24. RESEARCH OF THE ECONOMIC DEVELOPMENT STRUCTURE OF THE REGIONS OF UKRAINE WITH THE METHODS OF MULTI-DIMENSIONAL STATISTICAL ANALYSIS

Gybkina N., Sidorov M., Storozhenko O.

The classification problem of the regions of Ukraine on the main economic indicators in 2012 and 2017 is considered. The principal component method and cluster analysis methods were used for solving the problem. The obtained results graphically display the position of the regions of Ukraine on the plane of the first two main components grouped in selected clusters.

The economic situation of the regions of Ukraine. The variety of natural conditions, historically developed directions of development of certain areas and their economic orientation, which developed at the beginning of the 90-ies of the twentieth century, the difficulties associated with the shortcomings of planning and management of the political and economic system in the field for less than 30 years of independence of Ukraine have led to a significant imbalance in the levels of economic development of individual regions, which, in turn, significantly hinders the development of the whole country and requires constant increased attention to the allocation of funds and organization of management both locally and at the state level. In order to facilitate this task, it is necessary to clearly classify the individual regions in order to distinguish among them regions that are similar in economic terms. The problem here may be that funding in the social and economic spheres may require the distribution of funds between individuals in unequal proportions, since the level of development of the same entity may differ significantly in different directions.

The economy of the country and its individual parts is described by a large number of indicators, the values of which depend on the peculiarities of the geographical location and level of development and are constantly changing over time. Therefore, the issue of classification of economic objects is one of the tasks that can be considered as an initial issue when developing short and long-term programs for managing the economic development of the country. A comprehensive study of the issue in recent years has received increasing attention. In [9, 16, 21] the classification of regions of Ukraine according to the direction of their economic development is given, taking into account historical peculiarities and real existing natural conditions. Accordingly, four macro-regions are usually distinguished: Kyiv, industrial regions of Eastern Ukraine, coastal regions of the South, agrarian regions of the West and the Center.

In addition to using geographic considerations to solve the classification problem, you can use multidimensional statistical analysis methods, such as cluster analysis [1, 5, 8, 12, 14, 15]. In [10], the clustering of regions of Ukraine by the complete set of socio-economic features (41 socio-economic indicators) was performed, which allowed dividing the set of regions into eight clusters. The Kohonen network algorithm was used for clustering. On the basis of the results obtained in [10] in [11], an applied estimation of the growth potential of individual socio-economic indicators of the regions of Ukraine was carried out. The issue of grouping regions of Ukraine by the level of economic security is devoted to the work [3]. For classification, the author uses groups of socio-economic indicators characterizing socio-demographic, resource-energy, ecological-economic and financial-economic security of each region of Ukraine, and applies the k-means method. In [7], it is proposed to use the results of the cluster analysis to improve the budgeting potential of Ukrainian regions. The work [13] is devoted to the use of cluster analysis to assess the development of small business in the regions of Ukraine. It calculates a regional small business development index for each oblast, grouping the oblasts into separate clusters. In article [17], the different cluster analysis methods (hierarchical and iterative) to cluster regions by the level of economic potential are used. Three clusters were identified, the first of which consisted of Dnipropetrovsk, Donetsk, Zaporizhia, Kyiv, Lviv, Odesa and Kharkiv regions, the second cluster consists of Kyiv, and the third contains the rest of Ukraine's regions. The work [6] is devoted to the clustering of regions of Ukraine in order to identify disproportions in ensuring their financial security.

From the conducted review of works, it can be concluded that the authors use different groups of socio-economic indicators for classification, based on the purpose of their research. The desire to take into account as many indicators as possible in the selection of groups can lead to computational complications despite the use of electronic computing facilities. Therefore, it is promising to use cluster analysis methods in combination with dimension-reduction methods that will significantly reduce the amount of input information and present the results in an obvious way. So, let's look at methods that allow us to formalize and solve this problem.

Basic approaches to dimensionality reduction and classification of multidimensional observations. The economic status of an entity (business, city, region, country, etc.) can be described by a large number of indicators, different in nature, origin, and possibly stochastically related. As a result, the direct analysis and comparison (classification) of objects by all indicators is impossible and the additional task of reducing the dimensionality of many characteristics of the object. These problems can be solved with modern data analysis methods, including component and cluster analysis [1, 2, 5, 8, 12, 14, 15].

Component analysis (principal component method) allows for the formation of several integral indicators over the whole set of indicators (usually two indicators for the sake of clarity of interpretation of the results), which thus solves the problem of reducing the dimensionality of the initial set of indicators, and such compression of information ensures its minimum compression. This method is based on the study of the structure of a covariance (or correlation) matrix of a system of random variables and allows to distinguish linear combinations of initial features with the largest contribution to the total variance. Among the advantages of component analysis, it should be noted that no statistical assumptions are made to realize the probabilistic distributions of the system of economic indicators under consideration for its implementation.

Cluster analysis, in turn, solves the problem of splitting a given set of objects into groups (clusters) consisting of similar objects. The output for cluster analysis is presented as, for example, a matrix of distances between objects in the selected metric. The most common methods of cluster analysis are probabilistic (k-means, k-median, etc.) and hierarchical (agglomerative and di-visa algorithms) approaches.

