UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Learning beyond the spatial autocorrelation structure: A machine learning-based approach to discovering new patterns and relationships in the context of spatially contextualized modeling of voting behavior

Tiago Pinho da Silva

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Tiago Pinho da Silva

Learning beyond the spatial autocorrelation structure: A machine learning-based approach to discovering new patterns and relationships in the context of spatially contextualized modeling of voting behavior

Thesis submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Gustavo Enrique de Almeida Prado Alves Batista

USP – São Carlos
October 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

P654l

Pinho da Silva, Tiago
Learning beyond the spatial autocorrelation structure: A machine learning-based approach to discovering new patterns and relationships in the context of spatially contextualized modeling of voting behavior / Tiago Pinho da Silva; orientador Gustavo Enrique de Almeida Prado Alves Batista. -- São Carlos, 2023.
128 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Machine Learning. 2. Spatial Data Modeling. 3. Spatial Autocorrelation. 4. Electoral Behavior Modeling. 5. Electoral Geography. I. Enrique de Almeida Prado Alves Batista, Gustavo, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Tiago Pinho da Silva

Aprendendendo além da estrutura de autocorrelação espacial: Uma abordagem baseada em aprendizado de máquina para a descoberta de novos padrões e relações no contexto de modelagem espacialmente contextualizada do comportamento eleitoral

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Gustavo Enrique de Almeida Prado Alves Batista

USP – São Carlos
Outubro de 2023

Este trabalho é dedicado a minha esposa Marília de Lima Cirqueira,
e a minha mãe Aldenora Alves Pinho.

ACKNOWLEDGEMENTS

The following thesis is a compilation papers I developed during my Ph.D at the Instituto de Ciências Matemáticas e Computação (ICMC) from the University of São Paulo (USP) in Brazil. Although the major part of this is written in English I will indulge myself to write in Portuguese so it can be understood by the people who accompanied me and contributed, in many ways, to my personal growth and professional maturity.

O início de toda essa jornada ainda me aparece bem vívido na memória, sob relances de momentos que eu poderia jurar que aconteceram semana passada. Desses momentos ainda me lembro com alegria a primeira vez que conheci o LABIC e o sentimento de entusiasmo que senti quando me dei conta que trabalharia ali. Naquela época eu não fazia ideia do que viria, das inúmeras discussões, das amizades e dos desafios que enfrentei e me ajudaram a me tornar quem sou hoje. Além disso, tivemos que enfrentar uma pandemia que frustou planos e me fez rever pensamentos e ideais a muito tempo enraizadas na minha mente.

Sendo assim, é com sentimento de gratidão que chego no fim de mais uma etapa, dessa vez uma das mais longas e mais prematuramente planejadas da minha vida até agora. Apesar de todos os planos que eu poderia fazer jamais teria imaginado o caminho que segui. Nesse caminho, dentre as várias decisões e acasos, pude conhecer lugares e pessoas incríveis que me inspiraram, apoiaram e me encorajaram para que eu pudesse chegar até aqui orgulhoso do meu trabalho e com sentimento de missão cumprida.

Agradeço a minha esposa Marília de Lima Cirqueira por todo apoio, cumplicidade e amor. Nossas viagens para conhecer lugares novos, as noites de fim de semana regadas a boas cervejas e petiscos e sem falar nos vários momentos assistindo séries de comédia e doramas foram muito importantes pra mim. Você me faz querer ser a melhor versão de mim.

À minha mãe Aldenora Alves Pinho que foi a primeira a acreditar em mim e me apoiar em todas a minha decisões. Mãe, jamais teria alcançado minhas conquistas se não fosse seu amor incondicional e seu apoio. Pela qual também agradeço meu padastro Josias Lima pelo apoio e ajuda.

À minha segunda família que me acolheu com carinho, Maria Messias Cirqueira, Miguel Cirqueira, Vitor Cirqueira e Raul Cirqueira. Obrigado pelo apoio e me aceitarem como membro da família.

Ao meu orientador Gustavo Enrique de Almeida Prado Alves Batista, pela confiança, paciência e respeito na maneira como conduziu seus ensinanentos e me ajudou durante minha pesquisa. Sua orientação foi essencial para meu crescimento como pesquisador.

Aos meus avós Paula Moraes Alves, Antonia Lima da Silva e Raimundo Marques Pinho, que dedicaram suas vidas aos seus filhos e sem o quais eu não estaria aqui. Pelos quais agradeço a todos os meus familiares que de forma direta ou indireta me ajudaram até aqui.

Aos meus primos Elen Moreno, João Paulo, Ana Carolina, Priscilla e Luis Vitor as conversas e memes foram fundamentais nessa minha jornada.

Ao meu amigo Antonio Parmezan, pelas infinitas ajudas e pela paciência em trabalhar comingo. Você foi uma peça fundamental nesse trabalho.

Aos meus colegas e professores do LABIC em especial ao Brucce, Wackar e Vinícius Mourão, Professora Solange, Professor Ricardo e Professor Diego. Obrigado pelas discussões sobre meus projetos e por terem escutado minhas idéias mais absurdas e me encorajado a fazê-las.

Aos meu amigos do intercâmbio, Alison, Luis, Luma, Gustavo, Cris, Ramom, Gabriela, Zenival e Ney. Obrigado por sempre estarem por perto mesmo a Quilômetros de distância, saber que sempre posso contar com vocês me ajudou bastante nesse caminhada.

Aos meus amigos Carlos Soares e Fidel Marx, pelos finais de semana de jogatinas. Vocês foram essensiais nesse minha jornada.

Aos meus amigos Iara, Mariana e Pedro, pelos finais de semana regados a boa cerveja e conversa.

Aos meus amigos de tabuleiros, Maykon, Marcela e Eduardo. Obrigado pelas noites de jogatinas que aliviavam muito do estresse da vida acadêmica e me faziam sentir renovado.

Aos meus amigos de mestrado Eric, Joel, Danilo, Gean, Gerson e Eduarth pelas mais abusurdas conversas que me trouxeram risadas em momentos difíceis.

Ao meu psicólogo Pedro Almeida, pela paciência em ouvir meus problemas e pela ajuda em vários momentos difíceis.

Por fim, agradeço a todos os pesquisadores do Brasil. Ser pesquisador no Brasil é sobretudo um ato de resistência, é sobre resistir aos inúmeros ataques a ciência aos constantes cortes de gastos na educação e ao descrédito da sociedade. Obrigado por resistirem.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

学而不厌，诲人不倦。孔子

Nunca se canse de aprender, nunca se canse de ensinar. Confúcio

RESUMO

DA SILVA, T. P. **Aprendendendo além da estrutura de autocorrelação espacial: Uma abordagem baseada em aprendizado de máquina para a descoberta de novos padrões e relações no contexto de modelagem espacialmente contextualizada do comportamento eleitoral**. 2023. 128 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

As eleições são um pilar fundamental das sociedades democráticas, proporcionando aos cidadãos meios para eleger os seus representantes e moldar a direção de governos. No entanto, nos últimos anos, houve um aumento na preocupação com a integridade dos processos eleitorais em todo o mundo, com alegações de fraude e crescente polarização. Para compreender melhor o eleitorado e os fatores que influenciam suas escolhas, um número crescente de pesquisadores se voltaram para a modelagem do comportamento eleitoral, que lança luz sobre fenômenos políticos como a polarização e os contextos demográficos e socioeconômicos que moldam a natureza do eleitorado. A literatura sobre modelagem de comportamento eleitoral pode ser amplamente dividida em duas áreas principais: ciência política, que argumenta que apenas fatores individuais explicam o comportamento eleitoral usando principalmente dados de pesquisas eleitorais; e geografia eleitoral, que afirma que fatores contextuais, como localização, desempenham um papel crucial na determinação do comportamento eleitoral usando conjuntos de dados com informações agregadas espacialmente, como dados do censo. A ciência política tornou-se a abordagem dominante devido ao aumento da qualidade dos dados coletados nas pesquisas, mas a disponibilidade pública de tais dados é limitada e cara. Em contraste, os dados do censo, que fornecem informações detalhadas sobre as características socioeconômicas e demográficas de uma população, são disponibilizados publicamente por agências governamentais. No entanto, apesar de seu potencial para fornecer informações abrangentes e perspicazes sobre o eleitorado, esses tipos de dados são subutilizados na modelagem do comportamento eleitoral, principalmente devido às limitações dos principais métodos de modelagem do eleitorado em lidar com dados de alta dimensão e identificar relações não lineares. Para lidar com essas limitações, tem havido uma tendência crescente na utilização de métodos de aprendizado de máquina que podem lidar melhor com alta dimensionalidade e modelar relações não lineares. No entanto, a maioria desses trabalhos negligencia as características espaciais dos dados. Esta tese defende a importância de incorporar informações de dependência espacial no *pipeline* de aprendizado de máquina para a tarefa de modelagem do comportamento eleitoral usando dados do censo. O *pipeline* de aprendizado de máquina tradicional pode exibir viés em relação a modelos que aprendem a estrutura de autocorrelação espacial, dificultando a descoberta de novos padrões fora da estrutura de autocorrelação, o que contradiz o objetivo principal de identificar novos padrões. Nesta tese, o impacto da dependência espacial na tarefa de modelagem do comportamento eleitoral é estudado, e adaptações ao *pipeline* tradicional de aprendizado de máquina são propostas, desenvolvidas e avaliadas. Nesse sentido, propomos duas técnicas de validação cruzada espacial que levam em consideração os aspectos espaciais dos dados e fornecem cenários para a avaliação de modelos de aprendizado de máquina sem a influência da dependência espacial. Além disso, propomos uma abordagem de aprendizado de máquina baseada em *stacking* para modelar os dados com base em contextos geográficos e identificar relações locais e globais para entender os resultados das eleições. Os resultados desta tese indicam que as abordagens propostas são adequadas para a tarefa de modelagem espacialmente contextualizada do comportamento eleitoral. As técnicas de validação foram capazes de fornecer cenários mais realistas e menos tendenciosos quando comparadas às abordagens existentes na literatura, e a abordagem de aprendizado de máquina superou o estado da arte na literatura e forneceu resultados interpretáveis. No geral, esta pesquisa avança o estado da arte em modelagem de comportamento eleitoral e fornece uma nova metodologia na área, abrindo caminho para novas abordagens de aprendizado de máquina para ajudar a entender os resultados das eleições.

Palavras-chave: Aprendizado de Máquina, Modelagem de Dados Espaciais, Autocorrelação Espacial, Dependência Espacial, Particionamento de Dados Espaciais, Semivariograma, Erro de Predição, modelos de Ensemble, metaaprendizagem, votação preferencial, comportamento eleitoral, modelagem de comportamento eleitoral, geografia eleitoral.

ABSTRACT

DA SILVA, T. P. **Learning beyond the spatial autocorrelation structure: A machine learning-based approach to discovering new patterns and relationships in the context of spatially contextualized modeling of voting behavior**. 2023. 128 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Elections are a cornerstone of democratic societies, providing citizens with the means to elect their representatives and shape the direction of their government. However, we have seen in recent years an increase in concern about the integrity of electoral processes worldwide, with allegations of fraud and rising polarization. To better comprehend the electorate and the factors influencing its choices, an increase number of researchers have turned to electoral behavior modeling, which sheds light on political phenomena such as polarization and the demographic and socioeconomic contexts shaping the nature of the electorate. The literature on electoral behavior modeling can be broadly divided into two main areas: political science, which argues that only individual factors explain electoral behavior using primarily survey data; and electoral geography, which asserts that contextual factors, such as location, play a crucial role in determining electoral behavior using datasets with spatially aggregated information such as census data. Political science has become the dominant approach due to the increased quality of data collected from surveys, but the public availability of such data is limited and costly. In contrast, census data, which provides detailed information about a population's socioeconomic and demographic characteristics, is made publicly available by government agencies. However, despite its potential for providing comprehensive and insightful information on the electorate, large census datasets are underutilized in modeling electoral behavior, mainly due to the limitations of regression analysis in handling high-dimensional data and identifying non-linear relationships. To address these limitations, there has been a growing trend towards using machine learning methods that can better handle high-dimensionality and model non-linear relationships. However, most of these works neglect the spatial characteristics of the data. This thesis argues for the importance of incorporating spatial dependence information in the machine learning pipeline for the task of electoral behavior modeling using census data. The traditional machine learning pipeline can exhibit bias towards models that learn the spatial autocorrelation structure, hindering the discovery of novel patterns and relationships beyond this structure, which contradicts the main objective of identifying new patterns and relationships. In this thesis, the impact of spatial dependence on the task of electoral behavior modeling is studied, and adaptations to the traditional machine learning pipeline are proposed, developed, and evaluated for the considering task. In this regard, we propose two Spatial Cross-Validation techniques that take into account the spatial aspects of the data and provide scenarios for the evaluation of machine learning models without the influence of spatial dependence. Moreover, we propose a stacking-based machine learning approach to model the data based on geographical contexts and identify local and global relationships to understand the election results. The results in this thesis indicate that the proposed approaches are well-suited to the task of spatially contextualized modeling of electoral behavior. The validation techniques were able to provide more realistic and less biased scenarios when compared to existing approaches in the literature, and the machine learning approach outperformed the state-of-the-art in the literature and provided interpretable results. Overall, this research advances the state-of-the-art in electoral behavior modeling and provides a novel methodology in the electoral behavior area, paving the way for new machine learning approaches to help understand election results.

Keywords: Machine Learning, Spatial Data Modeling, Spatial Autocorrelation, Spatial Dependence; spatial data partitioning, Semivariogram, Prediction Error, Ensemble Learning, Metalearning, Preferential Voting, Voting Behavior, Electoral Behavior Modeling, Electoral Geography.

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF TABLES

CONTENTS

CHAPTER 1

INTRODUCTION

Elections are fundamental processes of any democratic society, allowing citizens to elect their representatives and shape the direction of their government. It presents a broad impact, from local elections for city councils to national elections for presidents and members of parliament. Moreover, when fairly conducted, elections reduce the likelihood of actions that undermine democratic foundations with destabilizing effects, allowing the best expression of the electorate's political preferences (LEHOUCQ, 2003). However, recently, there has been an increasing concern about the electoral processes worldwide. This concern regards factors causing distrust in the electoral processes, such as allegations of fraud and increasing polarization (NORRIS, 2013). Nonetheless, ascertaining the factors that influence elections is a complex task requiring a detailed and comprehensive analysis of the electoral process.

This obstacle, however, is not insurmountable. As a response, modeling the electorate's preferences behavior, also named *voting behavior* or *electoral behavior modeling*, allows us to comprehend the electorate and the aspects that guide its choices. For instance, it reveals meaningful insights into studies of political phenomena, such as the extensions of polarization and the demographic and socioeconomic contexts shaping the nature of the electorate (PINHEIRO-MACHADO; SCALCO, 2020). In this regard, there is a broad literature on building models to help understand the electorate preferences (ELKINK; FARRELL, 2021; LAGO, 2019; FOREST, 2018). This literature can be roughly divided into two main areas: *political science* and *electoral geography* (FOREST, 2018).

Political scientists argue that only individual factors explain electoral behavior. In other words, electoral results can only be explained when analyzing the individual characteristics of the electorate, such as age, incoming and ideological affinity. Therefore, they often rely on individual-level data from extensive surveys to build a detailed electorate profile (PINHEIRO-MACHADO; SCALCO, 2020). This perspective has become dominant in the literature mostly due to the increased quality of the data collected from surveys (FOREST, 2018). This rise in data quality has facilitated regression analysis, the primary method used to model electoral behavior, thereby reinforcing the dominant perspective in the literature.

In contrast, electoral geographers assert that not only individual characteristics but contextual factors, such as location, play a crucial role in determining electoral behavior, a perspective that we will name spatially contextualized modeling of electoral behavior to differentiate from the previous. In this perspective, researchers argue that nearby neighbors tend to vote similarly. A phenomenon commonly called the "Neighborhood Effect" (AGNEW, 1996; FOREST, 2018). Moreover, they usually use data that describes populations at different spatial aggregation levels, such as municipalities and districts. In general, such data provides information regarding general aspects of groups of populations, such as incoming per capita and the percentage of men and women. Thus, the spatially contextualized modeling of electoral behavior provides a more general perspective of the electorate when compared to the political science perspective. However, despite the works proposed in electoral geography over the last decades (MANSLEY; DEMŠAR, 2015; FOREST, 2018), political science has overshadowed this area (FOREST, 2018).

Although the political science perspective has become the dominant approach due to the quality of survey data which can provide specific information to facilitate the electorate modeling, the public availability of such data is scarce, and the data requires expensive resources to be collected and usually covers a small portion of the population. In contrast, datasets composed of spatially aggregated data information regarding population groups are made publicly available by government agencies. Therefore, they constitute official information from the electorate that is extensively curated. Moreover, it presents wide coverage and comprehensive information on the population. In particular, census data provide detailed information about a population's socioeconomic and demographic characteristics from thousands of features describing locations in different geographical levels of aggregation. These characteristics provide the ideal scenario for a comprehensively modeling of the electorate regarding the spatial and explanatory features domains (LI; PERRIER; XU, 2019). However, there are two major aspects to be considered, spatial autocorrelation and high-dimensionality.

Spatial autocorrelation regards the classical Tobler's first law of geography, which states that "everything is related to everything else, but near things are more related than distant things" (TOBLER, 1970). It describes the spatial dependence between observations of a variable at different locations, where nearby locations are more likely to present similar values than distant ones. For example, this behavior can be observed while comparing vote counts from a given candidate in a determined election (TERRON; SOARES, 2010). We expect nearby cities to share similar vote counts, while distant ones may present distinct results. It is a fundamental characteristic of spatial data (CLIFF; ORD, 1972), which became the basis of subsequent research in spatial data analysis and related areas (GETIS, 2010). The existence of spatial autocorrelation requires the model to incorporate the dependence assumption as a condition to avoid erroneous results interpretation. Regarding electoral behavior modeling, researchers generally propose regression methods incorporating spatial dependence in the model structure, such as mixed effect models and geographically weighted regression (WONG; WONG, 2022; MANOEL; COSTA; CABRAL, 2022).

The high dimensionality of census data is associated to the vast information it presents for describing the population. For example, the Brazilian census comprises approximately 4,000 variables at different aggregation levels, ranging from census tracts to states. Similarly, Australia has over 30,000 features describing its population. However, the curse of dimensionality problem arises when modeling the electorate preferences through census data, referring to the exponential increase in the modelling complexity as the data dimensionalities increase. To address this problem, researchers typically choose a few variables of interest manually and apply them to regression models (MANSLEY; DEMŠAR, 2015; FOREST, 2018).

In summary, electoral behavior modeling is a fundamental tool for understanding elections. It can provide meaningful insights into factors influencing the electorate's choices. However, the community has overlooked the potential of the electoral geography perspective by favoring survey data, despite the wide coverage and amount of information in spatial datasets such as census data. Conversely, the electoral geography community does not explore the high dimensionality of such datasets that presents a detailed information and describe entire regions at different geographical levels of aggregation. Thus, there is gap regarding the exploration of these datasets to uncover new insights regarding electorate preferences.

1.1 Motivation

The predominant approach in electoral geography is to use regression analysis, which cannot handle high-dimensional data, such as census data, and is limited to identifying linear relationships. From this perspective, there is a necessity for new methodologies capable of fully exploiting the potential of large spatial datasets such as census data in modeling electoral behavior.

Nonetheless, recently, there has been a growing trend towards using machine learning methods to model electoral behavior, as evidenced by several recent studies (ELKINK; FARRELL, 2021; YERO; SACCO; NICOLETTI, 2021; LI; PERRIER; XU, 2019). These methods have the advantage of being able to model both linear and nonlinear relationships and are better suited to handling high-dimensional data. This shift in methodology has led to a more exploratory approach, in contrast to the traditional hypothesis-driven approach (ELKINK; FARRELL, 2021). Most of the works try to exploit the information on census data (LI; PERRIER; XU, 2019; YERO; SACCO; NICOLETTI, 2021). However, some of these studies neglect the spatial characteristics of census and electoral data by considering the traditional machine learning pipeline. It is noteworthy that in this thesis, we consider the pipeline consisting of the steps in Figure 2: data acquisition, data pre-processing, modeling, validation, and results evaluation. In this pipeline, the modeling is done by traditional machine learning methods, and the validation step uses traditional validation techniques such as K-Fold Cross Validation (CV).

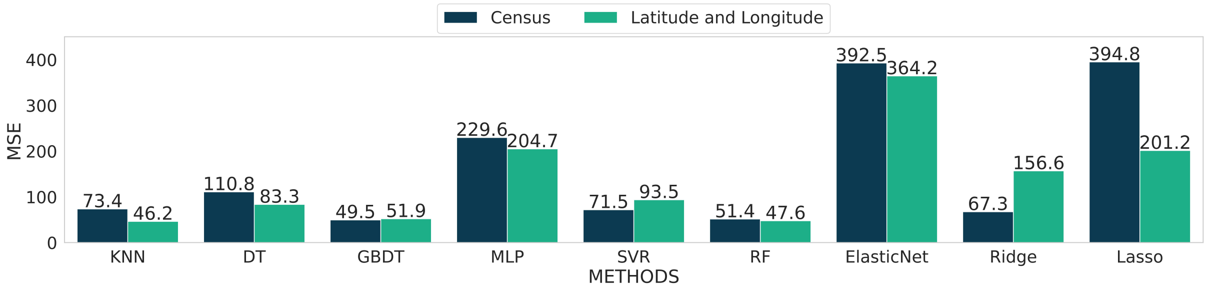
On this matter, the validation step is one of the processes in the traditional machine learning pipeline that needs more attention when dealing with spatial data. From a general perspective, CV and Hold-Out are the standard validation processes applied in the machine learning literature (PLOTON; MORTIER *et al.*, 2020). They can assess the model generalization on unseen data by splitting the dataset into training and test set. Consequently, they have been applied unconstrainedly to recent works on spatially contextualized electoral behavior modeling using machine learning, disregarding the spatial aspects of the data (YERO; SACCO; NICOLETTI, 2021; LI; PERRIER; XU, 2019). Such disregard can impact the quality of the research in the area by assuming that the results obtained by such validation techniques are not influenced by spatial autocorrelation.

Moreover, another step that needs to incorporate the spatial characteristics of the data is modeling. Unfortunately, several works often disregard spatial dependence during the modeling step (YERO; SACCO; NICOLETTI, 2021; LI *et al.*, 2019). Traditional methods are applied without incorporating spatial dependence, which results in models that reflect the spatial dependence structure of the data. This limitation hinders the discovery of new patterns and relationships that can be relevant from a local perspective.

In order to demonstrate the limitations of applying traditional machine learning methods to spatially contextualized electoral behavior modeling, a simple experiment is presented in Figure 1. The experiment evaluates the mean-squared error performance of various traditional machine learning methods for predicting city-level election results using two feature sets: one comprised of census data and the other *only containing latitude and longitude information*. In this experiment, latitude and longitude represent the set of features with extremely high spatial autocorrelation compared to census data. Results indicate that models trained with features with high spatial autocorrelation exhibit similar or superior performance relative to models utilizing census data. Moreover, methods that leverage spatial dependence information, such as KNN and Tree-based methods such as decision trees, random forests and gradient boosting (ROBERTS *et al.*, 2017; PLOTON; MORTIER *et al.*, 2020), outperform other techniques when using only latitude and longitude information. These results underscore the importance of incorporating spatial dependence information in the machine learning pipeline when modeling electoral behavior from spatial data.

In summary, when working with spatially dependent data, traditional machine learning pipelines can exhibit a bias towards models that learn the spatial autocorrelation structure rather than those that capture other patterns and relationships (ROBERTS *et al.*, 2017). However, when using machine learning methods to model electoral behavior with thousands of features, the approach is often exploratory in nature (ELKINK; FARRELL, 2021). Consequently, the traditional machine learning pipeline can hinder the discovery of novel patterns and relationships. This is because the models produced by such a pipeline typically explain electoral choices

Figure 1 – 10-Fold Cross Validation Mean Squared Error (MSE) results obtained from different machine learning methods trained on the census and latitude and longitude features for predicting the vote-shares from the winning party in the second round of the 2018 Brazilian Presidential Election. The acronyms for each machine learning methods are: k-Nearest Neighbors (kNN), Decision Tree (DT), Gradient Boosting DT (GBDT), Randon Forest (RF), Multiyear Perceptron (MLP), Support Vector Regression (SVR), Least Absolute Shrinkage and Selection Operator (LASSO).



Source: Elaborated by the author.

based on the well-explored phenomenon of the "Neighborhood Effect," which posits that nearby locations tend to vote similarly (AGNEW, 1996). From this perspective, there is a need for new validation techniques and machine learning methods that incorporate spatial dependence for the task of spatially contextualized electoral voting behavior modeling.

1.2 Justification

Machine learning applied to spatial data is a well-established area of research, with researchers proposing approaches to deal with such data for over a decade (ROBERTS *et al.*, 2017; PLOTON; MORTIER *et al.*, 2020; VALAVI *et al.*, 2019). However, recent advances in the modeling of spatial data have renewed interest in revisiting the electoral geography perspective of modeling electoral behavior, particularly in terms of the validation and modeling processes.

The validation of machine learning models trained on spatially dependent data is a topic of increasing interest in the research community (MILÀ *et al.*, 2022; PLOTON; MORTIER *et al.*, 2020; ROBERTS *et al.*, 2017). In this area, researchers have proposed and discussed various validation techniques that consider the spatial characteristics of the data. For the sake of simplicity, we refer to such techniques as Spatial Cross-Validation (SCV) in this thesis. Typically, SCV techniques divide the dataset based on the spatial domain, creating training and testing sets that correspond to spatial blocks. The rationale behind this approach is that, given positive spatial autocorrelation, nearby data points will not be separated into training and testing sets, thereby mitigating the influence of spatial dependence on the results. Thus, various SCV techniques have been recently proposed for diverse applications involving spatial data (MILÀ *et al.*, 2022; DEPPNER; CAJIAS, 2022; PLOTON; MORTIER *et al.*, 2020; ROBERTS *et al.*, 2017).

Nonetheless, the advance in SCV techniques is insufficient to cope with the spatially contextualized electoral behavior modeling task. Existing approaches still have limitations. For instance, some are based on Leave-One-Out, which can be time-consuming, particularly for electoral datasets with thousands of examples. Others assume that a removing buffer region can be defined to separate the test from the training set independently of the test set with the same size in all directions. Additionally, some approaches seek spatial independence disregarding the amount of data removed from the training set, which may lead to overfitting. Finally, some approaches assume that spatial dependence occurs contiguously, which may not hold in real-world datasets such as census and electoral data.

Moreover, regarding the modeling process, to the best of our knowledge, the first study to apply a machine learning method to model electoral behavior from census data was proposed by (LI; PERRIER; XU, 2019). The authors combined the hierarchical nature of census and election data with the capability of Graph-Convolutional Neural Networks (GCNNs) to learn local patterns, resulting in a model that was able to identify local relationships in the context of the 2019 New South Whales (NSW) congressional election. Lately, (ELKINK; FARRELL, 2021) proposed the use of decision trees to model voting behavior and identify new relationships regarding the 2020 Irish General Election. They train a decision tree model on data from the 2020 Irish National Election Study (INES) survey that describes, at the individual level, the socioeconomic and demographic characteristics of a population sample followed by the candidates of their choice. The authors encounter results that corroborate with other works that imply a major transformation in Irish electoral politics concerning an increasing polarization between left and right ideologies. Following a similar direction, (YERO; SACCO; NICOLETTI, 2021) proposed using decision trees trained on Human Development Indexes (HDI) from various domains, such as education, health, and age, at the municipality level of aggregation to explain the 2018 Brazilian Presidential election. The study identified a correlation between high HDI municipalities and votes for the winning party.

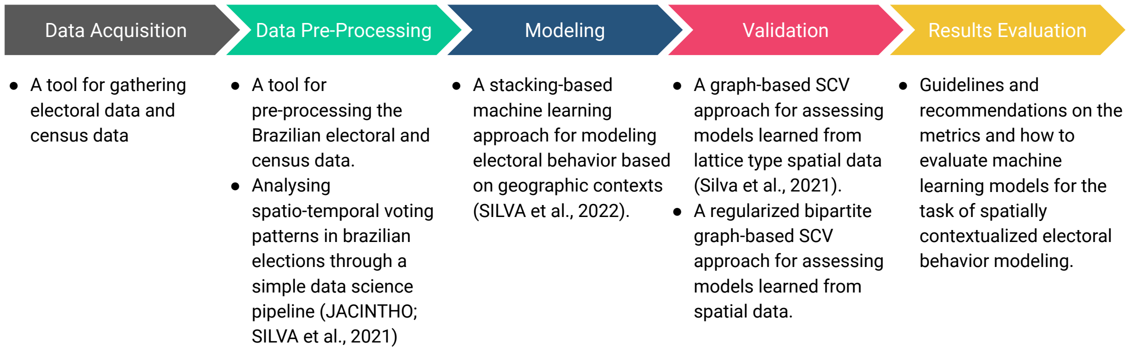
However, traditional machine learning methods ignore the spatial characteristics present in electoral data, such as spatial boundaries, clustering effects, and distance measures (GRAEFE; GREEN; ARMSTRONG, 2019; CHAUHAN; SHARMA; SIKKA, 2021). As a result, these methods tend to treat data separated into regions as independent and identically distributed, failing to capture the spatial dependence structure of the data. This approach, in turn, limits the ability of the models to identify patterns beyond the spatial data structure and thus impairs their ability to uncover new relationships that might aid in understanding the electorate's preferences.

Aligned with the issues mentioned above, the current thesis aims to study, propose, develop, and evaluate adaptations to the traditional machine learning pipeline making it suitable for spatially contextualized electoral behavior modeling. This research goal is expected to enhance the current status of machine learning applied to spatial data while also making a valuable contribution to the significant domain of electoral behavior modeling.

1.3 Main Contributions

The scientific challenges in machine learning applied to electoral behavior modeling are as interesting as their practical application. However, although the literature on electoral behavior modeling is extensive, the research community has only recently started using machine learning methods. From this perspective, we survey the current status of this research area and fill important gaps by proposing adaptations to the traditional machine learning pipeline and the guidelines for applying machine learning to the task of spatially contextualized electoral behavior modeling. From this perspective, Figure 2 presents a summary of the contributions of this research regarding each step of the traditional machine learning pipeline. Thus, this thesis advances the state-of-the-art in machine learning applied to spatial and electoral behavior modeling by proposing and developing tools, analysis and approaches that takes into consideration the spatial dependence when modeling the electorate preference using machine learning methods.

Figure 2 – A sumary of the contributions of this thesis based on each step of the traditional machine learning pipeline.



Data Acquisition

- A tool for gathering electoral data and census data

Data Pre-Processing

- A tool for pre-processing the Brazilian electoral and census data.
- Analysing spatio-temporal voting patterns in brazilian elections through a simple data science pipeline (JACINTHO; SILVA et al., 2021)

Modeling

- A stacking-based machine learning approach for modeling electoral behavior based on geographic contexts (SILVA et al., 2022).

Validation

- A graph-based SCV approach for assessing models learned from lattice type spatial data (Silva et al., 2021).
- A regularized bipartite graph-based SCV approach for assessing models learned from spatial data.

Results Evaluation

- Guidelines and recommendations on the metrics and how to evaluate machine learning models for the task of spatially contextualized electoral behavior modeling.

Source: Elaborated by the author.

Moreover, as far as we know, this thesis is the first to address the issues related to applying machine learning models learned from spatial data to model the electorate preferences and delineate the algorithms and guidelines to avoid erroneous results and missintepretation. Additionally, this research has a significant social impact by assisting on the identification of new patterns and relationships to understand election results. Finally, the authors have published nine scientific papers, listed in Chapter 6, where five were published and submitted as a direct result of this thesis.

In more detail, the main scientific novelty points in this thesis are summarized as follows:

- We developed tools for pre-processing and geocoding census data and electoral data. From these tools, we generated datasets with geocoding information regarding the polling places and cleaned and normalized datasets for all kinds of geographical aggregation levels regarding the census data. Our datasets present integrity measures that indicate confidence in the geocoding information. Finally, all the codes regarding the pre-processing and geocoding tools as well as the codes related to the validation and machine learning approaches, are publicly available at <https://github.com/tpinhoda>.

