



International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Principal component analysis and cluster analysis for evaluating the natural and anthropogenic territory safety

T.G. Penkova^{a,b,*}

^a*Institute of Computational Modelling of the Siberian Branch of the Russian Academy of Sciences,
50/44 Akademgorodok, Krasnoyarsk, 660036, Russia*

^b*Siberian Federal University, 79 Svobodny pr., Krasnoyarsk, 660041, Russia*

Abstract

This paper presents an approach to evaluating the natural and technogenic safety of the one of the largest regions in Siberia through the comprehensive analysis of territorial indicators in order to explore geographical variations and patterns in occurrence of emergencies by applying the data mining techniques – principal component analysis and cluster analysis – to data of the Territory Safety Passports. For data modeling, two principal components are selected and interpreted taking account of the contribution of the data attributes to the principal components. Data distribution on the principal components is analyzed at different levels of the territory detail: municipal areas and settlements. Two- and three- cluster structures are constructed in multidimensional data space; the main clusters features are investigated. The results of this analysis have allowed to identify the high-risk territories and rank them according to danger degree of occurrence of the natural and technogenic emergencies. This evaluation gives the basis for decision making and makes it possible for authorities to allocate the forces and means for territory protection more efficiently and develop a system of measures to prevent and mitigate the consequences of emergencies in the large region. The suggested in this work approach in terms of its stages, techniques and reasoning procedures can be considered as a model of comprehensive multidimensional analysis of the control objects in various areas.

© 2017 The Authors. Published by Elsevier B.V.

Peer-review under responsibility of KES International.

Keywords: comprehensive multidimensional analysis; principal component analysis; cluster analysis; data mining; territory safety evaluation; prevention of emergencies, territorial management.

* Corresponding author. Tel.: +7 913-516-9981; fax: +7-391-290-7476.

E-mail address: penkovatg@gmail.com

1. Introduction

Prevention of natural and technogenic emergencies is one of the major tasks of the territory management. Analytical support of decision-making processes based on modern technologies and efficient methods of data analysis is a necessary condition for improving the territorial safety system and management quality.

The Krasnoyarsk region is the second largest federal subject of Russia and the third largest subnational governing body by area in the world. The Krasnoyarsk region lies in the middle of Siberia and occupies an area of 2.4 million square kilometers, which is 13% of the country's total territory. This territory is characterised by heightened level of natural and technogenic emergencies which is determined by social-economic aspects, large resource potential, geographical location and climatic conditions¹. In order to improve the population and territory safety, a lot of monitoring systems for on-line observation and for operational control of the state of technosphere and environment objects are being actively introduced within the region^{2,3,4,5,6}. The Ministry of Emergency has enacted the structure and order of conducting the Territory Safety Passport, which defines a system of indicators to estimate the state of territory safety, the risk of emergencies and possible damages to create efficient prevention and mitigation actions⁷. At present, there are massive data collections about the state of controlled objects, occurred events and sources of emergencies^{8,9}. However, we have to admit that the processing of stored data, aimed at obtaining the new and useful knowledge, is insufficient. The local databases remain unused, while the emergencies prediction, reasonable decisions and comprehensive analysis are sorely needed. Thus, identification of risk factors of emergencies based on monitoring data and investigation of their impact on key indicators of human safety are topical and important tasks in territorial management.

Data Mining, as the extraction of hidden predictive information from large databases, is a powerful modern technology of intelligent data processing. Data mining techniques provide the effective tool for discovering previously unknown, nontrivial, practically useful and interpreted knowledge needed to make decisions^{10,11,12}. This paper presents an approach to evaluating the natural and technogenic safety of the one of the Krasnoyarsk region through the comprehensive analysis of territorial indicators in order to explore geographical variations and patterns in occurrence of emergencies by applying the data mining techniques – principal component analysis and cluster analysis – to data of the Territory Safety Passports.

The outline of this paper is as follows: Section 1 contains the introduction. Section 2 describes the initial data. Section 3 presents the principal component analysis: identification and interpretation of principal components; analysis of data distribution on the principal components at different levels of the territory detail. Section 4 presents the cluster analysis: construction of two- and three-cluster structures in multidimensional data space and analysis of their basic features. Section 5 draws the conclusion.

