Scientific Visualization, 2021, volume 13, number 5, pages 1 - 15, DOI: 10.26583/sv.13.5.01

Dendrograms in Regional Socio-Economic Analysis: Interpretation and Verification

V.I. Blanutsa[1]

V.B. Sochava Institute of Geography, Siberian Branch of the Russian Academy of Sciences

[1] ORCID: 0000-0003-3958-216X, blanutsa@list.ru

Abstract

Interest in regional socio-economic analysis has grown in the last decade. Various tools are used to visualize its results. One of them is the dendrogram. However, there is no generalization of the experience of using dendrograms to display the sequence of combining regions into socio-economic clusters in world science. Therefore, the goal of our study was to generalize such experience with an emphasis on the interpretation and verification of dendrograms. Based on eight bibliographic databases using a special semantic search algorithm, more than eighty journal articles published around the world in the last two decades have been identified. These articles contain one hundred and thirty dendrograms. Their analysis showed that the main purpose of tree diagrams is to fix the sequence of combining regions into clusters and to substantiate the number of clusters. Two new types of interpretation of dendrograms are proposed – the allocation of nuclei in clusters and the determination of the level of socio-economic cohesion of clusters. To test the validity of determining the number of clusters, the author's algorithm for identifying the optimal clustering option is proposed, based on the idea of the complexity of the tree in graph theory. The ten main problems of visualizing the results of regional socio-economic analysis using dendrograms are listed.

Keywords: regional economy, regional analysis, hierarchical cluster analysis, regional economic convergence, socio-economic regionalization, dendrogram, interpretation, verification.

1. Introduction

Regional socio-economic analysis is aimed at studying the functioning (interaction) of the administrative-territorial divisions of a state (parts of a state; sometimes administrative units of several states) to manage the socio-economic development of regions and substantiate regional economic policy. According to the set of estimated parameters and methods used, socio-economic analysis at the regional level differs somewhat from the same analysis at the local (within a locality), international (several states) and global (all states) levels [1–5]. If we operate with the number of articles published annually in scientific journals around the world, then in the 21st century there is an increase in interest in the problems of regional socio-economic analysis (Fig. 1). The results of such an analysis are visualized mainly using a cartographic scheme (a simplified geographical map that displays only the borders of regions). Interactive online mapping tools [6], cartogram series [7], cartograms in combination with self-organizing maps [8], cartodiagrams and anamorphosis maps [9] are used to visualize the identified interregional socio-economic differences much less often. Additionally, we can note such a rare method of visualization as clustergram [10]. When displaying the results of socio-economic clustering of regions, cartograms are also used, but in hierarchical cluster analysis (as opposed to non-hierarchical), the cartogram is supplemented or replaced by a dendrogram (tree diagram), which records the sequence of combining regions into groups (clusters). The dendrogram is used not only to visualize the

results of cluster analysis, but also the economic convergence of regions [11, 12] and socio-economic regionalization [13, 14].

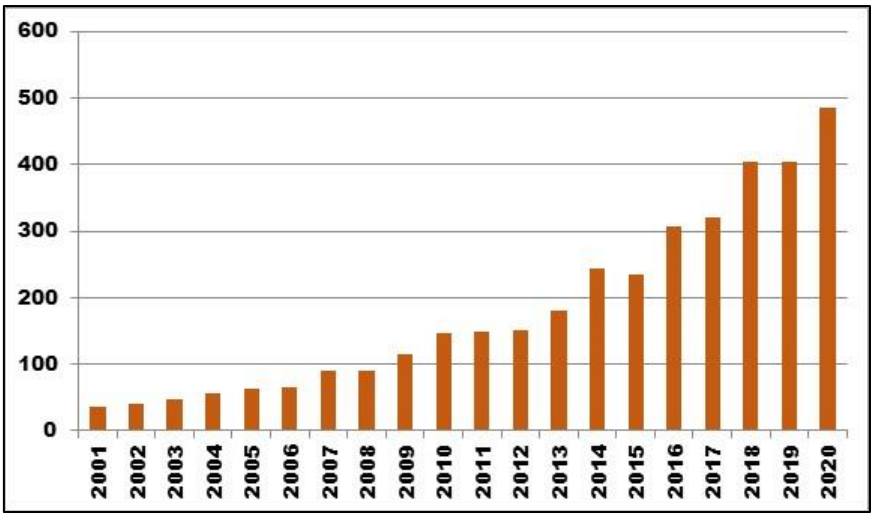

Fig. 1. Change in the annual number of articles on regional socio-economic analysis published in scientific journals around the world in 2001–2020 (according to Scopus; April 15, 2021)

In some scientific areas, the experience of using dendrograms to visualize research results has already been generalized (for example, in bioinformatics [15], plant breeding [16] and in studying the functional diversity of ecosystems [17]). However, such generalizations were not made in the regional economy. Therefore, the purpose of our study was to generalize the world experience of using dendrograms to visualize the results of regional socio-economic analysis. To achieve this goal, it was necessary to solve the following tasks: to identify the world array of publications on the subject under consideration; to extract a lot of dendrograms from it; to determine the general characteristics of the identified tree diagrams; to understand existing and outline promising ways of interpreting dendrograms; to diagnose the dendrograms used and propose new algorithms for verifying dendrograms. When implementing the tasks, the following restrictions were used: clustering of regions was analyzed (there is also clustering of features); a dendrogram was considered without its dependence on the similarity matrix or differences of regions in the feature space (in most publications there are no such matrices); the sequence of combining regions into groups was taken into account, regardless of the chosen similarity measure and clustering method (not all publications contain them); an agglomerative scheme for obtaining clusters (combining individual regions into groups) was discussed, since the divisional scheme (dividing all regions into groups) was not widespread in the regional economy; the sequence of combining regions in the space of economic or socio-economic characteristics was studied (without clustering regions based solely on social characteristics, which goes beyond the regional economy). The extraction of specific representations for machine learning from dendrograms [18] turned out to be outside of our research, since this promising direction relates to the problems of using artificial intelligence algorithms in the regional economy [19] and requires separate consideration.

2. Materials and Methods

When solving the first task, only journal articles were considered, since all texts with illustrations can be obtained from them, and not everything is available for other types of scientific publications – monographs, collections of articles and conference materials. Therefore, our conclusions relate only to the array of articles published in scientific journals around the world. The last twenty years (2001–2020) were chosen as a chronological limitation. Before that, articles on the subject under consideration were almost not published.

To identify articles with dendrograms related to regional socio-economic analysis, one domestic and seven international bibliographic databases were used (eLibrary.RU Scientific Electronic Library, Springer, Wiley, Elsevier and SAGE Publishers, Web of Science, Scopus and IDEAS databases). These databases contain the search for the necessary publications by keywords. However, such a search has many disadvantages. For example, using the keyword "regional socio-economic dendrogram" in the Scopus database, we managed to find only one article dedicated to the sustainable development of rural areas.

