CEUR-WS.org/Vol-3047/paper7.pdf

Data Mining for Estimating Living Standards in the Constituent Entities of the Russian Federation

Bary Ilyasov, Elena Makarova, Elena Zakieva, Elvira Gabdullina and Timur Teregulov

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

Abstract

The issues of increasing the efficiency of socio-economic management processes aimed at improving the living standards of the population in the Russian Federation are of great importance. To make rational, well-grounded management decisions, accurate information about the current situation in the constituent entities of the Russian Federation is required. Analytical information is formed by performing various studies to determine the index of living standard to assess regional differences and analyze the well-being of the population, taking into account the degree of its differentiation. Methods of data mining, in particular, the method of principal components and method of cluster analysis, are used as a toolkit for determining the index of the population living standards. The information material for the study is the indicators of the standard of living of the population published in the official bulletins of the Federal State Statistics Service of the Russian Federation. The analysis procedure is performed in the Statgraphics software package. The article presents a cluster structure of the standard of living of the population in the regions of the Russian Federation, including six clusters. The characteristics of each cluster are formed in terms of GRP, monetary income of the population, unemployment, housing provision, mortality, migration flows, etc. The results of the study make it possible to increase the validity of decision-making in the management of socio-economic systems from the standpoint of sustainable development of the standard of living of the population.

Keywords

Data mining, principal component analysis, cluster analysis, living standards of the population, sustainable development goals

1. Introduction

Various social reforms in the Russian Federation, resulting in the strengthening of independent national cooperation, at the same time cause a significant interregional differentiation in the living standards of the population of the Russian Federation constituent entities. The differentiation of constituent entities [1] is accompanied by a wide range of regional contrasts and manifests itself both in the form of objective differences (geographical location, climatic conditions, natural resources), and in the form of demographic, agglomeration, sectoral, social and other differences.

The most important task facing the bodies of the state, regional, and municipal administration is sustainable socio-economic development, manifested in improving the living standards of the population, which should be accompanied by the implementation of actions and processes to smooth out interregional differences in their living conditions. In order to make sustainable decisions at various levels of management, it is necessary to conduct scientific research on the analysis of regional differences and assessment of the well-being of the population.

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); 0000-0001-9593-2860 (T. Teregulov)

© 2021 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

2. Material and methods of research

The standard of living is an important indicator which reflects the measure of meeting the needs of the population both in terms of its material and financial capabilities, as well as social, economic and environmental spheres of life [2]. In international practice, there are many methods for assessing the living standards and further sustainable development of the population, which are mainly divided into two groups associated with the construction of either systems of indicators or integral indices of the living standards.

Thus, the first group of methods includes the System of Global Indicators for Achieving the Sustainable Development Goals (SDGs), developed by the UN (United Nations) [3]. This system is aimed at achieving an equilibrium combination of the well-being of the population, and protection of the natural environment, as well as compliance with social equality; currently it contains 17 global SDGs, specified in 231 unique indicators. The World Bank's Living Standards Measurement Study (LSMS) program, related to a multidisciplinary household survey, belongs to the same group [4]. The goal of the program is to form a Unified National Information System based on an approach to data management which integrates the production, protection, exchange and use of data in planning and decision-making processes of various state bodies. This World Bank's program is aimed at achieving the global SDGs set by the UN.

The disadvantage of these methods is the complexity of analyzing both the living standards of the population as a whole and the turning points of socio-economic development caused by the dynamic characteristics of each particular indicator. The second group of methods is aimed at constructing integral indices of the living standards, such as, for example, the human development index, quality of life index, happiness index, ecological footprint, real progress index, index of sustainable economic well-being, etc., which can be studied in dynamics. It should be noted that various international agencies use significant indicators within the specified categories (social, labor, demographic, natural and climatic, etc.) to perform an integral assessment of the living standards, but the main difficulty in constructing the living standard index as a weighted indicator is rather in the uncertainty when forming the coordinate information space of various statistical indicators and setting the weight coefficients of the initial indicators.

