The Methodological Approach to Multi-dimensional Classification of Regions by Investment Potential

Liliia V. Matraeva*

Russian State Social University, Russian Federation
**Corresponding Email: matraeva@rambler.ru*

Olga E. Bashina

Moscow University for the Humanities, Russian Federation

Doi:10.5901/mjss.2015.v6n6s7p194

Abstract

This article deals with the issues on classification of territories in terms of investment potential from the point of view of foreign investors. The cluster analysis approach is being used. In contemporary regional studies, this issue is covered by a sufficiently large number of scientific papers devoted to problems of development of the investment potential of the territories of the Russian Federation, which confirms the relevance and debatability of this problem. Current article raises questions on the use of hierarchical and non-hierarchical methods. The advantages and disadvantages of these methods are analyzed. The authors come out with a proposal of an integrated approach for classification purposes.

Keywords: *regional investment potential, multi-dimensional classification, complex approach*

1. Introduction

Over the recent years, a number of methods on evaluation of investment attractiveness have been developed by various organizations and consulting companies, which are being already widely used by Russian scientists. Among the most famous in this regard are the ratings of the agency "Expert", an approach of the Russian Institute for Urban Economics, Expert Institute of the Russian Union of Industrialists and Entrepreneurs, etc. (Bakhtizin & Akinfeeva, 2010). However, it is almost impossible to analyze the results of most of these techniques on a content level, as "…a rangeof the monitored indicators and the method of assigning weights remain closed to the general public. According to the developers, it is their commercial know-how" (Bakhtizin & Akinfeeva, 2010, p. 79). In this situation, it is the absence of a single perfect method causes the emergence of new techniques and constant refinement of their methodical positions.

Therefore, the main task in the development of the statistical evaluation of the investment potential of the territories of the Russian Federation is to achieve a holistic approach in building the model. The holistic approach (i.e. its complexity) in this study refers to the ability to objectively reflect the implicit and explicit investment resources of the region, as well as the efficiency of their use. Such an approach would not only characterize the level of development of investment potential, but also identify the most problematic areas that are subject to scrutiny and adjustment on behalf of the regional and national authorities. It will also enable a more rational allocation of resources in the framework of regional development and investment programs and will enhance the quality of managerial decisions in this area.

2. Map of Investment Potential Indicators

One of the main key points is the choice of factors to be included in the model. Thus, the total number of expert and statistical factors referred to in the literature are more than 140. The number of accounted indicators ranges from nine – used it Euromoney journal, to 381 – in the analysis of the Swiss Institute for Management Development (Burtseva, 2009; Folomiev & Revazov, 1999). It should be noted that the approach of Russian and foreign scientists in identifying groups of factors determining investment attractiveness varies considerably. The system of indicators, considered as part of this methodology has been constructed in accordance with the principles of international assessment, developed under the method of calculating the investment component of the index of global competitiveness. The proposed structure and coding system of indicators is presented in figure 1.

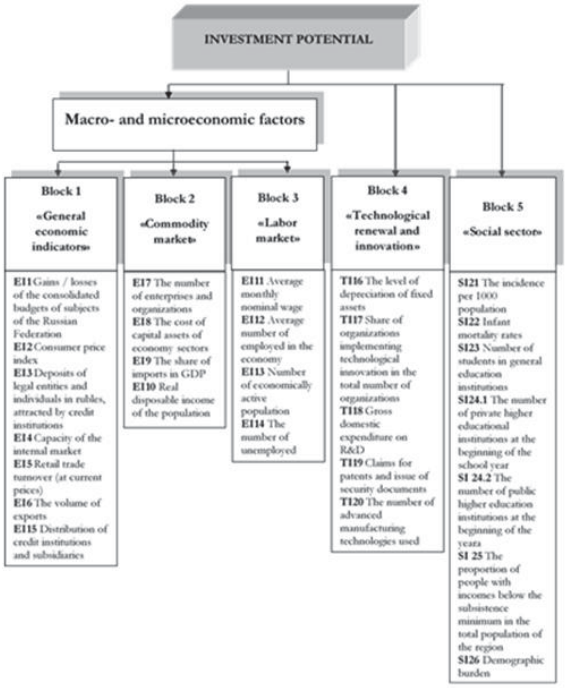

Figure 1. Map of statistical indicators of the investment potential of Russian regions

