Brazilian Presidential Elections: Analysing Voting Patterns in Time and Space Using a Simple Data Science Pipeline

L. H. M. Jacintho[1], T. P. da Silva[1], A. R. S. Parmezan[1], G. E. A. P. A. Batista[2]

[1] Universidade de São Paulo, Brazil
{lucasmantovani, tpinho, parmezan}@usp.br
[2] University of New South Wales, Australia
gbatista@cse.unsw.edu.au

Abstract. Since 1989, the first year of the democratic presidential election after a long period of a dictatorship regime, Brazil conducted eight presidential elections. This period was marked by short and long-term shifts of power and two impeachment processes. Such instability is a case of study in electoral studies, *e.g.*, the study of the population voting behavior. Understanding patterns in the population behavior can give us insight into factors and influences that affect the quality of democratic political decisions. In light of this, our paper focuses on analyzing the Brazilian presidential election voting behavior across the years and the Brazilian territory. Following a data science pipeline, we divided the analysis process into five steps: (i) data selection; (ii) data preprocessing; (iii) identification of spatial patterns, in which we seek to understand the role of space in the election results using spatial autocorrelation techniques; (iv) identification of temporal patterns, where we investigate similar trends of votes over the years using a hierarchical clustering method; and (v) evaluation of the results. It is noteworthy that the data in this work represents the election results at the municipal level, from 1994 to 2018, of the two most relevant parties of this period: the Brazilian Social Democracy Party (PSDB) and the Workers' Party (PT). Through the results obtained, we found the existence of spatial dependence in every electoral year investigated. Moreover, despite the changes in the political-economic context over the years, neighboring cities seem to present similar voting behavior trends.

CCS Concepts: • **Applied computing**;

Keywords: data mining, machine learning, preferential voting, spatio-temporal patterns, voting behavior

1. INTRODUCTION

After the historical period called re-democratization, from 1975 to 1985, Brazil has considered democracy as its political regime. However, it was only in 1989 that the population went to the ballots to vote for their candidates after a long term of a dictatorship regime, inaugurating a period called the New Republic. Since 1989, Brazil has conducted eight presidential elections with short and long-term shifts of power. Considered a new democracy, Brazil is a case of study in the electoral studies, especially in the voting behavior research field [Carvalho and Menezes 2015; Marzagão 2013]. In this area, the primary purpose is to understand how and why the population makes political decisions. Understanding voting behavior is a fundamental key to identifying processes and influences that can affect the quality of democratic decisions.

As a fact, elections are complex processes that can be influenced by several variables, from socioeconomic to geographic factors [Mansley and Demšar 2015]. The last one has been increasingly studied over the years [Mansley and Demšar 2015; Agnew 1996]. The role of space is an important factor in electoral processes, as pointed by [Agnew 1996] in his multidimensional place-centered perspective on political behavior, and later by [Mansley and Demšar 2015]. Regarding the Brazilian elections, there are a few publications that seek to understand the factors that contribute to the outcome of the elections [Carvalho and Menezes 2015; Marzagão 2013]. However, in general, the analyses are punctual, carried out on the data related to the year of interest. Thus, they do not consider the influence of previous elections.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, and the Brazilian National Council for Scientific and Technological Development [grant number 142050/2019-9].

Besides, the reproducibility of the results is made difficult since no data pipeline is described or made available. In this matter, the use of a data science pipeline has been strongly encouraged in many applications, from the stock market to medical purposes, as a tool for knowledge discovery that can ensure results reproducibility [Han et al. 2011]. It consists of five steps: (i) data collection, (ii) data preprocessing, (iii) data exploration, (iv) modeling, and (v) analysis of results. Such an approach is ideal for exploratory analyses as voting behavior discovery, especially in the case of Brazil, where the data available is not well structured.

Following a similar idea, this paper proposes a municipal level analysis of spatial and temporal patterns in the Brazilian presidential results from 1994 to 2018, regarding the two most relevant parties of this period: the Brazilian Social Democracy Party (PSDB) and the Workers Party (PT). To accomplish that and provide results reproducibility, we applied a simple data science pipeline in our methodology. Our goal is to understand the variations of the Brazilian electorate's behavior over the years, considering the spatial domain. Also, a part of this work consists of the production of datasets that can be reused for other analyses. Since this work was developed seeking the experiments' reproducibility, all developed code is available on Github[1] for consultation, use, and modification.