These problems can be solved by using data mining methods, the principal component method [5], aimed at reducing the multidimensional space of the initial features, the formation of complex features in the form of principal components representing linear combinations of indicators, and the identification of hidden patterns in a multidimensional data set, as well as using the cluster analysis method. The objectives of the research are the classification of the constituent entities of the Russian Federation by the living standards in accordance with the socio-economic indicators and the detection of problematic aspects of life of the population in the regions.

In this paper, we study the living standards of the population of the Russian Federation constituent entities using the data mining methods based on the indicators published in the official bulletins of the Federal State Statistics Service of the Russian Federation [6] and describing the most important key indicators of the population living standards. 7 of the 17 global Sustainable Development Goals were used to select the initial indicators: Goal 1 (SDG 1) – the universal elimination of poverty in all its forms; Goal 2 (SDG 2) – the elimination of hunger, ensuring food security and improving nutrition and promoting sustainable agricultural development; Goal 3 (SDG 3) – ensuring a healthy lifestyle and promoting well-being for all people at all ages; Goal 4 (SDG 4) – ensuring inclusive and equitable quality education and promoting lifelong learning opportunities for everyone; Goal 5 (SDG 8) – promoting progressive, inclusive and sustainable economic growth, full and productive employment and decent work for everyone; Goal 6 (SDG 10) – reducing inequality within and between countries; Goal 7 (SDG 11) – ensuring the openness, security, resilience and environmental sustainability of cities and human settlements.

At the first stage of the research, the data is analyzed by the method of main components, which allows one to identify the preliminary structure of data on the living standards of the population of the Russian Federation constituent entities. This method allows forming several significant main components, each of them consisting of a finite set of initial indicators ranked by the degree of influence on the component, and thus representing a new integral feature. Within new integral features, classes of objects can be distinguished which demonstrate the characteristic indicators of the living standards.

At the second stage of the research, the problem of clustering the regions is solved, and within this framework the features characteristic of each group of the constituent entities of the Russian Federation are finally determined. The cluster analysis methods make it possible to make automatic classification of observations based on the geometric proximity in a multidimensional space and build rules for assigning to a particular class by the coordinates of cluster centroids.

3. Results of the research

The initial data [6, 7] are 16 of the socio-economic indicators of the constituent entities of the Russian Federation: x_1 is the gross regional product (GRP) per capita (RUB) (SDGs 8); x_2 is the growth rate of GRP (%) (SDG 8); x_3 is the life expectancy (years) (SDG 3); x_4 is the level of education (%) (SDG 4); x_5 is the level of poverty (population with the income below the subsistence minimum, %) (SDG 1); x_6 is the unemployment rate (%) (SDG 8); x_7 is the overall mortality rate (SDG 3); x_8 is the number of crimes per thousand people (SDG 11); x_9 is the real income of the population (RUB) (SDG 10); x_{10} is the consumer expenses, total (RUB) (SDG 10); x_{11} is the real disposable income in % to the previous year (SDG 10); x_{12} is the Gini index (SDG 10); x_{13} is the share of consumer expenses on food, % (SDG 2); x_{14} is the change of savings per capita (index) (SDG 10); x_{15} is the provision of housing, sq. m on average per person (SDG 11); x_{16} is the net migration rate per 10,000 of the population (SDG 10).

Data mining by the principal component method is carried out in the Statgraphics software package. According to the "scree test" by R. Kettel, 6 significant principal components are selected, which describe, in general, about 85% of the variability of the initial data. Each of the principal components is a complex feature, which represents a linear combination of the initial features. Thus, the first principal component (PC_1) is a linear equation:

$$PC_1 = 0{,}35 \cdot x_1 - 0{,}296 \cdot x_5 - 0{,}2 \cdot x_6 + 0{,}43 \cdot x_9 + 0.39 \cdot x_{10} + 0{,}28 \cdot x_{11} + \\ +0{,}377 \cdot x_{12} \quad (1)$$

The greatest contribution to PC_1 is made by such socio-economic indicators as GRP, real income and consumer expenses of the population, as well as the Gini index, which indicates that in the regions with an increase in income and expenses, there is an increase in the differentiation of society. At the same time, negative coefficients at coefficients x_5 and x_6 indicate that in economically developed constituent entities, with an increase in the financial capabilities of citizens, there is a decrease in negative social indicators, such as poverty and unemployment.

