Regressão Espacial aplicado ao estudo eleitoral:

Caso do partido NOVO nas eleições de 2018

Afonso Henrique Barros Machado

Áquila Estevão da Silva Campos

Felipe Gontijo Fonseca

Marcos Antonio Alves Bezerra Júnior

Coordenador

Eduardo de Rezende Francisco

FGV Management

RESUMO

Em um contexto cada vez mais dominado pelo debate político nas redes sociais, qual a relevância das variáveis socioeconômicas demográficas como determinantes do voto nas eleições presidenciais? A variável geográfica é interessante no momento de traçar estratégias políticas de um partido? Procurando responder a essas perguntas, este artigo utilizou-se de dados de variáveis socioeconômicas, eleitorais e espaciais em nível municipal. O partido utilizado para a análise foi o partido NOVO, que disputou sua primeira eleição, o que torna relevante avaliar se, dados os objetivos do partido, a captação de votos possuiu características de eleitores de um grupo mais segmentado. Realizou-se uma análise exploratória dos dados, incluindo a distribuição espacial da quantidade de votos atribuída ao partido NOVO na eleição presidencial de 2018, buscando traçar o perfil dos munícipios onde o partido teve melhor desempenho. Posteriormente, utilizou-se técnicas geográficas exploratórias e modelos de Estatística Espacial, como o *Spatial Lag*, *Spatial Auto-Regressive Model* e o *GWR (Geographically Weighted Regression)*. Conclui-se que o aspecto geográfico possui grande relevância na explicação do desempenho político do partido e que os modelos estatísticos espaciais melhoram o desempenho preditivo do modelo.

INTRODUÇÃO

Antes de darmos início às abordagens específicas deste estudo, façamos uma revisitação em contexto histórico para melhor compreender a evolução do tema geografia eleitoral, afim de despertar um maior interesse ao arcabouço analítico espacial.

O início histórico do que conhecemos hoje como geografia eleitoral, remete ao ano de 1913, ano próximo a implosão da primeira grande guerra mundial (1914), quando representações cartográficas e estatísticas foram utilizadas em conjunto por André Siegfried, na França, mais especificamente para avaliar e compreender fenômenos políticos e sociais que guiavam as orientações de votos com a relação causal ao solo no Oeste da França na Terceira República. A hipótese de Siegfried, era de que

existia correlação direta entre as estruturas sociais pré-existentes no local e o comportamento dos eleitores, tal sugestão foi abordado em sua obra republicada em 1995, intitulada *Tableau Politique de la France de I'Ouest sous la IIIe République* (Siegfried, 1995).

Em 1949, o efeito de vizinhança (*friends and neighbors effect*) foi descrito por V. O. Key (Key, 1949), afim de verificar a preferência de candidatos locais em sacrifício aos demais e acabou ganhando espaço influenciando outros estudos futuros relacionados aos temas de dinâmica das eleições, pesquisas de geografia regional entre outros, e que até os tempos de hoje, o efeito de vizinhança está presente como objeto de pesquisa. Mesmo após 30 anos, sua obra ainda é bastante influente no assunto das ciências políticas estando presente como requisito obrigatório para qualquer estudante de política americana.

A avaliação de aspectos socioeconômicos como influenciadores na escolha eleitoral foi longamente abordada na década de 1960 no estudo *The American Voter* (Campbell, 1960), onde temas como características pessoais, econômicas e religião, investigadas sob o aspecto das teorias de vizinhança e de influência local e regional, ganharam bastante força.

Na década de 1970, Kevin Cox trouxe para a discussão do tema da geografia eleitoral um estudo em que abordava temáticas como a integração dos efeitos das instituições sociais, contextuais e ideológicas, integrando a distância geográfica, o círculo de convivência e a reciprocidade como temas de avaliação e objetos de investigação sobre a escolha eleitoral (Terron, 2012).

Com o grande avanço e evolução da análise espacial e tecnológica, dispomos atualmente de um conjunto de ferramentas bastante útil para elaboração de avaliações, investigações e pesquisas no campo do comportamento político-eleitoral, dando maior importância do "onde" ocorre determinado fenômeno e qual sua influência e peso no aspecto decisório.

Vimos então, que a geografia eleitoral é um campo com mais de cem anos de existência, porém com um número relativamente pequeno de publicações quando comparado a disciplinas de economia e

ciência política. No Brasil, onde até 2008 era exíguo o número de publicações relacionadas à geografia das eleições subsequentes ao regime militar (O'Loughlin, 2003 e Terron, 2012), o interesse por esse campo de conhecimento vem crescendo.

Os avanços tecnológicos, que possibilitam cada vez mais o surgimento de novas técnicas de análise espacial, aliados a necessidade de conhecer melhor o impacto do espaço brasileiro na escolha do candidato motiva a realização deste estudo.

Singer (1999) identifica a identidade ideológica entre eleitor e candidato como um fator central na decisão do voto e Carreirão (2002) confirma a identificação da identidade ideológica, porém adiciona a escolaridade do eleitor como fator que determina o impacto da identidade ideológica. Este trabalho usa a hipótese da identidade ideológica para escolher as variáveis de estudo e adiciona o fator geográfico para tentar aumentar a eficácia da análise.

Foi escolhido como alvo da análise proposta para este estudo a eleição presidencial do partido NOVO em 2018. Esta escolha foi feita pois o partido NOVO apresenta ideologia bem definida em seu estatuto e os valores do partido foram bem difundidos na sua primeira eleição.

OBJETIVO

Estudos de ciência política ressaltam que algumas variáveis eleitorais possuem significância em relação ao desempenho eleitoral. Por exemplo, Pereira & Rennó (2007), na previsão de determinantes da reeleição para câmara de deputados, encontraram significância em variáveis eleitorais, tais como: percentual de emendas executadas, base eleitoral, votação da eleição anterior, concentração eleitoral, presença de mesa diretora, relator de comissão e projetos aprovados. O partido NOVO, em razão de seu pouco tempo de fundação, não possui ainda variáveis eleitorais consolidadas. Espera-se que esse fato aumente o valor explicativo das variáveis socioeconômicas escolhidas para fazer parte desse estudo, bem como a significância da variável espacial no desempenho da eleição presidencial do partido NOVO.

As evidências empíricas demonstram que a identificação partidária é fator de peso nas escolhas dos eleitores, embora ainda não haja consenso sobre a forma mais adequada de mensuração do vínculo entre eleitores e partidos (Braga e Pimentel Jr., 2011; Cabello e Rennó, 2010; Carreirão e Barbetta, 2004; Peixoto e Rennó, 2011; Rennó, 2007.)

Nesse sentido, seria possível identificar que características socioeconômicas levam os municípios do Brasil a se identificarem com o partido NOVO? Qual a importância da variável geográfica nesse contexto? Neste estudo, utilizou-se a variável percentual de votos do partido NOVO para eleição presidencial de 2018 em relação a população do município para medir o desempenho da agremiação eleitoral no município.

Espera-se encontrar um padrão socioeconômico, bem como geográfico entre os munícipios que mais depositaram confiança no partido. Desse modo, o objetivo é analisar o poder de explicação do desempenho eleitoral a partir de modelos distintos comparados: (i) regressão clássica, que não considera a variável geográfica, e modelos que incluam a variável geográfica, (ii) em aspecto global (*Spatial Auto-Regressive model*, SAR), e (iii) em abordagem local, que diferencie os parâmetros por meio das características de cada região – clusters (*Geographically Weighted Regression*, GWR).

