COMPUTER RESEARCH AND MODELING
2020 VOL. 12 NO. 6 P. e1501–e1513
DOI: 10.20537/2076-7633-2020-12-6-1501-1513

Ки&М

MODELS OF ECONOMIC AND SOCIAL SYSTEMS

UDC: 519.237.8

Assessing the validity of clustering of panel data by Monte Carlo methods (using as example the data of the Russian regional economy)

I. L. Kirilyuk[1,a], O. V. Senko[2]

[1] Institute of Economics, Russian Academy of Sciences,
32 Nakhimovskii pr., Moscow, 117218, Russia

[2] Federal Research Center Computer Science and Control, Russian Academy of Sciences,
44/2 Vavilova st., Moscow, 119333, Russia

E-mail: [a] igokir@rambler.ru

Received 04.05.2020, after completion — 02.09.2020.
Accepted for publication 18.09.2020.

The paper considers a method for studying panel data based on the use of agglomerative hierarchical clustering — grouping objects based on the similarities and differences in their features into a hierarchy of clusters nested into each other. We used 2 alternative methods for calculating Euclidean distances between objects — the distance between the values averaged over observation interval, and the distance using data for all considered years. Three alternative methods for calculating the distances between clusters were compared. In the first case, the distance between the nearest elements from two clusters is considered to be distance between these clusters, in the second — the average over pairs of elements, in the third — the distance between the most distant elements. The efficiency of using two clustering quality indices, the Dunn and Silhouette index, was studied to select the optimal number of clusters and evaluate the statistical significance of the obtained solutions. The method of assessing statistical reliability of cluster structure consisted in comparing the quality of clustering on a real sample with the quality of clustering on artificially generated samples of panel data with the same number of objects, features and lengths of time series. Generation was made from a fixed probability distribution. At the same time, simulation methods imitating Gaussian white noise and random walk were used. Calculations with the Silhouette index showed that a random walk is characterized not only by spurious regression, but also by "spurious clustering". Clustering was considered reliable for a given number of selected clusters if the index value on the real sample turned out to be greater than the value of the 95% quantile for artificial data. A set of time series of indicators characterizing production in the regions of the Russian Federation was used as a sample of real data. For these data only Silhouette shows reliable clustering at the level $p < 0.05$. Calculations also showed that index values for real data are generally closer to values for random walks than for white noise, but it have significant differences from both. Since three-dimensional feature space is used, the quality of clustering was also evaluated visually. Visually, one can distinguish clusters of points located close to each other, also distinguished as clusters by the applied hierarchical clustering algorithm.

Keywords: clustering validity, panel data, mesoeconomics, regional economics

Citation: *Computer Research and Modeling*, 2020, vol. 12, no. 6, pp. 1501–1513 (Russian).

КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2020 Т. 12 № 6 С. e1501–e1513
DOI: 10.20537/2076-7633-2020-12-6-1501-1513

Ки&М

МОДЕЛИ ЭКОНОМИЧЕСКИХ И СОЦИАЛЬНЫХ СИСТЕМ

УДК: 519.237.8

Оценка качества кластеризации панельных данных с использованием методов Монте-Карло (на примере данных российской региональной экономики)

И. Л. Кирилюк[1,a], О. В. Сенько[2]

[1] Институт экономики Российской академии наук,
Россия, 117218, г. Москва, Нахимовский проспект, д. 32

[2] Федеральный исследовательский центр «Информатика и управление» Российской академии наук,
Россия, 119333, г. Москва, ул. Вавилова, д. 44/2

E-mail: [a] igokir@rambler.ru

Получено 04.05.2020, после доработки — 02.09.2020.
Принято к публикации 18.09.2020.

В работе рассматривается метод исследования панельных данных, основанный на использовании агломеративной иерархической кластеризации — группировки объектов на основании сходства и различия их признаков в иерархию вложенных друг в друга кластеров. Применялись 2 альтернативных способа вычисления евклидовых расстояний между объектами — расстояния между усредненными по интервалу наблюдений значениями и расстояния с использованием данных за все рассматриваемые годы. Сравнивались 3 альтернативных метода вычисления расстояний между кластерами. В первом случае таким расстоянием считается расстояние между ближайшими элементами из двух кластеров, во втором — среднее по парам элементов, в третьем — расстояние между наиболее удаленными элементами. Исследована эффективность использования двух индексов качества кластеризации — индекса Данна и Силуэта для выбора оптимального числа кластеров и оценки статистической значимости полученных решений. Способ оценивания статистической достоверности кластерной структуры заключался в сравнении качества кластеризации, на реальной выборке с качеством кластеризаций на искусственно сгенерированных выборках панельных данных с теми же самыми числом объектов, признаков и длиной рядов. Генерация производилась из фиксированного вероятностного распределения. Использовались способы симуляции, имитирующие гауссов белый шум и случайное блуждание. Расчеты с индексом Силуэт показали, что случайное блуждание характеризуется не только ложной регрессией, но и ложной кластеризацией. Кластеризация принималась достоверной для данного числа выделенных кластеров, если значение индекса на реальной выборке оказывалось больше значения 95%-ного квантиля для искусственных данных. В качестве выборки реальных данных использован набор временных рядов показателей, характеризующих производство в российских регионах. Для этих данных только Силуэт показывает достоверную кластеризацию на уровне $p < 0.05$. Расчеты также показали, что значения индексов для реальных данных в целом ближе к значениям для случайных блужданий, чем для белого шума, но имеют значимые отличия и от тех, и от других. Визуально можно выделить скопления близко расположенных друг от друга в трехмерном признаковом пространстве точек, выделяемые также в качестве кластеров применяемым алгоритмом иерархической кластеризации.