- We conducted a comprehensive analysis of the Brazilian electoral data at the municipal level from 1998 to 2018 using a standard data science pipeline consisting of five main steps: data selection, data pre-processing, identification of spatial patterns, identification of temporal patterns, and evaluation of results. The study focused on the presidential elections and analyzed the period's most prominent left and right parties: the Workers' Party (PT) and the Brazilian Social Democracy Party (PSDB). Additionally, we examined the election data of congressmen affiliated with parties ideologically positioned on the left and right of the political spectrum. Our findings suggest the existence of high spatial dependence in all the investigated electoral years. Furthermore, despite changes in the political and economic context over the years, neighboring municipalities presented similar voting behavior.

- We proposed and developed Spatial Cross-Validation techniques for the assessment of machine learning methods applied to the task of spatially contextualized electoral behavior modeling. Our approaches take advantage of the spatial graph structure provided in the dataset to define test-dependent buffer regions. Moreover, it allows the spatial folds to be defined as pre-existent geographical boundaries, facilitating the interpretation of the results by higher-level stakeholders. We evaluated the effectiveness of our approaches by considering three real datasets related to the 2018 Brazil Presidential Election, the 2019 New South Whales Congress Election, and the 2020 United States Presidential Election. Results show that our approaches contribute to the fair evaluation of models by enabling more realistic and local modeling.

- We proposed and developed a stacking-based machine learning approach for the task of spatially contextualized electoral behavior modeling. Our approach models data in spatial contexts of different dimensions and operates on them at two levels. First, it captures local patterns extracted from spatial contexts. Then, at the meta-level, it globally captures information from the K contexts nearest to a region we want to predict. We estimated the performance of our proposal concerning baseline and reference models taking data from the second round of the 2018 Brazilian presidential elections into account. Our approach presented the best overall performance, being able to generalize better than the compared ones. It also provided intelligible and coherent predictions in challenging regions, emphasizing its interpretability.

- We introduced the first guidelines for the assessment of machine learning methods learned from thousands of spatial features for the task of spatially contextualized electoral behavior modeling. Our guidelines involve recommendations on the use of four individual performance measures: Mean Squared Error (MSE), Mean Error Standard Deviation (MESD), SPearman correlation (SP), and SPearman correlation Standard Deviation (SPSD). MSE expresses the approaches' performance in predicting the correct vote-share scale. At the same time, SP assesses the order of the predictions yielded by the approaches. We also used a Multi-Criteria Performance Measure (MCPM) to combine the four metrics mentioned above and thus guide the choice of adequate models. Finally, to understand the predictive power of the models, we employed the SHapley Additive exPlanation (SHAP) Values technique to analyze the results generated and capture new relationships.

1.4 Practical Significance

We evaluated the approaches developed in this thesis using electoral and census data from different elections, including Brazil, Australia, and the United States of America. The collection of spatial geocoded datasets with census and electoral features in different levels of aggregation are other contributions of this research. To the best of our knowledge, the literature has a few benchmark datasets for the task o machine learning applied to spatially contextualized electoral behavior modeling.

Moreover, the findings of this study are important both from a theoretical-methodological point of view and a practical perspective. Our research provides guidelines for applying machine learning methods to spatially contextualized electoral behavior modeling. In addition, the approaches developed in this thesis can be generalized to other real-world problems involving spatial data.

Finally, from a social perspective, our findings can enrich our understanding of electoral processes. Our research aids in the exploratory analysis of the electorate preferences concerning their geographical context and socioeconomic characteristics. From this thesis, researchers will be able to expand their analysis and identify new patterns and relationships to help understand election results. We hope that the experimental protocols outlined and the different conclusions obtained can be used as a reference to guide future studies on the subject.

1.5 Thesis Outline

This dissertation is presented as a compilation of papers or manuscripts outlining our research contributions. The contents of each chapter are based on the original papers and have undergone minor adaptations to conform to the thesis format. While the chapters can be read in any order, their arrangement reflects the progress of scientific developments. In the following subsections, we provide a brief overview of each chapter presented in this thesis.

1.5.1 Chapter 2

Title: *Analyzing Spatio-Temporal Voting Patterns in Brazilian Elections Through a Simple Data Science Pipeline*.This chapter is based on a paper with shared co-first authorship (JACINTHO *et al.*, 2021). This chapter is based on a paper published on the Journal of Information Data Management. The contributions of this author regards: writing, revision, bibliometric analysis, data analysis, and plots production. It is important to emphasize that this publication involved collaboration with an undergraduate researcher, and the author of this thesis was the responsible for the research.

In this chapter, we undertake a comprehensive analysis of the Brazilian presidential and congressional elections since the first democratic election after a period of dictatorship. Our investigation adopts a spatial-temporal perspective, examining the election results at the municipal level. Specifically, we focus on the two most significant political parties in the period, namely the Workers' Party (PT) and the Brazilian Social Democracy Party (PSDB), analyzing the presidential election outcomes. In addition, we study the congressional election data by considering parties located ideologically to the left and right in the political spectrum. Our results reveal the presence of spatial dependence in every electoral year investigated, indicating the existence of interdependence among municipalities. Notably, despite the changes in the political-economic context over the years, neighboring cities exhibit similar voting behavior trends, highlighting the existence of spatial effects that shape electoral outcomes.

1.5.2 Chapter 3

Title: *A Graph-Based Spatial Cross-Validation Approach for Assessing Models Learned with Selected Features to Understand Election Results* (SILVA; PARMEZAN; BATISTA, 2021). This chapter is based on a paper published on the Proceeding of the International Conference of Machine Learning Applications. The contributions of this author regards: data collection, implementation, writing, revision, data analysis, and plots production.

In this chapter, we propose a novel approach to fairly evaluate machine learning methods for the task of spatially contextualized electoral behavior modeling. We introduce a graph-based spatial cross-validation technique that utilizes the spatial graph structure of lattice-type spatial objects to create local training sets for each test fold. Our approach removes highly correlated spatially close data and irrelevant distant data that may impact error estimates, thereby enabling more realistic and local modeling. Experiments using data from the second round of the 2018 Brazilian presidential election demonstrate that our approach leads to fairer evaluations of models. Our proposal can potentially improve the validity and reliability of the machine learning applied to spatially contextualized electoral behavior modeling area.

1.5.3 Chapter 4

Title: *Geographic Context-Based Stacking Learning for Election Prediction from Socio-Economic Data* (SILVA; PARMEZAN; BATISTA, 2022). This chapter is based on a paper published on the Proceeding of the Brazilian Conference on Intelligent Systems. The contributions of this author regards: data collection, implementation, writing, revision, data analysis.

In this chapter, we present a new machine learning approach to model electoral behavior by considering geographical contexts. Moreover, we introduce adaptations to the traditional machine learning pipeline that considers a Spatial Cross-Validation technique to evaluate our approach against existing literature and baselines fairly. Our approach models data in spatial contexts of different dimensions and operates at two levels. First, it captures local patterns extracted from spatial contexts, and then, at the meta-level, it globally captures information from the K contexts nearest to a region of interest. Our experiments demonstrate that our approach outperforms the compared methods and is able to generalize better. Furthermore, our proposal provides interpretable and coherent predictions in challenging regions, highlighting its potential use in supporting social research. The proposed approach is shown to be effective in modeling electoral behavior in a spatially contextualized setting.

1.5.4 Chapter 5

Title: *Learning Beyond the Spatial Dependence Structure: A Spatial Cross-Validation Approach for Assessing Models Learned from Census Data to Understand Elections*. This chapter is based on a manuscript submitted to the International Journal of Geographical Information Systems. The contributions of this author regards: data collection, implementation, writing, revision, data analysis, and plots production.

In this chapter, we present a novel and more general Spatial Cross Validation approach that allows for the assessment of machine learning models learned from various types of spatial data. Our proposed approach recognizes that pre-existing geographical boundaries may define spatial folds and that spatial dependence can exist non-contiguously across space. To isolate the test set from the training set, we calculate a buffer region by considering a bipartite graph structure and formalize the problem as a one-class transductive classification task. Additionally, we propose a label propagation method that integrates the semivariogram technique to classify nodes in the training set as part of the removing buffer region. To evaluate our approach, we use three case studies involving presidential and congressional elections in Brazil, Australia, and the United States. The experiments show that our approach yields less biased and more realistic results in the presence of spatial dependence, making it a suitable tool for identifying new patterns and relationships.

1.5.5 Chapter 6

In this chapter, we conclude our research work by summarizing the key findings discussed in the preceding chapters. We also highlight the limitations of our approaches and identify potential avenues for future research. Moreover, we finish this chapter by listing throughout all the papers produced during the thesis development period.

CHAPTER 2

ANALYZING SPATIO-TEMPORAL VOTING PATTERNS IN BRAZILIAN ELECTIONS THROUGH A SIMPLE DATA SCIENCE PIPELINE

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, this chapter 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.

2.1 Introduction

The constitution promulgated during the New Republic represents an important moment of re-democratization 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 hyper-inflationary 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; 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; DEMŠAR, 2015). The last one has been increasingly studied over the years (MANSLEY; DEMŠAR, 2015; AGNEW, 1996). The geographical space, in turn, plays a key role in electoral processes, as pointed out initially by Agnew (1996) in his multidimensional place-centered perspective on political behavior and later by 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; MENEZES, 2015; MARZAGÃO, 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; KAMBER; PEI, 2011). In this chapter'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; 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 chapter 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 chapter is organized as follows: section 2.2 reports a comprehensive review of the literature along with a meta-analysis of related work; section 2.3 describes our data science pipeline for analyzing spatio-temporal voting patterns in Brazilian elections; section 2.4 compiles the results and discusses them; finally, section 2.5 concludes the study with some comments on future research directions.

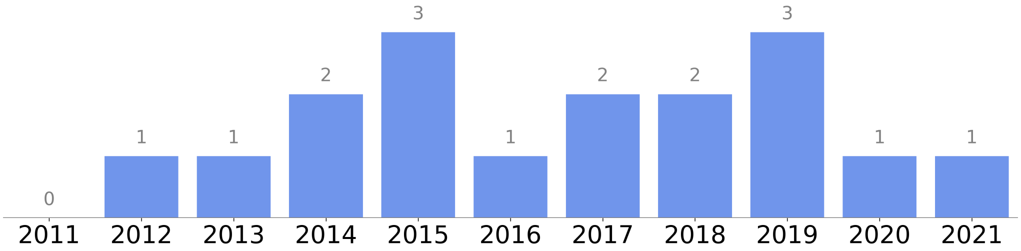
2.2 Related Work

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; 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; LEóN, 2020; OKUNEV; GORELOVA; GRUZDEVA, 2020; REID; LIU, 2019), the study of the role of space in elections (MOTA, 2019; MANSLEY; DEMŠAR, 2015), the analysis of ideological trends (ZUCCO; POWER, 2020; FAUSTINO *et al.*, 2019; POWER; RODRIGUES-SILVEIRA, 2019), and the voting behavior on social media (RECUERO; SOARES; GRUZD, 2020; PRACIANO *et al.*, 2018).

This chapter 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

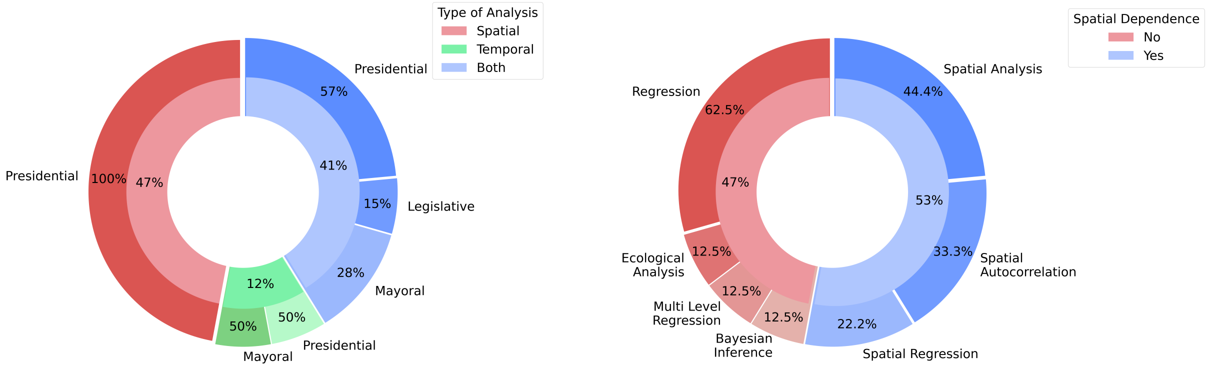
[1] <https://github.com/LucasManto/analyzing_brazil_presidential_elections>.

Figure 3 – Distribution of selected publications by year.



Source: Elaborated by the author.

Figure 4 – Overview of the meta-analysis results.



(a) Publications by type of analysis.

(b) Publications that regard spatial dependence.

Source: Elaborated by the author.

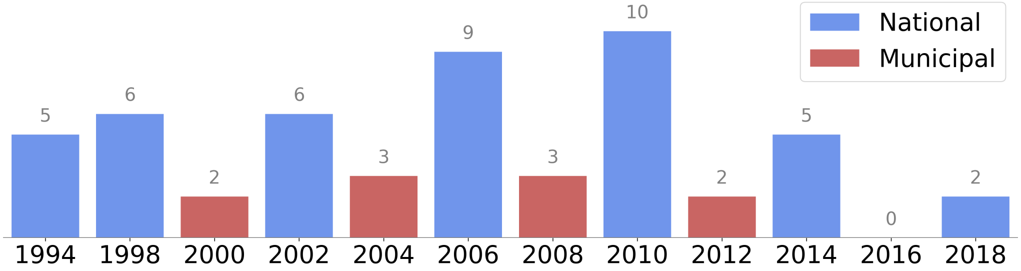
followed by a meta-analysis. We defined a global search expression[2] and submitted it to six electronic databases[3] to retrieve pieces of work published in the last ten years. We adopted two criteria for selecting publications: (i) the study must cover election results provided by the Brazilian Superior 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.

Figure 3 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 work 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.

[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.

Figure 5 – Number of scientific analyzes per election year.



Source: Elaborated by the author.

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. Figure 4a 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. Figure 4b 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, Figure 5 displays the number of times that the identified papers addressed a given electoral year. In addition 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 work. 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, Silva and Dias (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

[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.

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; 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 1 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.

Table 1 – 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; MENEZES, 2015)	2006 - 2010	✓	—	—	—
(MAGALHÃES; SILVA; DIAS, 2015)	2010	✓	—	—	—
(CORRÊA, 2015)	2006	✓	—	—	—
(MARZAGÃO, 2013)	2010	✓	—	—	—
(TERRON; SOARES, 2010)	1994 - 2006	✓	✓	—	—
This paper	1998 - 2018	✓	✓	✓	✓

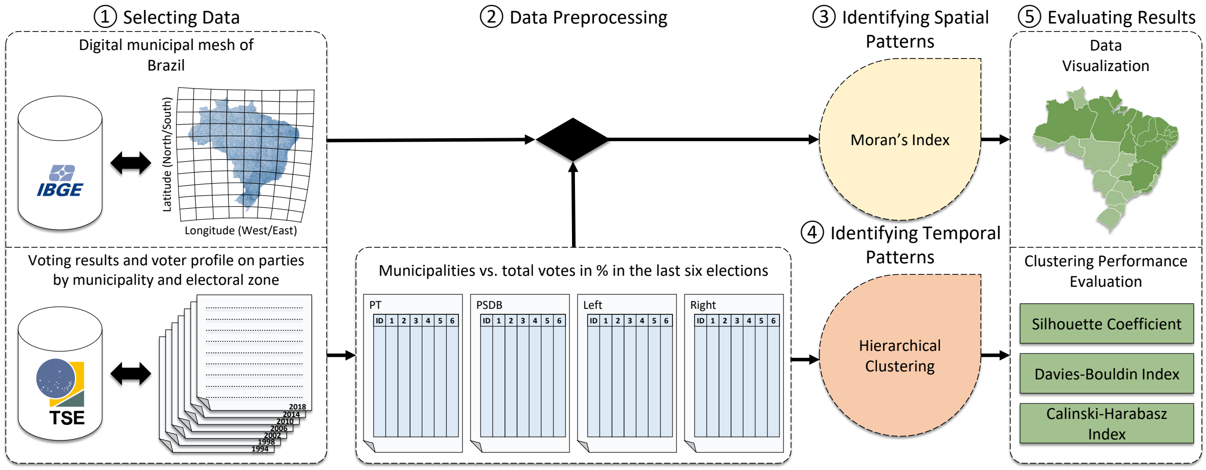
Source: Research data.

As summarized in the last row of Table 1, 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.

2.3 Materials 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. Figure 6 organizes the pipeline in five steps, which we will describe in detail throughout subsection 2.3.1–subsection 2.3.5.

Figure 6 – 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).



Source: Elaborated by the author.

Table 2 – 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	✓	✓	✓	✓	✓

Source: Research data.

We implemented the data science pipeline of Figure 6 employing the Python[5] programming language combined with the following libraries: Pandas[6], GeoPandas[7], SciPy[8], PySAL[9],

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

[6] <https://pandas.pydata.org/>.

[7] <https://geopandas.org/>.

[8] <https://www.scipy.org/>.

[9] <https://pysal.org/>.

and Scikit-learn[10]. Table 2 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].

2.3.1 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 Figure 6, 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 3 exhibits the attributes addressed in the present study.

Table 3 – 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

Source: Research data.

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.

[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/>.

2.3.2 Data Preprocessing

Our data preprocessing consists of four processes (Step 2 of Figure 6): (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 chapter, 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` 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 4).

Table 4 – 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

Source: Research data.

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; POWER, 2020). The classification of a given party's ideological spectrum is based on its ideological score (ZUCCO; 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 5 exhibits the analyzed parties and their respective classification.

Table 5 – 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)

Source: Adapted from Zucco and Power (2020).

2.3.3 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, which states that "everything is related to everything else, but near things are more related than distant things" (TOBLER, 1970). 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; GETIS, 1992). In this work, more precisely in Step 3 of Figure 6, we used one of the most popular methods developed to assess the spatial autocorrelation (LI; CALDER; CRESSIE, 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 2.1 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, $\bar{x}$ 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.

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.1}$$

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 2.2 shows how the index is calculated. Likewise the global method, x_i is the observed value at location i, $\bar{x}$ 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 2.3. Again, the locations are the Brazilian municipalities, and the weight strategy adopts the Queen method.

$$I_i = \frac{x_i - \bar{x}}{S_i^2} \frac{\sum_{j=1, j \neq i}^{n} w_{i,j}(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.2}$$

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^{n} (x_j - \bar{x})^2}{n - 1} \tag{2.3}$$

2.3.4 Identifying Temporal Patterns

In Step 4 of Figure 6, 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; 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 chapter 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; 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 (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.

2.3.5 Evaluating Results

In Step 5 of the pipeline illustrated in Figure 6, 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; Bouldin, 1979), and Calinski-Harabasz index (CALIńSKI; 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 2.4 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 2.5, 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.

$$S = \frac{b-a}{\max(a,b)} \qquad (2.4) \qquad\qquad DB = \frac{1}{n}\sum_{i-1}^{n} \max_{i \neq j} \left[\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right] \qquad (2.5)$$

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 2.6 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.

$$s = \frac{SS_B}{SS_W} \times \frac{n_E - k}{k-1} \qquad (2.6)$$

$$SS_W = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)^2 \quad SS_B = \sum_{q=1}^{k} n_q (c_q - c_E)^2 \qquad (2.7)$$

2.4 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:

1. In what extensions of Brazilian territory neighboring cities exhibit similar vote distribution?
2. 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.

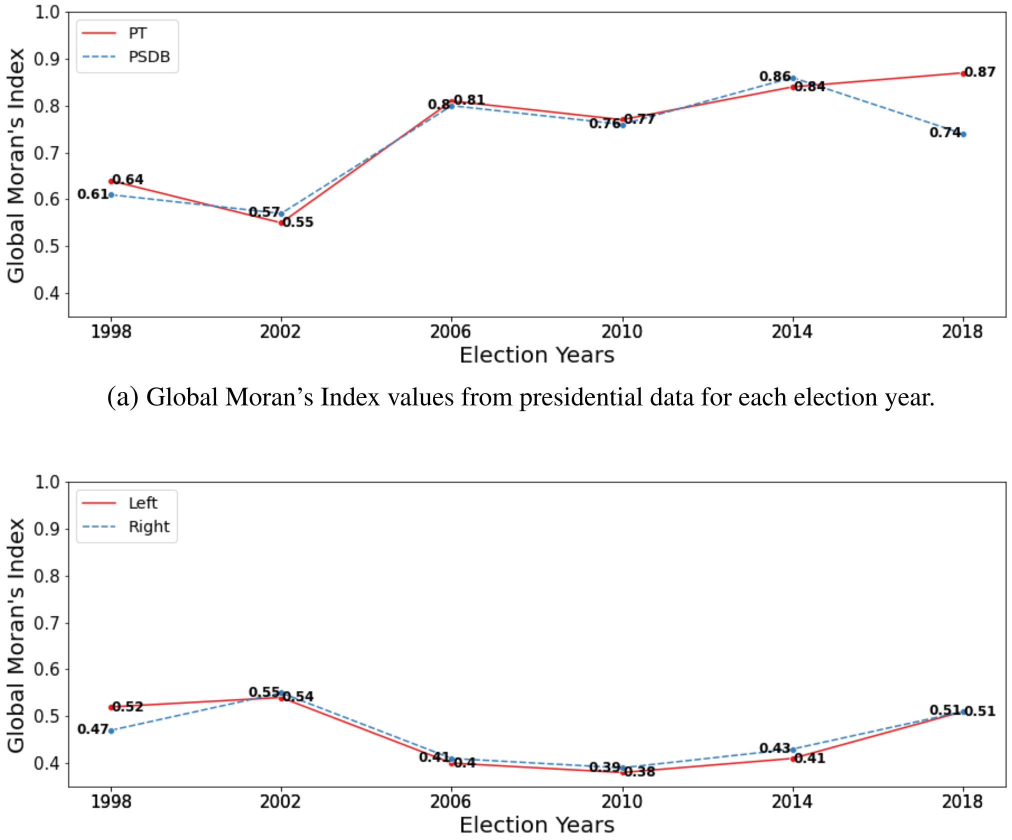
2.4.1 Analyzing Spatial Patterns

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 (Figure 7a) 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 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 (Figure 7b), 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. Figure 8 displays the results for presidential election data, while Figure 9 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



(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.

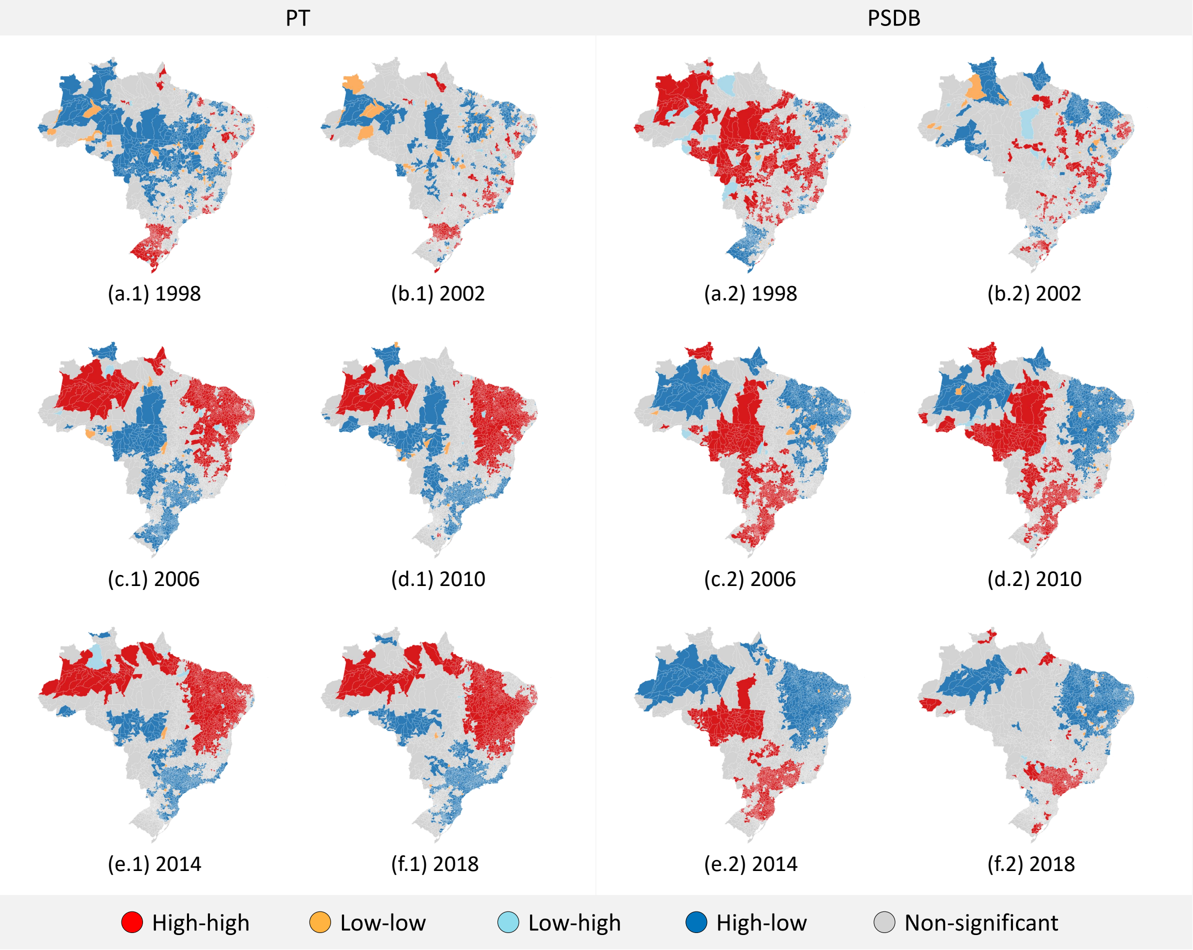
Figure 7 – Moran's Index results by electoral year.

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 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 (Figure 8 – 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

Figure 8 – Local Moran's Index plot of Workers' Party (PT) and Brazilian Social Democracy Party (PSDB) by election year.



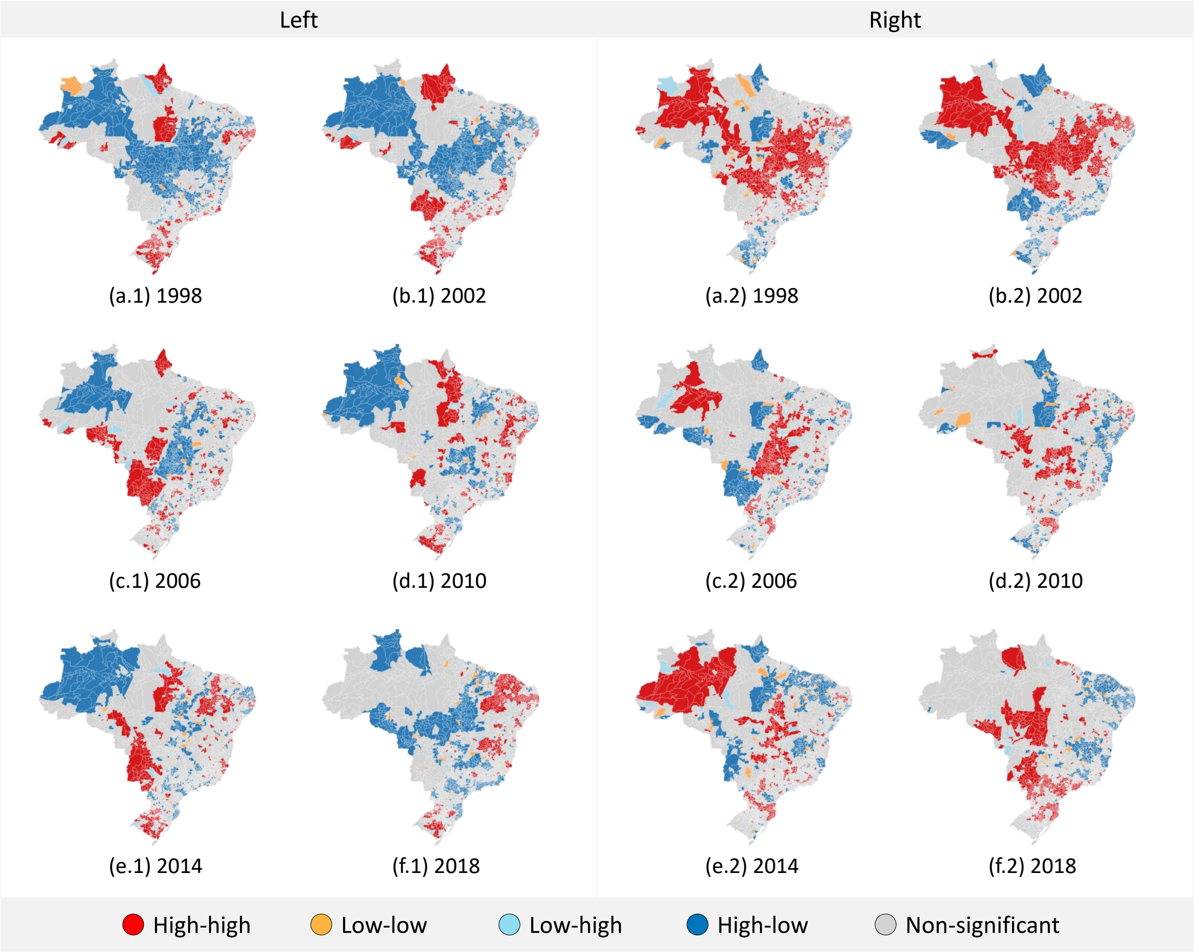
Source: Elaborated by the author.

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 (Figure 8 – 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 (Figure 8 – PSDB). The same regions presented a high number of cities colored in red for PT (Figure 8 – 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). Not surprisingly, it was the year when both parties were the most voted.

We followed the same approach applied in Figure 8 for the congress election data displayed in Figure 9. Concerning the parties on the right (Figure 9 – 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

Figure 9 – Local Moran's Index plot of left and right parties by election year.



Source: Elaborated by the author.

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 (Figure 9 – 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 (Figure 8 and Figure 9) over the years. On the other hand, PSDB and the right parties

Table 6 – 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

Source: Research data.

displayed an increase of hegemony in the Southeast and Southern. An inverse behavior can be observed in the Northeast with an increase of cities supporting PT and left parties and a reduction of cities supporting PSDB and right parties.

2.4.2 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 6 and 7 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 (Figure 10a.1 and Figure 10a.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 (Figure 10a.1), almost every city of the Northeast region belongs to the same group. On the other hand, considering the results for PSDB (Figure 10a.2), the majority of the Southeast cities belong to the same group.

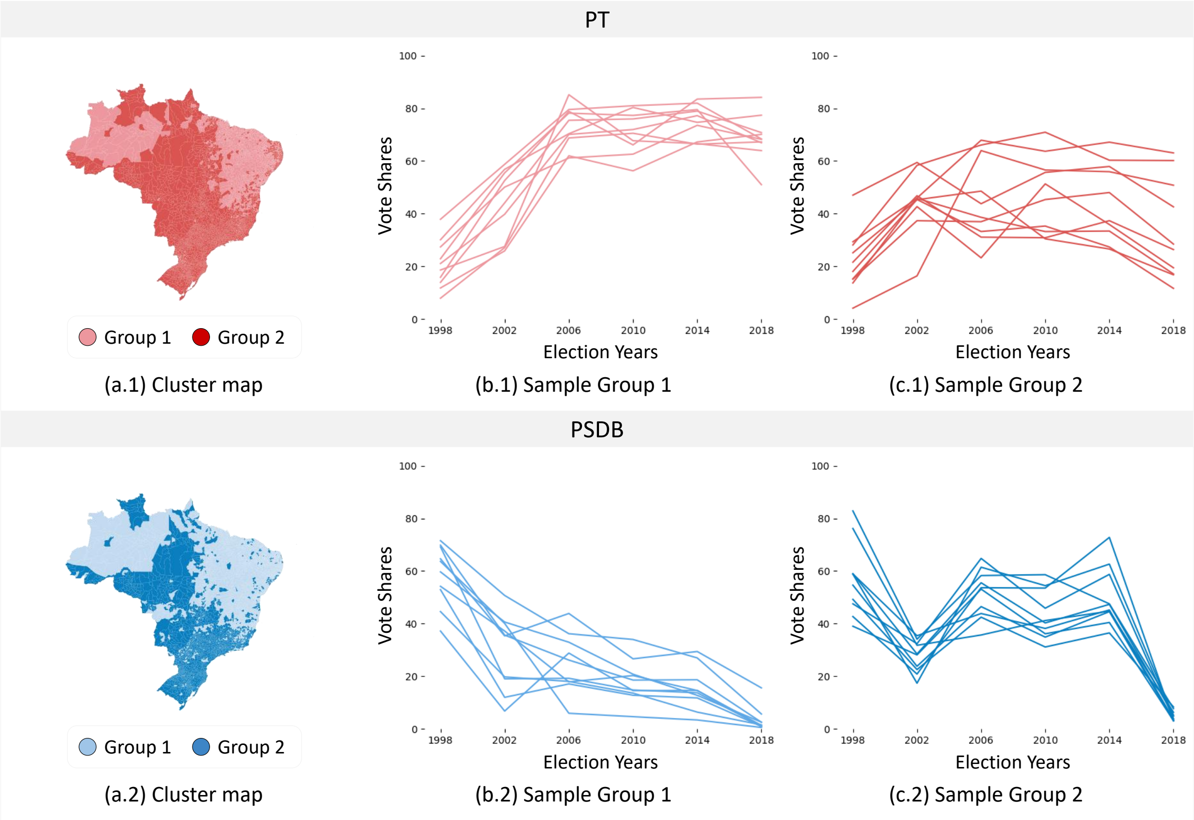
In more detail, Figure 10b.2 and Figure 10c.2 present randomly chosen samples of PSDB

Table 7 – 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

Source: Research data.