2. RELATED WORK

The analyses of voting behavior have always been a well-grounded research field in political science, and most recently, there has been an increase of interest in analyzing the outcomes of elections over the world [Norris and Grömping 2019]. Not surprisingly, the Brazilian elections are a hot research topic due to many factors that vary from the impeachment processes to increasing political polarization.

The study reported in [Power and Rodrigues-Silveira 2019] presents an ecological analysis of Brazilian elections at the municipal level from 1995 to 2018. They look for ideological alignments over space and time and try to identify four possible explanations regarding the outcomes of the elections, which were: the effect of incumbent alignments, social modernization, political pluralism, and social inclusion. The authors did not identify long periods of electoral realignment when examining local voting in presidential elections.

Using spatial econometrics techniques, the work of [Carvalho and Menezes 2015] analyses the Brazilian presidential elections of 2010 in a national scope. They considered data from the Bolsa Família Program, Gross Domestic Product, and Human Development Index to produce models that calculate their impact on the percentage of votes received by the Workers' Party.

Similar to our study, [Marzagão 2013] tests two alternative hypotheses. The first concerns social interaction between residents of neighboring municipalities, and the second seeks to assess the existence of concentration of electoral campaigns in certain regions.

Finally, the works described here present valuable analyses towards understanding Brazilian voting behavior considering the role of space and time. Our paper differs from the literature by evaluating the election results using techniques to identify spatial autocorrelation and clustering methods. Thus, this paper's results and analyses add more information regarding understanding the Brazilian population's electoral decisions.

3. MATERIAL AND METHODS

This study analyses voting patterns of the Brazilian presidential elections at the municipalities level regarding time and space domains following a simple data science pipeline. Fig. 1 organizes the pipeline in five steps, which we will describe in detail throughout Sections 3.1–3.5.

We implemented the data science pipeline of Fig. 1 using the Python[2] programming language

[1] https://github.com/LucasManto/analyzing_brazil_presidential_elections

[2] https://www.python.org/.

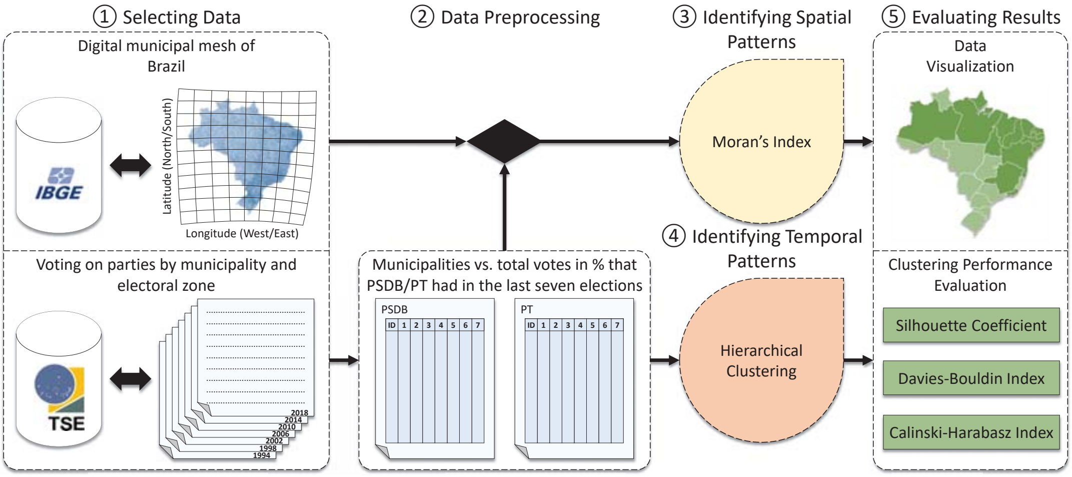


Fig. 1: Pipeline of the proposed analysis. The acronyms are: Brazilian Institute of Geography and Statistics (IBGE), Superior Electoral Court (TSE), Brazilian Social Democracy Party (PSDB), and Workers' Party (PT).

combined with the following libraries: GeoPandas[3], SciPy[4], PySAL[5], and scikit-learn[6]. We also adopted the Cookiecutter Data Science framework that provides guidelines to build reproducible projects[7]. All our code is available on the Github platform[8].

