Masters Program in Geospatial Technologies

SPATIAL PATTERNS AND IRREGULARITIES OF THE ELECTORAL DATA:

GENERAL ELECTIONS IN CANADA

Alexey Eskov

Dissertation submitted in partial fulfilment of the requirements
for the Degree of *Master of Science in Geospatial Technologies*

UNIVERSIDADE NOVA ISEGI
UNIVERSITAT JAUME·I
WESTFÄLISCHE WILHELMS-UNIVERSITÄT MÜNSTER

SPATIAL PATTERNS AND IRREGULARITIES

OF THE ELECTORAL DATA:

GENERAL ELECTIONS IN CANADA

Dissertation supervised by

PhD Jorge Mateu Mahiques,

PhD Marco Painho,

PhD Edzer Pebesma

March 2013

ACKNOWLEDGEMENTS

First of all, I would like to express my gratitude and appreciation to EU educational bodies, particularly EACEA and MSGT consortium, which made this study possible for me and many people from different countries.

I wish to acknowledge the help provided by Dr. Mateu, who initially raised my interest to statistics, being a teacher, and then kindly agreed to be my thesis supervisor. His guidance was enthusiastic and patient at the same time, making our work a great pleasure for me. My co-supervisors from WWU and ISEGI, Dr. Pebesma and Dr. Painho, have provided valuable suggestions and additional points of view on the problem, helping me to clarify things where necessary.

I would like to give special thanks to Dori Apanewicz who was helping us from the very first till the last days of our stay and made many things much easier than they could be for us.

And of course, I'm particularly grateful to my family and my partner for their continuous support.

SPATIAL PATTERNS AND IRREGULARITIES OF THE ELECTORAL DATA: GENERAL ELECTIONS IN CANADA

ABSTRACT

Democratic elections are one of the most important social phenomena of the last centuries. Countries which publish elections results on the polling station level provide a valuable source of data for different groups of scientists like geographers and statisticians. In this work, we combined geographical and statistical analysis, pursuing a goal of defining the spatial patterns and irregularities of the electoral data. From theoretical point of view, it will help to find out if the electoral behavior has any spatial dependency. From practical perspective, it can give a new insight about the electoral fraud detection. We have applied a set of statistical methods to estimate the distribution and variability of the electoral behavior in space and time for different geographic units. Canada was selected as a study area because it is an old democracy where the elections are considered being fair, and all the necessary data are available.

KEYWORDS

Electoral geography

Spatial analysis

Voter turnout

Party share

Electoral district

Polling division

Distribution

Variability

Correlation

Spatial autocorrelation

Moran's Index

Local Moran's Index

Cluster and outlier analysis

Clustering

Neighborhood

PostgreSQL

PostGIS

R

Canada

INDEX OF THE TEXT

INDEX OF THE TABLES

INDEX OF THE FIGURES

1 INTRODUCTION

1.1 Theoretical Framework

Democratic elections are one of the most important social phenomena of the last centuries. Since voting process requires personal presence of each voter and must be completed in a strictly limited period of time, there should be a great number of polling stations. The electoral agencies aggregate voting results from the stations to get final results, i.e. selected party standings in the Parliament. In some countries they publish data at polling station level, while in others they present only intermediate aggregation results. In any way, these are the valuable sources of data for different groups of scientists like geographers and statisticians.

One of the main research directions in the electoral science is the contextual study. Geographers and statisticians are trying to relate a certain voting behavior to socio-economic context of the particular geographic areas. This context can be very different, from the ethnicity to the level of income. Examples include, but are not limited to:

- Impact of Negro migration on the electoral geography of Michigan (P. Lewis, 1965)
- The electoral geography of recession: local economic conditions, public perceptions and the economic vote in the 1992 British general election (Pattie et al, 1997)
- Protestant support for the Nazi Party in Germany (J. O'Loughlin, 2002)
- The territorial variable in the analysis of electoral behavior in Spain (A. de Nieves and M. Docampo, 2013).

In general, the papers confirm a strong influence of the socio-economic context on the electoral behavior. For instance, in the last of the abovementioned papers the authors classify the territory of Spain into habitats by land use type: city, periphery, small urban, deactivated rural, agrarian rural and manualized rural. Then, they argue that each habitat has its own electoral portrait: “peripheries are the habitat that is more clearly oriented towards left-wing politics”, “...cities are clearly

characterized for their electoral support to green parties", "deactivated rural ... is certainly one of the habitats with highest percentages of electoral support to right-wing parties" (A. de Nieves and M. Docampo, 2013) and so on.

Moreover, the local context has a synergetic effect because the local majority affects the local minority: "people tend to vote in a certain direction based upon the relational effects of the people living in the neighborhood" (Cox, 1969). To describe this phenomenon, K. Cox introduced the term "neighborhood effect". His approach was actively developed by a group of British researchers, mainly R. Johnson, C. Pattie and W. Miller. Their findings "provided a very strong circumstantial evidence of neighborhood effects, local polarization produced through social interaction in which the area's majority political opinions is accentuated through processes of 'conversion through conversation'" (Johnson et al, 2000).

Going further, we could expect that neighborhoods tend to share similarity between each other. As so-called Waldo Tobler's first law of geography says, "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). This concept was developed for socio-economic data by L. Anselin in "Spatial Econometrics: Methods and models" (Anselin, 1988). We could expect that if the independent socio-economic variables are spatially determined, dependent variables like voting behavior could inherit such kind of distribution. Also, some people move between the neighborhoods, so the process of 'conversion through conversation' could work not only internally but externally as well. It raises a relevant question about the spatial determinance of the voting behavior.

Besides completely theoretical conclusions, research of the spatial determinance of the voting behavior could give a new insight about the electoral fraud detection. All of the existing methods are based on the assumption that if the electoral data is manipulated, statistical analysis might reflect the interference by disclosing certain anomalies. For example, application of Benford's Law to polling returns (Mebane, 2009, Deckert et al., 2009) and last digit testing (Beber and

Scacco, 2008) refer to the stability in distribution of the digits in real datasets. If the distribution is different, it is the evidence of manipulation. These methods are not very sensitive, and they were proven as non-effective for the electoral data by Deckert, Myagkov and Ordeshook: “Deviations from either the first or second digit version of that law [Benford’s Law] can arise regardless of whether an election is free and fair. In fact, fraud can move data in the direction of satisfying that law and thereby occasion wholly erroneous conclusions.” (Deckert et al, 2009). Another method, a more important one, is voter turnout and party share regression (Myagkov et al, 2009, Mebane and Kalinin, 2009, Klimek et al, 2012, Sonin, 2012). Researchers state that there is no or a very week correlation between the voter turnout and the winner party share in old democracies where the elections are considered being fair, while in developing democracies like Russia and Uganda such correlation can be observed clearly. Summarizing, we have to stress that none of the abovementioned methods deal with geographic data.

Nevertheless, we could find some papers dealing with the geographical context of the electoral fraud. We could mention the work Skye S. Christensen, who tried to explain the level of fraud in Afghanistan, Kazakhstan and Sierra-Leone by linking the modeled levels of the electoral fraud with population density, natural resources distribution and security events (Christensen, 2011). Although some valuable conclusions were made, the analysis was done on a very large scale (second-order administrative division) and it could not provide a detailed picture. Another remarkable paper dedicated to detecting the electoral fraud was written by J. Chen. He analyzed the distribution of the financial aid after 2004 Florida hurricane by Bush administration and its impact on the changes in turnout and vote shares on the electoral district level (Chen, 2008). Finally, he stated that Bush administration had concentrated the aid on core Republican districts, increasing the voter turnout and the Republican share, and claimed that votes were bought by using public funds. Anyway, the mentioned studies are contextual (i.e. population vs. fraud, aid vs. voting), and the question of detecting the fraud by analyzing the spatial patterns of the electoral data is still open. Logically, it could be expected that if there is some

spatial regularity in the voting patterns, spatial irregularities could point on the fraud.

Working on geographical analysis of the electoral data, we have to consider one of the fundamental problems of spatial statistics: so-called Modifiable Areal Unit Problem (MAUP). Per ESRI GIS dictionary, it is "a challenge that occurs during the spatial analysis of aggregated data in which the results differ when the same analysis is applied to the same data, but different aggregation schemes are used. For example, analysis using data aggregated by county will differ from analysis using data aggregated by census tract." This issue was described by S. Openshaw, 1983, who stated that "the areal units (zonal objects) used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating." An example of such statistical bias can be seen on Figure 1 below. It is clear that grouping of the observations in different ways can give absolutely different results:

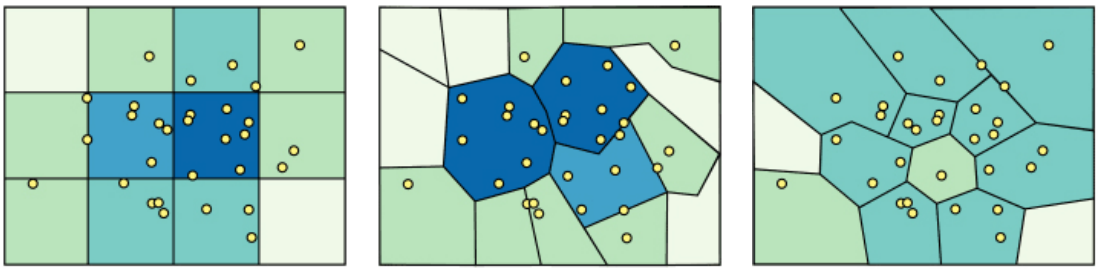


Figure 1. Modifiable Aerial Unit Problem.

Of course, given a set of polling divisions we can do nothing to change the division schema, but this is a problem not only in case of transforming the point data to the polygonal data. Also, when we group the polygons in one or another manner at higher levels of aggregation, we can get different results. An explaining example can be seen below:

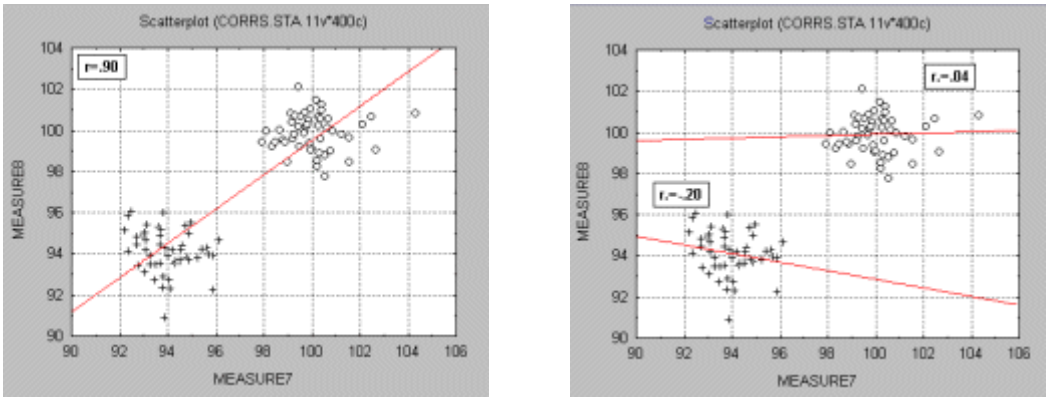


Figure 2. Aggregation problem.

When we group all the demonstrated observations into a single group there is a strong correlation pattern, but when we break them into two groups the result changes dramatically. Most of the researchers analyze the elections results at the level of the entire country or at the first level of administrative divisions (i.e. Canadian provinces). Meanwhile, aggregation of the data at the level of the administrative districts or cities could give unexpected results. Thus, we strongly believe all methods that are related to geographic data have to be tested at different aggregation levels.

1.2 Objectives

While the old democracies have made a substantial progress in developing free, fair and transparent election process which results are considered to be legitimate in most cases, the electoral outcomes in some of the young democracies are often quite questionable. For establishing any method of the electoral fraud detection, countries from the first category should be compared with ones from the second. We suggest that the initial step in this direction is to analyze the first category. Canada is a good example of such country, and we can easily access all the necessary data on polling division level (both tabular and geographic). This is why Canada is selected as the study area.

The main goal of the thesis is to investigate patterns and irregularities of Canadian electoral data at different geographic scales.

1.3 Assumptions

There is a set of assumptions regarding different parts of the research:

- Elections results are spatially determined;
- Electoral data has certain statistical and spatial characteristics, deviations from which would let the researcher expect the data manipulation;

1.4 General Methodology

Taking into account the pursued goal, we decided to use the following methods:

- Estimate the variability and distribution of the main variables by getting ranges, interquartile ranges, etc. and building plots and histograms,
- Evaluate the bivariate distribution of the voter turnout and party shares by building density scatterplots, convex hulls, etc.;
- Explore correlation between the voter turnout and party shares by getting correlation coefficients;
- Define spatial autocorrelation for the main variables by calculating Moran's Index;
- Perform cluster and outlier analysis by calculating Local Moran's Index for the main variables;
- Work on multivariate analysis of the data by using hierarchical clustering algorithms and estimating spatial distribution of the classes;
- Perform simulation studies to model the electoral fraud and repeat the abovementioned methods to estimate how the interference is reflected in the results.

1.5 Dissertation Organization

The thesis is organized accordingly to the selected methodology. In the second chapter, we described the data and methods of data access. The third chapter is dedicated to the exploratory analysis: distribution, variability and correlation analysis. The chapter ends with a simulation study where we modelled a

real-life situation of the electoral fraud and discussed how the exploratory analysis reveals the data manipulation. The fourth chapter is related to the spatial analysis. It is dealing with spatial autocorrelation and multivariate spatial analysis, and also ends with a simulation study. There is one annex which contains an example of a function written in R. Complete set of code and graphs set can be found in a digital attachment (DVD), along with a PostgreSQL database backup.

2 DATA DESCRIPTION

2.1 How are the elections organized

Since the electoral data describes a real-world process of elections, before making any research it is necessary to understand how are the elections organized.

Representation in the Canadian House of Commons is based on electoral districts, also known as constituencies or ridings ("electoral districts" further). Each electoral district elects one member to the House of Commons, and the number of electoral districts is established through a formula set out in the Constitution. Their boundaries are designed in a way that they contain similar amounts of people. This is done to represent of people's political preferences in the government in an equal way.

In 2011, there were 308 electoral districts in Canada. Their boundaries can be seen of Figure 3, within Canadian provinces:

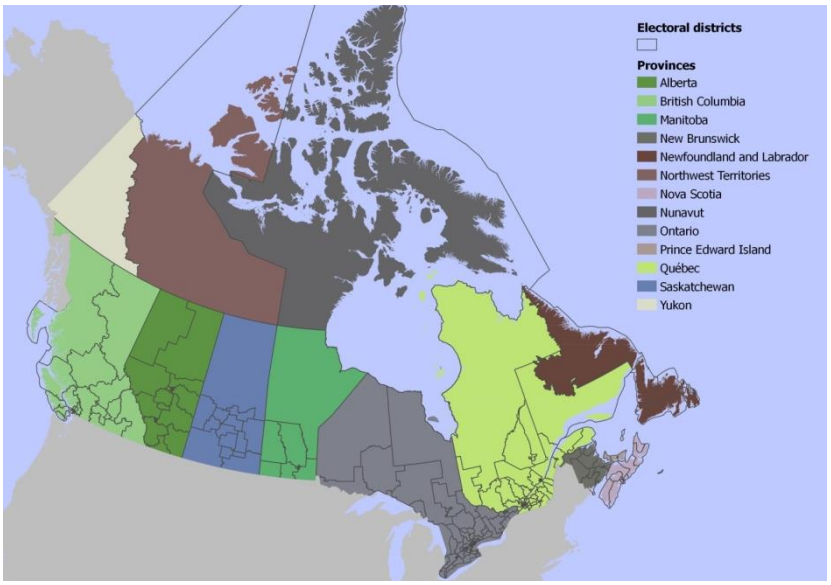


Figure 3. Electoral districts (Canada, 2011).

Some of the provinces, like Northwest Territories, include just one electoral district due to their low population, while heavily inhabited provinces, like Quebec, contain many of them. Each electoral district is divided into a set of polling divisions, again in accordance with the amount of population. Polling divisions can be seen on Figure 4:

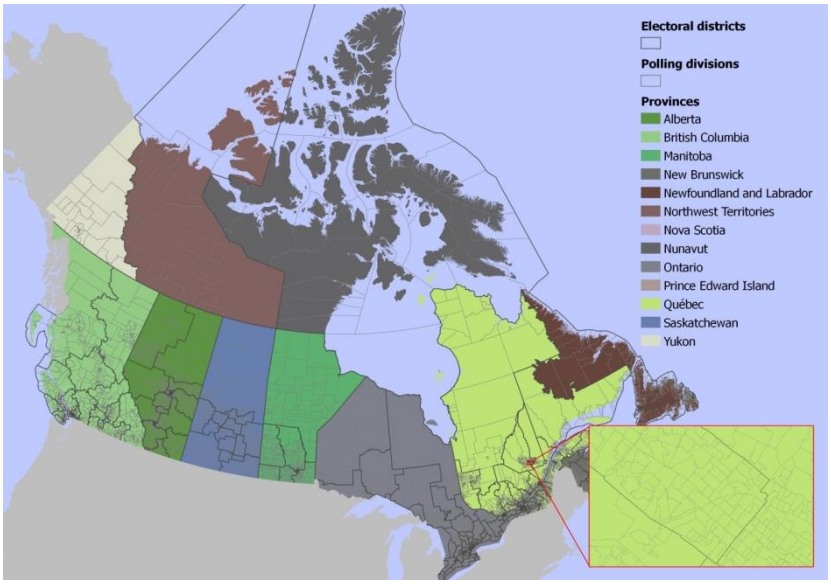


Figure 4. Electoral districts and polling divisions (Canada, 2011).

Each polling division has a certain number of citizens that live within its area and are eligible to vote, being older than 18. In most cases, this number is between 200 and 600:

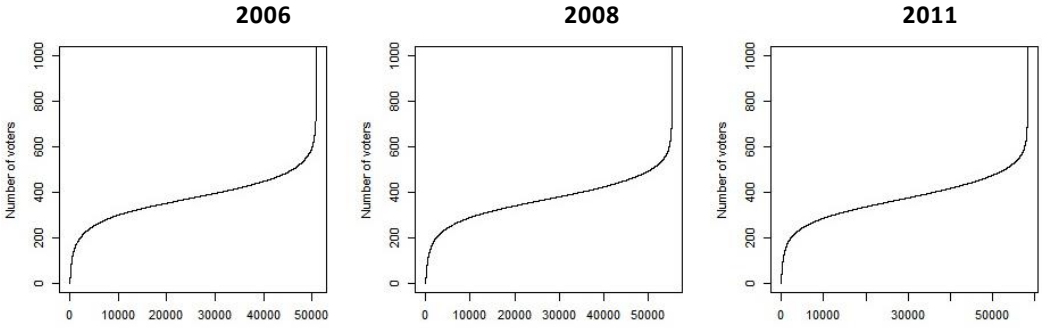


Figure 5. Number of voters in polling divisions of Canada.

In Canada, there is no obligatory participation in any kind of elections, so some of these people participate and some do not. Thus, the number of possible and actual voters on each station is different. Ratio between these two numbers is called voter turnout and is represented as a percentage between 0 (no actual electors) and 100 (when everybody participated). Each electoral district has its own set of candidates. People vote for different candidates which are associated with their political parties. About 20 parties have participated in General elections in 2010. Among these parties, there are three main parties: Conservative, New Democratic and Liberal, which were represented by their candidates in all ridings.

Also there is one strong regional party: Bloc Québécois, which is represented in French-speaking province Quebec. All the rest parties are minor parties represented in a small number of federal districts, except Green Party which had 304 candidates but was elected only in one electoral district.

The winning candidate has more votes within the electoral district than any other candidate. This is called “first-past-the-post” election, or “winner-takes-all”, or “simple plurality”. Party who gained majority of the chairs in the Parliament is called “Government”, the second is the “Official Opposition”, and there are “Third”, “Fourth” and “Fifth” parties. The Government was formed by Conservatives from 2006 to 2011, while Liberal party was the Official opposition in 2006 and 2008, replaced by New Democratic party in 2011. In that year, Liberal party became the third party, replacing Bloc Quebecois which was there in 2006 and 2008.

