Available online at www.sciencedirect.com
SciVerse ScienceDirect
Procedia Computer Science 17 (2013) 687 – 694
Procedia Computer Science

Information Technology and Quantitative Management (ITQM2013)

Pattern analysis in the study of science, education and innovative activity in Russian regions

Aleskerov F.[a]*, Egorova L[a]., Gokhberg L[a]., Myachin A[a]., Sagieva G[a].

[a]National Research University Higher School of Economics, 20 Myasnitskaya str., Moscow, 101000, Russia

Abstract

We describe the method of pattern analysis and the results of its application to the problem of analyzing the development of science, education and the success of innovative activity in the regions of the Russian Federation. We examine characteristics of the regions of Russia such as the level of socio-economic conditions and the potential and efficiency of science, education and innovative activity from 2007 to 2010. Also we obtain a classification of regions by the similarity of the internal structure of these indicators, construct trajectories of regional development over time, and find groups of regions carrying out similar strategies.

Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management

Keywords: pattern, cluster analysis, regions of the Russian Federation, science, education, innovative activity

1. Introduction

The successful development of the regions in the long run is strongly connected with economic and social factors determined by the degree of development of education and science in the region. Any innovative activity is based on a solid basis of science and is not possible without proper personnel - not only scientists, but also students of all levels of education, as well as qualified employees. All these factors are interconnected, and the analysis of this relationship is difficult, but interesting problem. To solve it we use the method of pattern analysis.

Pattern analysis is rather new area of data analysis associated with a search for relationships among objects, the construction of their classification and the study of objects' behavior in dynamics.

More formally, the method of pattern analysis consists of three stages:

- information search, primary statistical processing of data and selection the basic system of indicators,
- construction of the feature space of the objects and clustering them to find patterns,
- the study of the objects behavior in dynamics using dynamic analysis of patterns.

* Corresponding author. Tel.: +7-495-621-13-42
E-mail address: alesk@hse.ru.

Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management
doi:10.1016/j.procs.2013.05.089

Thus the task of pattern analysis is to construct a partition of a given sample into subsets, called patterns, so that each pattern consists of similar objects, whereas objects of different patterns are significantly different. This means that each pattern reflects the essential characteristics of a class of objects that specifies objects of this pattern from the rest of the sample. Analyzing the behavior of patterns in dynamics we examine the changes of objects in patterns over time by constructing trajectories showing to which pattern the object belongs to at some moment of time. Objects with similar trajectories form dynamic groups. Generally, sample is separated into groups of objects according to the degree of patterns stability - from absolutely stable to completely unstable objects. This method allows highlighting similarities and differences in the dynamics of the objects functioning, identify common trajectories of development, and identify those objects that behave in an atypical manner and require more attention from experts.

Pattern analysis was developed in applied problems, such as studies of banks business type models [1, 2], comparison of macroeconomic development of different countries [3, 4], competition between political parties in the electoral districts [5] or problem of staff distribution at the bank branches [6-8].

In this paper pattern analysis is used to analyze the data of science, education and innovative activity in the regions of Russia.

1.1. Basic system of indicators

On the first stage a basic system of indicators has been constructed. It is based on a system of indicators designed for ranking of regional innovative development [9]. That system of indicators is multi-layered and includes 4 groups: social and economic conditions of innovative activity (macroeconomic fundamentals, characteristics of the educational potential of population), scientific and technical potential of innovative activity (including personnel and financial potential, publication and patent activity), innovative activity (the activity in the field of technological and non-technological innovations, the development of innovative small business, the cost of technological innovation and the impact of innovation), and the quality of the regional innovative policy (quality of the regulatory framework and organizational support for innovative policy, volume of consolidated budget expenditures in regions).

In our study the system of indicators was modified. We construct the following sets of indicators

1. Socio-economic conditions in the region;
2. Educational potential of the region's population;
3. Potential for research;
4. Potential for innovative activity;
5. Efficiency of innovative activity.