Figure 10 – Presidential election clustering maps results and voting series samples by groups.



(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

Source: Elaborated by the author.

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 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 (Figure 10b.1) features an increasing trend from 1994 to 2006, stabilization in 2010 and 2014, and a decrease in 2018. For group 2 (Figure 10c.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 (Figure 11a.1 and Figure 11a.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.

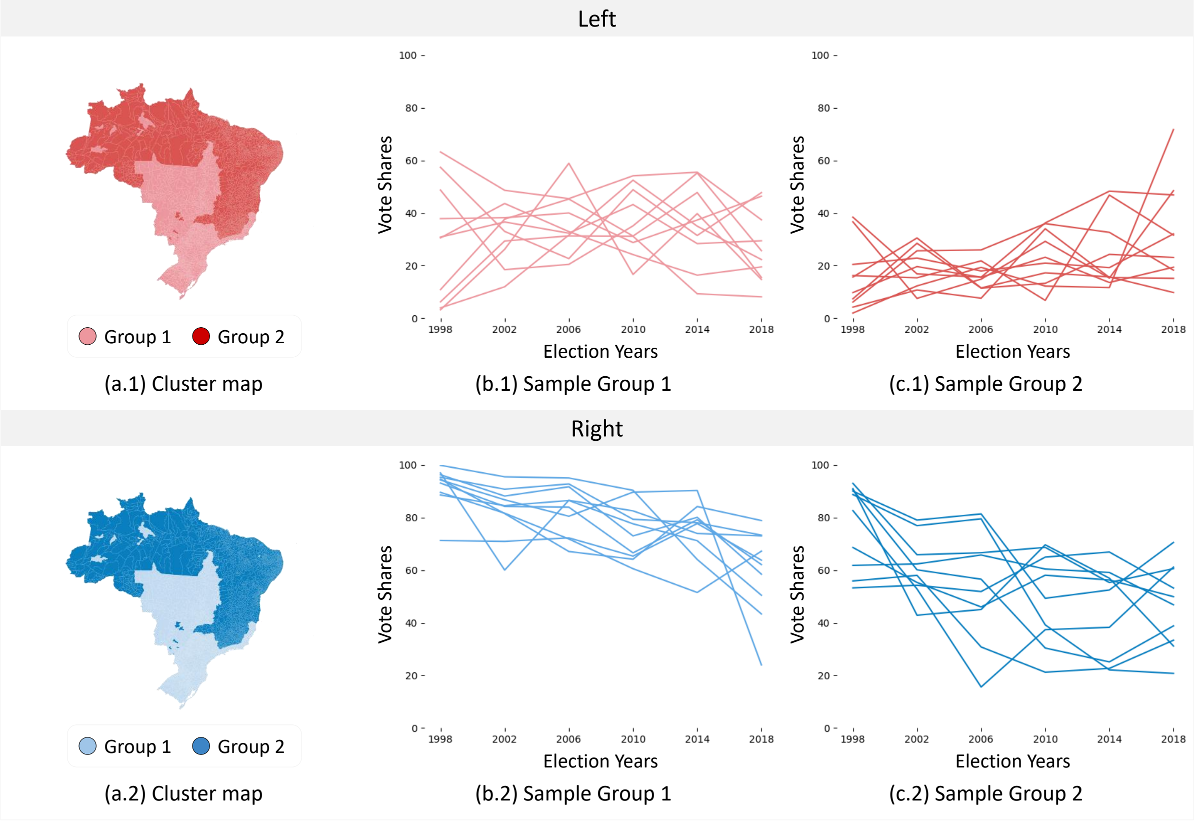
In more detail, Figure 11b.2 and Figure 11c.2 present randomly chosen samples of right parties' vote-shares time-series from cities belonging to groups 1 and 2, respectively, selected following the same criterion described for the presidential elections. The cities from group 1 exhibit a small decrease 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 (Figure 11b.1 and Figure 11c.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.

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.

2.5 Conclusion

This chapter 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

Figure 11 – Congress election clustering maps results and voting series samples by groups.



(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

Source: Elaborated by the author.

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 chapter 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.

CHAPTER

3

A GRAPH-BASED SPATIAL CROSS-VALIDATION APPROACH FOR ASSESSING MODELS LEARNED WITH SELECTED FEATURES TO UNDERSTAND ELECTION RESULTS

Elections are complex activities fundamental to any democracy. The contextualized analysis of election data allows us to understand electoral behavior and the factors that influence it. Multidisciplinary studies have been prioritized the predictive modeling of electoral features from thousands of explanatory features, considering geographic and spatial aspects inherent to the data. When building a model for such a purpose, it must be rigorously evaluated to understand its prediction error in future test cases. Although cross-validation is a widely used procedure for this task, it leads to optimistic results because the spatial independence between test and training data is not ensured in the resampling. On the other hand, alternatives to deal with spatial dependence may fall into a pessimistic scenario by assuming total spatial independence between the test and training sets regardless of the size of the first one, increasing the probability of overfitting. This chapter addresses these issues by proposing a graph-based spatial cross-validation approach to assess models learned with selected features from spatially contextualized electoral datasets. Our approach takes advantage of the spatial graph structure provided by the lattice-type spatial objects to define a local training set to each test fold. We generate the local training sets by removing spatially close data that are highly correlated and irrelevant distant data that may interfere with error estimates. Experiments involving the second round of the 2018 Brazilian presidential election demonstrate that our approach contributes to the fair evaluation of models by enabling more realistic and local modeling.

3.1 Introduction

Democratic countries define electoral laws to protect people's preferences and convert them into a set of objective social priorities. Under this condition, elections can provide the best expression of electoral opinion and party involvement. However, especially in new and fragile democracies, turning votes into political power can generate intentions to manipulate the majority decision, in the worst scenario leading to political polarization and allegations of electoral frauds (LEHOUCQ, 2003).

Rising allegations of political polarization and electoral fraud have boosted concerns about the integrity of democratic elections (NORRIS, 2013). In this context, multidisciplinary studies point to a growing body of research interested in understanding the factors influencing electoral results (JACINTHO *et al.*, 2021). Scientific efforts are concentrated, but not limited, to investigating external factors that impact the popular vote (REID; LIU, 2019), analyzing ideological trends (FAUSTINO *et al.*, 2019), and studying elections on social networks (RECUERO; SOARES; GRUZD, 2020).

Additionally, geographers consider that the electoral processes are explainable by the population characteristics of the locations at which they occur (MANSLEY; DEMŠAR, 2015). Thus, an electoral process comprises elements that indicate local patterns related to spatial autocorrelation and the so-called spatial non-stationarity (CHARNEY; MALKINSON, 2015). From this perspective, people belonging to the same region should present similar voting behavior, while distinct areas may have different vote distributions.

Considering how people are geographically contextualized and the data's spatial characteristics can enrich our understanding of electoral processes. With this in mind, few researchers have been dedicating themselves to developing non-linear models capable of predicting electoral features from thousands of explanatory features, taking geographic and spatial aspects intrinsic to the data into account (JACINTHO *et al.*, 2021). Such models, when deployed, need to be fairly evaluated to understand their prediction error in future test cases. Cross-validation is a widely used method for this task. However, it provides optimistic results because resampling does not guarantee spatial independence between test and training data (ROBERTS *et al.*, 2017). In contrast, alternatives to deal with spatial dependence may fall into a pessimistic procedure that assumes total spatial independence between the test and training samples regardless of the size of the first one, increasing the probability of overfitting (PLOTON; MORTIER *et al.*, 2020; VALAVI *et al.*, 2019; ROBERTS *et al.*, 2017).

In this chapter, we work around the limitations of such methods in the context of our application by introducing a graph-based spatial cross-validation approach designed to assess models learned with selected features from spatially contextualized electoral datasets. Our method differs from existing ones for the following reasons: (i) it considers that spatial folds can be defined using pre-existing geographical boundaries (*e.g.*, states and countries) that were not designed for such purpose. Hence, the data inside each fold may not present similar characteristics; (ii) it represents lattice-type spatial objects (*e.g.*, cities and districts) as undirected graphs. This structure provides neighborhood information regardless of the size and shape of the spatial object, facilitating the identification of spatial dependence; (iii) it defines to each test fold a local training set by removing spatially close data that are highly correlated and irrelevant distant data that may interfere with error estimates; (iv) it is publicly available[1], simple to implement, and has a quadratic-order time complexity.

By analyzing data from the second round of the 2018 Brazilian presidential election, we demonstrate that our approach defines buffer regions that decrease the spatial dependence between test and training samples. Furthermore, by selecting instances based on the spatial closeness and similarity in the explanatory feature space regarding the test fold, our method provides a more realistic and local data modeling.

The remainder of this chapter is organized as follows. section 3.2 introduces the fundamentals related to cross-validation under spatial dependence. section 3.3 describes our graph-based spatial cross-validation approach. section 3.4 presents the case of study involving data from the second round of the 2018 Brazilian presidential election. Finally, section 3.5 points out the achievements and remaining challenges of our research line.

3.2 Background

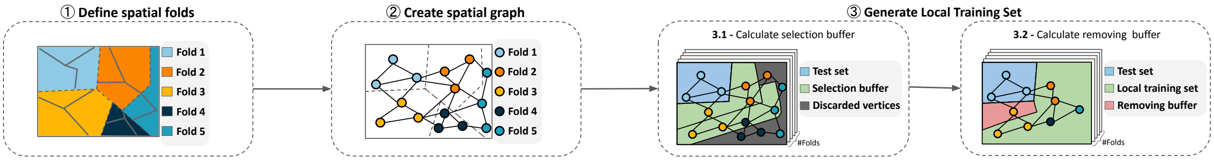
This section describes the mathematical setup and problems involved in assessing models learned from selected features to understand election results. It also discusses how to estimate these models with spatially dependent data.

3.2.1 Concepts and Definitions

First, let O be a set of lattice-type spatial objects (*e.g.*, neighborhoods, districts and cities), such that a polygon delimits each object in the spatial domain; the spatial intersection between two distinct objects o_i and $o_j \in O$ is the empty set ($\emptyset$), and the spatial union of O forms a contiguous study area. Now, let us assume D a spatial dataset composed by $Y \in \mathbb{R}$ the target attribute that describes the vote-shares for each spatial object in O from a given candidate or party, and $X \in \mathbb{R}^d$, where $d > 0$ the explanatory feature space that describes the spatial objects from O in other related explanatory contexts. Illustratively, O can be the set of cities (o_i), such that $\cup(O)$ forms a country, *i.e.*, the area of study. $Y \in \mathbb{R}$ is the outcome space containing vote-shares distribution for a given candidate, and $X \in \mathbb{R}$ is the explanatory feature space that can have the demographic information for the cities.

[1] <https://github.com/tpinhoda/Graph-Based_Spatial_Cross_Validation>.

Figure 12 – Pipeline of the proposed graph-based spatial cross-validation.



Source: Elaborated by the author.

Following the classical Tobler's first law of geography, which states that "everything is related to everything else, but near things are more related than distant things" (TOBLER, 1970), the spatial dataset D presents spatial dependency between observations of features at different locations, where nearby locations are more likely to present similar values than distant ones. This is a fundamental characteristic of spatial data, named spatial autocorrelation, which became the basis of subsequent research in spatial data analysis and related areas (GETIS, 2010). Its existence requires us to consider the dependency assumption as a condition in evaluating predictive models to avoid erroneous results.

Mindless of the spatial autocorrelation of the data in D, machine learning and feature selection methods can be evaluated using the traditional k-fold cross-validation approach. The evaluation takes place by generating k folds by randomly sampling the dataset. Each fold serves as a test set once, while the others are used as the training set. In this context, we might have neighboring spatial objects separated in the test and training sets. However, feature selection and machine learning algorithms will yield biased results if applied directly to election data due to the spatial autocorrelation characteristic they present.

Therefore, there is a need for sampling approaches that fairly assess learned models with selected features from spatially dependent data. Such approaches are crucial in analyzing massive electoral results since they lead to more reliable and realistic results.

3.2.2 Estimating Models from Spatially Dependent Data

In response to the problem of spatial autocorrelation, methods to evaluate models under spatial dependence were proposed (PLOTON; MORTIER *et al.*, 2020; VALAVI *et al.*, 2019; ROBERTS *et al.*, 2017). In general, they split the dataset into k spatial folds. The data from each fold can be a block of similar data in the spatial domain (VALAVI *et al.*, 2019) or a single spatial object (ROBERTS *et al.*, 2017). The definition of test and training sets occurs in the same manner as in the traditional cross-validation.

Typically, the spatial folds are defined automatically by running a clustering algorithm taking the target feature and the spatial information as the input. Another strategy considers the monitoring of the semivariogram on the residuals of the fitted model to define the size of the spatial folds (VALAVI *et al.*, 2019). Thus, the data in the spatial folds will be more related to each other, reducing the chances of having spatial dependence between folds. Furthermore, to increase independence between the test and training set, a removing buffer region can be calculated to isolate them (ROBERTS *et al.*, 2017). The buffer region is part of the training set that is too close spatially and highly correlated to the test data such that its presence in the training set can generate biased results. Usually, the removing buffer region is defined manually or using the semivariogram applied on the target attribute or the residuals of a fitted model.

The semivariogram calculates the spatial variance (γ) from a given feature regarding a determined lag distance. Equation 3.1 describes how this calculation is done, where $C(h)$ is the set of all spatial objects pairs $(o_i, o_i + h)$ separated by the distance h, and $z(o_i)$ is the value of feature z from spatial object o_i. The first lag distance h where γ reaches the sill—variance of feature z over the entire dataset—is generally used to define the size of the removing buffer.

$$\gamma(h) = \frac{1}{2|C(h)|} \sum_{i=1}^{C(h)} (z(o_i) - z(o_i + h))^2 \tag{3.1}$$

From our application perspective, when considering pre-existing geographical boundaries as delimitation for spatial folds, we cannot guarantee that the data distribution is similar inside each fold and present lower spatial dependence cross folds. This characteristic arises because such pre-existing boundaries were not defined for the purpose to delimit spatial folds. In this scenario, buffer regions play an essential role in ensuring independence between test and training sets. However, approaches that define buffer regions based on the residuals of a model imply having fitted the model already, which contrasts with our goal to have the model adjusted after determining the test and training sets. Besides, manually defining buffer regions requires domain knowledge and can be time-consuming. On top of it, the buffer regions present the same size for all test folds and isotropic property—same size in all directions—, and are created independent of the test set (ROBERTS *et al.*, 2017). These characteristics may lead to improper buffer regions by removing unnecessary data or leaving data with high correlation. When applied to data with high spatial autocorrelation, it prioritizes the independence of the test and training sets over the number of training instances, creating an ideal setting for the model to overfit.

3.3 Our Contributions

The existing model validation approaches do not suit our application correctly. Their four limitations are:

1. Spatial folds may be determined by a pre-existing spatial delimitation that was not designed for this purpose;
2. The lattice-type nature of spatial data makes defining distance lags for the semivariogram a challenge since the spatial objects can present different sizes and shapes;
3. Determining a test set independent buffer region that has the same size in all directions is not a realistic assumption;

4. For continental region spatial datasets, distant data from the test set can be irrelevant to the modeling.

From this perspective, we propose a graph-based spatial cross-validation method that takes advantage of the spatial graph structure from lattice-type spatial objects to define a local training set to each test fold by removing spatially close data that are highly correlated and irrelevant distant data that may interfere with error estimates. Figure 12 portrays the pipeline of the validation approach we propose. First, we consider pre-existing geographical delimitations to respond to higher-level stakeholders analysis (Step 1). Then, we create the spatial graph representation from the dataset to use the neighboring information as lag distance in the semivariogram, making the approach independent of the size and shape of the dataset's spatial objects (Step 2). Finally, we generate a local training set for each test fold using the dataset graph representation and the semivariogram (Step 3). Creating a local training set per fold is a two-step process: first, we calculate a selection buffer by considering the similarity in the explanatory feature space and proximity in the spatial domain (Step 3.1). Then, we calculate the removing buffer over the selection buffer to exclude spatially high correlated data regarding the outcome feature space (Step 3.2). Thus, the local training set is constituted by the data in the selection buffer region except those in the removing buffer region. The following subsections describe these steps in detail.

3.3.1 How to Define Spatial Folds?

This subsection discusses the definition of folds to evaluate models learned from spatial data in the context of our application (Step 1 of Figure 12). As reported throughout this chapter, this type of data requires unique treatment due to the spatial autocorrelation. Thus, a spatial fold is not defined in the same way as a fold in conventional cross-validation. The spatial folds are not composed of randomly sampled instances. Instead, existing methods design the boundaries of the spatial folds based on the data similarity in the spatial domain as a manner to diminish spatial dependence across folds (VALAVI *et al.*, 2019). Others use a single spatial object as a fold with a fixed size and isotropic buffer to ensure independence between the test and training sets (ROBERTS *et al.*, 2017).

Even though we can apply such techniques, they may be time-consuming in the case of a single spatial object or generate non-existing spatial boundaries. Concerning our application, we argue that the best spatial fold delimitations are the pre-existing ones because, at a higher level, stakeholders are more interested in understanding the election results considering a known geographic boundary, such as states and countries. In this regard, when evaluating models to understand election results, we highly recommend using such boundaries to facilitate the result's interpretation. The boundaries, however, are not created to delimit spatial folds. Thus, we must define test fold dependent buffer regions to isolate the test set from the training set correctly.

3.3.2 How to Represent Lattice-Type Spatial Objects?

In this subsection, we explain our approach to representing spatial data as an undirected graph (Step 2 of Figure 12). Traditionally, spatial data can be represented by three types of models (FISCHER, 2006): (i) point patterns, (ii) continuous surfaces, and (iii) lattices. Most recently, researchers began using graphs to describe such data to identify spatial patterns (LI *et al.*, 2019). According to this representation, a graph's vertex express a spatial object (*e.g.*, a state or a municipality). The edges, in turn, are built from the vertices' proximity in the geographical space or the explanatory feature space that describes the spatial objects.

In this work, we considered an undirected graph $G(V,E)$ in which V corresponds to the set of spatial objects under analysis and E is the set of edges indicating if two spatial objects share a geographical boundary. G contains no vertex with missing data and has a single major component, where it is possible to traverse G from one spatial fold to another. Such structure provides neighboring information independent of the spatial object size and shape. Besides, it does not rely on distance thresholds to define neighbors.

3.3.3 How to Generate Local Training Sets?

This subsection describes the procedures to generate the local training sets (Step 3 of Figure 12). Our approach comprises two steps: first, we apply a selection buffer to filter the training set based on distance and similarity on the explanatory feature space regarding the test set, allowing local modeling of the data. Then, we define a removing buffer to avoid spatially high-correlated data in the target feature that may generate biased results when evaluating models. This approach can be considered a more realistic strategy than the current ones since each buffer (selection and removing) depends on its test set spatial distribution rather than a unique buffer region template to all test sets.

Algorithm 1 describes the procedure to generate a local training set per test fold. First, we calculate the transformation of the explanatory feature space based on the Principal Component Analysis (PCA) by considering the component with the highest explained variance ratio (line 8). Then, for each test fold, we get the vertices from G related to the test and training sets (lines 10–11). Subsequently, we calculate the number of neighboring folds regarding the test fold, which will be used in the calculation buffer procedure (line 6). Next, we calculate the buffer selection by monitoring the PCA's first component values (lines 14–15) and the removing buffer by monitoring the target attribute values (lines 17–19). To calculate the removing buffer, we consider the induced graph G_s composed by the vertices in the selection buffer (B_S) and the test set V_{test}. Finally, the local training set for the test fold comprises all the data in the selection buffer minus the data in the removing buffer (lines 23–24).

The selection and removing buffer operate similarly. In summary, we propose a semivariogram approach based on the dataset spatial graph structure that considers the test set to compute

Algorithm 1 – Generate Local Training Sets

```
Require:
$D$: the dataset;
$G$: the dataset spatial graph;
$V$: the set of vertices;
$Y$: the set of target attribute values;
$X$: the set of explanatory attributes values;
$\kappa$: the relaxing factor for the number of folds in the buffer.

$PCA \leftarrow \text{PCA}(X, n = 1)$
For Each $fold \in D$ do
    $V_{test} \leftarrow V(fold)$
    $V_{training} \leftarrow V - V_{test}$
    $n_{test} \leftarrow \textsc{NeighboringFolds}(G, V_{test})$
    ** Calculating selection buffer **
    $Z \leftarrow \text{V}(PCA)$
    $B_S \leftarrow \textsc{CalculateBuffer}(..., type = \text{“selection”})$
    ** Calculating removing buffer **
    $G_s \leftarrow G[B_S \cup V_{test}]$
    $Z \leftarrow V(Y)$
    $B_R \leftarrow \textsc{CalculateBuffer}(..., type = \text{“removing”})$
    ** Save test and local training set **
    $D_{test} \leftarrow \textsc{GetData}(D, V_{test})$
    $D_{training} \leftarrow \textsc{GetData}(D, B_S)$
    $D_{training} \leftarrow \textsc{Drop}(D_{training}, B_R)$
    $\textsc{Save}(D_{test}, D_{training})$
end for
```

γ and uses the folds spatial boundaries to introduce direction. The procedure to calculate the buffers is shown in Algorithm 2.

Algorithm 2 starts by initializing the variable *growing*, which indicates whether the buffer is still growing, the variable h, which corresponds to the lag distance for the semivariogram calculation, and the buffer V_{buffer} (lines 9–12). Afterward, we compute the sill per fold in the training set (lines 13–16). We consider the mean between the variance of values in the training fold and test set and the values from the entire dataset. Then, if we calculate the selection buffer, the sill for all folds becomes the highest sill computed (lines 17–21). Next, we calculate the depth of the Breadth-First Search (BFS) tree considering the vertices in V_{test} as the root level (line 22). Finally, we initiate the buffer generation (lines 23–38).

We define a logarithmic decreasing sill that decreases as the lag distance increases, intending not to consider the training data distant from the test set (line 25). The traversal of G occurs in a BFS manner by getting $N_G(S)$, the vertices' open neighborhood in the buffer region and the teste set (S). The direction of the semivariogram depends on the folds presented in the neighborhood $N_G(S)$. For each fold, we estimate the semivariogram and build the buffer (lines

27-36).

To calculate the semivariogram, we count the number of folds present in $N_G(S)$ (line 28). Then, we generate $C(h)$, the set of all pairs distant by lag distance h. It corresponds to the cartesian product between V_{test} and V_{fold} (line 29). Next, we get z, values from the set of vertices in S, to measure the semivariogram (line 30). Finally, we estimate the semivariogram at a distance lag h according to Equation 3.1 (line 31). The neighboring vertices from the considered fold (V_{fold}) are appended to the buffer if the semivariogram value obtained is less than the fold sill, and the number of folds in the buffer times a relaxing factor κ is less than the number of the test set neighboring folds (lines 32–35). The distance lag, in turn, is added by one (line 30). The procedure stops if no vertex is added to the buffer. The output of this procedure is the buffer, which corresponds to the set of vertices in V_{buffer} (line 39).

3.4 Case of Study

To assess our approach, we consider the 2018 Brazil presidential election, focusing on the second round. The dataset describes 5570 Brazil's municipalities. The explanatory features correspond to the most recent census of 2010, data related to the analyzed election—*i.e.*, percentage of null and blank votes—, and the 2010 human development index. Thus, the dataset comprises 4023 explanatory features and one target feature, the latter being the winning party's vote-shares.

The main goal was to build a model learned from selected features to predict the winning party's vote-shares. We considered Gradient Boosting (FRIEDMAN, 2001) as the regression model and Correlation-based Feature Selection (CFS) (HALL, 2000) as the filter method for simplification. These methods generally achieve good results even using default parameter values. Besides, CFS finds a minimal optimal subset of features, not requiring to define the number of features to be selected.

Each spatial fold follows the geographical boundaries of the 27 Brazilian states. To enable a fair comparison, we defined a *baseline* and a *topline* for our approach. The *baseline*, called Ultra-Conservative, calculates the size of the buffer region by monitoring the target variable spatial dependence using the semivariogram over the entire dataset. The size is chosen as the lag distance where the semivariogram first reaches the sill. For comparison purposes, we specified the lag distance as the h-degree neighborhood. In other words, the buffer size was defined as the 27$^{\text{th}}$-degree neighborhood from the test fold subgraph. This approach prioritizes the total independence between test and training sets, disregarding the number of instances in the first one. On the other side, the *topline* method, named Optimistic, does not adopt a removing buffer region and prioritizes the model's performance over the independence between test and training sets. Furthermore, to evaluate the impact of the selection buffer, we consider a variation of our approach that only applies the removing buffer procedure, namely R Buffer. The proposal that

Algorithm 2 – Calculate Buffer

```
Require:
$G$: the dataset spatial graph;
$V$: the set of vertices;
$Z$: the dataset values to calculate the semivariogram;
$V_{test}$: the test set vertices;
$n_{test}$: the number of test set neighboring folds;
$\kappa$: the relaxing factor for the number of folds in the buffer;
$type$: buffer type (“selection” or “removing”).

$h \leftarrow 0$
$V_{buffer} \leftarrow \emptyset$
$growing \leftarrow True$
$V_{training} \leftarrow V - V_{test}$
For Each $fold \in V_{training}$ do
    $S \leftarrow V_{test} \cup V fold$
    $Sill[fold] \leftarrow \textsc{Mean}(VAR(S(Z)), VAR(V(Z)))$
end for
if $type =$ “selection” then
    For Each $fold \in Sill$ do
        $Sill[fold] \leftarrow \textsc{Max}(Sill)$
    end for
end if
$depth \leftarrow \textsc{Depth}(BFS(G, root = V_{test}))$
while $growing$ do
    $growing \leftarrow False$
    $dec \leftarrow log_{depth}(depth - h)$
    $S \leftarrow V_{buffer} \cup V_{test}$
    For Each $V_{fold} \in N_G(S)$ do
        $n_{folds} \leftarrow \textsc{CountFolds}(N_G(S))$
        $C(h) \leftarrow V_{test} \times V_{group}$
        $z \leftarrow S(Z)$
        $\gamma(h) \leftarrow \textsc{Semivariogram}(C(h), z)$
        if $\gamma(h) \leq Sill[fold] * dec$ & $n_{test} \leq n_{folds} * \kappa$ then
            $V_{buffer}.\textsc{Append}(V_{group})$
            $growing \leftarrow True$
        end if
    end for
    $h++$
end while
return $V_{buffer}$
```

considers both the selection and removing buffer procedures is referenced as SR Buffer. The values for the parameter κ in both approaches were set empirically as half of the number of fold.

3.4.1 The Impact of Buffer Regions

This subsection analyzes the impact of buffer regions generated by our approach in evaluating models based on election data. We sought to validate our approach as a response to avoid biased conclusions caused by ignoring the spatial dependence. Table 8 shows the mean of three statistics—Mean Squared Error (MSE), number of features selected (#FS), and number of training instances (#TI)—generated by the four approaches considered; standard deviations are in parentheses. On average, the two variants of our method stand between the *baseline* and the *topline*. These results are more realistic as they are not too close to the high biased scenario (Optimistic) and the Ultra-Conservative approach, the latter contributing to overfitting due to the low number of training instances. Besides, SR Buffer—the method that uses the selection buffer—produced a slightly lower MSE when compared to the one that does not use it. We can explain this result by the higher number of features selected in the SR Buffer approach and the filtering of relevant and spatially close training instances.

Table 8 – Mean statistics results.

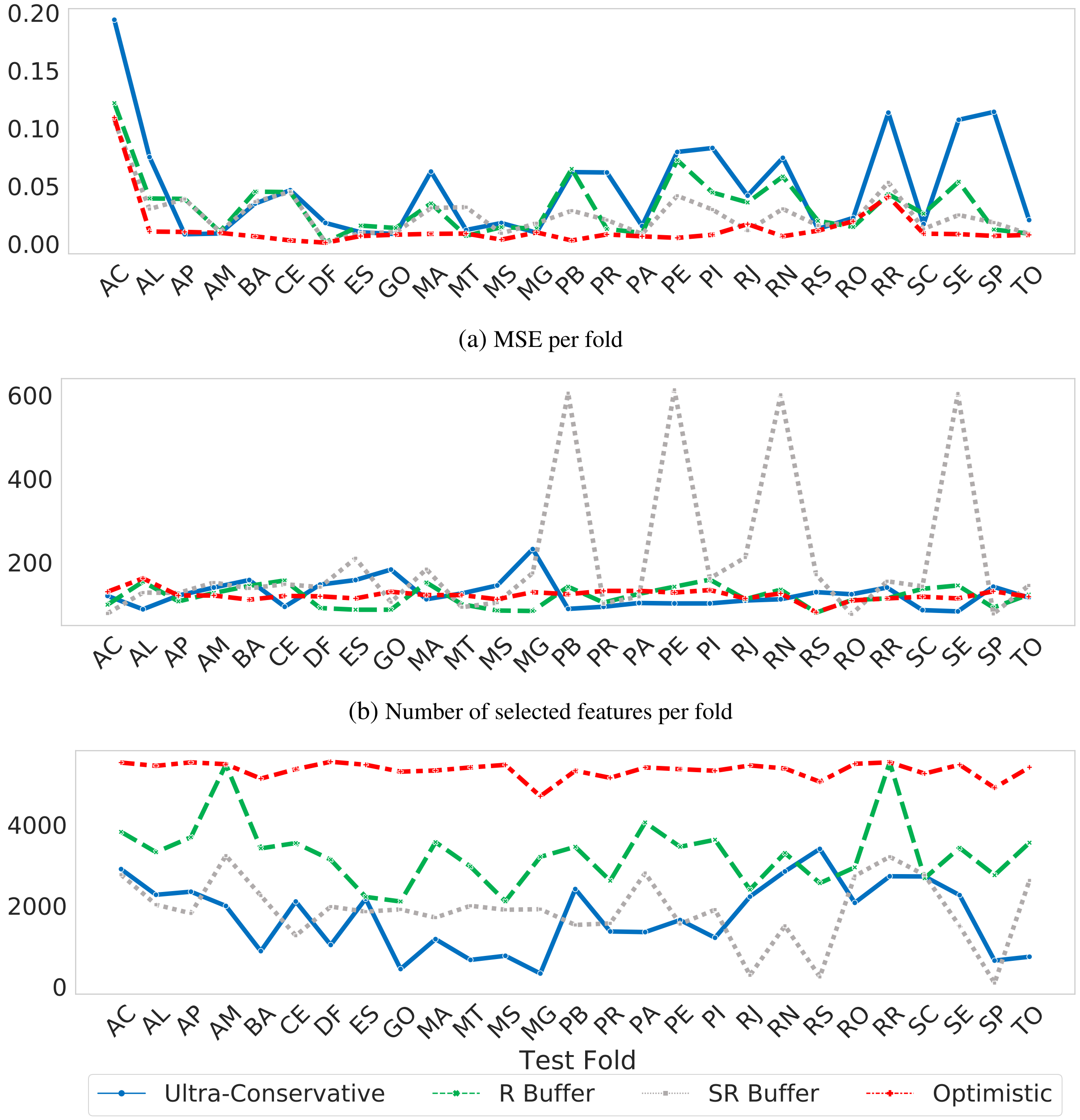
	Ultra-Conservative	R Buffer	SR Buffer	Optimistic
MSE	0.049 (0.045)	0.033 (0.026)	0.026 (0.021)	0.013 (0.021)
#FS	124.25 (33.25)	118.29 (25.68)	206.18 (173.59)	121.77 (13.34)
#TI	1732.92 (866.84)	3299.29 (831.86)	1888.92 (807.76)	5354.07 (202.74)

Source: Research data.