2. Data Description

Analysis of natural and technogenic safety indicators is based on data of the Territory Safety Passports of the Krasnoyarsk region for 2014 collected in Center of Emergency Monitoring and Prediction (CEMP). Original dataset contains 1,690 objects, essentially discrete settlements-level geographical entities of the Krasnoyarsk region, each with 12 measured attributes. Data attributes are listed in Table 1. One part of attributes characterizes the sensitivity of the territory to the risk factors effects (e.g. population density, the presence of industrial and engineering facilities) that is determined by the number of objects located on the territory (i.e. a number of potential sources of emergencies), it is so-called "object attributes". The other part of attributes characterizes the presence of potential factor that can damage the health of people, can cause irreversible damage to the environment that is determined by the statistic of events occurred in the territory (i.e. a number of emergencies), it is so-called "event attributes". In addition, some locational reference characteristics are used for data interpretation and map visualization.

The preliminary correlation analysis of original data has shown a fairly strong relationship between "object" and "event" attributes, therefore for further analysis we will consider the attributes that characterize population and events. The correlation coefficients are presented in Table 2.

Table 1. List of the data attributes of Territory Safety Passports.

No	Attributes	Description
1	Pop	Population
2	Soc_object	Number of important social facilities (e.g. educational, health, social, cultural and sports facilities)
3	Water_object	Number of dangerous water bodies
4	Indust_object	Number of potentially dangerous industrial objects (e.g. plants, factories, mines)
5	Oil_line	Number of pipeline sectors in 5 km. radius from borders of settlement
6	Munic_object	Number of municipal facilities (e.g. power supply, water supply and heating facilities)
7	Flood_event	Number of floods
8	NFire_event	Number of natural fires
9	TFire_event	Number of technogenic fires
10	Munic_event	Number of accidents at municipal facilities
11	Nat_event	Number of natural events (excluding natural fires and floods)
12	Tech_event	Number of technogenic events (excluding technogenic fires and accidents at municipal facilities)

Table 2. Correlation coefficients between data attributes.

No	2	3	4	5	6	7	8	9	10	11	12
1	0.97	0.39	0.96	0.04	0.28	0.29	0.08	0.96	0.95	0.08	0.60
2		0.36	0.96	0.01	0.25	0.25	0.05	0.91	0.94	0.06	0.59
3			0.39	-0.01	0.32	0.60	0.12	0.39	0.36	0.17	0.30
4				0.01	0.24	0.29	0.05	0.91	0.91	0.07	0.56
5					0.08	-0.02	0.06	0.07	0.02	0.05	0.14
6						0.29	0.08	0.31	0.43	0.13	0.48
7							0.06	0.33	0.30	0.13	0.28
8								0.10	0.06	-0.02	0.05
9									0.93	0.11	0.63
10										0.08	0.58
11											0.13

Within this research, the analysis and visualisation of multidimensional data are conducted using the ViDaExpert¹³. Data visualization on geographical maps is performed by applying the mapping tools «ArcGIS»¹⁴.

3. Principal Component Analysis

Principal Component Analysis (PCA) is one of the most common techniques used to describe patterns of variation within a multi-dimensional dataset, and is one of the simplest and robust ways of doing dimensionality reduction. PCA is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components¹⁵. The number of principal components is always less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance and each subsequent component, respectively, has the highest variance possible under the constraint that it is orthogonal to the preceding components.

3.1. Contribution of the Data Attributes to the Principal Components

One of the greatest challenges in providing a meaningful interpretation of multi-dimensional data using PCA is determining the number of principal components. In general, the method allows to identify k components based on k initial attributes. Table 3 shows the results of calculating the eigenvectors of the covariance matrix arranged in order of descending eigenvalues.

Table 3. Results of principal components calculation.

Components	1	2	3	4	5	6	7
Eigenvalues	0.404	0.249	0.141	0.116	0.075	0.010	0.005
Accumulated dispersion	0.405	0.652	0.793	0.909	0.985	0.995	1
Pop	0.509	0.109	0.111	0.113	0.227	0.182	0.787
TFire_event	0.513	0.083	0.061	0.088	0.171	0.616	-0.557
NFire_event	0.060	0.439	-0.876	0.186	-0.022	-0.033	0.012
Munic_event	0.503	0.096	0.120	0.084	0.251	-0.764	-0.263
Flood_event	0.235	-0.314	-0.325	-0.853	0.109	-0.004	0.029
Nat_event	0.086	-0.822	-0.311	0.458	0.103	-0.015	0.010
Tech_event	0.397	-0.072	0.019	0.013	-0.913	-0.051	0.024

Based on combination of Kaiser's rule and the Broken-stick model¹⁶, two principal components for data attributes were identified (PC1 and PC2) with 65% accumulated dispersion. Figure 1(a) illustrates the eigenvalues of components. As can be seen from Figure 1(a), Kaiser's rule determines two principal components – eigenvalues of first two components are significantly greater than the average value and the Broken-stick model gives also two principal components – the line of Broken-stick model also cuts the eigenvalues of first two components. The contribution of the reduced data attributes to principal components is presented in Figure 1(b).