2.2 Elections results data

Elections Canada is an organization responsible for conducting the federal elections. All data related to elections is available on their website. Election results are published for each polling division since 2000. At the same time, the representation format was changed throughout the years. For example, in 2004 we can find only the vote count for each candidate called by name (without related political party), which makes the analysis of party support impossible. This format is called "pollbypoll", and its structure is described in Table 1 (Annex 1). Since 2006, Elections Canada published CSV files with political affiliation of candidates included. The new format is called "pollresults". Its structure is described in Table 2 (Annex 2). Since 2006, data is published as two sets of CSV files, one per each electoral district in both formats. It was decided to use only “pollresults” tables because they contain all the necessary information. Tables for each election were created in PostgreSQL database, and a Python script was written to generate valid SQL scripts for data import. The script retrieved all file names in the folder and pasted them into COPY statements. All the data was imported to text columns. This is a default setting in PostgreSQL to avoid misinterpretations and errors during the import, importing data “as is”.

The next step was to get data into a format appropriate for the analysis and visualization. Only 5 parties won at least in 1 electoral district, so it was decided to summarize the data only for these parties, showing the total result of the rest parties in a single column. The structure of the summarized data can be found in Table 4 (Annex 1). The obtained structure described is more understandable and more suitable for the analysis, since the vote count is assigned to political parties and not to the individual politicians which are different for each electoral district. It contains both absolute (vote count) and relative (percentage) values. This table is the final attribute table holding the election results which will be used for the future analysis.

2.3 Geographic features

Geographic locations of the polling stations and/or polling divisions are necessary to perform spatial analysis of the electoral data. Since we have the necessary electoral data for 2006, 2008 and 2011 elections, we have to get related geographic information for these years. Electoral district and polling division boundaries for General Elections are available for downloading at the website of Elections Canada. They are represented by polygon datasets in common geospatial formats. The attribute structure of polling division data is described in Table 6 of Annex 1.

Since the electoral district number is assigned to each polling division, there is no need to operate with electoral district geometries during the analysis. We already have the necessary identifier in the table. Electoral district geometries will be used just for mapping purposes. For analysis, only polling division geometries will be used.

In addition, there are datasets containing point locations of mobile polling stations and single-building electoral divisions like hospitals, etc.

All the described shapefiles for 2006, 2008 and 2011 elections were imported into PostGIS-enabled PostgreSQL database with “PostGIS 2.0 Shapefile and DBF Loader Exporter” and named by using the following naming convention:

“country code”_”year”.”dataset abbreviation”. Thus, the tables were named like this: “ca_2011.ed” (electoral districts), “ca_2011.pd” (polling divisions) and “ca_2011.ps” (polling stations). Each combination of the country code and the year represents a database schema in this case. This is done for convenience, avoiding confusingly large number of tables in a single database schema.

In the electoral districts dataset, some districts were represented by several polygons, and these polygons were joined by SQL script. To enable the analysis which requires point geometries instead of polygonal, we derived point geometries from polygons and placed them in separate column. For doing this, we utilized PostGIS function ST_Pointonsurface instead of ST_Centroid, since it creates point that lies inside polygon even if it has a concave shape (see Figure):

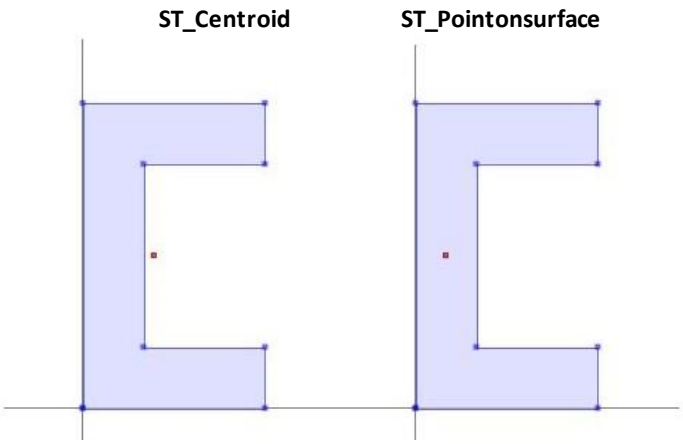


Figure 6. ST_Centroid and ST_PointOnSurface functions in PostGIS.

One of the concerns about the electoral districts is that their boundaries might not match the administrative ones. If not, it can be useful to perform the analysis in administrative division context as well. To do this, we should find the relationship between electoral and administrative division of Canada. So, what is the administrative structure of Canada?

The first level of Canadian administrative division is Provinces and Territories, and there are 13 of them. As already demonstrated on Figure 1 above, boundaries of the electoral districts perfectly match with provincial boundaries, and it is clear which province the electoral district belongs to.

The second level of Canadian administrative division is administrative district level. There are 293 administrative districts in Canada. This number is quite close to the number of the electoral districts (308). When overlaid, electoral district boundaries look quite arbitrary (see Figure 4 below). The administrative boundaries are likely to be less artificial than electoral boundaries because the first ones take into account the historical and geographic differences, while the second ones are designed to contain approximately the same amount of population. This is why the electoral behavior analysis in the administrative context could give us different results.

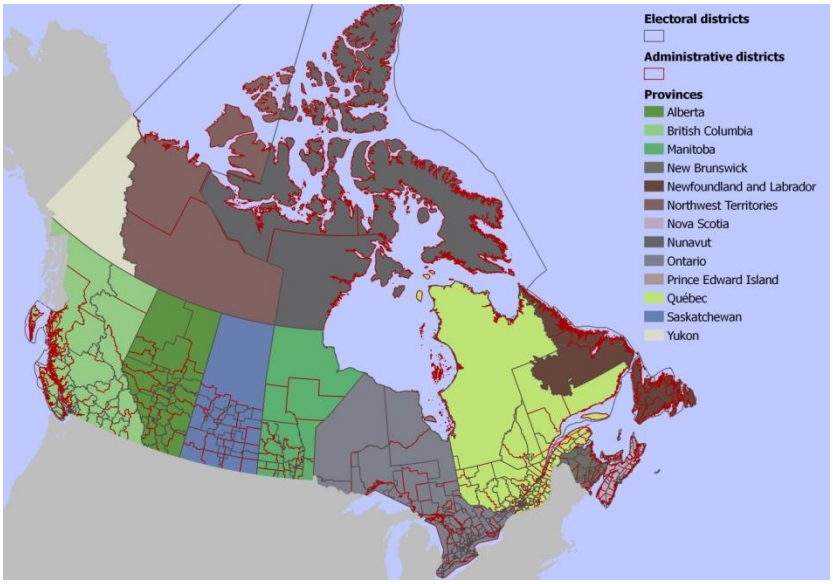


Figure 7. Administrative and electoral districts (Canada, 2011).

The third level is the municipal level. At this level, there are 5589 territory units. Municipal division is again quite different from electoral division, as can be seen on Figure 5. Municipalities vary in population greatly, the urban ones contain much more polling divisions than the rural ones. For example, on the insert in Figure 5 you can see municipality of Saskatoon city which partly intersects with 4 federal districts: #47002, #47009, #47010 and #47011 (thicker grey boundaries) and contains 411 polling divisions (thin light grey boundaries). Aggregation the polling data on the city level could give us an insight about how different is the electoral behavior within the cities. Analysis on the electoral districts level could not provide such information because they are not tied to specific cities.

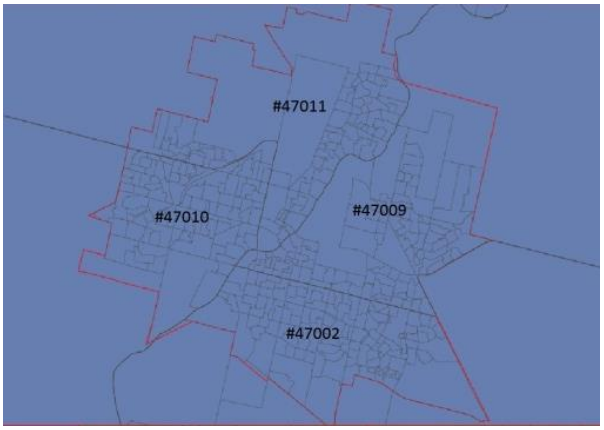


Figure 8. Urban municipalities and polling divisions (Canada, 2011).

We related polling divisions to each of the administrative units and stored these relations in a table for improving performance of spatial operations (it is faster to get precalculated relationships than performing spatial joins each time). The structure of the table is described in Table 7 (Annex 1).

2.4 Data access

Since we have very similar datasets for several years of elections and keep the geographic relationships in a single table, the code to query this data will look the same for all cases, with a couple of parameters. In this case, an efficient solution is a stored procedure (in PostgreSQL terminology they are called functions). The created functions will take the following parameters:

- level (integer) – specifies the aggregation level (see in the Table above)
- unit (varchar(150), default null) – territory unit filter. If not specified, returns all observations within the aggregation level

For example, getting all observations for Ontario province in 2011 is as easy as executing this line of code:

```
select * from ca_2011.getdata(1,"Ontario");
```

Another stored procedure was created to get the same data as the previous one does, and polling division geometries in addition:

```
select * from ca_2011.getgeodata(1,"Ontario");
```

This is very important because managing the SQL code in stored procedures helps to avoid incorrect code and SQL-related errors when programming in R.

Short summary of the data description looks like this:

- The main investigated *variables* are:
    - Voter turnout
    - Conservative party share
    - Liberal party share
    - New Democratic party share
    - Bloc Quebecois share (only for Quebec province);
- Each polling division is an *observation* consisting of ID and a vector of variables;
- Each electoral district and Canadian province, administrative district, city and neighborhood is a *territory unit*;
- A set of territory units is an *aggregation level*. If we analyze the distribution of some variable at province level, it means that we describe this distribution for 13 sets of observations (1 for each province).
- All data is imported to a PostgreSQL database and the access to this data is enabled by the stored procedures (PostgreSQL functions).

3 EXPLORATORY ANALYSIS

Investigating patterns and irregularities, it is very important to start by doing the exploratory analysis. The character of distribution and variability of the main variables is a key information in this case. A set of exploratory procedures should be done at all aggregation levels. First we are going to estimate global distribution of the variables (then the population is the entire country's data). Then, we will look closer to the local distributions (distributions for different territories smaller than the country). The next thing to look at will be the ranges and standard deviations for our variables at different aggregation levels, and in the last subchapter we will work on correlation analysis.

3.1 Distribution and variability

Starting with distribution, it is important to understand how many observations does each territory unit contain. These amounts can be very different:

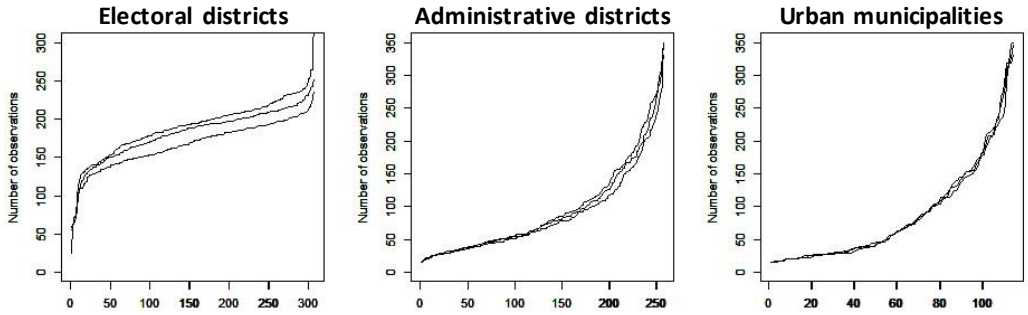
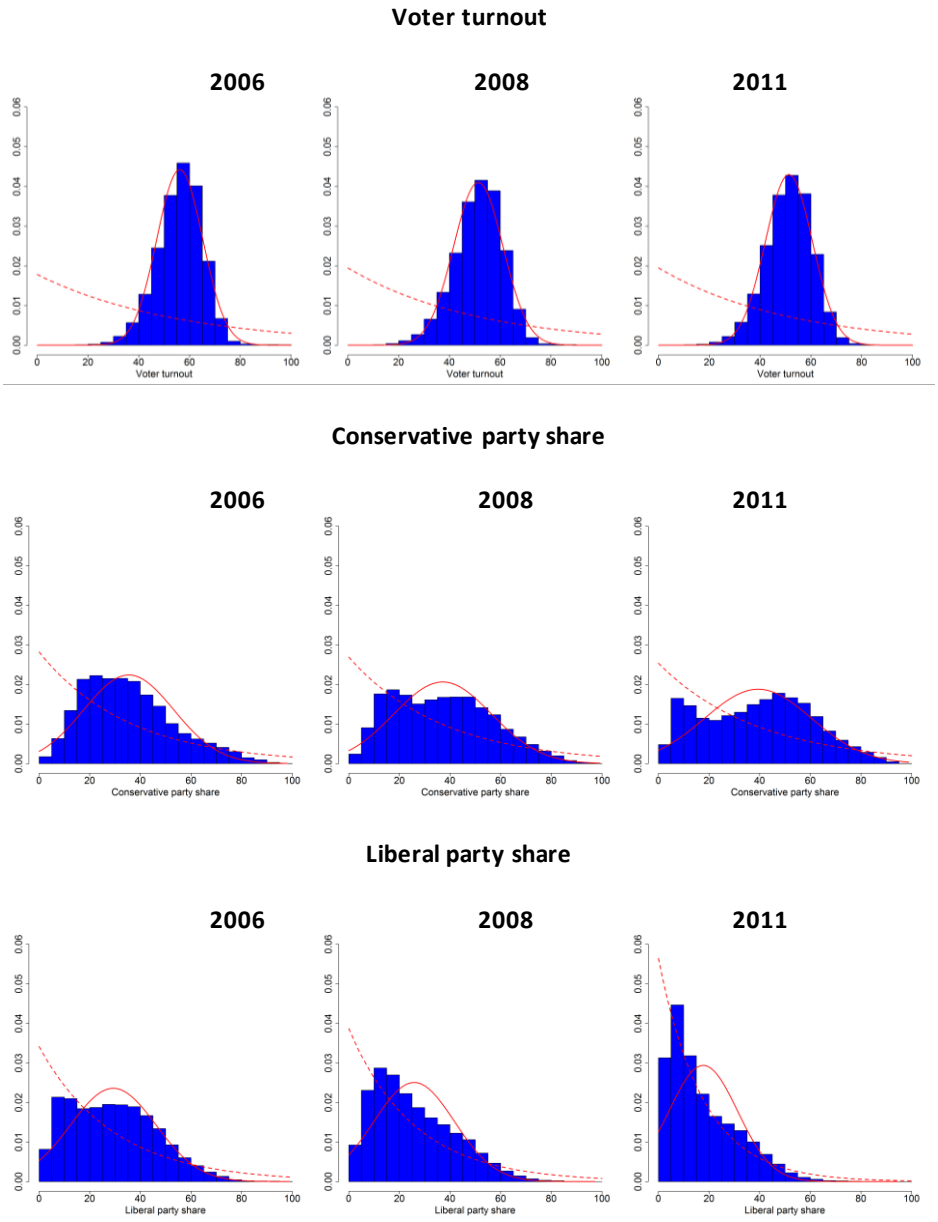


Figure 9. Number of polling division inside territory units at different aggregation levels (Canada, all years).

Electoral districts have the distribution close to normal, while administrative boundaries demonstrate the exponential growth in the amount of observations. Thus, any obtained indicators should be checked for positive correlation with the number of observations, because higher variation might be just the consequence of a larger number of the observations inside the units. Literally, more polling divisions are located within the unit, more different values we observe. On the other hand, a weaker dependency indicates meaningful results.

Distribution of the main variables for the entire Canada (except Bloc Quebecois which participated only in Quebec) can be observed on the histogram matrix on Figure 10. Solid red line is the normal distribution curve, while dashed red line reflects the modelled exponential distribution curve.



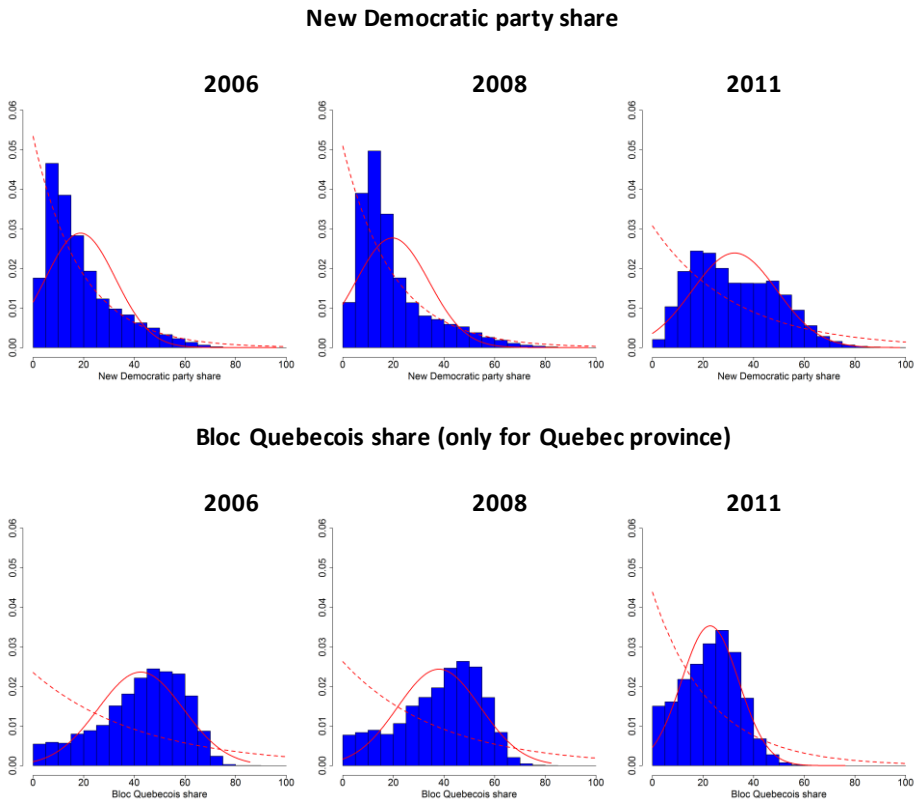


Figure 10. Global distribution of the main variables.

These histograms let us draw the following conclusions:

- Voter turnout demonstrates very evident normal distribution with most observations between 40 and 70% for all three years (a bit higher in 2006 but still close to 2008 and 2011). It means that the level of participation in parliamentary elections seems to be stable, regardless of changes in people's political preferences;
- There are no polling divisions with zero participation, i.e. having turnout value equal to 0%; and there is only a tiny fraction with complete participation, having turnout value >90%;
- 2011 was the most successful year for Conservative party. Its share had distribution close to normal in 2006 and since then it started to change its nature to bimodal with the main peak at 40-60%. Still, without a peak at lower values the distribution looks close to a normal one;
- Liberal party support was slowly decreasing from 2006 to 2011, and the distribution was changing from normal-like to a exponential-like during these years;

- New Democratic party (NDP) share has changed its distribution from exponential to having a plain top between 20 and 50%. Indeed, 2011 elections were the most successful for NDP, they became an Official Opposition in that year;
- Bloc Quebecois had a strong support in Quebec in 2006 and 2008, but in 2011 their support has decreased dramatically (probably, in favor to NDP). At the same time, the distribution was close to normal in all years;
- Parties with weaker support tend to have exponential distribution, while parties with stronger support usually demonstrate distribution close to normal.
- All the variables demonstrate the stability of change, gradually moving towards a one direction through time (3 years might not enough to make such conclusions).

Though we have made a set of important conclusions, these histograms provide just a general picture of voting patterns. They indicate the "global" distribution for the entire country, while there are many local distributions in different geographic regions that can be very different from each other. As a rough example, a bipolar "global" distribution can be a result of two normal local distributions. So, we have to analyze the local patterns.

