Analyzing Spatio-Temporal Voting Patterns in Brazilian Elections Through 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] University of 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. Short and long-term shifts of power and two impeachment processes marked such a period. These instabilities are a research case in electoral studies, mainly regarding the understanding of citizens' voting behavior. Comprehending patterns in the population behavior can give us insight into phenomena and processes that affect democratic political decisions. In light of this, our article analyzes Brazilian electoral data at the municipal level from 1998 to 2018 using a simple data science pipeline, which consists of five steps: (i) data selection; (ii) data preprocessing; (iii) identification of spatial patterns, where we seek to understand the role of space in the election results employing spatial autocorrelation techniques; (iv) identification of temporal patterns, in which we explore similar trends of votes over the years applying a clustering method; and (v) evaluation of results. We study the presidential elections focusing on the most relevant left and right parties for the period: Workers' Party (PT) and the Brazilian Social Democracy Party (PSDB). We also analyze the congressman election data concerning parties ideologically to the left and right in the political spectrum. From the obtained results, we found the existence of spatial dependence in every electoral year investigated. Furthermore, despite the changes in the political-economic context over the years, neighboring cities presented similar voting behavior trends.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications; J.4 [**Computer Applications**]: Social and Behavioral Sciences

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

1. INTRODUCTION

The constitution promulgated during the New Republic represents an important moment of redemocratization in Brazil after the civil-military dictatorship from 1964 to 1985. Citizens, however, only went to the polls to vote for their favorite candidates in 1989. Since then, Brazil has held eight presidential elections with short and long-term changes in power.

The establishment of a new democracy was accompanied by a strong political crisis and a hyperinflationary process whose effects permeate modern Brazilian society. Aspects like these make Brazil a recurrent research case in electoral studies, especially in the electoral behavior field [Carvalho and Menezes 2015; Marzagão 2013]. One of the goals of this area is to understand how and why the population makes political decisions. Analyzing electoral behavior is essential to identify phenomena and processes that can affect the quality of democratic decisions.

Elections are complex activities that can be influenced by several variables, including socioeconomic and geographic factors [Mansley and Demšar 2015]. The last one has been increasingly studied over the years [Mansley and Demšar 2015; Agnew 1996]. The geographical space, in turn, plays a key role in electoral processes, as pointed out initially by Agnew in his multidimensional place-centered perspective on political behavior [Agnew 1996] and later by Mansley and Demšar [Mansley and Demšar 2015]. Regarding Brazilian elections, the literature covers a few pieces of work that seek to understand the aspects that contribute to the outcome of the elections [Carvalho and Menezes 2015; Marzagão

Copyright©2021 Permission to copy without fee all or part of the material printed in JIDM is granted provided that the copies are not made or distributed for commercial advantage, and that notice is given that copying is by permission of the Sociedade Brasileira de Computação.

2013]. Because the analyzes are conducted on the data related to the year of interest, such studies are punctual and do not consider previous elections' influence. Furthermore, most of these papers do not provide a detailed description of the processes used to obtain the results, impairing reproducibility.

Reproducibility is an important criterion for assessing the quality of research and consistency of results. In the face of big data, we can easily meet such a criterion by applying systematic methodologies for knowledge extraction and intelligent data analysis [Han et al. 2011]. In this article's context, a data science pipeline is strongly encouraged as it has been vastly employed in general applications, from the stock market to medical purposes [Hand and Adams 2014]. A common data science pipeline generally comprises five steps: (i) data collection, (ii) data preprocessing, (iii) data exploration, (iv) analytical modeling, and (v) analysis of results. They are ideal for exploratory investigations, especially those with a lack of well-structured data, which is Brazil's case.

A preliminary version of this work is described in [Jacintho et al. 2020]. Therein, we explored the Brazilian presidential election results for the two most relevant parties from 1994 to 2018, searching for spatial and temporal patterns. Here, we extend the data science pipeline introduced in the referred paper to analyze the congressman election results for left and right parties compared to results at the presidential level. Our goal is to learn whether patterns similar to those from presidential elections are found in time and space.

The contributions of this article are listed as follows:

—We conducted a broad literature review followed by a meta-analysis of papers published in the past ten years covering Brazilian elections;

—We designed a flexible and straightforward social data science pipeline and proposed using it to collect, preprocess, and analyze spatial and temporal patterns in Brazilian elections;

—We provided access to the datasets and codes to ensure the reproducibility of our results. They are available on Github[1] for the community to review and inspect;

—From the perspective of spatial autocorrelation, our results revealed that neighboring municipalities tend to vote similarly with each other. This finding, in turn, proved to be consistent over time in terms of temporal dependence;

—To the best of our knowledge and research in the literature, our study is a pioneer in applying machine learning techniques to analyze Brazilian elections' temporal patterns.

The rest of this article is organized as follows: Section 2 reports a comprehensive review of the literature along with a meta-analysis of related work; Section 3 describes our data science pipeline for analyzing spatio-temporal voting patterns in Brazilian elections; Section 4 compiles the results and discusses them; finally, Section 5 concludes the study with some comments on future research directions.

2. RELATED WORK

Voting behavior analysis has always been a well-grounded research field in political science. In recent years, however, we have witnessed an increased interest in analyzing the outcomes of elections worldwide due to advancements in technologies that allow easy access to electoral data [Norris and Grömping 2019]. Thereby, researchers from the statistics and social science communities have contributed to several aspects of voting behavior analysis, such as understanding external factors that may influence electoral results [Hernández and León 2020; Okunev et al. 2020; Reid and Liu 2019], the study of the role of space in elections [Mota 2019; Mansley and Demšar 2015], the analysis of ideological trends [Zucco and Power 2020; Faustino et al. 2019; Power and Rodrigues-Silveira 2019], and the voting behavior on social media [Recuero et al. 2020; Praciano et al. 2018].

This article focuses on understanding the role of space and time in Brazilian elections. To provide an overview of related work, we performed a comprehensive review of the literature followed by a

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

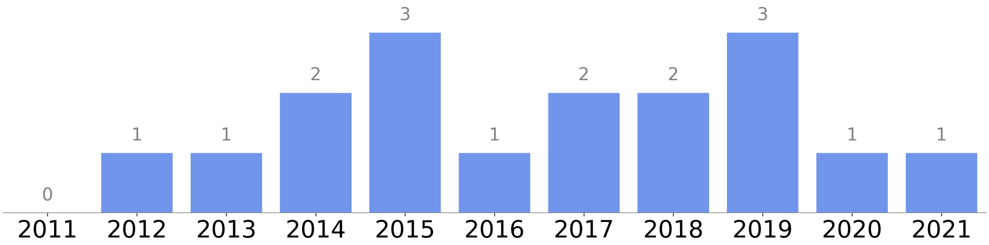
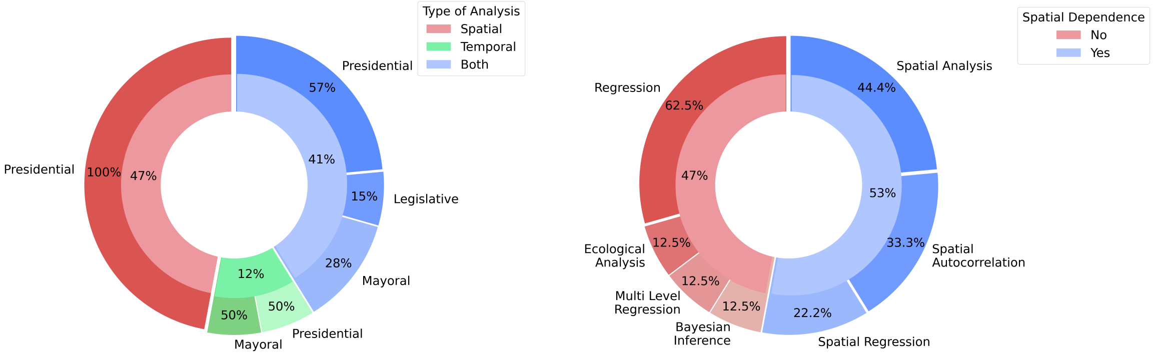


Fig. 1: Distribution of selected publications by year.



