital of Ukraine, it is positioned as a separate cluster.

Table 2. Clustering results for the k-means algorithm

Number of clusters		2				3				7			
Cluster number		1	2	1	2	3	1	2	3	4	5	6	7
Region	Vinnyts'ka	+		+				+					
	Volyns'ka	+		+									+
	Dnipropetrovs'ka		+		+						+		
	Donetska		+		+						+		
	Zhytomyrs'ka	+		+					+				
	Zakarpats'ka	+		+								+	
	Zaporiz'ka	+		+				+					
	Ivano-frankivs'ka	+		+								+	
	Kievs'ka		+		+						+		
	Kirovohrads'ka	+		+								+	
	Luhans'ka	+		+								+	
	L'vivs'ka		+		+					+			
	Mykolayivs'ka	+		+								+	
	Odes'ka		+		+					+			
	Poltavs'ka	+		+				+					
	Rivnens'ka	+		+					+				
	Sums'ka	+		+					+				
	Ternopils'ka	+		+									+
	Kharkivs'ka		+		+					+			
	Khersons'ka	+		+								+	
	Khmelnyts'ka	+		+					+				
	Cherkas'ka	+		+				+					
	Chernivets'ka	+		+					+				
	Chernivhivs'ka	+		+								+	
	city of Kiev		+			+	+						

The second cluster (Vinnitsa, Zaporozhia, Poltava, Cherkassk) contains fairly balanced enterprises (small, medium and large) and their indicators of the region, and is characterized mainly by industrial and agricultural profile.

The third cluster (Zhytomyr, Rivne, Sumy, Khmelnytsky, Chernivtsi) has a more pronounced agrarian profile with significant influence of the economic potential of small and medium enterprises.

The fourth cluster (Lviv, Odessa, Kharkiv) is represented primarily by regions in which all types of enterprises are harmoniously represented. In addition, significant developments in these areas of information industries and high-tech industries can be

noted, indicating their informational and post-industrial profile.

The fifth cluster (Dnipropetrovsk, Donetsk, Kiev) has the largest number of large enterprises (from 45 to 61) and is characterized by significant industrial potential.

The sixth cluster (Zakarpattia, Ivano-Frankivsk, Kirovograd, Luhansk, Mykolayiv, Kherson, Chernihiv) contains areas with a well-developed small and medium business and a widespread service area. They can be attributed equally to the industrial and agricultural profile, and to the agrarian-industrial.

The seventh cluster (Volyn, Ternopil) is represented by regions with lower socio-economic development and agronomic indicators than the average in Ukraine and the dominance of small and medium-sized enterprises of a non-industrial type.

8 Conclusions

The results of the model work showed the high efficiency of the developed clustering models on the novel inductive method of simulation of complex systems.

In this paper, the k-medium algorithm was used as the basic algorithm, while the effect of four internal criteria (Silhouette, Dunn's index, Calinski-Harabasz index, Entropy) on the quality of clustering was studied. This choice was determined by the simplicity of its implementation. The advantage of the proposed model lies in its stability, which is determined by using an external balance criterion on two coherent samples.

It should be noted that the use of inductive simulation methods does not eliminate the main disadvantage of the k-mean algorithm: the result of clustering depends on the selection of the source centers of the clusters, but with other things being equal, the proposed model gave better clustering results compared to the traditional k-medium algorithm implemented in the software environment R.

Interpreting the results of clustering using an inductive model showed a significant effect of using such a methodology to identify the degree of influence of small, medium and large enterprises on the socio-economic development of regions, which is largely confirmed by the results of other studies. However, some of the detected clusters (for instance 6) contain ambiguous characteristics that do not allow them to be clearly interpreted. This indicates the need to use a larger array of output data or to combine this technique with other approaches to provide the most rational and reliable result that could be the subject of further scientific research.