Therefore, to find the desired articles, a "Self-Organizing System for Publications' Searching on a Given Topic in a Bibliographic Database" [19] was used, which is a machine learning algorithm with a constant expansion of the semantic field. This algorithm was applied to each of the eight databases in an iterative mode: a semantic field identified in one database was applied and expanded in the next database, after which there was a return to the previous database for additional search of articles using the extended field. This happened until the size of the semantic field stabilized. A limitation of the algorithm is the selection of publications only in Cyrillic and Latin. Therefore, scientific articles using a different alphabet (for example, Chinese or Arabic) were left out of our analysis. Another limitation was the use of only eight databases, which cover most, but not all articles in the world.

Extraction of dendrograms from the identified articles (the second task) does not present any difficulty, except for the quality of illustrations, which is not equally high in all journals. When solving the remaining three tasks, dendrograms, texts of articles with their description, as well as author's developments on generalization, interpretation and verification of socio-economic regionalization schemes [14] were used.

3. Results and Discussion

The application of the semantic search algorithm [19] allowed us to find 81 journal articles on regional socio-economic analysis using dendrograms of combining territorial units into groups (clusters, convergence clubs, districts). These articles were distributed among 73 journals. Most of the articles (three) were published in "European Journal of Operational Research". When fixing the annual number of articles, there was a consistent (with fluctuations) increase in publication activity by the end of the period under review (Fig. 2), as a result of which about half of all articles accounted for the last four years (42 out of 81). At the same time, the share of publications in economic journals was less than 30%. Most of the articles contained one (60 publications) or two (16) figures, and each figure contained one (98 cases), two (3), three (4), four (2) or six (1) tree diagrams. As a result, 130 dendrograms were identified.

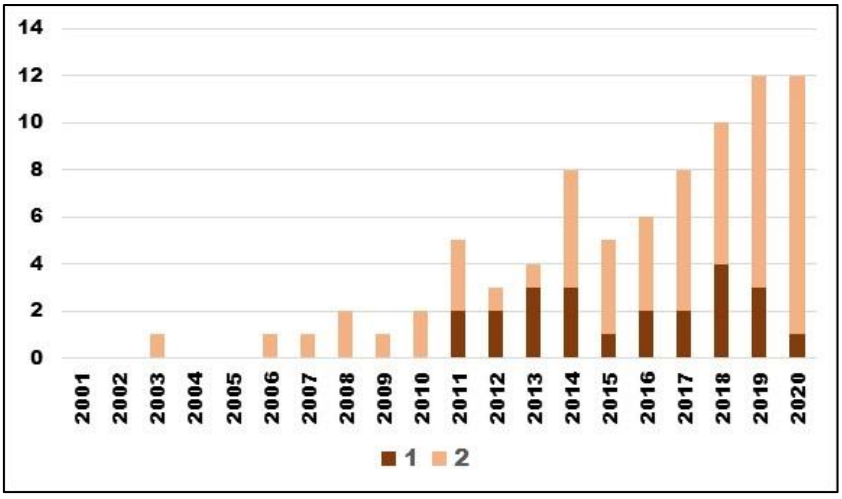

Fig. 2. Change in the annual number of articles with visualization of the results of regional socio-economic analysis using dendrograms published in economic (1) and other (2) scientific journals around the world in 2001–2020 (compiled by the author)

The main purpose of the dendrogram as a visualizer of the results of regional socio-economic analysis was to display the sequence of combining regions into groups according to given characteristics (Fig. 3) and to confirm the validity of the choice of the number of groups. The vast majority of the analyzed articles used both functions, but in 5 articles the tree diagram was used only to demonstrate the grouping of regions. Territorial units were compared with each other by one (most often the gross regional product per capita was estimated) or several characteristics. In the second case, the magnitude of similarity or difference between each pair of regions was calculated (preference was given to the Euclidean distance). On the basis of these distances, the regions were grouped into groups using various methods. Unfortunately, for 72 dendrograms, the method is either not specified, or a reference was given to standard software products that present several methods. Among the mentioned algorithms, the Ward's method [20] dominated (36 dendrograms).

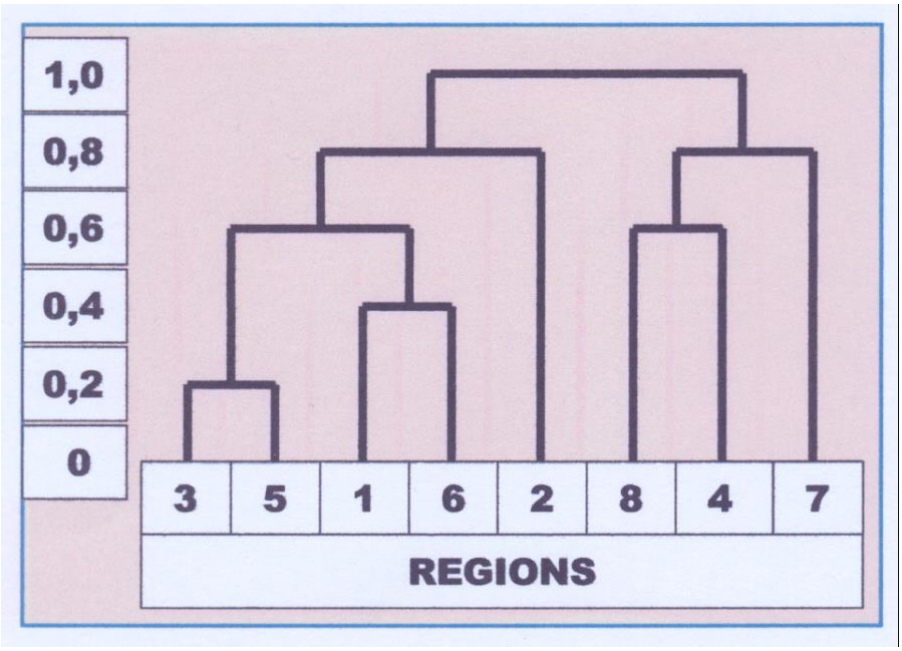

Fig. 3. A conditional example of the sequence of combining regions (1–8) into groups depending on the interregional distance in the feature space (0 –1.0)

The main visual difference between the dendrograms was manifested in the choice of the direction of tree structure convergence. There were five alternative directions: up (see Fig. 3), down, right, left or inside. The latter direction is associated with a circular (radial) dendrogram, in which the convergence is directed to the center of the circle (inside). In the analyzed array, preference was given to a graphical representation of the grouping of regions in two alternative directions (Fig. 4).

