On the measurement of population heterogeneity in electoral studies

Alla Tambovtseva

National Research University

Higher School of Economics

Draft, 13.03.2018

Introduction

Empirical quantitative research dedicated to elections usually requires handling large amount of aggregated data. The lower is the level of aggregation, the more difficult it is to consider the characteristics of population living within a chosen unit of analysis due to the lack of information. This difficulty causes a problem: within a certain low level of aggregation (electoral precinct) population is supposed to be homogeneous since it is impossible to identify distinct clusters of people. In some branches of research, in particular, in electoral fraud studies such an assumption calls forth serious debates: without careful testing it is not accurate to claim, for example, that turnout rate differs at two polling stations exclusively due to electoral malpractice and not because of natural reasons.

In most electoral studies the minimal unit of analysis is a district. This problem impedes the empirical studies of electoral processes in Russia as well since there is no data on social and demographic features of the population at a level of electoral precincts. However, recentlty with the rise of open-data resources the situation has turned better. The project *Reforma GKH* provides the house features on every living building such as the year of building, the operator or organization responsible for the maintainance, total square of living space and others. At the same time there are resources that aggregate the prices of flats in houses. Although such indicators cannot be proper proxies for socio-demographic data because there is no strong assosiation between house characteristics and its residents, the attempts to use such information still can be considered as a movement towards the more detailed approaches in political studies.

In this research I propose a method that allows to go deeper and choose an electoral precinct as a unit of analysis. Since there is no socio-economic and demographic data with high detail, some proxies are used. They include features of houses assigned to each precinct: year of building, average square of flats, average price of square meter. One important assumption is

made: people living in different types of houses differ in their social characteristics and, hence, in electoral behaviour. It is worth to note that this approximation is not completely new. It was used, for example, in [Makeeva, 2014]. However, in this research I want to use another approach - combine different levels of aggregation: district typology and precinct-level clustering.

This paper is aimed at answering the following questions. Firstly, does the clustering of electoral precincts based on house features correspond to the clustering based on spatial information? In other words, is it true that polling stations that are geographically close to each other refer to houses of the same type? Secondly, does the clustering of electoral precincts based on house features correspond to the clustering based on election results (turnout rate and percent of votes obtained by different parties?

In the research the data on State Duma elections in Moscow (2016) is analyzed. The year was chosen as a year of the recent federal elections to the State Duma.

Data and methods

The empirical part of this research is based on three main data sources: precinct-level electoral results taken from the official website of the Central Election Commission; district-level economic and demographic data from the Russian Federal State Statistics Service; house-level data obtained from the websites reformagkh.ru (house features) and mydata.biz (average prices per square meter in a house).

The research comprises three stages. At the first step I aggregate data by district and try to find distinct clusters of electoral precincts within each district. Three different types of clustering are obtained: based on house features, on spatial information, and on electoral results. As the house features I take the following indicators calculated by houses assigned to a particular electoral precinct: median year of house building, median average price per square meter, median average square of flats, and share of houses on the account of the regional operator. The spatial clustering is done based on geographical coordinates of polling stations, latitude and longitude. Electoral results used to get the clustering of the third type includes the turnout rate and the vote share for different parties. The parties are: Motherland-National Patriotic Union (*Rodina*), Communists of Russia, Russian Party of Pensioners for Social Justice, United Russia, The Greens, Civic Platform, Liberal Democratic Party of Russia, People's Freedom Party, Party of Growth, Civilian Power, Yabloko, Communist Party of the Russian Federation, Patriots of Russia, A Just Russia.

During the analysis I performed an iterative k-means clustering based on Euclidean distance with two clusters as a minimum possible number of clusters and ten as a maximum number of clusters. To find the optimal clustering the ensemble of 30 different methods [Charrad et al, 2014] is used, and then the recommended number of clusters is determined by the majority voting. Thus, the number of clusters suggested by most algorithms is chosen. The ensemble of methods seems to be a more appropriate approach to determine the number of clusters because of the two reasons. Firstly, it is a formal way to more accurate results that are more resistant to researcher's biases compared to techniques based on visual analysis (for instance, the Elbow method or Silhouette method). Secondly, it provides for more reliable results rather than the usage of a single method since the output of a certain algorithms might be unstable.

There were several problematic cases to handle. A few number of electoral precincts in a district did not allow to perform the ensemble procedure, so a more traditional technique was used. For such districts I performed a k-means clustering with one cluster as a minimum value and five clusters as a maximum value, plotted a graph and used the Elbow method. In the extreme case (*Molzhaninovsky district*) with only one electoral precinct the number of clusters is also one. This particular observation was excluded from the further analysis since it biases the distribution of similarity measures for comparing clusterings.

The results obtained by the procedure described above were checked against the substantial reasons for finding a certain number of clusters. I used hierarchical clustering to visualize the possible groupings and compared the results with ones given. The formal criteria were applied as well: Kruskal-Wallis test was used to verify whether the median values of features used for grouping vary significantly for different clusters.