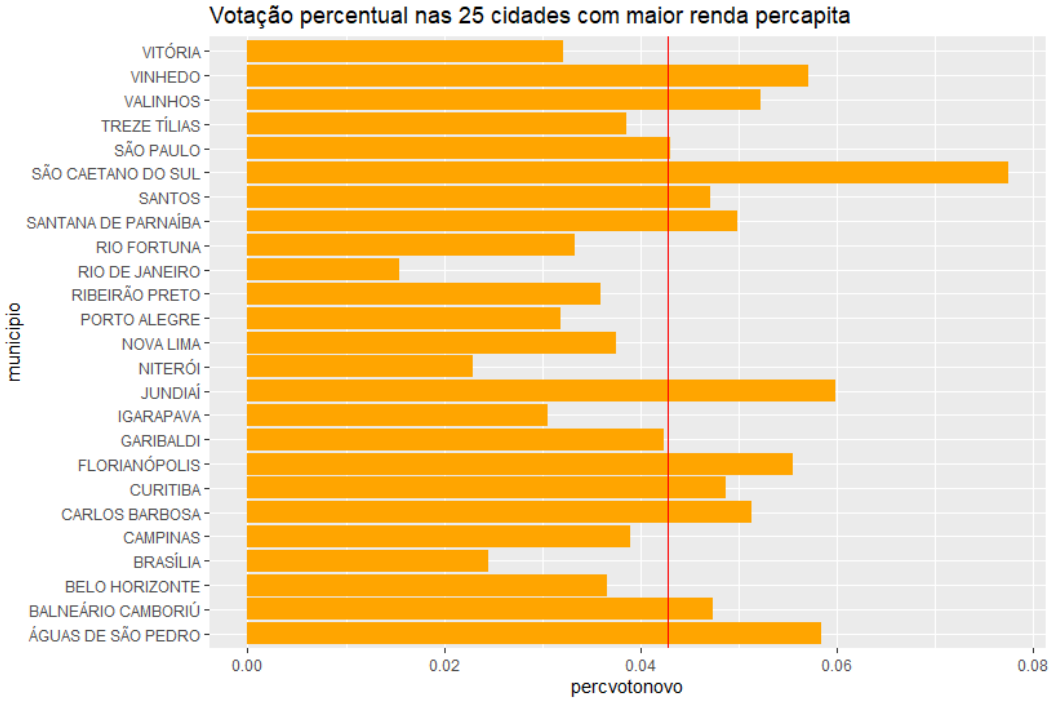
Burnham e Anderson (2004), ressaltam a importância da seleção de modelos em evidências cientificas. Dentre as alternativas disponíveis escolhemos o modelo de critérios de informação de Akaike (AIC) para comparar o desempenho dos modelos.

ANÁLISE ESPLORATÓRIA DE DADOS

Em primeiro lugar, foram escolhidas variáveis que expressam os valores do partido, comunicados em seu estatuto, utilizando a hipótese de identidade ideológica como fator decisivo do voto. Antes de se aplicarem as análises geográficas, serão exibidos, para algumas das variáveis analisadas, os 25 municípios do Brasil com maior índice daquela variável. E será feito um gráfico comparando o percentual de votos entre os municípios exibidos, de forma a orientar as ações do partido para as

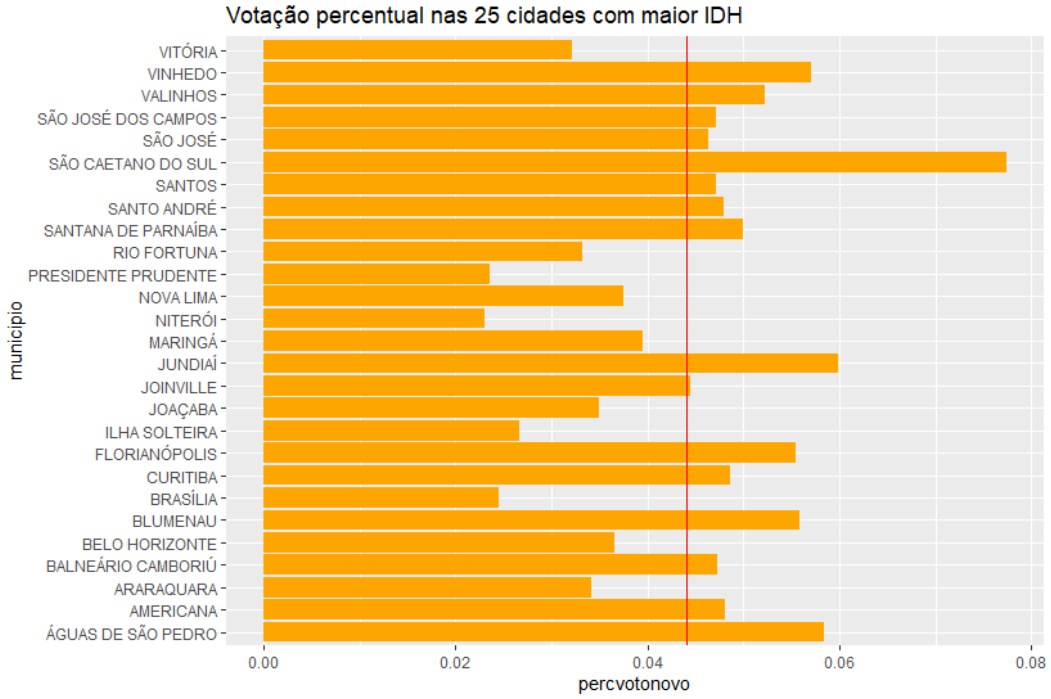
próximas eleições. Com os gráficos, o partido poderá vislumbrar onde deve concentrar seus esforços e qual assunto (variável) deve ser explorada em suas campanhas.

Gráfico 1: Votação percentual nas 25 cidades com maior renda per capita.



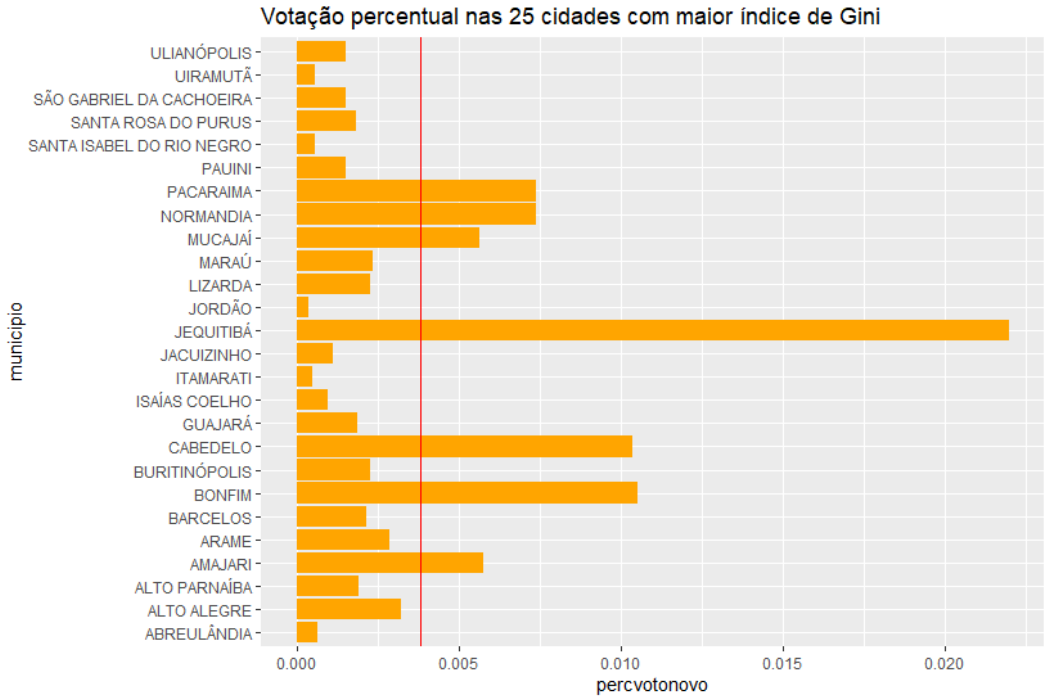
Fonte: Tribunal Superior Eleitoral (TSE) – Elaboração própria dos autores, a partir do software R.

Gráfico 2: Votação percentual nas 25 cidades com maior IDHM.



Fonte: Atlas Brasil – Elaboração própria dos autores, a partir do software R.

Gráfico 3: Votação percentual nas 25 cidades com maior índice de GINI.



Fonte: Atlas Brasil – Elaboração própria dos autores, a partir do software R.

Para descobrir as variáveis que poderiam ser utilizadas para o modelo, foram testadas suas correlações com a variável dependente que é percentual de votos no NOVO por município. A variável dependente (percvotonovo), assim como as variáveis consideradas, estão apontadas abaixo no (Quadro 1) com a descrição de cada variável.

Quadro 1: Variáveis, descrições e fontes.