According to the method of component analysis, the principal components are ranked by the degree of decreasing variance; so PC_1 is the most significant, implying that it explains most variability in the values of the studied data. However, despite the fact that the following principal components describe a smaller percentage of the initial data variance, they allow one to find the most important hidden patterns that are present in the data. For example, according to the principal component PC_4, a pattern was found that in the constituent entities with a high net migration rate there is a significant decrease in savings among the population, as well as poor housing provision; according to PC_5, in the constituent entities of the Russian Federation with low socio-economic indicators, the share of food costs is significant, and a fairly high mortality rate is observed; according to PC_6, in the constituent entities with an increase in the GRP growth rate, there is an increase in the level of education.

Visual analysis of the data on the scatterplot in the space of the principal components allows identifying a characteristic of the constituent entities which are similar in terms of the specified characteristics. Figure 1 shows a two-dimensional scatterplot in the space of PC_1 and PC_2, which demonstrates the irregular location of the constituent entities of the Russian Federation. The analysis of the scatterplot allows one to conclude that PC_1, which characterizes the socio-economic state of the constituent entities, has positive values only in 34 constituent entities, indicating that only 39% of the constituent entities have a prosperous socio-economic situation; even among the constituent entities with a favorable socio-economic situation, there is a significant differentiation. It is shown that 11 constituent entities exceed the average indicators of prosperous constituent entities; these are the following: Moscow and St. Petersburg; Moscow, Tyumen, Sakhalin Regions; Nenets, Khanty-Mansi, Yamalo-Nenets Autonomous Okrug; Kamchatka Krai; and in accordance with the center of mass, the "average lagging" constituent entity of the Russian Federation for this PC_1 differs from the "average developed" by 3.5 times.

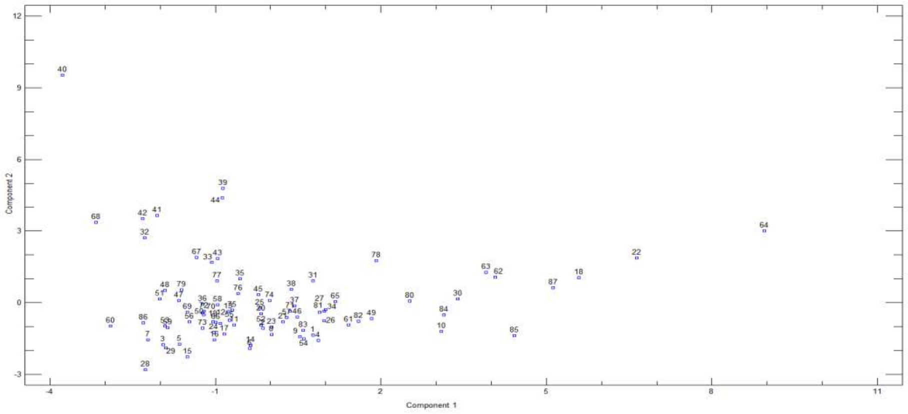

Figure 1: The scatterplot in the space of PC_1 and PC_2

It should be noted that the main component method, being a method of data visualization, requires the interpretation of the analysis results based on the subjective judgments of a decision-maker. In order to confirm these results, the cluster analysis method is often used as an automatic classification method, in which the number of clusters is set in advance. It is the rational choice of the number of clusters that is the main factor for grouping objects with the description of the complex specific of some of their properties and features. The cluster data analysis on the living standards was carried out by the Ward method aimed at minimizing the variance within the clusters, using the City-block metric. A dendrogram is constructed, which is a graphical result of the cluster data analysis, in which the numbers of the combined objects – regions of the Russian Federation – are indicated horizontally, and the distances at which the combinations occurred in the form of a sequential enlargement of the clusters are indicated vertically. Figure 2 shows the division of the data on the living standards in the constituent entities of the Russian Federation into six clusters. Based on the constructed data structure, taking into account the numerical values of the coordinates of the cluster centroids and patterns identified on the basis of the component analysis, the characteristics of the obtained clusters are determined.