As can be seen from the data shown in figure 1, all of the above indicators reflect the socio-economic and technological components of the investment potential of the territory, which is subject to the *objective* systematic quantification of official statistical bodies. These are the fundamental factors that determine the socio-economic component of the investment potential of the territory, affecting the volume and trend of FDI in the long run. These factors cannot be changed in the short or medium term perspective. In addition, is very difficult to influence them by the only regional instruments of influence. In terms of the particularities of their statistical processing, it is advisable to use classical parametric methods of statistics that account for specifics of the formation of the initial source database.

Since the foreign investors are driven by different motives when choosing investment object, one must split the set of regions of the Russian Federation on investment potential into classes within which investors can have similar preferences. One of the features of this objective is the complexity of the investigated statistical data set. Firstly, the data describing the investment potential of Russian regions is characterized by heterogeneity and a high degree of differentiation. Secondly, there is a large number of linkages between its constituent elements.

In order not to exclude the mutually correlated variables from the analysis system and to strengthen regular component in the time series, the procedure of reducing the dimension of the original space data was involved using principal component analysis. As to maintain the logic of the study by the principal component, each of the selected blocks of indicators structured into major components identified – synthetic indices constructed that are set as a linear function of the input variables. These synthetic indicators are linearly independent by themselves. The number of principal components in each block is determined by the SPSS system according to the following rule: significant are the factors with eigenvalues greater than 0.95. The rotation of the axes in this case does not apply because its use (i.e. Varimax method) did not lead to an improvement in the interpretation of results. The following multi-dimensional classification was based on the selected principal components.

3. Multidimensional Classification of the Investment Potential

The first step in the construction of a multi-dimensional classification is agglomerative clustering, which is used to study the structure of the considered set and determine *the amount of stable clusters* of regions in the space of selected

indicators. To do this, it is advisable to use the method of Ward. The choice of this method is due to the following reasons. Firstly, the method of Ward creates clusters combining those that result in the smallest intra-cluster sum of squares. In particular, this method enables to obtain compact clusters of spherical shape, and it allows to more accurately create the 'image' of a set of coordinates of its typical representative. Secondly, it works better in case of 'fouled' data (Enyukov, 1989; Muller et al., 2001; Fuller et al., 2005).

Euclidean distance is used as a measure. Initially, the classification was performed according to n=80 for the period from 2000 to 2013. The options of classification from 2 to 14 clusters were considered. Data processing was performed using the statistical package SPSS 20.0. However, the results obtained in the first stage did not offer a complete classification of regions. Division of a given set gave stable allocation of clusters consisting of one observation. In this regard, it is competent to suggest that these regions are essentially abnormal, 'smearing' the results of observations, thus, there is a need to re-classify the classification for all regions except for Moscow, Chukotka Autonomous District and the Chechen Republic.

Re-classification of n=77 regions was also carried out on 11 main components using hierarchical cluster analysis, Ward's method and Euclidean distance. The resulting summary table of frequencies of clustered solutions indicates that the size of the majority of clusters are approximately equal, which is characteristic of the method of Ward. However, a clear image of the cluster solution by visual analysis of dendrograms and frequency tables was not generated. Analysis of the agglomeration scheme and the coefficient that characterize the distance between clusters, suggest the advisability of splitting the results into three or four groups. With such partition the single or small clusters do not form in either case. Meanwhile, the methods takes into account that if the number of clusters are divided into a number from 7 to 11, the number of full-fledged clusters remain equal to 3 or 4.

Thus, in this case, when determining the optimal number of clusters, it is necessary to use several additional criteria apart from the cluster size and inter-cluster distance, they are:

1) *Criteria for testing equality of variances*. In this step, one should analyze the variance of the studied parameters.
2) *Criterion of solution stability*. The resulting solution must maintain a certain degree of stability over time. This technique is very well laid out by Mitchell (1994).
3) *Criterion of interpretability*. There must be sufficiently clear differences in behavior between the clusters.