3.1 Selecting Data

Most democratic countries make available the results right after the electoral process is over and maintain historical data from previous elections. Usually, there is enough information so that researchers can explore it within the context of the results. In Brazil, the government agency responsible for making election data available is TSE[9].

From the TSE data repository, we selected seven datasets related to the elections of 1994, 1998, 2002, 2006, 2010, 2014, and 2018. We need to highlight that electoral data before 1994 were only available at the state level, and since our analyses concern data at the municipalities level, we decided to discard them. Each selected dataset is organized in tables, where the lines represent an electoral zone of a certain municipality, and the columns are attributes that describe different characteristics of the zone.

As our analysis also requires geographical information regarding areas and geographical coordinates of Brazilian municipalities, we considered the digital meshes provided by IBGE[10]. To join the electoral data with the digital meshes, we use a dataset that relates the municipalities IDs assigned by IBGE with the IDs assigned by TSE.

3.2 Data Preprocessing

We arranged the data preprocessing step into four processes: (1) filtering raw data, (2) aggregation at municipality level, (3) transformation of vote counts into vote shares, and (4) filtering by parties.

[3] https://geopandas.org/.
[4] https://www.scipy.org/.
[5] https://pysal.org/
[6] https://scikit-learn.org/stable/.
[7] https://drivendata.github.io/cookiecutter-data-science/.
[8] https://github.com/LucasManto/analyzing_brazil_presidential_elections.
[9] http://www.tse.jus.br/eleicoes/estatisticas/repositorio-de-dados-eleitorais-1/.
[10] https://mapa.ibge.gov.br/bases-e-referencial/bases-cartograficas/malhas-digitais/.

First, we filtered the datasets to keep only the rows where the attribute CD_CARGO had the value 1, which corresponds to the position of president, and the attribute NR_TURNO had the value 1, resulting in seven presidential election datasets with the first-round results. Note that we chose only the first-round results to have the data for all years since some election years did not have a second round. Next, we aggregated the electoral results by municipalities summing the values of the attribute $QT\_VOTOS\_NOMINAIS$. In this way, we obtained seven datasets where each row represents a municipality. After, we calculate the vote share for each party concerning the municipalities. Finally, we filtered the datasets to keep only the rows where the attribute $SG\_PARTIDO$ has the acronym PT (Workers' Party) and PSDB (Brazilian Social Democracy Party). Further, we separated each party's rows and concatenated them, creating two datasets where the rows denote the municipalities, and the attributes are the vote share that the party had in the election years (Table I).

CD_MUNICIPIO	**1994**	**1998**	**2002**	**2006**	**2010**	**2014**	**2018**
72150	0.653289	0.724919	0.525627	0.663776	0.516710	0.624789	0.128623
92134	0.746534	0.635629	0.346545	0.528929	0.260935	0.372672	0.040883
12190	0.341180	0.297991	0.116617	0.276511	0.222201	0.188364	0.029210
28070	0.830481	0.684308	0.232916	0.202928	0.242993	0.171235	0.026018
71030	0.596187	0.527083	0.225716	0.694932	0.399485	0.524107	0.118832

Table I: Sample of the final dataset. In this case, 5 random examples were extracted from the PSDB dataset.

3.3 Identifying Spatial Patterns

In order to identify spatial patterns, we calculated the global autocorrelation for each election year to assess the continuity of Spatial Autocorrelation over the years. We considered Moran's Index, one of the most popular methods to generate a global measure of spatial autocorrelation [Li et al. 2007]. Its results vary from -1 to 1, where -1 indicates a negative spatial dependence, 1 relates to a positive autocorrelation, and 0 indicates spatial randomness. Equation 1 shows how the Moran's Index is measured, where n is the number of locations, x_i is an observed value at location i, $\overline{x}$ is the average of x, w_{ij} is a weight between locations i and j, and S_0 is the squared sum of all weights. In our analysis, the weights were calculated using a strategy called Queen, where, for each municipality, neighboring municipalities receive a value of 1 while the others receive a value of 0.