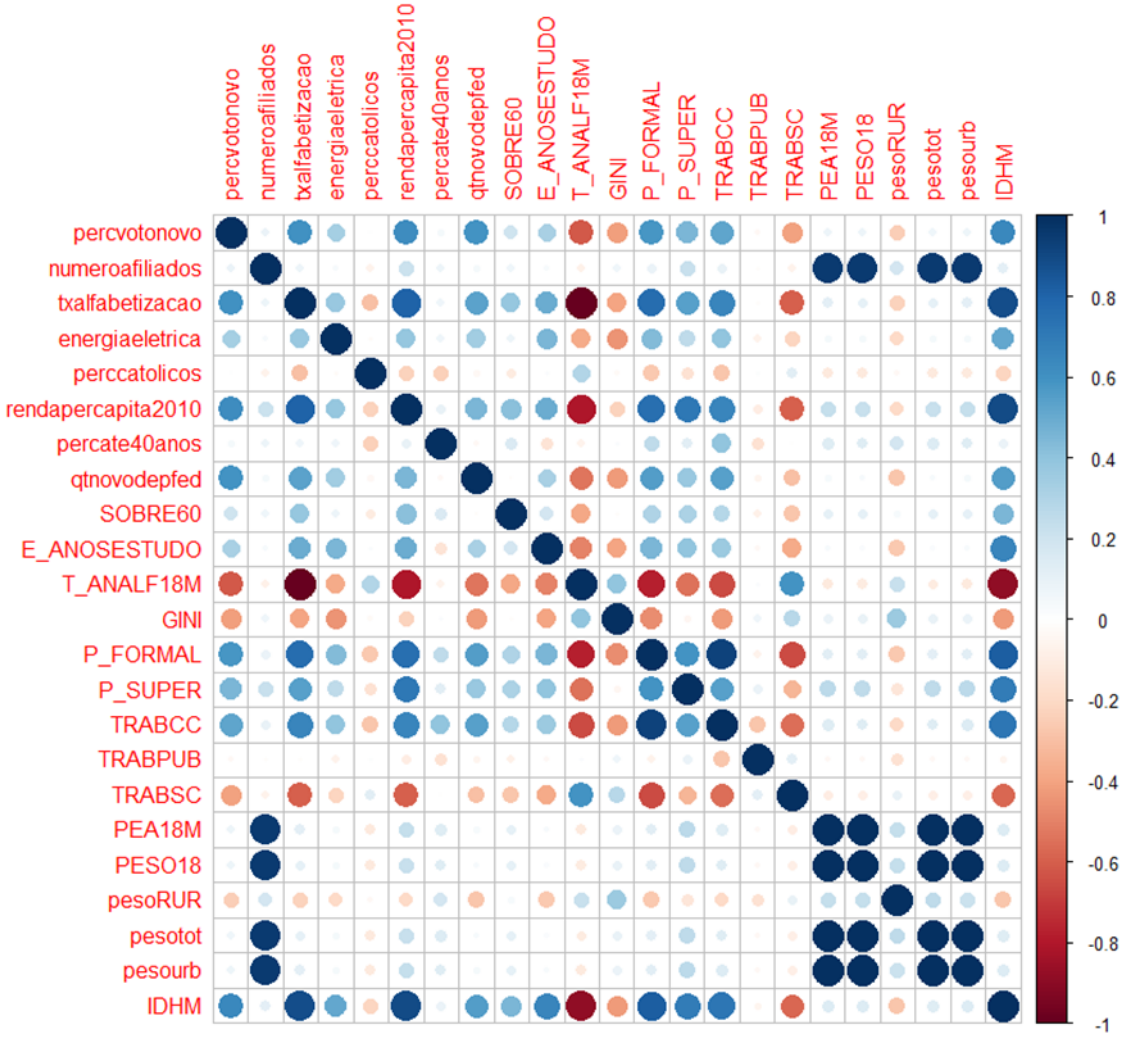
Variáveis	Descrição	Fonte
percvotonovo	Percentual de votos para presidente do partido NOVO em relação a população de 18 anos ou mais	TSE
numeroafiliados	Número de afiliados do partido NOVO no município	TSE
txalfabetizacao	Taxa de alfabetização	IBGE
energiaeletrica	Percentual de domicílios com energia elétrica	IBGE
perccatolicos	Percentual de católicos	IBGE
rendapercapita	Renda per capita (base salário mínimo de 2010: R$510)	IBGE
percate40anos	Percentual de pessoas de até 40 anos	IBGE
qtnovodepfed	Quantidade de candidatos a deputado federal do NOVO para o município	TSE
sobre60	Probabilidade de sobrevivência até 60 anos	Atlas Brasil
e_anosestudo	Expectativa de anos de estudo aos 18 anos de idade	Atlas Brasil
t_analf18M	Taxa de analfabetismo da população de 18 anos ou mais de idade	Atlas Brasil
GINI	Índice de Gini: grau de desigualdade existente na distribuição de indivíduos segundo a renda domiciliar per capita.	Atlas Brasil
p_formal	Grau de formalização do trabalho das pessoas ocupadas	Atlas Brasil
p_super	Percentual dos ocupados com superior completo	Atlas Brasil
trabcc	Percentual de ocupados de 18 anos ou mais que são empregados com carteira	Atlas Brasil
trabpub	Percentual de ocupados de 18 anos ou mais que são trabalhadores do setor público.	Atlas Brasil
trabsc	Percentual de ocupados de 18 anos ou mais que são empregados sem carteira	Atlas Brasil
pea18M	População economicamente ativa de 18 anos ou mais de idade	Atlas Brasil
peso18	População de 18 anos ou mais de idade	Atlas Brasil
pesoRUR	População rural	Atlas Brasil
pesoTOT	População total	Atlas Brasil
pesoURB	População urbana	Atlas Brasil
IDHM	Índice de Desenvolvimento Humano Municipal. Média geométrica dos índices das dimensões Renda, Educação e Longevidade, com pesos iguais.	Atlas Brasil

Fonte: Elaboração própria dos autores.

Os resultados da matriz de correlação estão descritos no (Figura 1). Percebe-se que a taxa de alfabetização, renda per capita, quantidade de candidatos a deputado federal pelo NOVO por estado, taxa de analfabetismo, índice de GINI, grau de formalização do trabalho, percentual dos ocupados com superior completo, percentual de ocupados com carteira, percentual de ocupados sem carteira e

o IDHM foram as variáveis com correlações mais significativas, merecendo ser consideradas para uma regressão.

Figura 1 – Matriz de correlação.



Fonte: IBGE, TSE e Atlas Brasil – Elaboração própria dos autores, a partir do software R.

Entretanto, muitas dessas variáveis possuem correlação elevada entre elas mesmas, como é o caso de taxa de alfabetização e taxa de analfabetismo, renda per capita e IDHM, entre outras. Para evitar que essa correlação entre variáveis explicativas cause multicolinearidade, a escolha das variáveis a serem utilizadas levou isso em conta.

Com as variáveis selecionadas, utilizamos a otimização de *Natural Breaks*, método de classificação de quebras naturais de Jenks (1977), para observar a distribuição e o agrupamento de dados no mapa.

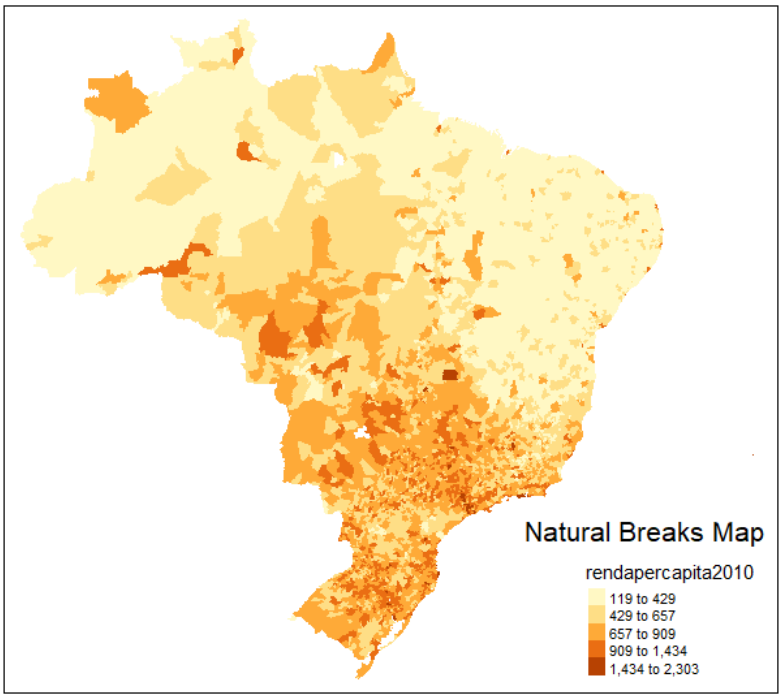
Pelo mapa da (Figura 2) é possível observar as classes de maior renda per capita concentradas nas regiões Centro-Oeste, Sudeste e Sul do Brasil. Semelhante a distribuição de renda per capita, a quantidade de deputados federais disputando um cargo pelo partido NOVO, tiveram maior concentração nos estados de São Paulo, Minas Gerais, Rio de Janeiro, Mato Grosso, Paraná, Santa Catarina e Rio Grande do Sul (Figura 3).

Já o coeficiente de GINI, índice criado pelo italiano Conrado Gini (1912) para medir a desigualdade de renda de uma população, costumeiramente sendo medido em 0 para igualdade e 1 para desigualdade, observado no mapa da (Figura 4) é possível verificar uma maior distribuição de renda nas regiões Sudeste e Sul, tendo as regiões Norte e Nordeste apresentado maiores índices de desigualdade de renda.

A taxa de analfabetismo da população de 18 anos ou mais de idade está mais elevada no extremo Norte e na região do Nordeste, indicando a falta de instrução de uma maior parte da população dessas regiões (Figura 5).