Ключевые слова: достоверность кластеризации, панельные данные, мезоэкономика, экономика регионов

1. Introduction

In many fields of knowledge, researchers are dealing with sets of time series, or panel data, grouped into subgroups based on some characteristic. Sometimes such a division of a set of time series into subgroups is not initially obvious, and its identification is the subject of interest of researchers. In this case, the problem of clustering time series arises [Liao, 2005; Aghabozorgi et al., 2015; Ivakhnenko et al., 2007]. Some publications also consider the problem of clustering panel data, for example, in [Kapetanios, 2006; Niu, 2012]. Traditional approaches in the analysis of panel data are largely focused on testing hypotheses about the equality of coefficients in panels; the task of clustering them is rarely posed, although in some cases it can provide more detailed essential information. And ignoring the cluster data structure can lead to incorrect statistical conclusions when studying them.

An important factor in clustering tasks is the assessment of its quality and reliability. To assess the quality of clustering, a variety of indices have been proposed (the literature contains lists of dozens of such indices [Halkidi et al., 2001; Charrad et al., 2014; Sivogolovko, 2011]). These indexes allow you to compare different clustering options. With their help, it is determined, the division into what number of clusters in the investigated volume of the feature space gives the most reliable and pronounced grouping. However, they do not allow directly drawing a conclusion about the reliability of the solutions obtained. There are various approaches to assessing the reliability of clustering. For example, a conclusion about its reliability can be made based on the opinion of experts or if clustering results corresponds to the values of some external factors that are not used in clustering. One of the important ways to establish the reliability of clustering is to evaluate its statistical significance in the sense of the probability of accidentally refuting the null hypothesis of the absence of clustering. Verification is an important element of scientific research. Lack of verification, or its incorrect implementation, can lead to unreasonable and often false conclusions. This statement should undoubtedly apply to all methods of searching for patterns in data, including clustering.

In order to meaningfully interpret the effects associated with clustering, it is necessary to make sure with an acceptable degree of confidence that they exist, that the data is concentrated in several separated areas of the feature space (that is, that the probability of matching the data with the null hypothesis, which implies the generation of all data from one and the same uniform or unimodal distribution that does not have a cluster structure, respectively, is small, for example, less than 0.05). A description of null hypotheses used in validation of clustering results can be found, for example, in [Gordon, 1996; Giancarlo, Utro, 2012].

A measure of the statistical significance of the assumption about the objective existence of clustering, obtained on real data, can be the probability of accidentally reaching or exceeding the value of the corresponding clustering quality assessment index over the values of the clustering quality assessment indices obtained on the data generated under the condition of the null hypothesis.

In statistics, such probabilities are usually called *p*-values. One of the ways to estimate *p*-values is the use of Monte Carlo methods, when the values of the test statistic on real data are compared with the values of the test statistic on data sampled from the distribution in accordance with the null hypothesis using random number generators. This approach is widely used to verify a variety of regression relationships. However, to verify the results of cluster analysis, random sampling is used in our opinion much less frequently, especially in tasks with panel data. At the same time, such tasks have significant specificity associated with the ambiguity of the choice of the null hypothesis.

In particular, the null hypothesis may consist in the fact that all considered time series for each feature are implementations of some random process with the same characteristics (mean, variance, series length, etc.). In this case, the time series actually form one cluster, and the identification of a larger number of clusters is an artifact.

This study demonstrates the verification methodology by a number of alternative ways of existence of more than one cluster in the data on the example of studying a set of regions of the Russian Federation in the space of features characterizing their production functions. The choice of the exam-

ple is due to the fact that earlier the authors used a similar methodology to study the production functions of Russian regions [Kirilyuk, Senko, 2020].

In the works of a number of researchers, for example, in [Aivazyan et al., 2016; Bakhitova et al., 2014; Magomadov, Shamilev, 2014] and in many others, there are options for classifying and clustering Russian regions into groups using production functions, or sets of some economic indicators. Examples of publications where hierarchical clustering is used for these tasks is [Nizhegorodtsev, Goridko, 2014; Sibukaev, 2019]. Interest in this topic is justified by the fact that the identification of sufficiently reliable clusters allows for more correct statistical calculations, and also involves further research to identify the mechanisms that led to their occurrence, which can increase the accuracy of forecasts. For regions from one cluster, it can be useful to develop common recommendations, develop common programs for balanced development.

Our approach allows us to give a mathematical assessment of the validity of the division of regions into clusters when they are produced on the basis of sets of quantitative features.

In publications examining a set of economic objects, in our opinion, the following approaches to the use of cluster analysis can be distinguished:

1) objects are considered without taking into account their heterogeneity;
2) objects are divided into groups, but without the use of cluster analysis;
3) cluster analysis is carried out, but the quality of the resulting clustering is not assessed using the appropriate indices;
4) the quality of clustering is assessed using the appropriate indices, but the problem of assessing the probability of accidental occurrence of high values of indices (which can be considered as false clustering) is not solved;
5) cluster analysis is carried out, clustering quality indices are calculated and the probability of random occurrence of the resulting values is estimated.

Our experience says that the number of publications that can be matched to the item numbers of the above list significantly decreases with the growth of the number.

2. Used data

We have made an assessment of the cluster structure in the space of indicators for 79 regions of the Russian Federation, for which there is the required data set for the period under consideration [Regiony Rossii…, 2017] (based on data for 1996–2014). The panel data used (the same that we used earlier in the aforementioned article [Kirilyuk, Senko, 2020] to construct the production functions of regions) include indicators: Y — gross regional product, I — investment in fixed assets, L — average annual number of people employed in the economy multiplied by the average monthly nominal gross wages employed in the economy. The values were brought to constant prices (a procedure that eliminates the distorting effect of inflation) using consumer price indices. All used features were logarithmized. The average values of the time series of indicators, their variances, trends and other similar characteristics form three-dimensional spaces, which makes it possible to visually assess their cluster structure.

3. Clustering methods

Clustering methods are divided into non-hierarchical, typical of which is, for example, the k-means method, and hierarchical. The authors use agglomerative hierarchical clustering, when, as a result of the work of the corresponding algorithm, a hierarchy (tree) of nested clusters is created. The advantage of hierarchical clustering over alternative approaches is that when using it, you do not need to make a priori assumptions about the number of clusters.