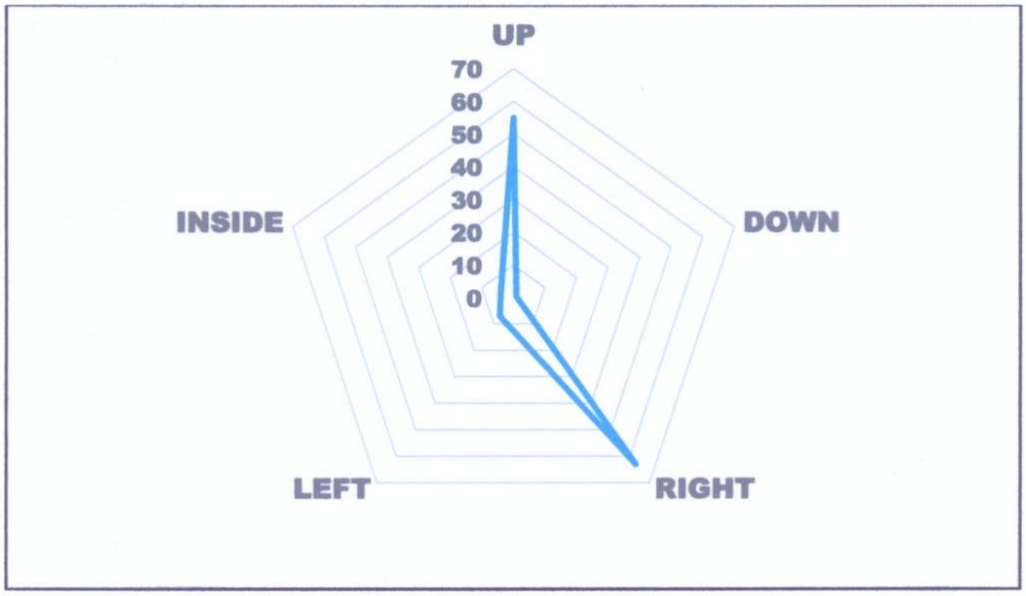

Fig. 4. The number of dendrograms from journal articles on regional socio-economic analysis (2001–2020), where the sequence of combining regions into groups is graphically presented in different directions

The last general characteristic of dendrograms (the third task) is the number of identified clusters. In the regional socio-economic analysis, regardless of the initial number of territorial units, four clusters were most often distinguished (Fig. 5).

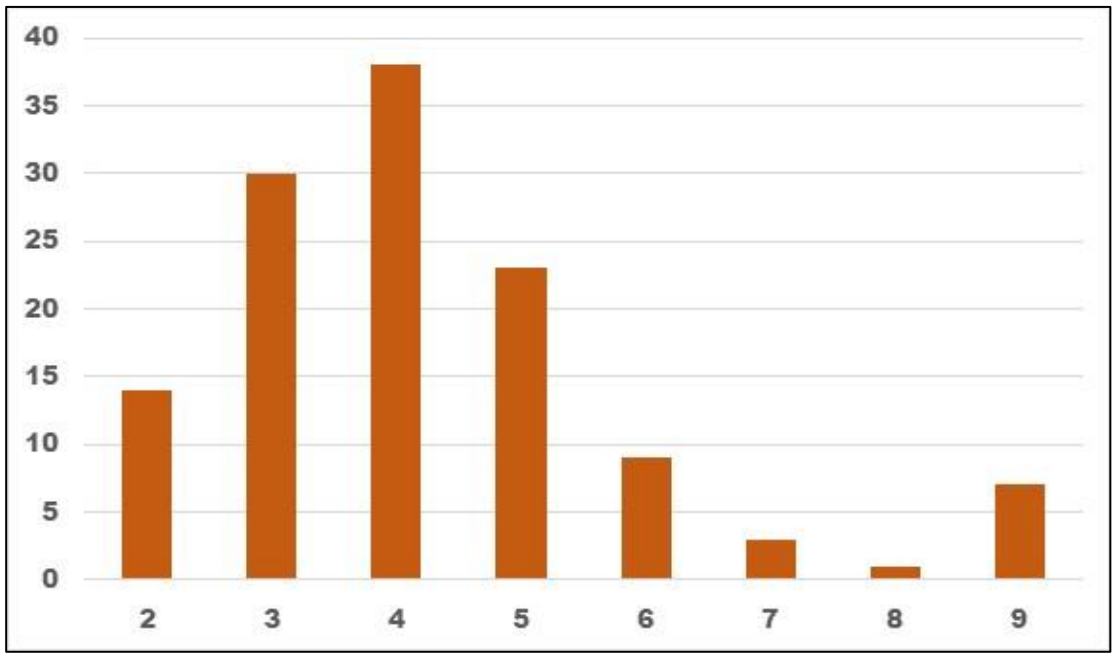

Fig. 5. The number of dendrograms with a different number of clusters (from 2 to 9) published in journal articles on regional socio-economic analysis (2001–2020)

As a result of solving the fourth task of the study, it was found that in the regional socio-economic analysis, the interpretation of dendrograms was reduced only to deciphering its two main functions – characterizing the sequence of combining regions into groups and confirming the validity of choosing a certain number of groups (clusters). At the same time, the characteristic of combining regions was to repeat what is already shown on the dendrogram – how one region was connected to another region at some step of grouping, and then a third region joined them, and so on. Such a description is not an interpretation, since it

does not carry new information and does not interpret the results from the position of another concept (approach). In such a situation, two new types of interpretation of dendrograms can be proposed: the allocation of cores in clusters and the determination of the level of socio-economic cohesion of clusters.

Based on the three rules for the allocation of cores in the types of territorial information and communication networks [21], it is possible to determine the cores according to the dendrogram of socio-economic analysis, understanding the core as a group of the most similar regions, to which the rest of the regions of the cluster under consideration join during the unification. At the same time, the smallest similarity or maximum distance is achieved at the last step of grouping, when all regions are combined into one cluster. The greatest similarity between regions is equal to one or zero distance. However, in the regional dimension of socio-economic processes, completely similar territorial units are rarely found [14]. Therefore, a certain amount of permissible similarity is established, not exceeding which the regions are considered the most similar. The justification of this value is the subject of a separate study. If we take into account the experience of the European Union on the economic convergence of regions [22], then this value is an acceptable deviation of 25% from the desired socio-economic situation. In relation to our problems, this is expressed in 75% similarity or 25% distance. The conditional example shows (see Fig. 3) that regions 3 and 5 are grouped at a distance between them of 0.20, and the remaining regions are grouped at distances greater than 0.25. Let's say that in the conditional example, two clusters are allocated: regions 1, 2, 3, 5, 6 and regions 4, 7, 8. Then the first cluster will be single-core, and the second will be non-core. A multicore cluster is also possible when there are two or more groups formed with a similarity of at least 75% (a distance of no more than 25%).

The identification of core-free, single-core and multi-core clusters gives a new meaning to the interpretation of the sequence of combining regions into groups. For example, when clustering 13 regions (peripheries) of Greece by 11 socio-economic indicators (the situation was assessed for three years – 1995, 2000 and 2007) [23], 5 clusters were identified (Fig. 6). If we take 25% of the maximum distance of 25 units, it turns out that the cores can only be formed at a distance of 6.25. In this case, the following interpretation can be obtained: one dual-core and four non-core clusters in 1995; three single-core and two non-core clusters in 2000; one dual-core, one single-core and three non-core clusters in 2007. When the requirements for the minimum allowable distance are tightened to 10% (distance of 2.5 for this case), a different interpretation is obtained: one single-core and four non-core clusters in 1995; one dual-core, one single-core and three non-core clusters in 2000; one dual-core and four non-core clusters in 2007. The choice of 25%, 10% or other maximum permissible distance depends on the specifics of the socio-economic situation in Greece and requires appropriate justification. Therefore, the distances of 6.25 and 2.5 units are given only as an illustration of the possibility of identifying cores. Examples of cluster allocation at smaller distances in the socio-economic regionalization of Sierra Nevada (Spain) [24] and business clustering of Romania's districts [25] serve as confirmation of the impossibility of applying a single permissible distance (up to 25% of the maximum) to all dendrograms. Therefore, in each specific case, it is necessary to justify not only the choice of the measure of similarity of regions and the method of cluster analysis, but also the permissible similarity (difference).

