Method of Data Mining on the State of Industries in Russian Regions with Regard to Social Factors

Bary Ilyasov, Elena Makarova, Elena Zakieva, Elvira Gabdullina and Margarita Mansurova

University 1, Address, City, Index, Country Ufa State Aviation Technical University, Karl Marks str., 12, Ufa, 450077, Russia

Abstract

A method of data mining on the state of industrial sectors in the regions of Russia, taking into account social factors, was developed. The method is based on the development of a hierarchical structure of initial features and application of the principal component analysis according to the constructed structure. The necessity of decomposing a set of the initial features is justified, due to the fact that the results of the preliminary principal component of a large number of features turned out to be cumbersome and difficult to interpret. A two-level decomposition tree of the set of the initial features was developed. The analysis is carried out in accordance with the developed tree structure of the initial features. Three samples of the lower level and one sample of the upper level were selected. The use of the principal component analysis of the lower-level samples allowed us to assess the state of regions from various viewpoints: the development of the extractive industry, manufacturing industry and electric power industry, taking into account social factors. Fuzzy rules for clustering the regions in the space of the features of the lower-level samples are formed. Based on the constructed principal components for the samples of the lower level, an integral sample of the upper level is formed, which allows assessing the level of development of the regions as a whole. The proposed method of data mining should be used to compile a database of the fuzzy rules of an intelligent decision support system for managing the socio-economic development of the regions.

Keywords

Data mining, principal component analysis, region, clustering, industry, social factors, decomposition tree, fuzzy rule

1. Introduction

In modern Russia, the study of industrial production as a factor in increasing competitiveness is due to the urgent need to ensure economic growth in all sectors of the economy. The government is developing concepts and strategies aimed at ensuring sustainable development of the state. The Strategy for the Spatial Development of the Russian Federation for the period up to 2025 emphasizes the importance both of the formation and strengthening of mineral resource centers, and development of industries in the branches of promising economic specializations [1]. The comparative analysis of the levels of development of the following industries is relevant: "Extraction of minerals" (EI); "Manufacturing industries" (MI); "Providing regions with electricity, gas and steam and air conditioning" (EP). To ensure the long-term sustainable development of the Russian economy, it is necessary to strive for the balanced development of these industries. In addition, it is worth considering the social aspect of the economic development. For a qualitative assessment of the structure of industrial production in the country, taking into account social factors, it is advisable to conduct a study in the regional context.

SibDATA 2021: The 2nd Siberian Scientific Workshop on Data Analysis Technologies with Applications 2021, June 25, 2021, Krasnoyarsk, Russia

EMAIL: ea-makarova@mail.ru (E. Makarova)
ORCID: 0000-0001-5642-0723 (B. Ilyasov); 0000-0003-0043-3693 (E. Makarova); 0000-0002-6921-7473 (E. Zakieva); 0000-0002-2421-1510 (E. Gabdullina); 0000-0002-8912-2754 (M. Mansurova)

A number of works are known to be related to the construction of regional clusters and analysis of regional differences, for example, both in terms of the level of development of the real sector [2], and in general, in terms of socio-economic development [3]. As part of the research, an approach is being developed to support decision-making in the management of regional systems, involving the use of tools for data mining on the state of sectors, industries, regions of the Russian Federation, as well as fuzzy systems and simulation [4]. The purpose of this study is to develop a method of data mining on the state of industry in Russian regions, taking into account social factors, which will make it possible to form clusters of the regions and clustering rules, taking into account the complementary or mutually exclusive development of enterprises of EI, MI and EP.

2. Research method

The information base of the study is the data on the state of industrial development of the regions of the Russian Federation for 2019 [5]. The peculiarity of the proposed method consists in the decomposition of the set of the initial features and subsequent formation of new integral features. The need to decompose the initial features is justified by the fact that when performing the component analysis of a large number of features without dividing them into subsamples of data, the results obtained are cumbersome and difficult to interpret.

According to the proposed method for generating samples of the initial data and conducting data mining, the set of the initial features is decomposed into three directions which determine the samples of the initial features of the lower level (see **Ошибка! Источник ссылки не найден.**). The choice of a set of the initial features is determined both by the purpose of the study being carried out and by the composition of the data available for the analysis [5].