Component and cluster analysis procedures and their application to solving the object classification problem. Consider the scheme of the method of principal components [1, 2, 5]. Let the economic system consist of n objects, each characterized by a vector of p metrics $x^{(1)}, x^{(2)}, ..., x^{(p)}$. Then the result of observing the system will be an array consisting of n p-dimensional vectors

$$X_1 = (x_1^{(1)}, x_1^{(2)}, ..., x_1^{(p)})^T, \; X_2 = (x_2^{(1)}, x_2^{(2)}, ..., x_2^{(p)})^T, \; \ldots, \; X_n = (x_n^{(1)}, x_n^{(2)}, ..., x_n^{(p)})^T.$$

Because indicators $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ can be measured in different scales, we assume that before applying the component analysis method, we make the transition to observations

$$\tilde{X}_1 = (\tilde{x}_1^{(1)}, \tilde{x}_1^{(2)}, ..., \tilde{x}_1^{(p)})^T, \; \tilde{X}_2 = (\tilde{x}_2^{(1)}, \tilde{x}_2^{(2)}, ..., \tilde{x}_2^{(p)})^T, \; \ldots, \; \tilde{X}_n = (\tilde{x}_n^{(1)}, \tilde{x}_n^{(2)}, ..., \tilde{x}_n^{(p)})^T$$

normalized values $\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots, \tilde{x}^{(p)}$ by formulas

$$\tilde{x}_i^{(j)} = \frac{x_i^{(j)} - \overline{x}^{(j)}}{\hat{\sigma}^{(j)}}, \; i = 1, 2, ..., n, \; j = 1, 2, ..., p, \tag{1}$$

where $\overline{x}^{(j)} = \frac{1}{n}\sum_{k=1}^{n} x_k^{(j)}$ and $\hat{\sigma}^{(j)} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k^{(j)} - \overline{x}^{(j)})^2}$ – accordingly, the sample mean and standard deviation of the *j*-th indicator, $j = 1, 2, ..., p$.

Then, by normalized observations $\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_n$ we find a sample correlation matrix $\Sigma = [\hat{\sigma}_{ij}]_{p \times p}$ of indicators $\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots, \tilde{x}^{(p)}$, where

$$\hat{\sigma}_{ij} = \frac{1}{n}\sum_{k=1}^{n} \tilde{x}_k^{(i)} \tilde{x}_k^{(j)} \ , \ i, \ j = 1, 2, ..., p \ . \qquad (2)$$

The first selective main component $y^{(1)}$ is the normalized linear combination of indicators $\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots, \tilde{x}^{(p)}$ which, among other normalized linear combinations of indicators $\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots, \tilde{x}^{(p)}$ has the largest variance.

The second selective main component $y^{(2)}$ is the normalized linear combination of indicators $\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots, \tilde{x}^{(p)}$ which, among other normalized linear combinations of indicators $\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots, \tilde{x}^{(p)}$ uncorrelated with $y^{(1)}$, has the largest dispersion. And so on.

The task of component analysis is to construct a linear transformation

$$y^{(1)} = l_{11}\tilde{x}^{(1)} + l_{12}\tilde{x}^{(2)} + \ldots + l_{1p}\tilde{x}^{(p)},$$

$$y^{(2)} = l_{21}\tilde{x}^{(1)} + l_{22}\tilde{x}^{(2)} + \ldots + l_{2p}\tilde{x}^{(p)},$$

$$\ldots \ \ldots \ \ldots \ \ldots \ \ldots \ \ldots \ \ldots \ \ldots$$

$$y^{(p)} = l_{p1}\tilde{x}^{(1)} + l_{p2}\tilde{x}^{(2)} + \ldots + l_{pp}\tilde{x}^{(p)},$$

for which

$$l_{j1}^2 + l_{j2}^2 + \ldots + l_{jp}^2 = 1, \ j = 1, 2, ..., p,$$

$$\mathsf{cov}(y^{(i)}, y^{(j)}) = 0, \text{ if } j < i,$$

$$\mathsf{D}y^{(1)} \geq \mathsf{D}y^{(2)} \geq \ldots \geq \mathsf{D}y^{(p)}.$$

Thus, the selective principal components $y^{(1)}, \ y^{(2)}, \ \ldots, \ y^{(p)}$, formed as linear combinations of the initial normalized features $\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots, \tilde{x}^{(p)}$, give a new set of features (generalized integral indicators) to the set of economic objects under consideration. In this case, the sample principal components are ordered by the degree of dispersion (by the magnitude of the variance).

