Cluster analysis of regional indicators using DBSCAN algorithm

Yulia Valentinovna Granitsa[11] and *Shokhjakhon Akmaljon ugli* Khujayev[2]

[1]Lobachevsky State University, Institute of Economics and Entrepreneurship, Nizhny Novgorod, Russia
[2]Tashkent State University of Law, Department Intellectual property law Department, Tashkent, Uzbekistan

Abstract. Regional economies are playing an increasingly important role in the development of the national economic complex of the country. The disproportionate development of economies at the mesolevel is associated with a number of risks affecting various markets and industries, which in turn necessitates the development of effective methods for identifying regional clusters and the search for effective methods for assessing the interconnections of regional economic determinants. To conduct the study, the authors accumulated data on 25 indicators reflecting the investment, resource, production and financial performance components of the socio-economic development of Russian regions. Applying machine learning algorithms such as XG Boost, Gradient Boosting, CART, we identified the most significant factor for assessing regional sustainability and established the regional development indicators associated with it by calculating the non-linear correlation coefficient Phi_K. The use of the DBSCAN algorithm allowed us to identify two regional clusters, while per capita consumption, the level of demographic load and urbanization were significant factors for the clustering of regions. The significance of the criteria for combining regions into clusters using the DBSCAN method was established using the construction of a classification tree.
Keywords: region, correlation coefficient Phi_K, clustering, DBSCAN algorithm, CART algorithm, classification tree

1 Introduction

In the current conditions of instability caused by the global pandemic, the deficit of federal budget funds due to the growth of public social spending, it is worth noting the growing role of the region as an independent business entity.

Disproportions in regional development, as well as numerous risks penetrating into different sectors of the economy at the mesolevel, necessitate the development and scientific justification of effective methods for clustering regions and tools for assessing the relationship of macroeconomic indicators that reflect the characteristics of regional dynamic development.

[1] Corresponding author: ygranica@yandex.ru

The author's approach to the analysis of regional indicators, proposed in the article, develops the ideas previously disclosed in the works of domestic and foreign researchers, in particular Statsenko [1], Barinova [2], Klimanov et al. [3], Rahmah [4], Szitasiova et al. [5], Rickman [6], Yeah [7], Billon and Marco [8], Fornaro and Wolf [9], Raven and Walrave [10], Michaels [11], Biau and Devroye [12], Arifovic and Petersen [13], Christopeit and Massmann [14], Cornand and Hubert [15], Beechey et al. [16], Granitsa [17] and others.

Data on 25 indicators reflecting the investment, resource, production and financial performance components of the socio-economic development of 82 constituent entities of the Russian Federation for 2019 was collected by the authors to conduct the study.

2 Methods

Studies of regional stability based on the analysis of these indicators using supervised machine learning methods such as the XG Boost algorithm, the gradient boosting method, the construction of classification trees, allowed us to identify significant factors for assessing the stability of regions (Table 1).

Table 1. Significant Factors for Assessing Regional Sustainability

Name of machine learning algorithm	List of significant factors
Gradient Boosting Method	investment risk (Risk), net financial result (SFR), budget expenditures per capita (Bd) investment risk (Risk), net financial result (SFR), budget expenditures per capita (Bd), employment rate (Uz)
Classification tree	
XG Boost	investment risk (Risk), employment rate (Uz), net financial result (SFR), share of shipped goods, works, manufacturing services in GRP (MI)

It should be noted, that in all the algorithms used, the most significant indicator is the investment risk, which characterizes the investment climate in the region and is defined by experts as the probability of losses and uncertainty in the financial result.

Investment risk is one of the most important indicators taken into account when creating new forms of interaction between the state and business, which helps to attract new sources of financing for the regions and gives regional structures competitive advantages.

Let us evaluate the relationship of investment risk with factors characterizing the resource, production and financial performance of the regional components using the correlation coefficient Phi_K.

The mechanism for using this coefficient is described in the work of M. Baaka et al. [18].

The correlation coefficient Phi_K captures non-linear relationships between two indicators and is equal to the Spearman correlation coefficient in the case when the indicators are distributed according to the normal law and there is a linear relationship between them.

The coefficient calculation algorithm contains a built-in noise reduction technique from statistical fluctuations, that is, the coefficient is not sensitive to outliers.

We calculated the Phi_K correlation coefficients in the Jupyter Notebook environment using the special phik library. The calculation results are shown in Figure 1.