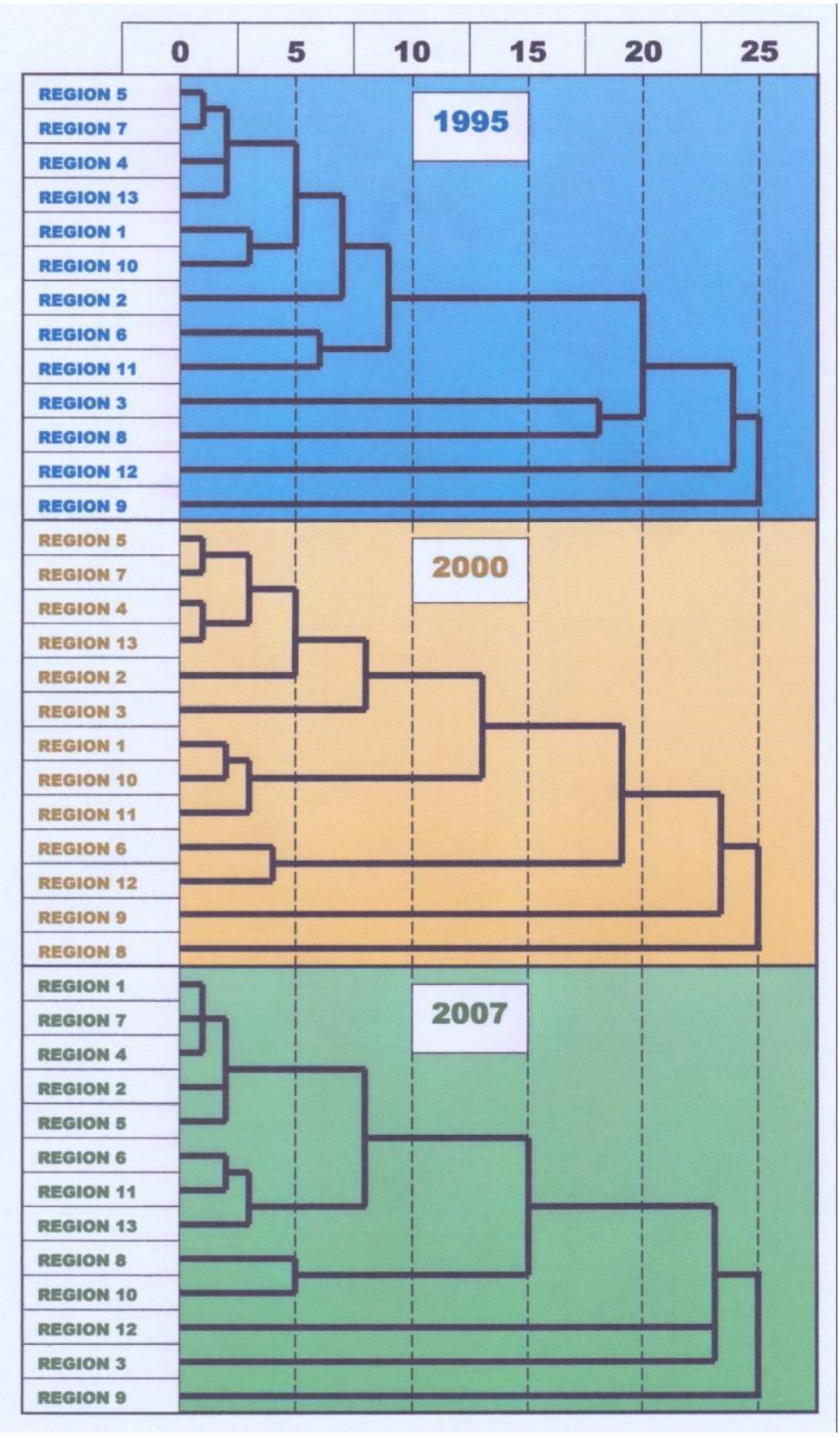

Fig. 6. Socio-economic clustering of Greek regions with a sequential increase in the interregional distance in the feature space from 0 to 25 units (1995, 2000 and 2007) [23]
Regions: 1 – Eastern Macedonia and Thrace, 2 – Central Macedonia, 3 – Western Macedonia, 4 – Epirus, 5 – Thessaly, 6 – Ionian Islands, 7 – Western Greece, 8 – Central Greece, 9 – Attica, 10 – Peloponnese, 11 – Northern Aegean Islands, 12 – Southern Aegean Islands, 13 – Crete.

Determining the level of cluster cohesion based on socio-economic characteristics can also be based on the maximum allowable distance. Here, the interpretation is related to understanding how heterogeneous (homogeneous) the resulting clusters are. This can be judged by the distance at which the line of "cutting" the dendrogram into clusters passes. If in our conditional example (see Fig. 3) two clusters are allocated, then the dendrogram is cut at a distance of 0.8. This value can be estimated taking into account the maximum allowable distance: at 25%, the entire interval (0–100%) is divided into four quartiles. The conditional example falls into the last (fourth) quartile. For comparison, 5 clusters of Greek regions (see Fig. 6) are characterized by the second quartile, which indicates their higher cohesion in relation to the conditional example. Instead of quartiles, you can compare dendrograms along the cutting line. Relative values of similarity or difference should be used only to ensure comparison. Comparing the level of regional cohesion obtained during the analysis with the same level in the control example opens up new possibilities of interpretation.

The last and rarely encountered type of interpretation of the sequence of combining territories occurs in the case of appearance of an "outlier" (a separate region that is so unlike other regions that it is not considered a cluster). Therefore, such a region is excluded from clustering as "information noise". However, from an epistemological point of view, the analysis of integral territorial formations should not contain exceptions [14]. The existence of very specific regions as clusters allows us to interpret the result of grouping as an unbalanced system, and ignoring these regions leads to the appearance of "white spots" in the studied territory, which distorts the real socio-economic situation. It is on this issue that the distinction between domestic and foreign regional analysis is made. In some Russian studies, the assumed outliers are removed a priori (before clustering), and in foreign studies this happens posteriori (after clustering). For example, the city of Moscow [26, 27] or Moscow together with the Moscow Region [28] are excluded from the Central Federal District, and the Chechen Republic is excluded from the North Caucasus District [29]. Abroad, for example, the Aosta Valley Region was not considered a separate cluster in the economic convergence of Italian regions [30], and Qinghai Province was excluded from the American study of social security regimes in China [31].

Another interpretation of dendrograms is related to determining the optimal number of clusters. This can be done on the basis of a matrix of similarities (differences) between regions according to specified characteristics (indicators, parameters), which allows us to calculate clustering efficiency indices (for example, the Caliński-Harabasz index [32, 33]). In addition, it is possible to make a comparison with a geographical map (preference is given to the option in which the regions within each cluster are neighbors [34]) or to conduct a visual analysis of the dendrogram. In the world array of articles on regional socio-economic analysis, the marked matrices are rarely given, and the requirement of geographical compactness of clusters applies only to regionalization. Therefore, in our study, we will limit ourselves to visual analysis.