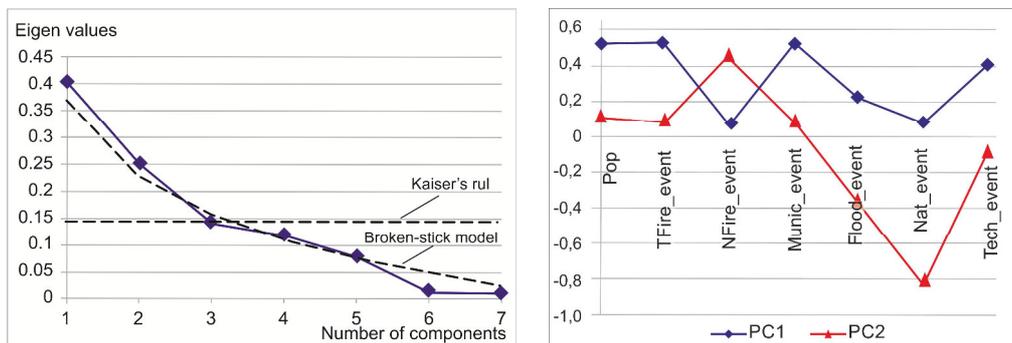


Fig. 1. (a) Eigenvalues of components; (b) contribution of the data attributes to the principal components

From Figure 1(b) we can see that the first principal component (PC1) is characterised by the following attributes: a high level of population, high proportions of technogenic fires, accidents at municipal facilities and other technogenic events, a low percentage of natural events including natural fires and floods. In combination, these characteristics present the big settlements (e.g. cities) with high levels of technogenic hazards. The second principal component (PC2) is characterised by the following attributes: a low level of population, a high proportion of natural fires, strong negative correlation with the percentage of natural events including floods and technogenic events including fires and accidents at municipal facilities. In combination, these characteristics present relatively small settlements (e.g. villages) with high levels of natural fires. This means that in comparison with other types of emergencies the technogenic and natural fires are the greatest threat for the Krasnoyarsk region.

3.2. Data Distribution on the Principal Components

The data can be divided into groups according to where the settlements are located in terms of Territory Standard. There are three standard levels of the territory detail: settlements, municipal areas and groups of municipal areas that give 1,690 objects, 65 objects and 8 objects respectively for the Krasnoyarsk region. Figure 2 shows the visualisation of standard groups (groups of municipal areas) on the geographic coordinates and the PCA plot, where: group 1 (green) – Angarsk Group; group 2 (rose) – Eastern Group; group 3 (purple) – Yeniseisk Group; group 4 (light blue) – Western Group; group 5 (yellow) – Central Group; group 6 (red) – Southern Group; group 7 (blue) – Taymyr Autonomous Okrug; group 8 (brown) – Evenk Autonomous Okrug.

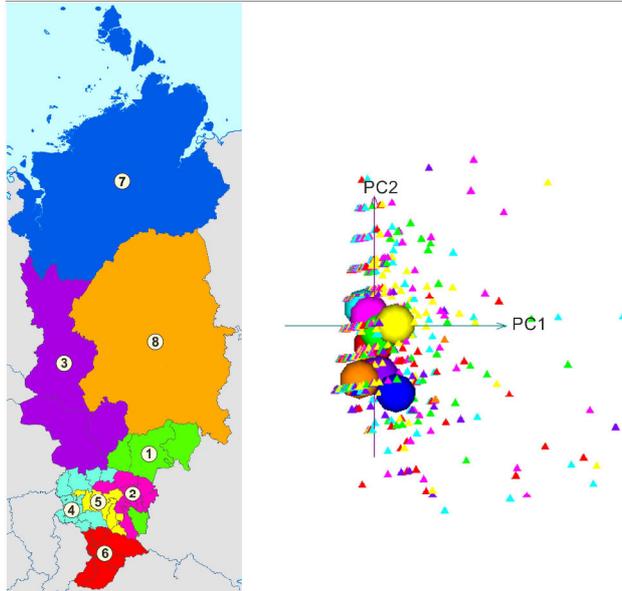


Fig. 2. Visualisation of territorial groups on the geographic map and the PCA plot.