There are a number of metrics that characterize the distance between time series (for example, dynamic time warping, Manhatton distance, etc.). In this article, for data characterizing objects that

develop almost synchronously in time and constitute a relatively short time series, the authors preferred to use the Euclidean metric. There are two alternative approaches using the Euclidean metric.

1. Calculation of the time average values of the series and the subsequent application of the cluster analysis algorithm to them. Clustering is performed in a space of three features, which are the average values over the observation interval of the indicators *Ln* (*Y*), *Ln* (*I*), *Ln* (*L*). The distances between regions *i* and *j* are calculated in this case by the formula:

$$d_{xij} = \left[\sum_{p=1}^{3} \left(\overline{x}_{jp} - \overline{x}_{ip} \right)^2 \right]^{1/2}, \quad (1)$$

where the bar over x_{ip}, x_{jp} means the time averaging of the values of the indicators *Ln* (*Y*) for $p = 1$, *Ln* (*I*) for $p = 2$, and *Ln* (*L*) for $p = 3$ for these regions.

2. Calculation of distances between three-component time series using the differences of their values for all 19 years of the considered observation period and clustering using these distances:

$$\left[d_{xij} = \sum_{p=1}^{3} \sum_{t=1}^{19} \left(x_{jpt} - x_{ipt} \right)^2 \right]^{1/2}, \quad (2)$$

where x_{ipt}, x_{jpt} are the values of the features of the *i*-th and *j*-th regions in the year *t*.

There are a number of alternative methods for determining the inter-cluster distance in agglomeration. In this work, three methods were used: “complete”, “average”, “single”, where the distance between clusters is defined as the distance between the most distant elements of two clusters, the average distance between all pairs of elements and the distance between the closest elements. Algorithms “complete” find more compact clusters, and “single”, on the contrary, clusters of complex shape, elongated, and they are more sensitive to noise.

4. Indexes for assessing the quality of clustering

Of the many existing indices for assessing the quality of clustering, we have selected for research (as the most popular) two: Dunn's index [Dunn, 1974] and Silhouette [Rousseeuw, 1987].

The Dunn index is used here in its original version (there are a number of its modifications) and is determined by the formula:

$$D = \min_{i,j \in \{1 \ldots c\}, i \neq j} \left\{ \frac{d(c_i, c_j)}{\max_{k \in (1 \ldots c)} \operatorname{diam}(c_k)} \right\}, \quad (3)$$

where d is the distance between clusters c_i, c_j; $\operatorname{diam}(c_k)$ — maximum distance between elements of one cluster.

The silhouette of the entire cluster structure (Silhouette Width Criterion — SWC) is determined by the formula:

$$swc = \frac{1}{N_x} \sum_{j=1}^{N_x} S_{xj}, \quad (4)$$

as divided by the number of elements in the clustered set N_x the sum of the Silhouettes of each individual element, determined by the formula:

$$S_{xj} = \frac{b_{pj} - a_{pj}}{\max(a_{pj}, b_{pj})}, \quad (5)$$

where a_{pj} is the average distance from an object to other objects in its cluster, b_{pj} is the average distance from an object to other objects in the nearest other cluster.

For Silhouette (4)–(5), as well as for Dunn's index (3), the rule is fulfilled: the higher the clustering quality for a given number of clusters, the higher the index value.

5. Method for assessing the statistical significance of clustering

Let us describe the algorithm used in this article for assessing the reliability of clustering using Monte Carlo methods. Pseudosamples are generated to simulate *Ln* (*Y*), *Ln* (*I*), *Ln* (*L*) (independently of each other). They correspond to two variants of the null hypothesis about the absence of clustering, or, which is the same, about the existence of one single cluster. There are 2 options for generating pseudo-samples with the row length equal to the length of the rows of the used real data (5000 pseudo-samples each):

1) rows defined by the formula:

$$x_t = e_t, \tag{6}$$

where e_t is the white noise *iid* with a normal distribution, the length of the series, the mean values and variances of the process realizations are taken equal to the regionally averaged values of the real investigated time series of features;

2) rows defined by the formula:

$$x_{t+1} = x_t + e_{t+1}, \tag{7}$$

which have the property of stochastic nonstationarity, can demonstrate the effect of false regression [Granger, Newbold, 1974] and are referred to as random walk processes. The initial values for them are generated from normal distributions with mean values and variances equal to the averaged mean values and variances for the aggregates of real data, the length of the series and the variance of the increments are equal to the corresponding regional-averaged values for the real series of features, the average value of the increments is zero. Random walks based on the results of a number of studies, for example, [Nelson, Plosser, 1982], describe many time series of economic data much better than stationary processes such as white noise.

The resulting pseudosamples, like real data, are investigated by the methods described above for assessing the quality of clustering.

For each of the described clustering options, as a result of calculations, graphs of the dependence of the values of the used indices for assessing the quality of clustering on the assumed number of clusters were obtained.

On each graph of the dependence of the index on the number of clusters, 7 types of data are plotted: the medians of the values of the indices and the boundaries of their confidence intervals at the 5% level for simulations with white noise and random walks, as well as indices corresponding to real data.

The reliability of clustering is estimated by comparing the values of the indices of real data with appropriate index quantiles from the pseudo-samples used. For example, if the indices of the real data are larger than the indices corresponding to the 95% quantiles of the pseudo-samples, the clustering is assumed to be reliable at $p = 0.05$.

6. Results of calculations

6.1. Data visualization

Since the used feature space has only three dimensions, the cluster structure in it can be easily assessed by direct visualization. In Fig. 1 shows three projections of a set of indicator values for all regions for all the years under consideration (left graphs) as well as three projections of a set of indicator values for all regions, averaged over the time interval used (right graphs).