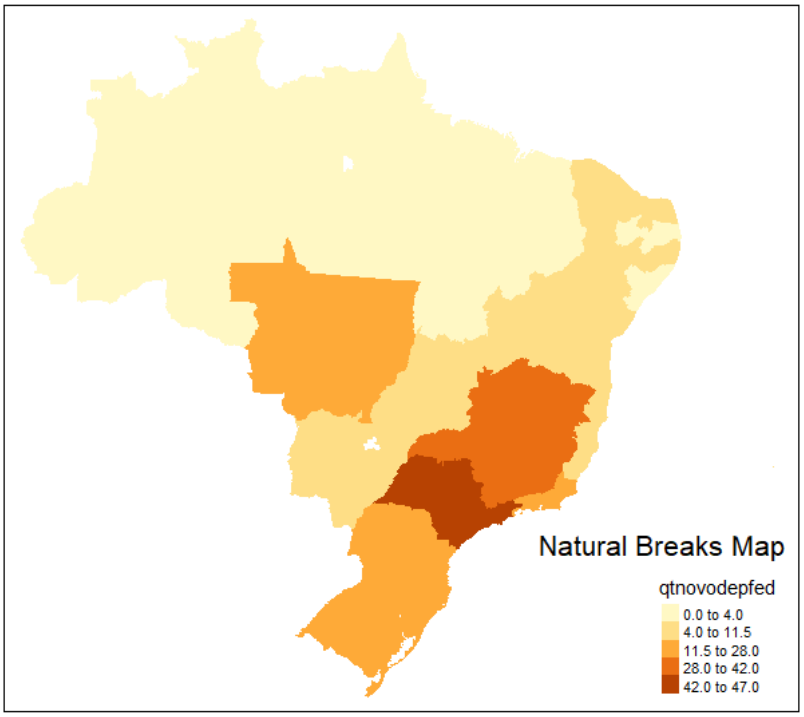
Visualmente, é possível verificar que, nas regiões com maior renda per capita, maior quantidade de deputados federais pelo partido NOVO, menor coeficiente de GINI e menores taxas de analfabetismo, estão as maiores taxas de distribuições de votos para o partido NOVO, ou seja, nas regiões Centro-Oeste, Sul e principalmente Sudeste do país (Figura 6), indicando assim, correlação entre as variáveis explicativas e a variável independente.

Figura 2: Mapa, renda per capita (Natural Breaks).



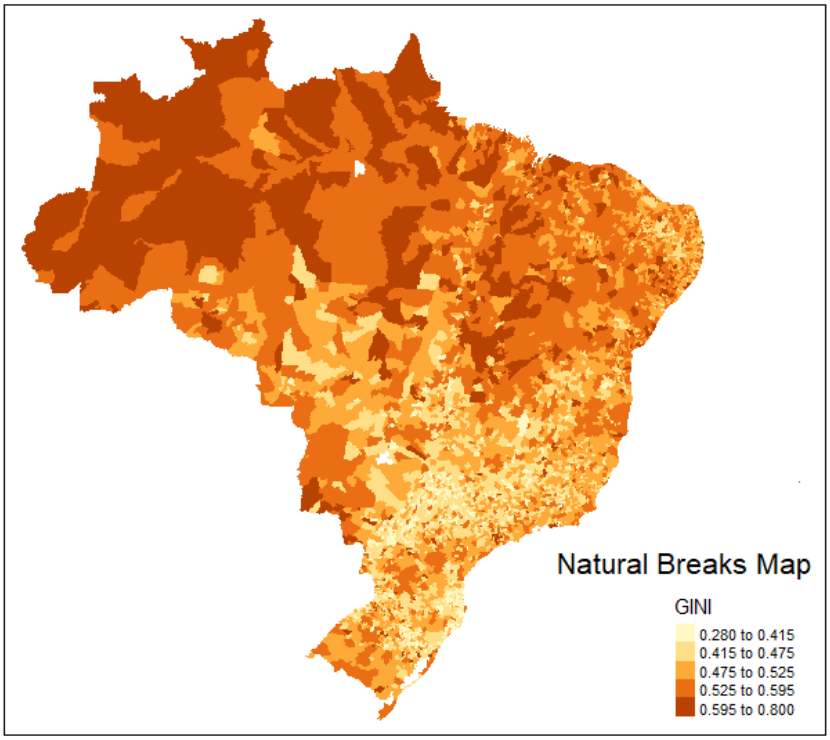
Fonte: IBGE – Elaboração própria dos autores, a partir do software R.

Figura 3: Mapa, quantidade de candidatos a deputado federal pelo partido NOVO (Natural Breaks).



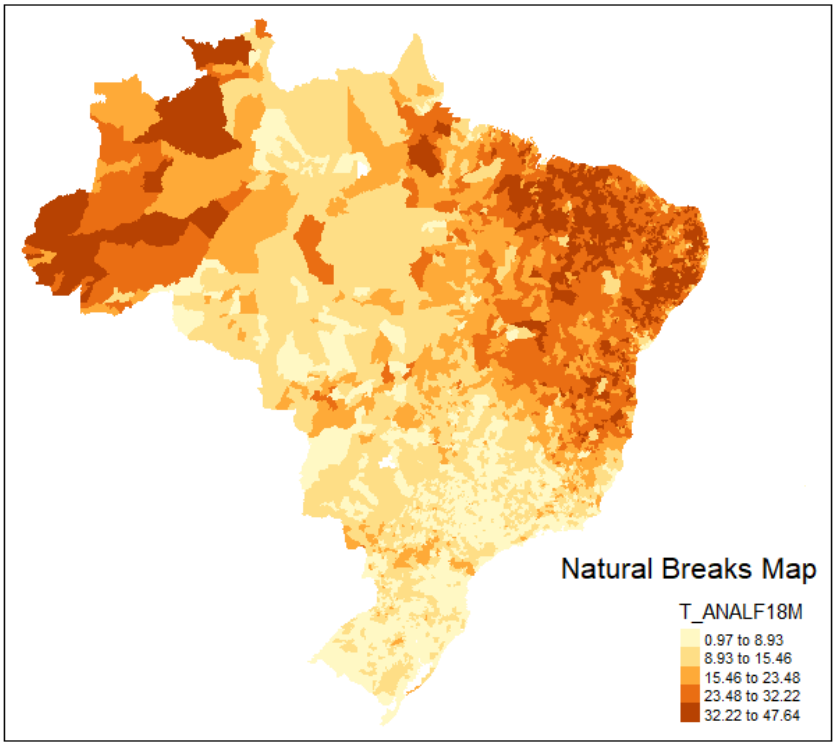
Fonte: IBGE e TSE – Elaboração própria dos autores, a partir do software R.

Figura 4: Mapa, índice de GINI (Natural Breaks).



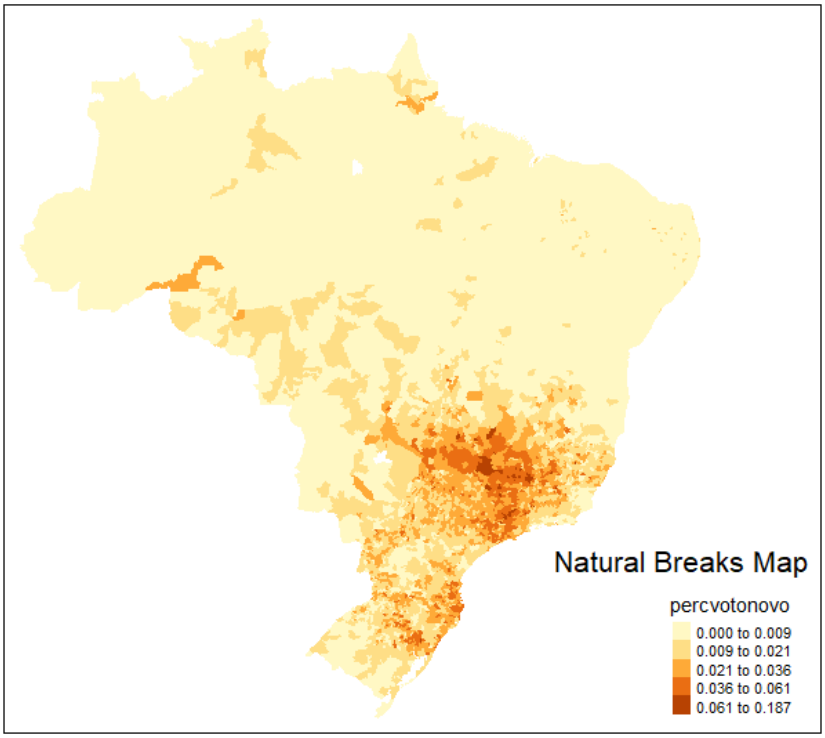
Fonte: IBGE e Atlas Brasil – Elaboração própria dos autores, a partir do software R.