It was obtained [1, 2, 5] that the vector $\mathbf{L}^{(j)} = (l_{j1}, l_{j2}, ..., l_{jp})^T$ is the j th normalized eigenvector of the sample correlation matrix Σ, which corresponds to the j th in size eigenvalue λ_j, that is λ_j is the j-th root of the characteristic equation $\det(\Sigma - \lambda E) = 0$ (E is the unit matrix of the p th order), and $\mathbf{L}^{(j)}$ is normalized by the condition $l_{j1}^2 + l_{j2}^2 + \ldots + l_{jp}^2 = 1$ of the solution of a homogeneous system $(\Sigma - \lambda_j E)\mathbf{L}^{(j)} = \mathbf{0}$, $j = 1, 2, ..., p$, with $\mathsf{D}y^{(j)} = \lambda_j$ and

$$\mathsf{D}\tilde{x}^{(1)} + \mathsf{D}\tilde{x}^{(2)} + \ldots + \mathsf{D}\tilde{x}^{(p)} = \mathsf{D}y^{(1)} + \mathsf{D}y^{(2)} + \ldots + \mathsf{D}y^{(p)} = \lambda_1 + \lambda_2 + \ldots + \lambda_p.$$

The contribution of the first p' main components ($1 \leq p' \leq p$) to the total variance is characterized by a value

$$q(p') = \frac{\mathrm{D}y^{(1)} + \mathrm{D}y^{(2)} + \ldots + \mathrm{D}y^{(p')}}{\mathrm{D}\tilde{x}^{(1)} + \mathrm{D}\tilde{x}^{(2)} + \ldots + \mathrm{D}\tilde{x}^{(p)}} = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_{p'}}{\lambda_1 + \lambda_2 + \ldots + \lambda_p},$$

analyzing which one can conclude how much is enough to allocate the first principal components to reduce the dimension of the space of output indicators of a given economic system.

Therefore, in order to get a clear idea of the structure of the economic system in question using the principal component method, it is necessary to design observation points $\tilde{X}_1$, $\tilde{X}_2$, …, $\tilde{X}_n$, from the original p-dimensional feature space into a one-dimensional (two-dimensional or three-dimensional) subspace stretched over the principal component $y^{(1)}$ (or major components $y^{(1)}$, $y^{(2)}$, or $y^{(1)}$, $y^{(2)}$, $y^{(3)}$), that is, each of the n objects of the economic system is represented by dots $(y_k^{(1)})$ ($(y_k^{(1)}, y_k^{(2)})$ or $(y_k^{(1)}, y_k^{(2)}, y_k^{(3)})$), $k = 1, 2, \ldots, n$, where

$$\begin{gathered} y_k^{(1)} = l_{11}\tilde{x}_k^{(1)} + l_{12}\tilde{x}_k^{(2)} + \ldots + l_{1p}\tilde{x}_k^{(p)}, \\ y_k^{(2)} = l_{21}\tilde{x}_k^{(1)} + l_{22}\tilde{x}_k^{(2)} + \ldots + l_{2p}\tilde{x}_k^{(p)}, \\ y_k^{(3)} = l_{31}\tilde{x}_k^{(1)} + l_{32}\tilde{x}_k^{(2)} + \ldots + l_{3p}\tilde{x}_k^{(p)},\ k = 1, 2, \ldots, n. \end{gathered} \tag{3}$$

The dimensionality of the p' subspace to which observations $\tilde{X}_1$, $\tilde{X}_2$, …, $\tilde{X}_n$, are designed, is determined based on the analysis of the behavior of the magnitude $q(p')$.

For the sake of certainty, let us assume that the design of the observations is made on a two-dimensional space, and, therefore, the observations are given as points $(y_k^{(1)}, y_k^{(2)})$, $k = 1, 2, \ldots, n$. To apply the cluster analysis procedure, you must select a metric by which the distances between objects will be measured. Since the initial observations were measured at interval scales, it is natural to use the Euclidean distance, which for points $(y_i^{(1)}, y_i^{(2)})$ and $(y_j^{(1)}, y_j^{(2)})$ is given by the formula

$$d_{ij} = \sqrt{(y_i^{(1)} - y_j^{(1)})^2 + (y_i^{(2)} - y_j^{(2)})^2}.$$

One of the hierarchical methods of cluster analysis is the method of minimizing Ward dispersion [8, 14]. This method allows obtaining the highest density clusters. Accordingly, in the first step of the algorithm, each cluster consists of a single object. The distance between the clusters is calculated as the increment of the sum of the squares of missing objects to the centre of the cluster resulting from their merging. This then combines the objects or clusters that give the smallest increase in the sum of squares. The results of hierarchical methods are usually presented in the form of a dendrogram - a treelike structure based on a measure of distance.

One of the most common non-hierarchical methods of cluster analysis is the k-means method, as well as its modification, which is more stable in the use of emission-observers, the k-median method [8, 14]. The essence of these methods is as follows. Upstream assumptions are made about the presence of clusters and on the plane randomly selected points (cluster centres) around which clusters are formed on the principle of minimizing the distance from the object to the centre of the cluster. Next, for the created clusters, the new centres are calculated as the arithmetic mean of the objects (or median signs) that formed the cluster, and several refinement iterations occur.