As can be seen from Figure 2, along the first principal component (PC1) the territorial groups are concentrated quite densely, it means that technogenic fires are general characteristic for all territorial groups of region, but along the second principal component (PC2) the territorial groups are distributed significantly and we can see that the natural fires are indicative of northern territorial groups.

The visualisation of the projections on the first and second principal components on the geographic map is displayed in Figures 3 and 4. On these figures, the negative values in the range $[-1, 0]$ correspond to Group 1 (blue), the highest positive values in the range $(0.5; 1]$ correspond to Group 2 (red). The color intensity of municipal areas corresponds to the number of settlements in the group.

The lowest values of projections on the first principal component (Figure 3, blue points) are observed for such settlements as: Ust-Kamo, Shigashet, Kasovo, Verhnekemskoe, Srednya Shilka, Komorowskiy, Noviy Satysh, Angutiha, Lebed. It can be explained by the fact that these settlements are very small villages and, at present, in these settlements there are no any socially significant objects and residents. The complete absence of the economic activity in these settlements leads to the lowest level (or absence) of technogenic fires. The highest values of the projections on the first principal component (Figure 3, red points) are observed for such large settlements as: Krasnoyarsk, Norilsk, Achinsk, Kansk, Minusinsk, Lesosibirsk, Nazarovo, Emelyanovo, Aban, Yeniseysk, Berezovka. These settlements present the big cities of the Krasnoyarsk region where the population and number of socially significant and industrial facilities are above average level in the region.

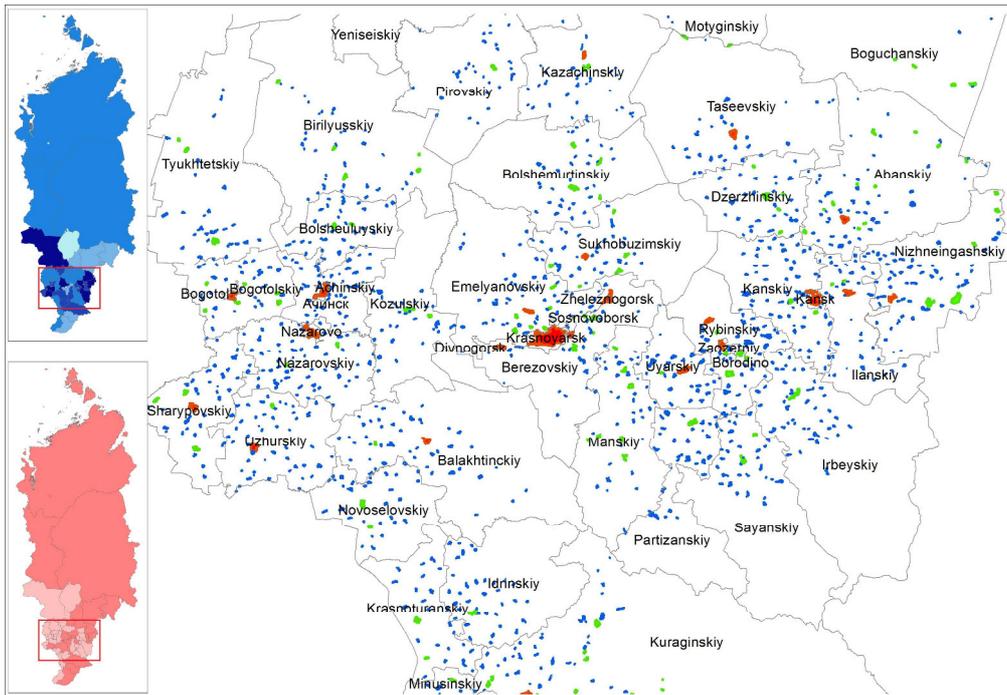


Fig. 3. Visualisation of the projections on the first principal component for municipal areas and settlements.