(a) Publications by type of analysis.

(b) Publications that regard spatial dependence.

Fig. 2: Overview of the meta-analysis results.

meta-analysis. We defined a global search expression[2] and submitted it to six electronic databases[3] to retrieve pieces of work published in the last ten years. We adopted two criteria for selecting publications: (i) the study must cover election results provided by the Brazilian Superior Electoral Court, and (ii) the analysis should explore or at least consider the role of space or time in the data. Following this bibliographic review protocol, we identified 17 papers that meet all predefined criteria.

Fig. 1 displays a distribution bar chart, by publication year, concerning the 17 selected pieces of work. In this figure, from January 2012 to February 2021, there was at least one publication correlated to this article per year. Furthermore, the two peaks (2015 and 2019) may be related to the 2014 and 2018 presidential elections. We need to emphasize that this amount of published work reflects the scarcity of studies that evaluate Brazilian elections taking into account the role of space and time.

The inspection of the 17 papers identified by our bibliographic review enabled the elaboration of a meta-analysis that aimed to answer which elections the studies addressed and whether they involved spatial analysis, temporal analysis, or both. Fig. 2a exhibits the relative frequency of both the type of analysis and the type of election scrutinized per analysis. We can see in this figure the predominance of spatial analysis regarding presidential elections, indicating that most related papers did not deal with legislative and municipal elections. Fig. 2b shows the relative frequency of pieces of work that considered or not the electoral data's spatial dependence [Schuhli 2018]. This figure also includes the relative frequency of the approaches taken by each of these two groups of studies. By looking at such information, we can observe that almost half of the selected publications disregarded the spatial dependence of the electoral data, thus compromising the scientific rigor of their analysis. Finally, Fig. 3 displays the number of times that the identified papers addressed a given electoral year. In addition

[2] Search string: "brazilian elections" **AND** ("ESDA" **OR** "spatio temporal analysis" **OR** "spatial analysis" **OR** "ecological analysis" **OR** "spatial econometrics" **OR** "spatial autocorrelation" **OR** "regional voting patterns" **OR** "clustering" **OR** "data science" **OR** "voting behavior").

[3] Sources: ACM Digital Library, DBLP Computer Science Bibliography, Google Scholar, Scopus, IEEE Xplore, and Web of Science.

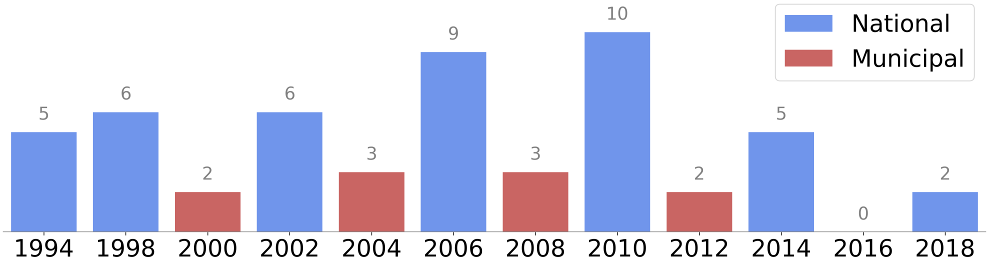


Fig. 3: Number of scientific analyzes per election year.

to comprising two peaks corresponding to the second election of President Lula (2006) and the first of President Dilma (2010), this figure reaffirms the interest in national elections rather than municipal ones. The disinterest in municipal elections is because they are pulverized across the territory and have different electoral disputes in terms of candidates and parties. Notably, analyzing data like these from a global perspective is a challenging task. It is also important to highlight that we did not identify any study on the 2016 Brazilian elections.

From the 17 papers selected at the end of the bibliographic review, we delved into six because they are strictly related to the purpose of this article. The research described in [Schuhli 2018] studied the spatial patterns in the 2010 election and the impact of catholic voting employing spatial autocorrelation techniques. Similarly, [Martins et al. 2016] applied spatial regression analysis to understand the factors that influenced the 2014 presidential election. Using spatial econometrics techniques, [Carvalho and Menezes 2015] and [Magalhães et al. 2015] analyzed the Brazilian presidential elections of 2010 in a national scope. They considered data from the *Bolsa Família Program*[4], Gross Domestic Product, and Human Development Index to build models that calculate their impact on the percentage of votes received by the Workers' Party. Likewise, adopting spatial autocorrelation techniques and regression analysis, [Corrêa 2015] investigated the impacts of the *Bolsa Família* Program in the 2016 presidential election. Finally, [Marzagão 2013] searched for spatial patterns in the 2010 presidential election. The authors tested two alternative hypotheses. The first one guided the understanding of social interaction between residents of neighboring municipalities. The second one sought to assess the existence of concentration of electoral campaigns in certain regions.

We take the liberty of adding to the six research pieces reported in the past ten years, the following study published in 2010: [Terron and Soares 2010]. Such a paper appears to be the first to investigate Brazilian presidential elections' spatial and temporal patterns employing spatial autocorrelation and regression techniques. Although the authors considered only presidential elections, they explored temporal dependency via regression analysis from 1994 to 2006. In contrast to them, we analyze the data distribution by applying clustering techniques to voting time series.

Table I compares our study with the seven most similar publications found. This comparison uses the following criteria: the electoral year(s) analyzed, whether the paper provided spatial analysis, whether there were any temporal analyses, if the analysis pipeline is made available, and if the dataset is made available. In general, most of the pieces of work assessed a single election, providing only spatial analyses. Although [Terron and Soares 2010] made a temporal analysis of the Brazilian elections, the verified period was short, and the authors examined the electoral years independently. Moreover, almost all the papers do not provide their evaluation pipeline, nor the dataset(s) created.

As summarized in the last row of Table I, our proposal differs from the literature not only on the number of electoral years scrutinized and the temporal analysis adopted but also by using machine learning methods and making publicly available the datasets and source codes necessary to reproduce our results. Such differences were made possible due to applying a data science pipeline that automates the decision-making process, from the preparation of datasets to their analysis.

[4] *Bolsa Família* (Family Allowance) is a social welfare program of the Government of Brazil, part of the *Fome Zero* network of federal assistance programs. *Bolsa Família* provides financial aid to poor Brazilian families; and if they have children, families must ensure that the children attend school and are vaccinated.

Table I: Properties of the most similar related work.

Paper	Electoral Year(s)	Spatial Analysis	Temporal Analysis	Pipeline Available	Dataset(s) Available
[Schuhli 2018]	2010	✓	—	—	✓
[Martins et al. 2016]	2014	✓	—	—	—
[Carvalho and Menezes 2015]	2006 - 2010	✓	—	—	—
[Magalhães et al. 2015]	2010	✓	—	—	—
[Corrêa 2015]	2006	✓	—	—	—
[Marzagão 2013]	2010	✓	—	—	—
[Terron and Soares 2010]	1994 - 2006	✓	✓	—	—
This paper	1998 - 2018	✓	✓	✓	✓

3. MATERIAL AND METHODS

This study analyzes voting patterns of the Brazilian presidential elections at the municipalities level concerning time and space domains following a simple data science pipeline. Fig. 4 organizes the pipeline in five steps, which we will describe in detail throughout Sections 3.1–3.5.