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{1}$$

Global methods, however, only indicate evidence of spatial dependence. To identify spatial patterns in our datasets, we need to apply a local spatial correlation method. For this paper, we considered the Local Moran's Index [Anselin 1995]. The results generated by this method have a similar interpretation as the method that it was inspired. That is, a location with a positive measure presents neighbors with similar values, while locations with negative indexes indicate that their neighbors have different values. Equation 2 presents how the Local Moran's Index is calculated, where n is the number of locations, x_i is an observed value at location i, $\overline{x}$ is the average of x, w_{ij} is a weight between locations i and j, and S_i^2 is calculated by Equation 3. The weights were calculated following the Queen strategy, as in the global Moran's Index.

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

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

3.4 Identifying Temporal Patterns

Intending to identify temporal patterns, we applied a clustering algorithm in the party datasets. It is a valid strategy since each row of our datasets represents a time-series of the percentage of votes that a given party received in Brazilian municipalities.

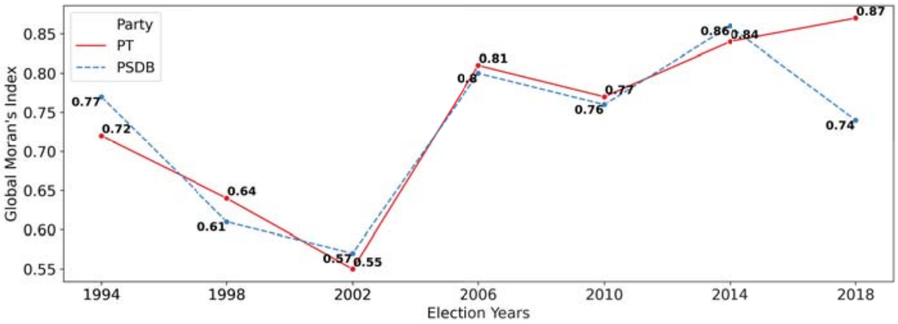


Fig. 2: Global Moran's Index values for each election year.

Clustering algorithms can be categorized according to their criteria to form groups [Rokach and Maimon 2005]. There are centroid-based, density-based, and connectivity-based methods (hierarchical clustering algorithms). Due to the exploratory nature of this work and the lack of information regarding the number of groups to be formed, we employed a hierarchical clustering algorithm with the Euclidean distance and the Ward method as the clustering criterion. At each iteration, the Ward method connects the groups with the smallest increase in their internal variance after the join [Jr. 1963].

3.5 Evaluating Results

We evaluated the results of this paper from two perspectives: (1) data visualization, *i.e.*, we produced graphic representations of the special patterns found in the election data; and (2) clustering performance evaluation, in which we evaluated the quality of the groups formed in order to identify temporal patterns in the data election data. In this context, we evaluated the quality of the clusters formed according to the following metrics: silhouette coefficient [Rousseeuw 1987], Davies-Bouldin index [Davies and Bouldin 1979], and Calinski-Harabasz index [Caliński and Harabasz 1974].

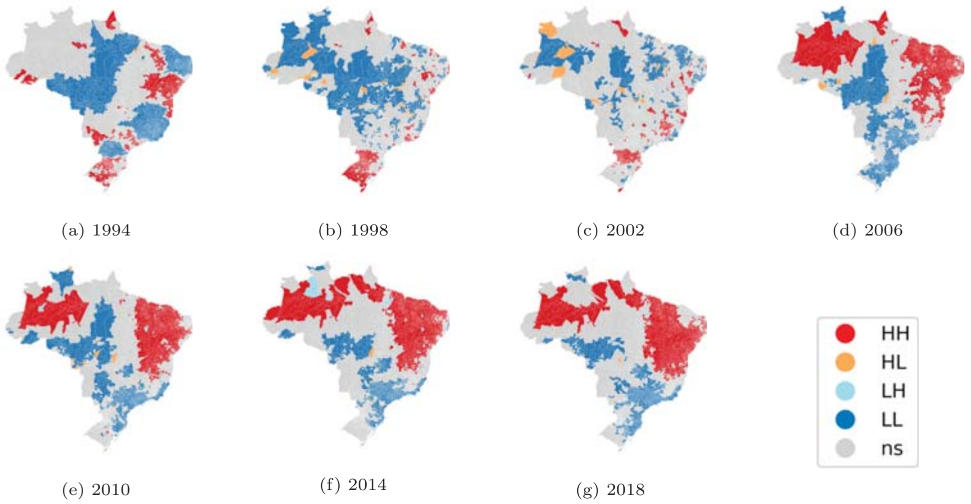
4. RESULTS AND DISCUSSION

We divided the analyses and evaluation of the results into two steps. First, seeking to identify spatial dependence, we discuss the results of techniques to measure spatial autocorrelation, globally and locally, considering each electoral year. Next, we evaluate and discuss the hierarchical clustering results obtained from the parties' vote shares time-series of each Brazilian municipality to identify temporal patterns. The results and discussions for each step are described in what follows.