The procedure of visual analysis is quite subjective and can manifest itself in the choice of the number of clusters, for example, according to "our understanding of the peculiarities of the Indian states" [35] or "the possibility of economic interpretation" [36]. However, in most cases, the dendrogram was cut along its longest segments without combining the regions into groups. In the case of grouping of the Greek peripheries (see Fig. 6) the cutting was carried out approximately at a distance of 10 units from the initial state. For 1995, the interval between the formation of the first cluster (regions 1, 2, 4, 5, 6, 7, 10, 11 and 13) and the union of regions 3 and 8 is indeed the longest, and for 2000 and 2007 the maximum distance without associations falls on the next segment, where four clusters are observed instead of the declared five groups. If we turn to the conditional example (see Fig. 3), then all the segments there are the same (0.20 units of distance each). These examples indicate that the selection of the longest segment on the dendrogram is not always carried out correctly or does not allow you to select the maximum segment. Therefore, a heuristic algorithm is needed that formalizes intuitive ideas about visualizing the optimal number of clusters.

Another type of interpretation of the number of clusters is associated with the incompleteness of the study. In most publications, the determination of the optimal number of regional groups completes the study (followed only by the characteristics of the selected clusters). However, in several cases, hierarchical cluster analysis was used to justify the number of groups in the subsequent non-hierarchical grouping of regions into clusters using the k-means method [37–39]. In such cases, it was interpreted not the distribution of regions by clusters, but only the number of clusters in a hierarchical grouping, which was sometimes interpreted as excessive for a non-hierarchical association [37].

In almost all the identified articles, the number and composition of clusters were determined by rectilinear cutting of the dendrogram. However, in one case, a curved cutting was observed, leading to an incorrect interpretation of the number of clusters. When studying the cyclical nature of regional housing prices in the United States, the possibility of the existence of 4 clusters was established (according to the dendrogram). However, the authors then decided to divide one large cluster into two sub-clusters on the grounds that this "allows us to highlight interesting details concerning the geographical distribution of housing price cycles" [40]. As a result, the dendrogram was cut along a curved line, and the subclusters were considered on a par with the remaining three clusters. It is not clear why the choice of 5 clusters was not made immediately (this can be done according to the dendrogram), but it was necessary to go to a four-cluster solution and then only within the first cluster to return to the option of grouping territories according to a five-cluster solution.

Dendrograms can be used not only to interpret the sequence of combining regions into clusters and choosing the optimal number of clusters, but also to verify the results of regional socio-economic analysis. Since one of the main results of such an analysis is the unification of territories into groups (clusters, convergence clubs, districts), it is first necessary to check the validity of such an association. At the same time, we will operate only with dendrograms as a visualizer for combining regions (as noted above, it is not possible to use matrices of similarity or difference of territories, since they are not given in most articles).

One of the most common ways to verify the results of grouping regions is associated with determining the optimal number of clusters using an alternative algorithm. This point of view is shared by many researchers in the field of regional analysis [32, 34, 36, 39, 41–46]. However, the algorithms they use cannot be called "alternative". For example, it was proposed to divide all the features (variables) into two groups and use the Ward's method to first cluster the Greek regions according to the first group of features, and then compare the results with the clustering of the same regions according to both groups of features [41]; to look at the differences between the groupings of Croatian districts by absolute and relative indicators [42]; to compare the groups of territories according to the regional index of information and communication technologies development obtained by applying the Ward's method and the complete linkage method [39]; to compare the groupings of Italian regions according to the trends of employment of disabled people obtained by three methods (complete, single and average linkage methods) [32], or the unification of Bangladesh cities as a result of using four methods that differ only in the criterion of joining the region to the group [43]; to compare geographical maps with the unification of French municipalities by socio-economic distance between them and their association taking into account the geographical distance between neighbors [34]. In the examples given, the algorithms were of the same type (the same source matrix or its variant), which allowed us to evaluate only some differences. Therefore, to verify the results of a regional analysis, it is advisable to use a method that is in no way related to the procedures of this analysis. Such an "independent" algorithm, which was not used in the identified array of articles, can be a heuristic method for determining the most complex tier of the forest in socio-economic regionalization [14].

The general structure of the alternative algorithm can be represented as the following sequence of actions: the tested dendrogram is transformed into a grouping tree; the absolute and relative values of the complexity of each tier of the forest (the step of grouping regions) are calculated; the most complex tier (clustering option) is selected and compared with the

cluster solution being tested. To visualize the choice of the optimal number of clusters, the original dendrogram is transformed into one of its varieties – a grouping tree in which the vertices are connected by edges without forming loops (cycles, closed contours). In our conditional example (see Fig. 3) we can distinguish 5 grouping steps (one step is equal to 0.20 units of distance), leading, respectively, to five options for combining 8 regions into 7, 6, 4, 2 and 1 cluster (Fig. 7). The latter option covers all regions and in this sense is not a cluster solution. There are four options left, among which it is necessary to choose the optimal solution.

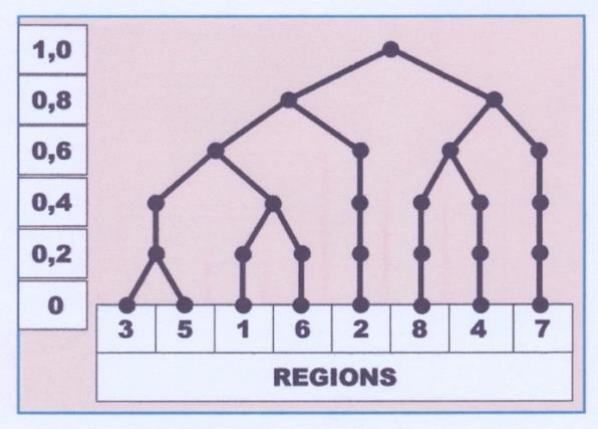

Fig. 7. Grouping tree corresponding to the dendrogram of the conditional example (see Fig. 3)

When choosing the optimal variant of socio-economic regionalization, the best correlation is sought between the requirements of maximum homogeneity of groups of neighboring regions and the minimum number of such groups [47, 48]. If each region is unique (the most common situation in regional analysis [14]), then the maximum homogeneity of groups of regions is achieved only in the initial situation (the number of groups or clusters corresponds to the number of regions). The subsequent unification of the regions reduces the homogeneity of the resulting groups, but allows you to get a small number of them, which is convenient for socio-economic management [1, 13, 14]. The problem is that these requirements contradict each other: maximizing uniformity leads to maximizing the number of groups or minimizing the number of groups leads to minimizing uniformity. Hence, the problem arises of finding some optimum at which a small number of groups is characterized by acceptable uniformity. The tree graph (grouping tree), reflecting the sequence of combining the largest number of regions into the smallest number of groups (clusters) in a minimum of steps, according to Yu.A. Schrader [49] is the most complex. The complexity of the tree can be estimated using the following recurrent formula [49]:

$$\sigma(x) = \alpha\gamma + \sum \sigma(y),$$

where $\sigma(x)$ is the complexity of vertex x; α is the number of edges coming down from vertex x; γ is the number of tree vertices; $\Sigma\sigma(y)$ is the total complexity of the vertices of the previous tier connected by edges to vertex x. According to this formula, the complexity of the entire grouping tree in our example (see Fig. 7) is 182 units (2×28+78+48). If there is not one tree, but a set of them (a forest), then the complexity of the forest $\sigma(D)$ corresponds to the total complexity of the trees of this forest:

$$\sigma(D) = \sum_{x=1}^{r} \sigma(x),$$

where r is the number of trees (clusters). So, in our example, the complexity of the forest after the first step (7 trees) was 18 units (2×3+2+2+2+2+2+2).