Comprehensive application of component and cluster analyses to the classification of regions of Ukraine in the 21st century. Consider the consistent application of component and cluster analysis methods to classify regions of Ukraine by their economic situation. To do this, consider the following socio-economic indicators traditionally used to characterize the economic situation:

$x^{(1)}$ – Gross regional product in actual prices, million UAH.;

$x^{(2)}$ – Gross regional product per capita, in actual prices, million UAH.;

$x^{(3)}$ – Consumer price index by region, % to the previous year;

$x^{(4)}$ – Industrial production index by region, % to the previous year;

$x^{(5)}$ – Index of agricultural production by regions, % to the previous year;

$x^{(6)}$ – Exports of goods by region, value, mln. US$;

$x^{(7)}$ – Imports of goods by region, value, mln. US$;

$x^{(8)}$ – Exports of services by region, cost, mln. US$;

$x^{(9)}$ – Imports of services by region, value, mln. US$;

$x^{(10)}$ – Wholesale turnover by regions, mln UAH.;

$x^{(11)}$ – Retail turnover in actual prices, mln UAH.;

$x^{(12)}$ – Registered unemployed, by region, % of the working-age population;

$x^{(13)}$ – Demand for labour, number of vacancies, thousand;

$x^{(14)}$ – Average monthly nominal salary by regions, UAH.

The values of the selected indicators for the regions of Ukraine are taken from [4, 18-20].

Let us consider the application of the clustering procedure with the previous reduction of the dimension by the principal component method according to 2012 data. We have $n = 27$ - number of regions (24 regions of Ukraine, Autonomous Republic of Crimea, Kyiv and Sevastopol); $p = 14$ - the number of indicators to be analyzed.

The preliminary stage of the statistical analysis is the normalization of the considered economic indicators in accordance with (1) since the raw data are given in different units (US dollars, interest, UAH, etc.). The basis for the application of the principal component method is the correlation matrix Σ, which is constructed by the formula (2) on the basis of the normalized input information (Table 1).

Table 1 – Correlation matrix of indicators Σ according to the 2012 data

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$	$x^{(6)}$	$x^{(7)}$	$x^{(8)}$	$x^{(9)}$	$x^{(10)}$	$x^{(11)}$	$x^{(12)}$	$x^{(13)}$	$x^{(14)}$
$x^{(1)}$	1	0,923	0,650	–0,157	–0,754	0,422	0,899	0,852	0,906	0,952	0,962	–0,557	0,889	0,914
$x^{(2)}$		1	0,636	–0,155	–0,844	0,325	0,952	0,902	0,956	0,917	0,823	–0,506	0,823	0,937
$x^{(3)}$			1	–0,267	–0,540	0,447	0,580	0,570	0,583	0,641	0,641	–0,716	0,488	0,752
$x^{(4)}$				1	0,111	–0,284	–0,147	–0,180	–0,159	–0,175	–0,181	0,458	–0,007	–0,287
$x^{(5)}$					1	–0,200	–0,857	–0,884	–0,862	–0,787	–0,691	0,425	–0,723	–0,777
$x^{(6)}$						1	0,294	0,256	0,283	0,329	0,437	–0,362	0,254	0,505
$x^{(7)}$							1	0,965	0,985	0,934	0,815	–0,552	0,820	0,869
$x^{(8)}$								1	0,964	0,915	0,808	–0,578	0,750	0,842
$x^{(9)}$									1	0,965	0,814	–0,526	0,797	0,884
$x^{(10)}$										1	0,890	–0,558	0,800	0,891
$x^{(11)}$											1	–0,611	0,848	0,859
$x^{(12)}$												1	–0,490	–0,627
$x^{(13)}$													1	0,745
$x^{(14)}$														1

Since the combination of the clustering method with a visual presentation of the results of grouping is chosen as a method of study, then we will use the two main components obtained by the method of component analysis. They make 81.15% of the contribution to the total variance. So, here are the first two maximal eigenvalues of the correlation matrix:

$$\lambda_1 = 9{,}85041,\ \lambda_2 = 1{,}51092$$

(note that the following eigenvalues λ_3 – λ_{14} do not exceed 1) and the corresponding eigenvectors:

$$\mathbf{L}^{(1)} = (0{,}306733;\ 0{,}304837;\ 0{,}230243; -0{,}0749314; -0{,}270995;\ 0{,}133939; 0{,}304549;\ 0{,}298144;\ 0{,}305175;\ 0{,}305669;\ 0{,}291656; -0{,}210277; 0{,}271502;\ 0{,}302963)^T;$$

$$\mathbf{L}^{(2)} = (0{,}0493388;\ 0{,}114377;\ -0{,}313895;\ 0{,}601219; -0{,}187837; -0{,}45862; 0{,}141146;\ 0{,}122458;\ 0{,}145224;\ 0{,}0819298;\ -0{,}0198077;\ 0{,}412784;\ 0{,}189012; -0{,}0982057)^T.$$

Using the above values, we use the formula (3) to calculate the values of the first two principal components:

$$y^{(1)} = 0,306733\, x_1 + 0,304837\, x_2 + 0,230243\, x_3 - 0,0749314\, x_4 -$$
$$-0,270995\, x_5 + 0,133939\, x_6 + 0,304549\, x_7 + 0,298144\, x_8 + 0,305175\, x_9 +$$
$$+0,305669\, x_{10} + 0,291656\, x_{11} - 0,210277\, x_{12} + 0,271502\, x_{13} + 0,302963\, x_{14};$$
$$y^{(2)} = 0,0493388 x_1 + 0,114377 x_2 - 0,313895 x_3 + 0,601219 x_4 - 0,187837 x_5 -$$
$$-0,45862 x_6 + 0,141146 x_7 + 0,122458 x_8 + 0,145224 x_9 + 0,0819298 x_{10} -$$
$$-0,0198077 x_{11} + 0,412784 x_{12} + 0,189012 x_{13} - 0,0982057 x_{14}.$$