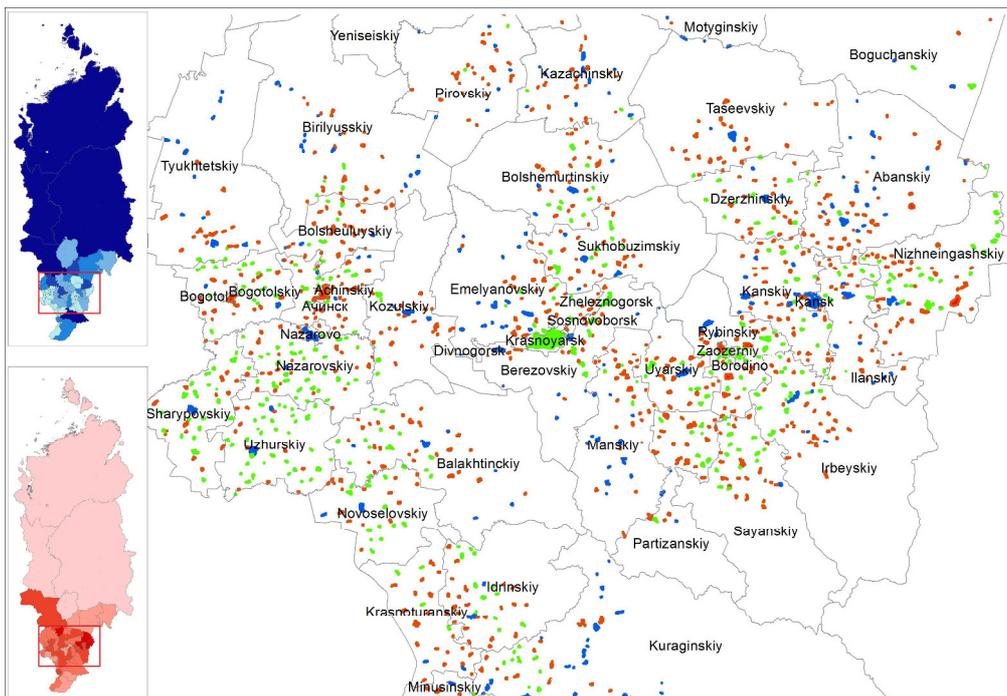


Fig. 4. Visualisation of the projections on the second principal component for municipal areas and settlements.

The lowest values of projections for the second principal component (Figure 4, blue points) are observed for such settlements as: Turuhansk, Cheremshanka, Tanzybey, Emelyanovo, Ermakovskoe, Nizhniy Ingash, Velmo, Kuragino and Uzhur. Low levels of natural fires can be explained by the following facts: the absence of vegetation as a source of emergency in steppe areas (e.g. Western and Southern groups) and the absence of settlements in forest zone (e.g. Evenk Autonomous Okrug, Yeniseysk and Turukhansky areas). The highest values of projections for the second principal component (Figure 4, red points) are observed for such settlements as: Startsevo, Tilichet, Kuray, Baikal, Glinniy, Udzhey, Abalakovo and Protochniy. The high risk of natural fires is observed in the large settlements that are located close to the forest zones. In addition, there is a probability of natural fires in the big cities where the forests constitute the part of their territories.

4. Cluster Analysis

Cluster analysis is a tool for discovering and identifying associations and structure within the data and typology development. Cluster analysis provides insight into the data by dividing the dataset of objects into groups (clusters) of objects, such that objects in a cluster are more similar to each other than to objects in other clusters. At present, there are many various clustering algorithms which are categorized based on their cluster model¹⁷. In this research, the centroid-based clustering method is used. *K*-means is a well-known and widely used clustering method which aims to partition objects based on attributes into *k* clusters. The *k*-means clustering is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. For the *k*-means clustering method the most important and difficult question is the identification of the number of clusters that should be considered. In this work, in order to determine the number of clusters the PCA technique was used: the number of clusters being dependent upon the number of principal components. Thus, referring back to the previous discussion, the first component forms two clusters, the second component forms three clusters. This means that the data has 2-3-cluster structures, where *k*=3, is the maximum number of informative clusters.

4.1. Two-Cluster Structure

In the two-cluster structure (*k*=2) Cluster 1 (blue) has 352 objects and Cluster 2 (red) has 1,338 objects. The difference between clusters is identified by the standard deviation of cluster averages of attributes. Figure 5 shows the distribution of the clustered data on the attributes in two-cluster structure.

As can be seen from Figure 5, the two clusters differ significantly on such characteristics as population and number of technogenic fires. In addition, Cluster 1 is characterized by high proportions of accidents at municipal facilities, natural fires and floods. Therefore Cluster 1 covers both large settlements with well-developed infrastructure that increases the risk of technogenic emergencies and large settlements with rich natural environment (e.g. forests, water bodies etc.) that increases the risk of natural emergencies. Cluster 2 combines small settlements with low risk of natural and technogenic emergencies. The distribution of the clustered data on the territories in the two-cluster structure is represented in Figure 6.