All data used has been taken from the Handbooks of the Federal State Statistics Service [10-12] and the Statistical Handbooks of HSE [13], the data covers 4 years from 2007 to 2010.

Then as in [9] the normalized values of each indicator for all regions are defined as the ratio of the difference between the value of the indicator in the region and the minimum value of it for all regions divided to the difference between the maximum and minimum values of this indicator for all regions

$$\widetilde{z}_{k,i}^{x} = \frac{z_{k,i}^{x} - \min\limits_{x}\left(z_{k,i}^{x}\right)}{\max\limits_{x}\left(z_{k,i}^{x}\right) - \min\limits_{x}\left(z_{k,i}^{x}\right)}, \tag{1}$$

where i stands for the number of indicator in the set k, x denotes the region, z_i^x denotes the value of indicator i in the region x. So, the normalized indicators are changed from 0 (the region with the minimum value of the indicator) to 1 (the region with the maximum value of the indicator).

Aggregate value z_k^x for a set k from the above list is calculated as the mean of the normalized values of the indicators from this set, defining 5 aggregated indices (below we will for short call them as indices). These indices have been used to characterize the region in the context of the development of science, education and innovative activity and form a description of the region $\left(z_1^x, z_2^x, z_3^x, z_4^x, z_5^x\right)$ in five-dimensional feature space.

1.2. Description of the pattern construction

Let X denotes a set of objects and Y denotes a set of numbers (names, labels) of clusters. A distance (metric) $\rho(x, x')$ between objects is specified for a formal description of 'closeness' of objects $x, x' \in X$. The set of objects X is split into disjoint subsets, called clusters, so that each cluster consists of objects that are close with respect to chosen metric, and objects of different clusters are significantly different. The cluster number y_i is assigned to each object $x_i \in X$.

The feature description of the regions we present in a system of parallel coordinates [14, 15], replacing the point $\left(z_1^x, z_2^x, z_3^x, z_4^x, z_5^x\right)$ in five-dimensional feature space by piece-wise linear functions. These functions are constructed as follows: on the x axis we put numbers of indices that characterize the structure of the objects; the y axis represents the values of these parameters. For each object we have a set of points that correspond to the values of these 5 indices. Piecewise-linear function is constructed by connecting these points with straight lines. This procedure is applied to each region.

The example of obtained piecewise linear functions are shown in Fig. 1.

In fact, we do not use the absolute values of indices in the pattern analysis and determine patterns not by the position of the points on the axis y but by the slopes of the corresponding lines, i.e., we use the slopes of piecewise functions in clustering. The formal description of the methods can be found in [5].

This approach solves the problem when the values of indices for different objects are different but have the same structure. For example, suppose there are two objects A and B, characterized by the following values of three indices (20, 50, 60) and (2, 5, 6). According to the traditional methods of cluster analysis these two objects will be assigned to the different clusters. At the same time it is obvious that their characteristics are similar, since the values of indices of the object A are equal to the values of the indices of object B multiplied by 10.