By these ratios, we will determine the coordinates of regions of Ukraine in the coordinate system $(y^{(1)}, y^{(2)})$ in 2012:

Vinnytsya Region: $(-1,822;\ 1,451)$;
Volyn region: $(-1,680;\ -0,520)$;
Dnipropetrovsk Region: $(3,320;\ 0,543)$;
Donetsk Region: $(4,259;\ -1,61842)$;
Zhytomyr Region: $(-2,130;\ 2,490)$;
Zakarpattya Region: $(-1,479;\ 0,046)$;
Zaporizhzhya Region: $(0,034;\ -0,376)$;
Ivano-Frankivsk Region: $(-1,418;\ 0,091)$;
Kyiv Region: $(1,015;\ -0,445)$;
Kirovograd Region: $(-1,857;\ 1,705)$;
Luhansk Region: $(0,919;\ -3,847)$;
Lviv Region: $(-0,034;\ -0,066)$;
Mykolayiv Region: $(-0,855;\ 0,468)$;
Odesa Region: $(1,330;\ -0,264)$;
Poltava Region: $(-0,402;\ 0,747315)$;
Rivne Region: $(-1,816;\ 0,185)$;
Sumy Region: $(-1,428;\ -0,402)$;
Ternopil Region: $(-2,124;\ 0,666)$;
Kharkiv Region: $(1,045;\ -0,0855)$;
Kherson Region: $(-1,503;\ 0,762)$;
Khmelnytsky Region: $(-1,710;\ -0,129)$;
Cherkasy Region: $(-1,477;\ -0,025)$;
Chernivtsi Region: $(-1,860;\ -1,243)$;
Chernihiv Region: $(-1,988;\ 0,413)$;
Kyiv City: $(13,719;\ 1,545)$;
Sevastopol City: $(-0,025;\ -2,242)$;
Autonomous Republic of Crimea: $(-0,033;\ 0,147)$.

Note that these coordinates can be considered as the values of the integral indicators obtained by the method of principal components that characterize the economic situation of each region of the country as a whole.

The final stage of the analysis is the clustering of the data obtained. In Fig. Figures 1 and 2 show the results of cluster analysis using non-hierarchical and hierarchical approaches.

In particular, Figure 1 shows the graphical location of Ukraine's regions by principal components $y^{(1)}$, $y^{(2)}$, using the k-median cluster analysis method (6 clusters). Fig. 2 shows a dendrogram (cluster tree) constructed by the Ward method. In both cases, the Euclidean distance was used as the metric.

Analyzing the results, we can draw several conclusions. Both methods distinguish Kyiv into a separate cluster, which is explained by the peculiarities of its political position as the capital of Ukraine and, as a consequence, by the values of economic indicators. The other 5 clusters from fig. 1, obtained by the k-median method, contains the following regions:

- Dnipropetrovsk, Donetsk, Kyiv, Odesa, Kharkiv;
- Lugansk, Sevastopol;
- Volyn, Chernivtsi;
- Vinnytsya, Zhytomyr, Kirovograd;
- AR of Crimea, Zakarpattya, Zaporizhzhya, Ivano-Frankivsk, Lviv, Mykolayiv, Poltava, Rivne, Sumy, Ternopil, Kherson, Khmelnytsky, Cherkasy, Chernihiv.

On the basis of the dendrogram (Fig. 2), it is natural to distinguish (except Kyiv City) clusters consisting of the following regions:

- Vinnytsya, Zhytomyr, Kirovograd;
- Volyn, Zakarpattya, Ivano-Frankivsk, Rivne, Sumy, Ternopil, Kherson, Khmelnytsky, Cherkasy, Chernivtsi, Chernihiv;
- Dnipropetrovsk, Donetsk, Lugansk, Sevastopol;
- AR of Crimea, Zaporizhzhya, Kyiv, Lviv, Mykolayiv, Odesa, Poltava, Kharkiv.