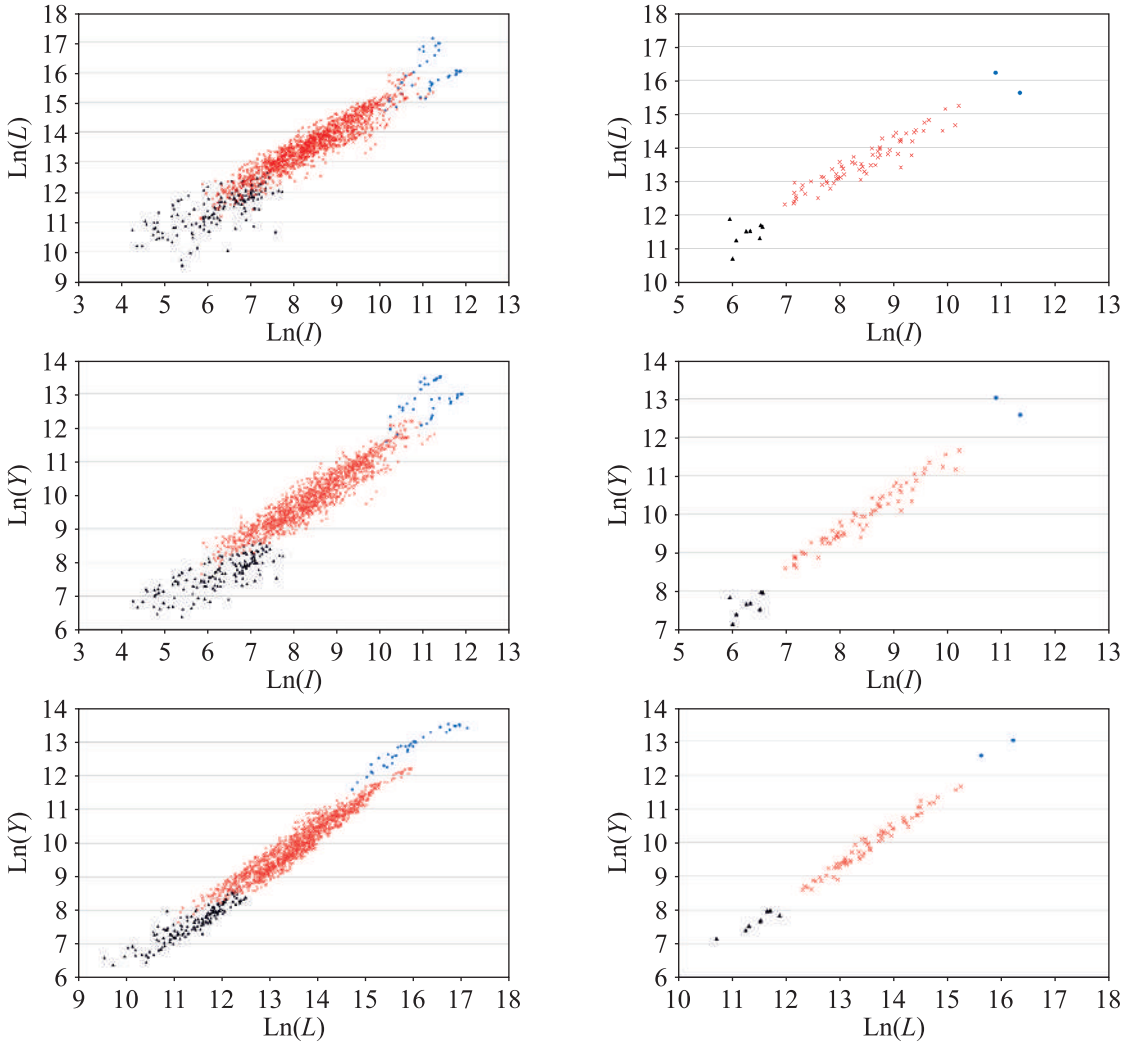

Fig. 1. Visualization of the cluster structure of Russian regions in the space of attributes *Ln* (*Y*), *Ln* (*I*), *Ln* (*L*) with one of the options for dividing into 3 clusters, indicated by symbols of different colors

As seen from Fig. 1, the data are located on all plots along oblique lines (due to the significant correlation between the considered features). At the same time, visually it is possible to distinguish subgroups of points located somewhat apart from the rest, and perceived subjectively as clusters. Let's evaluate the quality of clustering by calculating indices of real data and comparing them with indices of simulations. The results of this assessment are presented below.

6.2. Clustering by distances calculated by formula (1)

In Fig. 2 shows the dependence of the Dunn index on the assumed number of clusters. The results are presented only for the "average" and "complete" methods, since no qualitatively new effects were revealed for the "single" method.

In Fig. 2 and similar figures below, on the left are the results of calculation by the "average" method, on the right — by the "complete" method.

In Fig. 2 and in the following figures, the following symbols are adopted:

▲ — 250th in rank (that is, 95% quantiles), 2500th in rank (that is, median) and 4750th in rank (that is, 5% quantile) values of the clustering quality assessment index for simulation by the formula (6);

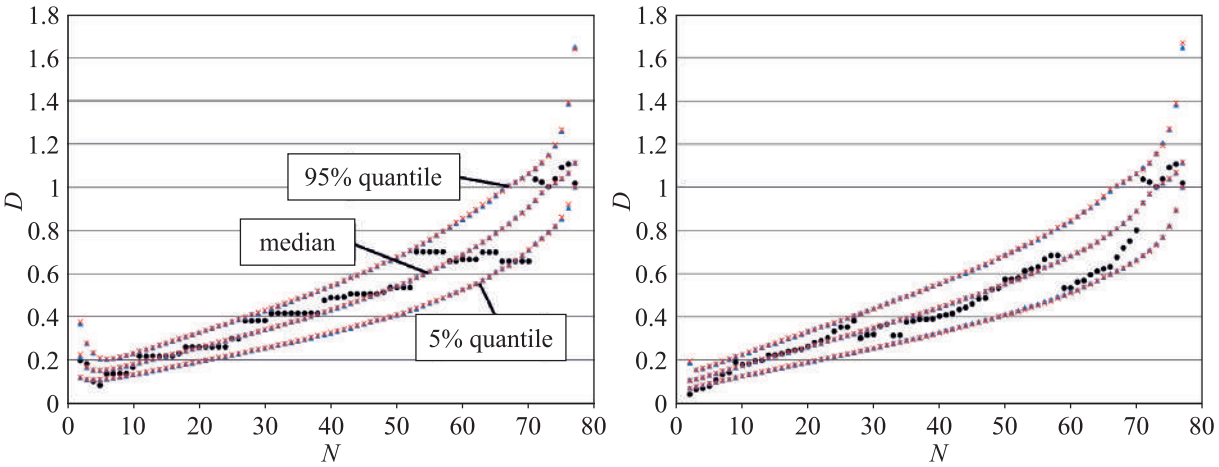

Fig. 2. Dependence of the Dunn index on the number of selected clusters for the case of clustering using distances (1)