Comparative analysis of selected criteria divided by 3 and 4 groups led to the following conclusions.

On the criterion of testing the equality of variance. Test for equality of dispersions in this case is advantageously carried out by a test of Leuven, since this test, in contrast to the F-test, is insensitive to the requirement of normality of the initial data. The data obtained from the test of equality of variances showed that in the division into 4 groups, the value of intergroup variances exceed the value of intra-group variance at $p < 0.05$ for all indications, except for 'inflation'. For this factor test value of Louvain performed at $p < 0.25$-0.4. When partitioning into three groups, the level of statistical significance test is considerably lower for most features (i.e. factors). Thus, in the division into four groups, each resulting cluster has a more 'recognizable' image, which is inherent only to it in the value of various indicators by which one can get an idea about the features of the formation and development of the investment potential of the territory.

According to the criterion of solution stability, the partition into four groups is also preferably, as in this case, the quantitative composition of the groups are more stable in dynamics. Furthermore, when splitting into four groups, the core of the typical representatives of the group is clearer. Thus, the percentage of regions clustered to the same group in 6 out of 10 possible observations (i.e. typical cluster) was 75.32%. On the whole, in the division into 4 groups generated the following results: the first cluster has a stable composition of 19 regions out of 77; the second – 25; the third – 7, and the fourth – 7. As many as 19 regions were not classified according to the criterion of stability to any of the clusters.

On criterion of interpretability. All profiles of the mean values of clusters obtained in the division into four groups have similar significant differences in all the years of the analyzed period.

Similar results on the number of group partitions were obtained by clustering using the distant neighbor method. However, the clusters formed this way have been difficult to interpret, although being similar to the profile of the mean values.

The next step was the classification by K-means for which the number of clusters previously identified by Ward was set. However, this method proved to be rather insufficient while applied to a given data set. As shown in the results of previous clustering, the initial data set is not easy in terms of solving this problem:

- Firstly, there is a high degree of differentiation of nearly all of the inputs that distort the results of classic methods based on the use of averages.
- Secondly, there is a clear overlap in clusters. That is the classic non-hierarchical algorithms (i.e. k-means) are met with constraints in areas of overlapping clusters. This fact is further proved by Korolev (2007), Su & Chou,

(2001). The results obtained also confirm this fact.

Therefore, in this case, it is expedient to use the advanced algorithm of k-means: EM-clustering algorithm. Detailed description of the algorithm is given in the papers of Korolev (2007), Su & Chou, (2001). These publications present numerical experiments showing that k-means algorithm can be advantageous when dealing with non-overlapping clusters, but completely inferior to the EM algorithm in the presence of the overlap. Calculations were performed using the Data Mining package, implemented under the program STATISTICA 8.0.

The results obtained are similar to those obtained by hierarchical methods. Cluster sizes correspond to the sizes obtained by Ward method. However, this method of clustering implies that Leven criterion is statistically significant for all factors without exception throughout the analyzed period.

Comparison the results of clustering performed via EM and Ward methods reveals the following patterns.

According to the method of Ward, the 1st cluster composed of leader-regions, the 2nd cluster – typical representatives, the 3rd cluster composed of industrialized remote regions, and the 4th cluster – backward regions of the Russian Federation. The main difference between the results obtained by EM method from the results obtained by the method of Ward, is that a cluster of leading regions (cluster №1) was divided into two: 1st and 3rd. While the 3rd had only the most advanced regions, such as Moscow region and St. Petersburg. The fourth cluster was unambiguously determined by both methods of clustering. While the 2nd and 3rd clusters identified by the method of Ward were united in the cluster №2 according to EM method. Thus, the image is more blurred, as clusters of typical representatives and remote regions were merged into one, and the leader-regions composed of only a small group of the most advanced regions of the Russian Federation. In addition, the classification of the data using the number of 'stable' (i.e. resistant) group representatives is slightly higher compared with the method of Ward: the first cluster has 13 stable regions out of 77; the second – 31; the third – 4, and fourth -15. As many as 14 regions were not classified according to the criterion of stability to any of the clusters.