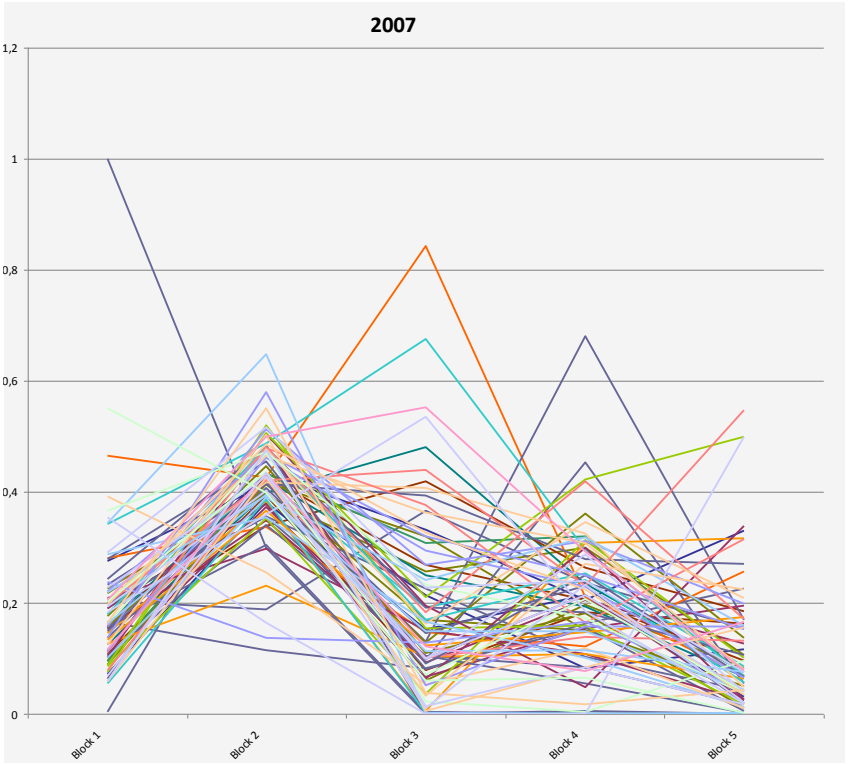

Fig. 1. The piecewise functions that represent 83 regions of Russian Federation, for this example the data was taken only for 2007

The seeming chaotic conglomeration of lines on Fig. 1 shows some regularities and one can see that many regions may have not equal values of indices, but the structures of them are very similar. It is reasonable to assume that if the regions have similar shapes of a piece-wise linear functions, then they have a similar structure of the indices and, therefore, these regions use similar models of science, education and innovative activity development.

1.3. Results

Taking into account that our purpose is an analysis of the dynamic behavior of the regions, the resulting piecewise linear functions for each year 2007-2010 were merged into the total sample. We had 83*4 = 332 objects for clustering.

We used two methods of clustering: k-means and hierarchical clustering method [16], in every method Euclidean metric was used. We conducted a multi-step procedure for cluster analysis, combining the above two methods and obtain 24 patterns involving more than two objects within the pattern, and 24 patterns with single object in each of them. Some of these patterns are shown on Figs. 3-5.

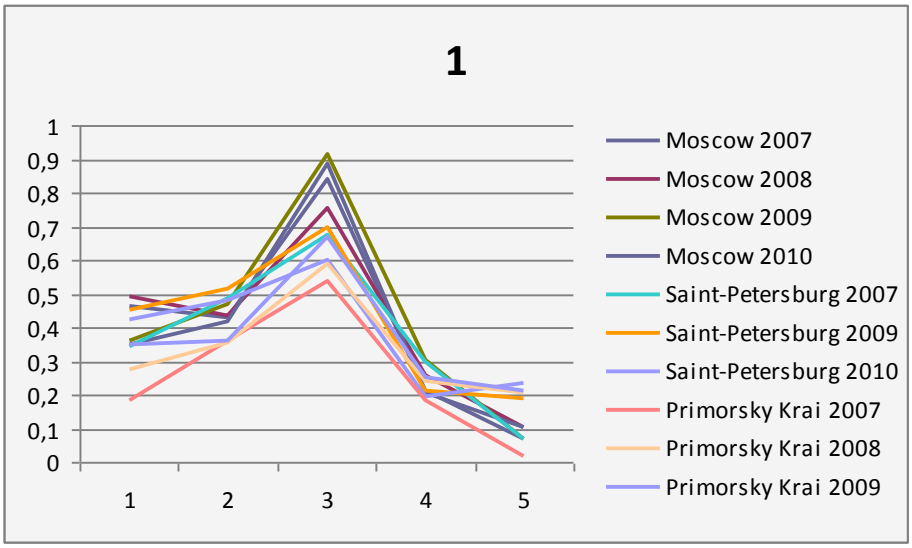

Fig. 2. Pattern 1

Pattern 1 contains 11 objects representing the city of Moscow and the Primorsky Krai for all 4 years from 2007 to 2010, and the city of St. Petersburg in 2007, 2009, 2010. This pattern is characterized by moderate (0.3-0.5) values of the first two indices ('Socio-economic conditions' and 'Educational potential'), high and somewhere extremely high (0.5-0.9) value of 'Effectiveness of research', and very low (less than 0.2) values for indices 4 and 5 responsible for the potential and effectiveness of innovative activity.