To estimate local distributions, we need to compute the number of samples, min, max, range, mean and standard deviation for each variable in each territory unit at all aggregation levels. If we presented each local distribution on a separate histogram, it would be hard to compare them, since we will have hundreds of them. Instead, we decided to create the representation which would have all local distributions as semitransparent histograms drawn on the same canvas. Each histogram's opacity was derived from the total number of histograms. For example, if we plot some variable at the electoral district level (n=308), the opacity will be 2 out of 255. Such overlay indicates both when many histograms share the same area and when there is an uncommon distribution nature. This was done for real data (the left column) and for normal and uniform simulations on the basis of that data (columns 2 and 3). To simulate normal distribution, we got the number of observations, mean and standard deviation from the real data and create variable

vector in the basis of these values. To simulate random distribution, we created random vectors for each unit, taking the number of samples and the minimum and maximum value of the variable. Also, we have added a histogram of p-values from Shapiro-Wilk test for each territory unit (the right column). When p-value is less than 0.05 it means that with 95% chance the distribution is normal.

The procedure described above was done for each of the main variables at the level of provinces, administrative districts, urban municipalities and electoral districts for each year. An example of Conservative party share in 2011 can be seen below:

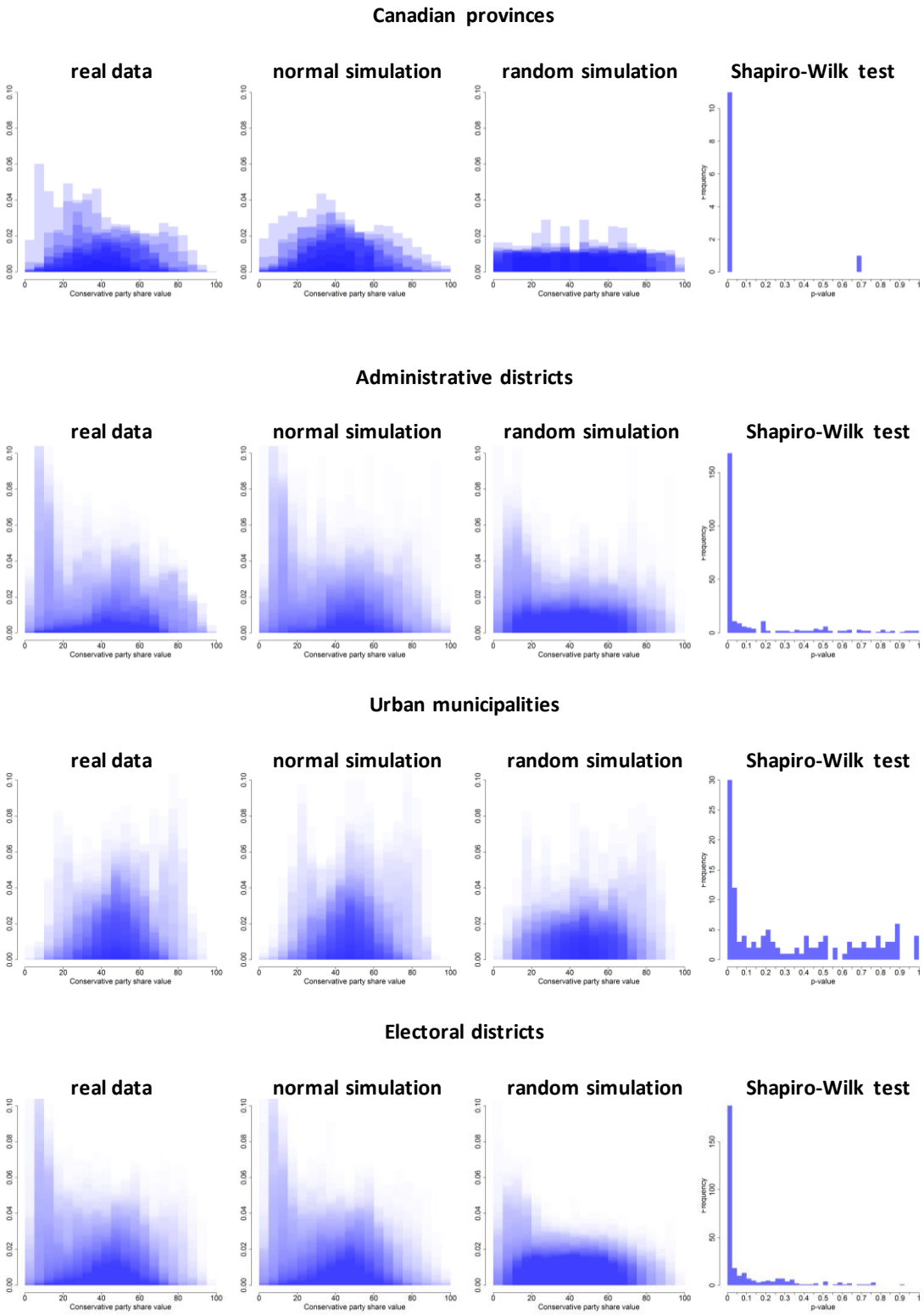


Figure 11. Local distributions of Conservative party share at main aggregation levels (Canada, 2011).

The graphs tell us the following:

- Regardless of how does the global distribution look (normal, exponential, bimodal, etc.) and regardless of how does it change through time, local distribution histograms for real data look very similar to normal simulation histograms. On the contrary, random simulation looks different;
- Shapiro-Wilk tests confirm normality of most of the local distributions, but at the same time there are many p-values higher than 0.05. For example, for Conservative party in 2011 they are 100 out of 306. At urban municipality level it is even more (86 out of 128), and this is related to more complicated political landscape in the cities;
- There is no difference in the normality of distribution for all variables. Voter turnout and any of the party shares have the same pattern of distribution: around 2/3 of the local distributions is normal and 1/3 is not normal, and vise versa for urban municipalities;
- Higher p-values for different variables are usually represented in different territory units, i.e. there is a very small number of units which have p-values >0.05 for all variables.

Conclusions confirm the assumption that the analysis of local patterns can give a lot of additional information to the global distribution analysis.

The next step is to analyze variability of the main variables at different aggregation levels. The most basic indicator of variability is the range which the difference between maximum and minimum values. Interquartile range (IQR), which is the range between the upper and lower quartiles (50% of values which lie around the mean), can give more meaningful information. As we already mentioned, it is very important to look at the significance of results by checking for a correlation between the variability indicator and the number of observations in each territory unit. In this case, scatterplots are the right method. Finally, 12 groups of graphs were produced (4 variables * 3 years). They are available in Annex II. Here we can see an example of one group for Conservative party share in 2011:

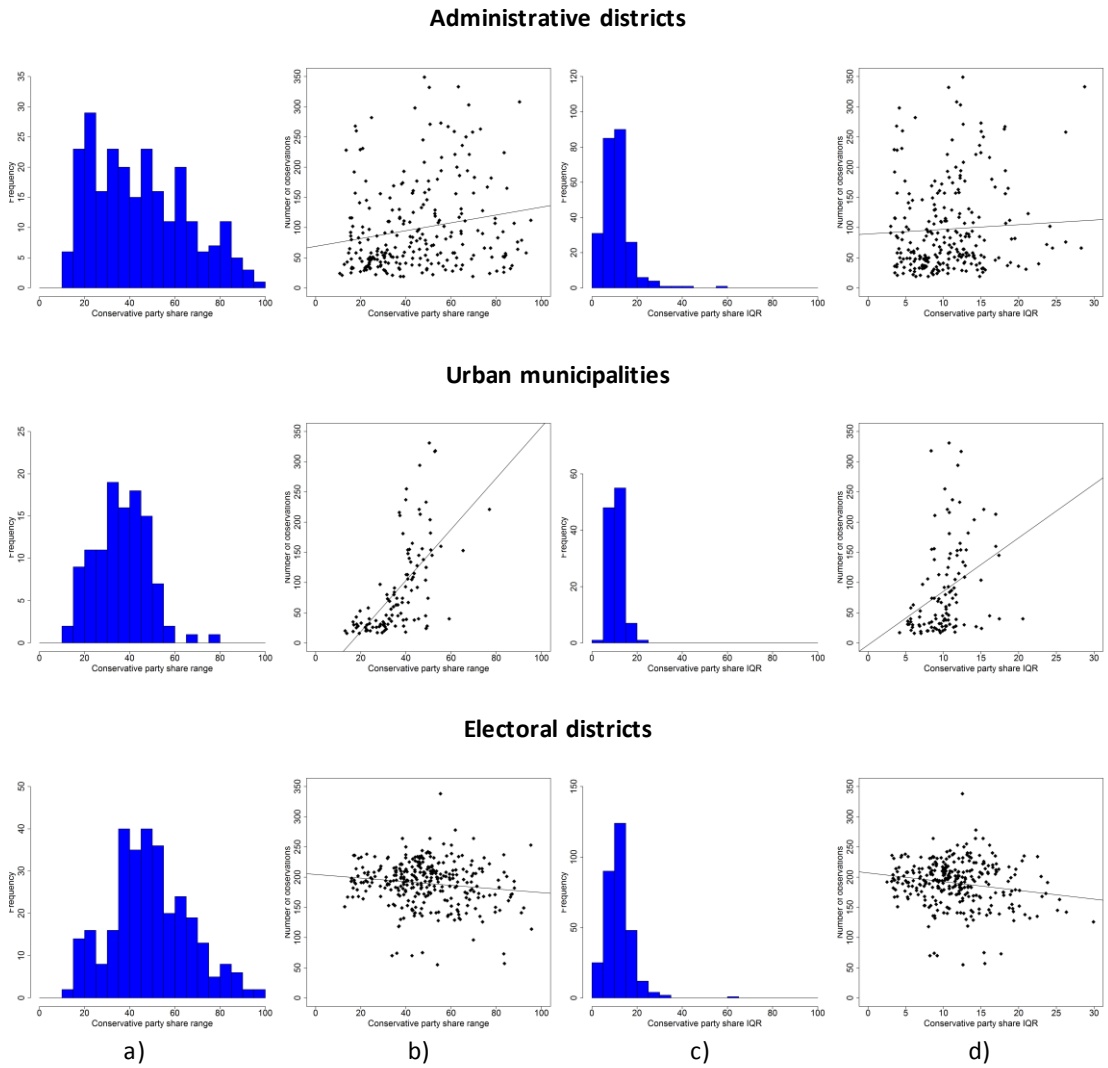


Figure 12. Simple and interquartile ranges for Conservative party share at main aggregation levels (Canada, 2011).

We decided to build graphs only for these aggregation levels since at province level variability is too high and not meaningful. From the entire set of graphs for all years, the observations are the following:

- Natural variability of the main variables is higher than expected. Units with simple range <10% and IQR <5% for any variable are extremely rare;
- For three aggregations levels, histograms look differently. There is a very weak or no correlation between variability indicators and the number of observations for administrative and electoral districts, while for urban municipalities it is very strong. Polling divisions belonging to heavily populated districts can have both similar and different results. Larger cities have more differentiated electoral behavior if their inhabitants, since they usually have more

social and economic contrast than the small ones. So, it was decided to exclude urban municipalities from the variability analysis. Large cities could be analyzed as the sets of neighborhoods, which is out of scope of this work;

- Voter turnout range varies from 20 to 80%, with peak at 35-50%, and its IQR varies from 5 to 15% (with is a couple of exceptions).
- Simple range for all parties can be very different, from 20 to 80%, while IQR mostly falls into the interval between 5 and 15% (generally, between 0 and 20%), with a couple of exceptions. It means that the electoral districts include the groups of polling divisions sharing the same behavior. These groups were not revealed by the simple range, while IQR helps to indicate them;
- Variability indicators of Conservative party and especially New Democratic party share increase from 2006 to 2011, with the overall growth of these parties' support. On the other hand, variability of Liberal party share decreases from 2006 to 2011, while its share had been decreasing. This tendency is reflected especially in IQR histograms. Probably, the larger is the overall party share, the larger the variability.

The next step is to calculate standard deviations. Standard deviation (SD) of the variable is an average difference between its values and its observed mean. It helps to measure the level of dispersion of the variable. Higher is the standard devotion, more dispersed the values are. SD also helps to find the outliers. These are the values which lie outside the symmetric intervals (-2*SD,2*SD) and (-3*SD,3*SD) from the mean. Finally, we produced the graph matrix with 4 columns:

a) SD distribution;

b) scatterplot of SD and the number of observations;

c) scatterplot of the number of outliers and the number of observations (empty circles for (-2*SD,2*SD) and solid rhombi for (-3*SD,3*SD));

d) scatterplot of the percentage of outliers and the number of observations (the same).

Again, this is done for the administrative districts, urban municipalities and electoral districts. The described graph matrix for Conservative party share in 2011 is shown on Figure 13 below:

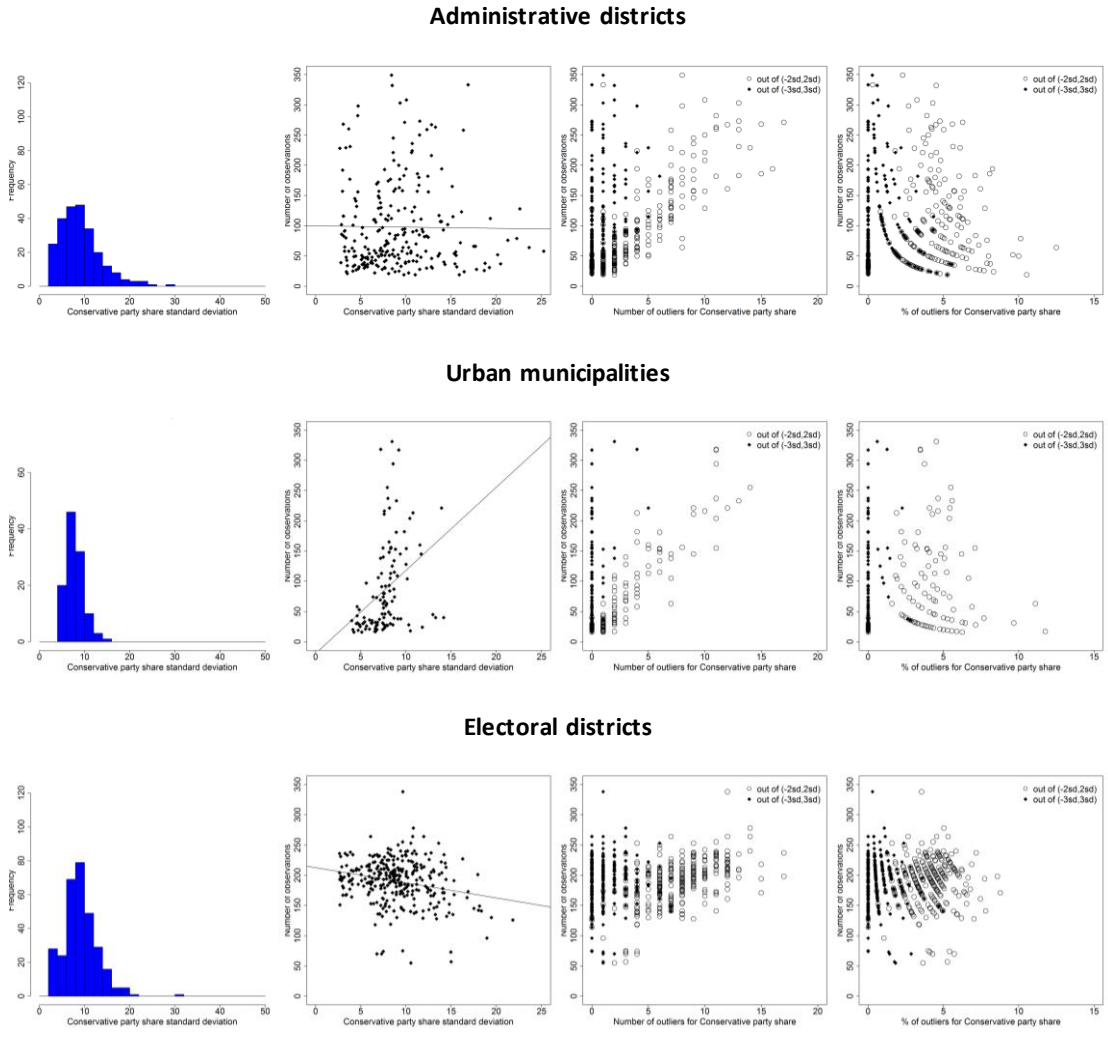


Figure 13. Standard deviations and outliers for Conservative party share at main aggregation levels (Canada, 2011).

Other graphs can be seen in Appendix 2. From what we can see on the obtained graphs, we can tell the following:

- Standard deviations of voter turnout are stable throughout the years and they are generally smaller than for party shares, almost all of them are between 6 and 10;
- Standard deviation of the party shares depends the overall party share, the same like the ranges are;

- Correlation between standard deviation and the number of observations is the same as for range statistics, so we exclude urban municipalities from the analysis;
- For party shares, the amount of the outliers is the largest in the electoral districts, varying from 5 to 15 for an interval (-2*SD,+2*SD) and less than 5, mostly 0, for an interval (-3*SD,+3*SD);
- There in an exponential dependency between the percentage of outliers and the number of observations. It means that when the number of observations increases, the amount of outliers remains stable.

Also if we have 3 years we could plot the percentage of outliers on a 3D scatterplot, where each axis stands for a year of elections. Thus, if the amount of outliers is stable for each territory unit, the point cloud will be oriented diagonally, from the coordinate zero point towards the maximum values. On Figure 14 below, there are such cubes for each of the main variables:

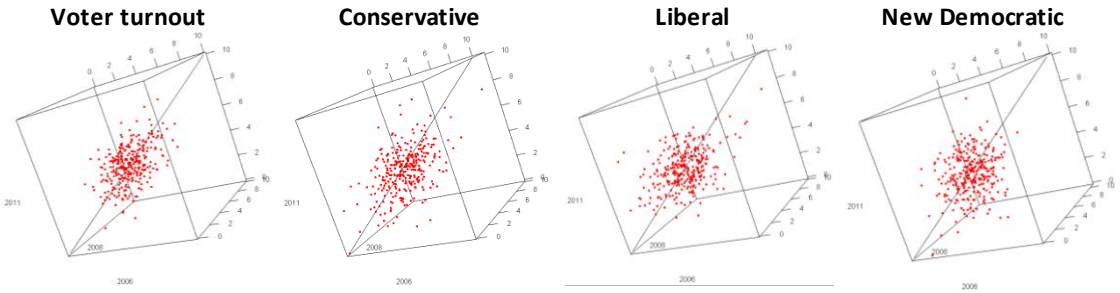


Figure 14. 3D plots showing the amount of outliers in 2006, 2008 and 2011.

Voter turnout and Conservative party share clouds are oriented as described above, while Liberal and New Democratic party clouds have more spherical nature. Still, they are located around the diagonal.

After all, the most important conclusions are:

- Voter turnout has a very strong normal pattern in global distribution and mostly in the local distribution;
- Winning and losing parties have their own characters of distribution and variability. For the winning party, global distribution is close to normal, and variability is higher. For the party which loses the elections, global distribution is

likely to be exponential, and variability is lower. There are intermediate stages of transition between these two conditions;

- Ranges and standard deviations in the electoral districts have their own distribution which has some stability and can be compared with the new data or a model.

3.2 Correlation between voter turnout and party shares

One of the key points of the exploratory analysis is the regression analysis. As discussed in the Theoretical background chapter, some authors state that high correlation between the voter turnout and the winning party share points to the electoral fraud (Myagkov et al, 2009, Mebane and Kalinin, 2009, Klimek et al, 2012, Sonin, 2012). In most cases, they investigate this correlation working only with the entire country without breaking the data into subsets for different geographic regions.