4.1 Analysing Spatial Patterns

A first assessment of the Global Moran's Index (2) shows that since 1994 the Brazilian presidential elections present evidence of clustered distributions over space, with a decay in the first years and an increase after the 2002 election, reaching values greater than 0.8 for both parties in 2014 and 2018. The election of 2002 marks the end of the PSDB Government and PT's ascension as the incumbent party. Not surprisingly, there is a drop in Moran's Index for both parties at this period. These values indicate the dispersion of previous years clustered distributions over the Brazilian territory, which can be seen as a raise of uncertainty concerning the Brazilian voters. However, only the values of Global Moran's Index are not sufficient to confirm spatial patterns and locate where they occur. To accomplish that we calculate the Local Moran's Index for all the municipalities of Brazil in each year (3, 4).

For a better understanding of the Local Moran's Index results, we plot them for each municipality. In this way, cities with positive local spatial autocorrelation, values closer to 1, are represented by the colors red and blue, where the red (HH) are cities with a high percentage of votes for the party



(a) 1994 (b) 1998 (c) 2002 (d) 2006

(e) 2010 (f) 2014 (g) 2018

Fig. 3: Local Moran's Index plot of *Workers Party* (PT) by election year.

surrounded by cities with an also high percentage of votes for the party. On the other hand, the blue regions (LL) represent cities with a low percentage of votes for the party surrounded by cities with an also low percentage of votes for the party. The light blue and orange cities can be seen as outliers. They are municipalities where the local spatial autocorrelation had negative values closer to -1. The light blue (LH) are the municipalities where the party presented a low percentage of votes, and the neighboring cities had a higher percentage of votes. The orange cities (HL) follow the opposite behavior. Finally, the cities in gray (ns) are those where the index value was closer to 0, indicating randomness in the spatial distribution.

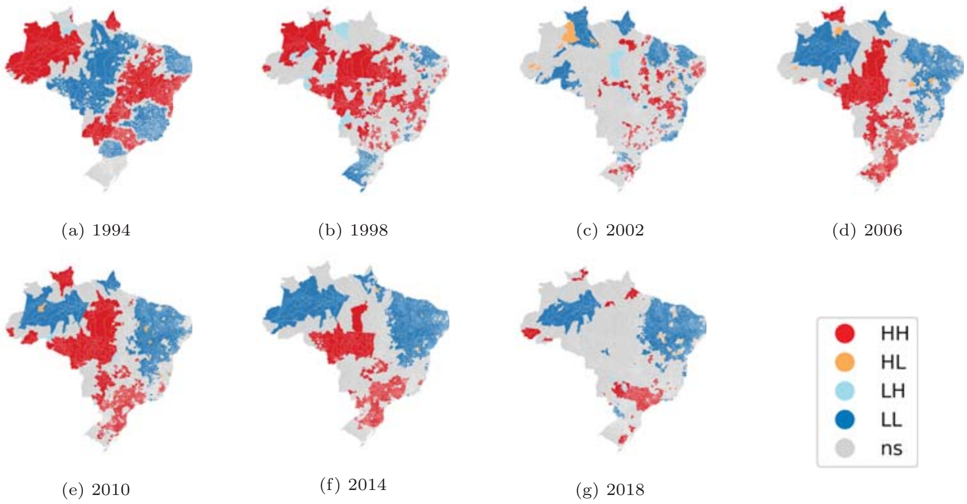
Comparing the local spatial autocorrelation plots from 1998 to 2002 of PSDB (4), it is possible to understand what caused the drop in the Global Moran's Index in 2002. Since 1994 PSDB presented a decrease of hegemony, with a considerable number of cities going from red to gray or even blue in 2002. The same phenomenon occurs inversely for PT (3), the number of blue points decreased, and the number of gray points increased, indicating a disperse growth.