Figura 5: Mapa, taxa de analfabetismo (Natural Breaks).



Fonte: IBGE e Atlas Brasil – Elaboração própria dos autores, a partir do software R.

Figura 6: Mapa, percentual de votos para presidente do partido NOVO (Natural Breaks).



Fonte: IBGE e TSE – Elaboração própria dos autores, a partir do software R.

ANÁLISE ESPACIAL

Dadas as proporções continentais do Brasil, assim como características socioeconômicas que variam bastante ao longo do território, é de se esperar que o espaço exerça influência na votação para presidente de um partido novo. Isso está de acordo com a primeira lei da geografia, "tudo está relacionado com tudo, mas coisas mais próximas estão mais relacionadas que as distantes" (Tobler, 1970).

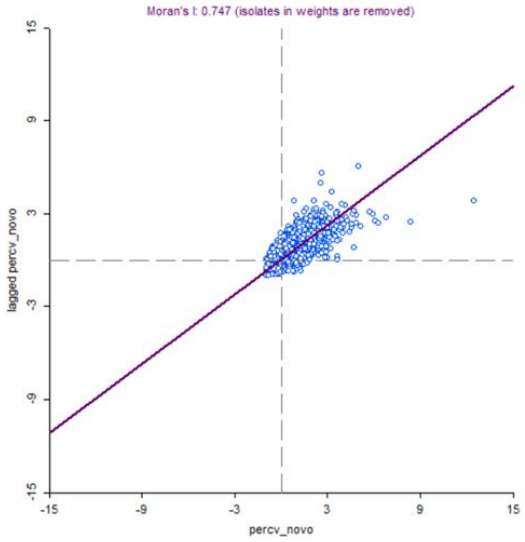
Adiciona-se também, o fato de que o NOVO, adotou uma estratégia de divulgação majoritariamente via redes sociais e seu foco para a eleição de deputados federais, estaduais, senadores e governadores foram apenas em alguns estados.

Para se adequar aos dados, foram excluídos do mapa dois municípios (Lagoa dos Patos – RS e Lagoa Mirim – RS) que aparecem na malha municipal de 2010 do IBGE, mas que não estavam nas estatísticas do Censo 2010, fonte da maior parte dos dados municipais brasileiro.

Sendo assim, realizou-se análise espacial por meio do teste espacial de *I de Moran Univariado Local*. O *I de Moran Univariado* e os *Indicadores Locais de Associação Espacial - LISA*, conforme proposto por Anselin et al (2007), são as técnicas mais comuns para explorar a existência de autocorrelação espacial entre amostras. E utilizou-se o local, porque ele mostra onde estão os clusters, ao contrário do global que responde se existem clusters espaciais.

O *I de Moran* obtido, a partir de uma matriz de vizinhança do tipo Rook de Ordem 1, está demonstrado no mapa de dispersão do (Gráfico 4) e foi de 0,747, o que demonstra uma autocorrelação positiva e significativa, ou seja, o percentual de votos no NOVO no Brasil tem algum tipo de padrão espacial.

Gráfico 4: I de Moran.

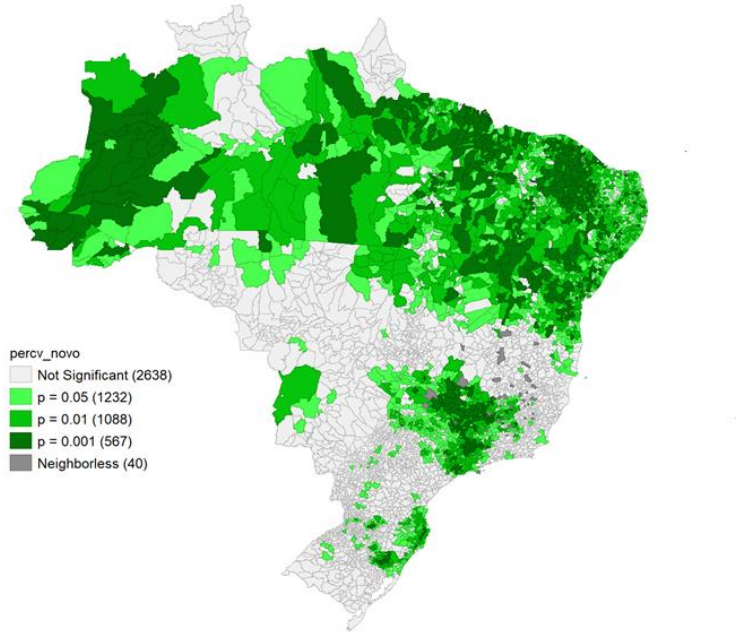


Fonte: TSE – Elaboração própria dos autores, a partir do software GeoDa.

Percebe-se grande concentração das observações no 1º quadrante, ou seja, de um município alto vizinho de outro município alto.

Foi elaborado, além disso, o Mapa de Significância - LISA para as observações, conforme a (Figura 7), evidenciando que os municípios significantes para a análise espacial estão concentrados no Norte, Nordeste, Sudeste e litoral Sul.

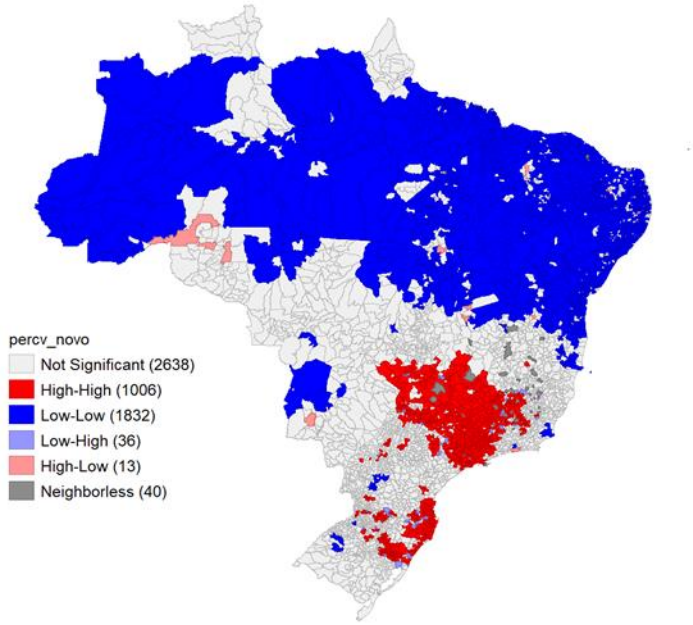
Figura 7: Mapa de Significância – LISA.



Fonte: IBGE e TSE – Elaboração própria dos autores, a partir do software GeoDa.

Também, foi realizado o Mapa de Cluster - LISA, evidenciando os clusters existentes para o Brasil, conforme a (Figura 8). Visualmente, é possível perceber a presença de um cluster *low-low* englobando grande parte das regiões Norte e Nordeste, enquanto as regiões Sudeste e litoral Sul indicam a presença de clusters *high-high*. Esse cluster do Sudeste é em linha com o que se podia esperar, já que as ações do partido foram mais focadas ali, com um candidato a governador do partido, inclusive, sendo eleito.

Figura 8: Mapa de Cluster – LISA.



Fonte: IBGE e TSE – Elaboração própria dos autores, a partir do software GeoDa.

METODOLOGIA

Verificada a importância de se considerar o espaço geográfico para a análise de determinantes do voto no candidato do partido NOVO para presidente, dividiu-se a análise em três partes:

- Regressão linear: antes da verificação do efeito espacial, foram testadas diversas variáveis que pudessem prever a votação em João Amoedo, candidato do partido NOVO para presidência em 2018;
- Regressão espacial global: inclusão do efeito do espaço, pelo teste da dependência espacial e utilização de lag espacial ou erro espacial, de acordo com a significância;
- Regressão geograficamente ponderada: feita uma análise espacial com regressões estimadas por região, ponderada pela distância.

Por meio da técnica *Stepwise* do tipo *backward*, foram verificadas quais variáveis eram significativas para prever o percentual de votos para o candidato do NOVO a presidente em um município. Além disso, testou-se o *Valor de Inflação da Variância – VIF* para verificar a severidade da

multicolinearidade em uma análise de regressão de mínimos quadrados ordinários (MONTGOMERY, PECK, VINING, 2006).