× — 250th in rank, 2500th in rank and 4750th in rank of the value of the clustering quality assessment index for simulation according to the formula (7);

● — real values of the index.

Fig. 2 that the values of the Dunn index, except for those corresponding to the small N, increase for all types of data with an increase in the number of clusters. For the "average" method, the violation of the monotonicity of the growth of the index is more pronounced at small N. The values of the index for simulations by formulas (6) and (7) almost merge on both graphs. The index values for real data are nowhere above the upper bounds of the simulated confidence interval. However, in this case, this should not be taken as an unambiguous indication that there is no data clustering. There are examples for the Dunn index when it does not distinguish clearly visible, but too close to each other, clusters.

In Fig. 3 shows the dependence of the values of the Silhouette index on the assumed number of clusters. The same patterns are seen as for the Dunn index: a gradual increase in the value of the index except for the smallest N, the merging of values for simulations according to formulas (6) and (7). The difference from the results for the Dunn index is that Silhouette, especially in the case of using the "average" method, shows a significantly better reliability of clustering of real data in the area of a small number of clusters.

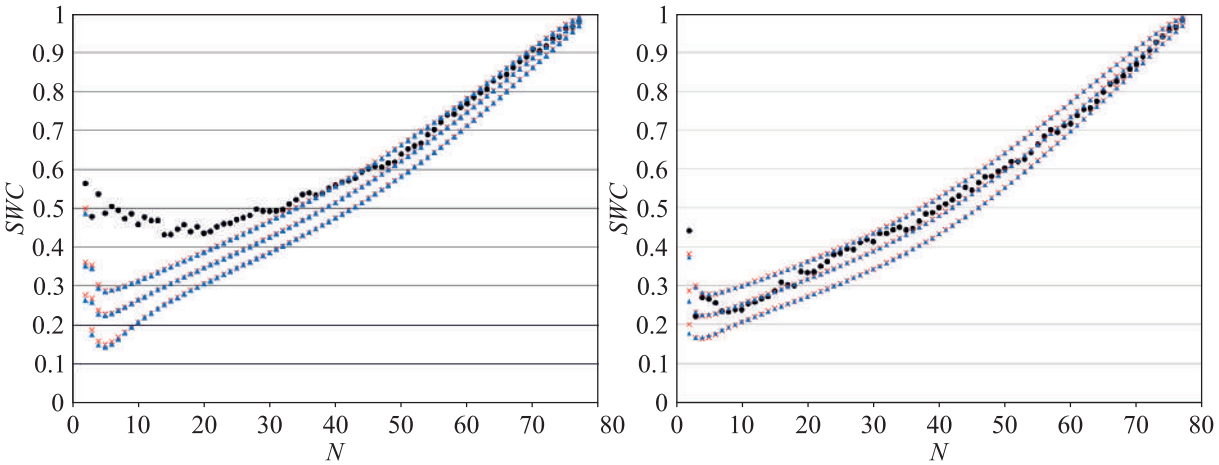

Fig. 3. Dependence of the Silhouette index on the number of selected clusters during clustering using distances (1)

6.3. Clustering by distances calculated by formula (2)

In this case, the distances between time series of real data differ significantly both from the distances for simulations (6) and from the distances for simulations (7), as it is seen in the Fig. 4.

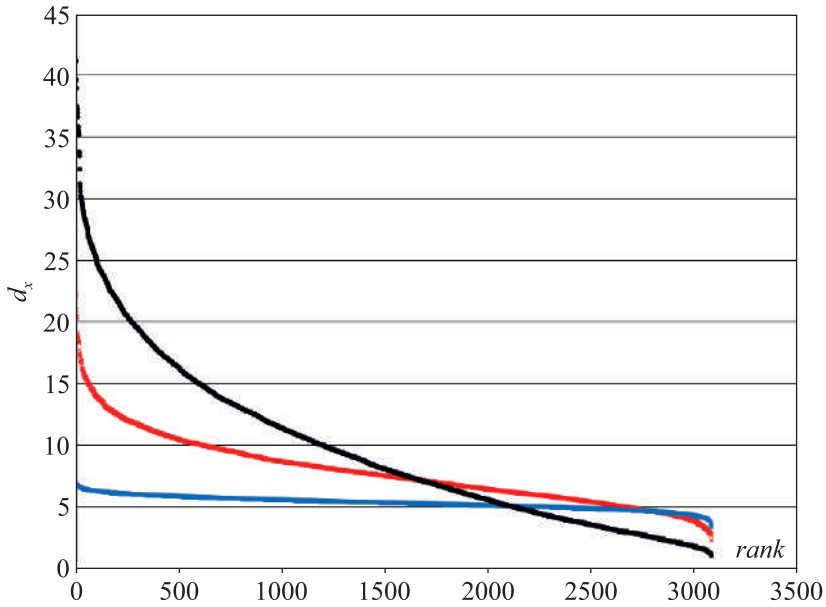

Fig. 4. Rank distributions of distances d_x, calculated by the formula (2)

In Fig. 4 shows the rank distributions of the distances between the three-component time series, calculated by formula (2) for the following types of data (listed in the order of the ordinate from top to bottom): real data, simulations by formula (7), simulations by formula (6). Fig. 4 demonstrate that the distribution for real data differs significantly not only from the distribution for simulations (6), but also from the distribution for simulations (7), which have an intermediate position in the range of values.