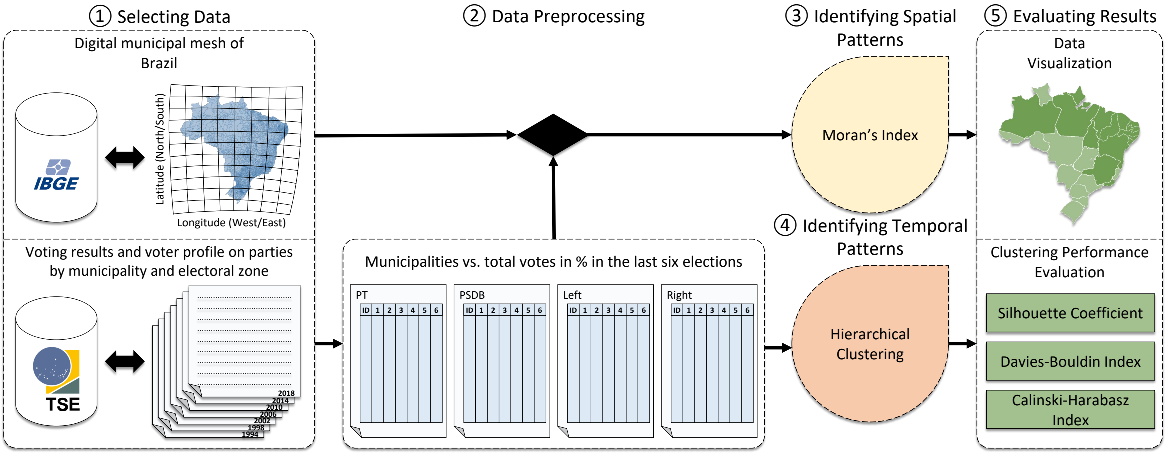


Fig. 4: Pipeline of the proposed analysis. The acronyms are: Brazilian Institute of Geography and Statistics (IBGE), Superior Electoral Court (TSE), Workers' Party (PT), and Brazilian Social Democracy Party (PSDB).

Table II: Python packages used in the pipeline implementation.

Step	Pandas	GeoPandas	PySal	ScyPy	Scikit-learn
Selecting data	✓	—	—	—	—
Data preprocessing	✓	✓	—	—	—
Identifying spatial patterns	✓	✓	✓	—	—
Identifying temporal patterns	✓	✓	✓	✓	✓
Evaluating results	✓	✓	✓	✓	✓

We implemented the data science pipeline of Fig. 4 employing the Python[5] programming language

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

combined with the following libraries: Pandas[6], GeoPandas[7], SciPy[8], PySAL[9], and Scikit-learn[10]. Table II summarizes the technologies applied at each step of the pipeline. We also adopted the Cookiecutter Data Science framework, which provides guidelines to build reproducible projects[11]. Our codes and supplementary material are available on the Github platform[12].

3.1 Selecting Data

Most democratic countries make available the voting results right after the electoral process is over and maintain historical data from previous elections. Usually, there is enough information, so researchers can explore it within the context of the results. In Brazil, the government agency responsible for making election data available is TSE[13].

For this research, and in agreement with Step 1 of the pipeline outlined in Fig. 4, we selected election data from 1998 to 2018. Elections before 1994 were discarded because their data were organized at the state level and thus could not be of use on our municipality-level analysis. Also, data from the elections of 1994 was removed since it presents a high percentage of missing data—around 50% of municipalities do not have any data. The collected datasets are tables with the columns being attributes that describe the zone characteristics and rows representing an electoral zone of a certain municipality. The number of columns and rows varies a little from year to year, but the dataset from 2018, for instance, presents 28 columns and 590,530 rows. Table III exhibits the attributes addressed in the present study.

Table III: Attributes of interest.

Attribute	**Description**
NR_TURNO	Round number
SG_UF	Federal unit abbreviation
CD_MUNICIPIO	Municipality identifier
CD_CARGO	Political office identifier
SG_PARTIDO	Party acronym
QT_VOTOS_NOMINAIS	Number of votes received by the party

As our analysis also requires geographical information regarding areas and geographical coordinates of Brazilian municipalities, we considered the digital meshes—data that describes the municipalities borders and location represented by polygons—provided by IBGE[14]. To join the electoral data with the digital meshes, we used a dataset that relates the municipalities IDs assigned by IBGE with the IDs assigned by TSE.

3.2 Data Preprocessing

Our data preprocessing consists of four processes (Step 2 of Fig. 4): (i) filtering raw data; (ii) aggregation at the municipality level; (iii) conversion of vote counts into percentage of votes received by the party, which we will reference as vote-shares; and (iv) selecting data by parties. First, we filtered raw data concerning the political offices for each electoral year. In this article, we focus on the presidential and legislative elections for congress. Regarding presidential elections, we kept the rows where the attribute `CD_CARGO` had the value equals 1, while for the congressmen's office, the rows with value 6 were selected. Next, we aggregated the data by municipalities summing the attribute `QT_VOTOS_NOMINAIS` for presidential election data and the attributes `QT_VOTOS_NOMINAIS`

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

and QT_VOTOS_LEGENDA for congress election data. QT_VOTOS_LEGENDA are non-nominal votes when the voter chooses a party instead of a specific candidate. The results are datasets for each electoral year with the vote counts aggregated per city. In the next step, we converted the vote counts into vote-shares based on the turnout, *i.e.*, the attribute QT_COMPARECIMENTO. Subsequently, we selected the data by parties or groups of parties. The final datasets were then concatenated to represent the party's vote-shares over the years. If selected a group of parties, the final dataset represents the group's total vote-shares over the years (Table IV).

Table IV: Sample of the final dataset with five random lines extracted from the right parties dataset. Column CD_MUNICIPIO corresponds to the identifiers of the municipalities, while the numbers of the year-columns are, in this sample, the vote-shares obtained by the right parties.

CD_MUNICIPIO	**1998**	**2002**	**2006**	**2010**	**2014**	**2018**
7315	0.831878	0.945746	0.750485	0.493275	0.702853	0.361684
35556	0.961531	0.874252	0.674680	0.628548	0.867617	0.608456
38750	0.910133	0.885119	0.797627	0.757765	0.445783	0.239770
77992	0.904682	0.750096	0.909836	0.848947	0.396799	0.776768
73334	0.986744	0.885215	0.940652	0.766229	0.917694	0.525755

For the presidential analysis, we selected the PT (Workers' Party) and PSDB (Brazilian Social Democracy Party) parties since they are the most predominant ones over the years. As for the congressman's office, we investigated the ideological concepts of left and right spectrum [Zucco and Power 2020]. The classification of a given party's ideological spectrum is based on its ideological score [Zucco and Power 2020]. To ensure comparison with the presidential elections dataset, we considered on the right dataset all the parties with the mean ideological score over the years greater or equal to the mean ideological score of PSDB. Thus, we categorized the remaining as left. Table V exhibits the analyzed parties and their respective classification.

Table V: Parties grouped by political classification.