From a more detailed perspective, since the spatial folds present different sizes and distributions, we conducted comparisons considering the results of each test fold (Figure 13). According to Figure 13a, the two variants of our approach maintained the MSE values between the *baseline* and *topline* in most of the folds. Ultra-Conservative showed a high variance of results indicating the presence of overfitting, produced by the low number of training instances in some test folds (Figure 13c). On the other hand, Optimistic presents a low variance of results. However, since the spatial autocorrelation causes a high bias, these results may lead to biased conclusions. Our approaches, in turn, presents low bias by diminishing the presence of spatial autocorrelation on the training set and variance lower than Ultra-Conservative. Thus, generating more realistic and robust models.

Concerning the number of features selected per fold (Figure 13b), we highlight the four peaks in the test folds PE, PB, RN and SE, regarding the SR Buffer approach. Although this amount is considered a much higher number when compared to the other three methods, it shows the complexity of the local training sets generated, which can be seen as a more realistic scenario. To provide a more detailed discussion, Figure 15 presents the vote-shares distribution and the buffers obtained from the Ultra-Conservative, R Buffer, SR Buffer, and Optimistic approaches regarding the test fold PE. According to Figure 15, the removing buffer applied by Ultra-Conservative provided a small number of training instances (Figure 13c) with a dominance

Figure 13 – Statistics generated from the folds regarding each validation approach.



(a) MSE per fold

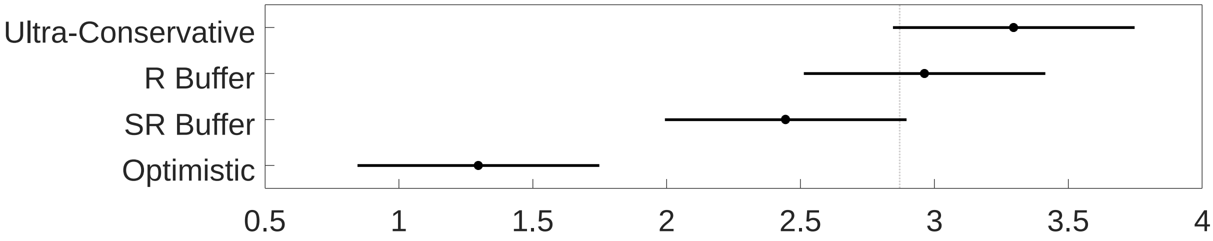
(b) Number of selected features per fold

(c) Number of training instances per fold

Source: Elaborated by the author.

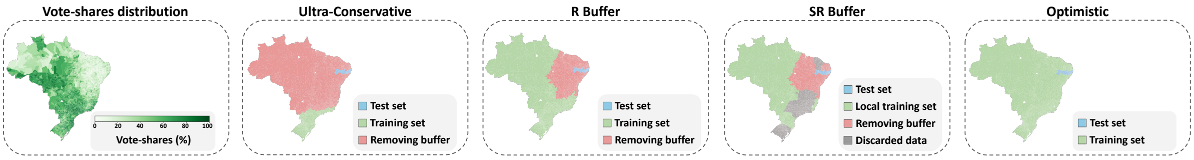
of vote-shares higher than 50%. These conditions provided the lowest number of features selected but the highest value of MSE (Figure 13a). On the other hand, R Buffer produced a training set with more data (Figure 13c) and heterogeneity concerning the distribution of vote-shares. Although the number of features selected was higher than Ultra-Conservative, the MSE value was closer (Figure 13a), indicating that the possible interference of distant and irrelevant data. Finally, SR Buffer discarded a neighbor fold and part of the southern folds in the selection buffer step, generating a local training set with higher complexity. This led the feature selection algorithm to select more features. However, the MSE evidences that the model learned from the

Figure 14 – Friedman test with p-value < 0.05 followed by Tukey-Kramer posthoc test concerning the MSE statistic.



Source: Elaborated by the author.

Figure 15 – Vote-shares distribution and determined training sets for each validation approach considering the state PE as test fold.



Source: Elaborated by the author.

selected features could predict the test fold better.

To support our discussion, we conducted significance tests on the MSE values using the following statistical methods available at MATLAB[2]: Shapiro-Wilk normality test, and Friedman test with p-value < 0.05 followed by Tukey-Kramer posthoc test. According to Figure 14, Optimistic was statistically different from the other three approaches, indicating that R Buffer and SR Buffer provided less biased results. However, both variants of our approach did not differ from Ultra-Conservative. However, since our approach also presents a conservative characteristic, this is an expected result. Nonetheless, the average rank position of SR Buffer and R Buffer was higher than Ultra-Conservative, with SR Buffer ranked between the *baseline* and the *topline*.

3.4.2 Assessing Dependence

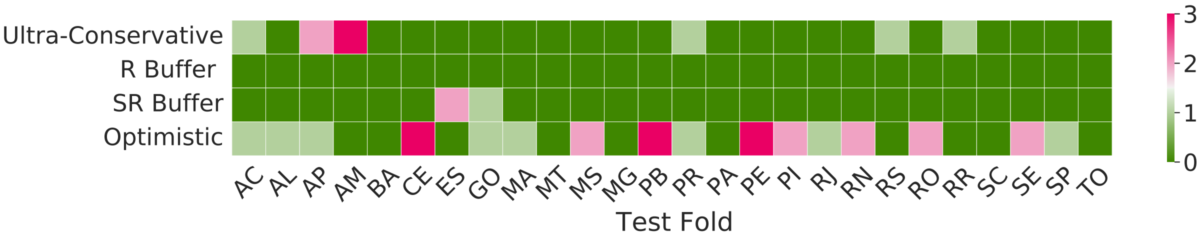
The effectiveness of our approach can be observed in Figure 15. The two variants of our validation method only removed the training data that presented a high spatial autocorrelation concerning the vote-shares and grew non-uniformly in different directions. To support this argument, we conducted a χ^2 homogeneity test between the test set and the three training folds on the boundary of the removing buffer most similar regarding the explanatory feature space to the test set. As this test uses categorical variables, we discretized the target feature so that values greater than 50% were defined as wins and the rest as losses.

Figure 16 displays a heatmap concerning the number of training folds where the null hypothesis was not rejected in the χ^2 independence test, indicating dependence with the test set.

[2] <https://www.mathworks.com/products/matlab.html>.

In this figure, green spots symbolize that one or none training folds present spatial dependence, while red spots denote that two or three training folds presented spatial dependence regarding the test set. We performed the statistical test considering only the test folds with the number of instances greater than one. Accordingly, the Optimistic approach presented 17 test folds with at least one neighbor training fold showing dependence. These results corroborate with the low value of MSE obtained. The other methods, in contrast, exhibited fewer test folds with dependence on the training set. We need to highlight here that R Buffer presented no test fold with dependence on the training set. Considering SR Buffer, there were only two test folds with dependence in the training set. As the training set is distant from the test fold, the distribution on the explanatory feature space can be different. Nevertheless, this does not pose problems for the data modeling.

Figure 16 – Number of training nearest folds dependent to the test set according to χ^2 independence test.



Source: Elaborated by the author.

3.4.3 Evaluating Processing Time

Another critical aspect of the validation approaches is the time spent to create folds. For simplicity but without any substantial loss, we can analyze the time complexity of the methods based on the semivariogram complexity time, which is $O(V^2)$, the same as the BFS used to traverse the graph to find the nodes at the distance lag h. Thus, the Ultra-Conservative approach has a time complexity of $O(V^3)$ because it runs the semivariogram for all the vertices in the graph. Differently, SR Buffer executes the semivariogram per fold two times to calculate the selection and removing buffer. Thus, it has a time complexity of $O(2F.V^2)$, where F is the total number of spatial folds. However, the term $2F$ presents little impact and can be removed, leading to a time complexity equals $O(V^2)$. Following this analysis, the time complexity of Optimistic is equal to $O(1)$ since no semivariogram is estimated. To strengthen our arguments, Table 9 shows the time spent by each approach as well as the number of times the semivariogram function was called. We ran the experiments on a computer with a Core i7 9th generation and 16GB RAM. In this scenario, our method (SR Buffer) and its variant (R Buffer) surpass the Ultra-Conservative approach providing a fast buffer calculation.

Table 9 – Method's processing time and number of times the semivariogram function was called.

	Ultra-Conservative	R Buffer	SR Buffer	Optimistic
Time (hours)	26	0.13	0.26	0.06
Semivariogram Calls	5570	27	54	0

Source: Research data.

3.5 Conclusion

The spatial nature of election datasets raises the problem of spatial dependence, which, if ignored by analysts, leads to biased results. In this direction, this chapter covers the initial efforts made in developing a graph-based spatial cross-validation approach originally proposed to assess models learned with selected features from spatially contextualized electoral datasets. By considering the second round of the 2018 Brazilian presidential election, we experimentally showed evidence about the impact of defining proper buffer regions to diminish spatial dependence between test and training sets. In addition, our experiments demonstrated that filtering the training set based on the spatial closeness to the test fold can provide more local and realistic data modeling.

Although we have tested the novel approach with a single dataset, the results show that it can guarantee independence between test and training sets and afford more realistic modeling. Our immediate plans include analyzing spatial datasets from other domains to provide more information on how our method can be generalized to other applications.

CHAPTER

4

GEOGRAPHIC CONTEXT-BASED STACKING LEARNING FOR ELECTION PREDICTION FROM SOCIO-ECONOMIC DATA

Voting behavior analysis involves understanding factors influencing an election to identify possible trends, new features, and extrapolations. A growing body of research has joined efforts to automate this process from high-dimensional spatial data. Although some studies have investigated machine learning methods, the capability of this artificial intelligence subarea has not been fully explored due to the challenges posed by the spatial autocorrelation structure prevalent in the data. This chapter advances the current literature by proposing a geographic context-based stacking learning approach for predicting election outcomes from census data. Our proposal models data in spatial contexts of different dimensions and operates on them at two levels. First, it captures local patterns extracted from spatial contexts. Then, at the meta-level, it globally captures information from the K contexts nearest to a region we want to predict. We introduce a spatial cross-validation-driven experimental setup to assess and compare the stacking approach with state-of-the-art methods fairly. This validation mechanism aims to diminish spatial dependence's influence and avoid overoptimistic results. We estimated a considerable multi-criteria performance of our proposal concerning baseline and reference models taking data from the second round of the 2018 Brazilian presidential elections into account. The stacking approach presented the best overall performance, being able to generalize better than the compared ones. It also provided intelligible and coherent predictions in challenging regions, emphasizing its interpretability. These results evidence the potential use of our proposal to support social research.

4.1 Introduction

Elections are non-trivial processes essential to any representative democracy, which can provide the best expression of public opinion and party involvement. A post-election data analysis allows us to describe voting behavior and the aspects that guide it (JACINTHO *et al.*, 2020). Understanding voting behavior is vital to identifying trends and factors influencing election results (LAYTON *et al.*, 2021; PINHEIRO-MACHADO; SCALCO, 2020).

Researchers consider that the electoral processes are associated with the population characteristics regarding the locations where they occur. Thus, an electoral process comprises aspects that indicate local patterns related to spatial autocorrelation and local relationships across space (FOTHERINGHAM; LI; WOLF, 2021). In this perspective, people from the same region tend to present similar voting behavior, while those from distinct areas may have different vote distributions.

Considering how people are geographically contextualized and the data's spatial characteristics can enrich our understanding of electoral processes. We have witnessed an increasing number of interdisciplinary studies aimed at predictive modeling election features from thousands of explanatory spatial features (SILVA; PARMEZAN; BATISTA, 2021; LI; PERRIER; XU, 2019). However, the high dimensionality and spatial autocorrelation structure inherent in such data limit the ability of conventional learning models to capture the relationships between spatial features completely.

Many econometric and machine learning methods, which can deal with the curse of dimensionality, totally ignore the geography present in electoral data, such as spatial boundaries, clustering effects, and distance measures (GRAEFE; GREEN; ARMSTRONG, 2019; CHAUHAN; SHARMA; SIKKA, 2021). Consequently, they treat data separated into regions as independent and identically distributed. In the opposite direction, recent studies have suggested using spectral and spatial filtering Graph Convolutional Neural Network (GCNN) methodologies to enrich election data modeling (LI; PERRIER; XU, 2019). Such methods seem to adequately fit the problem at hand, given the intrinsic graph structure of electoral data.

This work advances the literature on voting behavior analysis by proposing a geographic context-based stacking learning approach to describe election outcomes from thousands of census features. Our proposal models data in spatial contexts of different dimensions and operates on them at two levels: (i) at the base level, it captures local patterns extracted from spatial contexts; (ii) at the meta-level, it globally captures information from the K contexts nearest to a region we want to predict. Furthermore, we introduce a spatial cross-validation-driven experimental setup to assess and compare the stacking approach with state-of-the-art methods fairly. This validation mechanism can generate robust assessments by diminishing the spatial dependence's influence and consequently avoiding overoptimistic results (SILVA; PARMEZAN; BATISTA, 2021; PLOTON; MORTIER *et al.*, 2020).

We estimated a considerable multi-criteria performance of our proposal concerning two baselines and the state-of-the-art Hierarchical GCNN method taking data from the second round of the 2018 Brazilian presidential elections into account. The stacking approach exhibited the best overall performance, being able to generalize better than the compared ones. It also led to intelligible and coherent predictions in challenging regions, highlighting its interpretability. These results demonstrate the potential use of our proposal to support social research.

The rest of this chapter is organized as follows: section 4.2 introduces the background and related work. section 4.3 describes our geographic context-based stacking learning approach. Figure 4.3 reports the case study involving data from the second round of the 2018 Brazilian presidential election. Finally, section 4.4 concludes the study and highlights future work.

4.2 Background and Current Trends

This section defines the mathematical notation that models election voting behavior considering the spatial characteristics of the data. It also discusses related work covering the most recent advances in the literature.

4.2.1 Problem Definition and Research Challenges

We can formulate the problem in question as follows. First, let us specify a set of lattice-type spatial objects O, where each object o_i is a polygon that delimits a region in the spatial domain (*e.g.*, neighborhoods, districts and cities). Note that the spatial intersection between any distinct objects o_i and $o_j \in O$ is the empty set ($\emptyset$). Now, let us assume a spatial dataset $D = \{X, Y\}$ that characterizes each of the n objects in O. The target feature, $Y \in \mathbb{R}$, reflects the vote shares (vote percentage) for each spatial object in O from a given candidate or party. The explanatory features, $X \in \mathbb{R}^m$, where $m > 0$ is the number of characteristics, describes the spatial objects from O in another election-related domain. Let us also consider a set of spatial contexts C with boundaries that segment D in the geographic space, where C can be defined based on preexisting boundaries (*e.g.*, states and macro-regions). The objective is to generate a model $F(D, C)$ that learns local relationship patterns from the spatial contexts present in D.

Modeling local relationships between explanatory features and the target feature (vote shares) is not a trivial task. These relationships may vary across spatial contexts, meaning that a relevant characteristic that can describe the vote shares from one context may not be useful to another (FOTHERINGHAM; LI; WOLF, 2021). Furthermore, in a conventional machine learning approach, local relationships are disregarded in favor of those that describe the vote shares globally (SILVA; PARMEZAN; BATISTA, 2021; FOTHERINGHAM; LI; WOLF, 2021).

Another challenge in modeling voting behavior relates to using Spatial Cross-Validation (SCV) as a sampling technique. While it is the most suitable procedure for assessing machine learning models built from spatial data, it generates unseen correlated distributions in the test set. This scenario happens because spatial boundaries determine the folds, and a removing buffer region is defined as a strategy to diminish the spatial dependence between the test and training sets (SILVA; PARMEZAN; BATISTA, 2021; PLOTON; MORTIER *et al.*, 2020). Consequently, the test set distribution is not observed in the training set, and there are only correlated distributions.

Studies on applying machine learning methods for analyzing voting behavior are maturing through scientific debate. subsection 4.2.2 briefly summarizes some related work in this field, emphasizing the challenges they brought.

4.2.2 Related Work

The vast literature on voting behavior varies from standard econometric techniques (GRAEFE; GREEN; ARMSTRONG, 2019) to regression analysis (FOTHERINGHAM; LI; WOLF, 2021) and machine learning models (CHAUHAN; SHARMA; SIKKA, 2021; LI; PERRIER; XU, 2019). Econometrics and regression analysis studies usually focus on national-level estimators using surveys and economic features to understand election results. Although these methods are well established (GRAEFE; GREEN; ARMSTRONG, 2019), applying them to thousands of features in several locations is challenging. Conversely, machine learning models can deal with the curse of dimensionality more naturally. However, most works employs social media data and sentiment analysis to understand voting behavior. Their results are commonly explored on a national scale, and spatial aspects are not considered (CHAUHAN; SHARMA; SIKKA, 2021).

Recently, researchers recommended using a hierarchical GCNN-based approach that can be considered state-of-the-art in voting behavior analysis via machine learning (LI; PERRIER; XU, 2019). The authors combined the inherited hierarchical characteristic of the census and election data with the GCNN capability to learn local patterns and generate a model capable of predicting the vote shares from the 2016 Australia congress election with low error rates.

In contrast to existing analytical models, here we design a descriptive approach that considers thousands of socio-economic explanatory features and the involved spatial characteristic to analyze locally and comprehensively election outcomes across multiple locations.

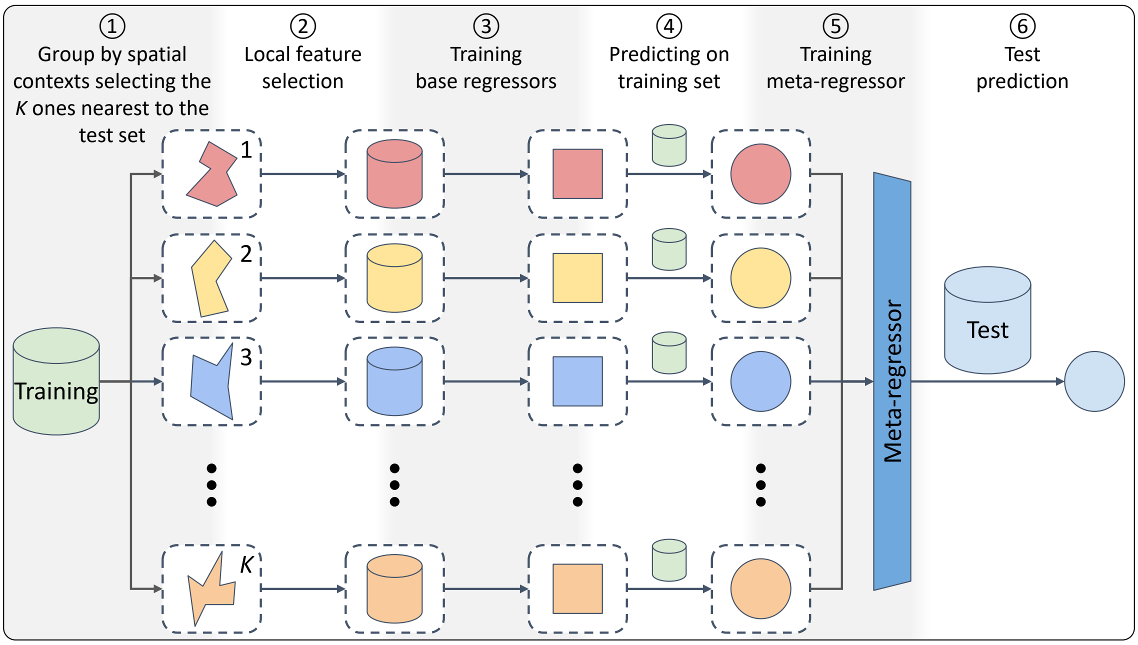
4.3 Proposed Approach

We have identified two main challenges linked to the problem formalized in Section 4.2.1: (i) capturing local patterns that are occluded when globally modeling the data; and (ii) building a model that can generalize over different spatial contexts. This chapter addresses these challenges by proposing a geographic context-based stacking approach to model local relationships at the ensemble level and globally capture information from contexts employing a meta-regressor.

When applied to regression tasks, the conventional stacking approach builds an ensemble using the entire training set to fit each base regressor. We typically choose regression algorithms from different paradigms to introduce diversity, generating a heterogeneous ensemble (DIETTERICH, 2000). The predictions of each base regressor on a validation set give rise to an attribute-value table, which is employed to train a meta-regressor. The meta-regressor, in turn, learns how to ponder the base regressors' predictions to issue final predictions.

Our approach differs from traditional ones in the following aspects. First, we define the ensemble by the K nearest spatial contexts to the test set region. Such a strategy is based on the first law of geography, which states that "everything is related to everything else, but near things are more related than distant things" (TOBLER, 1970). Second, we use spatial context to build the base regression models so that each model can capture local patterns related to the contexts. Lastly, the ensemble is homogeneous, meaning we adopt the same base regressor method. However, the diversity comes from the spatial contexts that present different dimensions (#instances × #features), following the idea that a different set of features may describe each context.

Figure 17 – Proposed approach.


① Group by spatial contexts selecting the K ones nearest to the test set
② Local feature selection
③ Training base regressors
④ Predicting on training set
⑤ Training meta-regressor
⑥ Test prediction
Training
1
2
3
K
Meta-regressor
Test

Source: Elaborated by the author.

Figure 17 outlines the steps of our approach. In Step 1, we group the training data in agreement with a pre-defined set of spatial contexts and select the K ones nearest to the test set, considering geographic proximity; the number of instances in each spatial context can vary. In Step 2, we run a feature selection method for each spatial context data, generating K spatial context with different dimensions. In Step 3, we build an ensemble where each base regressor is fitted to a spatial context. In Step 4, we use the base regressors that make up the ensemble to predict the entire training set, creating an attribute-value table. In Step 5, we employ the data table to train a meta-regressor that will seek to ponder the local knowledge learned by the base regressors to maximize the generalization potential. Finally, in Step 6, the meta-regressor uses the predictions from the base regressors on the test set to provide final predictions.

As can be seen, our approach operates on two levels. The first level learns local patterns from geographically contextualized data samples, *i.e.*, regions containing mutually exclusive instances described by an optimal feature subset. In a complementary way, the second level extracts global knowledge of the local patterns to predict a region of interest.

In 2018, Brazilians went to the polls to vote for their president. The final result was 55.13% for the Social Liberal Party (Jair Bolsonaro) and 44.87% for the Worker's Party (Fernando Haddad). This election, however, was marked by a highly polarized environment and flooded by distrust in the voting system (JACINTHO *et al.*, 2020). Understanding outcomes in this context is essential to discuss the external factors influencing voting decisions and identify the geographic and socio-economic extensions of the processes that undermine democratic foundations.

4.3.1 Brazilian Election Data

The dataset analyzed here expresses the second round of the 2018 Brazilian presidential election and portrays 5565 Brazilian municipalities. It has 3999 explanatory features that represents the 2010 census and one target feature, which is the vote share received by the winning party.

The census data was sourced from the Brazilian Institute of Geography and Statistics (*IBGE*). The data is available to the public via anonymized aggregated features that describe population groups delimited by geospatial areas, *e.g.* municipalities, which correspond to the aggregation level used in this study. To avoid erroneous results, we standardized all the features according to the city's population size or the number of domiciles.

We sourced the election data from the Superior Electoral Court (*TSE*), which provides vote count results regarding each voting machine called "*boletim da urna*". We aggregated the vote counts at a city level and calculated the vote shares as the percentual of valid votes for the winning party.

4.3.2 Machine Learning Approaches and Algorithms

Standard econometric methods are not comparable with our approach. They often focus on regression analysis and employ data at higher aggregation to provide national-level predictions (GRAEFE; GREEN; ARMSTRONG, 2019). The HIERARCHICAL GCNN model (LI; PERRIER; XU, 2019), in turn, can be used as a representative of state-of-the-art applied machine learning research. We adopted this method parameterized according to the best results in (LI; PERRIER; XU, 2019), named variation 2. Furthermore, we considered city-level data as the prediction layer and state-level data to create the second aggregation layer.

We also defined two baselines, GLOBAL and LOCAL MEAN, to be compared with our proposal, addressed from now on as LOCAL META. GLOBAL is the conventional approach that selects features and fits models favoring global relationships. LOCAL MEAN is an ensemble of contextual models that employs the average as a fusion function to compose the final predictions. These baselines can help us understand in which situations LOCAL META performs best and explain how the stacking strategy increases performance and improves generalization.

We investigated two configurations involving the number of spatial contexts (K) for LOCAL MEAN and LOCAL META. The first uses all the contexts in the training set ($K = C$), while the second employs the seven closest contexts ($K = 7$) to the prediction area. Note that 7 is the mean of each context's neighbors. This configuration, in particular, aims to answer whether filtering contexts based on the prediction area proximity can enhance results.

Concerning the base regressors, we considered nine belonging to different machine learning paradigms. Table 10 lists these algorithms and their parameters. As the meta-regressor for LOCAL META, we chose Ordinary Least Squares (OLS) since it is a simple and parameterless model. We adopted the Correlation-based Feature Selection (CFS) method to reduce the attribute space. CFS aims to find a minimal optimal subset of features that are highly correlated with the target and not very redundant with each other.

Table 10 – Base regressors and their parameters. The acronyms not yet defined are: k-Nearest Neighbors (kNN), Least Absolute Shrinkage and Selection Operator (LASSO), Decision Tree (DT), Gradient Boosting DT (GBDT), Multiyear Perceptron (MLP), and Support Vector Regression (SVR).

Base regressor	Parameter	Value
kNN	Number of nearest neighbors (k)	3
OLS	—	—
LASSO	Regularization strength (α)	1
Ridge	Regularization strength (α)	1
ElasticNet	Constant that multiplies the penalty terms (α)	1
	Mixing parameter ($l1\_ratio$)	0.5
DT	Split criterion	GINI
GBDT	Number of boosted trees to fit ($n\_estimators$)	100
	Learning rate (ε)	0.1
MLP	Hidden layers (h)	1
	Hidden layer size (n)	$M/2$
	Learning rate (ε)	0.001
SVR	Kernel	RBF
	Gaussian's width of the radial basis kernel function (σ)	$1/(M * X.var())$
	Regularization parameter ($\mathbb{C}$)	1

Source: Research data.

Finally, we defined the spatial contexts for the ensemble approaches as the 26 Brazilian states save the Federal District, which has only one city. This decision comprises the understanding that, at a higher level, stakeholders such as political scientists and journalists are more interested in analyzing the election results considering known spatial boundaries like states.

4.3.3 Evaluation Measures

We assessed the approaches described in subsection 4.3.2 using four individual performance measures: Mean Squared Error (MSE), Mean Error Standard Deviation (MESD), SPearman correlation (SP), and SPearman correlation Standard Deviation (SPSD). MSE expresses the approaches' performance in predicting the correct vote-share scale, while MESD reflects their stability regarding MSE. MSE does not indicate whether the order of the achieved predictions matches those in the ground truth. That is, if city *A* received more votes than city *B*, MSE does not tell us whether the approaches were able to capture this order. Thus, we employed SP to assess the order of the predictions yielded by the approaches. Finally, SPSD provides information on how SP is distributed over the folds.

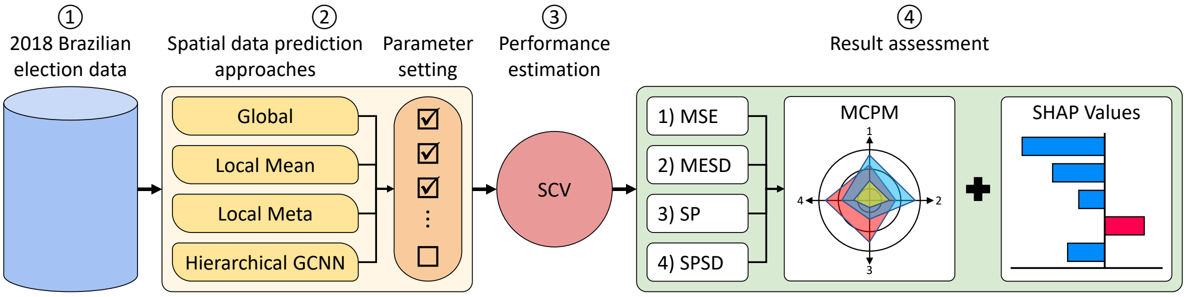
We also applied a Multi-Criteria Performance Measure (MCPM) (PARMEZAN; LEE; WU, 2017) to combine the four metrics mentioned above and thus guide the choice of adequate approaches. MCPM reflects the sum of the total area of an irregular polygon whose vertices comprise individual performance indexes. In this work, lower total area values indicate better predictive performances. Unlike the MSE and Standard Deviation measures, in which resulting values must be minimized, SP (ρ) must be maximized. Hence, we applied the SP complement: $1-\rho$.

To understand the predictive power of our proposal, we employed the SHapley Additive exPlanation (SHAP) Values technique to analyze the results in the best and worst fold scenarios (LUNDBERG; LEE, 2017). SHAP Values is a unified measure of feature importance widely used to comprehend predictions made by models. We believe that examining it within the scope of our application is indispensable, as it can reveal biased models and avoid misinterpretations. Especially for our approach, SHAP Values can be employed in the meta and base regressors to explain the most important spatial contexts to predict a given fold and the most relevant features of that context.

4.3.4 Experimental Setup

Figure 18 illustrates our experimental setup, which considers the space's role in evaluating models designed to predict election outcomes. As depicted in this figure, we used the dataset prepared in agreement with subsection 4.3.1 (Step 1) to assess the approaches parameterized according to subsection 4.3.2 (Step 2). We applied an SCV technique, which da Silva *et al.* (SILVA; PARMEZAN; BATISTA, 2021) explicitly proposed for the application in question, to estimate the performance of the models (Step 3). We reported these results via the evaluation metrics described in subsection 4.3.3 (Step 4). This experimental protocol assesses the investigated approaches more rigorously, as it avoids overoptimistic results by reducing the spatial dependence between test and training sets. While our multi-criteria analysis compares these models taking into account two important characteristics – the scale and the order of the vote shares –, our interpretability analysis is necessary to understand the patterns found and uncover biased models.

Figure 18 – Experimental setup.



Source: Elaborated by the author.

The main difference between the SCV adopted in this work and the standard cross-validation lies in the fold definition so that in the former, the folds are determined based on preexisting geographic boundaries. Here, each spatial fold follows the geographic boundaries of the 26 Brazilian states, also employed as spatial contexts to build contextual base models in the ensemble approaches. Unlike the traditional cross-validation, in the SCV, the folds may have different distributions and sizes, creating a more challenging scenario for the approaches.

This study did not use folds 12 (Acre) and 13 (Rondonia) to calculate MSE and MSDE values because they present ambiguous distributions (JIANG *et al.*, 2019); *i.e.*, they describe a population with similar socio-economic characteristics to the northeastern but with vote shares similar to the southern states. This fact requires the approaches to learn the opposite of what they observed in the training set. Scenarios like these are challenging and exhibit incredibly high error rates, impacting empirical assessments, specifically the choice of the best regressor to compose GLOBAL. We decided to keep the analysis of SP and SPSD on such folds considering that the raised issue is linked to scale and not to order.

Finally, we implemented the experimental setup of Figure 18 employing the Python programming language combined with the following libraries: Pandas, GeoPandas, SciPy, PySAL, and Scikit-learn. Our code and supplementary material are available on the GitHub platform[1].

4.3.5 Results and Discussion

As we can see from the averaged values of the individual performance metrics (Table 11), LOCAL META $K = 7$ achieved the best MSE and MESD results, LOCAL MEAN $K = C$ presented the highest SP values, and LOCAL MEAN $K = 7$ stood out in terms of SPSD. We obtained all these results using MLP as a base regressor. However, there was no consensus regarding the best model configuration – approach and base regressor combination – concerning all the metrics. To identify the most promising model, we evaluated the configurations under three

[1] <https://github.com/tpinhoda/Spatial_Context_Stacking_Approach>.

perspectives: (i) the MCPM to determine the best overall model; (ii) the performance per fold to identify the best context-level configuration; (iii) the model interpretability to understand the best configuration results.

Table 11 – Overall results of the approaches considering each base regressor and the following metrics: MSE, MESD, SP, and SPSD. Green cells symbolize the best results.