The decomposition tree is built, firstly, based on the purpose of the study and, secondly, on the results of the preliminary studies. According to the purpose of the study, the first two directions are highlighted, associated with the construction of clusters of the regions which differ in the level of economic development, as well as the third direction which takes into account regional differences in the level of social development. Based on the results of the preliminary studies by the method of principal components, the possibility of integrating the features characterizing a particular industry sector into a separate component is determined. The features characterizing the branches of MI and EP analyzed in the same EI sample (the first direction is highlighted) should be combined into one component; therefore, it was decided to conduct an additional analysis of the extended set of the features characterizing only MI and EP, in order to identify clusters of the regions with the predominant development of enterprises of either MI or EP (the second direction is highlighted).

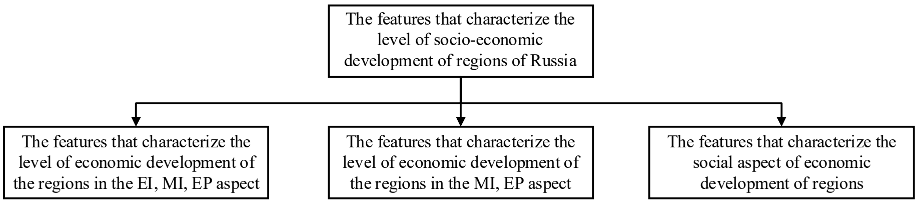
The features that characterize the level of socio-economic development of regions of Russia
The features that characterize the level of economic development of the regions in the EI, MI, EP aspect
The features that characterize the level of economic development of the regions in the MI, EP aspect
The features that characterize the social aspect of economic development of regions

Figure 1: The decomposition of the set of the initial features

The results of the data mining of the three samples of the lower level of decomposition are used to form an integral sample, which characterizes the level of development of the industry of the regions as a whole, taking into account social factors. The formed integral sample is also subjected to the data mining procedure. The results of the data mining analysis of the lower-level samples and integral sample at the upper level are the clusters of the regions and their characteristics, which are refined when moving up the tree of the initial features (see **Ошибка! Источник ссылки не найден.**).

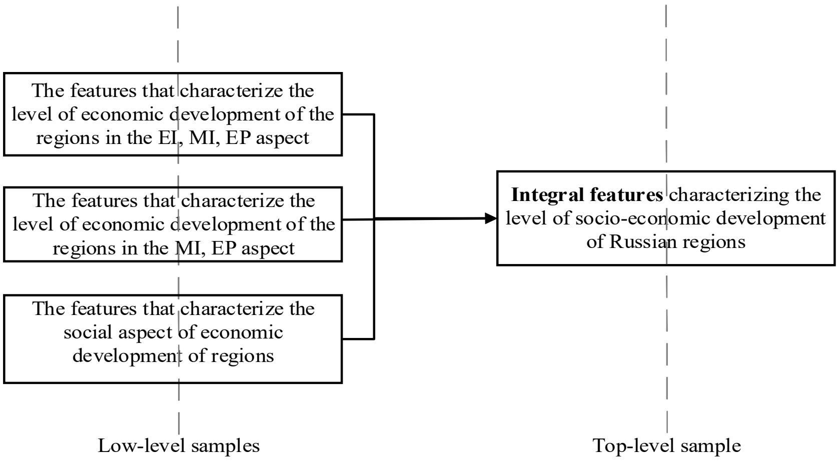
The features that characterize the level of economic development of the regions in the EI, MI, EP aspect
The features that characterize the level of economic development of the regions in the MI, EP aspect
The features that characterize the social aspect of economic development of regions
Integral features characterizing the level of socio-economic development of Russian regions
Low-level samples
Top-level sample

Figure 2: The composition of the initial features

The proposed method for conducting intelligent data analysis allows one to obtain information on the level of development of both individual industries in the regions, as well as to assess the state of the regions as a whole due to the introduction of the integral sample (top-level sample).

3. Results of data mining on the state of industries using low-level samples

For the first sample of the lower level of decomposition (**Ошибка! Источник ссылки не найден.**), using the method of principal components, the level of development of enterprises in the branches of EI, MI and EP was analyzed in the regions. The results of the analysis showed that in the first two principal components (PC), more than 91% of the variance was achieved. The weight coefficients of the features involved in the names of the principal components are presented in **Ошибка! Источник ссылки не найден.** and highlighted in bold type. The application of varimax rotation confirmed the composition of the identified principal components.