At the second step I analyse the correspondence between clusters of electoral precincts based on house data and geographical or electoral information, so I compare the pairs of clusterings: *house features versus spatial information* and *house features versus electoral results*. During the analysis I use different visualisation techniques and formal methods for comparing clusterings by calculating similarity measures: the Rand index and the adjusted Rand index. The Rand index was chosen because it allows to compare two different clusterings pairwise: for each pair of items it is checked whether its elements belong to the same cluster in both cases. The formula is the following:

$$Rand\ Index = \frac{a+b}{a+b+c+d},$$

where a is the number of pairs that are in the same group in both clusterings, b is the number

of pairs that are in different groups in both clusterings, c is the number of pairs that are in the same group in the first clustering and in different ones in the second clustering, and d is the number of pairs that are in the same group in the second clustering and in different ones in the first clustering. The values of the index range from 0 to 1, and the higher values correspond to the higher similarity between clusterings. Thus, if the Rand index is close to 1, it means that groups in both clusterings consists of approximately the same elements.

The adjusted version of Rand index corrected for chance is also used:

$$Adj.\ Rand\ Index = \frac{RI - Expected(RI)}{max(RI) - Expected(RI)},$$

where RI is the Rand index, and $Expected(RI)$ is its expected value calculated based on a contingency table. Unlike the simple Rand index, adjusted index can fall behind the range [0,1] and take negative values.

As in the research the pairs of clusterings *house features versus spatial information* and *house features versus electoral results* are compared, it is evident that the optimal number of clusters can vary for different types of clusterings. In other words, for the same district it is possible, for example, to get four groups based on house features and six groups based on geographical coordinates. Although the Rand index is not vulnerable to the difference in the number of clusters, to ensure the reliability of results, I compute the similarity measures for the clusterings with equal number of groups as well. Thus, I take the optimal number of clusters for the data on house features, and than use this number in as the number of centroids in k-means method applied to spatial data.

At the last step I define district types by performing cluster analysis on the district-level socio-economic indicators: share of people in working age, share of retired people, and share of budget-sphere workers. The budget sphere in this research was approximated by workers in the health, education, civil service and . The hierarchical clustering is applied first, and then the k-means method is used.

Finally, I group data by district type derived at the beginning and do partitioning within each district type using the same clusterings and algorithms as at the previous stage.

Results

[Detailed discussion of results here]

Table 1: Rand index: house features versus spatial information

Statistic	Mean	St. Dev.	Min	Median	Max
Rand Index	0.49	0.14	0.18	0.49	1.00
Rand Index (equal)	0.53	0.13	0.25	0.51	1.00
Adj. Rand Index	0.04	0.16	−0.13	0.00	1.00
Adj. Rand Index (equal)	0.05	0.16	−0.11	0.01	1.00

N = 117

Table 2: Rand index: house features versus electoral results

Statistic	Mean	St. Dev.	Min	Median	Max
Rand Index	0.50	0.13	0.18	0.50	1.00
Rand Index (equal)	0.54	0.13	0.23	0.51	1.00
Adj. Rand Index	0.04	0.16	−0.25	0.00	1.00
Adj. Rand Index (equal)	0.04	0.16	−0.25	0.00	1.00

N = 117

Based on the social and demographic data it is possible to determine three clusters of districts: 1) districts with the low share of people in the working age, the high share of retired people and the medium share of budget-sphere workers; 2) districts with the medium share of people in the working age, pensioners, and budget-sphere workers; 3) districts with the medium share of working-aged population, the low share of retired population and the high share of people working in the budget sphere.

[Histograms by clustres here]

Conclusion

In this paper I developed a method that allows to find groups of electoral precincts (polling stations) within districts. One of the main findings of this research is that clusters of precincts within each district do not differ significantly from clusters found within districts of a particular type. It means that it is not compulsory to perform cluster analysis for every district of interest; it is enough to perform it on data aggregated by district type. So, it is an economical and more general way of including exogenous information in electoral studies. Such an approach makes it easier to process large amount of data (on all Russian regions, for example) without increasing the unit of analysis.

It was found that there is no correspondence between the location of polling stations and types of house assigned to these stations. Electoral precincts that are geographically close to each other rarely consist of absolutely different house types. At the same time it was seen that clusters of precincts in Moscow based on election results do not coincide neither with house type clusters nor with district type clusters. It may serve as the evidence that population heterogeneity (if approximated by demographic and house data) is not seen at the lowest level of aggregation used in electoral studies and, thus, cannot be the serious reason explaining differences in voting behaviour at precincts in the neighbourhood.

References

Charrad M., Ghazzali N., Boiteau V., Niknafs A. (2014). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set, Journal of Statistical Software, 61(6), 1-36.

Makeeva A. Whethera natural pattern of spatial socio-economic segregation may explain a "mosaic" electoral behaviour in St. Petersburg (on the basis of a dataset of polling stations' reports for State Duma elections of December 4, 2011). Sociologiya v dejstvii – 2014. Izbrannye materialy VI sociologicheskoj mezhvuzovskoj konferencii studentov i aspirantov, SPb.: Otdel operativnoj polirafii NRU HSE – St.Petersburg. 2014.