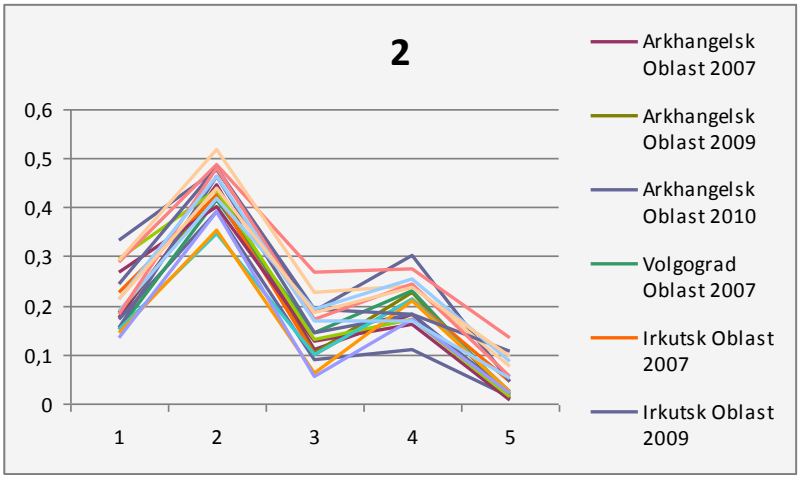

Fig. 3. Pattern 2

In Pattern 2 there are 19 objects. The main characteristics of this pattern are low (0.1-0.3) values for indices 1, 3, 4, 5, and medium (0.4-0.5) values of 'Educational potential'.

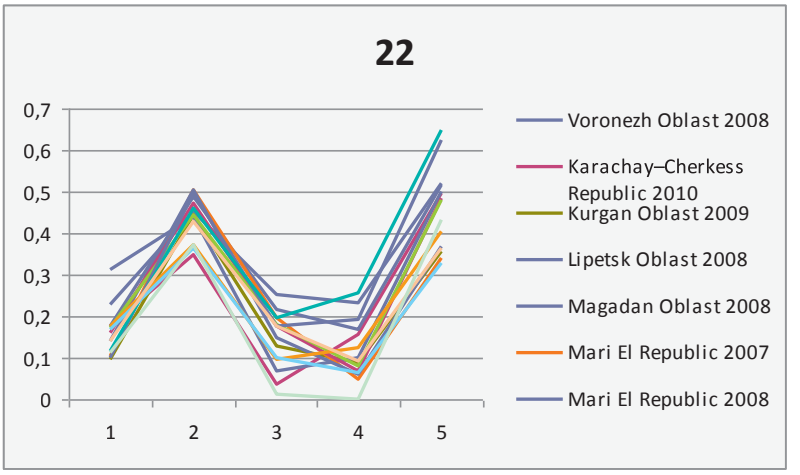

Fig. 4. Pattern 22

Pattern 22 contains 15 objects with very high effectiveness of innovative activity, with moderate educational potential and low level of the remaining indices.

Dynamic pattern analysis was conducted as well, which allows tracking what pattern each of the regions followed four years 2007-2010 based on the trajectories of the object. The trajectory of the object is an alternation of patterns, which describes changes of the object on the horizon of the analysis. Examples of such trajectories are shown on Fig. 5 and 6.