In contrast to the simulation graphs in Fig. 2–3, Fig. 5, the Dunn index values for simulations (6) and (7) are clearly separated from each other. Fig. 5 demonstrate that the curve of the index values for real data lies much closer to the curves obtained for simulations (7) than to the curves obtained for simulations (6), although in a significant number of cases it is below 5% quantile of simulations (7).

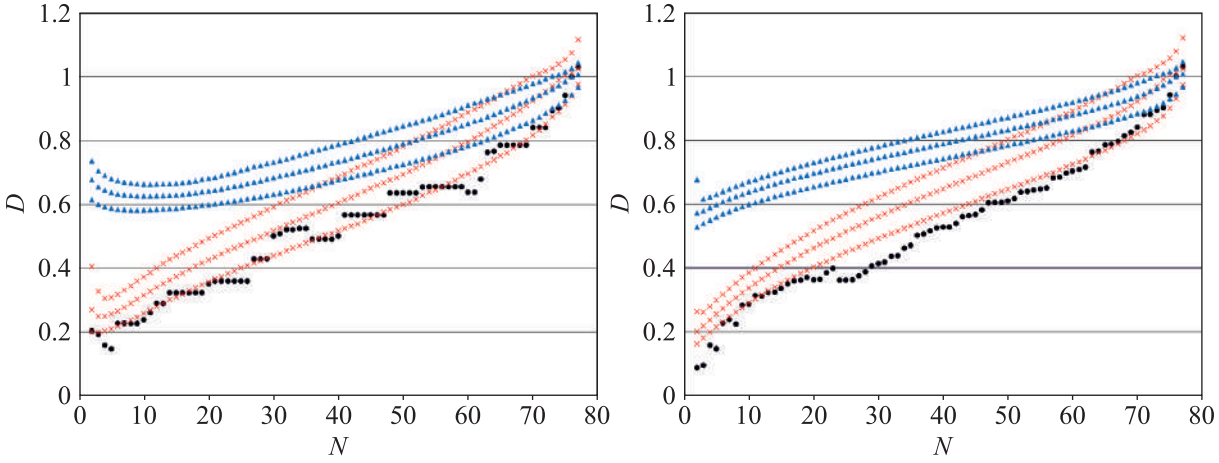

Fig. 5. Dependence of the Dunn index on the number of allocated clusters during clustering using distances (2)

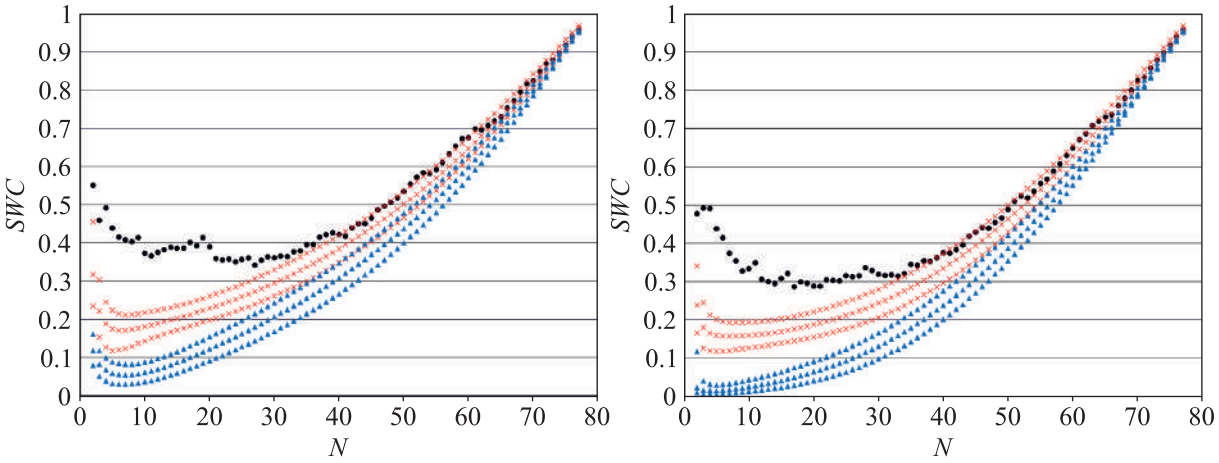

Fig. 6. Dependence of the Silhouette index on the number of selected clusters during clustering using distances (2)

Figure 6 shows the dependence of the Silhouette index values on the estimated number of clusters for the clustering option using distances (2). Both plots demonstrate reliable clustering at small N. It is unexpected that, in contrast to the case shown in Fig. 5 the indices for random walks are higher than those for white noise. This demonstrates that the application of clustering quality indexes to weakly clustered time series and panel data requires some caution, since the behavior of the indexes in this case may have nontrivial features that are not related to their purpose to assess the clustering quality.

6.4. Correspondence between regions and clusters

For all the clustering options described above, it was tested how the regions are distributed across two and three clusters. The following results were obtained:

When dividing into 2 clusters, Moscow and the Tyumen region are distinguished everywhere, which correspond to two points relatively far from the rest in the upper left parts of the graphs in Fig. 1. However, in cases using the “complete” method, the algorithm adds additional regions to their cluster.

For the case calculated by formula (1), these are: Krasnodar Territory, Krasnoyarsk Territory, Moscow Region, St. Petersburg, Sverdlovsk Region, Republic of Tatarstan.

For the case calculated by formula (2), in addition to the eight regions of the previous case, the following are added to the cluster: Arkhangelsk Region, Republic of Bashkortostan, Chelyabinsk Region, Khabarovsk Territory, Irkutsk Region, Republic of Sakha (Yakutia), Kemerovo Region, Komi Republic, Leningrad Region, Nizhny Novgorod Region, Novosibirsk Region, Omsk Region, Orenburg Region, Perm Region, Primorsky Region, Rostov Region, Sakhalin Region, Samara Region, Saratov Region, Stavropol Region, Volgograd Region, Vologda Region, Voronezh Region.