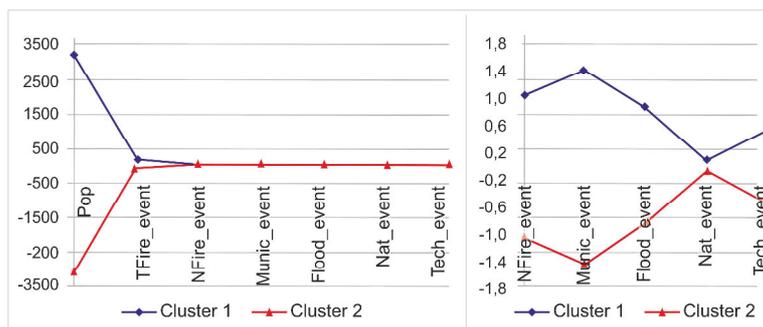


Fig. 5. Distribution of the clustered data on the attributes in two-cluster structure in small-scale (left) and large-scale (right) presentations.

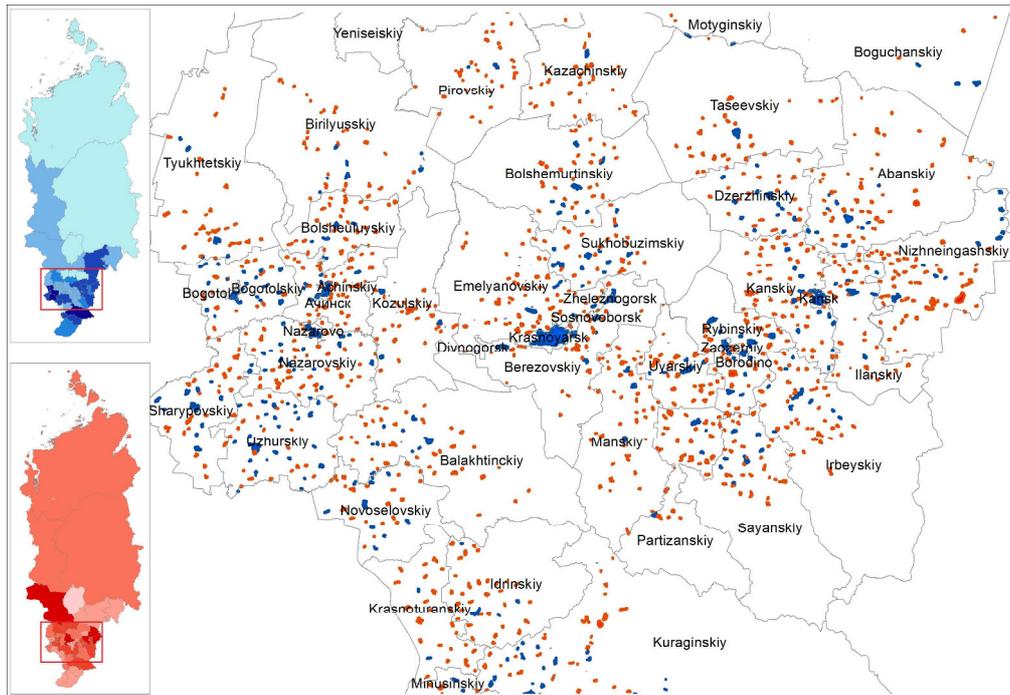


Fig. 6. Two-cluster structure on the geographic coordinates.

Representatives of Cluster 1 are the following biggest settlements: Krasnoyarsk, Norilsk, Achinsk, Kansk, Zheleznogorsk, Zelenogorsk, Minusinsk, Lesosibirsk, Sosnovoborsk and Nazarovo. Representatives of Cluster 2 are the following biggest settlements: Novohayskiy, Solnechniy, Kozulka, Podgorniy, Krasnoturansk, Zykovo and Krasnokamensk. A lot of industrial facilities and municipal facilities with high level of operation time in the big settlements lead to the high risk of technogenic emergencies; a lot of water bodies on these territories lead to the high risk of floods.

4.2. Three-Cluster Structure

In the three-cluster structure ($k=3$) Cluster 1 (blue) has 80 objects, Cluster 2 (red) has 720 objects and Cluster 3 (green) has 890 objects. Figure 7 shows the distribution of the clustered data on the attributes in the three-cluster structure.

As can be seen from Figure 7, the Cluster 1 differs significantly from Cluster 2 and Cluster 3 on such characteristics as population, number of technogenic fires, accidents at municipal facilities and other technogenic events. In contrast, Cluster 2 and Cluster 3 are characterized by low level of population and low proportions of natural and technogenic events in general but Cluster 3 demonstrates a trend to higher level of natural fires. Therefore, Cluster 1 combines the large settlements with well-developed infrastructure that increases the risk of technogenic emergencies. Cluster 2 combines the settlements with minimal risk of natural emergencies. Cluster 3 combines settlements where the basic threat is a natural fire. The distribution of the clustered data on the territories in the three-cluster structure is represented in Figure 8.