Left	**Right**
CID (Cidadania)	MDB (Brazilian Democratic Movement)
PCB (Brazilian Communist Party)	PFL (Liberal Front Party)
PCDOB (Communist Party of Brazil)	PL (Liberator Party)
PDT (Democratic Labour Party)	PMDB (Brazilian Democratic Movement Party)
PPS (Popular Socialist Party)	PP (Progressives)
PSB (Brazilian Socialist Party)	PPB (Progressive Party of Brazil)
PSOL (Socialism and Liberty Party)	PPR (Reform Progressive Party)
PT (Workers' Party)	PR (Party of the Republic)
	PSD (Social Democratic Party)
	PSDB (Brazilian Social Democracy Party)
	PTB (Brazilian Labour Party)
	PV (Green Party)

3.3 Identifying Spatial Patterns

A common approach for identifying spatial patterns is through the assessment of spatial autocorrelation. The term was formalized by [Cliff and Ord 1972] as being a fundamental feature of spatial data. It is grounded on Tobler's first law of geography [Tobler 1970], which states that "everything is related to everything else, but near things are more related than distant things". In other words, the measure indicates whether the location of the observation presents any influence on the registered value.

The simplest manner of estimating the spatial autocorrelation is to plot the observed values in maps, but there are also mathematical methods [Anselin 1995; Anselin and Getis 1992]. In this work, more precisely in Step 3 of Fig. 4, we used one of the most popular methods developed to assess the spatial autocorrelation [Li et al. 2007]. The Moran's Index varies from -1 to 1, where values different than 0 denote the existence of spatial dependence, meaning that data is not randomly distributed over space. Positive values indicate the dataset has a positive spatial dependence. In other words, closer locations have similar results. Negative values, instead, are an indication of negative spatial dependence. Equation 1 describes the index components, with n being the number of locations, w_{ij} a weight between locations i and j, x_i the observed amount at location i, $\overline{x}$ is the average of x, and S_0 the squared sum of all weights.

In our study, n is the number of municipalities and the locations are the municipality geographic centroid coordinates. As for the weights, we applied the Queen strategy, which assigns a value of 1 when locations share at least one common vertex and 0 when no vertex is shared. We discarded the distance-based weight methods to prevent distant municipalities from influencing each other results.

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

The Moran's index method only indicates whether spatial dependence is present or not on data, but it does not allow identifying patterns. Thus, to identify which locations present higher or lower spatial autocorrelation, a local method is necessary. These methods are grouped into a class called LISA (Local Indication of Spatial Autocorrelation) and determine the spatial autocorrelation value for each dataset's locality. For this research, we opted for the Local Moran's Index [Anselin 1995], a method inspired by its global variation presenting resembling interpretation of results. Equation 2 shows how the index is calculated. Likewise the global method, x_i is the observed value at location i, $\overline{x}$ is the average of x, n is the number of locations, w_{ij} is the weight between locations i and j, and S_i^2 is calculated by Equation 3. Again, the locations are the Brazilian municipalities, and the weight strategy adopts the Queen method.

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

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

3.4 Identifying Temporal Patterns

In Step 4 of Fig. 4, we clustered the parties' vote share time-series datasets for each municipality seeking to identify temporal patterns. With this approach, we aim to learn which Brazilian municipalities present similar voting trends through the analyzed years and whether they are located near or far from each other in space.

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). Holding for the reproducibility of the results and minimizing the number of parameters used in the pipeline, we employed a hierarchical clustering algorithm with the Euclidean distance and Ward's method as the clustering criterion. However, we emphasize that the pipeline developed in this article is flexible and allows other clustering methods combined with tools to find the optimum number of clusters [Campello et al. 2013].

A hierarchical clustering algorithm builds groups in an agglomerative or divisive way based on the distance between them [Rokach and Maimon 2005]. The agglomerative strategy initially considers each element as a group, and each iteration constructs new groups. The divisive approach is the opposite, as it considers the whole set as a group and divides it with each iteration. A distance metric is used as the connection criterion, and the possible ones are the longest distance between groups, the shortest distance between groups, and Ward's method [Ward Jr 1963]. The latter, at each iteration, connects the groups with the smallest increase in their internal variance after. In this study, we adopted the Euclidean distance and Ward's method as the clustering criterion.

3.5 Evaluating Results

In Step 5 of the pipeline illustrated in Fig. 4, we evaluated the results from two perspectives: (i) data visualization, *i.e.*, producing graphic representations of the spatial patterns found in the election data; and (ii) clustering performance assessment, in which we evaluated the quality of the groups formed in order to identify temporal patterns in the data election. In this context, we assessed 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].

Silhouette coefficient evaluates the quality of formed clusters, with values closer to 1 meaning better clusters and values closer to -1 indicating incorrect clusters. Equation 4 determines how the value is calculated for each point, where a is the average distance from one point to all points in the same group, and b is the average distance from that same point to all points in another nearest group.

Davies-Bouldin index measures the similarity between groups considering density and distance. The lower bound is zero and values closer to zero indicates better results. The index is computed as formalized by Equation 5, where n denotes the number of groups, c_x is the x group's centroid, σ_x expresses the average distance between elements of x and c_x, and $d(c_i, c_j)$ corresponds to distance between c_i and c_j.

$$S = \frac{b-a}{\max(a,b)} \qquad (4) \qquad\qquad DB = \frac{1}{n}\sum_{i-1}^{n} \max_{i \neq j} \left[\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right] \qquad (5)$$

Calinski-Harabasz index is the ratio between the inter-group dispersion average and intra-group dispersion average, with the dispersion as the sum of squares of distances. It is a comparative measure used to find the appropriate number of groups, with higher values meaning better results. Equation 6 defines the index, with SS_B as the inter-group dispersion, SS_W as the intra-group dispersion, n_E as the size of dataset E, k the number of groups, C_q as points from group q, c_q as the centroid of q, c_E as the centroid of E, and n_q as the number of points from q.

$$s = \frac{SS_B}{SS_W} \times \frac{n_E - k}{k-1} \qquad (6)$$

$$SS_W = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)^2 \quad SS_B = \sum_{q=1}^{k} n_q (c_q - c_E)^2 \qquad (7)$$

4. RESULTS AND DISCUSSION

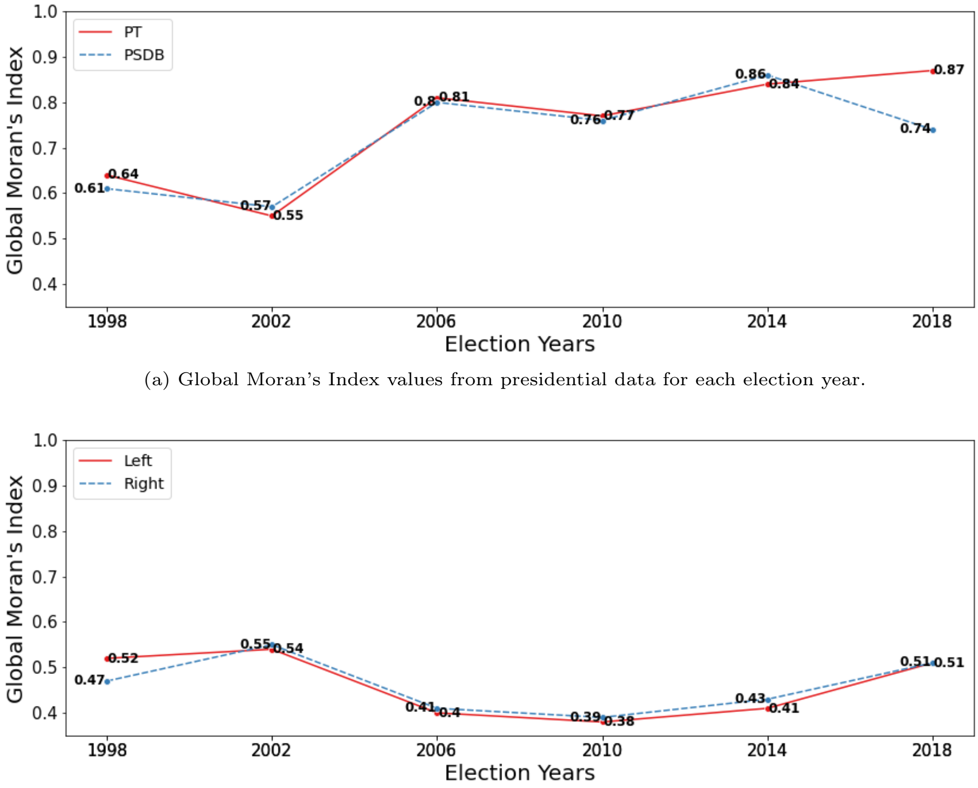
We arranged the analyses in two parts according to the datasets and techniques. First, we present and explain the results of global and local spatial autocorrelation measures from each electoral year to identify spatial dependence. Afterward, we describe and discuss the outcome of clustering the election results time-series of the Brazilian municipalities, seeking to detect temporal patterns. We performed both types of analysis twice; first with presidential election data, then with congress election data. The main goal behind these analyzes was to answer the following questions:

(i) In what extensions of Brazilian territory neighboring cities exhibit similar vote distribution?
(ii) Do neighboring cities present similar vote behavior over time?

The next subsections provide the answers to each question raised and the scientific background that supports them.

4.1 Analyzing Spatial Patterns

By analyzing the existence of spatial patterns, we expect to identify whether neighboring municipalities exhibit similar vote distribution. A first assessment of the Global Moran's Index (Fig. 5a) shows that since 1998 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's government



(a) Global Moran's Index values from presidential data for each election year.

(b) Global Moran's Index values from congressman data for each election year.

Fig. 5: Moran's Index results by electoral year.

and PT's ascension as the incumbent party. Not surprisingly, there is a drop in the Global 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 rise of uncertainty regarding the Brazilian voters.

As for congress elections (Fig. 5b), the overall results of the Global Moran's Index are lower than the obtained from the presidential data and follow an inverse behavior. It begins with an increasing tendency from 1998 to 2002 and presents a significant drop in 2006. The values remain low and decreasing until 2010, followed by a bit of increase in 2014 and a sharp increase in 2018. The lower values reflect the number of candidates, which increases the chances of having a random distribution over space, contributing to less spatial dependence.

Although the Global Moran's Index values can indicate spatial dependence, to identify the locations of the patterns, we calculated the Local Moran's Index of every municipality for each year. Fig. 6 displays the results for presidential election data, while Fig. 7 shows the results for congress election data.

For a better understanding of the Local Moran's Index results, we plotted 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 (high-high) are cities with a high percentage of votes for the party surrounded by cities with an also high percentage of votes for the party. On the other hand, the blue regions (low-low) 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 (low-high) are the municipalities where the party exhibited a low percentage of votes, and the neighboring cities had a higher percentage of votes. The orange cities (high-low) follow

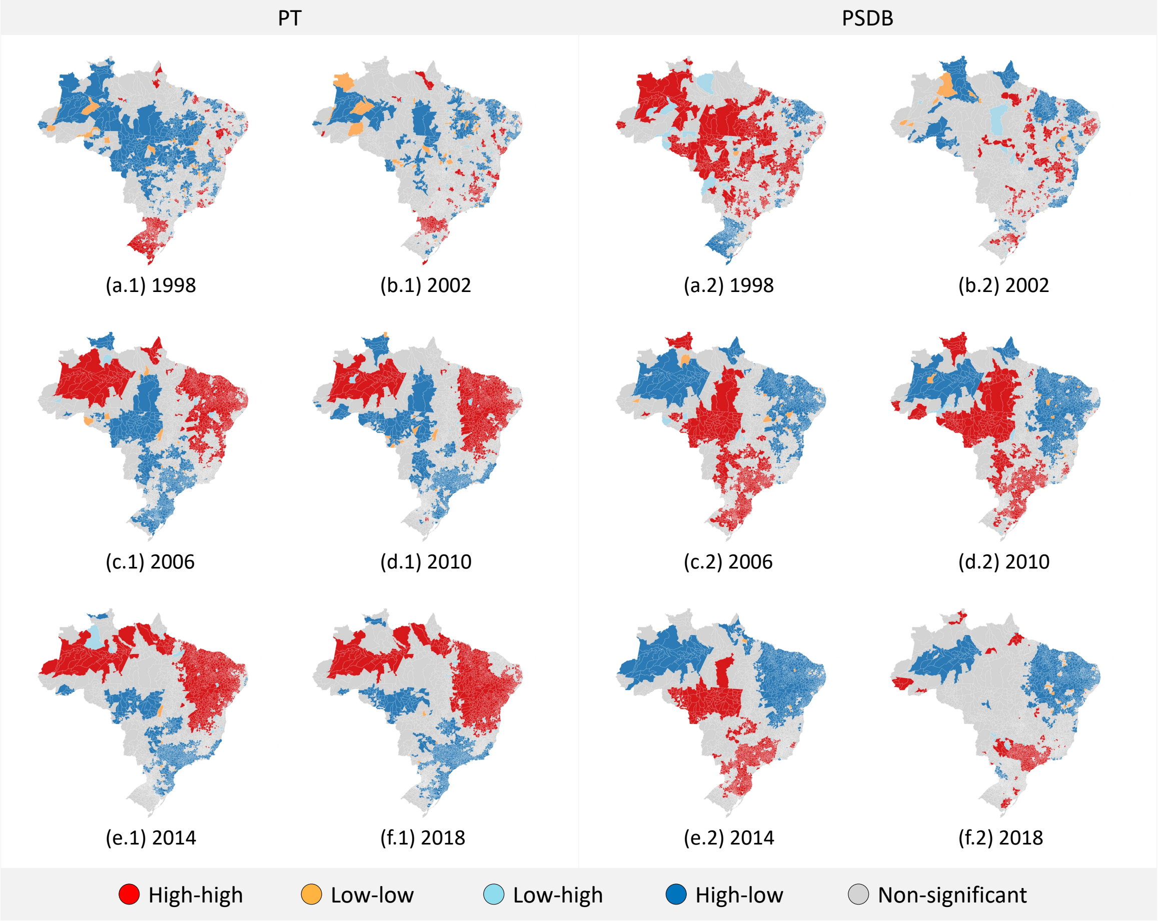


Fig. 6: Local Moran's Index plot of Workers' Party (PT) and Brazilian Social Democracy Party (PSDB) by election year.

the opposite behavior. Finally, the cities in gray (non-significant) are those where the index value was closer to 0, indicating randomness in the spatial distribution.

In order to achieve the discretization of municipalities into high-high, low-low, low-high, and high-low, we evaluated their results to be significant by the p-value (default of 0.05) obtained by the Local Moran's Index calculations with a Bootstrap method. For significant results, the municipalities with positive values for the normalized vote-shares and the average of the municipality's neighbors (spatially lagged variable) were classified as high-high. In opposite, the municipalities with negative values for both variables were the low-low locations. High-low is the label for the municipalities with positive vote-share and negative lagged vote-share, while low-high was the category for the opposite situation.

Comparing the local spatial autocorrelation plots from 1998 to 2002 of PSDB (Fig. 6 – PSDB), it is possible to understand what caused the drop in the Global Moran's Index in 2002. Since 1998, PSDB presented a decrease of hegemony, with a considerable number of cities going from red (high percentage of vote-shares) to gray or even blue (low percentage of votes) in 2002 (Fig. 6b.2). The same phenomenon occurs inversely for PT (Fig. 6 – PT), the number of blue regions decreased, and the number of gray regions increased, indicating a dispersed growth.