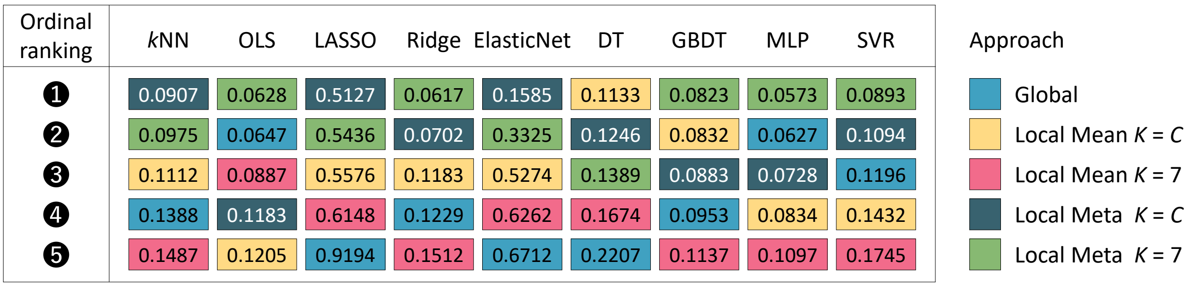
	GLOBAL				LOCAL MEAN K = ALL				LOCAL MEAN K = 7				LOCAL META K = ALL				LOCAL MEAN K = 7			
Base regressors	MSE	MESD	SP	SPSD	MSE	MESD	SP	SPSD	MSE	MESD	SP	SPSD	MSE	MESD	SP	SPSD	MSE	MESD	SP	SPSD
*k*NN	239.68	267.93	0.46	0.16	272.04	187.92	0.58	0.15	303.78	216.51	0.50	0.17	179.50	157.93	0.56	0.17	173.60	162.65	0.51	0.17
OLS	121.81	134.17	0.59	0.15	271.08	197.79	0.60	0.19	219.96	170.12	0.59	0.13	224.67	195.01	0.55	0.20	125.81	141.26	0.57	0.13
LASSO	965.26	465.45	0.02	0.50	949.42	457.43	0.38	0.24	949.24	456.68	0.26	0.24	876.07	405.93	0.35	0.25	837.60	427.85	0.27	0.25
Ridge	304.48	194.94	0.58	0.16	355.77	214.79	🟩 0.64	0.12	428.35	247.88	0.62	0.13	124.69	129.78	0.57	0.16	132.77	128.39	0.61	0.14
ElasticNet	964.61	465.43	0.23	0.29	961.42	463.94	0.48	0.27	961.99	464.01	0.36	0.35	299.24	216.62	0.50	0.21	528.13	317.23	0.43	0.34
DT	269.36	355.57	0.32	0.22	223.96	178.53	0.52	0.17	287.03	232.21	0.41	0.19	213.99	189.83	0.48	0.19	230.87	201.68	0.43	0.19
GBDT	171.56	159.91	0.51	0.17	196.32	153.97	0.61	0.16	249.46	186.00	0.56	0.17	159.09	149.73	0.54	0.18	142.39	140.13	0.56	0.18
MLP	133.87	138.18	0.61	0.14	240.22	170.30	🟩 0.64	0.13	311.43	207.35	0.62	🟩 0.12	127.01	132.60	0.57	0.17	🟩 111.22	🟩 127.19	0.59	0.14
SVR	243.97	206.58	0.55	0.18	398.58	238.68	0.62	0.14	447.63	261.98	0.59	0.15	168.97	149.09	0.52	0.23	164.13	146.32	0.55	0.18

Source: Research data.

4.3.5.1 Multi-Criteria Performance

Figure 19 shows, for each approach configuration, the MCPM values ranked in descending order of importance. LOCAL META $K = 7$ presented a more consistent behavior occupying the first and second positions for most configurations, with its lowest position being the third employing DT. On the other hand, the GLOBAL and LOCAL META $K = C$ approaches exhibited high variance across the multi-criteria ranks, indicating a sensibility to the choice of the base regressor. Furthermore, the ensemble approaches that adopted the average-based voting strategy (LOCAL MEAN $K = C$ and LOCAL MEAN $K = 7$) yielded the poorest MCPM values, occupying the fourth and fifth positions for most configurations.

Figure 19 – MCPM values ranked in descending order of importance for each approach regarding different base regressors.



Ordinal ranking	*k*NN	OLS	LASSO	Ridge	ElasticNet	DT	GBDT	MLP	SVR
❶	⬛ 0.0907	🟩 0.0628	⬛ 0.5127	🟩 0.0617	⬛ 0.1585	🟨 0.1133	🟩 0.0823	🟩 0.0573	🟩 0.0893
❷	🟩 0.0975	🟦 0.0647	🟩 0.5436	⬛ 0.0702	🟩 0.3325	⬛ 0.1246	🟨 0.0832	🟦 0.0627	⬛ 0.1094
❸	🟨 0.1112	🟥 0.0887	🟨 0.5576	🟨 0.1183	🟨 0.5274	🟩 0.1389	⬛ 0.0883	⬛ 0.0728	🟦 0.1196
❹	🟦 0.1388	⬛ 0.1183	🟥 0.6148	🟦 0.1229	🟥 0.6262	🟥 0.1674	🟦 0.0953	🟨 0.0834	🟨 0.1432
❺	🟥 0.1487	🟨 0.1205	🟦 0.9194	🟥 0.1512	🟦 0.6712	🟦 0.2207	🟥 0.1137	🟥 0.1097	🟥 0.1745

Approach	
🟦	Global
🟨	Local Mean $K = C$
🟥	Local Mean $K = 7$
⬛	Local Meta $K = C$
🟩	Local Meta $K = 7$

Source: Elaborated by the author.

To compare our proposal and the baseline approaches with the state-of-the-art model, we selected their best configurations in terms of base regressors and arranged their results in Table 12. LOCAL META $K = 7$ achieved the best performances in four out of five metrics, including MCPM, and presented the second-best SP result. HIERARCHICAL GCNN, in turn, presented the worst performance across all the metrics. We must emphasize that the method had parameter values following the best results reported in (LI; PERRIER; XU, 2019), which considered the 2019 Australian election and the traditional cross-validation. Therefore, the present work did not apply a fine-tuning step for Hierarchical GCNN since it is not a step performed in our experimental protocol.

Table 12 – Most promising approaches based on overall configuration performances. Green cells denote the best results.

Approach	Base regressor	MSE	MESD	SP	SPSD	MCPM
Global	MLP	133.87114	138.17511	0.5907	0.1531	0.0681
Local Mean $K = C$	GBDT	196.31742	153.97060	0.6090	0.1560	0.0832
Local Mean $K = 7$	OLS	219.95581	170.12371	0.5850	0.1298	0.0887
Local Meta $K = C$	Ridge	124.69428	129.78054	0.5701	0.1646	0.0702
Local Meta $K = 7$	MLP	111.22174	127.19046	0.5911	0.1355	0.0573
Hierarchical GCNN	GCNN	229.11080	209.11220	0.4917	0.1822	0.1279

Source: Research data.

In summary, our approach proved to be more stable against base regressors from different paradigms than the baselines. Additionally, Local Meta $K = 7$ configured with MLP culminated in the best overall results compared to the best configurations of the other approaches.

4.3.5.2 Performance per Fold

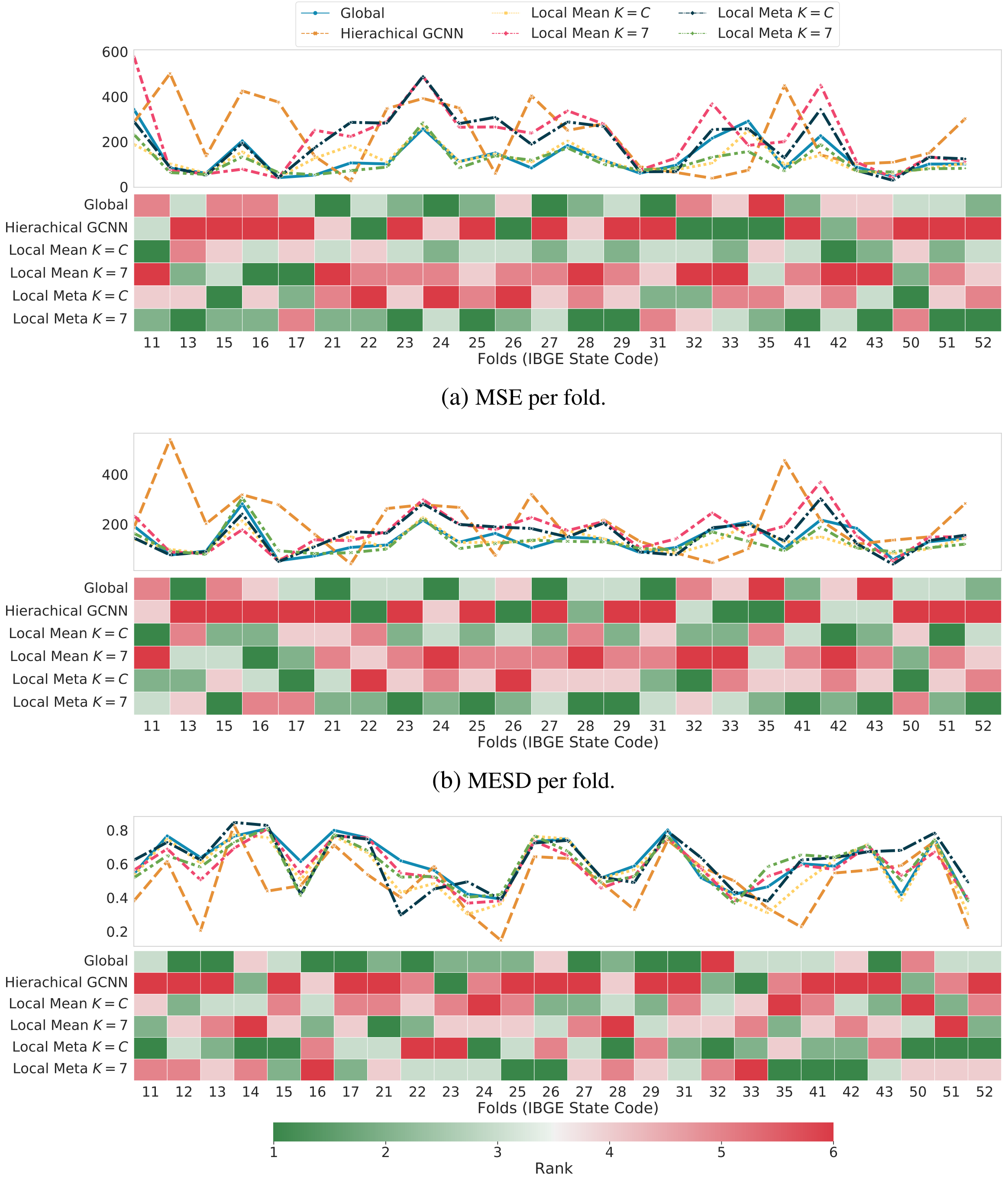
Figure 20 displays the fold-level results of the best-instantiated approaches indicated in Table 12. The performances are reported according to MSE, MESD, and SP. We disregard the SPSD metric here since we cannot produce its values per fold.

Concerning MSE (Figure 20a) and MESD (Figure 20b), Local Meta $K = 7$ behaved stably over the folds, followed by Global. Local Meta $K = 7$ also achieved the best results on most folds, performing exceptionally well in the northeastern states (samples 21 to 29), where it exhibited the best or second best MSE and MESD. Folds 17, 31 and 50, for which Local Meta $K = 7$ was ranked lower regarding MSE, demonstrated close values. Thus, there was no discrepant difference between the investigated approaches. We observed the same in folds 16, 17, and 30 concerning MESD.

In terms of SP (Figure 20c), the Global approach obtained better results than the two variations of our proposal in most folds. However, it was closely followed by Local Meta $K = 7$, specifically in the northeastern states (samples 21 to 29). This fact is reinforced by both approaches' relatively close average performances (Global: 0.61; Local Meta $K = 7$: 0.59).

As we can see, the per-fold analysis of the three individual performance measures corroborates the one guided by MCPM, indicating that the two variations of our approach perform better than the other baseline models. Our proposal exhibited better MSE and MESD results than the other approaches. However, despite presenting lower SP values when compared with Global, the two variations of our proposal showed close results in most folds and performed better in some folds from the Southeast and Northeast.

Figure 20 – Metrics per fold coming from the approaches whose configurations were considered the best by MCPM.



(a) MSE per fold.

(b) MESD per fold.

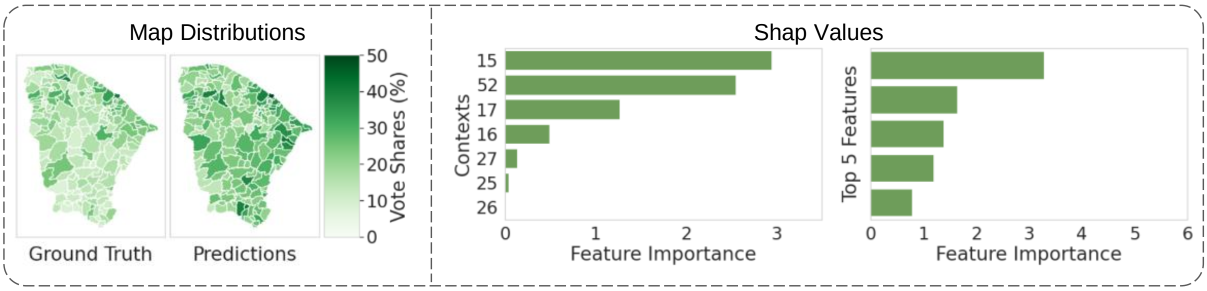
(c) SP per fold.

Source: Elaborated by the author.

4.3.5.3 *Model Interpretability*

Aiming to understand the predictions assigned by the best approach configuration (LOCAL META $K = 7$ with MLP), we considered the SHAP Values technique to produce an in-depth interpretability analysis. We sought to understand the most important context and the most important relevant features from this context in the best (sample 23 or Maranhão) and worst (sample 16 or Amapá) folds regarding MSE. Figure 21 and Figure 22 comprise four plots. The first is the actual vote share distribution, while the second is the predicted distribution. The third concerns the feature importance given by the meta-regressor for each spatial context when predicting the fold. The fourth and last plot presents the top-five most relevant features according to the base regressor fitted to the most important context. Besides, Table 13 and Table 14 list the top-five most important features of Figure 21 and Figure 22, respectively.

Figure 21 – Vote share distribution and SHAP Values for the best-case scenario (fold 23) in terms of LOCAL META $K = 7$ with MLP. The plots cover, from left to right, the following information: (a.1) actual vote share distribution, (a.2) predicted vote share distribution, (b.1) feature importance given by the meta-regressor for each spatial context, and (b.2) top-five most relevant features according to the base regressor fitted to the most important context (sample 15). The feature names are presented in the same order as in Table 13.



Source: Elaborated by the author.

Concerning the best-case scenario (Figure 21), our approach predicted vote shares slightly above the actual values, observed by the number of dark regions in the prediction map in relation to the ground-truth map. The meta-regressor chose the geographic context 15 (Amazonas) as the most important, and the most relevant feature from context 15 was `PessoaRenda V045`. The selection of Amazonas as the most important context to predict the vote shares in Maranhão is coherent since both states present similar vote shares and related socio-economic characteristics. We should note that the first and third most relevant features from context 15 describe the women with income per capita less than half of the minimum wage (Table 13). This result is in line with research that points to the relationship between low-income women and lower votes for the winning party in the 2018 Brazilian presidential election (LAYTON *et al.*, 2021; PINHEIRO-MACHADO; SCALCO, 2020). The remaining characteristics still need to be carefully analyzed to verify if they are proxies for other known related features such as poverty (`Domicilio02 V057` and `Domicilio02 V052`) or a local relationship (`Entorno05 V977`).

In the worst-case scenario (Figure 22), our approach predicted much higher vote shares than the ground truth, especially in the northern region. The meta-regressor chose context 24 (Rio Grande do Norte) as the most important, and the most relevant feature from context 24 was `Entorno04 V490` followed close by `Domicilio02 V040`. The selection of Rio Grande do Norte as the most important context to predict the vote shares in Amapá was not a good decision, given the high error rates. Its features were insufficient to provide a good performance and cannot

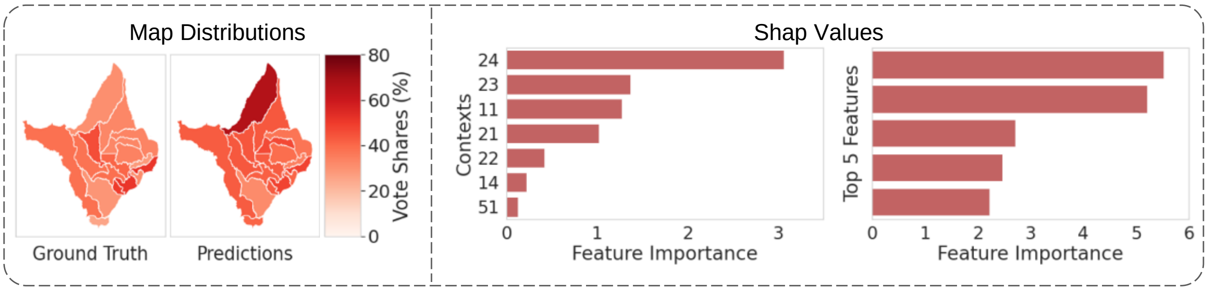
Table 13 – Top-five features from the most important context in the best-case scenario (fold 23).

Feature	Description
`PessoaRenda V045`	Women over ten years with a nominal monthly income of up to half minimum wage
`Domicilio02 V057`	Men living in permanent private households with water supply from a well or spring on the property
`ResplRenda V055`	Total nominal monthly income of responsible women with a nominal monthly income of up to 1/2 minimum wage
`Domicilio02 V052`	Men residing in rented permanent private homes
`Entorno05 V977`	Asian residents in permanent private homes with street lighting

Source: Research data.

deliver insights into Amapá's voting results. We can see from Table 14 that most of the top-five relevant features describe particularities related to rural regions of context 24. These proprieties may not be observed in fold 16 or present a different relationship with the target, causing the approach performance to deteriorate.

Figure 22 – Vote share distribution and SHAP Values for the worst-case scenario (fold 16) in terms of LOCAL META $K = 7$ with MLP. The plots cover, from left to right, the following information: (a.1) actual vote share distribution, (a.2) predicted vote share distribution, (b.1) feature importance given by the meta-regressor for each spatial context, and (b.2) top-five most relevant features according to the base regressor fitted to the most important context (sample 24). The feature names are presented in the same order as in Table 14.



Source: Elaborated by the author.

Despite the challenge in modeling local relationships from socio-economic and election data, the in-depth assessment of SHAP Values indicated that our proposal is intelligible and, at best, predictions are based on coherent features that can aid in understanding electoral outcomes.

4.4 Conclusion

This work proposed a geographic context-based stacking learning approach to predict election outcomes using socio-economic features. Our model is built in levels and dynamically selects contexts according to a data sample we want to predict. This modeling allows the generation of more realistic descriptive models whose relationships enable a more accurate understanding of voting behavior. We also introduced a spatial cross-validation-driven experimental setup to

Table 14 – Top-five features from the most important context in the worst-case scenario (fold 16).

Feature	Description
`Entorno03 V490`	Number of residents in private households without permanent public lighting with a well or spring on the property
`Domicilio02 V040`	Residents in permanent private households with electricity from other sources
`Domicilio01 V026`	Permanent private homes with two bathrooms for the exclusive use of residents
`Domicilio01 V162`	Permanent private dwellings, such as village houses or condominiums, without a bathroom for the exclusive use of residents
`Entorno04 V730`	Residents without nominal monthly household income per capita in permanent private households without sidewalks

Source: Research data.

fairly assess and compare geographically contextualized approaches. Despite the challenging nature of the problem, by considering the second round of the 2018 Brazilian presidential election, our proposal experimentally showed promising results, including intelligible and coherent predictions in the best-case scenario and stable performance over the remaining folds compared with the reference models.

However, there is still room for further improvement. Our approach does not deal with ambiguous distributions, an aspect that often appears in modeling voting behavior. Furthermore, this work was restricted to analyzing a single dataset, and studies with other election databases may be beneficial. Sophisticated machine learning methods such as Graph Neural Networks should also be better evaluated as they have shown satisfactory results for spatial data.

CHAPTER 5

LEARNING BEYOND THE SPATIAL DEPENDENCE STRUCTURE: A REGULARIZED GRAPH-BASED SPATIAL CROSS-VALIDATION APPROACH TO FAIRLY ASSESS CENSUS DATA-DRIVEN ELECTION MODELS

Spatial-contextualized modeling of voting behavior is an essential tool for understanding the electorate and the factors that shape its decision-making, including polarization and socioeconomic conjunctures. Machine learning models have outperformed classic approaches within this topic, especially in identifying patterns and relationships from high-dimensional census data. However, most reference studies do not account for the spatial dependence of the data when validating the models. Cross-validation, a widespread resampling method, limits the exploratory nature of the modeling by biasing it toward the already known spatial dependence structure. We propose RGraphSCV, a Regularized Graph-based Spatial Cross-Validation approach where spatial folds mirror pre-existing geographic boundaries and spatial dependence may occur non-contiguously across space. RGraphSCV uses a bipartite graph structure to determine a removing buffer region that isolates the test from the training set, formalizes the problem as a one-class transductive classification task, and introduces a novel label propagation method that integrates the semivariogram technique to classify nodes from the training set as part of the removing buffer. We evaluate RGraphSCV using three study cases related to recent presidential and congressional elections from Brazil, Australia, and the United States of America. Our experiments demonstrate that RGraphSCV yields less biased and more realistic results in the presence of spatial dependence, making it suitable for assessing machine learning models to identify new patterns and relationships beyond the spatial dependence structure of the data.

5.1 Introduction

The modeling of election voting behavior allows us to comprehend the electorate and the aspects that guide its choices, revealing meaningful insights into studies of political phenomena, such as the extensions of polarization and the demographic and socioeconomic contexts shaping the nature of the electorate (SILVA; PARMEZAN; BATISTA, 2022; PINHEIRO-MACHADO; SCALCO, 2020). In this regard, there is a broad literature on building models to help understand the electorate choices (ELKINK; FARRELL, 2021; LAGO, 2019; FOREST, 2018). This literature can be roughly divided into two main areas: political science and electoral geography (FOREST, 2018). The former argues that electoral behavior can be understood solely through individual factors. In contrast, the latter asserts that individual characteristics and contextual factors, such as location, are crucial in determining electoral behavior. This study focuses on the electoral geography perspective, specifically on utilizing census data to understand electoral behavior from a spatially contextualized perspective.

In particular, census data provide detailed information about a population's socioeconomic and demographic characteristics from thousands of features describing locations in different levels of aggregation. Thus, such data can provide comprehensive and detailed modeling of electoral behavior (SILVA; PARMEZAN; BATISTA, 2022; LI; PERRIER; XU, 2019). However, it comes with two major concerns: spatial autocorrelation and high-dimensionality. The former is a fundamental characteristic of spatial data in which electoral and census data are included. It relates to the spatial dependency between a feature's observations at nearby locations, requiring considering the dependency assumption when evaluating the models to avoid erroneous results and misinterpretation (GETIS, 2010). The latter regards the thousands of explanatory features in census data that comprise a high dimensional space, leading to the curse of dimensionality problem (KEOGH; MUEEN, 2017).

In general, electoral behavior modeling using census data relies mainly on multiple regression analysis using methods incorporating spatial dependence in the model structure, such as mixed effect models and geographically weighted regression (WONG; WONG, 2022; MANOEL; COSTA; CABRAL, 2022). Such modeling is typically used to find linear relationships, based on pre-defined hypotheses, between a small group of explanatory features and the target feature. However, regression analysis becomes a challenge when exploring the high-dimensionality of census data to identify new linear and non-linear relationships. Thus, there exists a need for new methodologies capable of handling high-dimensional data and effectively extracting valuable insights from such data.

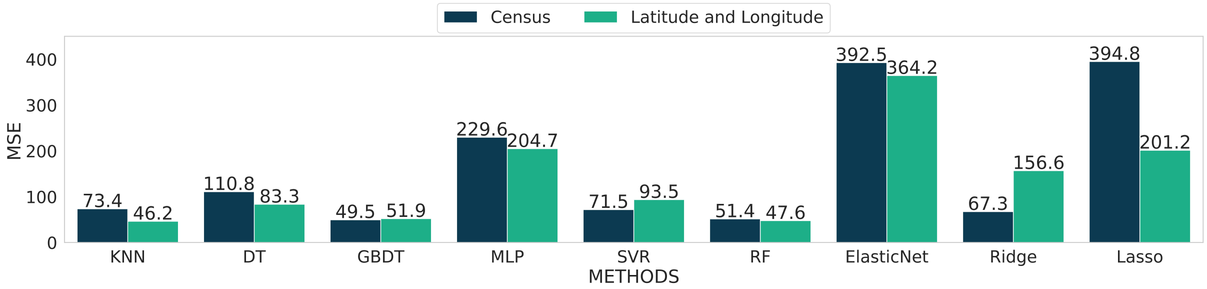
5.1.1 Motivation and Justification

Most recently, there has been an increasing number of works applying machine learning methods to model electoral behavior (SILVA; PARMEZAN; BATISTA, 2022; ELKINK;

FARRELL, 2021; LI; PERRIER; XU, 2019). These methods can model linear and non-linear relationships and deal with high-dimensional data more naturally. Such a shift in methodology has led to a change in the modeling objective, which has become more exploratory rather than based on preconceived hypotheses (SILVA; PARMEZAN; BATISTA, 2022; ELKINK; FARRELL, 2021). Nonetheless, in some works that employ spatial data, the spatial aspects are still neglected (YERO; SACCO; NICOLETTI, 2021; LI; PERRIER; XU, 2019), especially in the validation process, which may lead to overoptimistic results and misinterpretation (PLOTON; MORTIER *et al.*, 2020; ROBERTS *et al.*, 2017).

The standard validation processes applied in the machine learning literature is Cross-Validation (CV) and Hold-Out. They can assess the model generalization on unseen data by splitting the dataset into training and test set. Consequently, they have been applied unconstrainedly to recent works on spatially contextualized electoral behavior modeling using machine learning, disregarding the spatial aspects of data (YERO; SACCO; NICOLETTI, 2021; LI; PERRIER; XU, 2019). Unfortunately, such disregard can impact the quality of the research in the area. It assumes that the results obtained from CV are not influenced by spatial autocorrelation. However, a simple experiment can invalidate such an assumption, as shown in Figure 23. The experiment compares the CV mean-squared error performance of various machine learning methods for predicting city-level election results using two feature sets: one with census data and the other with *only* latitude and longitude. Results show that models learned from features with high autocorrelation information achieved better or comparable results to those learned from census data. Additionally, the methods that exploit spatial dependence information achieved the best results when using latitude and longitude as only features, such as KNN and DT.

Figure 23 – 10-Fold Cross Validation Mean Squared Error (MSE) results obtained from different machine learning methods trained on the census and latitude and longitude features for predicting the vote-shares from the winning party in the second round of the 2018 Brazilian Presidential Election. The acronym for each machine learning method is *k*-Nearest Neighbors (*k*NN), Decision Tree (DT), Gradient Boosting DT (GBDT), Randon Forest (RF), Multiyear Perceptron (MLP), Support Vector Regression (SVR), Least Absolute Shrinkage and Selection Operator (LASSO).



Source: Elaborated by the author.

In summary, when working with data that exhibit spatial dependence, traditional validation techniques can be biased towards models that learn the spatial autocorrelation structure over those that learn different patterns and relationships (ROBERTS *et al.*, 2017). However, when modeling electoral behavior from thousands of features using machine learning methods, the approach is often exploratory in nature (SILVA; PARMEZAN; BATISTA, 2022). Therefore, traditional validation techniques can limit the discovery of new patterns and relationships, as in most cases, it will select as the best model the one that explains electorate choices based on the spatial dependence structure present in the data, which is a well-explored characteristic of electoral data related to the phenomenon known as “Neighborhood Effect" that posits that nearby locations tend to vote similarly (AGNEW, 1996).

Nonetheless, evaluating machine learning models learned from data with spatial dependence is an increasing research topic (MILÀ *et al.*, 2022; PLOTON; MORTIER *et al.*, 2020; ROBERTS *et al.*, 2017). Researchers propose and debate validation techniques in this field that consider the data’s spatial characteristics. In this work, we name such techniques as Spatial Cross Validation (SCV) for simplicity. In general, SCV techniques splits the dataset according to the spatial domain. Thus, the test and training sets are defined by spatial blocks. The idea behind this strategy is that, under positive spatial autocorrelation, nearby data will not be separated into training and test set, diminishing the spatial dependence influence on the results.

Several SCV techniques have been proposed recently for various spatial data applications, including electoral behavior modeling (MILÀ *et al.*, 2022; DEPPNER; CAJIAS, 2022; SILVA; PARMEZAN; BATISTA, 2021; PLOTON; MORTIER *et al.*, 2020; ROBERTS *et al.*, 2017; VALAVI *et al.*, 2019). However, these approaches still have limitations. For instance, some existing SCV techniques are based on Leave-One-Out, which can be time-consuming, particularly for electoral datasets with thousands of examples. Others assume that the removing buffer region can be defined independently of the test set with the same size in all directions. Additionally, some approaches seek spatial independence disregarding the amount of data removed from the training set. Finally, some approaches assume that spatial dependence occurs contiguously, which may not hold in real-world datasets such as census and electoral data.

5.1.2 Main Contributions

In light of these limitations, this chapter contributes to the state-of-the-art in machine learning and spatially contextualized modeling of electoral behavior, proposing a novel SCV approach to asses models learned from thousands of spatial features. Our approach is based on defining a test-dependent removing buffer region to isolate the test from the training set. This allows spatial folds to be selected more freely without significant concerns about the spatial dependence structure. Thus, we present the main contributions of this work as the following:

- A novel SCV approach that produces removing buffer regions that are defined based on the test set allowing pre-existing geographical limitations to be used as spatial folds and assumes that spatial dependence may occur non-contiguously, which is a more realistic scenario;

- The formalization of the removing buffer region problem as a one-class transductive classification task;
- The proposal of a bipartite spatial graph structure to separate the test from the training set to facilitate the definition of the removing buffer;
- A novel label propagation method that integrates the semivariogram technique to classify nodes from the training set as part of the removing buffer based on the test set distribution.

Overall, these contributions improve upon existing approaches and advance the literature on machine learning applied to spatial contextualized modeling of electoral behavior. Thus, in this study, we evaluate the effectiveness of our proposed SCV approach by analyzing the results generated by various machine learning methods in three study cases related to the 2018 Brazilian Presidential Election, the 2019 New South Wales (NSW) Congress Election, and the 2020 United States of America (USA) Presidential Election. Furthermore, by considering a multi-criteria performance analysis, we compare the results obtained by the machine learning methods from our approach with those obtained from existing SCV methods and demonstrate that our approach yields more realistic results and is less biased in the presence of spatial dependence.

5.1.3 Outline

The rest of this chapter is organized as follows: section 5.2 introduces the related work. section 5.3 describes our SCV approach. section 5.4 reports the data and experimental setup. section 5.5 presents the results obtained. section 5.6 provide the discussion related to the results. section 5.7 describes the limitations and recommendations regarding our approach. Finally, section 5.8 concludes the study and highlights future work.

5.2 Related Work

Electoral behavior, or voting behavior, has long been a subject of academic study, with a vast body of literature having been published on the topic (PINHEIRO-MACHADO; SCALCO, 2020; STOKES, 1963). Moreover, most recently, there has been an increase in research on electoral behavior modeling as the general public has become more engaged in elections and concerned about issues that may threaten the foundations of democracy (SILVA; PARMEZAN; BATISTA, 2022; ELKINK; FARRELL, 2021; LI; PERRIER; XU, 2019). This section comprehensively reviews the existing literature on modeling electoral behavior, specifically concerning the electoral geography perspective. We begin by providing an overview of the relevant literature in the field, followed by a brief description of recent works on applying machine learning methods to electoral behavior modeling. Finally, we present the most recent advancements in validation techniques for spatial data proposed in the literature.

5.2.1 Spatially Contextualized Electoral Behavior Modeling

When modeling electoral behavior, two primary perspectives need to be considered: one from the field of political science and the other from electoral geography. These perspectives offer different interpretations and explanations for the factors that influence voting behavior and the implications of these factors on the electoral process (FOREST, 2018). This work focuses on the latter, specifically using census data to provide a spatially contextualized understanding of electoral behavior.

