The Problem of Clustering Countries of the World on the Geopolitical Data

I A Smirnov¹, O A Malafeyev¹, T E Smirnova¹, Yu V Yakovets², S A Strekopytov¹

¹Saint-Petersburg State University, Faculty of Applied Mathematics and Control Processes, 7/9 Universitetskaya nab., St. Petersburg, 199034, Russia
²Institute of P. Sorokin-N. Kondratiev, 6/1 Sretensky Boulevard, building 1, office 4, Moscow, 101000, Russia

E-mail: ivan.a.smirnov@inbox.ru

Abstract. This article deals with the problem of clustering countries by a specific set of geopolitical features, including demographic, geographical, economic, and other data. The K-means method was chosen as the clustering algorithm. This clustering allows us to identify the specifics and important trends in the development of the current geopolitical situation.

1. Introduction
The study of a country's geopolitical potential is one of the key issues in modern international relations. The correct measurement of this indicator makes it possible to clearly and clearly understand the balance of power in the international arena.

An important task in planning actions, making alliances, and identifying adversaries in geopolitics is to determine their specifics based on certain characteristics. However, as shown in [1], this task includes not one, but several features. This article deals with the problem of clustering the world's countries by specific demographic, economic, military, and technological characteristics. By assessing the dynamics of changes in these data in individual States, it is possible to identify important trends in their development.

For a better understanding of the specific potential, all these indicators should be taken into account at the same time. To work with multidimensional data, we have applied a number of methods that facilitate analysis and visualization. Namely, reducing the dimension by analyzing the main components, determining the optimal number of clusters using the elbow method, and the clustering algorithm itself by the K-mean method.

In the course of work [1-18], a cluster analysis was performed. For clarity, the results will be visualized and an overview of the composition of the final clusters will be presented.

2. Informal statement of the problem
The data is a table containing information about 219 countries by 32 criteria. For ease of displaying data, a unique three-letter code has been assigned for each country (See Annex 1).

The selected features include:
• economic: value of the territory (both General and urban and rural land separately), the population (both total urban and rural separately), the share of world GDP, the share index of production of food, the share of foreign exchange reserves, the share of the labor force and employment rates;
• military: absolute and percentage of the male population of military age in four groups (Men 20-24, Men 25-29, men 30-34, Men 35-39 years), the number of personnel of the Armed forces, military expenditures and their share of GDP, share of exports and imports of armaments, the mortality associated with military conflicts.

• scientific and technological: number of patents (for both residents and non-residents of the country separately), gross domestic expenditure on research and development, number of technical specialists and number of research and development researchers.

All data is taken from the world Bank's website and captures the state of Affairs for 2018.

Our task is to distribute the States listed above.

First of all, you need to build a mathematical data model, choose a clustering method, visualize the data, and get information for each state about its belonging to a cluster.

3. Using the elbow method to find the optimal number of clusters

One of the main difficulties in clustering a data set is determining the number of clusters that a set of samples can be divided into. Therefore, we need to use a method to quantify the quality of clustering. In the case of K-means clustering, it is better to use an internal metric, namely the intracluster sum of the quadratic error (SSE). It is also called the inertia of the cluster.

Based on this metric, we can calculate the measure of distortion for different numbers of clusters. For ease of interpretation of the result, we will use the elbow method. The idea of this method is that it can be used to easily identify when the distortion value becomes too large, and when it is acceptable. To illustrate and understand this method, we will plot the dependence of distortion on the number of clusters (See Fig. 1).

![Figure 1. Dependence of clustering distortion on the number of clusters.](image)

The figure shows that the elbow-the optimal number of clusters - for scientific, technical and economic data is quite early and ranges around 2-3 values. For military data, this value is greater than 4-5. For the entire set of attributes, the optimal number of clusters is close to 7-8.

4. The reduction of the dimensionality

We got a sample, each element of which is a vector of 32 features. When working with such multidimensional data, you must resort to methods of reducing the dimension. We used the principal component analysis (PCA) method as a dimensionality reduction algorithm. The essence of this method consists of two stages: a) finding the directions of maximum dispersion in high-dimensional data; b) projecting the data into a new subspace of equal or smaller dimensions. An important condition in this case is the orthogonality of the axes of new features to each other.