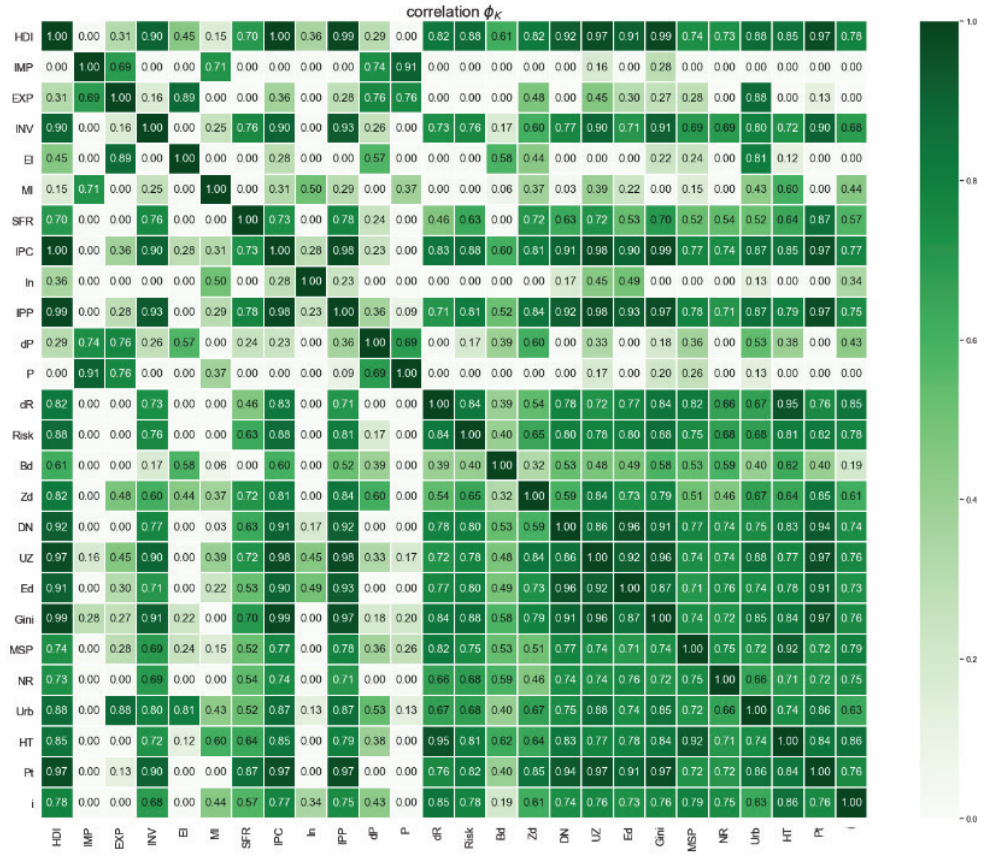

Fig. 1. Evaluation of the relationship of normalized regional economic determinants using the correlation coefficient Phi_K.

The correlation coefficient Phi_K takes a value from 0 to 1, and the closer the coefficient is to unity, the stronger the relationship between the indicators. Significant relationships of economic determinants in Figure 1 are highlighted in saturated green.

To form a conclusion about the dependencies between indicators, we will take into account not only the value of the correlation coefficients, but also their statistical significance, the calculation of which is shown in Figure 2.

A somewhat different picture emerges if we analyze the normalized values of economic indicators. In this case, we observe a close relationship between investment risk and human development index (HDI), fixed capital investment (INV), net financial result (SFR), consumer price index (IPC), industrial production index (IPP), employment rate (UZ), level of education (Ed), Gini coefficient (Gini), consumption level (Pt), dependency ratio (DN). However, all relationships are significant.

At the next stage of our study, we will solve the problem of clustering Russian regions.

Clustering is one of the problems of unsupervised learning when there are only input variables without corresponding labels and it is necessary to apply an algorithm that will find a pattern to solve the problem.

As a tool, we use the DBSCAN algorithm (Density-based spatial clustering of applications with noise) – a density spatial clustering algorithm using noise. The analysis of

this algorithm is contained in the works of Golomidov [19], Alireza Latifi-Pakdehi, Negin Daneshpour [20], Igor de Moura Ventorim et al. [21].