To determine the relative complexity of forest C, it is necessary to correlate the absolute value with the maximum possible complexity [14]:

$$C = \frac{\sigma(D)}{\sigma(D)_{max}} = \frac{\sigma(D)}{n(nh + k + 1)},$$

where C is the degree of complexity of the forest; h is the number of tiers of the forest; n is the number of regions; k is the coefficient of connection with the number of tiers of the forest, determined by the recurrent rule ($k = 0$ for $h = 1$, $k_i = k_{i-1} + h_i$ for $h \geq 2$). According to the conditional example, the following values of C were obtained for the first and subsequent tiers: 0.25; 0.24; 0.30; 0.38; 0.40. Discarding the last value, the fourth tier turned out to be the most difficult (combining 8 regions into 2 clusters).

Based on the presented algorithm, it is possible to check the validity of determining the number of clusters from dendrograms from journal articles. Let's take as an example three dendrograms of clustering of the peripheries of Greece (see Fig. 6). We transform them into three grouping trees by entering a step of combining regions into clusters equal to one unit of distance (the minimum step according to the dendrogram). Then you will get 25 steps (tiers). Recall that in the corresponding article [23], 5 clusters were allocated. For 1995, the highest degree of complexity of the forest was achieved with a two-cluster solution ($C = 0.2275$), and the five-cluster solution (0.2032) was also inferior in complexity to the variant with three clusters (0.2094). In 2000, the identification of 5 clusters ($C = 0.2041$) was less difficult than the identification of four (0.2192), three (0.2255) and two (0.2327) clusters. By 2007, the five-cluster solution ($C = 0.1857$) was worse than grouping regions into four (0.1863), eight (0.189) and two (0.2056) clusters. Thus, the verification of three dendrograms using an alternative algorithm showed that 13 Greek regions are combined into two clusters, and not into five groups.

We can check the other dendrograms in the same way. For example, 14 regions of the Czech Republic were grouped into two clusters according to socio-economic characteristics [50], but verification showed that this option ($C = 0.3986$) is significantly inferior to the three-cluster solution (0.6667). According to another example related to the identification of regional trends in the employment of disabled people in Italy [32], it was found that the authors' association of 20 regions into two clusters ($C = 0.1910$) is optimal compared to the formation of three (0.1763) or four (0.1734) groups. When analyzing 16 regions of Poland according to labor market indicators in 2005 and 2014, 5 clusters were identified [51], but in both cases the grouping into 8 clusters is optimal: with a degree of complexity of 0.2132 versus 0.1946 (5 clusters) in 2005 and 0.2059 versus 0.1620 in 2014. When grouping 32 provinces of China by domestic investment [52], a three-cluster solution was obtained ($C = 0.1198$), which contradicted the optimal combination of these provinces into 8 clusters (0.3731).

The examples given indicate contradictions in the choice of the optimal number of clusters. One of the reasons for the inefficiency of hierarchical cluster analysis is a non-hierarchical result, which manifests itself in the selection of only one type of taxa. For example, in socio-economic regionalization, several hierarchically ordered taxa are identified – zone, subzone, province and district [14], each of which is characterized by an increase in the degree of complexity of forest C compared to the previous and subsequent step of grouping.

The examples given indicate contradictions in the choice of the optimal number of clusters. One of the reasons for the inefficiency of hierarchical cluster analysis is the non-hierarchical result, which manifests itself in the allocation of only one type of taxa. For example, in socio-economic regionalization, several hierarchically ordered taxa are identified, such as a zone, a subzone, a province and a district [14]. In this case, each taxon is distinguished by an increase in the degree of complexity of forest C compared to the previous and subsequent step of grouping. This approach can be extended to cluster analysis if, with a gradual decrease in the number of clusters, there are not one, but several "peaks" of magnitude C. The number of such peaks indicates the number of taxa. Then, for example, with three peaks corresponding to an increase in the degree of complexity from the first to the

third peak, you can get a hierarchical result in the form of clusters (the third peak), sub-clusters (the second peak) and groups (the first peak). It is advisable to call the taxon with the highest degree of complexity a “cluster”, and part of the cluster can be called a “sub-cluster”, “group” or “subgroup”. In this case, the cluster union can be called a “super-cluster”, “cohesion”, etc. An alternative to the hierarchical result can be a fuzzy cluster solution, when the cluster cores are allocated, and the remaining regions relate to each core with a different degree. In the analyzed array of articles, clustering of regions was carried out without taking into account the hierarchy of taxa and the theory of fuzzy sets.

4. Conclusion

The generalization of the world experience of using dendrograms to visualize the results of regional socio-economic analysis, carried out taking into account only journal articles over the past two decades, allowed us to identify the following problems: (1) rare use (dendrograms are found in about 1 out of 45 articles; see Fig. 1 and 2); (2) use only for displaying the sequence of combining regions into clusters and determining the number of clusters; (3) the absence of a divisional scheme for obtaining clusters (dividing the entire studied territory into clusters; such a division could be used to verify the results of combining regions into clusters or an agglomerative scheme); (4) an infrequent representation in the form of a radial dendrogram (see Fig. 4; enumeration of regions in a circle allows you to fix significantly more regions than enumeration by diameter, which is one of the sides in ordinary dendrograms, but this advantage was not used in many studies with a large number of regions, when instead of the sequence of combining regions into clusters, a truncated grouping of some sets of regions into clusters was given); (5) lack of ways to identify non-core, single-core and multi-core clusters; (6) misunderstanding of the level of socio-economic cohesion of groups of regions; (7) ignoring abnormal regions (“outliers”); (8) subjectivity of visual selection of the optimal number of clusters; (9) lack of alternative clustering algorithms for verifying dendrograms; (10) non-hierarchical result of hierarchical cluster analysis. In our study, the first four problems are only listed, and for the remaining six problems, only the contours of future solutions are outlined.

Acknowledgments

The work was carried out at the V.B. Sochava Institute of Geography of the Siberian Branch of the Russian Academy of Sciences at the expense of the State task (registration number of the topic AAAA-A17-117041910166-3).

References

1. Soltes V., Stofkova K.R., Kutaj M. (2016) Socio-economic analysis of development of regions. *Global Journal of Business, Economics and Management*, vol. 6, no. 2, pp. 171–178. doi: 10.18844/gjbem.v6i2.1382

2. Bross U., Walter G.H. (2000) *Socio-Economic Analysis of North Rhine-Westphalia*, Karlsruhe: Fraunhofer Institute for Systems and Innovation Research.