However, the results of both classifications cannot be considered fully abortive. Both classifications quite clearly formed 'core' of the groups. Moreover, the method of Ward showed itself more efficient in recognizing clear images, and the method of EM – in recognition of overlapping areas. For a more accurate picture, considering the specifics of the source data identified that is associated with overlapping boundaries of the groups and high differentiation by the values of attributes, it is expedient to use neural networks in the process of classification.

The indisputable advantage of using neural networks for solving the problem of clustering is that is was initially focused on the processing of multidimensional data. Research in this area shows that methods using neural networks are superior to conventional methods by combining the features of iterative and vast in parallelism of algorithms for problem solving (Su & Liu, 2005). To solve this problem, one can use an algorithm of multilayered perceptron or self-organizing networks based on competition. They are implemented within the Statistica 8.0 software package and the algorithm can be used both with or without a teacher. The advantage of self-organizing networks is that they do not require the assignment of classes and separate them in the learning process. However, practical experience of the study has shown that networks using the mechanism of learning or 'cramming' are better able to cope with multi-dimensional problems, because the results of the classification in this case are loosely defined by the features involved in the classification.

One of the main problems of the use of the algorithm is a learning procedure. This refers to the definition of the parameters for the neurons of the input layer, which provide the highest order when displayed for each of the attributes of the input into the network. The neural network shows better results than traditional methods, if a set of training vectors was adequately formed. As an educational group, in this case, are the clusters, *consisting of stable representatives belonging to the same group as a result of both classifications (Ward method and EM), as these groups of regions differ* in unique and sustainable combination of characteristics inherent to the investment potential.

Those regions of the list above that have the number of assignments to the group of more than 70% were taken as training. The rest of unambiguously defined regions were assigned to the test group. Undecided regions were classified as validation. Clustering was carried out on 11 main components.

To compare the characteristics of the typological groups, the average values of the clusters and their ratio in percentage to the average population were calculated. Correlation of the average values of investment potential of a specific group to the average of the entire set in 2013 are represented in figure 2. For clarity in representation of the results, a logarithmic scaling was used, since the classical range hampered visual interpretation due to high degree of dispersion of the results.

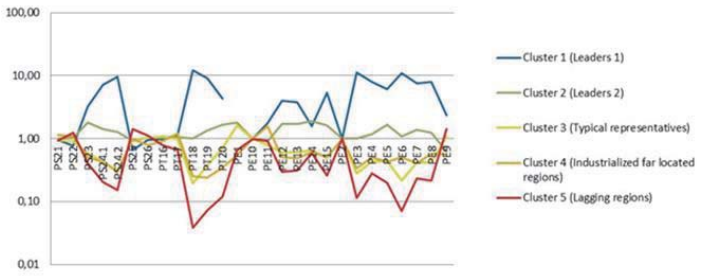

Figure 2. Average values in clusters of 2013 using the method of neural networks

During the period of 2000-2013, there has been quite a stable trend in the disparities of indicators characterizing the investment potential of Russian regions. The curves show the average values of the priority of the first cluster and significantly lagging position of the fifth cluster. However, the serial number does not indicate whether the characteristics of included territories are best and worst. Let us systematize individual characteristics inherent in each cluster in the context of the indicators characterizing the investment potential of the territory. This analysis will be carried out taking into account the sustainability of the trend:

- If the average values for the group are 25% higher than the average for the totality of this year for 8 of the 14 years in question, the value of this indicator is characterized as 'significantly above average';
- If the average values for the group are 25% lower than the average for the totality of this year for 8 of the 14 years in question, the value of this indicator is characterized as 'significantly below average';
- If the average values for the group deviates from the average for the aggregate of the present year for 8 of the 14 years considered are more than 5%, but not more than 25%, the values are treated as both 'above average' and 'below average'.

Final results of the componentwise dynamic analysis of clustering are presented in Table. 1.