Electoral geographers assert that individual characteristics and contextual factors, such as location, are crucial in determining electoral behavior. They argue that individuals living in nearby locations tend to vote similarly. A phenomenon commonly called the “Neighborhood Effect" (AGNEW, 1996; FOREST, 2018). In this perspective, they usually use data that describes populations at geographical aggregation levels, such as districts and municipalities. Moreover, the objective is to develop hypotheses based on expected relationships to model the data. However, given the inherently spatial nature of the data, spatial models are often employed in the modeling, such as geographically weighted regression. These models consider the presence of spatial dependence in the data, allowing for a more comprehensive understanding of the spatial factors influencing voting (WONG; WONG, 2022; MANOEL; COSTA; CABRAL, 2022).

The general approach to model electoral behavior involves defining a set of hypotheses that describes relationships with a feature of interest. For instance, researchers might hypothesize that low-income populations vote for candidates on the left concerning the ideological spectrum, while high-income population vote for candidates on the right. To test such hypotheses, they often utilize a limited number of features and apply multiple regression analysis as a statistical methodology (WONG; WONG, 2022; MANOEL; COSTA; CABRAL, 2022). However, this approach is not well-suited for handling high-dimensional data, which is the case of census data (SILVA; PARMEZAN; BATISTA, 2021; LI; PERRIER; XU, 2019). In light of this, researchers have begun to employ machine learning methods to fully explore such type of data related to elections and identify new relationships. This shift in methodology has also led to a change in the objective of the modeling, which has become more exploratory rather than testing preconceived hypotheses (SILVA; PARMEZAN; BATISTA, 2022; YERO; SACCO; NICOLETTI, 2021; ELKINK; FARRELL, 2021; LI; PERRIER; XU, 2019).

5.2.2 Machine learning in Electoral Behavior Modeling

To the best of our knowledge, the first study to apply machine learning methods to model electoral behavior from census data was proposed by (LI; PERRIER; XU, 2019). The authors combined the hierarchical nature of census and election data with the capability of Graph-Convolutional Neural Networks (GCNNs) to learn local patterns, resulting in a model that could identify local relationships in the context of the 2019 New South Wales (NSW) congressional election.

Lately, (YERO; SACCO; NICOLETTI, 2021) proposed using decision trees trained on Human Development Indexes (HDI) from various domains, such as education, health, and age, at the municipality level of aggregation to explain the 2018 Brazilian Presidential election. The study identified a correlation between high HDI municipalities and votes for the winning party. Following a similar idea, (SILVA; PARMEZAN; BATISTA, 2022) proposed a geographical context-based stacking learning approach to model the electorate in the context of the 2018 Brazilian Presidential Election from the Brazilian census data. The authors utilized the concept of “Neighborhood Effects" to model the data and mitigate the impact of spatial dependence. Their finding corroborates with other works that point to the impact of women’s vote in the studied election.

While utilizing machine learning methods for modeling the electorate from census data can provide detailed and comprehensive information about the elections, many studies fail to consider the spatial characteristics of the data, particularly regarding evaluation (SILVA; PARMEZAN; BATISTA, 2022; SILVA; PARMEZAN; BATISTA, 2021). Most studies evaluate their models using traditional validation techniques such as CV and hold-out, which are known to produce over-optimistic results when applied to spatial data (PLOTON; MORTIER *et al.*, 2020). Furthermore, in this application, such approaches can bias the selection of models by favoring methods that exploit the spatial dependence present in the data (ROBERTS *et al.*, 2017). As a result, the explanation of electoral results becomes limited to the spatial dependence structure in the data, indicating that nearby locations vote similarly, counterpointing the objective of identifying new relationships. Thus, it becomes necessary to propose new validation techniques that enable a fair evaluation of the models and facilitate identifying new relationships to model the electorate behavior better.

5.2.3 Spatial cross-validation approaches

The advent of SCV techniques is relatively recent, as evidenced by the limited number of methods and discussions on its application (ROBERTS *et al.*, 2017; PLOTON; MORTIER *et al.*, 2020). It has emerged in the map prediction research area as a response to the careless evaluation of machine learning models learned from spatial data that generated overly optimistic results (ROBERTS *et al.*, 2017). The rationale behind this strategy is that data with positive spatial dependence will not be separated into training and test sets, reducing the likelihood of inflated results influenced by spatial dependence (PLOTON; MORTIER *et al.*, 2020). Under this idea, some techniques were proposed in the literature (MILÀ *et al.*, 2022; SILVA; PARMEZAN; BATISTA, 2021; PLOTON; MORTIER *et al.*, 2020; VALAVI *et al.*, 2019; ROBERTS *et al.*, 2017). These methods can be broadly classified based on the traditional validation techniques that they are derived from, specifically K-Fold CV and Leave One Out (LOO).

Following a K-Folds CV setup, the most straightforward method automatically defines spatial folds by applying a clustering algorithm on the target and spatial features (e.g., latitude and longitude) (PLOTON; MORTIER *et al.*, 2020). Yet, another strategy involves using a semivariogram on a fitted model's residuals to define the delimitation of the spatial folds (VALAVI *et al.*, 2019).

Inspired by the Leave One Out (LOO) approach, other SCV approaches, (MILÀ *et al.*, 2022; ROBERTS *et al.*, 2017), define a removing buffer region to isolate the test example from the training set. The buffer region is defined as part of the training data that is close and highly correlated to the test example such that it can cause optimistic results. There are three strategies to define the buffer region. The most straightforward, although time-consuming, is to define manually with the assistance of a specialist (ROBERTS *et al.*, 2017). Another is to consider the semivariogram applied to the target feature or the residuals of a fitted model and define the *range* as the size of the removing buffer to all test folds (PLOTON; MORTIER *et al.*, 2020), generating a test-independent removing buffer region that with an isotropic property – same size in all directions–. Finally, the last strategy involves defining a method that considers information from the test set to automatically define the extensions of the buffer region for each test fold (SILVA; PARMEZAN; BATISTA, 2021; MILÀ *et al.*, 2022).

Regarding electoral behavior modeling, (SILVA; PARMEZAN; BATISTA, 2021) proposed an SCV method to evaluate machine learning models learned from thousands of spatial features in the task of predicting vote-shares. To the best of our knowledge, it is the first utilization of an SCV approach to address this issue. The method follows the K-Fold CV setup. However, the spatial folds are defined by pre-defined geographical delimitations to enhance the interpretability of the results. Since the delimitations were not created for the purpose of reducing spatial dependence, the authors propose a method to generate a buffer region for each of the test folds. The method considers the graph structure provided by lattice-type spatial objects, and for each test fold, it employs the semivariogram to define the buffer regions using the graph neighborhood as the distance metric. The approach was analyzed on a dataset regarding the second round of the 2018 Brazilian Presidential Election and produced more realistic and less biased results than the baselines with which it was compared.

Although some approaches have yielded promising strategies for dealing with spatial data during the validation step, they are unsuitable for electoral behavior modeling. As a response, we proposed an SCV approach inspired by the work of (SILVA; PARMEZAN; BATISTA, 2021). However, our approach is more generalist, with fewer requirements regarding the characteristic of the datasets used. Moreover, it is parameter-free, not requiring the method adjustment for each dataset. Finally, our approach assumes that spatial dependence can occur non-contiguously, a more realistic approach.

5.3 Proposed Method

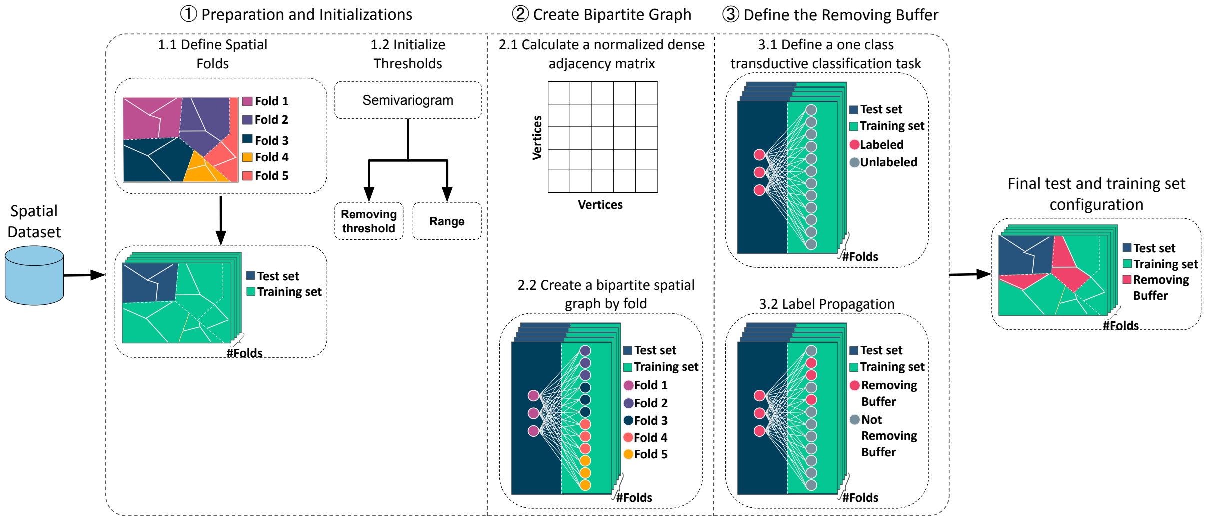
The current validation methods commonly used in machine learning are inappropriate for our specific application. There are several limitations to these methods that need to be addressed. These limitations include:

1. The use of leave-one-out (LOO) based techniques can be computationally expensive in our application since the dataset may present thousands of examples;
2. It is not a realistic assumption that a buffer region can be defined independently of the test set with the same size in all directions.;
3. Applying total independence between test and training regardless of the training set size may fall into a scenario prone to overfitting;
4. Most techniques do not allow the use of pre-existing geographical delimitation as spatial folds, a relevant characteristic to increase results interpretability;
5. Allowing only datasets formed by lattice-type spatial objects limits the modeling possibilities.
6. Requiring that a single component spatial graph should represent the dataset is not a realistic assumption;
7. Assuming that spatial autocorrelation forms a contiguous region may unnecessarily remove non-similar close data;

Given the limitations of the existing validation methods for our specific application, we propose a novel approach called Reg-GBSCV, based on the K-Fold validation technique, to reduce the computational cost associated with LOO-based techniques. Additionally, our method allows pre-existing spatial boundaries as spatial folds, which can improve the interpretability of the results. Moreover, to mitigate the effects of spatial dependence, our method takes a more realistic approach and defines buffer regions dependent on the test set. Such buffer regions can have different sizes and shapes and may not be contiguous. Finally, our approach can be applied to all spatial data types and does not require a single component graph to represent the dataset.

Figure 24 presents an overview of the proposed validation approach. The approach can be divided into three steps for each spatial fold: the preparation and initialization, the bipartite graph building, and the definition of the removing buffer. In the first step, we define the spatial folds (Step 1.1) and the thresholds to define the removing buffers (Step 1.2). Next, we calculate a normalized adjacency matrix based on the similarity between the spatial objects' target feature (Step 2.1). After that, we construct a bipartite graph based on the considered spatial test fold (Step 2.2). Using a one-class transductive classification task, we identify the buffer region for the test fold in the third step. In this task, we seek to identify the vertices in the training set that are part of the removing buffer. To do so, we labeled the test set vertices as part of the removing buffer (Step 3.1). Subsequently, we apply a label propagation method that we propose for this task (3.2). Our method utilizes the semivariogram equation to propagate the test set labels to the vertices in the training set. This approach maintains a consistent definition for buffer regions based on the semivariogram, similar to existing methods. The final training set is thus composed of the data that were not labeled as part of the removing buffer region. The following subsections provide a detailed explanation of these steps.

Figure 24 – Overview of our approach.



Source: Elaborated by the author.

5.3.1 Preparation and Initialization

The first step of our approach can be seen as a preparation step in which we define the spatial folds and the thresholds involved in the process. The definition of spatial folds is a straightforward and single-time step where pre-existing spatial boundaries are defined as the fold's spatial delimitation (SILVA; PARMEZAN; BATISTA, 2021). On the other hand, the threshold initialization steps occur for each test fold. Thus, before calculating the removing buffer for a given test fold, we need to initialize two thresholds: the removing threshold (σ) and the *range* (ρ). These thresholds are derived from the semivariogram technique, which is the core of our method.

$$\gamma(h) = \frac{1}{2|C(h)|}\sum_{i=1}^{C(h)}(z(o_i)-z(o_j))^2 \tag{5.1}$$

The semivariogram is a widely used technique that provides information regarding the spatial dependence structure in spatial datasets (CURRAN, 1988). It calculates the spatial variance (γ) from a given feature regarding a determined lag distance. To clarify, let us consider O a set of spatial objects where each object o_i is a location delimited in the spatial domain (*e.g.*, pooling places, districts and cities). Thus, we can describe the semivariogram calculation as in Equation 5.1, where $C(h)$ is the set of all spatial object pairs (o_i, o_j) separated by a distance h, and $z(o_i)$ is the value of feature z from spatial object o_i. Moreover, the first lag distance h where γ reaches the *sill*—variance of feature z over the entire dataset—is called *range* and corresponds to the distance where the autocorrelation is not present in the dataset.

$$\sigma = \frac{sill(z(O)) + sill(z(O^{Te}))}{2} \tag{5.2}$$

The first threshold we calculate is the *range* (ρ). We use it to identify the distance limit we can search for spatial dependence. It can be simply calculated by applying the semivariogram on the entire dataset's target feature. Thus, ρ is the same for all test folds. Next, we calculate the removing threshold that will be used to define the removing buffer region (Equation 5.2). Following the semivariogram property that indicates that the *sill* is the threshold where the spatial dependence does not exist. We calculate a removing threshold (σ) based the semivariogram's *sill* for each test fold (Equation 5.2). Our threshold is the average between the variance of the target feature regarding the entire dataset ($sill(z(O))$) and the variance of the target feature regarding the test-fold ($sill(z(O^{Te}))$). By defining the threshold as the average of the global *sill* and the test fold *sill*, we relax the test fold sill by adding information from the global variance allowing the identification of locations with weak spatial dependence, that can still interfere in the modeling.

In brief, the specifications for the thresholds are closely linked to the semivariogram methodology, incorporating its fundamental characteristics. Subsequently, after determining the values for the *range* (ρ) and the removing threshold (σ), the construction of the bipartite graph follows as the subsequent stage.

5.3.2 Creating the Bipartite Graphs

Spatial data can be represented as a graph, where the vertices of the graph correspond to a spatial object such as a state or municipality, while the edges are based on the proximity of the vertices in geographical space or the feature space that describes the spatial objects (LI *et al.*, 2019). Our approach takes advantage of this representation by defining a bipartite spatial graph that can be mathematically described as follows.

First, consider a complete and undirected bipartite graph $G(Te, Tr, W)$ in which Te and Tr are two disjoint and independent sets of vertices. Thus, Te corresponds to the set of vertices representing the spatial objects in O that compose the test set, while Tr is the set of vertices representing the spatial objects in O that composes the training set. Moreover, W is composed by the weights on the edges connecting vertices from the two sets Te and Tr, with the property that all vertices from one set are linked to vertices in the other set. However, there are no connections between vertices within the same set.

Now, consider the weight $w_{te,tr} \in W$, which represents the connection between the vertices $te \in Te$ and $tr \in Tr$. The calculation of $w_{te,tr}$ is defined in Equation 5.3, where $z(te)$ and $z(tr)$ are the target feature associated with the vertices te and tr. Additionally, a penalization factor $P_{te,tr}$ is incorporated to penalize edges with vertices whose spatial distance is greater than the *range* (ρ).

$$w_{te,tr} = (z(te) - z(tr))^2 + P_{te,tr} \tag{5.3}$$

The penalizer factor $P_{te,tr}$ is calculated as expressed by Equation 5.4. The calculation of $P_{te,tr}$ involves the spatial distance $g_{te,tr}$ between the vertices te and tr. The penalizer factor aims to impose a penalty on edges that connect vertices whose spatial distance exceeds the *range* (ρ). Vertices whose spatial distance is less than ρ receive no penalization, whereas vertices whose distance is greater than ρ receive a penalization equal to the *sill* times ratio of the distance to the *range*. This mechanism ensures that we only consider spatial dependence within the specified *range*.

$$P_{te,tr} = \begin{cases} 0, \text{ if } g_{te,tr} < \rho \\ \sigma * g_{te,tr}/\rho, \text{ if } g_{te,tr} \geq \rho \end{cases} \tag{5.4}$$

Finally, normalization and inversion of the weights are performed as described in Equation 5.5, such that vertices with weights closer to 1 indicate a high degree of similarity between the vertices, while vertices with weights closer to 0 indicate low similarity. Additionally, the removing threshold is normalized to be in the same scale as the weights (Equation 5.6). This inversion of the weights also results in an inverted behavior of the removing threshold compared to the semivariogram. Values higher than the threshold are now considered to represent spatial dependence, while values lower than the threshold indicate the absence of spatial dependence. These transformations are required to apply the label propagation to identify buffer regions, but they do not alter the fundamental idea behind the semivariogram.

$$W = 1 - W/max(W) \tag{5.5}$$

$$\sigma = \sigma/max(W) \tag{5.6}$$

In summary, for each spatial fold, a bipartite graph is constructed that connects the test set locations to the training set locations. This graph structure emphasizes the relationship between the test and training sets, enabling a more accurate evaluation of the spatial dependence between these two sets by eliminating irrelevant information, such as connections between locations within the same group. After constructing the bipartite graph, the subsequent step involves the removing buffer definition.

5.3.3 Defining the Buffer Region

This section addresses the challenge of ensuring independence between the test and training sets by defining a removing buffer, which is a crucial step as pre-existing spatial delimitations used to define the folds do not guarantee such independence. We first define the problem as a transductive one-class classification approach. Then we apply a novel label propagation method that incorporates the semivariogram properties to identify vertices in the training set exhibiting spatial dependence with the test set and label them as part of the removing buffer.

5.3.3.1 The One Class Transductive Classification Task

In the context of our proposed approach, the initial step towards identifying the removing buffer involves defining a specific classification task. Due to its nature, this step has two distinct characteristics: the presence of only one class of interest and all the vertices to be classified are observable. Since all the vertices to be classified are observed, this task can be categorized as a transductive classification problem, allowing the usage of methods that use information from vertices to be classified to improve classification accuracy. Thus, this task aims to develop a transductive one-class classification model to classify vertices that exhibit spatial dependence regarding the test set.

In this task, we inverted the roles for the test (Te) and training sets (Tr). Thus, all vertices in Te are labeled as part of the removing buffer class, so the model can classify the unlabeled vertices in Tr as belonging to the removing buffer based on the observed class distribution from Te. It is important to note that this inversion is exclusive to this stage and is not used in further steps. Moreover, this operation focuses solely on the target attribute, which is the attribute we are seeking to diminish spatial dependence. Thus no other attribute is used in this step.

In summary, we characterize the removing buffer definition step as a transductive one-class classification task, allowing a variety of machine learning techniques to be used. In the next step, we describe our proposed label propagation approach incorporating the semivariogram technique to identify vertices presenting spatial dependence in the training set.

5.3.3.2 The Novel Label Propagation Approach

Label propagation is an iterative, less computationally costly solution for methods based on graph regularization (SANTOS *et al.*, 2020). Graph regularization is a category of transductive graph classification methods that relies on class information (ZHU; GHAHRAMANI; LAFFERTY, 2003). Class information, in this context, refers to the pertinence of a given vertice concerning a particular class. Specifically, vertices with higher class information are more likely to belong to the corresponding class. Thus the goal in such methods is to minimize an objective function satisfying two premises: (i) the class information of neighboring vertices must be similar; (ii) the predicted class information of vertices must be similar to their true class information. These premises can be expressed as a function called the General Regularization Framework, as shown in Equation 5.7.

$$Q(\mathbf{F}) = \frac{1}{2} \sum_{v_i,v_j \in \mathscr{V}} w_{v_i,v_j} \Omega(\mathbf{f}_{v_i}, \mathbf{f}_{v_j}) + \mu \sum_{v_i \in \mathscr{V}^L} \Omega'(\mathbf{f}_{v_i}, \mathbf{y}_{v_i}) \tag{5.7}$$

Regarding the General Regularization Framework (Equation 5.7), $\Omega(.)$ corresponds to the first premise and computes the proximity between the class information of each pair of vertices in the graph using a distance or dissimilarity function. On the other hand, $\Omega'(.)$ represents the second premise and computes the proximity between the predicted class information and the true class information from a labeled vertex. In this way, $\mathscr{V}$ is the set of all vertices in the graph, $\mathscr{V}^L$ is the set of all labeled vertices in the graph, w_{v_i,v_j} indicates the weight of the relationship between vertices v_i and v_j, μ is a parameter that indicates the importance of true class information, $\mathbf{f_{v_i}}, \mathbf{f_{v_j}}, \in F$, and $\mathbf{y_{v_i}} \in Y$ are class information vectors, where F presents the class information of all vertices, and Y presents the labeled vertices true class information.

Based on General Regularization Framework, we propose a method that considers a bipartite graph structure for identifying a removing buffer to ensure independence between the test and training set. The proposed approach is summarized by the objective function shown in Equation 5.8, which has two terms. The first term propagates the class information from the test vertices (Te) to the training vertices (Tr). The second term ensures that the vertices' labels in Te do not change, we use the limit as μ approaches infinity to force the difference between the class information from vertices in Te and their true class information to be zero. This way, $w_{te,tr} \in W$ presents the relationship between a $te \in Te$ and vertice $tr \in Tr$, where W is calculated as in subsection 5.3.2. Moreover, $y_{te} \in Y^{te}$ is the vector representing the true class information of vertice te. Finally, $f_{te} \in F^{te}$ and $f_{tr} \in F^{tr}$ are class information vectors from vertices te and tr, respectively, where all the vertices from F^{te} present class information equal to 1, while the vertices in F^{tr} are calculated according to Equation 5.9.

$$Q(\mathbf{F}) = \frac{1}{2} \sum_{te \in T_e} \sum_{tr \in T_r} w_{te,tr} (\mathbf{f}_{te} - \mathbf{f}_{tr})^2 + \lim_{\mu \to \infty} \mu \sum_{te \in T_e} (\mathbf{f}_{te} - \mathbf{y}_{te})^2 \tag{5.8}$$

The training vertices' class information is the sum of all its neighbors' class information multiplied by the respective weights and divided by its degree. The idea is that we are calculating the semivariogram for each vertex in the training set. From this perspective, we can draw a parallel between the semivariogram equation (Equation 3.1) and Equation 5.9. Thus, we calculate the sum of the differences between the neighboring vertices from tr as in the semivariogram, which can be seen in Equation 5.3 on how we calculate the weights and divide them by the number of pairs involved, which in our case corresponds to the degree of vertice tr. By incorporating the semivariogram into the class information equation, we allow its properties to be considered to classify vertices as presenting spatial dependence regarding the test set.

$$\mathbf{f}_{tr} = \frac{\sum_{te\in Te} w_{tr,te} \cdot \mathbf{f}_{te}}{2 * \sum_{te\in Te} \lceil w_{tr,te} \rceil} \tag{5.9}$$

We use a label propagation approach described in Algorithm 3 to solve the proposed equation. This approach requires the test and training sets, Te and Tr, respectively, the vector Y^{te} representing the true class information of labeled vertices, the diagonal matrix D with the degree information of each vertice, and the removing thresholds σ calculated as shown in subsection 5.3.1. The output of the approach is the set of vertices that composes the removing buffer. The algorithm starts by calculating the matrix P, corresponding to the weights divided by the degree of each vertice (line 1). Next, while a stop criterion is not satisfied (line 2), the vector of class information is calculated (line 3), and the values from labeled vertices are maintained (line 4). The final step corresponds to the removing buffer, consisting of the vertices from Tr with class information higher than the removing threshold (line 6). To determine the stop criterion, we calculate the difference between the previous class information calculated f'_{tr} and the current class information f_{tr} from the vertices in Tr (Equation 5.10). The stop criterion is satisfied when the difference equals or near zero. In our application context, since we apply our method on a bipartite graph, the method converges with one iteration. The difference between the class information from the first iteration to the second equals zero.

Algorithm 3 – Label Propagation for the Removing Buffer

```
Require: $Te, Tr, \mathbf{Y}, \mathbf{W}, \mathbf{D}, \sigma$
1: $\mathbf{P} \leftarrow (2 \cdot \mathbf{D})^{-1} \cdot \mathbf{W}$
2: while stop criterion do
3:     $\mathbf{F} \leftarrow \mathbf{P} \cdot \mathbf{F}$
4:     $\mathbf{F^{te}} \leftarrow \mathbf{Y^{te}}$
5: end while
6: return $Tr(\mathbf{F^{tr}} > \sigma)$
```

$$\sum_{tr\in Tr} \left| f'_{tr} - f_{tr} \right| \tag{5.10}$$

In summary, we propose a label propagation approach that utilizes the semivariogram technique to identify the vertices composing the removing buffer. Our method is designed to automatically classify the vertices that exhibit spatial dependence regarding a group of labeled vertices. Finally, the empirical evaluation section further evaluates the proposed approach, where we define the evaluation setup to assess its performance in the Spatial Cross Validation context.

5.4 Empirical Evaluation

The suitability of our approach is evaluated using three case studies: the 2018 Brazilian Presidential Election, the 2019 New South Wales (NSW) Congressional Election, and the 2020 United States Presidential Election. These elections have garnered significant attention from researchers due to the highly polarized political climate in which they took place and concerns about the integrity of democratic institutions (LAYTON *et al.*, 2021; CAMERON; MCALLISTER, 2019; NORRIS, 2019). Additionally, these elections demonstrate high levels of spatial dependence, suggesting that voting patterns are similar among neighboring locations (JACINTHO *et al.*, 2020; LI; PERRIER; XU, 2019). As such, they provide an ideal scenario to test and evaluate our method.

5.4.1 The Datasets

The datasets used in this study combine census data and election data based on the considered aggregation level. Detailed information about each dataset is presented in Table 15, where it can be observed that the Brazilian election dataset comprises 5565 municipalities, 3999 explanatory features, and the target feature is the vote-shares received by the Jair Bolsonaro party. The Australian election dataset, on the other hand, focuses only on the state of New South Wales and consists of 2262 postal areas, 4000 explanatory features, and the target feature is the vote-shares received by the Labor Party. Finally, the United States election dataset encompasses 3143 counties, 400 explanatory features, and the target feature is the vote-shares received by Joe Biden.

Table 15 – Datasets properties

Dataset	Election	Year	Aggregation Level	#Instances	#Census	Target
Brazil	Presidential	2018	City	5165	3999	Jair Bolsonaro vote-shares
NSW	Congress	2019	Postal Areas	2262	4000	Labour Party vote-shares
USA	Presidential	2020	County	3153	400	Joe Biden vote-shares

Source: Research data.

The census data utilized in this study was obtained from official sources such as the Brazilian Institute of Geography and Statistics (*IBGE*), the Australian Bureau of Statistics (*ABS*), and the United States Census Bureau (*USCB*). These data sets were made available to the public through anonymized aggregated features that describe population groups defined by geographic regions, such as municipalities, postal areas, and counties, corresponding to the aggregation levels employed in this research. Moreover, to ensure accurate and reliable results, we standardized all the features concerning the population size or the number of households, depending on the aggregation level.

The election data used in this study was obtained from the Superior Electoral Court (*TSE*) for the Brazilian presidential election, the Australian Election Commission (AEC) for the

New South Wales Congressional Election, and the MIT Election Data Science Lab for the United States Presidential Election. The data includes vote count results for each ballot box, which were aggregated according to the chosen level of aggregation. The target features are the valid votes percentual received by the chosen parties, named vote-shares.

The datasets analyzed in this study are pertinent as they pertain to elections in countries currently of significant interest in the research community. Furthermore, these datasets exhibit a pronounced spatial dependence on the vote-shares, the target feature. This characteristic necessitates using an SCV approach to assess the performance of the machine learning models derived from the datasets.

5.4.2 Spatial Cross-Validation Approaches

The use of SCV in evaluating models trained on spatial data for modeling electoral behavior is a recent development. To our knowledge, only one method has been proposed for this task in the literature (SILVA; PARMEZAN; BATISTA, 2021). Therefore, we follow a similar evaluation setup as in (SILVA; PARMEZAN; BATISTA, 2021) and consider four validation techniques for comparing our method's results: K-Fold CV, Conservative Approach, Optimistic Approach, and the RBuffer SCV proposed in (SILVA; PARMEZAN; BATISTA, 2021).

In this study, the K-Fold Cross Validation and Optimistic approach are employed as baselines to compare against our proposed method. The K-Fold CV is included because it is known to be affected by the spatial dependence structure in the data, and we aim to avoid such bias. The Optimistic approach, on the other hand, uses pre-existing spatial delimitations to define folds but does not employ a removing buffer region and prioritizes model performance over independence between test and training sets. This approach is crucial in understanding the role of removing buffers. Additionally, we adopt the Conservative approach, which prioritizes spatial dependence over model performance by calculating a uniform size for buffer regions across all spatial folds. This is achieved by applying a semivariogram on the entire dataset and computing the *range*. Finally, we compare our proposed method against the RBuffer approach, with parameters set according to the original paper.

The selection of the baselines and approaches in this study was based on the application's specific characteristics, aiming to provide the most suitable scenarios to compare and evaluate the proposed approach's performance. Although SCV techniques inspired by Leave One-Out are also available, they were not considered due to their potential for computational cost and time consumption.

5.4.3 Machine Learning Algorithms and Evaluation Measures

Assessing a wide range of paradigms is crucial to obtain a comprehensive understanding of their behavior in all validation approaches. Furthermore, comparing machine learning models across different validation approaches is necessary to establish a fair and accurate assessment. Finally, employing appropriate performance metrics from the specific application under evaluation is crucial to provide a comprehensive and reliable comparison.

In this study, nine machine learning algorithms belonging to different paradigms were considered. The algorithms and their parameters are listed in Table 16. To reduce the attribute space, we adopted the Correlation-based Feature Selection (CFS) method. The objective of CFS is to identify a minimal subset of features that are highly correlated with the target attribute and are not very redundant with each other.

Table 16 – Machine learning algorithms and their parameters. The acronyms not yet defined are: k-Nearest Neighbors (kNN), Least Absolute Shrinkage and Selection Operator (LASSO), Decision Tree (DT), Gradient Boosting DT (GBDT), Random Forest (RF), Multiyear Perceptron (MLP), and Support Vector Regression (SVR).

Base regressor	Parameter	Value
kNN	Number of nearest neighbors (k)	3
LASSO	Regularization strength (α)	1
Ridge	Regularization strength (α)	1
ElasticNet	Constant that multiplies the penalty terms (α)	1
	Mixing parameter ($l1\_ratio$)	0.5
DT	Split criterion	GINI
GBDT	Number of boosted trees to fit ($n\_estimators$)	100
RF	Number of boosted trees to fit ($n\_estimators$)	100
	Learning rate (ε)	0.1
MLP	Hidden layers (h)	1
	Hidden layer size (n)	$M/2$
	Learning rate (ε)	0.001
SVR	Kernel	RBF
	Gaussian's width of the radial basis kernel function (σ)	$1/(M*X.var())$
	Regularization parameter ($\mathbb{C}$)	1

Source: Research data.