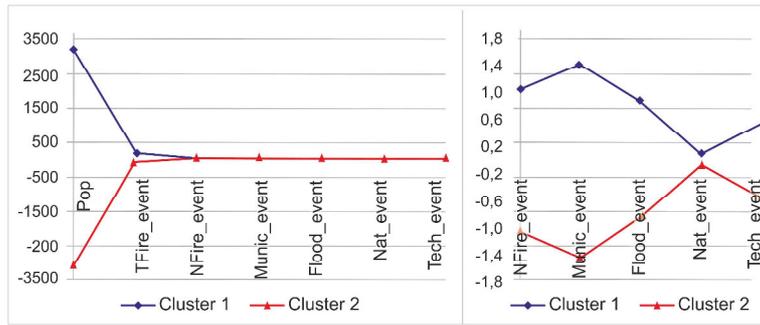


Fig. 7. Distribution of the clustered data on the attributes in three-cluster structure in small-scale (left) and large-scale (right) presentations.

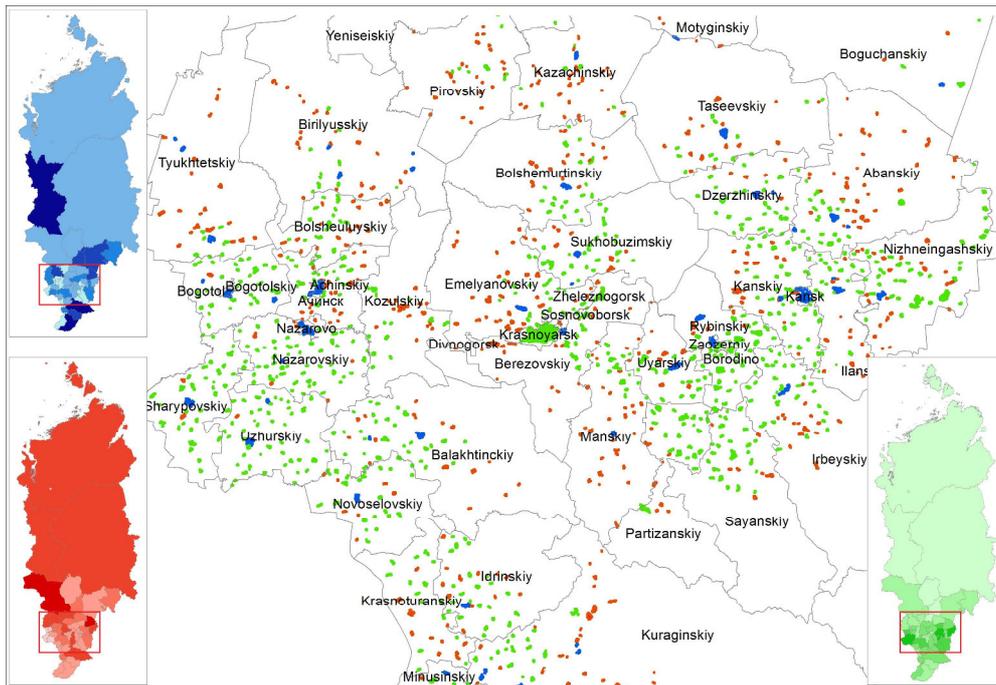


Fig. 8. Three-cluster structure on the geographic coordinates.

Representatives of Cluster 1 are the following biggest settlements: Achinsk, Kansk Zelenogorsk Lesosibirsk, Minusinsk, Sharypovo, Nazarovo, Norilsk; representatives of Cluster 2 are the following biggest settlements: Divnogorsk Kozulka, Severo-Yeniseisk, Podgorny, Krasnoturansk, Kedrovyy, Koshurnikova, Verhnepashino, Baykit; representatives of Cluster 3 are the following biggest settlements: Krasnoyarsk, Zheleznogorsk, Sosnovoborsk, Borodino, Shushenskoye, Kodinsk, Aginskoe. The high risk of natural fires is observed in the small settlements that are located close to the forest zones and in the large settlements where the forests are an integral part of their territory.

5. Conclusion

In this paper the evaluating of natural and technogenic safety of the Krasnoyarsk region in the context of settlements is carried out for the first time by applying the data mining techniques – principal component analysis

and cluster analysis – to data of the Territory Safety Passports. For data modelling, two principal components are selected and interpreted taking account of the contribution of the data attributes to the principal components. Data distribution on the principal components is analysed at different levels of the territory detail: municipal areas and settlements. Two- and three- cluster structures are constructed in multidimensional data space; the main clusters features are analyzed.