The first step to do for an overview is to build a correlation coefficient matrix for the entire dataset. It is shown below:

	Voter turnout	Conservative	Liberal	New Democratic
Voter turnout	1.00000000	0.07187695	-0.02489927	-0.1067010
Conservative	0.07187695	1.00000000	-0.22397221	-0.7158480
Liberal	-0.02489927	-0.22397221	1.00000000	-0.3356734
New Democratic	-0.10670100	-0.71584805	-0.33567338	1.0000000

Table 1. Matrix of correlation coefficients for the main variables (Canada, 2011).

It is obvious that correlation between different party shares is negative because these are the percentages from the entire amount of voters. The more votes one party gets, the less votes are left for the others. The brightest example is -0.7 between Conservative and New Democratic parties. This is somewhat natural, and in fact, the only relationship that is less natural is the relationship between voter turnout and party shares. It deserves a special investigation.

At the entire country level (without any aggregation) it is better to build a density scatterplot because there are too many observations for a typical scatterplot. We have created such scatterplots for the main parties, and in most cases they demonstrate the smooth bivariate distribution with a single hot spot in the center and the density falling towards its periphery. The only exclusion is the scatterplot for Conservative party share in 2011 (Figure 15a) which looks the same as one in Klimek et al, 2011:

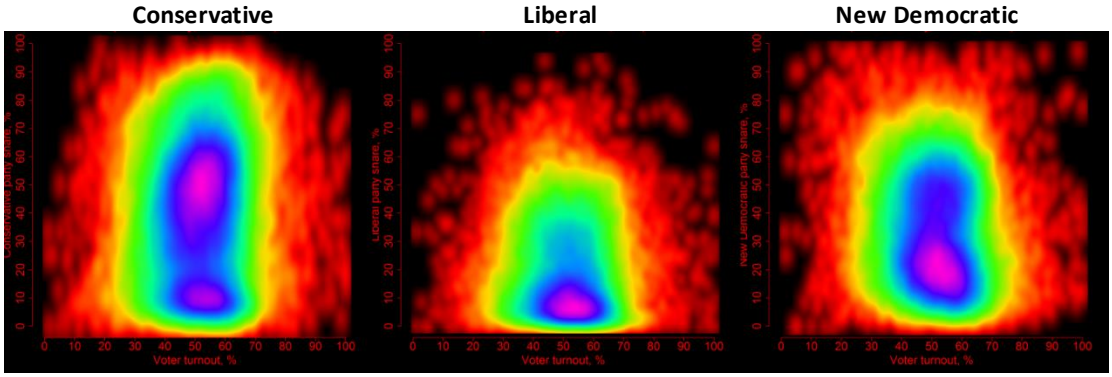


Figure 15. Voter turnout against party shares for all polling divisions (Canada, 2011).

Klimek et al state that a smaller area at the bottom stands for French Canada (Quebec province) and a larger area on top is for English Canada (all the rest Canadian provinces and territories). This assumption was checked by looking at province-level results which are published as well: "Looking at their results by province, they [Conservatives] tallied 16.5% of votes cast in Quebec but more than 40% of votes cast in 8 of the remaining 12 other provinces." (Klimek et al, 2011). This can be enough but since we have defined the relationship between the provinces and polling divisions, we could visualize this on the same scatterplot. To do this, we are plotting semi-transparent white points above the existing graph for the selected provinces: Ontario (the largest English-speaking province) and Quebec:

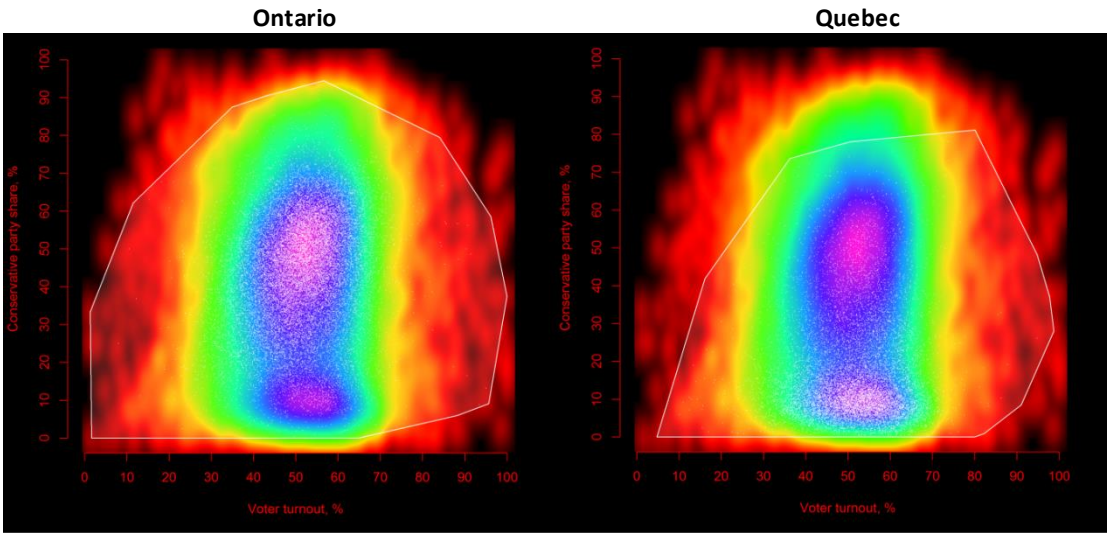


Figure 16. Voter turnout against Conservative party share at polling division level (entire country, Canada, 2011), combined with point clouds and convex hulls for selected Canadian provinces.

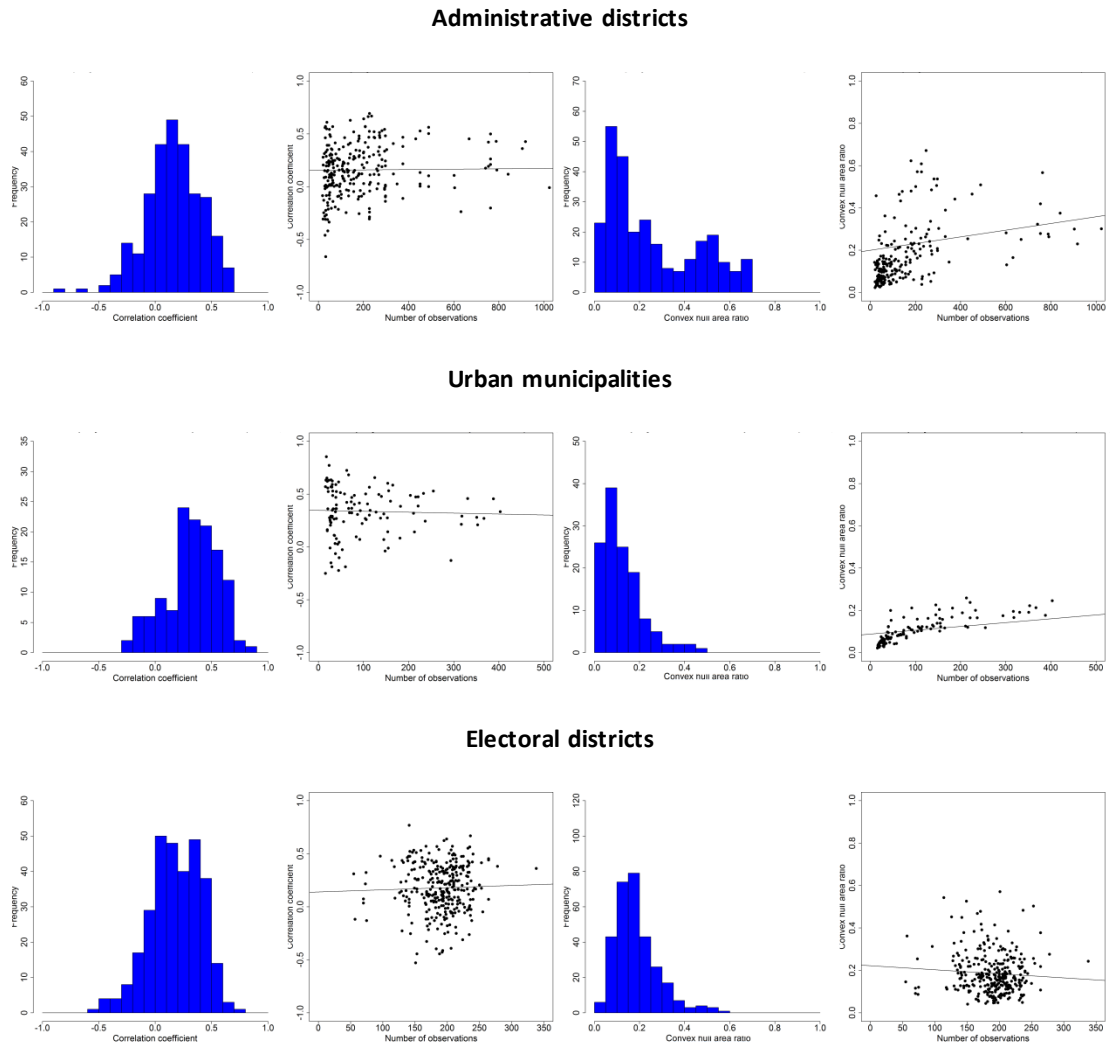
The plots above confirm the given statement. We can see that the areas of higher concentration of white points are located on respective hot spots of the density scatterplot. At the same time, they give more information: we can observe that even though the points are highly concentrated, there are the outliers which are very different from the main pattern. Convex hull shows the character of the local bivariate variance and how does it match with the global bivariate variance. This is a very important outcome because if we are looking for data irregularities to detect fraud, we should take the presence of such outliers into account. In Annex II, the remaining scatterplots can be found.

Using the same method, we could build the graph for any geographic unit, whether it is an urban municipality or an electoral district. Of course, it is hard to give estimation of each graph (i.e. 308 graphs for the electoral districts). Instead, we did the following:

- get correlation coefficient for each geographic unit and draw a histogram with their distribution, along with the plot for correlation coefficient and the number of observations;
- calculate the area of the convex hull for each territory unit and divide it by the area of the convex hull for the entire dataset. If the ratio is closer to one, the variability of voter turnout and selected party share combinations is close to that for the entire country. On the other hand, values closer to zero mean similar behavior within the unit.

Correlation coefficients for Conservative party share and voter turnout in 2011 look like shown below:

a) distribution of local correlation coefficients;
b) scatterplot of correlations coefficient and the number of observations;
c) distribution of convex hull area ratios;
d) scatterplot of convex hull area ratios and the number of observations.



a) b) c) d)

Figure 17. Distribution of the correlation coefficients for voter turnout and Conservative party share at main aggregation levels (Canada, 2011).

The remaining graphs are available in a digital annex. From all graphs, the outcomes are stated below:

- Though the global correlation coefficient might be very small, there can be large local coefficients. For example, the global coefficient for Conservative party share in 2011 is 0.07, while the local values for the electoral districts can go above 0.5;
- Correlation coefficients distribution changes between years but still each party has its own main range where most of the coefficients fall into:
    - Conservative party: -0.1 – 0.5,
    - Liberal party: -0.3 – 0.3,

- New Democratic party: -0.5 – 0.2;

- Unlike variability indices, correlation coefficients for turnout and party shares do not follow the overall party success or failure in time;
- As opposed to variability indices, there is no relationship between the correlation coefficients and the number of observations at any aggregation level. Least-squares equation lines on all graphs are mostly horizontal;
- Relationship between the convex hull area ratios and the number of observations is very weak. It means that variability of voter turnout and party shares does not depend of territory population. For example, New Brunswick province demonstrates the most similar results within itself, while Saskatchewan province, having the same population, has much higher variability;
- For the electoral districts, convex hull area ratio higher than 0.4 is extremely rare. This means that each electoral district has its own set of combinations of voter turnout and party shares but this set is always not as full as the entire country's set;
- We can see that convex hull area ratios follow the global party share. Better is the result of the elections for a party, more dispersed is the behavior within the territory unit, and vice versa.

Everything we did before was done on the aggregated data, i.e. polling divisions data grouped into subsets according to some geographies. The result was a set of indicators, like correlation coefficients, etc. At the same time, it is necessary to summarize data variables within to the same geographies, i.e. have 1 value for each territory unit. For instance, by dividing the total amount of participating voters by the total amount of possible electors for each unit we get the summarized voter turnouts. Doing the same with party shares, we can estimate their regression. An example of such summarization for Conservative party share at the electoral district level in 2011 can be seen below:

a) distribution of summarized voter turnout values;
b) distribution of summarized party share values;
c) scatterplot of A against B.

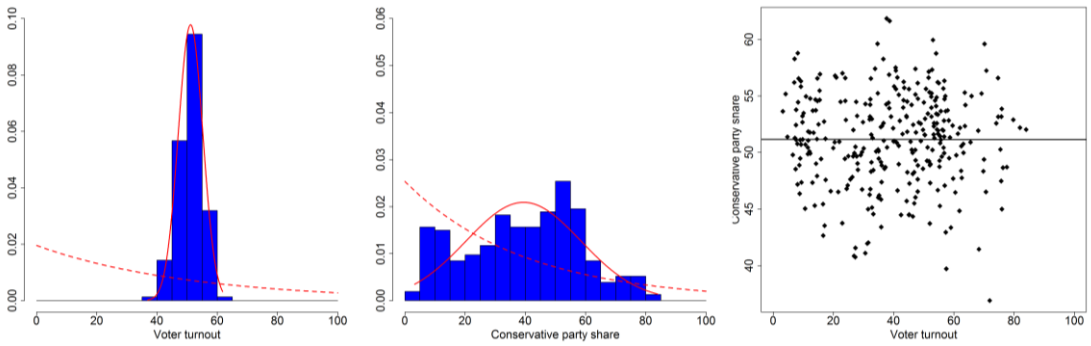


Figure 18. Correlation between summarized voter turnout
and Conservative party share (electoral districts, Canada, 2011).

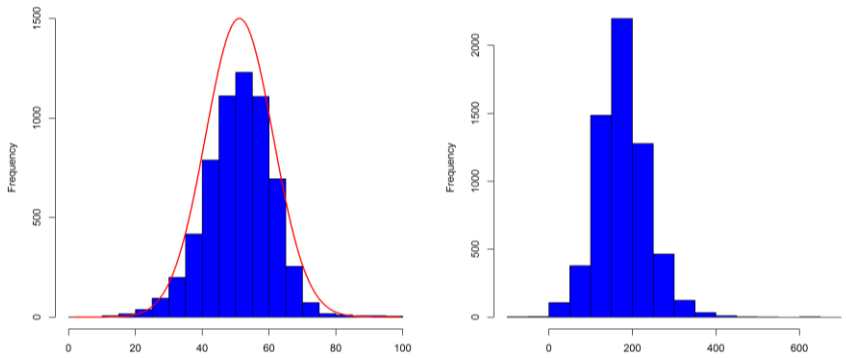
We can see that summarized data does not demonstrate any correlation patterns in any years for any variables.

As a conclusion, we can confirm that there is no expressed relationship between voter turnout and party shares. Though we observed local correlation in some of the territory units, the strongest pattern is the independence between the specified variables. Another important outcome is that the electoral districts are the best aggregation level for the study because they contain similar number of observations. When the number of observations is very different, i.e. there are very small and very big cities, it leads to a statistical bias in the analysis. Thus, using the administrative boundaries is possible for the exploratory analysis but it is not recommended in spatial analysis that is boundary-sensitive.

3.3 Electoral fraud modelling: a simulation study (I)

Doing the exploratory analysis, we have found a set of patterns. Our conclusions are valuable only if the detected patterns help to find out when the data is manipulated. The easiest way to check this is to model some data. In our case, we could change some of the results, for example by imitating the ballot stuffing. According to Uslegal.com dictionary, "Ballot stuffing is a type of electoral fraud whereby a person permitted only one vote submits multiple ballots.". Ballot stuffing elevates the share of some party, as well as the voter turnout. The first value increases because all stuffed ballots contain votes for a single party in favor to which the ballot stuffing is committed, and the second one grows since each ballot (even the stuffed one) is accounted as an actual voter.

There can be several scenarios to model. For example, if Liberal party support had been decreasing from 2006 to 2011, we can model ballot stuffing process for 2011 on the basis of 2008 results. For modelling, we selected 6099 out of 73862 polling divisions which belong to the electoral districts where Liberal party was elected in 2006, but was not elected in 2008, i.e. lost the chairs. This is around 8.25% from the total amount, so it can be a good number for performing the simulation. In these polling divisions, voter turnout has the normal distribution, as usual:



a) b)

Figure 19. Voter turnout (a) and empty ballot count (b) in polling divisions where Liberal party lost its chairs in 2008.

Turnout values mainly fall into the range between 40 and 60, and the amount of unused ballots is generally around 200 in each station. It looks sufficient for using

ballot stuffing techniques without the risk of overstuffing, when the turnout comes close to 100%. Also, we selected only those polling districts where the number of empty ballots is more than 50 to avoid too clear evidence of stuffing. The number of stuffed ballots was calculated as a random number between 50 and 75% of the empty ballot count in 2008. To clarify the process, we provided is step-by-step example:

- a given polling division has 300 electors and 200 voters in 2008, i.e. turnout 66.6% and the number of empty ballots is 100;
- between 50 and 75% (in this case, 60) extra ballots having Liberal party vote are going to be used in ballot stuffing in 2011;
- in 2011, the given polling district has 310 electors (its population has slightly increased) and 220 voters participate (a bit more than in the last year), i.e. real turnout is 71% and a real number of empty ballots is equal to 90;
- 60 ballots are stuffed in the ballot box, making the turnout increase to (220+60)/310=90%;
- Real share of Liberal party was 50 votes, or 22%, while after ballot stuffing it grew to 50+60=110, or (50+60)/(220+60)=39%.

Full results of the modelling can be checked by using the query from a digital annex (3.3 - fraud simulation). As soon as we had the manipulated data, we ran the analysis functions that we used before. The most important changes can be observed in voter turnout distribution: instead of a strong normal pattern observed for real data, we could see a new group of observations with higher turnout both for global and local distributions:

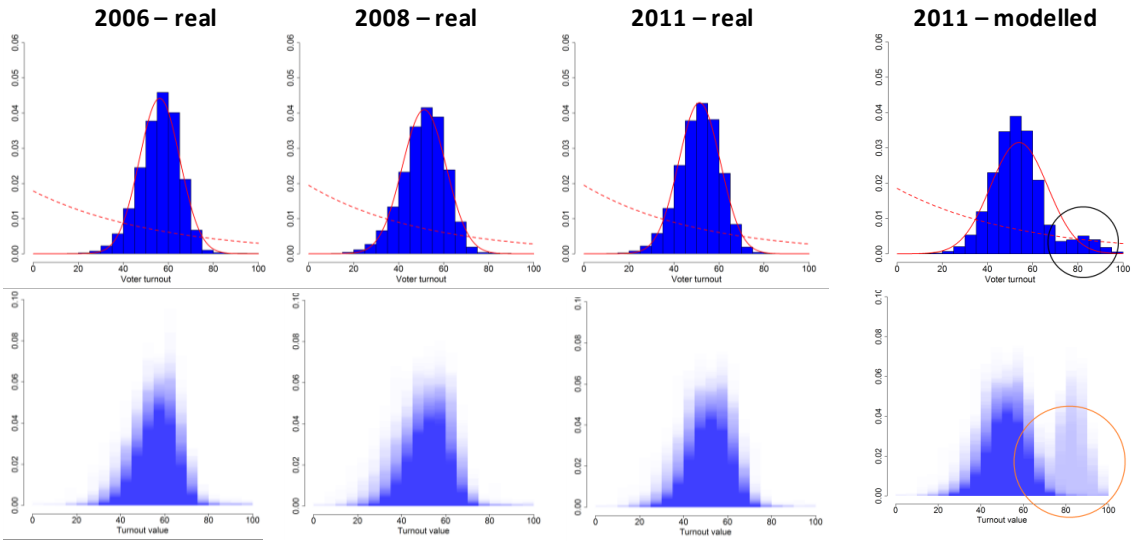


Figure 20. Distribution of voter turnout for real and modelled data.