Após essas verificações, chegou-se em um modelo conforme demonstrado abaixo:

$$percvotonov = \beta_0 + \beta_1 * rendapercapita2010 + \beta_2 * qtnovodepfed + \beta_3 * T_{ANALF18M} + \beta_4 * GINI + \varepsilon$$

As comparações entre os modelos serão feitas pelo Critério de Informação de Akaike (AIC), que é uma medida relativa da qualidade de ajuste de um modelo estatístico estimado e fundamenta-se no conceito de entropia, oferecendo uma medida relativa das informações perdidas quando um determinado modelo é usado para descrever a realidade. Akaike encontrou uma relação entre a esperança relativa da K-L informação e a função suporte maximizada, permitindo uma maior interação entre a prática e a teoria, em seleção de modelos e análises de conjuntos de dados complexos (Emiliano, 2009).

O critério, muito utilizado para escolher o melhor modelo, funciona da seguinte maneira: quanto menor o seu índice, melhor é o modelo.

RESULTADOS DA REGRESSÃO LINEAR MÚLTIPLA

Os resultados da regressão linear múltipla com as variáveis renda per capita, total de deputados federais candidatos ao partido NOVO, taxa de analfabetismo e o índice de GINI, podem ser observados na (Figura 9) abaixo.

Figura 9: Regressão linear múltipla.

```
Call:
lm(formula = percvotonovo ~ rendapercapita2010 + qtnovodepfed +
    T_ANALF18M + GINI, data = base)

Residuals:
      Min        1Q    Median        3Q       Max
-0.025459 -0.004961 -0.001157  0.002871  0.161870

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.662e-02  1.241e-03  13.396  < 2e-16 ***
rendapercapita2010 2.114e-05  8.202e-07  25.773  < 2e-16 ***
qtnovodepfed       2.730e-04  9.854e-06  27.703  < 2e-16 ***
T_ANALF18M        -6.791e-05  2.194e-05  -3.095  0.00198 **
GINI              -3.713e-02  2.217e-03 -16.752  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.009432 on 5560 degrees of freedom
Multiple R-squared:  0.5441,    Adjusted R-squared:  0.5437
F-statistic:  1659 on 4 and 5560 DF,  p-value: < 2.2e-16
```

Fonte: IBGE, TSE e Atlas Brasil – Elaboração própria dos autores, a partir do software R.

Para esse modelo, o valor-p associado com a estatística-F do modelo é menor do que o α de 0,05, ou seja, o modelo é significante para explicar a variável dependente. Verifica-se também, que todas as variáveis são significativas para o modelo a um nível de significância de 0,01, ou seja, não serão iguais a 0 os valores.

Ainda com essas variáveis, também foi testado o VIF, que pode ser verificado na (Figura 10) abaixo:

Figura 10: Valor de Inflação da Variância – VIF.

```
rendapercapita2010       qtnovodepfed          T_ANALF18M               GINI
          2.983086           1.532750            3.447281           1.341407
```

Fonte: IBGE, TSE e Atlas Brasil – Elaboração própria dos autores, a partir do software R.

Uma vez que o VIF quantifica qual a correlação de um preditor com outros preditores de um modelo, quanto maior o VIF, maior a probabilidade de ocorrer multicolinearidade. Seu menor valor possível é 1, que é quando não há correlação com outras variáveis, enquanto valores maiores que 4 ou 5 podem apontar correlação moderada a alta. Valores maiores que 10 podem ser caracterizados como possuindo correlação muito grande.

Para o modelo em questão, nenhuma variável apresentou fator maior que 4, o que aponta para uma baixa correlação entre as variáveis explicativas.

Devido ao tamanho da amostra analisada (5.565 observações), não será testada a normalidade já que, para grandes amostras, até uma pequena variação vai levar a um resultado significativo e que rejeita a normalidade (Field, 2009). Isso não será um problema pois como o teorema do limite central afirma, a média amostral converge para uma distribuição normal.

Realizou-se a verificação de existência de autocorrelação, ou seja, se há alguma correlação dos valores de uma mesma variável ordenados no tempo ou no espaço. Dado o valor do teste de *Durbin-Watson* próximo de 2, que revela ausência de autocorrelação, assim como o p-valor que não rejeita a hipótese nula de ausência de autocorrelação, supõe-se ausência de autocorrelação. Esse resultado já era esperado, já que não há ligação entre observações sem a dimensão tempo.

Foi então testado se o modelo apresenta heterocedasticidade, problema comum em dados transversais, que é quando o erro apresenta variância desigual. A heterocedasticidade pode resultar de várias razões, como a presença de valores discrepantes nos dados, uma forma funcional incorreta do modelo de regressão, a transformação incorreta de dados ou a mistura de observações com diferentes medidas de escala (Gujarati, 2019). Por essa razão, foi feita a verificação por meio do teste de *Breusch-Pagan* (Figura 11).

Figura 11: Teste de *Breusch-Pagan*.

```
studentized Breusch-Pagan test
BP = 153.05, df = 4, p-value < 2.2e-16
```

Fonte: Elaboração própria dos autores, a partir do software R.

Considerando a hipótese nula de que a variância do erro é homocedástica, para o p-valor significativo que foi obtido podemos rejeitar a hipótese nula, ou seja, há presença de heterocedasticidade. Ainda segundo Gujarati (2019), a heterocedasticidade não altera as propriedades dos estimadores de MQO de serem não viesados e consistentes, mas eles deixam de ser de variância mínima ou eficientes.

Como resultado os testes t e F com base nas premissas padrão podem não ser confiáveis, resultando em conclusões errôneas a respeito da significância estatística dos coeficientes de regressão estimados.

Essa heterocedasticidade identificada pode ter sua raiz no espaço geográfico, já que para um país de grandes dimensões e diversificado como o Brasil, possuindo um elevado número de municípios, é de se supor a grande variação entre as unidades territoriais mais distantes.

CONSIDERAÇÕES PARA A ANÁLISE ESPACIAL

Por essa razão, para avaliar a viabilidade de testar modelos espaciais nos nossos estudos, foram analisados alguns testes emitidos pelo software GeoDa, quando da análise da regressão simples (Tabela 1).

Tabela 1: Teste Jarque-Bera, teste Breusch-Pagan, teste Koenker-Bassett e teste White, GeoDa.

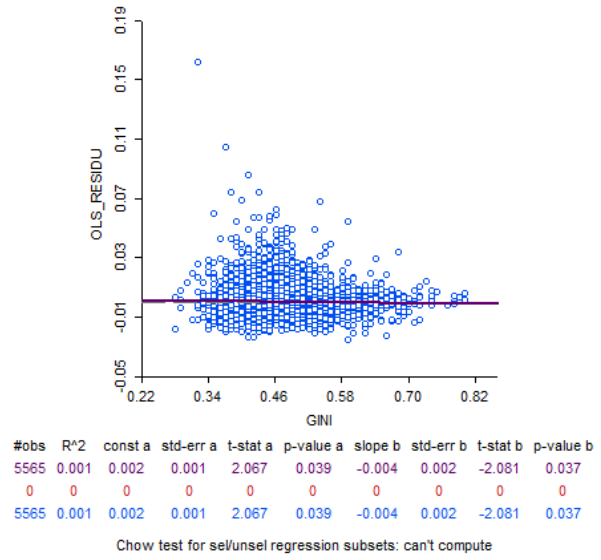
Test	***DF***	***Value***	***Prpb***
Jarque-Bera	2	168798.4617	0.00000
Breusch-Pagan test	4	1888.7113	0.00000
Koenker-Bassett test	4	133.5337	0.00000
White	14	252.2147	0.00000

Fonte: Elaboração própria dos autores, a partir do software Geoda.

O primeiro verificado foi o teste de não normalidade (Jarque-Bera). Esse teste considera como hipótese nula que a distribuição é normal. O resultado aponta para a rejeição da hipótese nula, ou seja, os erros não seriam distribuídos normalmente.

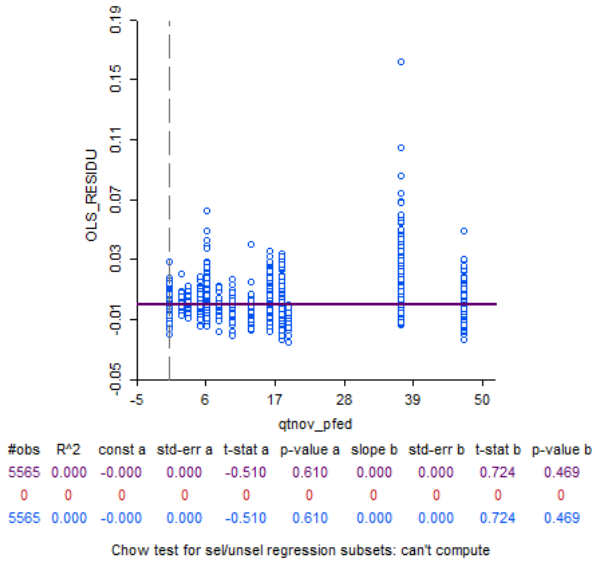
Os próximos testes verificam a questão da heterocedasticidade (*Breusch-Pagan*, *Koenker-Bassett* e teste *White*). Nesse caso, os testes consideram como hipótese nula que a variância dos erros é constante (homocedasticidade). Novamente, verificando o valor da probabilidade, deve-se rejeitar a hipótese nula. Os gráficos (5 a 8) abaixo, mostram como os resíduos se comportam em relação as variáveis.

