On the measurement of population heterogeneity in electoral studies

Introduction

Empirical quantitative research dedicated to elections usually includes working with 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 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. No direct meaningful information on precincts or, at least, streets, is available, so the only data that can be analysed at such a level are namely electoral results. 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][1]. However, in this research I want to use another approach - combine different levels of aggregation: district typology and precinct-level clustering.

Research questions

The research questions are the following. 1) 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

[1] Makeeva A.Whether a 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 poligrafii NRU HSE – St.Petersburg. 2014.

same type? 2) 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)?

Methods and data

In the research the data on Moscow (2016) is analyzed. The year was chosen as a year of the recent federal elections to the State Duma. The empirical part of this research is based on three main data sources: 1) precinct-level electoral results taken from the official website of the Central Election Commission; 2) district-level economic and demographic data from the Russian Federal State Statistics Service; 3) house-level data obtained from the website *reformagkh.ru* and *mydata.biz.*

Throughout the whole research hierarchical cluster analysis is applied first so as to learn the number of clusters, and then the k-means clustering is used. The empirical part comprises three stages. At **the first step** I work with district-level socio-economic data so as to define types of districts. Cluster analysis is based on the following indicators: share of people in working age, share of retired people, unemployment rate, and share of budget-sphere workers. At **the second step** I aggregate data by district and try to find distinct clusters of electoral precincts within each district. Cluster analysis is based on house data (all 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, share of houses on the account of regional operator. At **the last step** I group data by district type derived at the beginning and do clustering within each district type using the same data as at the previous stage.

To answer the research questions two additional clusterings are obtained: geographical and electoral. The former involves spatial clusters within each district. The latter comprises several variants – groups derived from turnout rate and vote share for different parties. To analyse the correspondence between clusters of electoral precincts based on house data and geographical/electoral information I use different visualisation techniques and formal methods for comparing clusterings by calculating similarity and dissimilarity measures (the Rand index, Jaccard measure, etc).

Results

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 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.