Experimentally, we set the external parameters – the number of neighbors, that is, the data density and the radius within which the objects will belong to one cluster. We perform the clustering procedure in the Jupyter Notebook environment using the DBSCAN library. Clustering was preceded by data normalization

As a result, two regional clusters were obtained (Fig. 3)

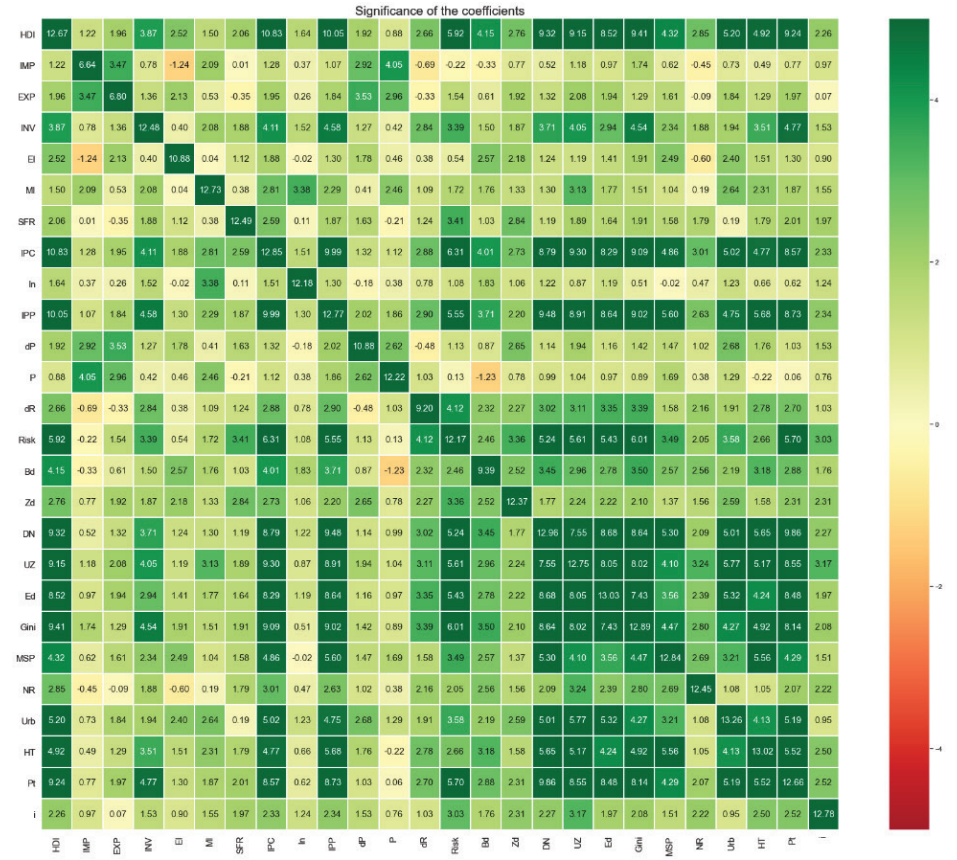

Fig. 2. Significance levels of correlation coefficients.

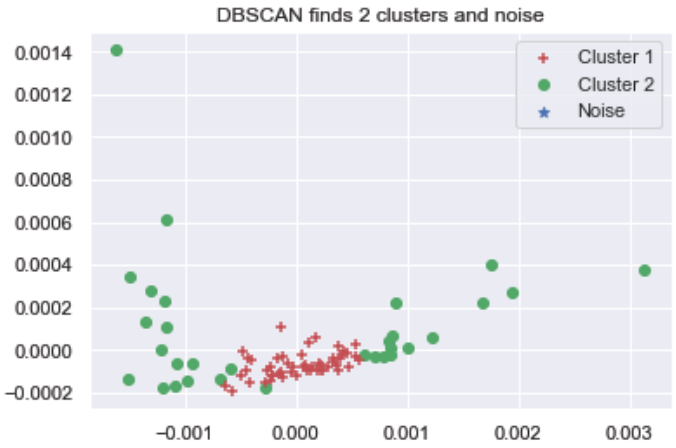

Fig. 3. Results of clustering Russian regions using the DBSCAN method.

Let us evaluate the significance of factors for the selection of regional clusters using the DBSCAN method using the CART algorithm (classification and regression trees).

The classification tree is a decision-making method based on the use of functions for dividing the original data set by setting external constraints in the form of threshold rules.

The difficulty in using this method is to determine the moment of stopping "set splitting".

The following stopping rules are distinguished: the measure of "contamination" is less than a certain value; limit on the number of nodes or layers of the tree; parent node size; child node size

In Figure 4, each numbered tree node defines:

• panel indicators, according to which the regions are classified, indicating their serial number in square brackets;

• the proportion of observations that fell into the node (samples);

• evaluation of each class (value)

• informative criterion (Gini coefficient)

The Gini coefficient is estimated as the probability that when choosing a class according to the distribution in a tree branch, a random sample will be classified incorrectly.

The calculation of the Gini coefficient is carried out according to (1):

$$\text{Gini coefficient} = 1 - \sum p_i^2, \quad (1)$$

where $p_{i\_}$ the probability of the system being in the i-th state.

