An Analysis of the Results of Polls with the Aim of Classifying the Regions of the Russian Federation[1]

E. A. Dem'yanov*, E. V. Dyukova, A. S. Inyakin**, and N. V. Peskov****

** Moscow State University, Vorob'evy gory, Moscow, 119234 Russia*

*** Dorodnitsyn Computing Center, Russian Academy of Sciences, ul. Vavilova 40, Moscow, GSP-1, 119991 Russia*

e-mail: egor@fom.ru, djukova@ccas.ru, andre_w@mail.ru, nick@motor.ru

Abstract—The use of cluster analysis in processing the results of polls carried out in regions of the Russian Federation in order to determine the basic characteristics of the regions is considered.

INTRODUCTION

We consider the following problem: to define the basic characteristics of the regions on the basis of the results of polls carried out in the regions of the Russian Federation. These characteristics should contain the essence of the results of the poll in each region and, consequently, represent the situation related to the problem under study in this region.

Two approaches are proposed.

The first approach consists in clusterization of the set of initial parameters (versions of the answers to the questions of a questionnaire) describing the regions of Russia and in the construction in each cluster of an aggregated parameter most strongly correlated with the other parameters. A heuristic method of the choice of initial partition of the set of parameters is proposed that allows one to increase stability and to improve the interpretability of results. A heuristic is proposed for defining the optimal number of aggregated characteristics.

The second approach consists in the following: a clusterization is carried out in such a way that respondents with the same political opinions get into the same cluster. Then, the regions of the Russian Federation are investigated in order to find out which part of the respondents of the region fall into some cluster. The task is specified by a large number of objects (respondents) and features (alternatives of the answers to the questions of the questionnaire); thus, the problem of the effective clustering of respondents arises.

The considered approaches were approved in the processing of the poll "Georating-2" carried out by the Public Opinion Foundation in 2003.

[1] This work was supported in part by the Russian Foundation for Basic Research, project no. 04-01-00795, and by a grant from the President of the Russian Federation for Scientific Schools, no. NSh 1721.2003.1.

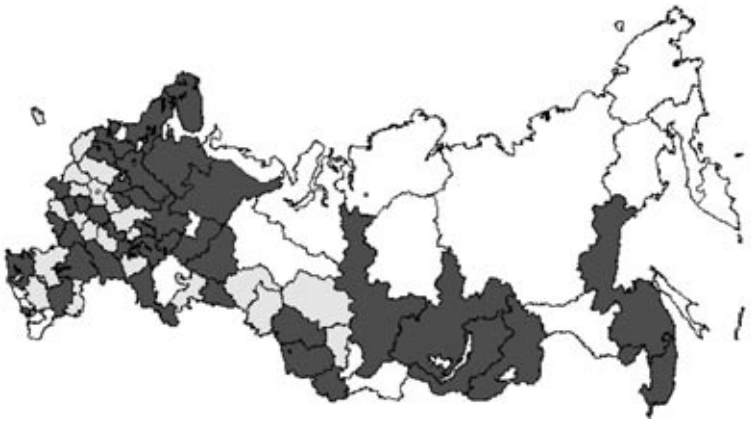
Received October 25, 2004

MAIN RESULTS

Linear correlation was used as a measure of the association between the parameters. The initial partitioning of the parameters into groups (clusters) was constructed in a random manner. Then, the maximization of a functional that is the sum of moduli of the correlations of each parameter with the factor of its group was carried out by the transfer of the parameters between the groups. For each group, the factors were chosen in such a manner as to maximize the sum of moduli of the correlations of parameters of the group with its own factor. These factors were used as the desired aggregated characteristics. The selected functional is appealing due to the fact that "good" partitioning of the parameters of regions into groups, from the point of view of an analyst, usually corresponds to the extremal (or close to extremal) value of the functional.

The use of this method yields the following difficulties: (i) the obtained clusters (and aggregated characteristics) are not always interpretable; (ii) the choice of the number of clusters into which the parameters should be partitioned is relatively laborious; and (iii) the method is very sensitive to the initial partition. In this paper, we propose a solution to these problems. Some modification of the model of the poll is carried out in order to join the informatively equivalent alternatives of the answers and to eliminate informatively uninteresting alternatives. The choice of the number of clusters and preparation of the initial partition are carried out in the following manner. The clusterization of the initial parameters into two clusters is carried out several times. According to the results of experiments, the large groups of parameters that always belong to a single cluster are separated. The parameters that "jump" from cluster to cluster are also separated. The required number of clusters is chosen according to the number of large groups. These large groups form the basis of the initial partition for the subsequent optimization of the functional. The remaining parameters are put in the clusters with factors with which they correlate most strongly. In most cases, this heuristic allows one to obtain a larger value of a functional than that corresponding to random initial partition. As a result, one

Pattern Recognition and Image Analysis, Vol. 15, No. 2, 2005, pp. 536–538.



The value of the "Supporters of the CPRF" factor in the regions of the Russian Federation.

can determine informative aggregated characteristics describing the political situation in the regions of the Russian Federation. One can use the chosen characteristics for a subsequent analysis of the regions of the Russian Federation, for example, for the creation of maps (see figure).

The clusterization of respondents is carried out in two stages due to the large volume of initial data. At the first stage, one uses a sufficiently rough method known as the method of *k*-means, and the "centers" of the clusters (the center is an object with feature values typical for the cluster) serve as the objects of further investigation. At the second stage, the clusterization of the centers is carried out by hierarchical grouping.