Fig. 1. k-median clustering, 2012

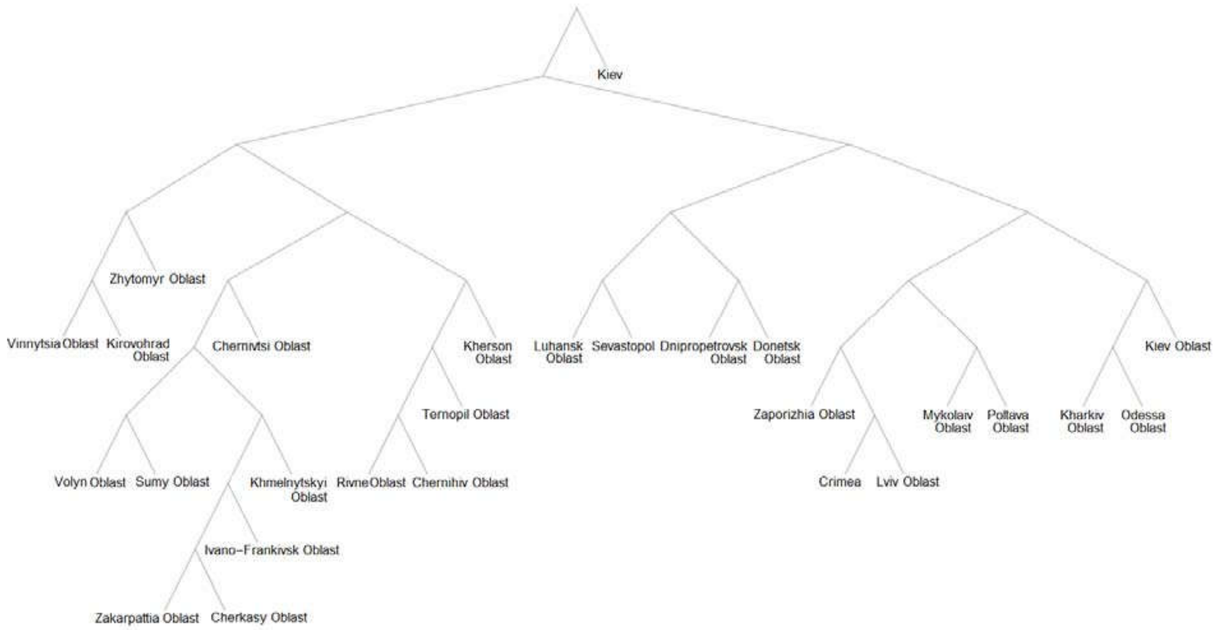

Fig. 2 Dendrogram of Ukrainian regions, 2012

As can be seen, in both cases the structure of the clusters is in good agreement (with some exceptions). The grouping differences that follow in Figs. 1 and 2 are explained by different mathematical approaches that use different clustering methods. By increasing or decreasing the number of clusters in the k-median method, it is possible to obtain other clusters (more detailed or more general) that are more suited to the purposes of the analysis. The use of hierarchical methods allows you to choose the right number of clusters of output from the appearance of the dendrogram.

Let us consider further what changes have taken place in the economic situation of Ukraine's regions in recent years. To do this, we apply the above approach for clustering according to 2017 data. As a result of component analysis, the following values of the first two components were obtained:

$$y^{(1)} = 0,322749x_1 + 0,311181x_2 + 0,137372x_3 - 0,045178x_4 -$$
$$-0,304181x_5 + 0,292846x_6 + 0,326835x_7 + 0,314691x_8 + 0,312756x_9 +$$
$$+0,324434x_{10} + 0,147145x_{11} - 0,114256x_{12} + 0,235248x_{13} + 0,317537x_{14};$$
$$y^{(2)} = 0,031930x_1 + 0,209915x_2 - 0,041099x_3 + 0,735189x_4 +$$
$$+0,013171x_5 - 0,030827x_6 - 0,034352x_7 - 0,014739x_8 - 0,060335x_9 -$$
$$-0,024382x_{10} + 0,318432x_{11} + 0,511370x_{12} + 0,200729x_{13} - 0,052707x_{14},$$

which make up 74.78% of the total variance.

The results of clustering by the k-median method provided that 6 clusters are shown in

Fig. 3. The dendrogram, constructed by the Ward method, is shown in Fig. 4.