From 2006 onward, the number of cities highlighted in red increased, mostly in North and Northeast regions, contributing to the global Moran's Index growth. Comparing the plots from 2014 (3f and 4f), when both parties were the most voted, the maps are almost the inverse of each other, that is, cities, where the highlight color is blue, are red on the other. Finally, it is possible to visualize that despite all the political-economic context changes, the regions kept a constant voting behavior until the most recent presidential election.

4.2 Analysing Temporal Patterns

To confirm this evidence and identify the cities with similar voting behavior across time, we run a hierarchical clustering method with the time-series of votes shares each party received in all the cities as input. To identify the best number of groups to analyze, we evaluate the results from 2 to 10 groups considering the metrics: Silhouette, Calisnk-Harabasz, and Davies-Bouldin. Table II exhibit the values of the evaluation metrics obtained for each party. In general, the best results – in bold – were achieved by considering two groups for both parties. Thus, from now on, we will focus our analyses considering the clustering results of two groups.

From 5a and 6a, it is possible to identify a spatial characteristic in the clustering results, even though no spatial information was given. In these figures, cities belonging to the same group present the same color. The results indicate that neighboring cities in some regions of Brazil presented similar voting behavior over the years. For instance, considering the results for PSDB (5a), the majority of



(a) 1994 (b) 1998 (c) 2002 (d) 2006

(e) 2010 (f) 2014 (g) 2018

Fig. 4: Local Moran's Index plot of *Brazilian Social Democracy Party* (PSDB) by election year.

	PSDB			PT		
Clusters	Silhouette	Calinski Harabasz	Davies Bouldin	Silhouette	Calinski Harabasz	Davies Bouldin
2	**0.36**	**3367.91**	1.20	**0.38**	**4746.80**	**1.02**
3	0.34	2969.28	**1.01**	0.30	3166.98	1.17
4	0.30	3028.89	1.14	0.22	2743.53	1.46
5	0.29	2810.47	1.16	0.21	2379.85	1.49
6	0.27	2592.52	1.21	0.17	2162.90	1.46
7	0.25	2390.76	1.19	0.15	2013.25	1.43
8	0.22	2255.85	1.31	0.16	1920.13	1.40
9	0.19	2131.52	1.37	0.14	1812.03	1.58
10	0.19	2030.41	1.39	0.15	1702.22	1.50

Table II: Clustering evaluation metrics results

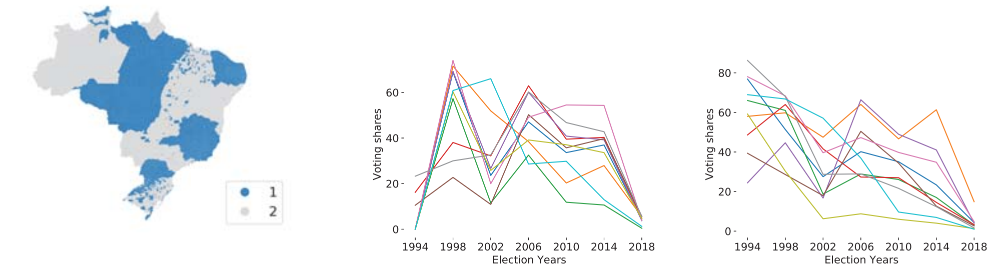
the southeast cities belong to the same group. On the other hand, considering the results for PT (6a), almost every city of the northeast region belong to the same group.

In more detail, 5b and 5c present randomly chosen samples of PSDB vote shares time-series from cities belonging to groups 1 and 2, respectively. The series from group 1 presents a low percentage of votes in 1994, followed by a peak in sequential years with a decreasing trend after. In group 2, as the opposite, the series starts with a high percentage of votes in 1994, followed by a decreasing trend with some peaks between 2006 and 2014. Differently from PSDB, the PT vote shares time-series for group 1 (Fig. 6b) features an increasing trend from 1994 to 2006, stabilization in 2010 and 2014, and a decrease in 2018. For group 2 (Fig. 6c), the series begins in an increasing trend as well, but with a decrease in 2006, resuming growth in 2010, and decreasing again in 2014 and 2018.