The subsequent years exhibited a higher number of cities highlighted in blue for PSDB, mainly in North and Northeast (Fig. 6 – PSDB). The same regions presented a high number of cities colored in red for PT (Fig. 6 – PT). Both situations contribute to the increase of the Global Moran's Index. It is worth mentioning that the plots from 2014 are almost the opposite of each other (Figs. 6e.1–e.2).

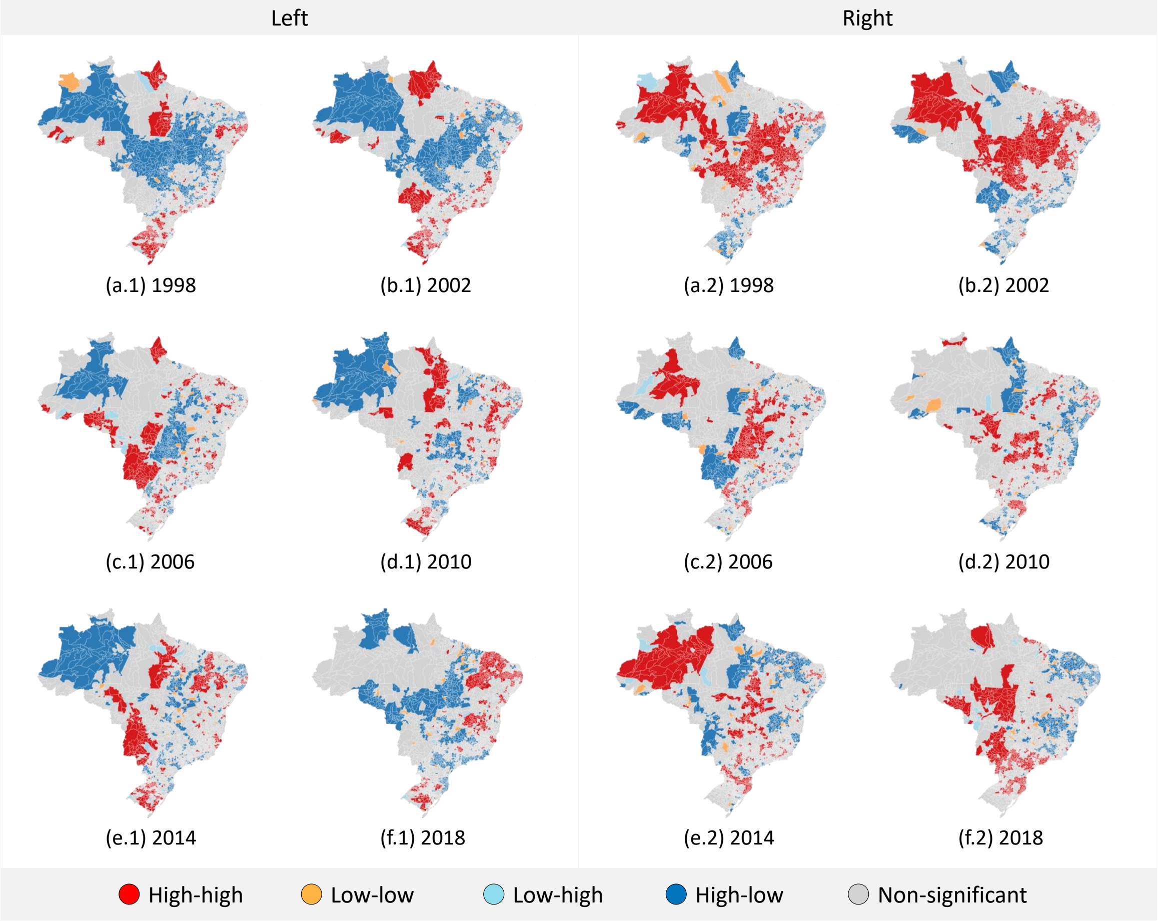


Fig. 7: Local Moran's Index plot of left and right parties by election year.

Not surprisingly, it was the year when both parties were the most voted.

We followed the same approach applied in Fig. 6 for the congress election data displayed in Fig. 7. Concerning the parties on the right (Fig. 7 – Right), from 1998 (Fig. 7a.2) to 2002 (Fig. 7b.2) the number of municipalities in red, meaning high vote-shares, was high in the Northern, Midwest, and part of the Northeast. In contrast, especially in the Southern, there were many blue regions, meaning low vote-shares. Moreover, there was also a high amount of orange and light blue cities, indicating outliers' presence. In the following years, from 2006 (Fig. 7c.2) to 2010 (Fig. 7d.2), the number of cities in gray indicating spatial randomness grew substantially, which can explain the Global Moran's Index decay in this period. Finally, in 2014 (Fig. 7e.2), there was an increase in red regions, especially in the Northern, and an increase of regions in blue in the Northeast. However, in 2018 (Fig. 7f.2), the spatial distribution in the Northern becomes random, while the Midwest, Southeast, and Southern present an increase of regions in red.

Similar changes that impact the Global Moran's Index can be observed on left parties maps (Fig. 7 – Left). The early years present more municipalities highlighted in red/blue, with many cities in orange/light blue. The number of gray municipalities grows until 2010 (Fig. 7d.1). In 2018 (Fig. 7f.1), the year with a higher global index value, there is an expressive increase of blue municipalities.

In general, the presidential and congress elections maps are comparable in some regions. For instance, in the Southern, it is possible to visualize the decrease of cities supporting PT and left parties (Fig. 6 and Fig. 7) over the years. On the other hand, PSDB and the right parties displayed an increase of hegemony in the Southeast and Southern. An inverse behavior can be observed in the

Table VI: Clustering evaluation metrics results from presidential analysis.

	PT			**PSDB**		
Clusters	Silhouette	Calinski Harabasz	Davies Bouldin	Silhouette	Calinski Harabasz	Davies Bouldin
2	**0.35**	**3645.71**	**1.01**	**0.19**	**1433.57**	1.90
3	0.21	3036.03	1.33	0.12	1132.51	**1.85**
4	0.22	2729.05	1.31	0.12	987.50	2.02
5	0.16	2348.92	1.67	0.07	821.82	2.08
6	0.15	2097.02	1.71	0.07	732.34	2.32
7	0.14	1856.44	1.72	0.06	653.02	2.67
8	0.13	1685.98	1.69	0.06	593.79	2.69
9	0.09	1578.14	1.74	0.05	539.31	2.51
10	0.09	1475.37	1.77	0.05	490.93	2.82

Table VII: Clustering evaluation metrics results from congress elections.