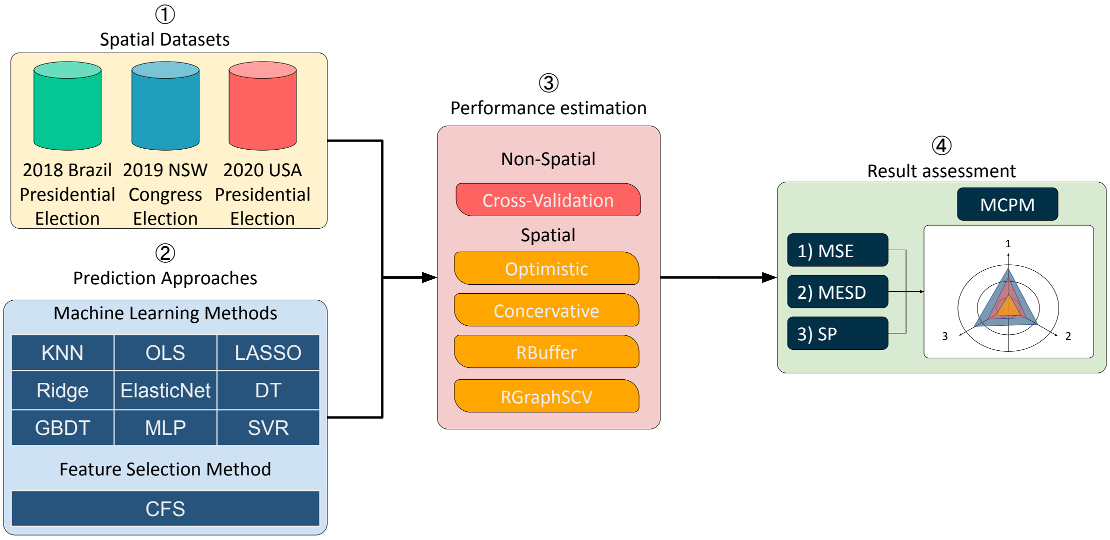
We employed three well-established performance measures, namely Mean Squared Error (MSE), Mean Error Standard Deviation (MESD), and Spearman correlation (SP), to evaluate the performance of the SCV approaches presented in subsection 5.4.2. In addition, we used a Multi-Criteria Performance Measure (MCPM) (PARMEZAN; LEE; WU, 2017) to provide an overall view obtained from different SCV approaches by combining the four metrics. The MCPM is calculated as the sum of the total area of an irregular polygon whose vertices comprise individual performance indexes, and lower total area values indicate better predictive performances. It is worth noting that unlike MSE and Standard Deviation measures, where the resulting values must be minimized, SP (ρ) must be maximized, and hence, we applied the SP complement: $\hat{\text{SP}} = 1 - \rho$.

In summary, this study considered machine learning approaches from different paradigms widely used in the literature to ensure a comprehensive evaluation of validation techniques. The chosen approaches are the traditional and most commonly used ones. In terms of performance measures, those selected are commonly used in evaluating regression models in the machine learning literature. They were also appropriate for assessing models in the task of electoral behavior modeling (SILVA; PARMEZAN; BATISTA, 2022).

5.4.4 Experimental Setup

In this study, we designed an experimental setup to evaluate the results generated by different validation techniques and compare them against our proposed approach. The experimental procedure consists of four main steps, as depicted in Figure 25. Firstly, we consider three datasets following the description in subsection 5.4.1 (Step 1). Then, machine learning methods from different paradigms and a feature selection technique were then considered and parameterized according to Table 5.4.3 (Step 2). Next, the machine learning methods and datasets were applied to different validation techniques described in subsection 5.4.2 (Step 3), and the obtained results were reported via the evaluation metrics described in Table 5.4.3 (Step 4). To ensure a fair comparison, the evaluation process for a given fold starts by filtering the features using the CFS on the training set, followed by fitting the models on the training set to predict the test fold using machine learning methods. The spatial folds were defined equally for all validation techniques in this study, except for the CV that defines its folds randomly.

Figure 25 – Experimental setup.



Source: Elaborated by the author.

In this work, we defined the spatial folds for each dataset as the 26 Brazilian states, the 47 Australian electoral divisions within the state of NSW, and the 49 states of the USA. In contrast, we consider the commonly used ten folds for the CV. The spatial folds defined pose challenging scenarios for the SCV approaches since they were not defined for such purposes. However, they facilitate the analysis of the results by higher-level stakeholders. Moreover, the results obtained from the considered metrics are utilized to evaluate our approach's applicability toward the spatially contextualized modeling of electoral behavior. While the ultimate aim is to identify novel patterns and relationships, comparing the results obtained against those of CV aids in determining whether the SCV approach is influenced by spatial dependence. Thus, results similar to those obtained from CV indicate that the SCV approach may be constraining the discovery of new relationships and patterns in favor of the spatial dependence structure of the data.

Nonetheless, the current experimental setup exhibited some exceptions that warrant explication. Specifically, two aspects of the experimental design necessitated elaboration: the selection criteria for spatial folds and the validation techniques applied to the datasets. We opt to include only spatial folds with a size greater than 10 to ensure the reliability of Spearman correlation measures. Consequently, the 2020 USA Election dataset was the only dataset affected by this criterion. The states of Hawaii, Rhode Island, New Hampshire, and Delaware were excluded from consideration as spatial folds. Moreover, the application of the RBuffer to the 2019 NSW election dataset was not feasible due to the dataset's failure to meet the requisite criteria for the approach to be used. Notably, the dataset is represented by more than one component, posing a challenge for the RBuffer approach.

Finally, for implementing the experimental setup described in Figure 25, we utilized the Python programming language and several libraries, including Pandas, GeoPandas, SciPy, PySAL, and Scikit-learn. The source code and supplementary materials related to this study are available on GitHub[1].

5.5 Results

Table 17 presents the overall results regarding the metrics MSE, MESD, and $\hat{\text{SP}}$ obtained by the machine learning methods in each validation approach regarding the datasets, where the values are mapped from red to yellow to green. Thus, cells in dark green indicate the lowest values, cells in yellow indicate median values, and cells in dark red indicate the highest values. The results indicate that the CV technique yielded the lowest values regarding all the metrics for most of the machine learning methods, followed by the Optimistic approach with several light green cells in the MSE and MESD metrics. The Conservative approach produced the highest values in most scenarios, shown by the high number of red cells in the MSE and $\hat{\text{SP}}$. Finally, our proposed approach and RBuffer presented median values, with our approach demonstrating slightly higher values than RBuffer.

Fig. 26 provides a better glance at the performance outcomes by displaying the MCPM values across all datasets and machine learning methods, offering a visual representation of the results patterns identified in Table 17. Concerning the 2018 Brazilian Election dataset (Figure 26a), all machine learning methods performed with the lowest MPCM values when

[1] anonymized.

Table 17 – Overall results of the approaches considering each machine learning method for all the datasets regarding the following metrics: MSE, MESD, $\widehat{SP}$. Dark green cells symbolize the lowest results, yellow cells symbolize median values, and dark red cells symbolize the highest values.

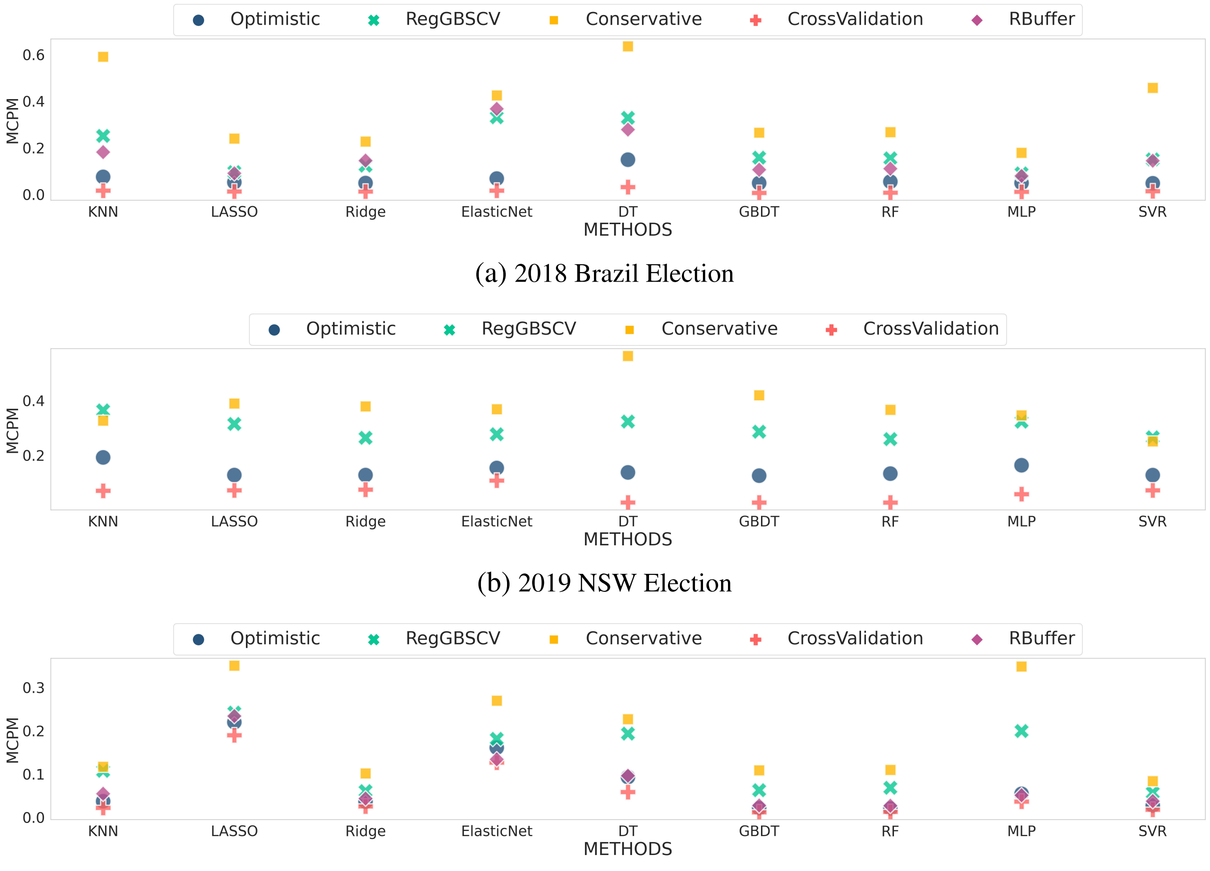
	ML Methods	Optimistic			RGraphSCV			Conservative			Cross-Validation			RBuffer		
		MSE	MESD	$\widehat{SP}$	MSE	MESD	$\widehat{SP}$	MSE	MESD	$\widehat{SP}$	MSE	MESD	$\widehat{SP}$	MSE	MESD	$\widehat{SP}$
2018 Brazilian Election	KNN	137.61	151.43	0.45	366.43	295.44	0.62	700.19	526.00	0.78	73.42	157.55	0.11	267.30	285.77	0.54
	LASSO	122.87	115.90	0.39	220.75	169.97	0.43	424.29	283.52	0.49	67.76	128.84	0.10	190.80	167.43	0.43
	Ridge	119.70	112.75	0.37	298.61	194.42	0.42	453.24	271.77	0.51	67.31	126.64	0.10	338.50	211.62	0.42
	ElasticNet	170.22	139.62	0.39	632.70	332.71	0.49	690.57	370.62	0.62	89.65	137.00	0.11	705.30	357.64	0.49
	DT	191.57	235.05	0.58	415.93	420.99	0.64	608.00	619.37	0.82	110.74	210.50	0.15	317.95	369.42	0.68
	GBDT	112.71	117.22	0.36	297.90	225.96	0.49	421.10	301.52	0.63	49.46	91.71	0.07	208.11	177.86	0.49
	RF	115.22	116.66	0.36	303.37	222.87	0.48	438.00	284.18	0.64	51.13	99.78	0.08	213.48	175.41	0.48
	MLP	123.86	117.47	0.36	219.79	174.96	0.40	354.83	240.61	0.53	63.24	120.07	0.10	187.91	160.64	0.39
	SVR	119.72	113.14	0.35	290.69	205.42	0.48	746.13	384.06	0.63	71.54	134.42	0.10	284.13	223.66	0.45
2019 NSW Election	KNN	197.73	241.13	0.79	406.30	357.72	0.84	328.10	307.32	0.94	142.40	261.40	0.30	-	-	-
	LASSO	161.88	187.71	0.72	394.20	285.10	0.79	395.78	295.18	1.04	153.02	210.44	0.37	-	-	-
	Ridge	166.36	185.74	0.72	333.60	259.23	0.78	392.51	289.81	1.04	162.50	211.81	0.37	-	-	-
	ElasticNet	212.82	200.23	0.72	353.71	265.32	0.79	393.11	290.33	1.03	216.68	262.86	0.40	-	-	-
	DT	150.92	207.17	0.71	381.78	349.20	0.80	520.15	410.76	1.04	88.69	164.68	0.19	-	-	-
	GBDT	141.64	194.10	0.72	350.43	314.11	0.77	415.68	342.60	1.00	88.60	163.69	0.19	-	-	-
	RF	145.63	197.94	0.74	328.21	305.37	0.75	379.45	311.34	1.01	88.70	163.65	0.19	-	-	-
	MLP	170.56	198.30	0.78	384.08	293.30	0.78	390.49	287.58	0.96	132.96	205.31	0.31	-	-	-
	SVR	164.39	186.92	0.72	346.48	279.90	0.77	287.99	223.66	0.97	156.71	208.77	0.36	-	-	-
2020 USA Election	KNN	89.45	127.40	0.31	223.06	224.24	0.36	210.88	224.00	0.43	71.48	131.60	0.20	127.75	154.78	0.34
	LASSO	295.27	337.77	0.55	395.71	328.40	0.46	430.78	344.12	0.79	242.08	353.50	0.59	312.33	335.59	0.56
	Ridge	108.64	128.69	0.24	180.11	157.36	0.25	227.63	192.28	0.32	83.20	137.80	0.21	131.29	135.55	0.25
	ElasticNet	260.76	283.77	0.43	360.94	286.78	0.36	411.70	319.96	0.52	202.78	301.41	0.45	268.75	273.49	0.35
	DT	152.80	230.76	0.41	293.41	328.48	0.50	293.88	336.04	0.55	121.56	235.09	0.29	175.04	233.01	0.40
	GBDT	68.27	98.25	0.23	163.70	166.70	0.28	198.53	197.08	0.38	51.60	101.97	0.14	92.30	111.46	0.23
	RF	72.75	98.13	0.22	188.96	179.00	0.28	203.44	189.95	0.38	55.03	104.19	0.15	95.94	109.16	0.22
	MLP	123.08	153.17	0.27	311.80	257.97	0.41	377.61	355.49	0.65	99.27	169.59	0.24	148.48	147.86	0.26
	SVR	84.13	110.33	0.24	163.92	151.91	0.26	181.67	181.41	0.36	65.79	116.42	0.18	114.12	123.96	0.25

Source: Research data.

subjected to CV, while the Optimistic approach reported slightly higher values. Notably, the proposed method achieved comparable outcomes to RBuffer, whereas the Conservative approach yielded the highest values. Moreover, we observed identical patterns in the 2019 NSW Election dataset (Figure 26b), where the proposed approach is relatively similar to the Conservative and slightly superior regarding KNN and SVR. Finally, concerning the 2020 USA Election dataset (Figure 26c), the RBuffer method reported results closer to the Optimistic and CV approaches. Conversely, the proposed technique achieved slightly higher values across most machine learning methods.

In more detail, Table 18 presents the lowest and highest MCPM values obtained in each validation technique across the datasets. A first assessment of the results shows that the GBDT produced the lowest values when subject CV in all the datasets. This implies that the GBDT algorithm could effectively exploit the dependence structure in the datasets, thereby yielding superior performance outcomes. The GBDT method also generated the lowest MCPM value in the Optimistic approach for the 2019 NSW Election dataset. Moreover, regarding the SCV approaches, the machine learning methods that produced the lowest MCPM results were MLP, SVR, and RF, which are considered more sophisticated models. Lastly, the highest values were reported by the KNN, DT, and LASSO methods. The first two are considered methods that take advantage of the spatial dependence structure presented in the data. Notably, the LASSO method

Figure 26 – MCPM statistics generated for each machine learning method regarding the validation approaches for all datasets.



(a) 2018 Brazil Election

(b) 2019 NSW Election

(c) 2020 USA Election

Source: Elaborated by the author.

performed the worst across all validation techniques for the 2020 USA Election dataset.

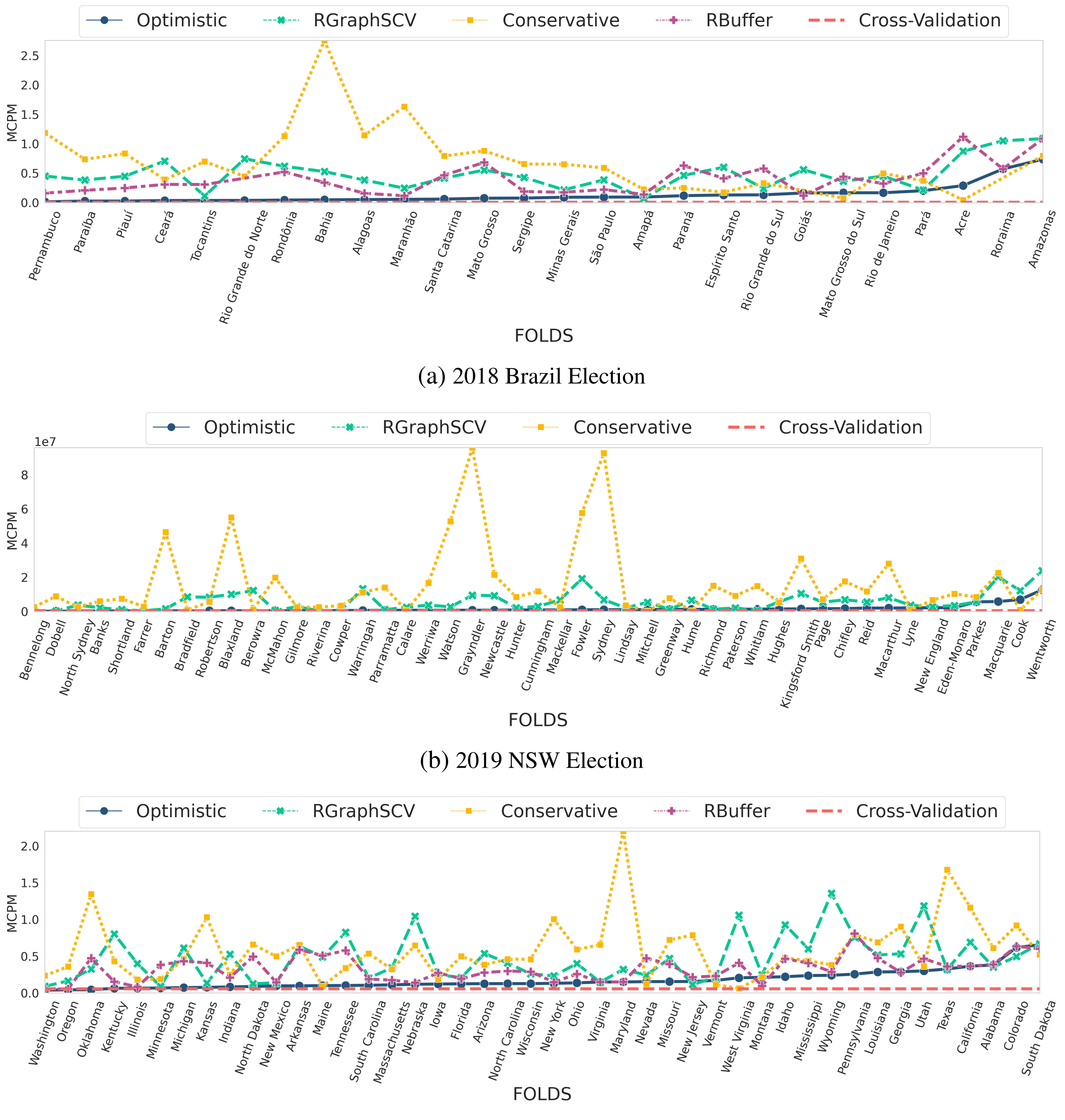
Table 18 – Machine learning models that achieved the lowest and the highest MCPM values regarding each validation approach for all the datasets.

	2018 Brazil Presidential Election				2019 NSW Congress Election				2020 USA Presidential Election			
	Lowest		Highest		Lowest		Highest		Lowest		Highest	
Approaches	ML Method	MCPM	ML Method	MCPM	ML Method	MCPM	ML Method	MCPM	ML Method	MCPM	ML Method	MCPM
Optimistic	SVR	0.0476	DT	0.1489	GBDT	0.1256	KNN	0.1922	SVR	0.0569	LASSO	0.2419
RGraphSCV	MLP	0.0900	ElasticNet	0.3311	RF	0.2591	KNN	0.3640	SVR	0.0836	LASSO	0.3505
TraditionalSCV	MLP	0.1787	DT	0.6362	SVR	0.2510	DT	0.5633	RF	0.0274	LASSO	0.2340
Cross-Validation	GBDT	0.0066	DT	0.0315	GBDT	0.0274	ElasticNet	0.1077	GBDT	0.0119	LASSO	0.1897
RBuffer	MLP	0.0791	ElasticNet	0.3673	-	-	-	-	RF	0.0204	LASSO	0.2198

Source: Research data.

In order to comprehensively understand the performance of each validation technique across spatial folds, Figure 27a, Figure 27b and Figure 27c present the MCPM outcomes per fold for the GBDT algorithm that effectively modeled the spatial dependence structure in all the datasets. Additionally, to explicitly identify scenarios where spatial autocorrelation may bias the results, we used only latitude and longitude features to predict the vote-shares. Finally, we arranged the folds according to the Optimistic approach to facilitate comparison and added a red dashed line representing the mean results obtained from CV. Thus, in general, our proposed approach exhibited more stable behavior than the Conservative approach and displayed similar behavior to the RBuffer approache. Furthermore, our approach was able to maintain a considerable distance from the Optimistic and CV outcomes across most spatial folds.

Figure 27 – MCPM statistics generated using the machine learning method GBDT for each fold of the datasets used in this study regarding each validation approach.



(a) 2018 Brazil Election

(b) 2019 NSW Election

(c) 2020 USA Election

Source: Elaborated by the author.

For a more comprehensive evaluation of the spatial dependence bias in the validation approaches, Table 19 displays the MCPM outcomes and training set size obtained by each approach for the spatial folds where GBDT produced the lowest results in the Optimistic approach.

Specifically, we focus on the spatial folds more likely to exhibit spatial dependence regarding the training set. For the 2018 Brazilian Election dataset, the fold selected was *Pernambuco*. In this fold all SCV approaches demonstrated MCPM outcomes higher than the CV mean. Moreover, the training set sizes for our proposed approach and the RBuffer were comparable. In the case of the 2019 NSW Election dataset, the chosen spatial fold was Bennelong. In this fold only our approach and the Conservative produced MCPM results above the CV mean. However, the MCPM metric for the Conservative approach was considerably high and the training set size was significantly small. Finally, the spatial fold chosen for the 2020 USA Election dataset was Washington. Here, the Optimistic and RBuffer approaches yielded MCPM outcomes lower than the CV mean, while only our proposed approach and the Conservative approach obtained higher results. However, the MCPM outcomes obtained by the Conservative approach were significantly higher than the other approaches.

Table 19 – MCPM and training set size for each of the validation approaches regarding the fold with the lowest MCPM generated by the GBDT subjected to the Optimistic approach.

	2018 Brazil Presidential Election			2019 NSW Congress Election			2020 USA Presidential Election		
Approaches	Fold	MCPM	#Training set	Fold	MCPM	#Training set	Fold	MCPM	#Training set
Optimistic	Pernanbuco	0.0131	5380	Bennelong	0.3073	2714	Washington	0.0372	3072
RGraphSCV	Pernanbuco	0.4500	3441	Bennelong	0.8287	1487	Washington	0.0903	2911
Conservative	Pernanbuco	1.1837	2385	Bennelong	5.0293	802	Washington	0.2360	2420
RBuffer	Pernanbuco	0.1604	3447	-	-	-	Washington	0.0476	3066
Cross-Validation	Mean	0.0071	5009	Mean	0.5374	2486	Mean	0.0573	2800

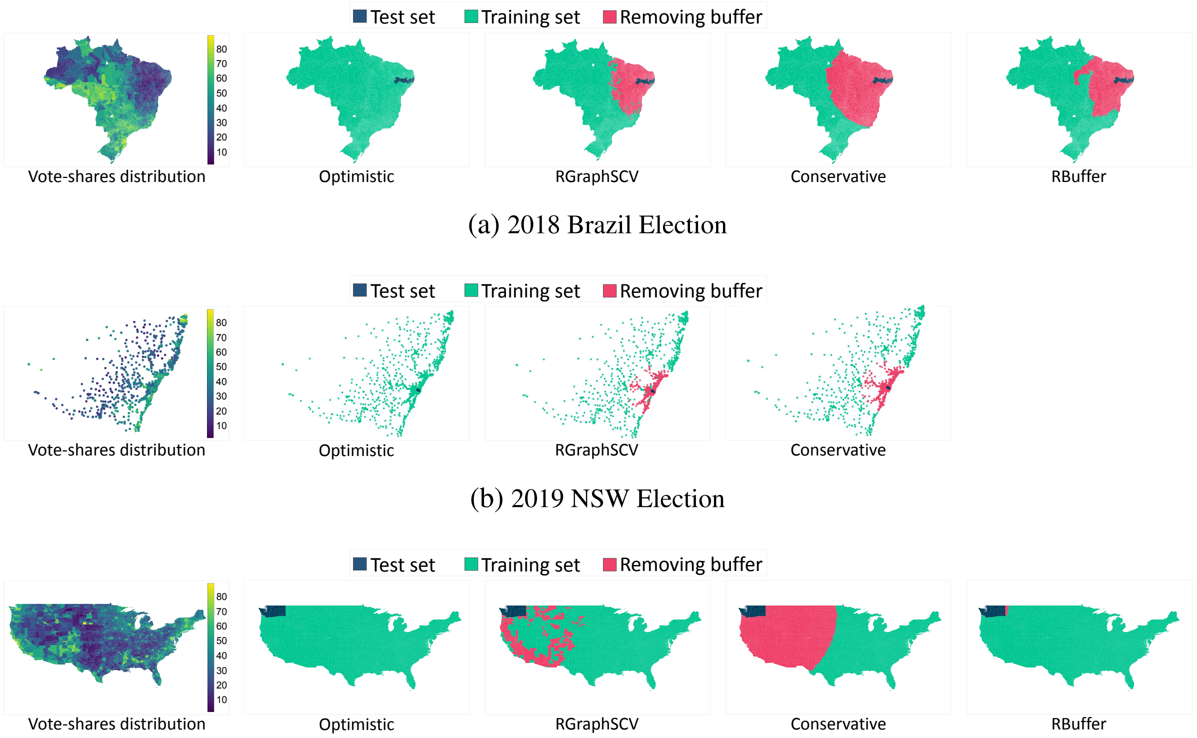
Source: Research data.

Finally, to provide a qualitative evaluation of the SCV approaches considered in this work, Figure 28 shows the spatial distribution of vote-shares and the test, training, and removing buffer sets for the folds listed in Table 19 regarding each SCV approach. Concerning the 2018 Brazil election dataset, Figure 28a shows that our approach produced a non-contiguous removing buffer that is more similar to the spatial distribution where the test set is part of, while the Conservative and RBuffer have extended their removing buffer beyond this distribution. In the 2019 NSW election dataset (Figure 28b), our approach identified a removing buffer that only includes the similar distribution near the test set, while the Conservative removed data beyond this distribution. Finally, for the 2020 USA dataset (Figure 28c), the RBuffer could not identify an adequate removing buffer, while the Conservative removed half of the dataset due to high spatial autocorrelation. Our approach, however, was able to define a non-contiguous removing buffer by removing data similar to the test set.

5.6 Discussion

The SCV technique is an essential machine learning validation process that accounts for spatial dependence in the data. Research has highlighted the necessity of SCV to avoid erroneous interpretations and over-optimistic results caused by spatial autocorrelation (PLOTON;

Figure 28 – Vote-shares spatial distribution and test, training and removing buffer set for each of the validation approaches regarding the fold with the lowest MCPM generated by the GBDT subjected to the Optimistic approach.



(a) 2018 Brazil Election

(b) 2019 NSW Election

(c) 2020 USA Election

Source: Elaborated by the author.

MORTIER *et al.*, 2020; VALAVI *et al.*, 2019; ROBERTS *et al.*, 2017). However, despite its potential, SCV remains underutilized in electoral behavior modeling as a tool to provide models that are able to learn relationships beyond the spatial dependence structure in the data (SILVA; PARMEZAN; BATISTA, 2021; SILVA; PARMEZAN; BATISTA, 2022). This study addresses this gap by proposing an SCV approach for the fair evaluation of machine learning models learned from thousands of spatial features in the context of electoral behavior. We evaluated our approach following a strict experimental protocol and a multi-criteria analysis.

A first assessment of the results in Table 17 indicates that the CV approach tends to yield over-optimistic outcomes, particularly concerning the MSE metric, which can be attributed to the strong spatial dependence between the training and test sets in the datasets. This assertion is consistent with prior researches (PLOTON; MORTIER *et al.*, 2020; ROBERTS *et al.*, 2017). Conversely, the conservative approach produces exceedingly high results, generating a pessimistic scenario due to the limited size of the remaining training set. This issue was first highlighted in (WADOUX *et al.*, 2021). Thus, a good SCV approach should strive to achieve a balance between the test and training sets spatial dependence and the size of the training set to generate fairly and realistic results as pointed by (SILVA; PARMEZAN; BATISTA, 2021).

Based on Figure 26, our proposed SCV approach was demonstrated to be suitable for evaluating machine learning methods in all datasets, as it did not generate overly optimistic or pessimistic results. Similar results were also observed when applied different SCV approaches in other domains (MILÀ *et al.*, 2022; VALAVI *et al.*, 2019; PLOTON; MORTIER *et al.*, 2020). Our approach produced similar results to the RBuffer approach in the 2018 Brazilian Election dataset, even though the RBuffer was specifically designed for this type of dataset and required parameter tuning, while our approach is generalist regarding the datasets to be used and parameter-free (SILVA; PARMEZAN; BATISTA, 2021). Furthermore, in the 2019 NSW Election dataset, our approach generated median results between the optimistic and conservative scenarios for most of the machine learning methods, except for KNN, which is known to be sensitive to spatial dependence and can benefit from the data's spatial structure (SCHRATZ *et al.*, 2019). Thus, generating a pessimistic scenario for this method could be considered positive, since we want to avoid methods that exploit the spatial dependence structure of the data. Finally, for the 2020 USA election dataset, our approach generated better results than the RBuffer, which may be attributed to the need for fine-tuning parameters for the RBuffer approach.

Regarding machine learning algorithm selection, our findings from Table 19 indicate that the GBDT algorithm outperformed other machine learning methods in all datasets when using the CV approach. This result is consistent with previous studies that have suggested that CV is biased towards tree-based methods when the data contains spatial dependence (SCHRATZ *et al.*, 2019; PLOTON; MORTIER *et al.*, 2020). However, when using SCV approaches, more complex models like MLP and SVR tended to perform better. In contrast, DT and KNN, which typically leverage spatial dependence in the data, performed poorly. One exception was observed in the 2019 NSW Election dataset, where GBDT was the best method under the Optimistic approach, indicating the influence of spatial dependence. In contrast, our approach selected MLP, SVR, and RF as the best methods. Although RF was the best method in the 2019 NSW election dataset, the results obtained were still higher than those obtained with the Optimistic and CV approaches, as shown in Figure 26.

Concerning the performance of our approach across the spatial folds, Figure 27 revealed that our approach produced more stable results throughout the spatial folds in contrast to the Conservative approach. The Conservative approach's unstable behavior can be associated with the small size of the training set in some of the spatial folds, leading to a higher risk of overfitting and high error rates. For instance, Table 19 reports that the Conservative approach left only 802 examples in training set for the Bennelong spatial fold in the 2019 NSW Election dataset. This phenomenon has been documented previously in the literature, with authors describing the pessimistic scenarios generated by SCV approaches (WADOUX *et al.*, 2021). By contrast, our approach produced more stable outcomes by reducing the sufficiency of the training set to the extent that spatial dependence was diminished, thereby avoiding overly optimistic results.

When comparing the buffer regions generated by different approaches, our method generates non-contiguous buffer regions, which better reflect the real-world spatial dependence structure that may not occur continuously or have a clear boundary. This behavior is supported by existing research that analyses the spatial dependence of vote-shares in different elections (JACINTHO *et al.*, 2020; FOREST, 2018). Thus, the results from Figure 28a, reveal that our approach generated a buffer region similar to the actual spatial dependence structure in the Northeast of Brazil, where the test fold is located. In contrast, the Conservative and RBuffer approaches produced larger buffer regions that extended beyond the Northeast. In the case of the 2019 NSW Election dataset (Figure 28b), our approach produced a more realistic buffer region with a smaller training data removal than the Conservative approach, yet still achieved higher results than the Optimistic and CV approaches (as shown in Table 19). Similarly, for the 2020 Election dataset (Figure 28c), our approach produced the best buffer region among the approaches, while the Conservative approach removed half of the dataset, and the RBuffer approach failed to generate a buffer region that resulted in performance similar to the Optimistic approach (as seen in Table 19), indicating the influence of spatial dependence.