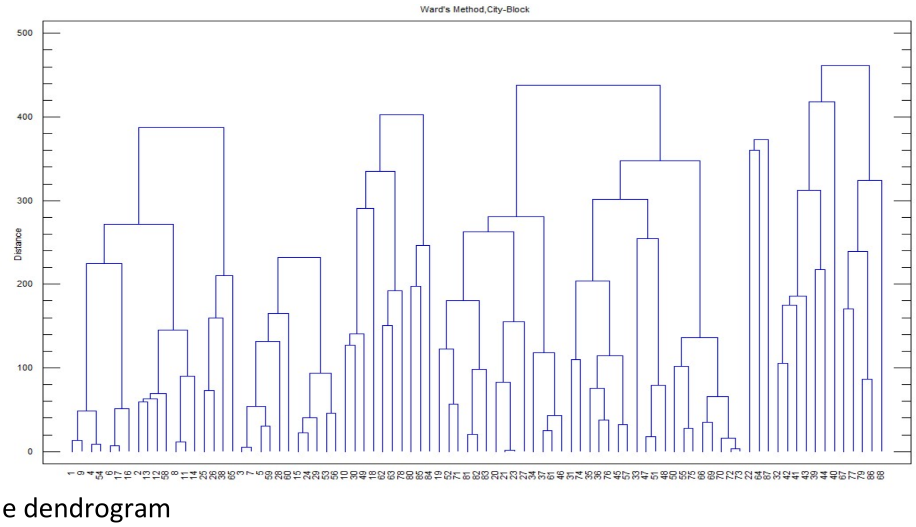

Figure 2: The dendrogram

Cluster 1 (10 constituent entities, 11.49% – Moscow, Sakhalin, Tyumen Regions; Moscow, St. Petersburg, Khanty-Mansiysk Autonomous Okrug, Republic of Tatarstan, Kamchatka Krai) includes the leading regions with high living standards. This cluster is characterized by the highest level of education of the population in the country, a smaller number of people with monetary income below the subsistence minimum, the lowest unemployment rate (3.6%), and a positive migration coefficient. The constituent entities of this cluster have high values of the average real monetary income of the population (46 thousand rubles), but at the same time, the cost of living in these entities is also significant; on average people spend 36.5 thousand roubles a month. This cluster is characterized by a high level of differentiation of the population, as well as positive dynamics of changes in the financial capabilities of the population and growth rate of GRP.

Cluster 2 (4 constituent entities, 4.59% – Nenets, Yamalo-Nenets, Chukotka Okrugs, Magadan Region) includes a small number of entities which have the highest GRP per capita and GRP growth rates, the highest level of education and the lowest level of poverty and unemployment, as well as high average real income of the population (73.9 thousand roubles). The peculiarity of the cluster is that despite the high monetary income, the living standards of the constituent entities cannot be considered as high, since the consumer expenses are quite high (31.8 thousand roubles), 38% of which is spent on food. In addition, this cluster has a high degree of differentiation of society, the lowest life expectancy and housing provision among all the regions, as well as a significant outflow of population. Thus, despite high income, it is impossible to determine the living standards of these entities as above average.

Cluster 3 includes 19 constituent entities (21.83%) (Belgorod, Bryansk, Voronezh, Kaluga, Kursk, Lipetsk, Oryol, Ryazan, Smolensk, Tambov, Tula, Yaroslavl, Kaliningrad, Leningrad, Saratov, Tyumen (without autonomous okrugs) Regions; Republic of Bashkortostan, Sevastopol), having the average living standards: average values of the indicators on the financial capabilities of the population, low unemployment and low crime. The regions included into this cluster have fairly high indicators of GRP per capita and its growth rate. A special feature of the cluster is the presence of positive migration flows which makes these areas attractive for people who would like to change their residence.