	Left			**Right**		
Clusters	Silhouette	Calinski Harabasz	Davies Bouldin	Silhouette	Calinski Harabasz	Davies Bouldin
2	**0.02**	**98.54**	**6.91**	**0.05**	**314.59**	**3.87**
3	-0.00	64.39	9.03	0.02	171.44	8.31
4	-0.03	57.19	8.59	-0.02	127.09	8.72
5	-0.03	53.68	9.38	-0.03	106.81	7.84
6	-0.04	50.87	7.99	-0.03	94.74	7.10
7	-0.04	68.58	7.01	-0.04	91.65	6.85
8	-0.07	60.23	7.57	-0.06	82.52	6.64
9	-0.06	73.01	6.53	-0.06	72.82	8.11
10	-0.06	65.29	8.29	-0.06	73.32	7.81

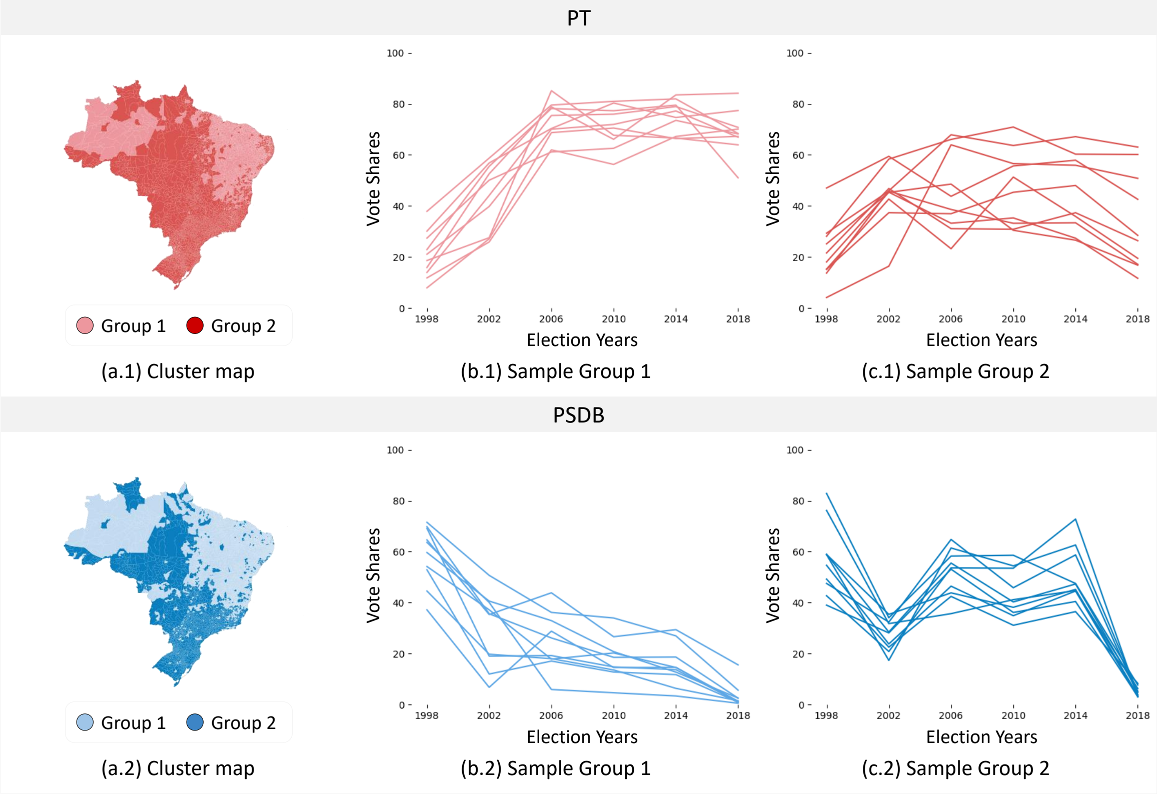
Northeast with an increase of cities supporting PT and left parties and a reduction of cities supporting PSDB and right parties.

4.2 Analyzing Temporal Patterns

Following the previous indications of neighboring municipalities sharing a similar voting behavior, we now aim to assess whether these municipalities maintain a similar voting pattern through time. To evaluate the behavior of regions over the years, we ran a hierarchical clustering method with the time-series of votes shares per city considering the four datasets being analyzed: PT, PSDB, left parties, and right parties. To identify the best number of groups to analyze, we evaluated the results from 2 to 10 groups considering three metrics: Silhouette, Calisnk-Harabasz, and Davies-Bouldin. Tables VI and VII show the results for presidential and congress elections, respectively. In general, the best results are in bold. However, it is noteworthy that while the silhouette's low values revealed a lack of cluster structure on the feature space, we are more interested in the geographical space. In other words, our main focus is to investigate whether neighboring cities are placed on the same group. Thus, we selected the number of clusters with the best metrics results to investigate it more deeply. Thus, from now on, we will focus our analyzes considering the clustering results of two groups.

Regarding presidential results (Fig. 8a.1 and Fig. 8a.2), 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 exhibited similar voting behavior over the years. For instance, considering the results for PT (Fig. 8a.1), almost every city of the Northeast region belongs to the same group. On the other hand, considering the results for PSDB (Fig. 8a.2), the majority of the Southeast cities belong to the same group.

In more detail, Fig. 8b.2 and Fig. 8c.2 present randomly chosen samples of PSDB vote-shares time-series from cities belonging to groups 1 and 2, respectively. In other words, we randomly selected ten municipalities classified as belonging to group 1 and plotted their vote-share on PSDB as a time series, beginning in 1998 and ending in 2018. The same criterion was used to produce the plots for municipalities of group 2 and the PT vote-shares. The series from group 1 displays a low percentage



(a.1) Cluster map (b.1) Sample Group 1 (c.1) Sample Group 2

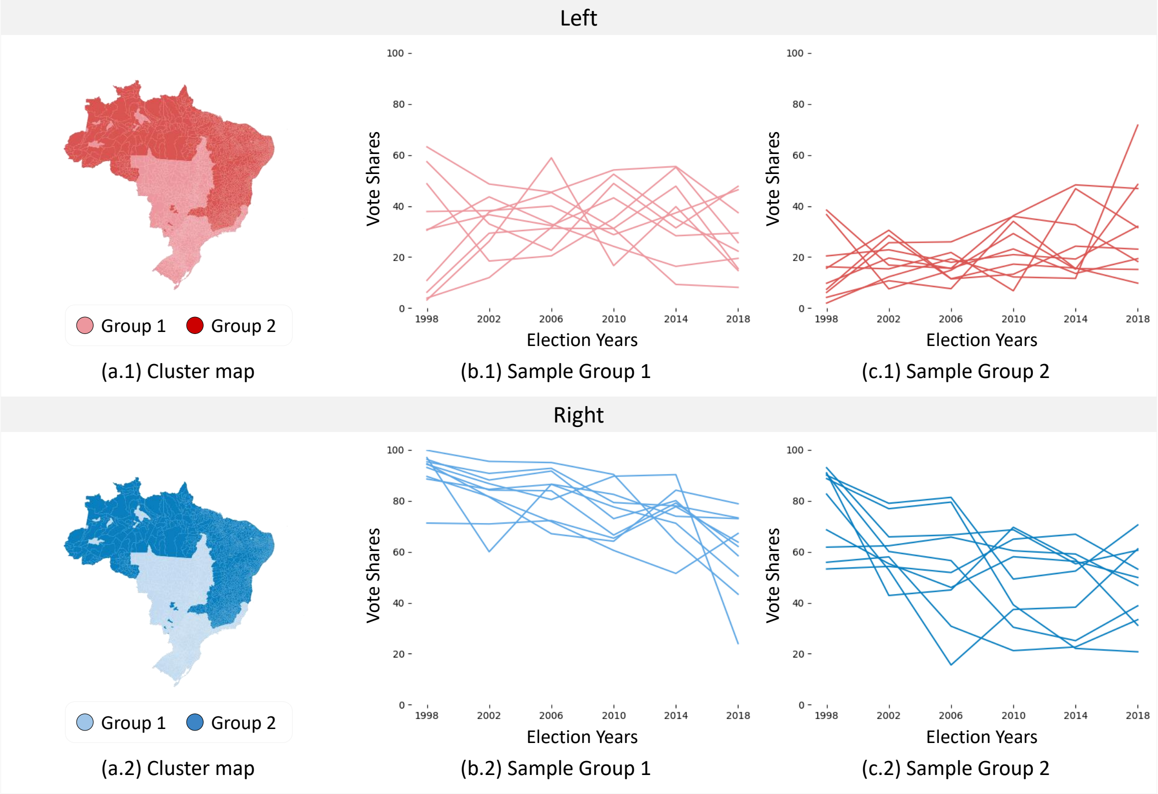
(a.2) Cluster map (b.2) Sample Group 1 (c.2) Sample Group 2

Fig. 8: Presidential election clustering maps results and voting series samples by groups.

of votes in 1994, followed by a peak in sequential years with a decreasing trend after. In contrast, for group 2, 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. 8b.1) features an increasing trend from 1994 to 2006, stabilization in 2010 and 2014, and a decrease in 2018. For group 2 (Fig. 8c.1), 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.