These methods were used in processing the poll "Georating-2," carried out by the Public Opinion Foundation in 2003 in the framework of a series of polls under general name "Georating." These polls investigate the political situation in the Russian Federation and are unique due to the fact that, for the first time, the sample of the polls is representative not only of Russia as a whole but also of each of its 65 regions. The total sample of each poll exceeds 32 000 respondents.

The use of heuristics for determining the optimal number of clusters in the problem of the aggregation of parameters reveals that the best choice is four clusters. The following interpretation of groups was obtained as a result of the aggregation procedure: the supporters of "United Russia" (ER), supporters of the Communist Party of the RF (CPRF), supporters of "Yabloko" and the "Union of the Right Forces" (SPS), and supporters of the "Liberal Democratic Party of Russia" (LDPR).

The black areas in the figure represent the regions with a high value of the factor "Supporters of the CPRF", gray represents a low value of the factor, and light gray represents neutral regions.

1500 groups were obtained by clusterization of the respondents by the method of "*k*-means." Then, the center (as a vector of typical values of features in the group) was taken instead of each group. After this, the clusterization of the centers of the groups into eight clusters was carried out by hierarchical grouping. As a result, each cluster contained from 40 to 400 centers. Finally, it was necessary to give an informational description to each cluster. This was carried out by calculating the percentage of specific answers to the questions by the respondents of the cluster.

Tables 1–8 present the interpretations of the obtained clusters of respondents. For each cluster, the features that allow one to interpret it are presented with

Table 1. Supporters of the LDPR

Feature	Value
Zhirinovskii/For whom would you vote in presidential elections?	0.75
LDPR/For which party will you vote in December 2003?	1
LDPR/Which parties do you trust?	0.92
LDPR/Which parties you do not trust?	0
Zhirinovskii/Which leaders do you trust?	1
Yes/Do you allow the possibility of your voting for the LDPR?	1

Table 2. Supporters of the CPRF

Feature	Value
Zyuganov/For whom did you vote in the last elections?	0.63
CPRF/For which party will you vote in December 2003?	0.96
CPRF/Which parties do you trust?	1
CPRF/Which parties you do not trust?	0
Zyuganov/Which leaders do you trust?	0.92
Yes/Do you allow the possibility of your voting for the CPRF?	1

Table 3. Supporters of democratic parties.

Feature	Value
SPS/Which parties do you trust?	0.6
Yabloko/Which parties you do not trust?	0.8
Yes/Do you allow the possibility of your voting for the SPS?	0.6
Yes/Do you allow the possibility of your voting for Yabloko?	0.8

Table 4. Supporters of Putin inclined to the CPRF

Feature	Value
Putin/For whom did you vote in the last elections?	0.87
CPRF/For which party will you vote in December 2003?	0.52
"Edinaya Rossiya"/Which parties do you trust?	0.06
CPRF/Which parties do you trust?	0.58

Table 5. Supporters of Putin inclined to Edinaya Rossiya

Feature	Value
Putin/For whom did you vote in the last elections?	1
CPRF/For which party will you vote in December 2003?	0.02
Edinaya Rossiya/Which parties do you trust?	0.87
Yes/Do you allow the possibility of your voting for Edinaya Rossiya?	0.91

Table 6. Those who found difficulty in answering

Feature	Value
It is difficult to answer/For whom would you vote in presidential elections?	0.54
It is difficult to answer/For which party will you vote in December 2003?	0.86
It is difficult to answer/Which parties do you trust?	0.66
It is difficult to answer/Which parties you do not trust?	0.68

Table 7. Those who refused

Feature	Value
Would not vote/For whom would you vote in presidential elections?	0.88
No/Will you vote in December 2003?	0.86
I do not trust any party/Which parties do you trust?	0.78
Do not trust anybody/Which leaders do you trust?	0.78

Table 8. Supporters of Putin who do not trust other parties

Feature	Value
Putin/For whom would you vote in presidential elections?	0.92
I do not trust any party/Which parties do you trust?	0.66
I do not trust anybody/Which leaders do you trust?	0.62

the arithmetic mean of the feature among the respondents of a cluster.

CONCLUSIONS

For the solution of the problem of relative analysis of the regions of the Russian Federation on the basis of the results of a poll of public opinion, first, methods of aggregation of the initial parameters (features) specifying the regions of the Russian Federation (objects) and, second, the clusterization of respondents were carried out. The following results were obtained.

Practical recommendations were made for modification of the initial data, which allow one to obtain results that can be better interpreted.

A heuristic is proposed for the determination of the optimal number of aggregated characteristics.

An approach is implemented that allows us to base the initial partition on certain considerations in clusterization of parameters rather than choose it in a random manner, which resulted in a higher quality of clusterization.

Informationally justified clusters of respondents, i.e., supporters of similar political opinions, are constructed.

The developed methods turned out to be applicable and allowed us to obtain results that could be easily interpreted.

REFERENCES

1. E. M. Braverman, *Structural Methods of Processing Empirical Data* (Nauka, Moscow, 1983).
2. E. G. Galitskaya, M. I. Levin, I. B. Muchnik, and V. G. Teterin, *Theory of Machine Learning: The Methods of the Analysis of Empirical Data* (Moscow, 1982).
3. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis* (Wiley-Interscience, New York, 1973).
4. Tuevo Kohonen, "Automatic Formation of Topological Maps of Patterns in a Self-Organizing System," in *Proceedings of the 2nd Scandinavian Conference on Image Analysis, Espoo (Helsingfors), Finland, 1981.*