Table 1
The weight coefficients of the features (analysis of the level of development of the regions in terms of EI, MI and EP)

	PC 1	PC 2
Volume of goods shipped (MI)	**0,387183**	-0,125757
Number of enterprises and organizations (MI)	**0,374812**	-0,247001
Turnover of organizations (MI)	**0,376474**	-0,190449
Volume of goods shipped (EP)	**0,390321**	0,0776813
Number of enterprises and organizations (EP)	**0,364406**	-0,0901584
Turnover of organizations (EP)	**0,375591**	-0,12932
Volume of goods shipped (EI)	0,173318	**0,858454**
Number of enterprises and organizations (EI)	0,3341	**0,345072**

The weight coefficients of the features obtained as a result of the analysis show that: PC 1 allows identifying the regions with the high and low level of economic development of MI and EP, taking into account the number of enterprises and organizations of EI; PC 2 allows dividing the regions according to the level of economic development of EI. When conducting the component analysis, it is revealed that the most developed region in terms of the level of economic development of MI and EP is Moscow, and the most developed one in terms of economic development of EI is the Tyumen region. The leading regions stand out significantly in the scatter diagram and do not allow the visual assessment of the location of the remaining regions; therefore, the leading regions were removed from the initial data sample. The component analysis was repeated, and the composition of the main components did not change. The scatter diagram obtained during the analysis without taking into account the leading regions is shown in **Ошибка! Источник ссылки не найден.**.

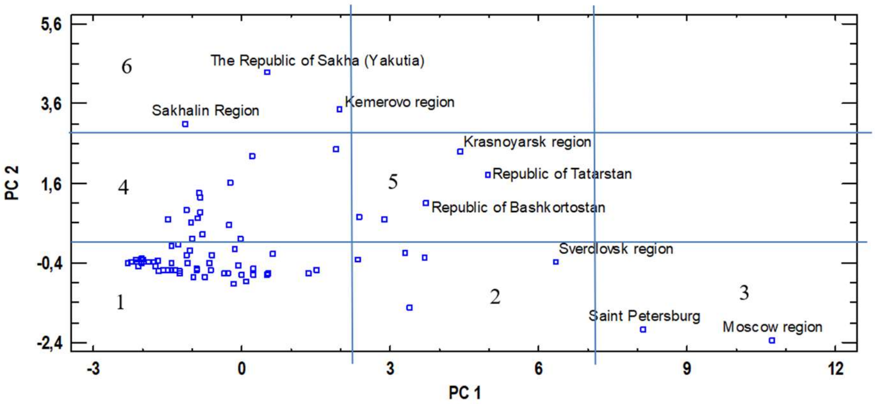

Figure 3: The scatter diagram obtained in the analysis of the first sample of the lower level concerning the state of EI, MI, EP (sample without the leading regions)

Six clusters are highlighted in the scatter diagram. To clarify the boundaries of the clusters, the cluster analysis was made (Ward's method, Squared Euclidean metric), which, in general, confirmed the composition of the clusters identified during the component analysis of the data.

Most regions of Russia are characterized by the low level of development of EI, MI and EP. The city of St. Petersburg, and the Moscow region are distinguished by the high level of development of MI and EP in the region. Such regions as the Republic of Sakha (Yakutia) and Sakhalin Oblast are characterized by the high level of development of EI. A number of regions, for example, the Republic of Tatarstan, and Krasnoyarsk Region, are in the area of the balanced development of industries.

Fuzzy rules for clustering the regions in the space of features for all clusters are formed. As an example, a rule for cluster 1 is given: IF the volume of goods shipped MI = small AND the number of enterprises and organizations MI = small AND the turnover of organizations MI = small AND the volume of goods shipped EP = small AND the number of enterprises and organizations EP = small AND the turnover of organizations EP = small AND the volume of goods shipped EI = small AND the number of enterprises and organizations EI = small, THEN Cluster = 1.

The analysis of the first sample of the lower level shows that the level of development of enterprises of MI and EP is separated into a separate principal component, while it is impossible to divide the regions with the predominant development of enterprises of either MI or EP. Therefore, the second sample of the lower level of decomposition was formed (**Ошибка! Источник ссылки не найден.**). The analysis of the level of development of the regions in terms of MI and EP was made in order to single out individual PCs for MI and EP. The analysis by the method of principal components shows that in the first three main components, more than 94% of the variance is achieved. The resulting weights of the signs (see **Ошибка! Источник ссылки не найден.**) allow us to conclude that: PC 1 characterizes the level of development of MI and EP in the region; PC 2 allows identifying the regions with the priority development of either MI or EP; PC 3 makes it possible to identify the regions in which both MI and EP are developed.