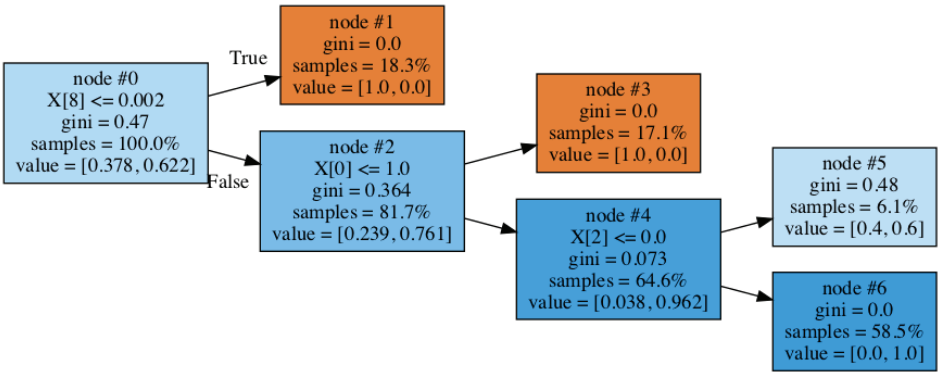

Fig. 4. Evaluation of regional clustering criteria using normalized data.

It should be noted that in the resulting classifier, we were able to identify significant variables for determining regional clusters using the DBSCAN method (Table 2).

Table 2. Significant classification features when constructing a classification tree of Russian regions using normalized indicators

Feature number	Indicator	Value
0	Consumption per capita (Pt)	0.57
8	Dependency ratio (DN)	0.39
2	The share of the urban population in its total number (Urb)	0.04

If the data normalization procedure is not carried out, the classification tree will look like this (Fig. 5).

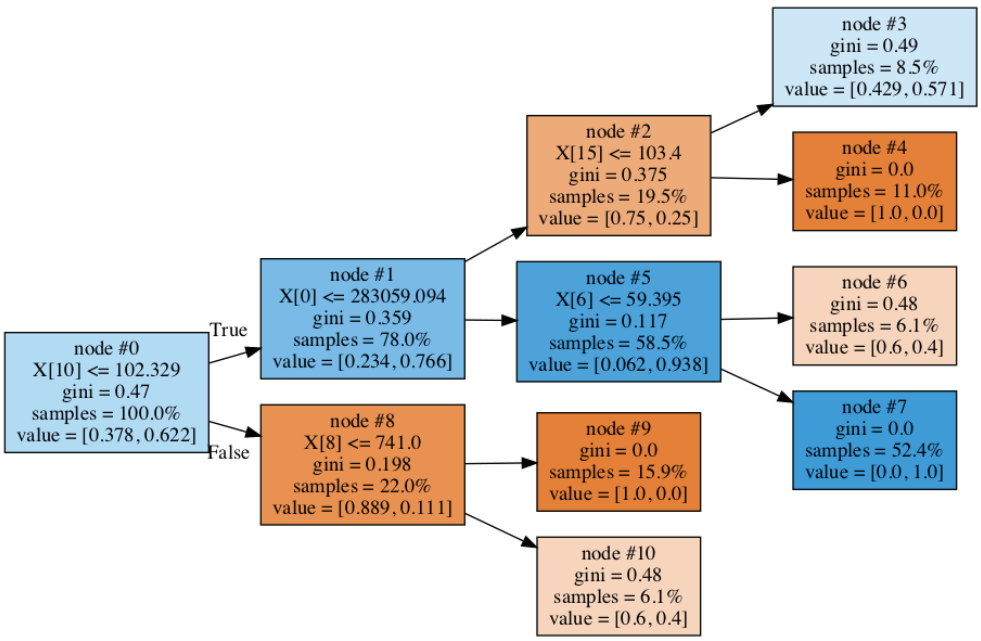

Fig. 5. Classification tree of Russian regions using absolute values of indicators.

Significant variables for the selection of regional clusters by the DBSCAN method in this case are indicated in Table 3.

Table 3. Significant classification features when constructing a classification tree of Russian regions using absolute values of indicators

Feature number	Indicator	Value
10	Budget expenditure per capita (Bd)	0.4
0	Consumption per capita (Pt)	0.37
6	Level of Education (Ed)	0.11
15	Industrial Production Index (IPP)	0.08
8	Demographic load (DN)	0.04

3 Results

Thus, in the course of the study, the following results were obtained:

1. A set of significant panel indicators has been established to assess the sustainability of regional development
2. Relationships between panel indicators characterizing regional stability using the Phi_K correlation coefficient are determined.
3. Clustering of regions by 25 regional economic indicators was performed using the DBSCAN method
4. It is possible to find significant factors for the selection of regional clusters using the CART algorithm.