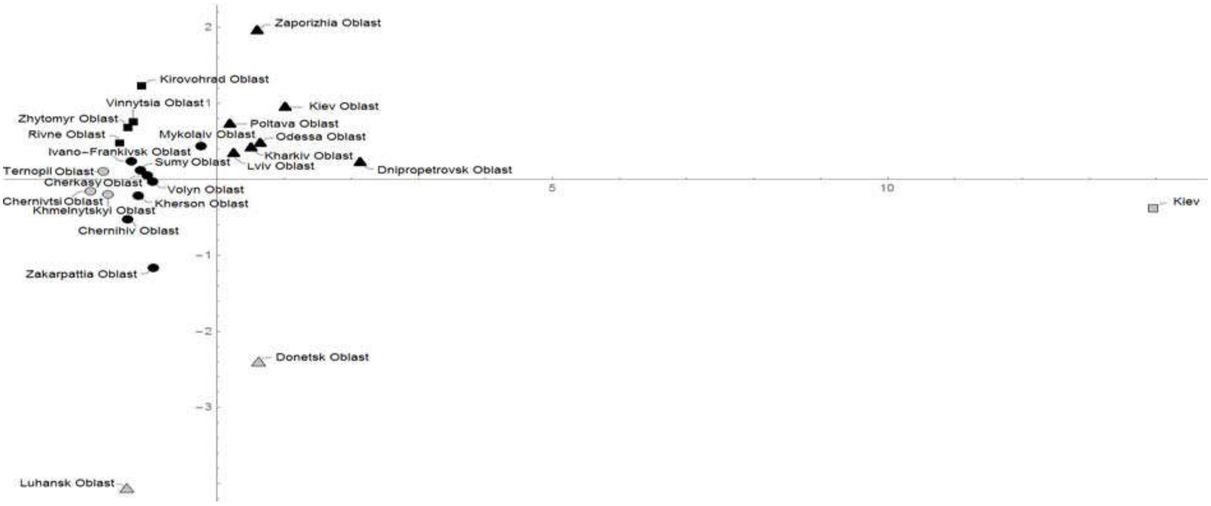

Fig. 3. k-median Clustering, 2017

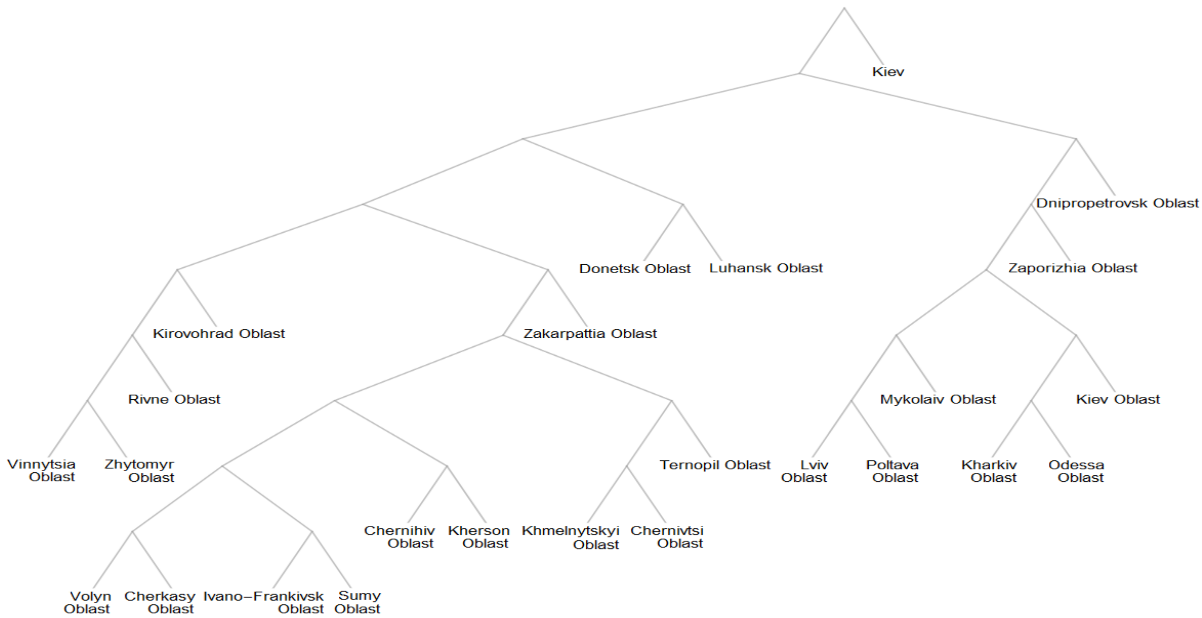

Fig. 4. Dendrogram of regions of Ukraine, 2017

As can be seen from the above figures, the special position of Kyiv as a separate cluster is preserved. Other clusters obtained by the k-median method (Fig. 3) consist of the following domains:

- Donetsk, Lugansk;
- Dnipropetrovsk, Zaporizhzhya, Kyiv, Lviv, Odesa, Poltava, Kharkiv;
- Ternopil, Chernivtsi, Khmelnytsky;
- Vinnytsya, Zhytomyr, Kirovograd, Rivne;

- Volyn, Zakarpattya, Ivano-Frankivsk, Mykolayiv, Sumy, Kherson, Cherkassy, Chernihiv.

The dendrogram (Fig. 4) allows you to group the areas into the following clusters (Kyiv is still allocated to a separate cluster):

- Vinnytsya, Zhytomyr, Kirovograd, Rivne;
- Donetsk, Lugansk;
- Dnipropetrovsk, Zaporizhzhya, Kyiv, Lviv, Mykolayiv, Odesa, Poltava, Kharkiv;
- Volyn, Zakarpattya, Ivano-Frankivsk, Sumy, Ternopil, Kherson, Khmelnytsky, Cherkasy, Chernivtsi, Chernihiv.

As before, overall, with some exceptions, the structure of clusters is maintained when using the k-median and Ward method.

Comparing the clustering results obtained by the same method for 2012 and 2017, we can conclude what changes have occurred in the economic situation of individual regions of Ukraine in the considered period of time.

Analysis of the received results. This procedure allows to group the regions with similar values obtained by the method of the main components of the integral indicators. Its advantage is that due to the large variety of natural economic characteristics we cannot only move to two or three summaries and merge the objects into homogeneous groups but also visually arrange the grouped objects in two-dimensional or three-dimensional space.