Local distributions of party shares did not reveal a significant change, the same for variability indices, ranges and standard deviations which changed just slightly.

Another visible change can be noticed in density scatterplots of voter turnout and party shares where the new hot spots appeared (highlighted by white circles on Figure):

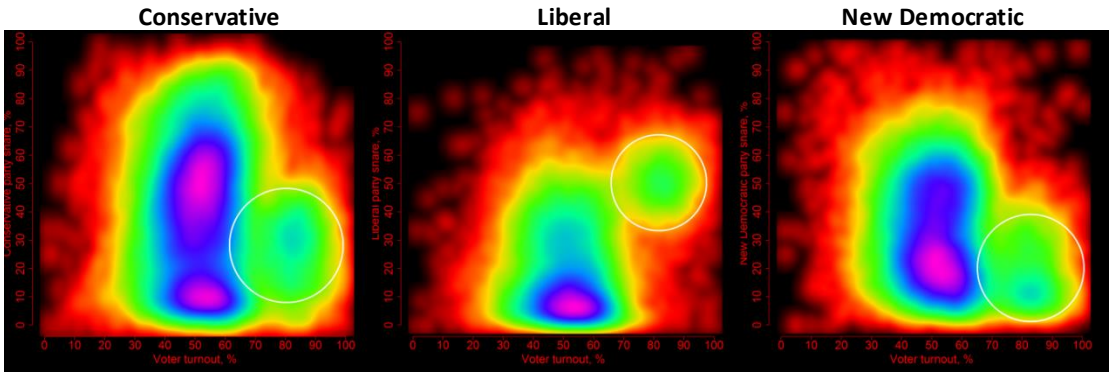


Figure 21. Density scatterplots for voter turnout and party shares (modelled data).

Anomalous hot spots tell that there are many polling divisions where bivariate distribution of voter turnout and party share deviates from the general pattern. It is also clear that these anomalies are in favor to Liberal party because its hot spot is the highest by y axis, which means larger party shares for higher turnout values. At the same time, these anomalies can not be explained by the geographic context, like it was with the real data. When we overlay our density scatterplots with

individual plots and convex hulls, we can not distinguish regions with predominant concentration of points around any of the new hot spots:

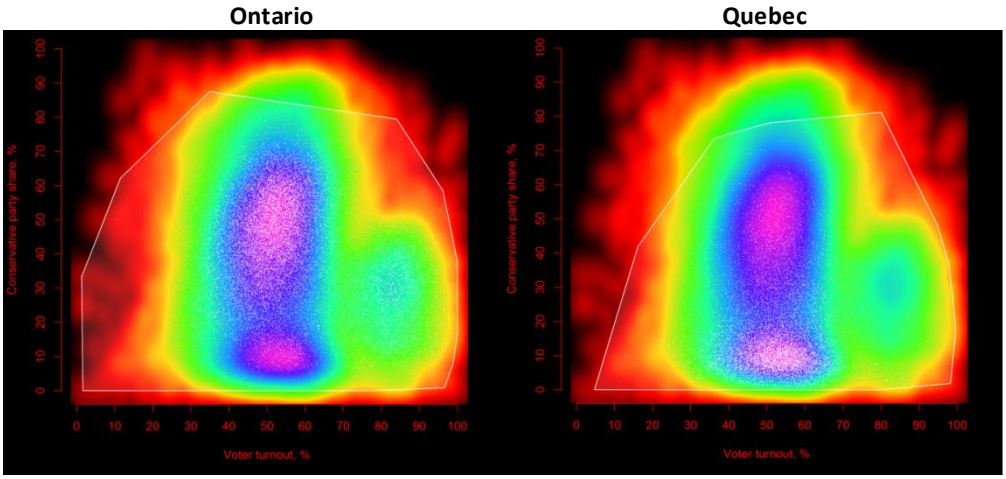


Figure 22. Density scatterplots for voter turnout and Conservative party share (modelled data).

Summarized values indicate the interference as well. On the histogram of summarized voter turnout distribution (Figure on the left) we can see the same artificial spike, while the exponential distribution of party share (in the middle) is broken. On the scatterplot we can see that values breaking the real data patterns belong to the same observations, as indicated by the outlying group.

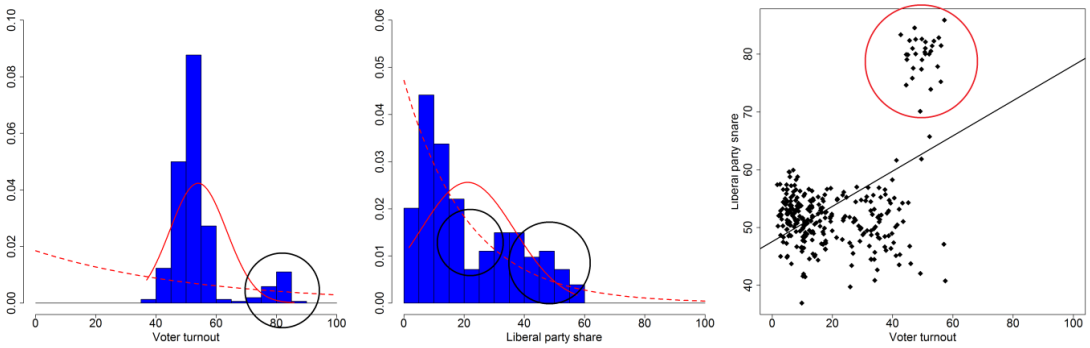


Figure 23. Correlation between summarized voter turnout and Liberal party share (modelled data, electoral districts, Canada, 2011).

Finally, we can say that the exploratory analysis techniques are good at indicating the electoral data manipulation when it happens compactly in a set of the electoral districts.

4 SPATIAL ANALYSIS

4.1 Spatial autocorrelation.

This subchapter is dedicated to discussion about the level of spatial autocorrelation of the electoral data. This is a critical point of the work because confirmed spatial autocorrelation is an evidence of data's geographic determinance.

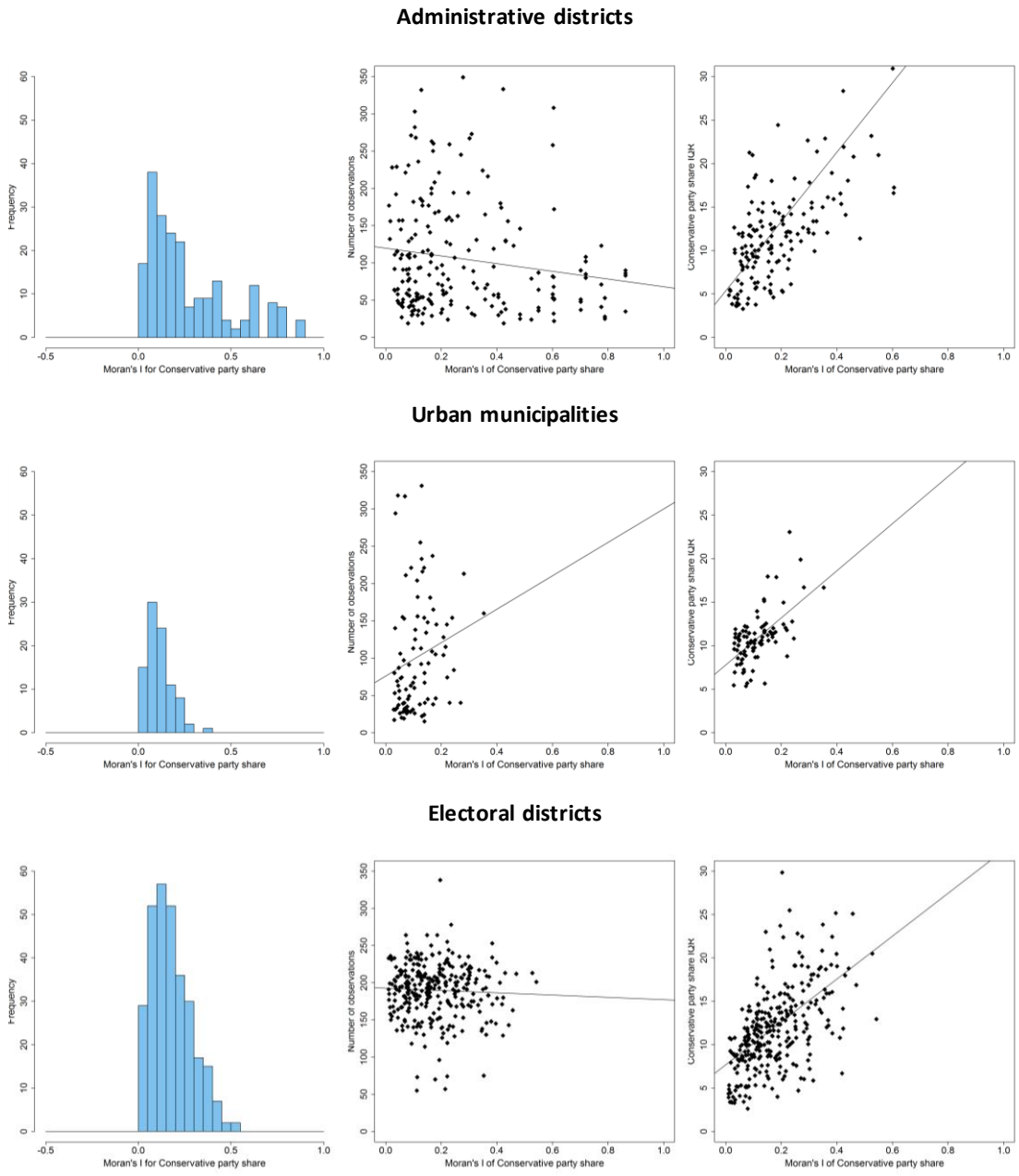
In the previous chapter we have found that there are some groups sharing the similar behavior in many territorial units. The next step of the analysis is to understand whether polling divisions belonging to these groups are geographically dispersed or they are located close to each other, forming groups, or clusters. As specified at the beginning of the work, we expect them to form groups. When observations with similar values form groups in space, and observations with different values tend to be faraway from each other, we observe spatial autocorrelation. There is a set of mathematical indices designed to measure the level of spatial autocorrelation, and Moran's Index is one o f them. It is defined as:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2} \quad (1)$$

where *N* is the number of spatial units indexed by *i* and *j*; *X* is the variable of interest; $\bar{X}$ is the mean of *X*; and ω_{ij} is an element of a matrix of spatial weights. The index varies from -1 (perfect dispersion) to 1 (perfect autocorrelation). Random distribution is indicated by Morans's I equal to 0.

Distribution of Moran's indices for the electoral districts is given in a graph matrix with the following structure:

a) histogram of Moran's index values for territory units;
b) scatterplot of Moran's index values against the number of observations;
c) scatterplot of Moran's index values against the interquartile ranges of variable.



a) b) c)

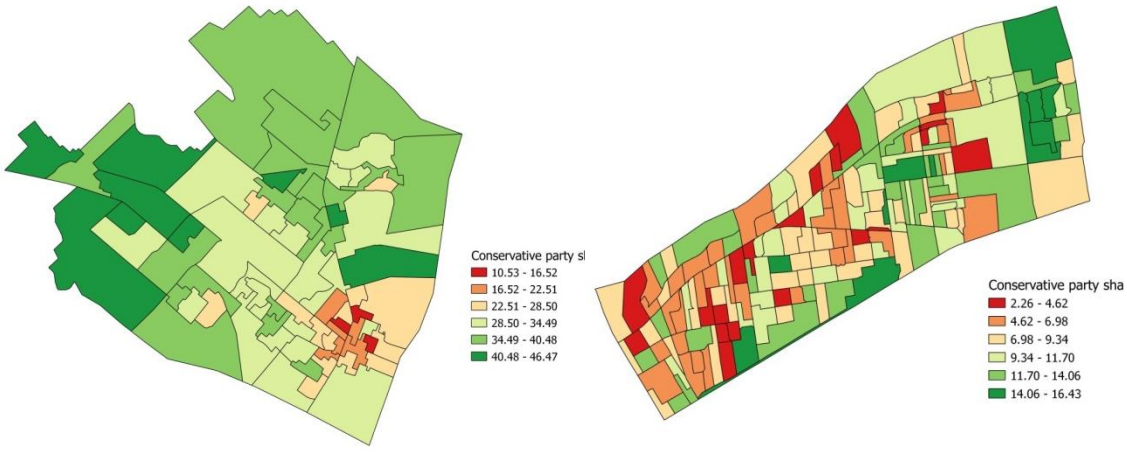
Figure 24. Moran's Index for Conservative party share (Canada, 2011).

As always, the remaining graphs are available in digital annex. The summary is given below:

- Histograms of Moran's I values for each variable only slightly change between years, demonstrating stability of geographic distribution;
- Among the aggregation levels, the highest values are observed for the administrative districts. This happens because some of the administrative districts

contain several electoral districts with very different results. Thus, we can say that on a higher level of geographical division variables are more determined;

- There are no negative values, except a couple, so there is no dispersion pattern in the data;
- For voter turnout in the electoral districts, Moran's I mainly falls in a range between 0 and 0.15, meaning random distribution of this variable;
- For Conservative party share in the electoral districts, Moran's I is above 0.15 for 50% of districts and above 0.20 for 30%. This is a good result, taking what we can see on Figure . On Figure a there is a map showing the spatial distribution of the Conservative party share at Moran's I equal to 0.22. Lower values are concentrated in one area, while higher shares can be observed in the periphery. This likely indicates urban and rural division. On Figure b, an example of the electoral district having Moran's I close to 0 can be seen. In general, such figures confirm the spatial determinance hypothesis;



a) b)

Figure 25. Distribution of Conservative party share within the electoral district (Canada, 2011).

- For Liberal and New Democratic party shares, indices are smaller (around 30% above 0.15 and 20% above 0.20);
- Relationship between Moran's I and the number of observations is not expressed. So, more heavily populated territories can have both spatially determined and random voting patterns;

- There is a strong relationship between Moran's I and the interquartile range of party shares. Most of the correlation coefficients for these two values exceed 0.30. Taking this fact, we can say that more dispersed is the variable, more spatially determined this dispersion is. In other words, more different are the political preferences in the area, more they tend to form groups in space.

Besides Moran's Index which describes spatial relationship between the components for the entire geographic unit, there are local indicators of spatial association (LISA). They are designed to examine relationships between the closest neighbors. Each observation has its own value of LISA. For example, Local Moran's index is defined as:

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^{n} w_{i,j}(x_j - \bar{X}) \tag{2}$$

where x_i is the variable value, X is the mean of that variable, ω_{ij} is a spatial weight between neighboring observations i and j, and

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^{n} (x_j - \bar{X})^2}{n-1} - \bar{X}^2 \tag{3}$$

with n equal to the total number of observations. Z-scores are derived using this formula:

$$z_{I_i} = \frac{I_i - \mathrm{E}[I_i]}{\sqrt{\mathrm{V}[I_i]}} \tag{4}$$

Local Moran's statistics allow indicating hot spots, cold spots and outliers in geographic data. Positive z-scores indicate 1.96 clusters of two types: hot spots, called HH (high-high) associations, where both the core and the neighbors have values higher than the mean value, and cold spots, named LL (low-low) associations, where values are lower than the mean value for all features. An observation is classified as HH or LL according to the difference between its value and the population mean. If the value is higher than the mean, it is marked as HH, and, logically, if the value is lower, it is marked as LL. Negative z-scores point on the

outliers of two types: LH (low-high) and HL (high-low), where the central observation has a magnitude different from its neighbors. Again, the type is selected on the basis of the difference between its value and the population mean. For all cases, p-values below 0.05 are necessary to confirm the statistical significance of the result.

Since our hypothesis is that variables are spatially autocorrelated, we expect to see some clusters and no or a very small number of outliers.

Function to calculate Local Moran's statistics is implemented in R. Its main parameters are a vector of values (for example, voter turnout for each polling division within a given electoral district) and a spatial weights list which contains description of spatial relationships between polling divisions. This list can be obtained from a spatial weights matrix which is a matrix of n rows and n columns, where n is the number of observations. Each cell of this matrix contains the value from 0 to 1 which shows the level of interaction based on the length of the common border. For example, if polling division A shares 50% of its boundary with polling division B, spatial weight of B for A is equal to 0.5. We have completed the following procedure for each electoral district:

- derived the spatial weights by a stored procedure in PostgreSQL (available in Digital Annex I), according to the length of common borders with first level neighbors, and imported them to R, along with the vector of variable values;
- constructed and filled spatial weights matrix;
- passed the obtained matrix and the vector of variable values to a function calculating Local Moran's statistics;
- written results back into PostgreSQL database;
- checked results for significance;
- prepared a set of histograms and plots.

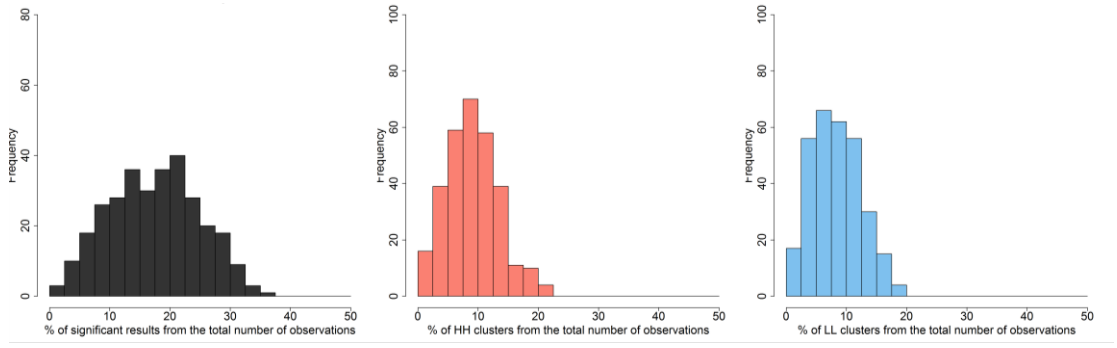
We did not find any HL and LH associations at all. None of the results had a combination of negative z-score and p-value below 0.05. It means that there are no irregular spots in the data and confirms the assumptions we did before. Examples of significant and insignificant Local Moran statistics can be seen below:

Type	LL	LH
Value	12.22	26.67
Population mean	56.41	52.62
Neighbors	32.34 33.74 27.65 37.39 43.94 38.03	80.25 62.50 84.93 57.06 45.32 55.97 58.97 56.57 57.07
Neighborhood mean	35.52	62.07
Local Moran statistic	14.438	-4.427
Expectation	-0.004	-0.004
Variance	0.251	0.14
Z-score	**28.793**	**-11.605**
p-value	**0.000**	**1.000**

Table 2. Examples of Local Moran statistics for voter turnout (Canada, 2011).

Even if there are no statistically significant outliers, we still have to estimate the number of HH and LL associations. In this case, histograms showing percentages of HH- and LL-classified observations are helpful. They have the following structure, shown on Figure :

a) percentage of the observations having significant p-values and z-scores;
b) percentage of the observations classified as HH associations;
c) percentage of the observations classified as LL associations.