In conclusion, our proposed approach was demonstrated to be well-suited for comparing machine learning methods in the context of spatially contextualized modeling of electoral behavior. Its parameterless nature and ability to generate more realistic removing buffers allow for less biased results and the discovery of new relationships beyond the spatial dependence structure of the data for different electoral datasets. Furthermore, although proposed for electoral behavior modeling, this approach can be applied to other domains with similar characteristics and objectives. In summary, it contributes to filling an important gap in electoral behavior analysis by bridging the gap between machine learning and electoral behavior modeling. Thus, it represents a new trend in understand of election results, leveraging exploratory analysis to identify new relationships and patterns regarding electoral behavior.

5.7 Limitations, Recommendations, and Practical Implications of the Outcomes

The area of machine learning applied to electoral behavior modeling is still recent, presenting complex and challenging scenarios to be explored. In light of this, this study does not explore our approach's role in identifying new relationships beyond spatial dependence in the spatially contextualized modeling of electoral behavior. Our findings demonstrate our approach's suitability for such a task. Moreover, the results obtained in this study are only relevant to the regression task. Finally, although our approach produced more realistic scenarios for most of the spatial folds when compared to existing approaches, it remains vulnerable to generating inadequate buffer regions in complex scenarios, such as those with highly heterogeneous or small test fold distributions.

In light of this, we highly recommend that researchers carefully inspect the removing buffer regions generated by the SCV approaches. This step is crucial for ensuring the results' reliability, as the removing buffer can significantly impact the performance of machine learning methods. Additionally, we emphasize the importance of ensuring a significant size for each spatial fold and a minimal level of homogeneity in the distribution of data within the spatial folds, despite the freedom to define these boundaries according to pre-existing geographical delimitations.

Finally, the demonstrated effectiveness of our approach suggests that it may be generalized to other domains, including but not limited to species suitability prediction (MILÀ *et al.*, 2022) and forest AGB predictions (PLOTON; MORTIER *et al.*, 2020). However, the application of the approach to these domains may require some modifications to ensure better performance.

5.8 Conclusion

This chapter proposes a novel Spatial Cross-Validation (SCV) technique to evaluate machine learning models that utilize thousands of spatial features for the spatially contextualized modeling of electoral behavior. The proposed technique accounts for pre-existing geographical limitations to be used as spatial folds, as well as non-contiguous spatial dependence. The technique defines a removing buffer region to isolate the test from the training set, which is determined by modeling the test and training sets as a bipartite graph structure and formalizing the problem as a one-class transductive classification task. A novel label propagation method, integrating the semivariogram technique, is proposed to classify nodes from the training set as part of the removing buffer. Our approach is evaluated using three study cases related to recent presidential and congressional elections. The results show that the proposed approach yields more realistic and less biased results allowing the discovery of new relationships beyond the spatial dependence structure of the data.

Future works for our research will focus on evaluating the efficacy of our proposed approach in uncovering new and significant relationships. Additionally, we plan to explore the potential of our approach in the classification task and devise adaptations accordingly. Finally, further investigation into the application of our approach across different domains will also be conducted.

CHAPTER 6

CONCLUSION

The study of electoral behavior is critical to understanding the underlying factors that shape decision-making among the electorate, offering valuable insights into political phenomena such as polarization and the impact of demographic and socioeconomic contexts. However, the election process is complex, and numerous factors can influence the results, ranging from global economic trends to local structural issues. Such complexity needs to be understood and we can use tools such as publicly available data, which is high-dimensional, such as census data. Such data can provide detailed and comprehensive information on the population from local and global perspectives. However, regression analysis, the prevailing methodology in electoral behavior modeling, is inadequate for exploratory analyses of high-dimensional data. On the other side, recent efforts to model electoral behavior using machine learning methods that can better deal with high-dimensional data have largely disregarded the spatial aspects of the data, leading to over-optimistic results and models biased towards the spatial dependence structure, which merely explains elections through the "Neighborhood Effect." Consequently, this thesis presents adaptation to the traditional machine learning pipeline that addresses these issues by providing approaches for the fair assessment of the machine learning models learned from spatial data for the spatially contextualized electoral behavior modeling. Moreover, we propose a machine learning approach to learn relationships beyond the spatial dependence structure of the data, expanding the possibilities of identifying new factors influencing election results. Thus, this thesis contributions can be categorized into three major steps of the machine learning pipeline, the data, the modeling and the validation.

One of the foremost contributions of this thesis is the development of tools for data acquisistion, pre-processing, and geocoding. Our approach facilitates the generation of datasets that span all geographical levels of aggregation available, with added measures of data integrity that aid in determining the reliability of the geocoded information. Notably, the Brazilian census and electoral data form a comprehensive collection that we can provide, with datasets available for different geographical levels of aggregation, election candidate, election year, region of interest, and census category. We expect this tools to assist researchers by facilitating the acquisition of different pre-processed datasets for specific analysis.

In the model assessment step, we propose and develop SCV techniques that take into account the spatial structure of the data to define test-dependent buffer regions and allow pre-existing geographical boundaries to be defined spatial folds. The guidelines for assessing machine learning methods are also introduced, including individual performance measures such as Mean Squared Error (MSE), Mean Error Standard Deviation (MESD), Spearman correlation (SP), and Spearman correlation Standard Deviation (SPSD). A Multi-Criteria Performance Measure (MCPM) is also proposed to guide the choice of adequate models by combining the four metrics mentioned above. Furthermore, the Shapley Additive exPlanation (SHAP) Values technique is used to analyze the results generated and capture new relationships. The effectiveness of the proposed SCV approaches is evaluated using three real datasets related to the 2018 Brazil Presidential Election, the 2019 New South Whales Congress Election, and the 2020 United States Presidential Election. The results show that the approaches contribute to scenarios where the spatial spatial presents less influence on the results, which enhances the chances of models that learn relationships beyond the data spatial dependence structure.

Moreover, in this research we present a novel contribution to spatially contextualized electoral behavior modeling. Specifically, we propose and develop a stacking-based machine learning approach that operates on spatial contexts of different dimensions at two levels. Firstly, it captures local patterns from spatial contexts, and then, at the meta-level, it globally captures information from the K nearest contexts to a region of interest. Our approach demonstrates significant improvements in performance compared to baseline and reference models, as evaluated using data from the second round of the 2018 Brazilian presidential elections. Additionally, our model's predictions are intelligible and coherent in challenging regions, underscoring its interpretability. Thus, our work provides a valuable contribution to the field, as it presents an effective and interpretable approach to spatially contextualized electoral behavior modeling using machine learning methods.

The outcomes obtained from this study are significant as they broaden the scope of potential applications for modeling electoral behavior. In addition, our research can serve as a foundational step in recognizing novel patterns and relationships that may be subject to further scrutiny via conventional electoral behavior techniques, such as regression analysis. Therefore, our research advances the current status of electoral behavior modeling, providing powerful tools to expand the identification of factors influencing election outcomes.

6.1 Limitations

Although the main limitations regarding the application of machine learning to spatial data for spatially contextualized modeling of electoral behavior were dealt with in this research, which regards the consideration of spatial dependence in the validation and modeling steps, other limitations are imposed in the approaches proposed in this research.

In the present study, the SCV approaches proposed were evaluated solely with regard to the regression task. It is noteworthy that the semivariogram technique employed as the basis of our approaches only accounts for continuous variables, thereby limiting the approach's applicability to other tasks such as classification. Hence, adaptations to the proposed approaches should be made to enable their use in other types of tasks. Although our approach produced realistic scenarios with less spatial dependence in most of the spatial folds as compared to existing approaches, it may still generate inadequate buffer regions in complex scenarios such as those with highly heterogeneous or small test fold distributions. However, it is imperative to note that such scenarios should not be considered since the definition of spatial boundaries for the spatial folds should be based on prior information regarding data distribution homogeneity. Additionally, smaller-sized spatial folds provide less information on their distribution, and we therefore recommend employing LOO SCV approaches to handle such situations more effectively.

In regards to the proposed machine learning approach for modeling electoral behavior, it is important to note some limitations. Firstly, our approach does not address ambiguous distributions, which is a common aspect in modeling electoral behavior. Secondly, we only evaluate our approach on a single dataset, and further studies with different election databases may prove beneficial. Another limitation is the parameter K, which corresponds to the number of neighboring contexts considered in the modeling process. While we use the average of neighbors in each context present, there is still potential for optimizing this parameter. Finally, it is crucial to carefully investigate the relationships and patterns identified by our approach to ensure the effectiveness of our approach in discovering new relationships and patterns for the spatially contextualized electoral behavior modeling task.

Finally, it is important to acknowledge a limitation inherent to the problem of machine learning applied to spatially contextualized modeling of electoral behavior. Specifically, there is an assumption that the relationships modeled by machine learning methods may constitute significant or real relationships. In many cases, however, machine learning methods may model feature relationships based solely on the performance metric, leading to spurious correlations favoring good performance. Thus, the machine learning perspective on electoral behavior modeling should be exploratory in nature, serving as an initial step to further analysis rather than a definitive answer.

6.2 Future Works

In addition to the limitations highlighted in the previous section and the approaches proposed in this thesis, there is a need to explore potential future research avenues. We identified three main areas of investigation, which include the development of novel machine learning techniques using state-of-the-art methods, the incorporation of explainable AI in machine learning methods learned from spatial data for the purpose of spatially contextualized electoral behavior modeling, and the evaluation of our pipeline's ability to uncover relationships and patterns beyond spatial dependence. These potential avenues of research can contribute to a better understanding of electoral behavior modeling and pave the way for more accurate and interpretable models.

Proposing and developing machine learning approaches that incorporate state-of-the-art methods for modeling electoral behavior is pertinent. Among these methods, Graph Neural Networks (GNNs) have gained considerable popularity in the spatial modeling community for their ability to model local and global relationships from high-dimensional data without requiring feature selection. As such, incorporating GNNs in electoral behavior modeling can offer significant advantages and benefits. Thus, further research in this direction can be undertaken to explore the potential of GNNs in modeling spatially contextualized electoral behavior.

Another field of research that demands attention is the explainable AI for models learned from spatial data in the context of electoral behavior modeling. It is necessary to propose novel approaches to understand the model's results better, considering both local and global relationships. In addition, it is crucial to consider the spatial aspect since features may have different importance across space, and this information can provide valuable insights into the model's decision-making process. By developing explainable AI approaches for spatially contextualized electoral behavior modeling, we can increase transparency and trust in the models' predictions and facilitate their interpretation by domain experts.

Finally, an immediate area of interest pertains to the evaluation of the machine learning pipeline's capacity to identify novel relationships and patterns beyond spatial dependence. To this end, our proposed approaches will be leveraged in conjunction with the pipeline to discern hitherto unknown relationships. A crucial step in this regard would be to obtain domain experts' assistance to determine if the identified relationships are substantive and contribute to meaningful discussion and understanding of electoral results. Such an endeavor holds significant promise in expanding the scope of research and providing valuable insights into the phenomenon of electoral behavior.

6.3 Thesis Related Publications

This thesis encompasses a comprehensive research effort within the domain of electoral voting behavior modeling using machine learning, resulting in four publications, listed bellow. These publications cover a wide range of topics, including spatial and temporal analysis of vote-shares, the development of novel evaluation pipelines to ensure fair model assessment, and the introduction of a new machine learning method capable of capturing both local and global relationships and patterns.

It is important to note that two of the publications involved collaboration with an undergraduate researcher, as indicated by the citations Jacintho *et al.* (2021) and Jacintho *et al.* (2020). However, it should be emphasized that the intellectual contributions for the research lie solely with the author of this thesis.

- Geographic Context-Based Stacking Learning for Election Prediction from Socio-economic Data, BRACIS, Nov 2022, Tiago P. da Silva, Antonio R.S. Parmezan, Gustavo E.A.P.A Batista
  DOI: 10.1109/10.1007/978-3-031-21686-2_44

- A Graph-Based Spatial Cross-Validation Approach for Assessing Models Learned with Selected Features to Understand Election Results, ICMLA, Dec 2021, Tiago P. da Silva, Antonio R.S. Parmezan, Gustavo E.A.P.A Batista
  DOI: 10.1109/ICMLA52953.2021.00150

- Analyzing spatio-temporal voting patterns in Brazilian elections through a simple data science pipeline, JIDM, Aug 2021 Lucas H.M. Jacintho, Tiago P. da Silva, Antonio R.S. Parmezan, Gustavo E.A.P.A Batista
  DOI: 10.5753/jidm.2021.1932

- Brazilian Presidential Elections: Analysing Voting Patterns in Time and Space Using a Simple Data Science Pipeline, KDMILE, Oct 2020 Lucas H.M. Jacintho, Tiago P. da Silva, Antonio R.S. Parmezan, Gustavo E.A.P.A Batista
  DOI: 10.5753/kdmile.2020.11979

6.4 Other Publications

During the tenure as a Ph.D. student, the author of this thesis made significant contributions to various research projects, assuming the role of either a contributor or first author. These projects encompassed diverse domains, including but not limited to data streams and recommendation systems. As a result, several publications were produced as listed bellow, showcasing the author's expertise and involvement in multidisciplinary research endeavors beyond the scope of this thesis.

- A Fuzzy Approach for Classification and Novelty Detection in Data Streams Under Intermediate Latency, BRACIS, Oct 2020 Andre L. Cristiani, Tiago P. da Silva, Heloisa A. Camargo
  DOI: 10.1007/978-3-030-61380-8_12

- Possibilistic Approach For Novelty Detection In Data Streams, FUZZIEEE 2020, Jul 2020 Tiago P. da Silva, Heloisa A. Camargo
  DOI: 10.1109/FUZZ48607.2020.9177582

- A Fuzzy Classifier for Data Streams with Infinitely Delayed Labels, CIARP 2019, Nov 2019 Tiago P. da Silva, Vinicius M. de Souza, Heloisa A. Camargo, Gustavo E.A.P.A Batista
  DOI: 10.1007/978-3-030-13469-3_34

- Evaluating Vector Representations from User's Reviews in a Recommendation Task, ENIAC 2019, Oct 2019 Vitor R. Tonon, Tiago P. da Silva, Vinícius Ferreira, Gean T. Pereira, Solange O. Rezende
  <https://sol.sbc.org.br/index.php/eniac/article/view/9291/9193>

- Evaluating stream classifiers with delayed labels information, BRACIS 2018, Oct 2018 Vinicius M. de Souza, Tiago P. da Silva, Gustavo E.A.P.A. Batista
  DOI: 10.1109/BRACIS.2018.00077

BIBLIOGRAPHY

AGNEW, J. Maps and models in political studies: a reply to comments. **Political Geography**, Pergamon, v. 15, n. 2, p. 165–167, 1996. Citations on pages 26, 29, 38, 92, and 94.

ANSELIN, L. Local indicators of spatial association–LISA. **Geographical Analysis**, Wiley Online Library, v. 27, n. 2, p. 93–115, 1995. Citations on pages 46 and 47.

ANSELIN, L.; GETIS, A. Spatial statistical analysis and geographic information systems. **The Annals of Regional Science**, Springer, v. 26, n. 1, p. 19–33, 1992. Citation on page 46.

CALIńSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics**, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974. Available: <https://www.tandfonline.com/doi/abs/10.1080/03610927408827101>. Citation on page 48.

CAMERON, S.; MCALLISTER, I. 2019 australian federal election: results from the australian election study. **Australian Election Study**, Australian Election Study, 2019. Citation on page 104.

CAMPELLO, R. J.; MOULAVI, D.; ZIMEK, A.; SANDER, J. A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies. **Data Mining and Knowledge Discovery**, Springer, v. 27, n. 3, p. 344–371, 2013. Citation on page 47.

CARVALHO, R.; MENEZES, T. Uma análise espacial das eleições presidenciais brasileiras de 2010. **Pesquisa e Planejamento Econômico**, v. 45, n. 3, p. 436–495, 02 2015. Citations on pages 38, 41, and 42.

CHARNEY, I.; MALKINSON, D. Between electoral and urban geography: Voting patterns and socio-spatial dynamics in Tel Aviv. **Appl. Geogr.**, Elsevier, v. 58, p. 1–6, 2015. Citation on page 58.

CHAUHAN, P.; SHARMA, N.; SIKKA, G. The emergence of social media data and sentiment analysis in election prediction. **J. Ambient Intell. Humaniz. Comput.**, Springer, v. 12, n. 2, p. 2601–2627, 2021. Citations on pages 30, 74, and 76.

CLIFF, A.; ORD, K. Testing for spatial autocorrelation among regression residuals. **Geographical Analysis**, Wiley Online Library, v. 4, n. 3, p. 267–284, 1972. Citations on pages 26 and 46.

CORRÊA, D. S. Os custos eleitorais do bolsa família: Reavaliando seu impacto sobre a eleição presidencial de 2006. **Opinião Pública**, SciELO Brasil, v. 21, n. 3, p. 514–534, 2015. Citation on page 42.

CURRAN, P. J. The semivariogram in remote sensing: an introduction. **Remote sensing of Environment**, Elsevier, v. 24, n. 3, p. 493–507, 1988. Citation on page 98.

Davies, D. L.; Bouldin, D. W. A cluster separation measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, PAMI-1, n. 2, p. 224–227, 1979. Citation on page 48.

DEPPNER, J.; CAJIAS, M. Accounting for spatial autocorrelation in algorithm-driven hedonic models: A spatial cross-validation approach. **The Journal of Real Estate Finance and Economics**, Springer, p. 1–39, 2022. Citations on pages 29 and 92.

DIETTERICH, T. G. Ensemble methods in machine learning. In: SPRINGER. **International workshop on multiple classifier systems**. [S.l.], 2000. p. 1–15. Citation on page 76.

ELKINK, J. A.; FARRELL, D. M. **Predicting vote choice in the 2020 Irish general election**. [S.l.]: Taylor & Francis, 2021. 521–534 p. Citations on pages 25, 27, 28, 30, 90, 91, 93, and 94.

FAUSTINO, J.; BARBOSA, H.; RIBEIRO, E.; MENEZES, R. A data-driven network approach for characterization of political parties' ideology dynamics. **Applied Network Science**, Springer, v. 4, n. 1, p. 1–15, 2019. Citations on pages 39 and 58.

FISCHER, M. M. Spatial analysis in geography. In: **Spatial Analysis and GeoComputation: Selected Essays**. [S.l.]: Springer, 2006. p. 17–28. Citation on page 63.

FOREST, B. Electoral geography: From mapping votes to representing power. **Geography Compass**, Wiley Online Library, v. 12, n. 1, p. e12352, 2018. Citations on pages 25, 26, 27, 90, 94, and 115.

FOTHERINGHAM, A. S.; LI, Z.; WOLF, L. J. Scale, context, and heterogeneity: A spatial analytical perspective on the 2016 us presidential election. **Annals of the American Association of Geographers**, Taylor & Francis, v. 111, n. 6, p. 1602–1621, 2021. Citations on pages 74, 75, and 76.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Ann. Stat.**, JSTOR, p. 1189–1232, 2001. Citation on page 65.

GETIS, A. Spatial autocorrelation. In: **Handbook of applied spatial analysis**. [S.l.]: Springer, 2010. p. 255–278. Citations on pages 26, 60, and 90.

GRAEFE, A.; GREEN, K. C.; ARMSTRONG, J. S. Accuracy gains from conservative forecasting: Tests using variations of 19 econometric models to predict 154 elections in 10 countries. **PloS One**, Public Library of Science San Francisco, CA USA, v. 14, n. 1, p. e0209850, 2019. Citations on pages 30, 74, 76, and 78.

HALL, M. A. Correlation-based feature selection for discrete and numeric class machine learning. In: **ICML**. [S.l.: s.n.], 2000. p. 359–366. Citation on page 65.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: Concepts and techniques**. 3. ed. California: Morgan Kaufmann, 2011. Citation on page 38.

HAND, D. J.; ADAMS, N. M. Data mining. **Wiley StatsRef: Statistics Reference Online**, Wiley Online Library, p. 1–7, 2014. Citation on page 38.

HERNáNDEZ, V.; LEóN, L. Geografía de la participación electoral y diferenciación socioespacial en Ciudad Juárez, Chihuahua (México). **Geopolítica(s). Revista de Estudios sobre Espacio y Poder**, v. 11, p. 145–172, 06 2020. Citation on page 39.

JACINTHO, L. H. M.; SILVA, T. P.; PARMEZAN, A. R. S.; BATISTA, G. E. A. P. A. Analysing spatio-temporal voting patterns in brazilian elections through a simple data science pipeline. **Journal of Information and Data Management**, SBC, p. 1–16, 2021. ISSN 2178-7107. Citations on pages 34, 58, and 121.

JACINTHO, L. H. M.; SILVA, T. P. da; PARMEZAN, A. R. S.; 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**. Porto Alegre: SBC, 2020. p. 217–224. ISSN 0000-0000. Available: <https://sol.sbc.org.br/index.php/kdmile/article/view/11979>. Citations on pages 38, 74, 78, 104, 115, and 121.

JIANG, Z.; SAINJU, A. M.; LI, Y.; SHEKHAR, S.; KNIGHT, J. Spatial ensemble learning for heterogeneous geographic data with class ambiguity. **ACM Trans. Intell. Syst. Technol.**, ACM, v. 10, n. 4, p. 1–25, 2019. Citation on page 81.

JR, J. H. W. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, Taylor & Francis, v. 58, n. 301, p. 236–244, 1963. Available: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500845>. Citation on page 47.

KEOGH, E.; MUEEN, A. Curse of dimensionality. In: ____. **Encyclopedia of Machine Learning and Data Mining**. Boston, MA: Springer US, 2017. p. 314–315. ISBN 978-1-4899-7687-1. Citation on page 90.

LAGO, I. A research agenda in elections and voting behavior in a global and changing world. **Frontiers in Political Science**, Frontiers Media SA, v. 1, p. 1, 2019. Citations on pages 25 and 90.

LAYTON, M. L.; SMITH, A. E.; MOSELEY, M. W.; COHEN, M. J. Demographic polarization and the rise of the far right: Brazil's 2018 presidential election. **Res. Politics**, SAGE Publications Sage UK: London, England, v. 8, 2021. Citations on pages 74, 85, and 104.

LEHOUCQ, F. Electoral fraud: Causes, types, and consequences. **Annu. Rev. Polit. Sci.**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 6, n. 1, p. 233–256, 2003. Citations on pages 25 and 58.

LI, H.; CALDER, C. A.; CRESSIE, N. Beyond Moran's I: Testing for spatial dependence based on the spatial autoregressive model. **Geographical Analysis**, Wiley Online Library, v. 39, n. 4, p. 357–375, 2007. Citation on page 46.

LI, M.; PERRIER, E.; XU, C. Deep hierarchical graph convolution for election prediction from geospatial census data. In: **AAAI Conference on Artificial Intelligence**. [S.l.: s.n.], 2019. v. 33, p. 647–654. Citations on pages 26, 27, 28, 30, 74, 76, 78, 82, 90, 91, 93, 94, and 104.

LI, Y.; FANG, Y.; CHENG, R.; ZHANG, W. Spatial pattern matching: a new direction for finding spatial objects. **SIGSPATIAL Special**, ACM, v. 11, n. 1, p. 3–12, 2019. Citations on pages 28, 63, and 99.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. **Adv. Neural. Inf. Process. Syst.**, v. 30, 2017. Citation on page 80.

MAGALHÃES, A. M.; SILVA, M. E. A. d.; DIAS, F. d. M. Eleição de Dilma ou segunda reeleição de Lula? Uma análise espacial do pleito de 2010. **Opinião Pública**, SciELO Brasil, v. 21, n. 3, p. 535–573, 2015. Citations on pages 41 and 42.

MANOEL, L.; COSTA, A. C.; CABRAL, P. Voter turnout in portugal: a geographical perspective. **Papers in Applied Geography**, Taylor & Francis, v. 8, n. 1, p. 88–111, 2022. Citations on pages 27, 90, and 94.

MANSLEY, E.; DEMŠAR, U. Space matters: Geographic variability of electoral turnout determinants in the 2012 london mayoral election. **Electoral Studies**, Elsevier, v. 40, p. 322–334, 2015. Citations on pages 26, 27, 38, 39, and 58.

MARTINS, D. J. D.; MANSANO, F. H.; PARRÉ, J. L.; PLASSA, W. Fatores que contribuíram para a reeleição da presidente Dilma Rousseff. **Política & Sociedade**, v. 15, n. 32, p. 145–170, 2016. Citations on pages 41 and 42.

MARZAGÃO, T. A dimensão geográfica das eleições brasileiras. **Opinião Pública**, SciELO Brasil, v. 19, n. 2, p. 270–290, 2013. Citations on pages 38 and 42.

MILÀ, C.; MATEU, J.; PEBESMA, E.; MEYER, H. Nearest neighbour distance matching leave-one-out cross-validation for map validation. **Methods in Ecology and Evolution**, Wiley Online Library, 2022. Citations on pages 29, 92, 95, 96, 114, and 116.

MOTA, A. M. S. **Modelling abstention rate using spatial regression**. Master's Thesis (Master's Thesis) — NOVA Information Management School, 2019. Citation on page 39.

NORRIS, P. The new research agenda studying electoral integrity. **Elect. Stud.**, Elsevier, v. 32, n. 4, p. 563–575, 2013. Citations on pages 25 and 58.

____. Do perceptions of electoral malpractice undermine democratic satisfaction? the us in comparative perspective. **International Political Science Review**, SAGE Publications Sage UK: London, England, v. 40, n. 1, p. 5–22, 2019. Citation on page 104.

NORRIS, P.; GRÖMPING, M. **Electoral integrity worldwide**. 2019. Sydney: Electoral Integrity Project. Available at <https://www.electoralintegrityproject.com/>. Citation on page 39.

OKUNEV, I. Y.; GORELOVA, J. S.; GRUZDEVA, E. Regional disparities of electoral behaviour in poland: Comparative spatial analysis. **Comparative Politics Russia**, v. 12, n. 1, p. 149–160, 2020. Citation on page 39.

PARMEZAN, A. R. S.; LEE, H. D.; WU, F. C. Metalearning for choosing feature selection algorithms in data mining: Proposal of a new framework. **Expert Syst. Appl.**, Pergamon, Tarrytown, United States of America, v. 75, p. 1–24, 2017. Citations on pages 80 and 106.

PINHEIRO-MACHADO, R.; SCALCO, L. M. From hope to hate: The rise of conservative subjectivity in brazil. **HAU: J. Ethnogr. Theory**, The University of Chicago Press Chicago, IL, v. 10, n. 1, p. 21–31, 2020. Citations on pages 25, 74, 85, 90, and 93.

PLOTON, P.; MORTIER, F. *et al.* Spatial validation reveals poor predictive performance of large-scale ecological mapping models. **Nat. Commun.**, Nature Publishing Group, v. 11, n. 1, p. 1–11, 2020. Citations on pages 28, 29, 58, 60, 74, 75, 91, 92, 95, 96, 113, 114, and 116.

POWER, T. J.; RODRIGUES-SILVEIRA, R. Mapping ideological preferences in Brazilian elections, 1994–2018: a municipal-level study. **Brazilian Political Science Review**, SciELO Brasil, v. 13, n. 1, p. e0001–1–27, 2019. Citation on page 39.

PRACIANO, B. J. G.; COSTA, J. P. C. L. da; MARANHÃO, J. P. A.; MENDONÇA, F. L. L. de; JÚNIOR, R. T. de S.; 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**. Singapore: IEEE, 2018. p. 1355–1360. Citation on page 39.

RECUERO, R.; SOARES, F. B.; 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**, v. 14, n. 1, p. 569–578, May 2020. Available: <https://ojs.aaai.org/index.php/ICWSM/article/view/7324>. Citations on pages 39 and 58.

REID, B.; LIU, G.-J. One nation and the heartland's cleavage: an exploratory spatial data analysis. In: GRANT, B.; MOORE, T.; LYNCH, T. (Ed.). **The Rise of Right-Populism: Pauline Hanson's One Nation and Australian Politics**. Singapore: Springer, 2019. p. 79–102. ISBN 978-981-13-2670-7. Available: <https://doi.org/10.1007/978-981-13-2670-7_5>. Citations on pages 39 and 58.

ROBERTS, D. R.; BAHN, V.; CIUTI, S.; BOYCE, M. S. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. **Ecography**, Wiley Online Library, v. 40, n. 8, p. 913–929, 2017. Citations on pages 28, 29, 58, 60, 61, 62, 91, 92, 95, 96, and 113.

ROKACH, L.; MAIMON, O. Clustering methods. In: MAIMON, O.; ROKACH, L. (Ed.). **Data Mining and Knowledge Discovery Handbook**. Boston: Springer, 2005. p. 321–352. Citation on page 47.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53 – 65, 1987. ISSN 0377-0427. Available: <http://www.sciencedirect.com/science/article/pii/0377042787901257>. Citation on page 48.

SANTOS, B. N. dos; ROSSI, R. G.; REZENDE, S. O.; MARCACINI, R. M. A two-stage regularization framework for heterogeneous event networks. **Pattern Recognition Letters**, Elsevier, v. 138, p. 490–496, 2020. Citation on page 101.

SCHRATZ, P.; MUENCHOW, J.; ITURRITXA, E.; RICHTER, J.; BRENNING, A. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. **Ecological Modelling**, Elsevier, v. 406, p. 109–120, 2019. Citation on page 114.

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**, v. 21, n. 1, p. 156–167, 2018. Citations on pages 41 and 42.

SILVA, T. P. D.; PARMEZAN, A. R.; BATISTA, G. E. Geographic context-based stacking learning for election prediction from socio-economic data. In: SPRINGER. **Brazilian Conference on Intelligent Systems**. [S.l.], 2022. p. 641–656. Citations on pages 35, 90, 91, 92, 93, 94, 95, 107, and 113.

SILVA, T. P. D.; PARMEZAN, A. R. S.; BATISTA, G. E. A. P. A. A graph-based spatial cross-validation approach for assessing models learned with selected features to understand election results. In: IEEE. **International Conference on Machine Learning and Applications**. [S.l.], 2021. p. 909–915. Citations on pages 34, 74, 75, 80, 92, 94, 95, 96, 98, 105, 113, and 114.

STOKES, D. E. Spatial models of party competition. **American political science review**, Cambridge University Press, v. 57, n. 2, p. 368–377, 1963. Citation on page 93.

TERRON, S. L.; SOARES, G. A. D. As bases eleitorais de lula e do pt: do distanciamento ao divórcio. **Opinião Pública**, SciELO Brasil, v. 16, n. 2, p. 310–337, 2010. Citations on pages 26 and 42.

TOBLER, W. R. A computer movie simulating urban growth in the detroit region. **Economic Geography**, Taylor & Francis, v. 46, n. sup1, p. 234–240, 1970. Citations on pages 26, 46, 60, and 77.

VALAVI, R.; ELITH, J.; LAHOZ-MONFORT, J. J.; GUILLERA-ARROITA, G. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. **Methods Ecol. Evol.**, v. 10, n. 2, p. 225–232, 2019. Citations on pages 29, 58, 60, 62, 92, 95, 96, 113, and 114.

WADOUX, A. M.-C.; HEUVELINK, G. B.; BRUIN, S. D.; BRUS, D. J. Spatial cross-validation is not the right way to evaluate map accuracy. **Ecological Modelling**, Elsevier, v. 457, p. 109692, 2021. Citations on pages 113 and 114.

WONG, M. Y.; WONG, S. H.-W. Income inequality and political participation: A district-level analysis of hong kong elections. **Social Indicators Research**, Springer, p. 1–19, 2022. Citations on pages 27, 90, and 94.

YERO, E. J. H.; SACCO, N. C.; NICOLETTI, M. do C. Effect of the municipal human development index on the results of the 2018 brazilian presidential elections. **Expert Systems with Applications**, Elsevier, v. 168, p. 114305, 2021. Citations on pages 27, 28, 30, 91, 94, and 95.

ZHU, X.; GHAHRAMANI, Z.; LAFFERTY, J. Semi-supervised learning using gaussian fields and harmonic functions. In: **Proceedings of the 20th International Conference on Machine Learning**. [S.l.]: AAAI Press, 2003. p. 912–919. Citation on page 101.

ZUCCO, C.; POWER, T. Fragmentation without cleavages? Endogenous fractionalization in the Brazilian party system. **Comparative Politics**, City University of New York, 01 2020. Citations on pages 39, 45, and 46.

ICMC USP

SÃO CARLOS