Gráfico 5: Renda per capta (Residual).



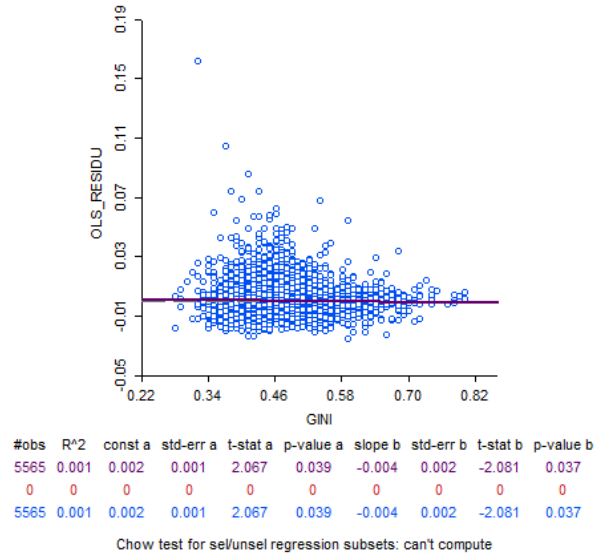
Fonte: Atlas Brasil – Elaboração própria dos autores, a partir do software GeoDa.

Gráfico 6: Dep. Federais P.NOVO (Residual).



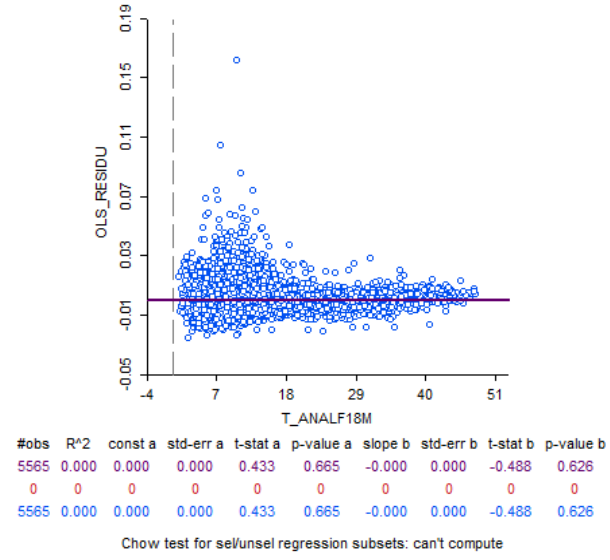
Fonte: TSE – Elaboração própria dos autores, a partir do software GeoDa.

Gráfico 7: Índice de GINI (Residual).



Fonte: Atlas Brasil – Elaboração própria dos autores, a partir do software GeoDa.

Gráfico 8: T. de Analfabetismo (Residual).



Fonte: Atlas Brasil – Elaboração própria dos autores, a partir do software GeoDa.

Os testes acima e os gráficos indicam que há algo que as variáveis escolhidas não estão conseguindo explicar completamente. Mas o que determina de forma objetiva a pertinência da utilização de modelos espaciais são os multiplicadores de Lagrange. Em outras palavras, esse teste verifica se vale ou não a pena testar o modelo espacial com nossos dados (Tabela 2).

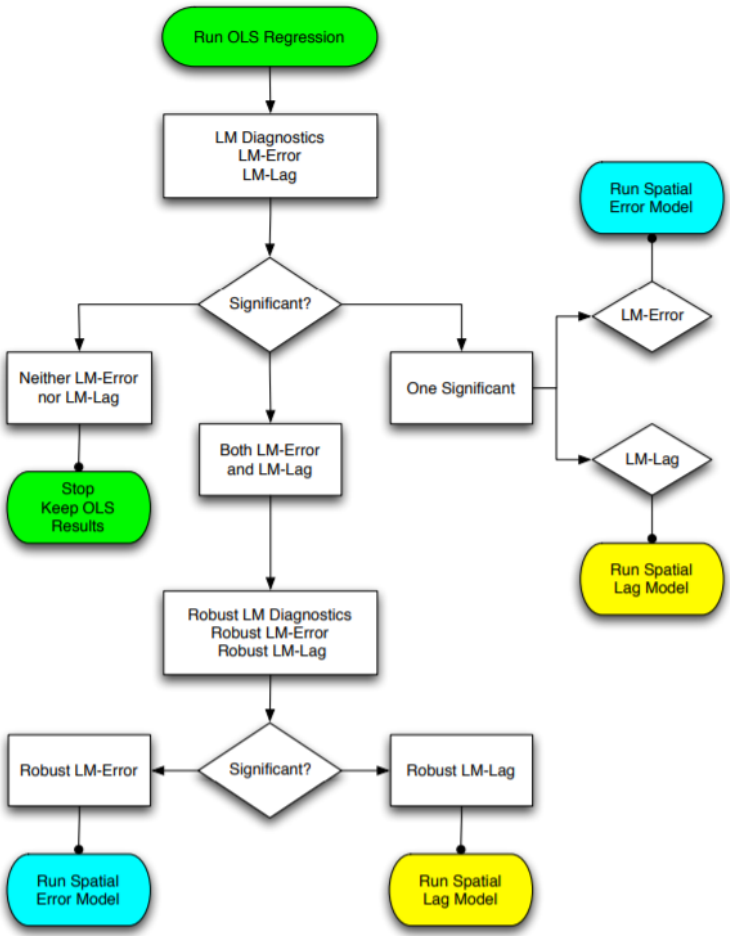
Tabela 2: Diagnostics for Spatial Dependence.

Test	*MI/DF*	*Value*	*Prob*
Moran's I (error)	0.5014	59.8621	0.00000
Lagrange Multiplier (lag)	1	2957.9804	0.00000
Robust LM (lag)	1	5.1300	0.02352
Lagrange Multiplier (error)	1	3563.7592	0.00000
Robust LM (error)	1	610.9088	0.00000
Lagrange Multiplier (SARMA)	2	3568.8892	0.00000

Fonte: Elaboração própria dos autores, a partir do software R.

O primeiro teste é *Moran's I* erro. Obteve-se um resultado de 0.5014 e com uma probabilidade muito baixa de estar errado. Esse resultado indica que é grande a chance de existir uma autocorrelação dos resíduos e é um indício forte que pode ser interessante utilizar modelos espaciais para diminuir esta autocorrelação. Na análise dos multiplicadores de Lagrange, foi considerado o fluxograma (Figura 12) *Spatial Regression Decision Process* (Anselin, p.198, 2005).

Figura 12: Fluxograma – Spatial Regression Decision Process.



Fonte: Spatial Regression Decision Process (Anselin, p.198, 2005).

Em resumo, foram obtidos resultados significantes em todos os testes, inclusive os testes robustos. Nesse caso, é necessário executar os dois modelos (Spatial Error e Spatial Lag) e verificar qual dos dois modelos é o melhor. Será utilizado o modelo de critérios de informação de Akaike (AIC) para comparar o desempenho dos modelos.

A ideia de realizar uma regressão espacial é incorporar aspectos geográficos ao modelo com o intuito de aumentar o poder preditivo dele.

Basicamente, a técnica pode ser aplicada de maneira global ou de forma local. Segundo Francisco (2010), a primeira maneira utiliza indicadores sumarizados para região como um todo, destacando as similaridades no espaço e "leis" gerais. A segunda forma utiliza a desagregação local das estatísticas globais, buscando as diferenças no espaço, procurando as exceções ou "hot" spots locais.

No tocante aos modelos globais, foram aplicadas duas técnicas: O *SAR – Modelo Espacial Auto Regressivo* e o *Spacial Rrror Model*. Com relação aos modelos locais, aplicou-se *Regressão Geograficamente Ponderada* (GWR).

SAR – Modelo Espacial Auto Regressivo

O ponto principal dos modelos SAR, é que a variável dependente de um elemento do modelo, possui influência no valor da variável dependente dos vizinhos. Segundo Carvalho Ywata e Albuquerque (2011), "*A ideia dos modelos SAR é utilizar a mesma ideia dos modelos AR (autorregressivos) em séries temporais, por meio da incorporação de um termo de lag entre os regressores da equação*".