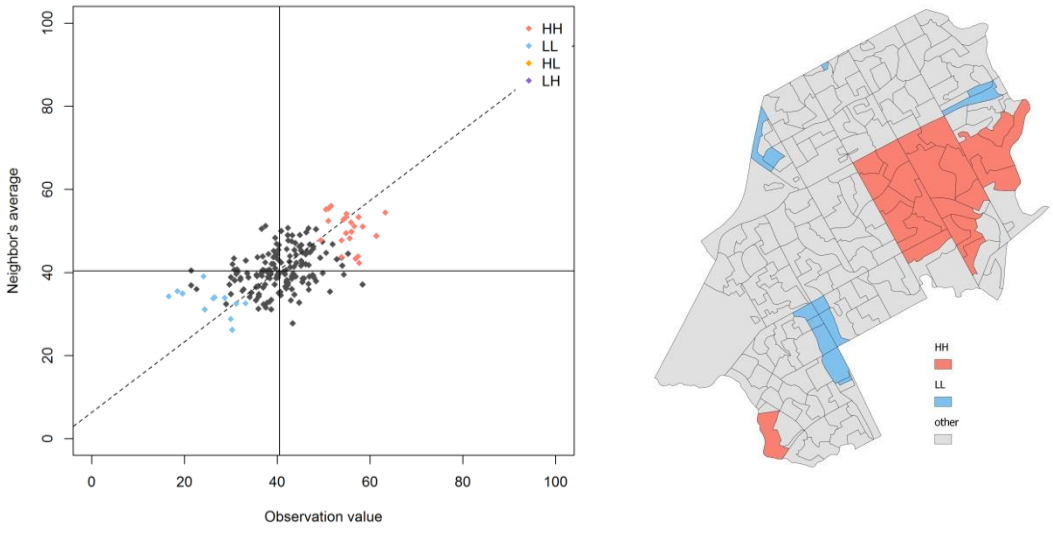


a) b) c)

Figure 26. Percentage of significant results of Local Moran statistics for Conservative party share (Canada, 2011).

Remaining histograms are available in a digital annex. For all histograms, we can tell that the electoral districts have mainly between 7.5 and 22.5% of statistically significant results. To show what do these numbers mean in details, we can select

an electoral district where the percentage of statistically significant results is close to an average (for example, 17.42% for Conservative party share in 2011). Data for this electoral district will be visualized with an exploratory plot showing observation values on x axis and neighborhood means on y axis (see Figure 27a). The scatterplot is divided into quarters by vertical and horizontal lines crossing the population mean. Thus, the upper right quarter contains observations that can possibly be included into HH associations because their value and the average value of their neighborhood is higher than the population mean, and so on for other types of local spatial associations. Grey points show for all polling divisions within a given electoral district, while statistically significant HH associations are marked with red color and LL groups are highlighted with blue. There are some points in the upper left and lower right quarters that could probably be marked as HL and LH associations but they are not because their p-values are lower than 0.05.



a) b)

Figure 27. Exploratory plot of Local Moran's statistics for voter turnout (electoral district #53022, Canada, 2011).

We can see that the orientation of the point cloud and the trendline indicate positive correlation between the variable value of the cores and their neighborhoods. So, we can observe similarity in polling divisions located close to each other. These are typical scatterplot and map, for other variables in other years

and in other electoral districts this looks much the same. On Figure b, you can find locations of the defined HH and LL spots on the map of the electoral district.

In general, results obtained from Moran and Local Moran tests confirm our hypothesis regarding the spatial determinance of the electoral data.

4.2 Multivariate spatial analysis.

The previous subchapter where we worked with Global and Local Moran statistics was dealing with univariate data, i.e. the investigation of a single variable. We studied voter turnout and each party share separately. To have a better picture, multivariate methods of the analysis should be used. Hierarchical clustering is one of such methods. If we group polling divisions into clusters, we can analyze the distribution of the qualitative class identifiers. To produce the clusters, we can use all the main variables. If the instances of the same class are located together, and this class is related to a certain electoral behavior, we can say that such behavior has concentrated nature. The amount of outliers (instances of one class surrounded by the instances of other classes) will let us know how random is the distribution of the main variables.

An appropriate geographic level for performing the analysis can be the entire country, but due to computational limitations (8Gb RAM was not enough to work with entire Canada) we have selected province level, particularly Quebec and Ontario provinces.

The first step is to find an appropriate clustering algorithm. We want to the clustering to reflect the patterns (there should be many classes with a large amount of observations), and to indicate the outliers (there should be some classes with a small number of instances).

Distance-based algorithms like complete distance algorithm (see Figure 28a) produce very uneven structure of classes:

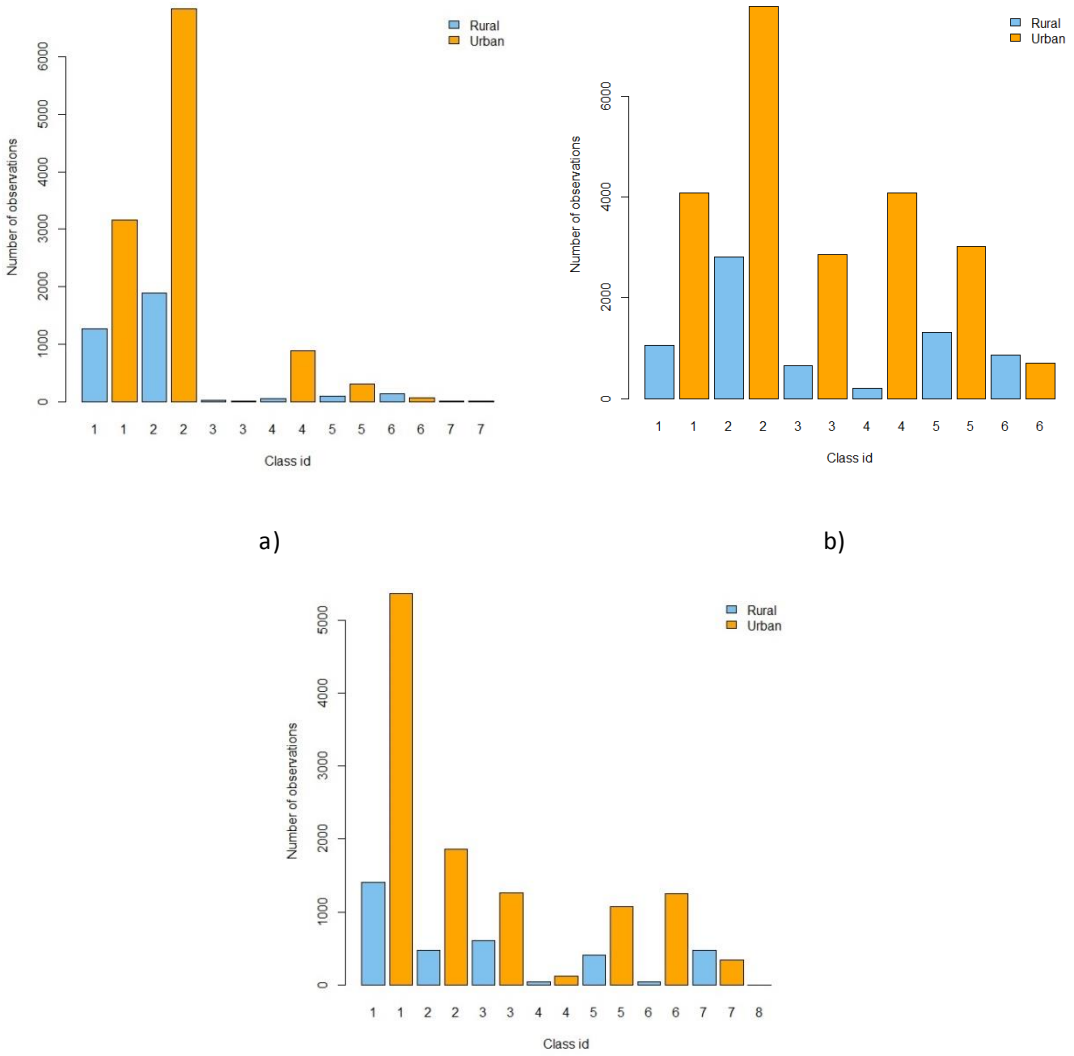


Figure 28. Distribution of the observations among the clusters with urban and rural indicators for different clustering algorithms (Quebec, Canada, 2011).

We can see that there are 2 totally dominating classes, while the other classes are represented by too small number of instances. This algorithm could work great for defining the outliers but at the same time it is poor for demonstrating the patterns because it has just 2 dominating classes and they will probably form similar neighborhoods just because of their amount.

Variance-based algorithms like Ward's algorithm produce a quite balanced structure of classes (see Figure 28b). Such algorithms could work well for defining the patterns, while they are not helpful for defining the outliers.

McQuitty's Similarity Analysis is an approach that produces the cluster structure which is good both for getting both the patterns and the outliers (see

Figure 28c). Besides one dominating class (#1), we have several classes with a large number of instances (#2,#3,#5,#6 and #7) and some classes of outliers (#4,#8,#9 and #10).

It was expected that classification should indicate the difference between rural and urban areas, i.e. some classes will be completely urban and some classes will be completely rural, but the obtained results don't confirm this assumption. The ratio between the number of rural and urban divisions in Quebec is 14737/3458=4.26, and, as can be seen from Figure 29, this ratio is what we see for the main classes. The most remarkable exclusions are #6 which has significantly larger amount of urban divisions (ratio 1233/41=30.07) and #7 which has a larger amount of rural divisions than urban ones, with ratio 344/471=0.73.

To learn the difference between the classes, we created a bar chart showing the average party shares and the turnout for each class:

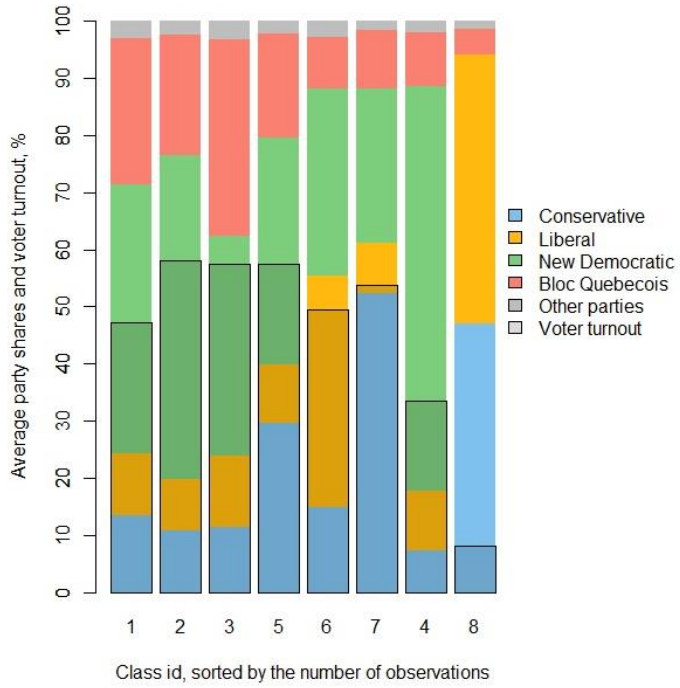


Figure 29. Average party shares and voter turnout for polling division classes, ordered by the amount of observations (Quebec, Canada, 2011).

Turnout values are displayed by semitransparent bars on top of the colored bars for shares. We can see that class #6 is characterized by highest support of Liberal party among all classes (we exclude class #8 because it has just 1 instance),

while class #7 reflects the largest Conservative share, and many other differences. Combining these observations with the urban/rural structure of the classes, we can say that Liberal party was mostly successful in urban areas, while Conservative party had better results in rural areas than in urban.

The next question is how are the classes distributed geographically. To answer this question, we have created a spatial view in PostgreSQL and produced a qualitative map based on the class identifier in QGIS. The first step here is to make the visual estimate of how are the instances of each class distributed in space. In addition to visual analysis, there should be some quantitative method to estimate the level of clustering for each class. We can not use the standard methods like Moran's index because they deal with quantitative values, while class identifiers are qualitative. In this case, the neighbor analysis is the way to go. We built the list of first-level neighbors. Only polling divisions within the same electoral district can become neighbors. If two polling divisions have a common border but are located in different electoral districts, they are not considered as neighbors. Then we found the spatial weight of similar neighbors. This number is a ratio between the total length of the common borders with neighbors of the same class and the perimeter of the polling division minus the common border with polling division from other electoral districts. If ratio is equal to 1, all surrounding divisions have the same class, and we can say that the class instances are clustered. If the ratio is equal to 0, all surrounding divisions have different classes, indicating an outlier (see Figure 30):

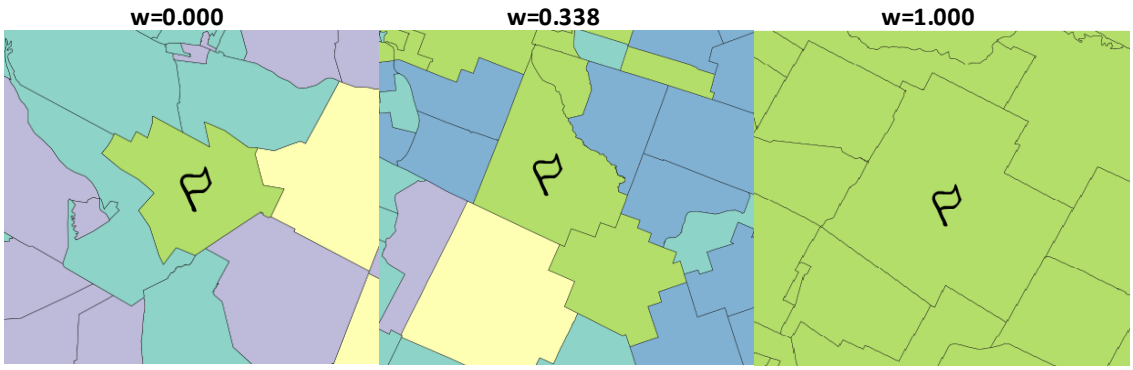


Figure 30. Examples of polling divisions with different similarity weights.

To estimate the level of clustering for all of the instances of each class, we built histograms showing the distribution of weights:

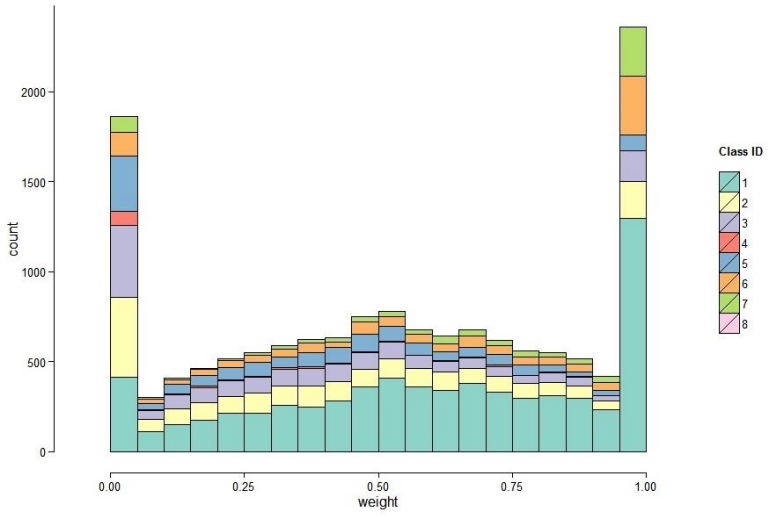


Figure 31. Stacked histogram of the similar neighbors weights (Quebec, Canada, 2011).

Looking at this histogram, we can say the following:

- Distribution of the similarity weights is close to normal, except two spikes at 1 and 0 meaning clusters and outliers, respectively;
- There are more clustered groups than the outliers;
- At the same time, the number of outliers is very large;
- Balance between clusters and outliers is not the same for all classes: #1, #6 and #7 include more clusters than outliers, while #2, #3, #4 and #5 are the opposite.

Below there are some examples illustrating the distribution which include the map and an individual histogram of the similarity weight distribution. Class #1 is the most common class, having 6771 instances. We can see that they tend to form groups of neighbors, like shown on Figure 32:

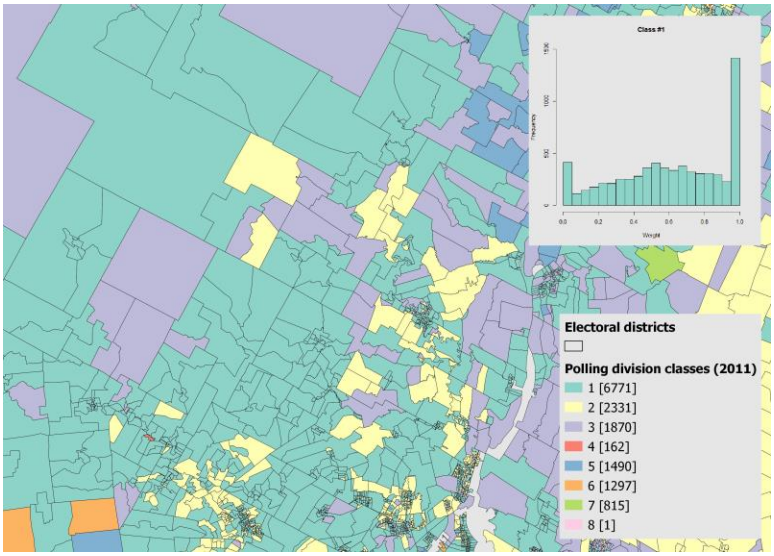


Figure 32. Instances of class #1 in middle-South Quebec.

As already been said above, classes #2, #3, #4 and #5 demonstrate less spatial homogeneity. An example can be seen below:

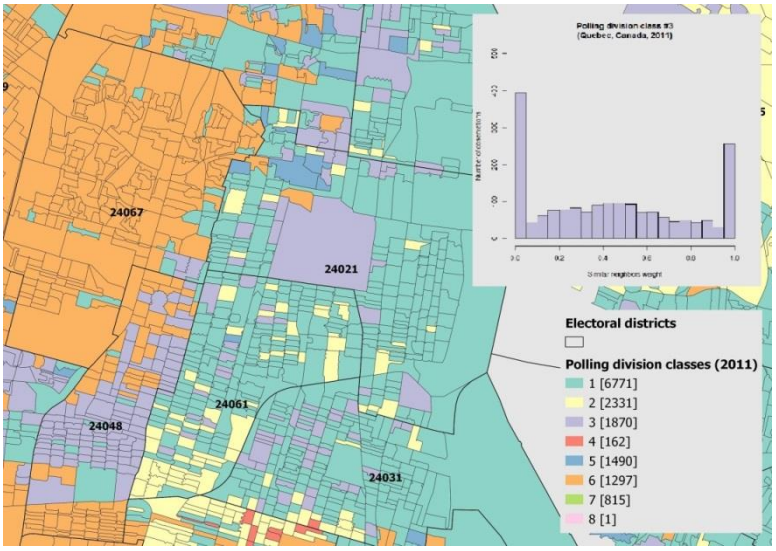


Figure 33. Instances of class #3 in Montreal (Quebec, Canada, 2011).

On the contrary, class #6 is represented by highly concentrated group in Montreal and its suburbs (Chambly, Mont-Saint-Hilaire, Vaudreuil-Dorion and others). 93% of its instances (1193 from 1274) are located there. On the Figure 33 above we can see clusters of the orange polygons in that area. Class #7 forms a compact group in a rural area in south-east of the province (the map is available in a digital annex).

The same procedure was completed also for Ontario province and for 2006 and 2008 years for these two provinces. The outcomes are the same, with even more similarity than in Quebec in 2011. For example, here is the same stacked histogram of similar neighbors weights for Quebec in 2008:

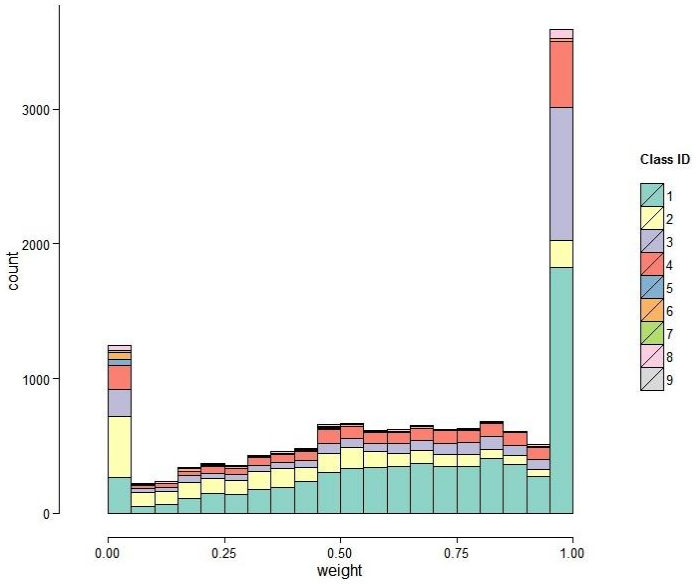


Figure 34. Stacked histogram of the similar neighbors weights (Quebec, Canada, 2008).