3. Zhang J., He X., Yuan X.-D. (2020) Research on the relationship between urban economic development and urban spatial structure – A case study of two Chinese cities. *PLoS ONE*, vol. 15, no. 7, e0235858. doi: 10.1371/journal.pone.0235858

4. Fura B., Wang Q. (2017) The level socioeconomic development of EU countries and the state of ISO 14001 certification. *Quality and Quantity*, vol. 51, no. 1, pp. 103–119. doi: 10.1007/s11135-015-0297-7

5. Lensen M., Li M., Malik A., Pomponi F., Sun Y.-Y., Wiedmann T., Faturay F., Fry J., Gallego B., Geschke A., Gómez-Paredes J., Kanemoto K., Kenway S., Nansai K., Prokopenko M., Wakiyama T. Wang Y., Yousefzadeh M. (2020) Global socio-economic losses and environmental gains from the Coronavirus pandemic. *PLoS ONE*, vol. 15, no. 7, e0235654. doi: 10.1371/journal.pone.0235654

6. Smith D.A. (2016) Online interactive thematic mapping: Applications and techniques for socio-economic research. *Computers, Environment and Urban Systems*, vol. 57, pp. 106–117. doi: 10.1016/j.compenvurbsys.2016.01.002

7. Sova M., Vukosav B. (2017) A contribution to measuring and visualizing regional disparities in the Czech Republic. *Kartografija i Geoinformacije*, vol. 16, no. 28, pp. 26–44.

8. Hajek P., Henriques R., Hajkova V. (2014) Visualizing components of regional innovation systems using self-organizing maps – Evidence from European regions. *Technological Forecasting and Social Change*, vol. 84, pp. 197–214. doi: 10.1016/j.techfore.2013.07.013

9. Gordeev S.S. (2016) Ocenka ustojchivosti prostranstvennogo socio-ekologo-ekonomicheskogo razvitiya v srede geoinformatiki [Assessment of the sustainability of spatial socio-ecological-economic development in the environment of geoinformatics]. *Vestnik Chelyabinskogo gosudarstvennogo universiteta. Ekonomicheskie nauki*, no. 11, pp. 37–50 [in Russian].

10. Schonlau M. (2004) Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. *Computational Statistics*, vol. 19, pp. 95–111. doi: 10.1007/BF02915278

11. Balash V., Balash O., Faizliev A., Chistopolskaya E. (2020) Economic growth patterns: Spatial econometric analysis for Russian regions. *Information*, vol. 11, e289. doi: 10.3390/info11060289

12. Blanutsa V.I. (2020) Klasterizaciya regionov Sibiri i Dal'nego Vostoka po perspektivnym ekonomicheskim specializaciyam [Clustering regions of Siberia and the Far East by promising economic specializations]. *Vestnik Omskogo universiteta. Seriya "Ekonomika"*, vol. 12, no. 2, pp. 80–90 [in Russian]. doi: 10.24147/1812-3988.2020.18(2).80-90

13. Mitchell W., Watts M. (2010) Identifying functional regions in Australia using hierarchical aggregation techniques. *Geographical Research*, vol. 48, no. 1, pp. 24–41. doi: 10.1111/j.1745-5871.2009.00631.x

14. Blanutsa V.I. (2018) *Social'no-ekonomicheskoe rajonirovanie v epohu bol'shih dannyh* [Socio-Economic Regionalization in the Era of Big Data], Moscow: INFRA-M Publ. [in Russian]. doi: 10.12737/monography_59f81ac5ede918.09423566

15. Schroeder M., Gilbert D., van Helden J., Noy P. (2001) Approaches to visualization in bioinformatics: From dendrograms to Space Explorer. *Information Science*, vol. 139, no. 1-2, pp. 19–57. doi: 10.1016/S0020-0255(01)00156-6

16. Arief V.N., DeLacy I.H., Basford K.E., Dieters M.J. (2017) Application of a dendrogram seriation algorithm to extract pattern from plant breeding data. *Euphytica*, vol. 213, e85. doi: 10.1007/s10681-017-1870-z

17. Petchey O.L., Gaston K.J. (2007) Dendrograms and measuring functional diversity. *Oikos*, vol. 116, no. 8, pp. 1422–1426. doi: 10.1111/j.0030-1299.2007.15894.x

18. Chehreghani M.H., Chehreghani M.H. (2020) Learning representations from dendrograms. *Machine Learning*, vol. 109, pp. 1779–1802. doi: 10.1007/s10994-020-05895-3

19. Blanutsa V.I. (2020) Regional'nye ekonomicheskie issledovaniya s ispol'zovaniem algoritmov iskusstvennogo intellekta: sostoyanie i perspektivy [Regional economic research using artificial intelligence algorithms: state and prospects]. *Vestnik Zabajkal'skogo gosudarstvennogo universiteta*, vol. 26, no. 8, pp. 100–111 [in Russian]. doi: 10.21209/2227-9245-2020-26-8-100-111

20. Ward J. (1963) Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244.

21. Blanutsa V.I. (2016) *Razvertyvanie informacionno-kommunikacionnoj seti kak geograficheskij process (na primere stanovleniya setevoj struktury sibirskoj pochty)* [Deployment of an Information and Communication Network as a Geographic Process (On the Example of the Formation of the Network Structure of the Siberian Post)], Moscow: INFRA-M Publ. [in Russian]

22. Voronov V.V. (2014) Konvergenciya regionov Evropejskogo soyuza: osobennosti i ocenka [Convergence of regions of the European Union: features and assessment]. *Ekonomicheskie i social'nye peremeny: fakty, tendencii, prognoz*, no. 6, pp. 85–99 [in Russian]. doi: 10.15838/esc.2014.6.36.7

23. Goletsis Y., Chletsos M. (2011) Measuring of development and regional disparities in Greek periphery: A multivariate approach. *Socio-Economic Planning Sciences*, vol. 45, no. 4, pp. 174–183. doi: 10.1016/j.seps.2011.06.002

24. Martín-Lȯpez B., Palomo I., García-Llorente M., Imesta-Arandia I., Castro A.J., Del Amo D.G., Cȯmez-Baggethun E., Montes C. (2017) Delineating boundaries of social-ecological systems for landscape planning: A comprehensive spatial approach. *Land Use Policy*, vol. 66, pp. 90–104. doi: 10.1016/j.landusepol.2017.04.040

25. Chivu L. (2019) Local entrepreneurship and social services in Romania. Territorial analysis. *European Research on Management and Business Economics*, vol. 25, no. 2, pp. 79–86. doi: 10.1016/j.iedeen.2019.04.001

26. Sogacheva O.V. (2016) Klasternyj analiz kak instrument upravleniya social'no-ekonomicheskim razvitiem regiona (na primere Central'nogo federal'nogo okruga) [Cluster analysis as a tool for managing the socio-economic development of the region (On the example of the Central Federal District)]. *Teoriya i praktika servisa: ekonomika, social'naya sfera, tekhnologii*, no. 1, pp. 43–46 [in Russian].