It should be noted that, unlike the traditional classification of regions of Ukraine as industrial, agricultural, resource, etc. [16, 21], the method of clustering proposed by the authors in combination with the method of principal components allows to integrate a large number of different economic indicators and perform groupings by the relevance of the general economic situation of the regions, not their economic profile. It should be noted that despite this peculiarity of the applied approach, the two cluster analysis methods used have allowed distinguishing groups that generally coincide with some traditionally considered ones: in particular, a group of large regions of eastern and southern Ukraine with great industrial potential; a group of regions mainly in western Ukraine, small in size and population; Kyiv as a separate cluster; the separation of Lugansk and Donetsk regions into a separate group because of the peculiarities of their current economic and political situation. It should be expected that the methodology discussed in this paper will also be useful for obtaining the classifications cited in general economic studies, for example, [21]. If the researcher is to obtain the grouping of regions by specific criteria (for example, breakdown into industrial/agricultural regions, etc., or classification by the level of development of social services), then indicators of this aspect of socio-economic status will have to be used as initial information. The selection of the required indicators and methods of cluster analysis is the prospect of applying the methodology in practice to obtain classifications that meet different economic research objectives.

Summarizing the results of the study, we can conclude that the consistent use of the principal component method and cluster analysis methods can help solve the problems of classification of elements of the economic system by a large set of indicators and significantly simplify the visual analysis. However, it should be noted that the use of component analysis procedure, which allows moving to a small number of integral indicators, at the same time leads to some loss of informativeness of the initial economic data that characterize the studied objects. For this reason, it is necessary to carefully consider the number and appearance of the indicators selected as initial when describing the economic status of the objects and to ensure that the compression of information does not become excessive. Taking into account the above comments allows us to consider this approach promising for visual statistical analysis of complex, in particular, economic, systems.

References

1. Aivazian S.A., Bejaeva Z.I., Staroverov O.V. Classification of multidimensional observations. - M.: Statistics, 1974. - 240 p.

2. Anderson T. An Introduction to Multivariate Statistical Analysis. - M.: GIFML, 1963. - 500 p.

3. Vavrin M.R. Economic Security of Ukraine: Clustering of Regions // Bulletin of the Dnipropetrovsk University. Series: Economics. - 2014. - № 22, issue. 8 (4). - P. 41-46.

4. State Statistics Service of Ukraine. [Electronic resource]. - Access mode: http://www.ukrstat.gov.ua.

5. Dubrov A.M., Mkhitaryan V.S., Troshin L.I. Multidimensional statistical methods. - M.: Finance and Statistics, 2003. - 352 p.

6. Ivanova N.S. Grouping of regions of Ukraine by financial security indicators using multidimensional classification methods // An effective economy. – 2014. – № 3.

7. Ignatenko V.Yu. Improvement of the mechanism of formation of the budget potential of the region on the basis of the clustering method // Scientific Bulletin of PUET. - 2012. - № 2. - P. 70-75.

8. Kim J.O., Müller C.H., Kleck U.R. Factor, discriminant and cluster analysis. - M.: Finance and Statistics, 1989. - 215 p.

9. Klinovy D.V., Pepa T.V. Deployment of productive forces and regional economy of Ukraine: textbook. // Scient. ed. L.G. Chernyuk. - K.: Center for Educational Literature, 2006. - 728 p.

10. Koroluk Yu.G. Cluster analysis of regional systems in the context of state regulation of socio-economic development // State and Regions. Series: Public Administration. - 2009. - № 4. - P. 174-179.

11. Korolyuk Yu. G. Estimation of the growth potential of socio-economic indicators of the region in the aspect of analysis and management of its development process // Investments: practice and experience. - 2011. - № 19. - P. 105-109.

12. Litvin VV, Pasichnyk VV, Nikolsky Yu.V. Data and knowledge analysis. - Lviv, "Magnolia 2006", 2015. - 276 p.

13. Lukan L., Cegelik G. Application of cluster analysis for the assessment of small business development in the regions of Ukraine // Formation of a market economy in Ukraine. - 2009. - № 19. - P. 73-80.

14. Mandel I.D. Cluster analysis. - M .: Finance and Statistics, 1988. - 176 p.

15. Applied Statistics: Classification and Dimension Reduction / S.A. Aivazian, V.M. Buchstaber, I.S. Enyukov, L.D. Meshalkin; Ed. S.A. Aivazian. - M.: Finance and Statistics, 1989. - 607 p.

16. Regional Economics: A Textbook / Ed. prof. E.P. Rocking. - Ternopil: Ternopil. Nat. Econom. Univ., 2008. - 800 p.

17. Sinitsa S.M., Vakun O.V. Clustering of regions by the level of economic potential // Economy and Society. - 2017. - Iss. 12. - P. 776-784.

18. Statistical collection «Regions of Ukraine». 2018 / Ed. I.E Werner. - Part 1. - K.: "August Trade" LTD, 2018. - 316 p.

19. Ukraine in Figures in 2012: Stat. Col. / Ed. O.G. Osaulenko. - K.: "August Trade" LTD, 2013. - 248 p.

20. Ukraine in Figures in 2017: Stat. Col. / Ed. O.G. Osaulenko. - K.: "August Trade" LTD, 2018. – 241 p.

21. Shabliy O.I. Ukraine - an object of political geography research // Social geography. Book Three. Problems of theoretical researches, geohumanistics, geographical education [Selected works] / Oleg Shabliy. - Lviv: Ivan Franko Lviv National University, 2017. - P. 154-164.