The remaining graphs can be found in digital annex. After getting all the graphs and maps and verifying the predominant similarity of neighborhoods, it is necessary to perform relevant simulations to estimate the significance of the obtained results. There was some chance that dominating classes like class #1 tend to form similar neighborhoods just because they have too many instances and they eventually group together. The first simulation was the complete randomization of the class identifiers. After performing clustering with the same height as before, all class identifiers were shuffled within the entire province. Thus, we had the same amount of observations assigned to each class but they were randomly distributed in space. For this simulation, similar neighbor weights look like this:

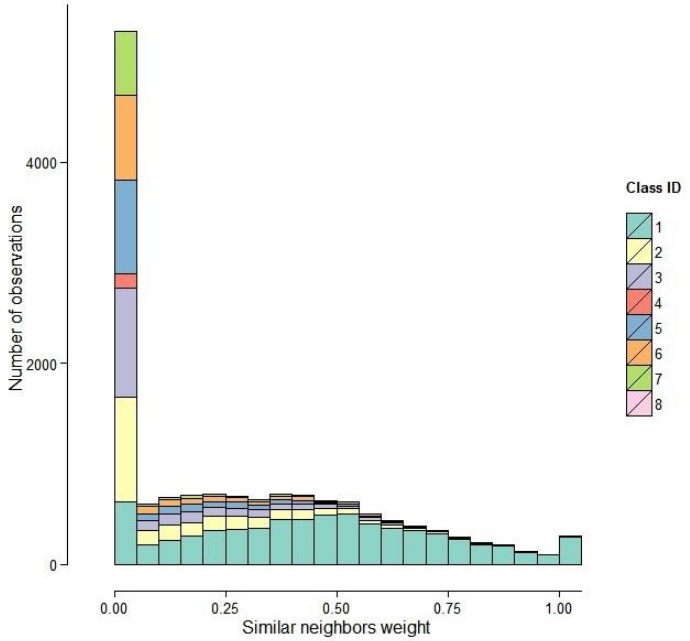


Figure 35. Stacked histogram of the similar neighbors weights (complete randomization of classes, Quebec, Canada, 2011).

We see that the results are very different. There is a dominating peek on 0.0 value where all classes can be recognized. Only class#1 demonstrates some similarity, but from weights higher 50% are less and less popular, ending with a very slight peak at 1.00. From the previous figures we can see that this is not the case for the real distribution where weights higher that 50% form a well distinguishable plateau. At the same time, the complete randomization is not the best way of simulation because it does not take the electoral district borders into account. The second way to go is to randomize the class identifiers by the electoral district. In this case, districts having mostly one class will remain almost unchanged, instead of being spread throughout the entire province. We expected the stacked histogram to look as something between the real one and a completely random one. Indeed, it looks like this:

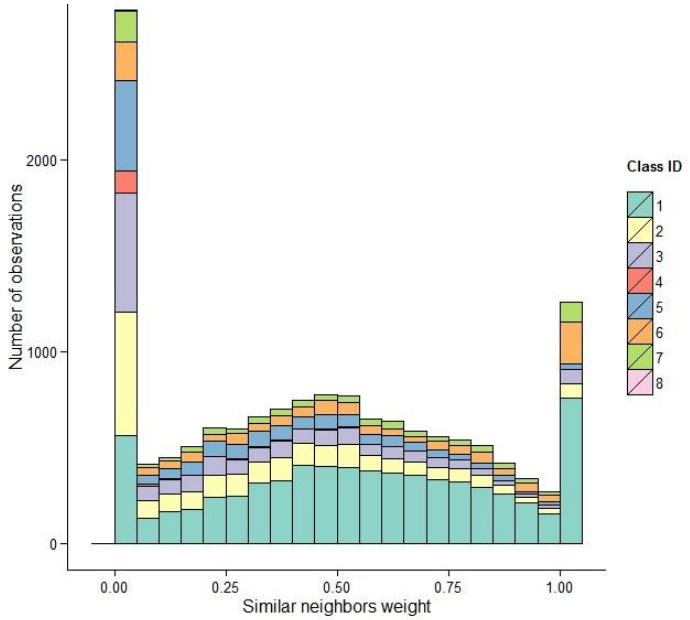


Figure 36. Stacked histogram of the similar neighbors weights (randomized by electoral district, Quebec, Canada, 2011).

We can see that the share of similar neighborhoods has increased, but still it is very far from what we observed with the real data.

Finally, results of multivariate analysis confirm the expected identity of polling divisions and their neighborhoods.

4.3 Electoral fraud modelling: a simulation study (II)

In chapter 3.3, we have modelled a scenario where a political party commits the electoral fraud in all polling divisions of the selected electoral districts. 8% of the influenced observations were located in 10% of the electoral districts. Thinking about a proper modelling scenario to check the outcomes Chapter 4, we have to reject the described scenario because if we perform some manipulation on all observations in the electoral district, the modelled data will inherit the geographical pattern of the original data. Instead, we have to introduce the manipulated data sporadically. For example, we could randomly select from polling divisions which have Conservative party share less than 33% in 2008. This is 28951 out of 63416 polling divisions. We could get 6300 polling divisions from that list to get the right sample size.

Again, distribution of voter turnout and quantity of empty ballots is normal, with turnout mostly between 40 and 60% and the empty ballot count mainly around 200. This data enables modelling of ballot stuffing. So, we extracted a data sample of size 6300 and performed modelling with the same conditions like the first scenario (50-75% from the empty ballot count in 2008 is used in ballot stuffing in 2011), but this time for Conservative party. Density scatterplots confirm the interference, as for the previous scenario:

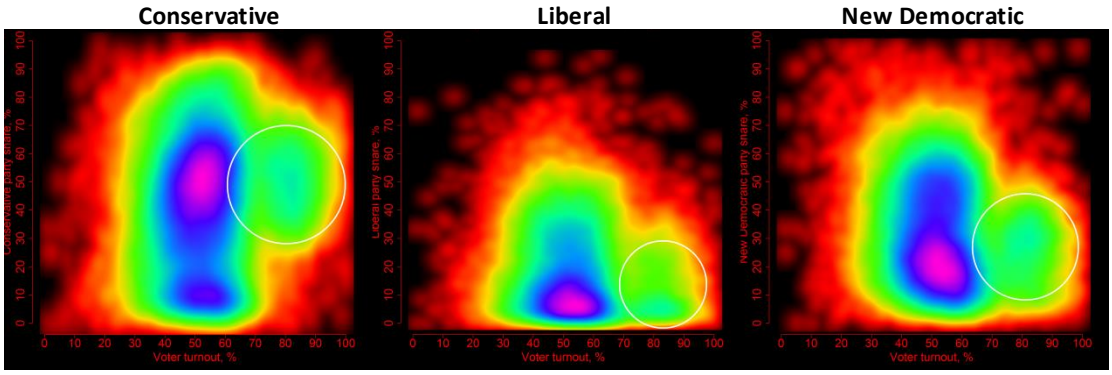


Figure 37. Density scatterplots for voter turnout and party shares (modelled data).

This confirms the conclusions already made in Chapter 3.3, and now the main question is how did the simulation affect the results of the spatial analysis. We can

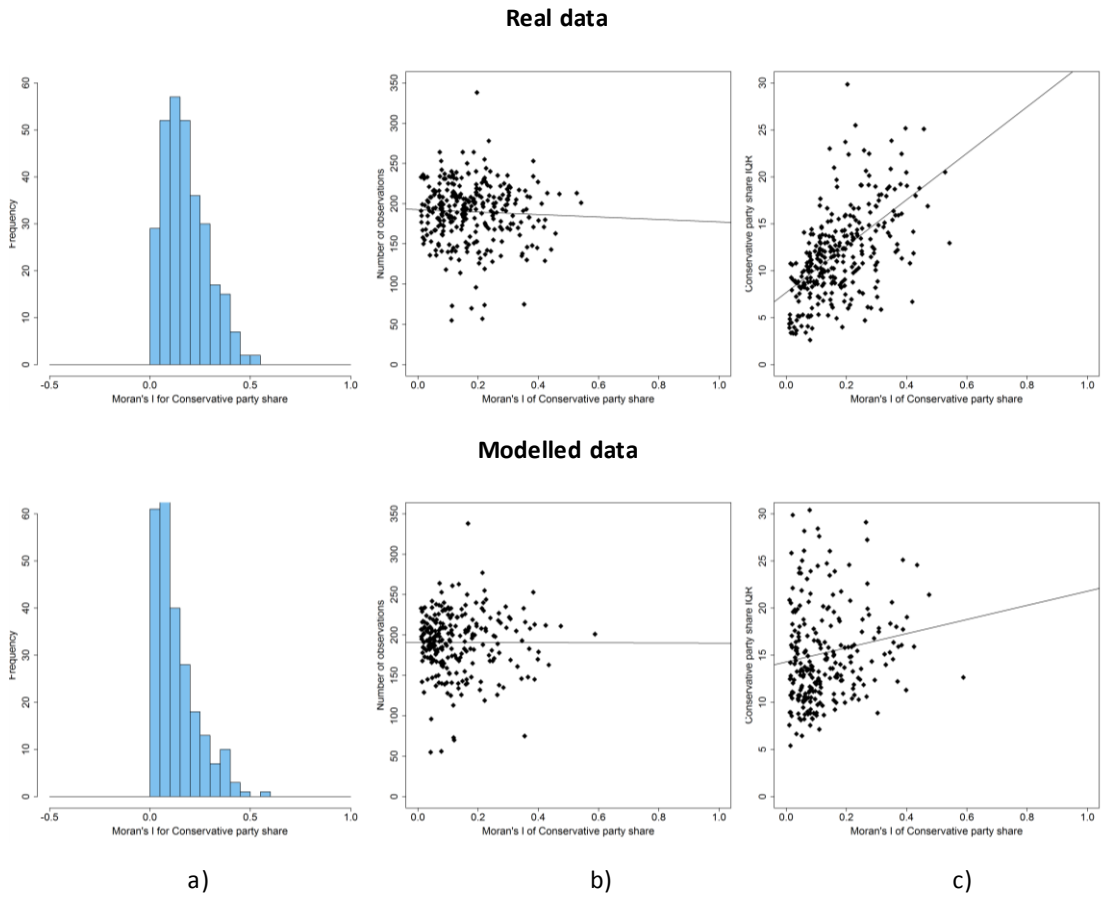
see that Global Moran's Indices changed dramatically for Conservative party share and voter turnout. The structure of the graph matrix is as before:

a) histogram of Moran's index values for territory units;

b) scatterplot of Moran's index values against the number of observations;

c) scatterplot of Moran's index values against the interquartile ranges of variable.

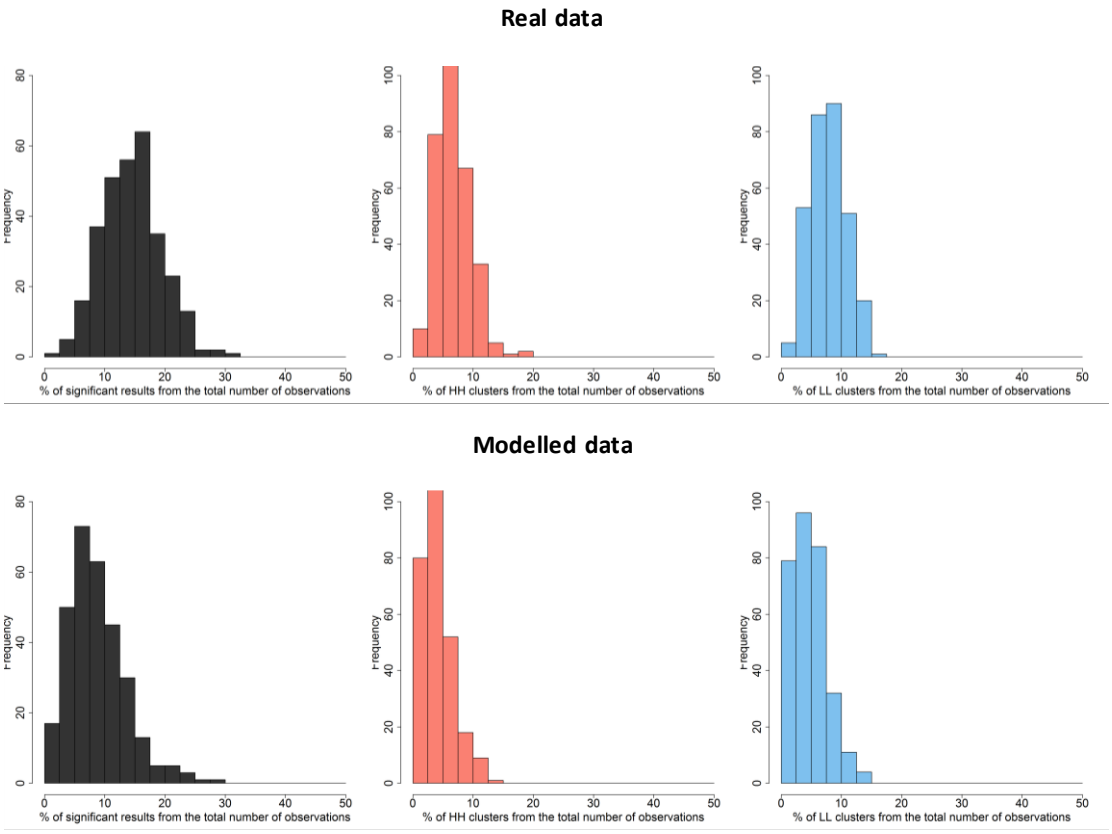
We can see how different are the values of Local Moran's I, and how the simulation also affected the relationship between the index and the interquartile ranges:



a) b) c)

Figure 38. Global Moran's Index for Conservative party share (Canada, 2011).

Thus, Moran's index looks like a sensitive measure of the interference. Less than 10% of the observations were affected, but most of the indices shifted into a range between 0 and 0.1 which means spatial randomness. Local Moran's test shows the same differences for voter turnout and Conservative party share, for which the percentage of statistically significant clusters decreased. This can be seen below:



a) b) c)

Figure 38. Percentage of significant results of Local Moran statistics for voter turnout (Canada, 2011).

Looking at the exploratory plots, we can see a significantly shifted group of observations for voter turnout (12 in total, marked by red circle):

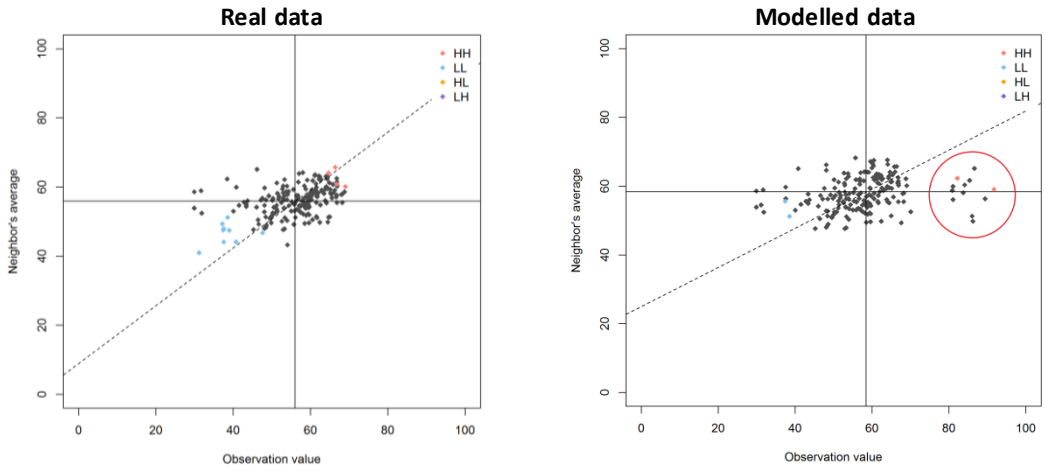


Figure 39. Exploratory plots of Local Moran statistics for voter turnout (Canada, 2011).

They are likely the observations where we have introduced ballot stuffing. We can see that they have higher own values on x axis, while neighborhood means on y axis

are the same as for most of the observations. In fact, there are 14 polling divisions where the data were influenced, so the graph can not detect all of them. As we see from the graph, significant Local Moran's result are occasional and they do not reflect that outstanding group. Also, there are no statistically significant HL and LH associations indicated, as before. So, we can say than even if Local Moran statistics indicate the interference in general, it can not detect specific examples. There can be three possible explanations of this:

- there are no outliers indeed (which we would like to reject because we can see points in LH and HL zones of the exploratory plot that have quite big difference between their values and neighborhood means);
- algorithms were used for calculations (probably, poor methodology for calculating spatial weights matrix based only on common border lengths);
- methodological approach (Local Moran's I is not a suitable method for detection of such data).

To check the second hypothesis, we have run Local Moran's index tool in ArcGIS. We mapped voter turnout values for real and modelled data using the same color scheme. Then, we ran Anselin Local Moran's I tool and overlaid identifiers of local spatial association types upon this map to see the changes:

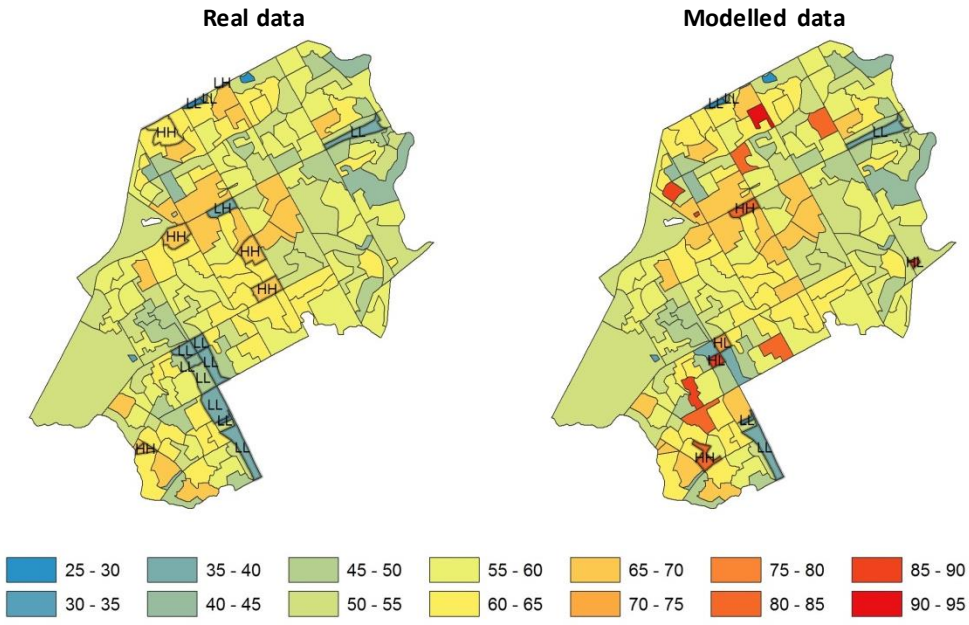


Figure 40. Local Moran statistics for voter turnout (electoral districts #53022, Canada, 2011).

First, we can clearly recognize only 14 out of 19 simulated changes by dark orange and red color. Some of the observations that were classified as LL and were not changed during the modelling process remained being LL, while some were changed and transformed from LL to HL and insignificant. One observation in the center, which was a LH-outlier, was transformed into HH after modelled ballot stuffing. This happened because the real value was low and the neighboring values were high, while after modelling the value increased (see the map on Figure 40). At the same time, some observations changed their class when neither their values nor the neighbors values had changed. But the most surprising thing was that some of the changed observations did not appear among significant results neither in real data nor in modelled data, though it is clear that their values are very different from neighboring values. Looks like this happens because both the observation and the neighbor need to have values different from population mean. In our case, only the value has is different, while the neighborhood is close to an average. Thus, even if Local Moran's statistics help to detect the interference, like shown above, it does only the general estimate if the data was changed. To enable detection of specific observations that were manipulated, we need to look for another approach. Coming back to Figure 39, we can expect that comparing values and neighborhood means without comparing them with population mean, like Local Moran's I does, could give us better results. The workflow is the next:

- get values and neighborhood means for selected electoral district from the database;
- get id list of polling divisions affected by modelling for the same district, this will be a model list;
- calculate ratio between value and neighborhood mean for each observation;
- calculate distances between these ratios;
- get the list of observations which make up an upper decile (10-quantile) with largest distances, this will be a method list;
- match the model list with the method list.