Applying this algorithm for reducing the dimension to 2 of two components to the source data, we get the following result (See Fig. 2). As you can see from figure 3, the relative positions of countries in the new subspaces differ from each other. So, in economic terms, the United States and India are
opposite poles, which can be explained by their greatest difference between them on these indicators: the share of GDP and the level of urbanization of the United States is significantly ahead of India. China is in an intermediate position between them: on the one hand, China is comparable to USA's economic might due to the high GDP and the largest gold reserves and a large population density, share of urban population China together with its Eastern neighbor. Against this background, Russia still looks quite weak in economic terms: its economic characteristics relate to Japan and Canada with digitalization, and on the other hand, to Brazil with its dependence on natural resources (although these data are not included in the general data set).

**Figure 2.** Projection of data after dimension reduction using principal component analysis: economic data.

5. The solution to the problem of clustering the States
For clustering, we use the most popular method – k-means (K-means method). This algorithm searches for a predefined number of clusters in an unmarked multidimensional dataset. This is achieved through a fairly simple representation of clusters. This method is based on two assumptions: the cluster center is the arithmetic mean of all points belonging to this cluster; each point is closer to the center of its cluster than to the centers of other clusters.

Now we will perform clusterization of countries by geo-political characteristics. At the same time, we will take into account the recommendations for the number of clusters for each data group and for the entire set of features as a whole (See Fig. 3).
As you can see from figure 3, the relative positions of countries in the new subspaces differ from each other.

5.1. Clustering of states by economic data

So, in economic terms, the United States and India are opposite poles, which can be explained by their greatest difference between them on these indicators: the share of GDP and the level of urbanization of the United States is significantly ahead of India. China is in an intermediate position between them: on the one hand, China is comparable to USA’s economic might due to the high GDP and the largest gold reserves and a large population density, share of urban population China together with its Eastern neighbor. Against this background, Russia still looks quite weak in economic terms: its economic characteristics relate to Japan and Canada with digitalization, and on the other hand, to Brazil with its dependence on natural resources (although these data are not included in the General data set).
For economic data, according to the elbow method, the entire set of countries was divided into two clusters (See Fig. 3a). The figure shows that India and China are in a common cluster, which is quite logical considering the above. The United States, forming its own separate cluster, is very different from all countries in this respect. Russia, on the contrary, is very close to the top four countries, including Brazil, Canada and Australia. It is surprising that the algorithm did not place Japan in the same cluster.

5.2. Clustering of states by military data
Relative to the subspace of military attributes (See Fig. 3b), we can say that the countries are distributed in a slightly more similar way: the three most distinguished countries in economic terms – China, India and the United States – and here form a separate group. On the other hand, there are three groups of countries on the left that have similar characteristics. Russia is in an intermediate position between the group of leading countries and medium-sized countries.

In military terms, the disposition of countries for 2018 is not as obvious as it might seem at first glance. It is logical that we have identified the top group of countries, led by China, India and the United States. It is also logical that a number of countries on the left are defined in a separate cluster, marked in red (See Fig. 3b). It includes such diverse countries as Monaco, Kosovo, Liechtenstein, etc., which do not have a particularly significant military potential. Afghanistan is a separate cluster. Other countries, including Russia, are represented in a large single cluster. Further increase in the number of clusters does not bring significant changes to the situation in Russia.

5.3. Clustering of states by scientific and technical data
In the subspace of scientific and technical features (See Fig. 3c) there is also some similarity: the US and China are again the designated leaders of technical innovation. Both are the largest patent registrars, but the structural difference is that China makes most of it from domestic patents, while the US makes most of it from official foreign borrowing.

In the subspace of scientific and technical features (See Fig. 3c) the countries were divided into 5 clusters. The former includes the Czech Republic and Korea. China and the United States, which are far from each other, have a very strong technological potential, but due to the above-mentioned difference between resident and non-resident patents, they were identified in two different clusters. Russia, Japan and Canada entered the common cluster as strong scientific and technological countries.

5.4. Clustering of states by whole data
The subspace of all features, as you can see from figure 6, takes into account all these features evenly, in proportion to their representation in the data. This is due to its greatest similarity with the military data.

In the general space of all indicators, all countries were divided into 8 clusters (see Fig. 10). It is noteworthy that a number of countries-Russia, Korea, Japan and the Czech Republic– were identified in separate clusters, which indicates that they are separate and unlike all others. Their common feature is their scientific and technical potential. The United States, China, and India were not United, which indicates that they are separate from each other in the economic, military, and scientific and technological spheres. A number of countries in the Middle East - the United Arab Emirates, Bahrain, Maldives, Oman, and Qatar–fell into a separate cluster.