4 Discussion

The conducted research allowed the authors to come to the following conclusions:

1. Significant non-linear relationships were found between regional investment risk and human development index (HDI), fixed capital investment (INV), net financial result (SFR), consumer price index (IPC), industrial production index (IPP), employment rate (UZ), level of education (Ed), Gini coefficient (Gini), consumption level (Pt), dependency ratio (DN).
2. The most reliable result is given by the procedures for analyzing normalized indicators due to the fact that data normalization ensures their comparability.
3. The most important factors for identifying regional clusters are per capita consumption, demographic load and the share of the urban population in its total population.

5 Conclusion

The authors see the prospect of this study in the development of scenarios for the dynamic development of regional structures in the post-pandemic period, in respect of which, taking

into account the identified clusters and significant determinants, the probabilities of an optimistic and pessimistic trend will be determined.

The study was carried out within the framework of the basic part of the state assignment of the Ministry of Education and Science of the Russian Federation, project 0729-2020-0056 "Modern methods and models for diagnosing, monitoring, preventing and overcoming crisis phenomena in the economy in the context of digitalization as a way to ensure the economic security of the Russian Federation".

References

1. I.M. Statsenko, Econ. Reg., **14(3)**, 927-940 (2018). https://doi.org/10.17059/2018-3-17
2. V.A. Barinova, S.P. Zemtsov, Region: Econ. Soc., **1**, 23-46 (2019). https://doi.org/10.15372/REG20190102
3. V.V. Klimanov, S.M. Kazakova, A.A. Mikhaylova, Econ. Iss., **5**, 46-64 (2019). https://doi.org/10.32609/0042-8736-2019-5-46-64
4. N. Rahmah, I.S. Sitanggang, IOP Conf. Ser.: Earth Environ. Sci., **31**, 012012 (2016)
5. V. Szitasiova, M. Sipikal, M. Siserova, Innovation support, resilience and regional development in Slovakia., in T. Baycan, H. Pinto (eds.), Resilience, crisis and innovation dynamics. New horizons in regional science, 90-111 (Cheltenham: Edward Elgar, 2018)
6. D.S. Rickman, Rev. Reg. Stud., **44(1)**, 1-12 (2014)
7. K.L. Yeah, Adv. Econ. Bus., **5(3)**, 109-128 (2017). https://doi.org/10.13189/aeb.2017.050301
8. M. Billon, R. Marco, F. Lera-Lopez, Empir. Econ., **53(3)**, 1083-1108 (2017). https://doi.org/10.1007/s00181-016-1153-x
9. L. Fornaro, M. Wolf, COVID-19 Coronavirus and Macroeconomic Policy. Barcelona Graduate School of Economics. Working Paper No. 1168 (2020)
10. R. Raven, B. Walrave, Techn. Forecast. Soc. Change, **153**, 119297 (2020). https://doi.org/10.1016/j.techfore.2018.05.008
11. G. Michaels, Econ. J., **121(551)**, 31-57 (2011). https://doi.org/10.1111/j.1468-0297.2010.02402.x
12. G. Biau, L. Devroye, J. Multivar. Analysis, **101**, 2499-2518 (2010)
13. J. Arifovic, L. Petersen, J. Econ. Dynam. Control, **82(C)**, 21-43 (2017). https://doi.org/10.1016/j.jedc.2017.04.005
14. N. Christopeit, M. Massmann, Econ. Theory, **34**, 68-111 (2018). https://doi.org/10.1017/S0266466616000529
15. C. Cornand, P. Hubert, J. Econ. Dynam. Control, **110**, 103746 (2020). https://doi.org/10.1016/j.jedc.2019.103746
16. M. Beechey, B. Johannsen, A.T. Levin, Amer. Econ. J.: Macroecon., **3(2)**, 104-129 (2011). https://doi.org/10.1257/mac.3.2.104
17. Yu.V. Granitsa, *Analysis of the relationship of economic indicators as a tool for predicting regional financial instability*, in Proceedings of the International Scientific Conference (ISCFEC 2020), 2772-2779 (2020). https://doi.org/10.2991/aebmr.k.200312.394

18. M. Baak, R. Koopmana, H. Snoek, S. Klous, A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics, arXiv:1811.11440v2 [stat.ME] (2019)
19. Yu.K. Golomidova, V.S. Kireev, Economics, **9(30)** (2017)
20. A. Latifi-Pakdehi, N. Daneshpour, Data Knowl. Eng., **135** (2021). https://doi.org/10.1016/j.datak.2021.101922.