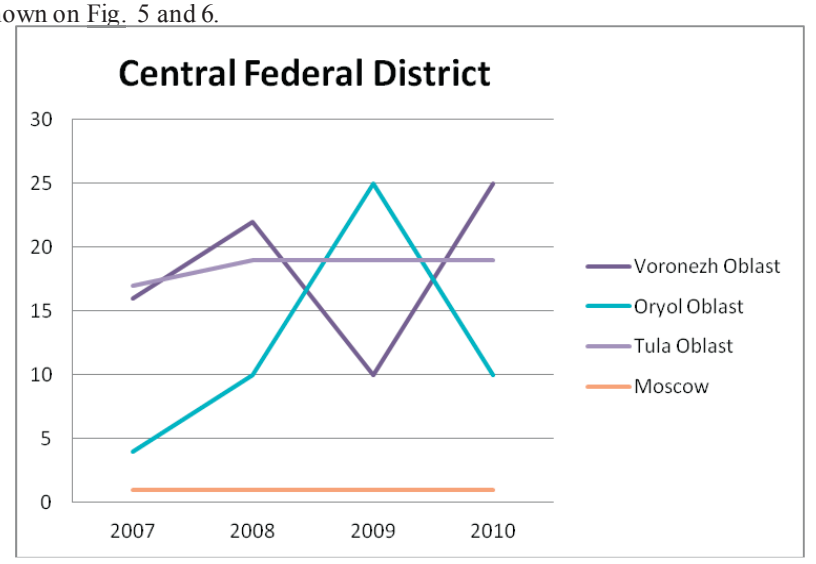

Fig. 5. Trajectories of some regions in Central Federal District over patterns. The numbers of patterns are on the y axis

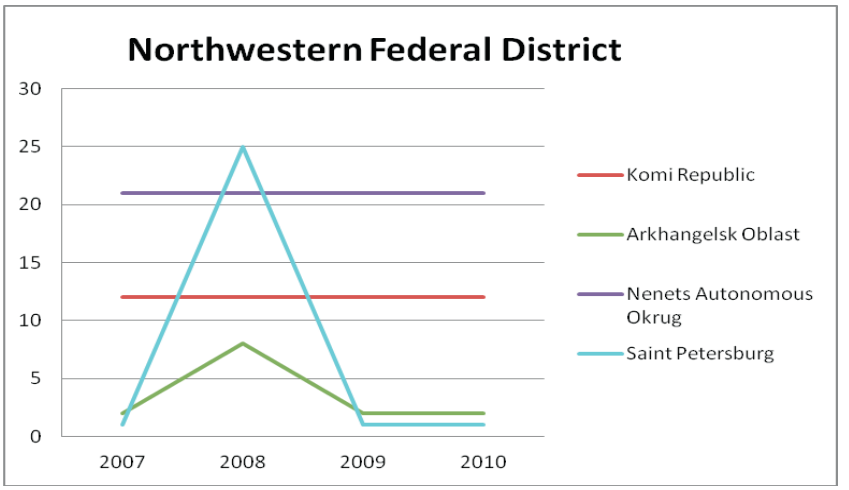

Fig. 6. Trajectories of some regions in Northwestern Federal District

All other figures of patterns and trajectories can be found in [17]. In this research the regions of Russia described with such characteristics as the level of socio-economic conditions and the potential and effectiveness of science, education and innovative activity were examined for 4 years from 2007 to 2010. A classification of regions was introduced with respect to the similarity of the internal structure of these indicators, trajectories of regional development over time were constructed as well, and a group of regions keeping the chosen strategy was found.

Acknowledgements

This work is a part of a more general project of data analysis of science, education and innovative activity performed by National Research University Higher School of Economics under the state contract No. 07.514.11.4144 "Development of an experimental sample of statistical analysis of science, education and innovation software using advanced techniques: pattern analysis and data ontological modeling" with Ministry of Education and Science, code 2012-1.4-07-514-0041.