6. Conclusion
As you can see from the examples above, the principal component analysis methods, the elbow method, and the k-means clustering algorithm together become good tools for analyzing and visualizing a large set of multidimensional geo-political data for countries around the world. Splitting multiple countries into clusters is quite meaningful and useful from the point of view of analyzing the geopolitical situation. Further use of the proposed set of tools can be used for research in modern economic, scientific, technical, military and other fields.
7. References
[1] Drozdov G D, Malafeyev O A, Nemnyugin S A 2015 Multicomponent dynamics of competitive single-sector economy development 2015 International Conference on "Stability and Control Processes" in Memory of V.I. Zubov, SCP 2015 Proceedings 7342188 pp 457-459
[2] Grigorieva X and Malafeev O 2014 A Competitive many-period postman problem with varying parameters (Applied Mathematical Sciences) 8(145-148) pp 7249-7258
[3] Kirjanen A I and Malafeyev O A and Redinskikh N D 2017 Developing industries in cooperative interaction: Equilibrium and stability in processes with lag Statistics, Optimization and Information Computing 5(4) 341-347
[4] Malafeev O A 1974 Equilibrium situations in dynamic games (Cybernetics and System Analisys) 10(3) 504-513
[5] Malafeev O and Nemnyugin S 1996 Generalized dynamic model of a system moving in an external field with stochastic components (Theoretical and Mathematical Physics) 107(3) 770 p
[6] Malafeev O A 1974 Equilibrium situations in dynamic games Cybernetics 10(3) pp 504-513
[7] Malafeev O A Stationary strategies in differential games USSR Computational Mathematics and Mathematical Physics 17(1) pp 37-46
[8] Malafeyev O A and Nemnyugin S A and Ivaniukovich G A 2015 Stochastic models of social-economic dynamics SCP Proceedings 483 – 485
[9] Malafeyev O A and Redinskikh N D 2018 Compromise solution in the problem of change state control for the material body exposed to the external medium, AIP Conference Proceedings
[10] Malafeyev O A and Rylow D and Zaitseva I and Ermakova A and Shlaev D 2018 Multistage voting model with alternative elimination AIP Conference Proceedings http://doi.org/10.1063/1.5043756
[11] Malafeyev O and Lakhina J and Redinskikh N and Smirnova T and Smirnov N and Zaitseva I 2019 A mathematical model of production facilities locatio Journal of Physics: Conference Series
[12] Malafeyev O and Farvazov K and Zenovich O and Zaitseva I and Kostyukov K and Svechinskaya T 2018 Geopolitical model of investment power station construction project implementation AIP Conference Proceedings
[13] Malafeyev O and Onishenko V and Zubov A and Bondarenko L and Orlov V and Petrova V and Kirjanen A and Zaitseva I 2019 Optimal location problem in the transportation network as an investment project: A numerical method AIP Conference Proceedings https://doi.org/10.1063/1.5114525
[14] Malafeyev O A, Nemnyugin S A 2016 Model of the optimal parameters choice for the charged particles beam 25th Russian Particle Accelerator Conference, RuPAC 2016
[15] Malafeyev O A, Redinskikh N D, Nemnyugin S A, Kolesin I D, Zaitseva I V 2018 The optimization problem of preventive equipment repair planning AIP Conference Proceedings 1978 100013
[16] Smirnov I A, Malafeev O A, Yakovets T Yu 2019 The Problem of Clustering Countries of the World On The Geopolitical Data (Integration Of The Greater Eurasian Partnership And The "One Belt, One Road" Initiative: Arctic Strategies, Programs, Projects) pp 246-258
[17] Vlasov M A and Glebov V V and Malafeyev O A and Novichkov D N 1986 Experimental study of an electron beam in drift space Soviet journal of communications technology & electronics 31(3) 145-149
[18] Zaitseva I, Malafeev O, Poddubnaya N, Vanina A, Novikova E 2018 Solving a dynamic assignment problem in the socio-economic system Journal of Physics: Conference Series 1172(1) 012092

Acknowledgment
The work is partially supported by the RFBR grant # 18-01-00796.