When dividing into 3 clusters, a cluster of regions appears in the lower left part of the graphs in Fig. 1. It is common for all used clustering options. It includes the following regions: Republic of Adygea, Republic of Altai, Chukotka Autonomous Okrug, Jewish Autonomous Region, Republic of Ingushetia, Republic of Kalmykia, Karachay-Cherkess Republic, Republic of Tyva.

Thus, the results obtained using formulas (1) and (2) differ from each other only when using the “complete” method. The “complete” method selects 2 clusters in a significantly different way than the “single” and “average” methods.

In all cases, plotting visually confirmed the adequacy of clustering. That is, even when the reliability of clustering is not confirmed by indices (in our case, the Dunn index), the distribution of regions by clusters can be generally recognized as adequate.

7. Conclusion

As a result of the study of the quality of agglomerative hierarchical clustering of panel data characterizing production processes in the regions of the Russian Federation by a number of alternative methods, it was revealed that the calculated degree of clustering reliability significantly depends on the indices used, the choice of the type of distances and methods for calculating the distance between clusters. Let us formulate the most significant conclusions in our opinion with recommendations for further research.

- When working with panel data, the possibility of their cluster structure is often not taken into account. However, ignoring the cluster structure can lead to significant distortions in econometric modeling.
- When processing data, one should not only find out the optimal partitioning of them into clusters, but also establish its reliability.
- It is advisable to use a set of different indexes for assessing the quality of clustering, and not be limited to any one index, so as not to come to false generalizations in the conclusions.
- When choosing null hypotheses to test the quality of clustering of time series, it is necessary to take into account the fact of their (non) stationarity, select null hypotheses about the absence of clustering, corresponding to the nature of the studied time series. The nonstationarity of time series can lead not only to false regression investigated by the authors, for example, in [Kirilyuk, Senko, 2020], but also to false clustering. Not knowing the problem of false regression [Granger, Newbold, 1974] before K. Granger's studies led to countless fake results. The use of additional verification methods made it possible to limit the flow of such "results". Taking into account the possibility of false clustering should increase the scientific significance of cluster analysis as an evidence-based research method. Currently, this is primarily a exploratory method.
- At the same time, the Silhouette and the Dunn indices give different answers to the question which has more pronounced clustering — a set of realizations of white noise (6), or random walks (7). In our opinion, this indicates a certain conventionality of the intuitive concept of the "clarity" of clustering in cases when this clarity is weak. For data with a more pronounced clustering, which corresponds, for example, to a significant excess of the distances between clusters over their characteristic sizes, both indices in our calculations gave the expected peaks corresponding to the objective number of clusters.
- Using clustering indices, it is possible not only to assess the clustering quality of the tested set of time series, but also to assess the degree of its compliance with alternative null hypotheses (for example, make an assumption about whether the series are stationary).
- The investigated empirical dataset differs significantly by properties from both typical realizations of white noise (6) and from typical realizations of random walks (7).
- Using the Silhouette index confirms the presence of a reliable division of regions into several clusters, which can also be assessed visually. Therefore, in our opinion, we can talk about the presence of a cluster structure, although not very pronounced, for the considered set of features that characterize production processes in the regions of the Russian Federation.

It is of interest to continue research using the approach described in the article with the joint use of a larger number of indexes for assessing the quality of clustering, other sets of features, including multidimensional ones, where checking the validity of clustering through visual assessment is difficult.

All calculations, the results of which are used in this article, were carried out using the R language, in particular, the packages NbClust [Charrad et al., 2014] and TSclust [Montero, Vilar, 2014].

References

Айвазян С. А., Афанасьев М. Ю., Кудров А. В. Метод кластеризации регионов РФ с учетом отраслевой структуры ВРП // Прикладная эконометрика. — 2016. — Т. 41. — С. 24–46.
Aivazyan S. A., Afanas'ev M. Yu., Kudrov A. V. Metod klasterizatsii regionov RF s uchetom otraslevoi struktury VRP [The method of clustering regions of the Russian Federation taking into account the sectoral structure of the GRP] // Prikladnaya ekonometrika [Applied econometrics]. — 2016. — Vol. 41. — P. 24–46 (in Russian).

Бахитова Р. Х., Ахметшина Г. А., Лакман И. А. Панельное моделирование объема выпуска продукции для регионов России // Управление большими системами. — 2014. — Т. 50. — С. 99–109.
Bakhitova R. Kh., Akhmetshina G. A., Lakman I. A. Panel'noe modelirovanie ob"ema vypuska produktsii dlya regionov Rossii [Panel modeling of output for regions of Russia] // Upravlenie bol'shimi sistemami [Large-scale systems control]. —2014. — Vol. 50. — P. 99–109 (in Russian).

Ивахненко А. А., Каневский Д. Ю., Рудева А. В., Стрижов В. В. Выявление групп объектов, описанных набором многомерных временных рядов // Математические методы распознавания образов. — 2007. — Т. 13 (1). — С. 134–137.
Ivakhnenko A. A., Kanevskii D. Yu., Rudeva A. V., Strizhov V. V. Vyyavlenie grupp ob"ektov, opisannykh naborom mnogomernykh vremennykh ryadov [Identification of groups of objects described by a set of multidimensional time series] // Matematicheskie metody raspoznavaniya obrazov [Mathematical methods for pattern recognition]. — 2007. — Vol. 13 (1). — P. 134–137 (in Russian).

Кирилюк И. Л., Сенько О. В. Выбор моделей оптимальной сложности методами Монте-Карло (на примере моделей производственных функций регионов Российской Федерации) // Информатика и ее применения. — 2020. — Т. 14, вып. 2. — С. 111–118.
Kirilyuk I. L., Sen'ko O. V. Vybor modelei optimal'noi slozhnosti metodami Monte-Karlo (na primere modelei proizvodstvennykh funktsii regionov Rossiiskoi Federatsii) [Selection of optimal complexity models by methods of nonparametric statistics (on the example of production functionmodels of the regions of the Russian Federation)] // Informatika i ee primeneniya [Informatics and Applications]. — 2020. — Vol. 14, iss. 2. — P. 111–118 (in Russian).