F. Aleskerov and L. Egorova express their sincere gratitude to the Laboratory of Decision Choice and Analysis NRU HSE for partial financial support. The study was undertaken in the framework of the Program of Fundamental Studies of the Higher School of Economics in 2012.

References

[1] Aleskerov F.T., Martynova Y.I., Solodkov V.M. Assessment and analysis of the efficiency of banks and banking systems. Proceedings of the 2d International Conference "Mathematical Modeling of Social and Economic Dynamic", (MMSED - 2007), Moscow, 2007, 13-15 (ISBN 978-5-209-02632-7).

[2] Aleksashin P.G., Aleskerov F.T., Belousova V.Yu., Popova E.S., Solodkov V.M. Dynamic Analysis of Russian Banks' Business Models in 2006–2009. Working paper WP7/2012/03; Moscow: Publishing House of the University "Higher School of Economics", 2012. – 64 p. (in Russian).

[3] Aleskerov F., Alper C.E. Inflation, Money, and Output Growths: Some Observations. Bogazici University Research Paper, #SBE 96-06, 1996.

[4] Aleskerov F., Alper C.E. A clustering approach to some monetary facts: a long-run analysis of cross-country data. The Japanese Economic Review, v.51, no.4, 2000, 555-567.
[5] Aleskerov F., Nurmi H. A Method for Finding Patterns of Party Support and Electoral Change: An Analysis of British General and Finnish Municipal Elections. Mathematical and Computer Modelling, 2008, 1225-1253.
[6] Aleskerov F., Ersel H., Gundes C., Minibas A., Yolalan R. Environmental Grouping of Bank Branches and their Performances. Yapi Kredi Discussion Paper Series, No: 97-03, 1997, Istanbul, Turkey.
[7] Aleskerov F., Ersel H., Gundes C., Yolalan R. A Multicriterial Method for Personnel Allocation among Bank Branches. Yapi Kredi Discussion Paper Series, No:98-01, 1998, Istanbul, Turkey.
[8] Aleskerov F., Ersel H., Yolalan R. Multicriterial Ranking Approach for Evaluating Bank Branch Performance. International Journal of Information Technology and Decision Making, v.3, no.2, 2004, 321-335.
[9] Gokhberg et al. Innovative development rating of regions of Russian Federation. Moscow: Publishing House of the University "Higher School of Economics", 2012, 104 p. (in Russian).
[10] Statistical handbook of the Federal Service of State Statistics "Russian Statistical Yearbook". Internet Sources: http://www.gks.ru/bgd/regl/b11_13/Main.htm
[11] Statistical handbook of the Federal State Statistics Service, "Household Survey on employment". Internet Sources: http://www.gks.ru/bgd/regl/b12_30/Main.htm
[12] Statistical handbook of the Federal State Statistics Service, "Regions of Russia. Socio-economic indicators". Internet Sources: http://www.gks.ru/bgd/regl/b11_14p/Main.htm
[13] Indicators of innovative activity: 2008. Statistical Yearbook. Moscow: Publishing House of the University "Higher School of Economics", 2008. - 424 p. (in Russian). Internet Sources: http://issek.hse.ru/news/49369377.html
[14] Few S. Multivariate Analysis Using Parallel Coordinates, 2006. Internet Sources: http://www.perceptualedge.com/articles/b-eye/parallel_coordinates.pdf
[15] Inselberg A. Parallel Coordinates: Visual Multidimensional Geometry and Its Applications. Springer, 2009.
[16] Mirkin B. Clustering for Data Mining: A Data Recovery Approach, Chapman and Hall/CRC, Francis and Taylor, Roca Baton, Fl., 2005.
[17] Aleskerov F., Gokhberg L., Egorova L., Myachin A., Sagieva G. Study of science, education and innovation data using the pattern analysis. Working paper WP7/2012/07, Moscow: Publishing House of the University "Higher School of Economics", 2012. – 72 p. (in Russian).