For voter turnout in the electoral district we were looking at above, the number of matches was 12 – the same as the number of outstanding points on the plot. For 14 modelled observations, this is a good number, but at the same time method list contained 19 observations because it was made up from 10-quantile. So, the true effect of the procedure is about 63% for that polling division. If we do the same for remaining polling divisions, we get results which detect the affected observations much better than Local Moran's I did:

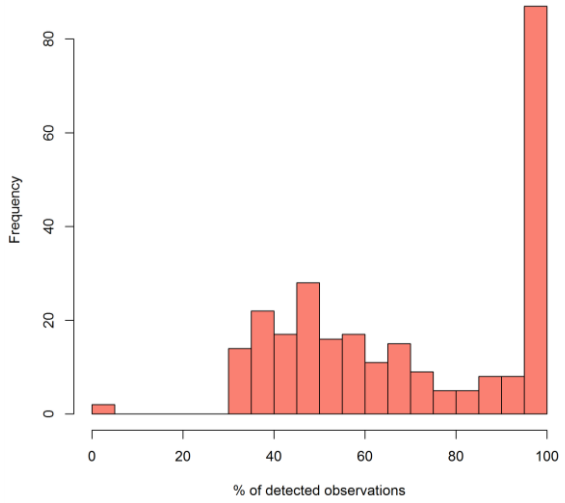


Figure 41. Percentage of detected observations for the electoral fraud modelling scenario in the electoral districts (Canada, 2011).

Thus, we can confirm that spatial analysis of geographic distribution of the electoral data can give valuable information regarding the electoral fraud detection. After modelling a real-life scenario we could reveal the places of interference by using statistical methods.

CONCLUSION AND FUTURE WORK

During the analysis, we have applied a set of statistical techniques to answer the question about the presence of spatial patterns of the electoral data. The analysis provided strong evidences of large-scale regional patterns in voting behavior, and gave an insight about the small-scale local patterns of neighborhood similarity. Simulation studies confirmed significance of the obtained results.

The offered methods and techniques have a lot of space for improvement. First, the given approach should be tested on other countries and territories. Second, computational methods might be improved, especially for time-series data analysis. Third, large cities can be analyzed as sets of neighborhoods, which gives new aggregation levels for the study. And finally, the next elections in Canada are coming in 2015, and for us it means new data which might be used as a control for the current results.

BIBLIOGRAPHIC REFERENCES

1) Peirce f. Lewis. *Impact of negro migration on the electoral geography of Flint, Michigan, 1932–1962: a cartographic analysis.* Annals of the Association of American Geographers (1965), 55 (1)

2) C. Pattie, D. Dorling, R. Johnston. *The electoral geography of recession: local economic conditions, public perceptions and the economic vote in the 1992 British general election.* Transactions of the Institute of British Geographers (1997), 22(2): 147-161

3) John O'Loughlin. *The Electoral Geography of Weimar Germany: Exploratory Spatial Data Analyses (ESDA) of Protestant Support for the Nazi Party.* Political Analysis (2002), 10(3): 217-243

4) Arturo de Nieves Gutiérrez de Rubalcava, Manuel García Docampo. *The territorial variable in the analysis of electoral behavior*. Report on Spain National Congress of the Asociación Española de Ciencia Política (2013).

5) Kevin R. Cox. *The voting decision in a spatial context.* Progress in Geography 1.1 (1969): 81-117.

6) Ron Johnson, Charles Pattie, Danny Dorling, Iain MacAllister, Helena Tunstall and David Rossiter. *The Neighborhood Effect and Voting in England and Wales: Real or Imagined?* Philip Cowley, David Denver, Andrew Russell. British Elections and Parties review (2000), 10(1): 47-63

7) Waldo Tobler. *A computer movie simulating urban growth in the Detroit region.* Economic Geography (1970) 46(2): 234-240

8) Luc Anselin. *Spatial econometrics: Methods and models.* Dordrecht: Kluwer Academic (1988)

9) Walter R. Mebane, Jr. Kirill Kalinin. *Comparative Election Fraud Detection*. Report on Annual Meeting of the American Political Science Association (2009)

10) Joseph Deckert, Mikhail Myagkov and Peter C. Ordeshook. *The Irrelevance of Benford's Law for Detecting Fraud in Elections.* Caltech/MIT Voting Technology (2010) 9

11) Bernd Beber and Alexandra Scacco. *What the Numbers Say: A Digit-Based Test for Election Fraud.* Political Analysis (2012)

12) Mikhail Myagkov, Peter C. Ordeshook and Dimitri Shakin. *The Forensics of Election Fraud: Russia and Ukraine.* Cambrige Press, 2009.

13) Peter Klimek, Yuri Yegorov, Rudolf Hanel, and Stefan Thurner. *Statistical detection of systematic election irregularities.* Proceedings of the National Academy of Sciences of the United States of America (2012)

14) Konstantin Sonin. *Presidential Elections in Russia: Massive Vote Fraud Ensures that Legitimacy is in Doubt, but the Policy Direction is not.* Report for the Forum for Research on Eastern Europe and Emerging Economies (2012)

15) Skye Christensen. Mapping Manipulation: Digital Observation of Electoral Fraud. Thesis Submitted to Uppsala University, Department of Government (2011)

16) Jowei Chen. Voter Partisanship and the Effect of Distributive Spending on Political Participation. American Journal of Political Science (2013) 57(1): 200-217

17) ESRI GIS Dictionary: http://support.esri.com/en/knowledgebase/GISDictionary/term/MAUP

18) US Legal Dictionary: http://definitions.uslegal.com/b/ballot-stuffing/

Annex 1: Data structure tables

#	Field Name	Value	Example
1	Electoral District Number	The number of the electoral district	10001
2	Electoral District Name	The name of the electoral district	Avalon/Avalon
3	Polling Station Number	The number assigned to the polling station	1
4	Polling Station Name	A name that generally represents the locality of the polling division boundary	Grates Cove
5	[Candidate 1 name]	The number of valid votes for the first candidate on the ballot at this polling station only	[Scott Andrews] 45
6	[Candidate 2 name]	The number of valid votes for the second candidate on the ballot at this polling station only	[Matt Crowder] 0
7	[Candidate 3 name]	The number of valid votes for the third candidate on the ballot at this polling station only	[Randy Wayne Dawe] 0
8	[Candidate 4 name]	The number of valid votes for the fourth candidate on the ballot at this polling station only	[Matthew Martin Fuchs] 10
…	…	…	…
9	[Candidate n name]	The number of valid votes for the n-th (if any) candidate on the ballot at this polling station only	[Fabian Manning] 16
10	Rejected Ballots	The number of rejected ballots at this polling station	0
11	Total Votes	The total number of ballots counted at this polling station	71
12	Electors	The number of electors on the list of electors for this polling station	165

Table 1. “pollbypoll_bureauparbureau” CSV format of General elections results data.

#	Field Name	Value
1	Electoral District Number	Electoral district number
2	Electoral District Name_English	The English name of the electoral district
3	Electoral District Name_French	The French name of the electoral district
4	Polling Station Number	The number assigned to the polling station
5	Polling Station Name	A name that generally represents the locality of the polling division boundary
6	Void Poll Indicator	Indicates that a poll exists but has no electors
7	No Poll Held Indicator	Indicates that the returning officer intended to hold this poll, but unforeseen circumstances prevented it
8	Merge With	Indicates the number of the polling station with which the results of this poll were merged
9	Rejected Ballots for Polling Station	The number of rejected ballots at this polling station
10	Electors for Polling Station	The number of electors on the list of electors for this polling station
11	Candidate's Family Name	The family name of the candidate
12	Candidate's Middle Name	The middle name of the candidate
13	Candidate's First Name	The first name of the candidate
14	Political Affiliation Name_English	The short-form English name of the candidate's political affiliation
15	Political Affiliation Name_French	The short-form French name of the candidate's political affiliation
16	Incumbent Indicator	"Y" if candidate was the incumbent, "N" otherwise
17	Elected Candidate Indicator	"Y" if candidate was elected, "N" otherwise
18	Candidate Poll Votes Count	The number of valid votes the candidate obtained at this polling station

Table 2. "pollresults_resultatsbureau" CSV format of General elections results data.

Since "pollresults" format contains several rows for each electoral division, it is better to show an example in a separate table:

1	2	4	5	6	7	9	10
10001	Avalon	1	Grates Cove	N	N	0	165
10001	Avalon	1	Grates Cove	N	N	0	165
10001	Avalon	1	Grates Cove	N	N	0	165
10001	Avalon	1	Grates Cove	N	N	0	165
10001	Avalon	1	Grates Cove	N	N	0	165
11	**13**	**14**	**16**	**17**	**18**		
Andrews	Scott	Liberal	Y	Y	45		
Crowder	Matt	Green Party	N	N	0		
Dawe	Randy Wayne	Independent	N	N	0		
Fuchs	Matthew Martin	NDP-New Democratic Party	N	N	10		
Manning	Fabian	Conservative	N	N	16		

Table 3. Example of "pollresults" format of General elections results data.

#	Field Name	Data type	Description	Example
1	ed_num	integer	Electoral district number	10001
2	ed_name	character varying(150)	Electoral district name in English	Avalon
3	pd_num	character varying(10)	Polling division number	1
4	pd_name	character varying(150)	Polling division name in English	Grates Cove
5	pd_type*	character varying(3)	Polling division type	pol
6	pd_agg**	boolean	Specifies whether the polling division results were aggregated or not	false
7	ed_incub	varchar(50)	Which party member was holding the chair by the moment of the elections	Liberal
8	ed_res	varchar(50)	Which party member was had the majority within the electoral district	Liberal
9	pd_res	varchar(50)	Which party member had the majority within the polling division	Liberal
10	total_electors	integer	Total number of people who are eligible to vote on the station	165
11	total_votes	integer	Number of people who actually came and voted	71
12	voter_turnout	numeric(6,2)	Ratio between the actual voters and total number of electors	43.03
13	conservative_v	integer	Number of votes for a candidate who belongs to Conservative party	16
14	conservative_p	numeric(6,2)	Percentage of votes for a candidate who belongs to Conservative party	22.54
15	liberal_v	integer	Number of votes for a candidate who belongs to Liberal party	45
16	liberal_p	numeric(6,2)	Percentage of votes for a candidate who belongs to Liberal party	63.38
17	new_democratic_v	integer	Number of votes for a candidate who belongs to New Democratic party	10
18	new_democratic_p	numeric(6,2)	Percentage of votes for a candidate who belongs to New Democratic party	14.08
19	bloc_quebecois_v	integer	Number of votes for a candidate who belongs to Bloc Québécois	null
20	bloc_quebecois_p	numeric(6,2)	Percentage of votes for a candidate who belongs to Bloc Québécois	null
21	green_v	integer	Number of votes for a candidate who belongs to Green Party	0
22	green_p	numeric(6,2)	Percentage of votes for a candidate who belongs to Green Party	0.00
23	other	integer	Total number of votes for candidates from other parties	0

Table 4. Structure of the summarized data.

* If the pd_num is between 0 and 499, it is a polling division which is represented as a polygon in polling division shapefile. Such rows have 'pol' value in pd_type field. If pd_num is between 500 and 599, this is a mobile polling station or a single building station which is represented as a point in ca_2011_ps dataset. Such rows have 'mob' value in pd_type field. If pd_num is 600 and more, it

stands for the advanced poll which is the place were people vote a few days before elections if they are not able to be there at the elections day. Such rows have 'adv' value in pd_type field.

** Some of the pd_num values include letters, like '1A', '1B', etc. These identifiers can not be related to polling districts polygons because for '1A' and '1B' there is only one polygon named '1'. To relate the tabular data and geographic features, we had to aggregate results from these specific id's. It means that we have inserted additional rows in the table, and this can affect the analysis results. To separate the aggregated and primary results, we had to add a marker field. An example of such aggregation can be seen below:

ed_num	pd_num	pd_agg	total_electors	total_votes	voter_turnout
35045	71-1A	false	448	134	29.91
35045	71-1B	false	328	107	32.62
35045	71-1	true	776	241	31.06

Table 5. An example of data aggregation.

Field Name	Description	Example	Years
PD_ID	Elections Canada unique identifier for Polling Divisions.	28921	all
PD_NUM	Polling Division number.	1	all
PD_NBR_SFX	Polling Division suffix.	0	all
PD_TYPE	Type of the Polling Division. Value is "N" for Normal.	N	all
ADV_POLL	Number of the Advance Poll the Polling Division refers to.	600	all
FED_ID	Elections Canada unique identifier for the electoral district.	1180	2011 only
FED_NUM	Electoral district number.	10001	all
A_UPDT_DTE	Date stamp of the last attribute modification.	2006-12-01	2011 only
G_UPDT_DTE	Date stamp of the last geometric modification.	2003-05-29	2011 only
EMRP_NAME	Concatenation of PD_NUM, "-" and PD_NBR_SFX.	1	2006 missing
POLL_NAME	Polling Division name.	Grates Cove	2011 only
PN_UPDT_DT	Date stamp of the last Polling Division name modification.	null	2011 only
AD_UPDT_DT	Date stamp of the last Advance Polling District modification.	null	2011 only
URBAN_RURA	Urban–Rural indicator.	R	2011 only

Table 6. The attribute structure of polling division data.

Field Name	Description	Example
AGG_LEVEL	Territory unit type: 1 - province 2 - administrative district 3 - city 4 - electoral district	1
UNIT_ID	Territory unit ID	11
UNIT_NAME	Territory unit name	Quebec
ED_ID	Electoral district ID	24009
PD_NUM	Polling Division ID	19

Table 7. Geographic relations table structure.

Annex 2: An example of R function

Most of R functions created for this work have the same structure. To give readers a chance to look how they work in general, one function is attached. Complete set of functions is available in a digital annex on a DVD.

```
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
#     #import necessary libraries and set database connection #     #
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

library(DBI)
library(RPostgreSQL)
drv<-dbDriver("PostgreSQL")
con<-dbConnect(drv, dbname="Thesis", user="postgres")

# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
#     #range and interquartile range distribution #     #
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

rgs<-function(var,year) { #function for calculating ranges

#the function builds histograms for simple range
#and interquartile range at several aggregation levels,
#as well as their scatterplots against the number of observations

#"var" specifies the short name of the variable
#and selects the correspondent column IDs

if (var=="turn") {
var.name<-"Turnout" #full name used for graph labeling
var.col<-8                #column id of values column
                          #in ca_year.getdata() stored procedure
} else if (var=="con") {
var.name<-"Conservative party share"
var.col<-9
} else if (var=="lib") {
var.name<-"Liberal party share"
var.col<-10
} else if (var=="dem") {
var.name<-"New Democratic party share"
var.col<-11
} else {
stop("Incorrect 'variable' option. The first parameter should be 'con',
'lib' or 'dem'.")
}

#"year" is the year of elections and it is directly used
#in the database query strings and graph label strings

if (year%in%c(2006,2008,2011)) {
rs1<-dbGetQuery(con,paste("select * from ca_",year,".var(1)",sep=""))
rs2<-dbGetQuery(con,paste("select * from ca_",year,".var(2)",sep=""))
rs3<-dbGetQuery(con,paste("select * from ca_",year,".var(3)",sep=""))
rs4<-dbGetQuery(con,paste("select * from ca_",year,".var(4)",sep=""))
} else {
stop("Incorrect 'year' option. The second parameter should be '2006', '2008'
or '2011'.")
}
```

```
#prepare a list of datasets, their names
#and vertical ranges for their plots

rs_list<-list(rs2,rs3,rs4)
rs_namelist<-c("Administrative districts","Urban municipalities","Electoral
districts")
rs_ylims<-c(120,75,150,35,25,50)

#open writing session for the graph matrix file and create a canvas
filename<-paste("D:/graphics/ranges/",var,"_ranges_",year,".png",sep="")
png(filename, units="in", width=28, height=21, res=300)
par(mfrow=c(3,4))

#get data for building the graphs

rs.id=2
for(rs in rs_list) { #aggregation level loop

       res.mat<-matrix(ncol=4,nrow=0) #results matrix
       colnames(res.mat)<-c("Unit_name","N_obs","Range","IQR")

       #go through each territory unit and get the variable vector

       for(i in 1:nrow(rs)) { #territory unit loop

              unit.id<-rs[i,1]
              unit.name<-rs[i,2]

              #this is a fix for names containing apostrophe symbol
              if (length(grep("'",unit.name))==0) {

                     #get values
                     unit.values<-dbGetQuery(con,paste("select * from
                            ca_",year,".getdata(",rs.id,",'",unit.name,"')
                            where unit_id=",unit.id,";",sep=""))
                     unit.values<-unit.values[,var.col]
                     unit.values<-na.omit(unit.values)
                     unit.count<-length(unit.values)

                     #filter units having between 15 and 350 divisions
                     if (unit.count>15&unit.count<350) {

                            #calculate ranges (or whatever needed)
                            unit.r<-max(unit.values)-min(unit.values)
                            unit.iqr<-IQR(unit.values, na.rm=TRUE, type=7)

                            #append results to results matrix
                            res.mat<-rbind(res.mat,
                                   c(unit.name, unit.count, unit.r, unit.iqr)
                                   )

                            } #end if

                     } #end if

              } #end territory unit loop
```

```
        #build graphs from the obtained results matrix

        #histogram of simple range
        hist(as.numeric(res.mat[,3]), col="blue"
               ,ylim=c(0,rs_ylims[rs.id+2]), breaks=5*(0:20)
               ,cex.main=2.8, cex.axis=2, cex.lab=2
               ,xlab=paste(var.name,"range"), ylab="Frequency"
                      ,main=rs_namelist[rs.id-1]
               )

        #scatterplot of simple range vs. amount of observations
        plot(as.numeric(res.mat[,2])~as.numeric(res.mat[,3]), pch=18
               ,cex=2, cex.main=2.8, cex.axis=2, cex.lab=2
               ,xlim=c(0,100), ylim=c(0,350)
               ,xlab=paste(var.name,"range"),ylab="Number of observations"
               ,main=""
               )
        out<-lm(as.numeric(res.mat[,2])~as.numeric(res.mat[,3]))
        abline(out)

        #histogram of the IQR
        hist(as.numeric(res.mat[,4]), col="blue"
               ,breaks=5*(0:20), ylim=c(0,rs_ylims[rs.id-1])
               ,cex.main=2.8, cex.axis=2, cex.lab=2
               ,xlab=paste(var.name,"IQR"),ylab="Frequency",main=""
               )

        #scatterplot of IQR vs. amount of observations
        plot(as.numeric(res.mat[,2])~as.numeric(res.mat[,4]), pch=18
               ,cex=2, cex.main=2.8, cex.axis=2, cex.lab=2
               ,xlim=c(0,30), ylim=c(0,350)
               ,xlab=paste(var.name,"IQR")
               ,ylab="Number of observations", main=""
               )
        out<-lm(as.numeric(res.mat[,2])~as.numeric(res.mat[,4]))
        abline(out)

        rs.id<-rs.id+1

        } #end aggregation level loop

#close the file writing session
dev.off()

} #end function rgs

#execution example
rgs("con",2011)

#prepare the variable space for "rgs" function
vars<-c("turn","con","dem","lib")
years<-c(2006,2008,2011)

#execute "rgs" function with all combinations of variables
for (v in 1:length(vars)){
        for (y in 1:length(years)){
               rgs(vars[v],years[y])
        }
}
```