Finally, as shown throughout the analysis from a municipality perspective, the Brazilian population shows a related voting behavior in a spatial and temporal aspect. In other words, voting trends in one party are usually followed by neighboring cities. Such characteristic generates spatial clusters, with different vote distribution over the regions.

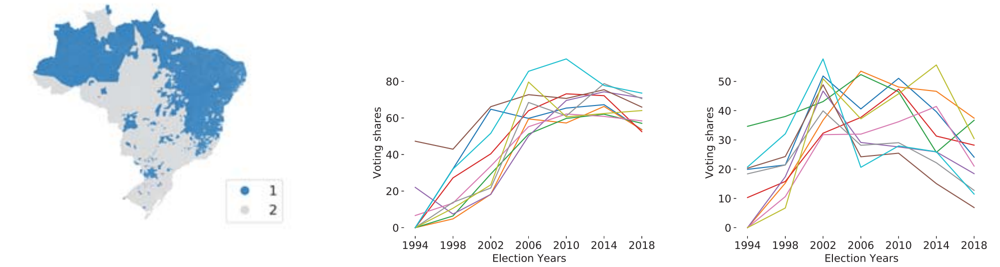
5. CONCLUSION

This paper presents additional efforts to understand the role of space in Brazilian voters' behavior and assesses the maintenance of the voting patterns found over the years. We applied a simple data science pipeline to identify and evaluate the Brazilian presidential election's spatial and temporal patterns from 1994 to 2018 at a municipal level. From the analysis of spatial autocorrelation, we identified spatially cluster distribution, which corroborates the hypothesis that neighboring cities are more likely to present similar voting behavior. Furthermore, when analyzing the hierarchical clustering results, we found that neighboring cities similarly change their electoral behavior.



(a) Clustering results for 2 groups (b) Group 1 sample series (c) Group 2 sample series

Fig. 5: PSDB clustering map and voting series samples



(a) Clustering results for 2 groups (b) Group 1 sample series (c) Group 2 sample series

Fig. 6: PT clustering map and voting series samples

The experiments held in this study were done to build more knowledge in the analyses of Brazilian elections. Our findings can be used as the basis for developing Machine Learning models, whose objective is to understand the electoral behavior. Also, part of this work consisted of preparing datasets that can be reused for further analysis.

REFERENCES

Agnew, J. Maps and models in political studies: a reply to comments. *Political Geography* 15 (2): 165–167, 1996.

Anselin, L. Local indicators of spatial association—lisa. *Geographical analysis* 27 (2): 93–115, 1995.

Caliński, T. and Harabasz, J. A dendrite method for cluster analysis. *Communications in Statistics* 3 (1): 1–27, 1974.

Carvalho, R. and Menezes, T. Uma análise espacial das eleições presidenciais brasileiras de 2010. *Pesquisa e Planejamento Econômico* 45 (3): 436–495, 02, 2015.

Davies, D. L. and Bouldin, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1 (2): 224–227, 1979.

Han, J., Kamber, M., and Pei, J. *Data mining: concepts and techniques*. Morgan Kaufmann, California, 2011.

Jr., J. H. W. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58 (301): 236–244, 1963.

Li, H., Calder, C. A., and Cressie, N. Beyond moran's i: testing for spatial dependence based on the spatial autoregressive model. *Geographical Analysis* 39 (4): 357–375, 2007.

Mansley, E. and Demšar, U. Space matters: Geographic variability of electoral turnout determinants in the 2012 london mayoral election. *Electoral Studies* vol. 40, pp. 322–334, 2015.

Marzagão, T. A dimensão geográfica das eleições brasileiras. *Opinião Pública* 19 (2): 270–290, 2013.

Norris, P. and Grömping, M. Electoral integrity worldwide, 2019. Sydney: Electoral Integrity Project. Available at https://www. dropbox. com/s/csp1048mkwbrpsu/Electoral% 20Integrity% 20Worldwide. pd f.

Power, T. J. and Rodrigues-Silveira, R. Mapping ideological preferences in brazilian elections, 1994-2018: a municipal-level study. *Brazilian Political Science Review* 13 (1): e0001–1–27, 2019.

Rokach, L. and Maimon, O. Clustering methods. In *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach (Eds.). Springer, Boston, pp. 321–352, 2005.

Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* vol. 20, pp. 53 – 65, 1987.