Table 1. Final componentwise results of the clustering of 2000-2013 (using the method of neural networks)

Factors	Group				
	1	2	3	4	5
	Leaders № 1	Leaders № 2	Typical representatives	Industrialized remote regions	Lagging regions
SI21	L	L	L	H	L
SI22	SL	L	L	H	SH
SI23	SH	SH	SL	SL	SL
SI24.1	SH	SH	SL	SL	SL
SI24.2	SH	SH	SL	SL	SL
SI25	SL	L	UT	L	SH
SI26	L	H	H	L	L
TI16	L	H	H	L	L
TI17	UT	UT	L	L	SL
TI18	SH	H	SL	SL	SL
TI19	SH	SH	SL	SL	SL
TI20	SH	SH	SL	SL	SL
EI1	SH	UT	SL	SL	SL
EI10	UT	UT	UT	L	UT
EI11	SH	L	L	SH	L
EI12	SH	SH	SL	SL	SL
EI13	SH	SH	SL	SL	SL
EI14	SH	SH	SL	SL	SL
EI15	SH	SH	SL	SL	SL
EI2	UT	L	UT	UT	UT
EI3	SH	H	SL	SL	SL
EI4	SH	H	SL	SL	SL
EI5	SH	SH	SL	SL	SL

EI6	SH	H	SL	SL	SL
EI7	SH	SH	SL	SL	SL
EI8	SH	SH	SL	SL	SL
EI9	SH	SL	L	SL	SH

Note: UT – unsteady trend; L – low; SL – significantly lower; H – high; SH – significantly higher;

4. Conclusion

The research results present a hierarchical classification held using the methods of hierarchical and non-hierarchical cluster analysis of the territories in terms of investment potential from the point of view of foreign investors. The economic characteristics of the obtained clusters is given.

The analysis showed that a set of indicators tracked by Rosstat that define the component of investment potential cannot determine the clearly defined regional clusters because of the close matrix distances between regions. In this case, the classic methods do not allow to identify clear boundaries between the clusters and further division is rather subjective. Therefore, in order to address the objective of this study, a complex clustering method was used that is based on hierarchical and non-hierarchical methods and classification methods as well as the classification methods based on neural networks. This enabled to prevent distortion of the original data space and perform classification without loss of original data.

The results of the clustering presents a statistically sound delimitation of investment potential into classes for a totality of regions of the Russian Federation, within which the investors can have similar preferences, as well as to develop a differentiated investment classification of the regions based on the differences in the development of their investment potential. Thus, the alignment of the existing disparities in regional investment development and provision of the sustained economic growth requires the development and implementation of differentiated, cluster-specific regional investment policy in relation to foreign investors.

References

Bakhtizin, A.R., & Akinfeeva, E.V. (2010), Comparative evaluation of innovative potential of regions of the Russian Federation. *Problems of Forecasting*. 3, 73-81.

Burtseva, T.A. (2009), The indicative model of monitoring the investment attractiveness of the region. *Voprosy Statistiki*, 6, 37-45.

Enyukov, I.S. (1989), *Factor, discriminant and cluster analysis*. Moscow: Finance and statitiska.

Folomiev, A.N., & Revazov, V.G. (1999), Investment climate of the regions of Russia and ways of its improvement. *Voprosy Statistiki*, 9, 57-68.

Fuller, D., Hanlan, J., & Wilde, S.J. (2005), *Market segmentation approaches: do they benefit destination marketers?*. Coffs Harbour: Center for Enterprise Development and Research Occasional, Southern Cross University.

Korolev, V.Yu. (2007), *EM algorithm6 its modifications and their application to the problem of the distribution of probability distributions. Theoretical review*. Moscow: IPIRAN.

Mitchell, V.-W. (1994), How to Identify Psychographic Segments. *Marketing Intelligence & Planning*, 12 (7), 4-10.

Muller, K.-R., Mika, S., Ratsch, G., Tsuda, K., & Schölkopf, B. (2001), An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12 (2), 181-201.

Su, M.C., & Chou, C.H. (2001). A modified version of the K-means algorithm with a distance based on cluster symmetry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6), 674-680.

Su, M.C., & Liu, Y.C. (2005). A new approach to clustering data with arbitrary shapes. *Pattern Recognition*, 38, 1887-1901.