The data analysis results show that the technogenic and natural fires are a greatest threat for territory of the Krasnoyarsk region. The high risk of technogenic fires is observed in large settlements where the population and number of socially significant and industrial facilities are above average level in the region. The high risk of natural fires is observed in the large settlements that are located close to the forest zones and in the big cities where the forests are part of their territories. The explored geographical variations and patterns allow to identify the high-risk municipal areas and rank the territories according to danger degree of occurrence of the natural and technogenic emergencies. The results of this research, as a part of great work of emergency risk assessment make, it possible for specialists of CEMP to develop a system of measures to prevent and mitigate the consequences of emergencies in the Krasnoyarsk region.

The suggested in this work approach in terms of its stages, techniques and reasoning procedures can be considered as a model of comprehensive multidimensional analysis of the control objects in various areas.

Acknowledgements

The reported study was funded in part by the Russian Foundation for Basic Research according to the research project No 16-37-00014.

References

1. Report of the State of Natural and Anthropogenic Emergencies Protection of Territory and Population in the Krasnoyarsk Region. *Annual Report of Ministry of Emergency*. Krasnoyarsk; 2016 (in Russian).
2. Penkova T.G., Korobko A.V., Nicheporchuk V.V., Nozhenkova L.F. On-line Control of the State of Technosphere and Environment Objects in Krasnoyarsk region. *International Journal of Knowledge-Based and Intelligent Engineering Systems*; 2016, 20(2), p. 65-74.
3. Korobko A.V., Penkova T.G., Nicheporchuk V.V., Mihalev A.S. The Integral OLAP-Model of the Emergency Risk Estimation in the Case of Krasnoyarsk Region. *Proceedings of 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*; 2013, p.1456-1461.
4. Yronen Y.P., Yronen E.A., Ivanov V.V., Kovalev I.V., Zelenkov P.V. The Concept of Creation of Information System for Environmental Monitoring Based on Modern GIS-technologies and Earth Remote Sensing Data. *IOP Conf. Series: Materials Science and Engineering*; 2015, 94, 012023.
5. Shaparev N.Y. Environmental Monitoring of the Krasnoyarsk Region in Terms of Sustainable Environmental Management. *Informational and Analytical Bulletin (Scientific and Technical Journal)*; 2009, 18(12), p. 110-113 (in Russian).
6. Bryukhanova E.A., Kobalinskiy M.V., Shishatskiy N.G., Sibgatulin V.G.: Improvement of Environmental Monitoring Information Maintenance as an Instrument for Sustainable Social and Economic Development (on the Example of the Krasnoyarsk Region). *Informatization and Communication*; 2014; 1, p. 43-47 (in Russian).
7. The Standard Territory Passport of Regions and Municipal Areas. *The Regulation of Ministry of Emergency*; 25/10/2004, No.484 (in Russian).
8. Regional Organizational System of Emergency Monitoring and Prediction. *The Regulation of the Krasnoyarsk Region*; 09/02/2011; N 80-p (in Russian).
9. Nicheporchuk V.V. Information Support of the Natural and technogenic safety Management. *Siberian Fire and Rescue Journal*; 2016, 1, p.49-54 (in Russian).
10. Giudici P. Applied Data Mining. *Statistical Methods for Business and Industry*, John Wiley & Sons; 2005.
11. Williams G.J., Simoff S.J. Data Mining. *Theory, Methodology, Techniques, and Applications*; 2006.
12. Schürer K., Penkova T. Creating a typology of parishes in England and Wales: mining 1881 census data. *Historical Life Course Studies*; 2015, 2, p. 38-57.
13. Gorban A., Pitenko A., Zinovyev A.: ViDaExpert: User-friendly Tool for Nonlinear Visualization and Analysis of Multidimensional Vectorial Data. *Cornell University Library*. <http://arxiv.org/abs/1406.5550>.
14. Using ArcViewGIS: The Geographic Information System of Everyone, ESRI Press; 1996.
15. Abdi H., Williams L. Principal Components Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*; 2010, 2(4), p. 439-459.
16. Peres-Neto P., Jackson D., Somers K. How Many Principal Components? Stopping Rules for Determining the Number of Non-trivial Axes Revisited. *Computational Statistics & Data Analysis*; 2005, 49(4), p. 974-997.
17. Jain A., Dubes R. Algorithms for Clustering Data. Michigan State University: Prentice Hall; 1988.