Concerning the congress elections (Fig. 9a.1 and Fig. 9a.2), the results are identical, meaning that the same spatial clusters were created in both datasets. These results occurred because the two datasets are almost complementary, *i.e.*, the sum of vote-shares per city is almost equal to 1. Another visual characteristic of the spatial clusters obtained is the groups' boundaries following the same limits as the state boundaries. We evaluated that the year 1998 has a strong impact on the generation of these results. There is a right parties hegemony in 1998, with vote-shares closer to 100%, while the left parties obtained voting shares closer to 0. Nevertheless, we can still observe a separation between North and Northeast regions from South, Southeast, and Central-West regions.

In more detail, Fig. 9b.2 and Fig. 9c.2 present randomly chosen samples of right parties' vote-shares time-series from cities belonging to groups 1 and 2, respectively, selected following the same criterion described for the presidential elections. The cities from group 1 exhibit a small decrease tendency starting with high vote-shares, closer to 100%, in 1998 and decreasing to a value closer to 60%. This tendency indicates that the right parties did not lose their hegemony in the group 1 regions. On the other hand, cities from group 2 present a more marked decrease tendency with vote-shares from 2018 lower than 50% indicating loss of hegemony in the cities from group 2. Moreover, the samples of vote-shares time series from left parties groups (Fig. 9b.1 and Fig. 9c.1) indicate that cities from group 1 exhibited an almost constant behavior over the years with vote-shares under 50%. Differently, cities from group 2 presented low vote-shares in the early years, but an increasing tendency in the last years.



(a.1) Cluster map (b.1) Sample Group 1 (c.1) Sample Group 2

(a.2) Cluster map (b.2) Sample Group 1 (c.2) Sample Group 2

Fig. 9: Congress election clustering maps results and voting series samples by groups.

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 article 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 and congress election's spatial and temporal patterns from 1998 to 2018 at a municipal level. From the spatial autocorrelation analysis, 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. Furthermore, the congress elections seem to be a slightly different process in comparison with presidential elections. It exhibits a hegemony of right parties over the years and a random component that diminishes spatial dependence.

The main difficulty faced in this work regards the lack of information concerning geographic units lower than municipalities, not allowing a more detailed analysis. Moreover, the data from the 1994 election presented a high frequency of missing data, improper to be used in the study. Finally, the results obtained in this article can only be discussed on a municipal level. Attempts to discuss them on lower levels will fall on the ecological fallacy problem.

Part of this study aimed to produce datasets that enable further work on the subject in question. Also, our findings can be the starting point for both broader and deeper analysis. Future research could be centered on refining the definitions of parties' location in the ideology spectrum, including more parties in the analysis, and reapplying the pipeline to compare the results. Besides, machine learning models focused on understanding and predicting electoral behavior could be explored.

ACKNOWLEDGMENTS

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].

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.

Anselin, L. and Getis, A. Spatial statistical analysis and geographic information systems. *The Annals of Regional Science* 26 (1): 19–33, 1992.

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

Campello, R. J., Moulavi, D., Zimek, A., and Sander, J. A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies. *Data Mining and Knowledge Discovery* 27 (3): 344–371, 2013.

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

Cliff, A. and Ord, K. Testing for spatial autocorrelation among regression residuals. *Geographical Analysis* 4 (3): 267–284, 1972.

Corrêa, D. S. Os custos eleitorais do bolsa família: Reavaliando seu impacto sobre a eleição presidencial de 2006. *Opinião Pública* 21 (3): 514–534, 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.

Faustino, J., Barbosa, H., Ribeiro, E., and Menezes, R. A data-driven network approach for characterization of political parties' ideology dynamics. *Applied Network Science* 4 (1): 1–15, 2019.

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

Hand, D. J. and Adams, N. M. Data mining. *Wiley StatsRef: Statistics Reference Online*, 2014.

Hernández, V. and León, L. Geografía de la participación electoral y diferenciación socioespacial en Ciudad Juárez, Chihuahua (México). *Geopolítica(s). Revista de Estudios sobre Espacio y Poder* vol. 11, pp. 145–172, 06, 2020.

Jacintho, L. H. M., da Silva, T. P., Parmezan, A. R. S., and Batista, G. E. A. P. A. Brazilian presidential elections: Analysing voting patterns in time and space using a simple data science pipeline. In *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*. SBC, Porto Alegre, pp. 217–224, 2020.

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.

Magalhães, A. M., Silva, M. E. A. d., and Dias, F. d. M. Eleição de Dilma ou segunda reeleição de Lula? Uma análise espacial do pleito de 2010. *Opinião Pública* 21 (3): 535–573, 2015.

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.

Martins, D. J. D., Mansano, F. H., Parré, J. L., and Plassa, W. Fatores que contribuíram para a reeleição da presidente Dilma Rousseff. *Política & Sociedade* 15 (32): 145–170, 2016.

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

Mota, A. M. S. *Modelling abstention rate using spatial regression*. M.S. thesis, NOVA Information Management School, 2019.

Norris, P. and Grömping, M. Electoral integrity worldwide, 2019. Sydney: Electoral Integrity Project. Available at `https://www.electoralintegrityproject.com/`.

Okunev, I. Y., Gorelova, J. S., and Gruzdeva, E. Regional disparities of electoral behaviour in poland: Comparative spatial analysis. *Comparative Politics Russia* 12 (1): 149–160, 2020.

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.

Praciano, B. J. G., da Costa, J. P. C. L., Maranhão, J. P. A., de Mendonça, F. L. L., de Sousa Júnior, R. T., and Prettz, J. B. Spatio-temporal trend analysis of the Brazilian elections based on twitter data. In *Proceedings of the IEEE International Conference on Data Mining Workshops*. IEEE, Singapore, pp. 1355–1360, 2018.

Recuero, R., Soares, F. B., and Gruzd, A. Hyperpartisanship, disinformation and political conversations on twitter: the Brazilian presidential election of 2018. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (1): 569–578, May, 2020.

Reid, B. and Liu, G.-J. One nation and the heartland's cleavage: an exploratory spatial data analysis. In *The Rise of Right-Populism: Pauline Hanson's One Nation and Australian Politics*, B. Grant, T. Moore, and T. Lynch (Eds.). Springer, Singapore, pp. 79–102, 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.

Schuhli, G. T. O Partido dos Trabalhadores e o voto católico no segundo turno da eleição presidencial de 2010: uma análise espacial a nível municipal. *Revista da FAE* 21 (1): 156–167, 2018.

Terron, S. L. and Soares, G. A. D. As bases eleitorais de lula e do pt: do distanciamento ao divórcio. *Opinião Pública* 16 (2): 310–337, 2010.

Tobler, W. R. A computer movie simulating urban growth in the detroit region. *Economic Geography* 46 (sup1): 234–240, 1970.

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

Zucco, C. and Power, T. Fragmentation without cleavages? Endogenous fractionalization in the Brazilian party system. *Comparative Politics*, 01, 2020.