Table 2

The weight coefficients of the features (analysis of the level of development of the regions in terms of MI and EP)

	PC 1	PC 2	PC 3
Volume of goods shipped (MI)	**0,418823**	0,0592664	0,0287141
Volume of goods shipped (EP)	**0,407367**	-0,11086	0,0284133
Share of MI in gross value added	0,0771204	**0,790076**	**0,579898**
Share of EP in gross value added	-0,0657783	**-0,574228**	**0,811001**
Number of enterprises and organizations (MI)	**0,410502**	-0,00192621	-0,0139553
Number of enterprises and organizations (EP)	**0,3902**	0,0242087	0,014693

Turnover of organizations (MI)	**0,407853**	-0,0516212	0,0214603
Turnover of organizations (EP)	**0,401563**	-0,164273	-0,0591032

The analysis of the obtained scattering diagram (**Ошибка! Источник ссылки не найден.**) makes it possible to form the following regularities. The city of Moscow is distinguished by the greatest development of MI and EP. In the Lipetsk and Kaluga regions, the strongest predominance of MI over EP is observed; in the Chukotka Autonomous Okrug, the strongest predominance of EP over MI is observed. The regions are generally evenly distributed along the second main component. Characterized by the balanced development of MI and EP is an impressive number of regions, including the Tyumen region, St. Petersburg, the Moscow region, the Republic of Tatarstan and other.

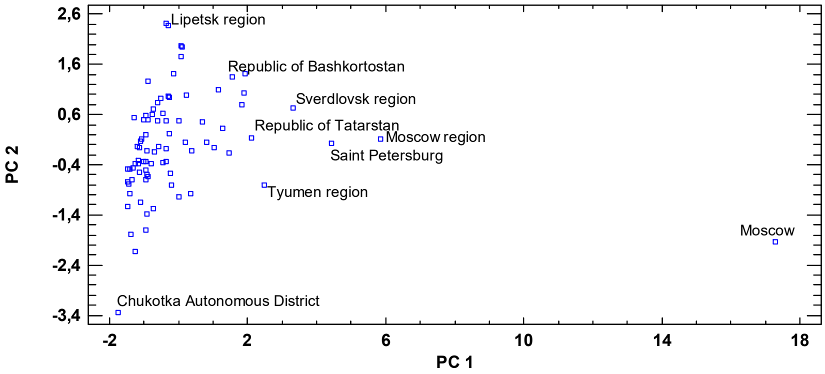

Figure 4: The scatter diagram (analysis of the second sample of the lower level on the state of MI, EP)

For the third sample of the lower level of decomposition (see **Ошибка! Источник ссылки не найден.**), the analysis of social factors was made. Based on the weights obtained for the signs, the following conclusion was drawn: the first PC characterizes the material well-being of the population, taking into account the level of employment in the region; the second PC characterizes the level of unemployment and the coefficient of tension in the labor market (with negative signs). These results are presented in the scatter diagram (**Ошибка! Источник ссылки не найден.**).

It is revealed that the greatest material well-being is observed in Moscow. The Republic of Ingushetia is distinguished by the highest level of unemployment, coefficient of tension in the labor market and the lowest material well-being of the population. Most regions in Russia have the medium or low material well-being and low tension on the labor market.

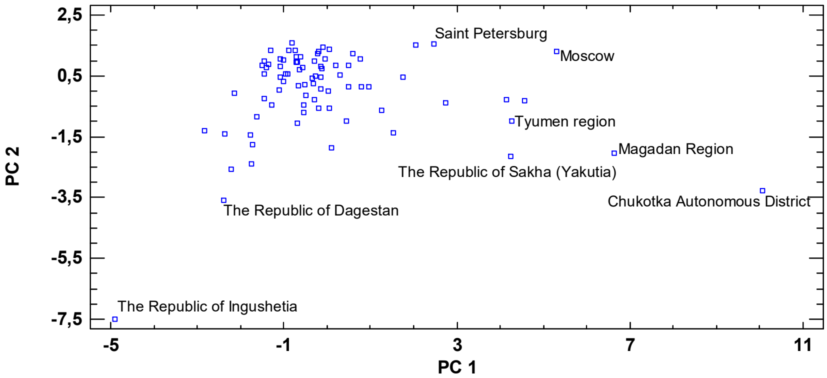

Figure 5: The scatterplot (analysis of social factors for the third sample)

4. Results of data mining on the state of industries using the top-level sample

The component analysis of the integral sample of the upper level is performed. The obtained weight coefficients of the features (see **Ошибка! Источник ссылки не найден.**) indicate that: PC 1 characterizes the structural features of the development of MI and EP in the regions; PC 2 characterizes the social aspect of economic development in the region; PC 3 characterizes the level of development of EI in the region.