Cluster 4 (32 constituent entities, 36.78% – Amur, Arkhangelsk, Murmansk, Astrakhan, Volgograd, Rostov, Orenburg, Samara, Sverdlovsk, Chelyabinsk, Irkutsk, Kemerovo, Novosibirsk, Omsk, Tomsk Regions; the Republics of Adygea, Karelia, Komi, Crimea, Mari El, Mordovia, the Chuvash Republic, the Udmurt Republic, Khakassia; Perm, Altai, Krasnodar, Krasnoyarsk, Stavropol, Khabarovsk, Primorsky Krai) is characterized by the living standards below the national average: the real monetary income of the population does not exceed the national average, and it amounts to 27.1 thousand roubles, while its average value is 29.6 thousand roubles; there is the highest value of the crime rate and a large number of people with the monetary income below the subsistence minimum (14.41%); there is a large value of unemployment (6%), as well as a significant decline in the population along with the lack of positive dynamics in changing the socio-economic situation.

Cluster 5 (11 constituent entities, 12.64% – Vladimir, Ivanovo, Kostroma, Tver, Vologda, Yaroslavl, Novgorod, Pskov, Kirov, Kurgan, Ulyanovsk Regions) includes regions with the reduced living standards. This cluster includes the constituent entities in which there are unfavorable trends, such as: a large outflow of population, the highest mortality rate of the population, lower level of education in comparison to other regions, a large number of people with the monetary income below the subsistence minimum (14.61%), and small monetary income of the population (23.6 thousand roubles).

Cluster 6 (12 constituent entities, 13.79% – Kalmykia, Dagestan, Ingushetia, Kabardino-Balkaria, Karachay-Cherkessia, North Ossetia - Alania, the Chechen Republic, Altai, Tyva, Buryatia; Zabaykalsky Krai, Jewish Autonomous Region) is characterized by the poorly developed social infrastructure, accompanied by a number of negative characteristics: the lowest values of GRP and rate of its development, the highest values of the poverty (24.74%) and unemployment (12.27%) indicators; the population has the worst housing provision and the lowest monetary income (20.7 thousand roubles), while a third of this small average income, as a rule, is spent on buying food. The combination of these social factors causes the strongest migration outflow of citizens, forming the lowest living standards in the Russian Federation.

4. Conclusion

The use of data mining methods allows one to build a holistic view of the living standards of the population in the country, taking into account the sustainable development goals. The component analysis method is used to form a preliminary idea of the character of the initial data distribution and to determine the approximate number of future clusters, as well as to identify the most characteristic features in the data. The use of the cluster analysis method allows one to finally build the structure of the cluster data and describe the features of each cluster and rules for assigning the constituent entities of the Russian Federation to it. The methods of component and cluster analysis, therefore, can become an important and necessary analytical tool in the formation of strategic decisions on the management of macroeconomic systems. They allow solving the problems of data structuring, identifying a unique set of characteristics for the selected groups of objects, which makes it possible to make balanced and informed decisions.

5. Acknowledgements

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

6. References

[1] O. M. Goch, On the Differentiation of Income of the Population in Modern Russia. Izv. Saratov University New ser. Ser. Economy. Management. Law. (2013) 3–2. URL: https://cyberleninka.ru/article/n/o-differentsiatsii-dohodov-naseleniya-v-sovremennoy-rossii, last accessed 2021/05/19.

[2] V. M. Zherebin, A. N. Romanov, The Living Standards of the Population. The Main Categories, Characteristics and Methods of Assessment, UNITY, Moscow, 2002.

[3] Transforming Our World: The Agenda for Sustainable Development up to 2030. The System of Global Indicators for Achieving the Sustainable Development Goals and for Executing the Tasks, Adopted by the UN General Assembly in Document A / RES / 71/313, URL: https://unstats.un.org/sdgs.

[4] Materials of the World Bank. LSMS Living Standards Research, URL: https://www.worldbank.org/en/programs/lsms.

[5] B. G. Ilyasov, E. A. Makarova, E. Sh. Zakieva, E. R. Gabdullina, The Quality of Life of the Population in the Regions of the Volga Federal District, Quality and Life 2 (10) (2016) 74–78.

[6] Materials of the Federal State Statistics Service. Regions of Russia. Socio-economic Indicators, URL: https://www.gks.ru/folder/210/document/13204.

[7] Materials of the Federal State Statistics Service. Distribution of the Monetary Income Total Amount and Characteristics of Monetary Income Differentiation of the Population, URL: https://www.gks.ru/folder/13723?print=1.