Магомадов Н. С., Шамилев С. Р. Анализ динамики ВРП регионов РФ производственными функциями // Современные проблемы науки и образования. — 2014. — № 6.
Magomadov N. S., Shamilev S. R. Analiz dinamiki VRP regionov RF proizvodstvennymi funktsiyami [Analysis of the dynamics of GRP of the regions of the Russian Federation by production functions] // Sovremennye problemy nauki i obrazovaniya [Modern problems of science and education]. — 2014. — No. 6 (in Russian).

Нижегородцев Р. М., Горидько Н. П. Инновационные факторы экономического роста регионов России: кластерный анализ // Труды XII Всероссийского совещания по проблемам управления (ВСПУ-2014, Москва). — М.: ИПУ РАН, 2014. — С. 6088–6093.
Nizhegorodtsev R. M., Gorid'ko N. P. Innovatsionnye faktory ekonomicheskogo rosta regionov Rossii: klasternyi analiz [Innovative factors of economic growth in the regions of Russia: cluster analysis] // Trudy XII Vserossiiskogo soveshchaniya po problemam upravleniya (VSPU-2014, Moskva) [Proceedings of XII All-Russian Conference on Control Problems]. — Moscow: ICS RAS, 2014. — P. 6088–6093 (in Russian).

Регионы России. Социально-экономические показатели. 2017 // Стат. сб. / Росстат. — М., 2017.
Regiony Rossii. Sotsial'no-ekonomicheskie pokazateli [Regions of Russia. Socio-economic indicators]. 2017 // Stat. sb. / Rosstat. — Moscow, 2017 (in Russian).

Сибукаев Э. Ш. Изучение регионов России посредством иерархического метода кластерного анализа и данных о производстве // Университетская наука. — 2019. — № 2 (8). — С. 86–93.
Sibukaev E. Sh. Izuchenie regionov Rossii posredstvom ierarkhicheskogo metoda klasternogo analiza i dannykh o proizvodstve [Study of Russian regions by means of hierarchical method of cluster analysis and production data] // Universitetskaya nauka [University science]. — 2019. — No. 2 (8). — P. 86–93.

Сивоголовко Е. В. Методы оценки качества четкой кластеризации // Компьютерные инструменты в образовании. — 2011. — № 4. — С. 14–31.
Sivogolovko E. V. Metody otsenki kachestva chetkoi klasterizatsii [Hard clustering validation methods] // Komp'yuternye instrumenty v obrazovanii [Computer tools in education]. — 2011. — No. 4. — P. 14–31 (in Russian).

Aghabozorgi S., Shirkhorshidi A. S., Wah T. Y. Time-series clustering — A decade review // Information Systems. — 2015. — Vol. 53. — P. 16–38. — https://doi.org/10.1016/j.is.2015.04.007

Charrad M., Ghazzali N., Boiteau V., Niknafs A. NbClust: An *R* Package for Determining the Relevant Number of Clusters in a Data Set // Journal of Statistical Software. — 2014. — Vol. 61, No. 6. — P. 1–36. — https://doi.org/10.18637/jss.v061.i06

Dunn J. Well Separated Clusters and Optimal Fuzzy Partitions // Journal Cybernetics. — 1974. — Vol. 4, No. 1. — P. 95–104. — https://doi.org/10.1080/01969727408546059

Giancarlo R., Utro F. Algorithmic paradigms for stability-based cluster validity and model selection statistical methods, with applications to microarray data analysis // Theoretical Computer Science. — 2012. — Vol. 428. — P. 58–79. — https://doi.org/10.1016/j.tcs.2012.01.024

Gordon A. D. Null Models in Cluster Validation // Gaul W., Pfeifer D. (eds) From Data to Knowledge. Studies in Classification, Data Analysis, and Knowledge Organization. — New York: Springer, 1996. — P. 32–44. — https://doi.org/10.1007/978-3-642-79999-0_3

Granger C. J., Newbold P. Spurious regressions in econometrics // Journal of Econometrics. — 1974. — Vol. 2. — P. 111–120. — https://doi.org/10.1002/9780470996249.ch27

Halkidi M., Batistakis I., Vazirgiannis M. On Clustering Validation Techniques // Journal of Intelligent Information Systems. — 2001. — Vol. 17, No. 2/3. — P. 107–145. — http://dx.doi.org/10.1023/A:1012801612483

Kapetanios G. Cluster analysis of panel data sets using non-standard optimisation of information criteria // Journal of Economic Dynamics and Control. — 2006. — Vol. 30, No. 8. — P. 1389–1408. — https://doi.org/10.1016/j.jedc.2005.05.010

Liao T. W. Clustering of time series data — a survey // Pattern Recognition. — 2005. — Vol. 38, No. 11. — P. 1857–1874. — https://doi.org/10.1016/j.patcog.2005.01.025

Montero P., Vilar J. A. TSclust: An R Package for Time Series Clustering // Journal of Statistical Software. — 2014. — Vol. 62, No. 1. — P. 1–43. — https://doi.org/10.18637/jss.v062.i01

Nelson Ch. R., Plosser C. I. Trends and random walks in macroeconomic time series: some evidence and implications // Journal of Monetary Economics. — 1982. — Vol. 10. — P. 139–162. — https://doi.org/10.1016/0304-3932(82)90012-5

Niu J. H. The Cluster Analysis of Multivariable Panel Data and Its Application // Applied Mechanics and Materials. — 2012. — Vols. 220–223. — P. 2668–2671. — https://doi.org/10.4028/www.scientific.net/amm.220-223.2668

Rousseeuw P. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis // Journal of Computational and Applied Mathematics. — 1987. — Vol. 20. — P. 53–65. — https://doi.org/10.1016/0377-0427(87)90125-7