Table 3
The feature weights (top-level sample analysis)

	PC 1	PC 2	PC 3
The level of economic development of MI and EP, taking into account the number of enterprises and organizations of EI	0,472551	0,0533564	-0,185048
The level of economic development of EI	0,164738	-0,375136	-0,872834
The level of economic development of MI and EP	0,468272	0,17044	0,058021
The regions with the priority development of either MI, or EP	0,460726	0,198851	0,116246
The regions in which both MI and EP are developed	0,468362	0,169958	0,0568524
Material well-being of the population, taking into account the level of employment in the region	0,301347	-0,50257	0,331057
The unemployment rate and labor market tension coefficient	-0,0885287	0,711598	-0,27241

The scatter diagram (see **Ошибка! Источник ссылки не найден.**) allows one to form the following regularities. During the component analysis of the integral sample (sample of the top-level) along PC 1, the regions are distributed according to the structural features of the development of MI and EP in the region. The most developed region is Moscow, and the Tyumen region, Moscow region, St. Petersburg are characterized by the high level of development. As concerns PC 2, the regions are distributed according to the level of social well-being in the region, and the highest values of this indicator are characteristic of such regions as Moscow, the Moscow region, the Sverdlovsk region, the Republic of Tatarstan and others, including the Republic of Bashkortostan, while the least high is the Chukotka Autonomous district.

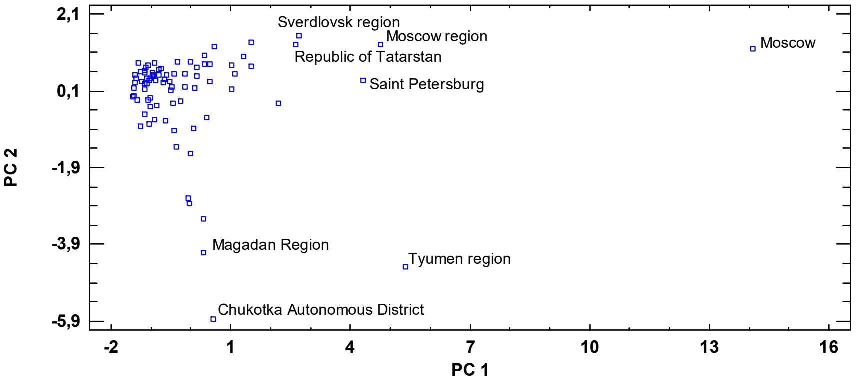

Figure 6: The scatter plot obtained from the analysis of the top-level sample (space PC 1 - PC 2)

5. Conclusion

A method for data mining is proposed, according to which a set of the initial features is decomposed to analyze the level of industrial development in Russian regions from various viewpoints: the development of the extracting industry, manufacturing industry and electric power industry, taking into account social factors. The composition of the formed principal components is used to form a sample of the top-level and its analysis, which makes it possible to assess the level of industrial development in the regions as a whole. It is advisable to use the proposed method of data mining to compile a base of fuzzy rules for an intelligent decision support system in managing the socio-economic development of the regions.

6. Acknowledgements

This work was supported by the RFBR (The Russian Foundation for Basic Research), grant No 20-08-00796 "Intellectual control of an industrial complex as a dynamic multi-agent object based on cognitive modeling and machine learning methods".

7. References

[1] On the approval of the Strategy for the Spatial Development of the Russian Federation for the period up to 2025: order of the Government of the Russian Federation of 13.02.2019, №. 207-p.

[2] E. E. Kolchinskaya, A. L. Kalishenko, I. M. Lementa, Study of the dynamics of development of the real sector of the regions of Russia. Regional economy: theory and practice 41(368), (2014) 47-60.

[3] R. M. Nizhegorodtsev, E. I. Piskun, V. V. Kudrevich, Forecasting indicators of socio-economic development of the region, Economy of the region 13(1) (2017) 38-48.

[4] B.G. Ilyasov, E. A. Makarova, E. Sh. Zakieva, E. S. Gizdatullina, Assessment of data on the income of the population in the regional context by the method of principal components, Economy of the region 15(2) (2019) 601-617.

[5] Materials of the Federal State Statistics Service. URL: https://rosstat.gov.ru/folder/210/document/13204.