References

1. Ivakhnenko, A.G.: Group method of data handling – a competitor of stochastic approximation method // Automatics, 1968. - №3. - P. 58-72. [In Ukraine] (1968)
2. Ivakhnenko, A.G.: Inductive method for self-organization of complex systems models.– Kiev: Scientific Thought, 1982.– 296 p. [In Russian] (1982)
3. Ivakhnenko, A.G.: Objective self-organization based on the theory of self-organization models // Automatics, 1987. - №5. - P. 6-15. [In Russian] (1987)
4. Bergman, E.M., Feser, E.J.: Industrial and Regional Clusters: Concepts and Comparative Applications, Regional Research Institute, WVU (1999)
5. Marshal, A. : Principles of Economics. London: Macmillan (1999)
6. Humphrey J., Schmitz H.: Governance and upgrading: linking industrial clusters and global value chain research. IDS Working Paper 120. Institute of Development Studies (2000)

7. Gordon, I. R., McCann, P. : Industrial Clusters: Complexes, Agglomeration and/or Social Networks? *Urban Studies*, *37*(3), pp. 513–532 (2000) https://doi.org/10.1080/0042098002096.
8. Andersson T., Schwaag-Serger, S., Sorvik, J.: Emily Wise Hansson. The Cluster Policies Whitebook, IKED (2004)
9. Porter, M. E. :The Competitive Advantage of Nations. New York: Free Press, 1990. (Republished with a new introduction, 1998.) (1990)
10. Madala, H.R., Ivakhnenko, A.G.: Inductive Learning Algorithms for Complex Systems Modeling.− CRC Press, 1994. – 365 p (1994)
11. Stepashko, V.S.: Theoretical aspects of GMDH as a method of inductive modeling // Managing Systems and Machines, 2003. - №2. - P. 31-38. [In Russian] (2003)
12. Stepashko, V.S.: Elements of the inductive modeling theory / State and prospects of informatics development in Ukraine: Monograph / Team of authors. - Kiev: Scientific Thought, 2010. - 1008 p. - P. 471-486. [In Ukraine] (2010)
13. Osypenko, V.V.: Two approaches to solving the problem of clustering in the broad sense from the standpoint of inductive modeling // Power and Automation, 2014. - №1. - P.83-97. [In Ukraine](2014)
14. Osypenko, V.V., Reshetjuk, V.M.: The Methodology of Inductive System Analysis as a Tool of Engineering Researches Analytical Planning / V.V. Osypenko, V.M. Reshetjuk // Ann. Warsaw Univ. Life Sci. – SGGW. – № 58, 2011. – P. 67-71 (2011)
15. Sarycheva, L.V.: Objective cluster analysis of the data on the basis of the Group Method of Data Handling // Problem of Management and Informatics, 2008. - №2. - P. 86-104. [In Russian] (2008)
16. Babichev, S., Lytvynenko, V., Taif M.: Estimation of the inductive model of objects clustering stability based on the *k*-means algorithm for different evels of data noise // Radio electronics, informatics, management. - Zaporizhzhya, ZNTU. – 2016. - №4, C. 55-60 (2016)
17. Reid, R., Schein, S., Wilson, H.: Industrial Clusters in Jackson and Josephine Counties. – School of Business Southern Oregon University, November 30, 2006. URL: http://www.ashland.or.us/files/SOU_IndustrialClustersReport.pdf (2006)
18. Kaufman, L., Rousseeuw, P. :Finding Groups in Data. An Introduction to Cluster Analysis. Wiley (2005)
19. Bezdek, J.C., Dunn, J.C.: Optimal fuzzy partitions: A heuristic for estimating the parameters in a mixture of normal dustrubutions // IEEE Transactions on Computers, 835–838pp (1975)
20. Calinski, R.B., Harabasz, J.: A dendrite method for cluster analysis // Comm. in Statistics, 3:1.27 (1974)
21. Sripada, S. Ch., Rao, M. S.: Comparison of purity and entropy of k-means clustering and fuzzy c means clustering , Indian journal of computer science and engineering; Vol 2 no.3 June 2011; ISSN:0976-5166 (2011)