27. Zholudeva I.I., Mel'nichenko N.F., Kozlov G.E. (2014) Primenenie klasternogo analiza dlya ocenki social'no-ekonomicheskogo razvitiya regionov na primere CFO i Yaroslavskoj oblasti [The use of cluster analysis to assess the socio-economic development of regions on the example of the Central Federal District and the Yaroslavl Region]. *Ekonomika, Statistika i Informatika*, no. 1, pp. 144–148 [in Russian].

28. Filonova E.S., Bukreeva Yu.V. (2013) Analiz sostoyaniya regionov Central'nogo federal'nogo okruga po pokazatelyam predpriyatij malogo i srednego biznesa [Analysis of the state of the regions of the Central Federal District by indicators of small and medium-sized businesses]. *Vestnik Finansovogo universiteta*, no. 6, pp. 35–47 [in Russian].

29. Uzdenov U.A. (2014) Intellektual'naya sistema ocenki kreditosposobnosti regionov. Chast' 1. Mnogomernyj statisticheskij analiz [Intelligent system for assessing the creditworthiness of regions. Part 1. Multivariate statistical analysis]. *Nauchnyj zhurnal KubGAU*, no. 104, pp. 957–968 [in Russian].

30. Brida J.G., Garrido N., Mureddu F. (2014) Italian economic dualism and convergence clubs at regional level. *Quality and Quantity*, vol. 48, pp. 439–456. doi: 10.1007/s11135-012-9779-z

31. Ratigan K. (2017) Disaggregating the Developing Welfare State: Provincial social policy regimes in China. *World Development*, vol. 98, pp. 467–484. doi: 10.1016/j.worlddev.2017.05.010

32. Agovino M., Rapposelli A. (2017) Regional performance trends in providing employment for persons with disabilities: Evidence from Italy. *Social Indicators Research*, vol. 130, pp. 593–615. doi: 10.1007/s11205-015-1186-0

33. Dutta I., Das A. (2019) Exploring the dynamics of spatial inequality through the development of sub-city typologies in English Bazar urban agglomeration and its peri urban areas. *GeoJournal*, vol. 84, pp. 829–849. doi: 10.1007/s10708-018-9895-y

34. Chavent M., Kuentz-Simonet V., Labenne A., Saracco J. (2018) ClustGeo: An R package for hierarchical clustering with spatial constraints. *Computational Statistics*, vol. 33, pp. 1799–1822. doi: 10.1007/s00180-018-0791-1

35. Mookherjee D., White J. (2011) Urban-regional dualism in India: An exploration of developmental indicators across urban size classes. *Asian Geographer*, vol. 28, no. 1, pp. 21–31. doi: 10.1080/10225706.2011.577976

36. Kovanova E.S. (2013) Klasternyj analiz v reshenii zadach tipologii regionov Rossii po urovnyu i intensivnosti vnutrennej trudovoj migracii [Cluster analysis in solving problems of

the typology of Russian regions by the level and intensity of internal labor migration]. *Vestnik NGUEU*, no. 4, pp. 166–175 [in Russian].

37. Abaev V.A., Shahov A.V. (2010) Metodicheskij podhod k opredeleniyu tipichnyh sel'skohozyajstvennyh rajonov Lipeckoj oblasti [Methodical approach to the definition of typical agricultural areas of the Lipetsk Region]. *Vestnik FGOU VPO MGAU*, no. 6, pp. 94–99 [in Russian].

38. Cziráky D., Sambt J., Rovan J., Puljiz J. (2006) Regional development assessment: A structural equation approach. *European Journal of Operational Research*, vol. 174, no. 1, pp. 427–442. doi: 10.1016/j.ejor.2005.03.012

39. Novo-Corti I., Barreiro-Gen M. (2015) Public policies based on social networks for the introduction of technology at home: Demographic and socioeconomic profiles of households. *Computers in Human Behavior*, vol. 51, pp. 1216–1228. doi: 10.1016/j.chb.2014.12.040

40. Flor M.A., Klarl T. (2017) On the cyclicity of regional house prices: New evidence for U.S. metropolitan statistical areas. *Journal of Economic Dynamics and Control*, vol. 77, pp. 134–156. doi: 10.1016/j.jedc.2017.02.001

41. Kyriakopoulos G.L., Arabatzis G., Tsialis P., Ioannou K. (2018) Electricity consumption and RES plants in Greece: Typology of regional units. *Renewable Energy*, vol. 127, pp. 134–144. doi: 10.1016/j.renene.2018.04.062

42. Krišto J., Dumičić K., Curković M. (2014) Banking business indicators in Croatian economic surroundings. *Economy of Eastern Croatia: Yesterday, Today, Tomorrow*, vol. 3, pp. 572–581.

43. Rahman S., Mohiuddin H., Kafy A.-A., Shell P.K., Di L. (2019) Classification of cities in Bangladesh based on remote sensing derived spatial characteristics. *Journal of Urban Management*, vol. 8, no. 2, pp. 206–224. doi: 10.1016/j.jum.2018.12.001

44. Litvinova G.P., Lisicyn A.E. (2018) Ocenka social'no-ekonomicheskogo potenciala regionov Sibirskogo federal'nogo okruga [Assessment of the socio-economic potential of the regions of the Siberian Federal District]. *Vestnik Kemerovskogo gosudarstvennogo universiteta. Seriya: Politicheskie, sociologicheskie i ekonomicheskie nauki*, no. 2, pp. 114–121 [in Russian]. doi:10.21603/2500-3372-2018-2-114-121

45. Jašková D., Havierniková K. (2020) The human resources as an important factor of regional development. International Journal of Business and Society, vol. 21, no. 3, pp. 1464–1478.

46. Andrejiova M., Kimakova Z. (2020) The comparison of transport infrastructures in individual Slovak regions by applying PCA and cluster analysis. *Acta Logistica*, vol. 7, no. 4, pp. 225–234. doi: 10.22306/al.v7i4.182

47. Berry B.J.L. (1961) A method for deriving multi-factor uniform regions. *Przeglad Geograficzny*, vol. 33, no. 2, pp. 263–282.

48. Cliff A.D., Haggett P. (1970) On the efficiency of alternative aggregation in region-building problems. *Environment and Planning A: Economy and Space*, vol. 2, no. 3, pp. 285–294. doi: 10.1068/a020285

49. Shrader Yu.A. *Ravenstvo, skhodstvo, poryadok* [Equality, similarity, order], Moscow: Nauka, 1971 [in Russian].

50. Kvičalová J., Mazalová V., Široký J. (2014) Identification of the difference between the regions of the Czech Republic based on the economic characteristics. *Procedia – Economics and Finance*, vol. 12, pp. 343–352. doi: 10.1016/S2212-5671(14)00354-2

51. Tatarczak A., Boichuk O. (2018) The multivariate techniques in evaluation of unemployment analysis of Polish regions. *Oeconomica Copernicana*, vol. 9, no. 3, pp. 361–380. doi: 10.24136/oc.2018.018

52. Ma Y., Zhuang X., Li L. (2011) Research on the relationships of the domestic mutual investment of China based on the cross-shareholding networks of the listed companies. *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 4, pp. 749–759. doi: 10.1016/j.physa.2010.10