Na sua forma mais simples, o modelo SAR tem expressão:

$$y = \rho W y + X\beta + \varepsilon \qquad 1.0$$

Segundo Francisco (2010), pode-se interpretar a fórmula acima da seguinte maneira: (W) é a matriz de proximidade espacial, o produto (Wy) representa a dependência espacial em (y) e (ρ) é o

coeficiente espacial autorregressivo. A hipótese nula para a não existência de autocorrelação, é que (ρ) é igual a zero. A ideia básica deste modelo é incorporar a auto correlação espacial como componente do modelo. O modelo também é conhecido como *Spatial Lagged Autoregressive Model*.

Spatial Error Model

No *Spatial Error*, o "erro" de um elemento influencia nos erros dos vizinhos. A ideia é que existe uma "variável escondida" que não se consegue descobrir qual, pois não se possui o dado ou informação, mas ela se distribui no espaço e tentou-se simular por meio dos erros da vizinhança. A formulação é a seguinte:

$$y = X\beta + u \quad \text{, onde } u = \lambda W + u + \varepsilon \qquad 1.1$$

Analisando a fórmula acima Carvalho Ywata e Albuquerque (2011) teceram o seguinte comentário:

> *O vetor de resíduos ε possui distribuição normal multivariada, com média nula e matriz de covariância σ^2 I ¬. O coeficiente escalar λ indica a intensidade da autocorrelação espacial entre os resíduos da equação observada. Mais especificamente, esse parâmetro mensura o efeito médio dos erros dos vizinhos em relação ao resíduo da região em questão. Note-se que, ao contrário dos modelos SAR, os modelos SEM não apresentam a variável resposta como uma função direta dos seus lags espaciais. A autocorrelação espacial nos modelos SEM aparece nos termos de erro. Outra diferença dos modelos SEM em relação aos modelos SAR é que os coeficientes no vetor , podem ser estimados consistentemente via mínimos quadrados ordinários. (Rev. Bras. Biom., São Paulo, v.29, n.2, p.273-306, 2011.).*

GWR - Geographically Weighted Regression

A ideia principal do modelo GWR, é que os parâmetros vão mudando ao longo do mapa. O foco é achar as diferenças no Mapa, com o intuito de prover um modelo mais eficiente.

Segundo Francisco (2010) a regressão ponderada geograficamente descreve uma gama de modelos de regressão em que os coeficientes, parâmetros (β), variam de acordo com a localidade. Ela ajusta

um modelo de regressão a cada ponto observado, ponderando todas as demais observações como função da distância (ou de qualquer medida de vizinhança) deste ponto. Em outras palavras, têm-se regressões diferentes para cada observação, assim a contribuição (valor do parâmetro) de cada variável explicativa ao modelo é diferente para cada ponto.

$$y(g) = \beta_0(\mathrm{g}) + \beta_1(\mathrm{g})x_1 + \beta_2(\mathrm{g})x_2 + \ldots + \beta_p(\mathrm{g})x_p + \varepsilon \qquad 1.2$$

Segundo Francisco (2010, p.123), "*g é um vetor dos n pontos, no espaço bidimensional, os parâmetros do vetor β(g) são específicos para cada observação i de localização g(i)=(u_i,v_i) e o termo de erro ε é suposto e independente e de comportamento ε ~ N(0,σ^2 I). Temos na realidade, um conjunto de n regressões diferentes, um para cada ponto g_i no espaço*".

RESULTADOS DAS REGRESSÕES ESPACIAIS

A primeira coisa que verificou-se é que o *Spatial Lag Parâmetro* é significativo. O (ρ) tenta explicar a influência dos valores da variável dependente dos vizinhos na variável dependente e se o efeito é positivo ou negativo. Nota-se, um efeito positivo de 0.69176. Prosseguindo com a análise dos estimadores da regressão, não se pode olhar para os resultados da (Figura 13) e tomá-lo como efeitos marginais, uma vez que existe um feito de "feedback global". Quando uma variável é alterada (renda per capita, por exemplo), afeta não somente a variável dependente (Y), mas também seus vizinhos que, por conseguinte afetam a variável (Y) novamente, em um "loop" sem fim. Por isso, não é possível verificar se os estimadores são estatisticamente significantes, sendo necessário aplicar o comando "impacts" regressão. Conforme descrito na documentação do software R, com essa função é possível interpretar os impactos dos estimadores da regressão corretamente, devido às repercussões entre os termos nesses processos de geração de dados (diferentemente do modelo de erro espacial).

Figura 13: Resultado da Regressão Spatial Lag.

```
Call:lagsarlm(formula = percvotonovo ~ qtnovodepfed + rendapercapita2010 +
    GINI + T_ANALF18M, data = setores_juntos, listw = vizinhanca_pesos)

Residuals:
       Min         1Q     Median         3Q        Max
-0.02163268 -0.00322131 -0.00046487  0.00187491  0.13406661

Type: lag
Coefficients: (asymptotic standard errors)
                    Estimate  Std. Error z value  Pr(>|z|)
(Intercept)        4.9342e-04  9.7272e-04  0.5073  0.611975
qtnovodepfed       6.6885e-05  7.9961e-06  8.3648 < 2.2e-16
rendapercapita2010 1.2527e-05  6.5675e-07 19.0740 < 2.2e-16
GINI              -1.0988e-02  1.7469e-03 -6.2903 3.168e-10
T_ANALF18M         5.1888e-05  1.6640e-05  3.1183  0.001819

Rho: 0.69176, LR test value: 2539.4, p-value: < 2.22e-16
Asymptotic standard error: 0.011453
    z-value: 60.399, p-value: < 2.22e-16
Wald statistic: 3648, p-value: < 2.22e-16

Log likelihood: 19322.45 for lag model
ML residual variance (sigma squared): 5.063e-05, (sigma: 0.0071154)
Nagelkerke pseudo-R-squared: 0.71118
Number of observations: 5563
Number of parameters estimated: 7
AIC: -38631, (AIC for lm: -36093)
LM test for residual autocorrelation
test value: 12.616, p-value: 0.0003825
```

Fonte: IBGE, TSE e Atlas Brasil – Elaboração própria dos autores, a partir do software R.

Segundo Bivand, Roger e Piras, Gianfranco (2015), modelos espaciais estimados fornecem maneiras de inferir sobre a importância do lado direito das variáveis. Quando a variável dependente espacialmente defasada está presente, o coeficiente (β) valores e seus erros padrão não fornecem uma base satisfatória para inferência se ($\rho Lag \neq 0$). Este problema pode ser resolvido retirando amostras do modelo estimado, usando uma distribuição normal multivariada centrada nos valores ajustados de [ρLag, β], sua covariância matriz e, em seguida, calculando os impactos dos coeficientes da amostra (Figura 14).

Figura 14: Impactos dos coeficientes da amostra (Regressão espacial lag).

```
> impacts(regressao_espacial_lag, listw = vizinhanca_pesos)
Impact measures (lag, exact):
                         Direct      Indirect         Total
qtnovodepfed        7.581701e-05  1.411723e-04  0.0002169894
rendapercapita2010  1.419979e-05  2.644021e-05  0.0000406400
GINI               -1.245586e-02 -2.319298e-02 -0.0356488377
T_ANALF18M          5.881720e-05  1.095185e-04  0.0001683357
```

Fonte: IBGE, TSE e Atlas Brasil – Elaboração própria dos autores, a partir do software R.

Analisando os resultados acima, verifica-se que as variáveis total de deputados federais candidatos pelo partido NOVO e a taxa de analfabetismo, respectivamente (qtnovodepfed e T_ANALF18M), possuem impacto direto maior que o indireto. Por outro lado, as variáveis índice de GINI e renda per capita, respectivamente (GINI e rendapercapita2010), possuem impacto indireto maior que o direto, ou seja, a variáveis dos vizinhos tiveram mais contribuição, mais impacto do que o valor das variáveis do próprio elemento. Com relação ao impacto total, a variável que possui maior impacto é o índice de GINI com -0.0356488377.

A primeira tentativa de melhorar a performance dos resultados, foi por meio da execução do Modelo Espacial Auto Regressivo (SAR). Houve uma melhora significativa no resultado do (R^2), com o valor chegando ao patamar de 0.725904 e o AIC atingindo o valor de -38403.3. Dessa forma, o *Spatial Lag* aumentou a qualidade do poder de explicação do modelo. Avançando para a execução do *Spatial Error*, obteve-se uma pequena melhora em ralação ao SAR com um (R^2) chegando ao valor de 0.746666 e o AIC atingindo -38690.9. Em relação às variáveis utilizadas no estudo, permaneceram significativas nos três modelos executados, com exceção da variável “Taxa de analfabetismo da população de 18 anos ou mais de idade” que perdeu a significância na execução do modelo *spatial lag*. Para comparar a importância das variáveis em cada modelo será utilizado o coeficiente padronizado, ou seja, dividem-se os coeficientes pelo desvio padrão (Tabela 3). A ideia é verificar o quão é importante para o modelo cada uma das variáveis.

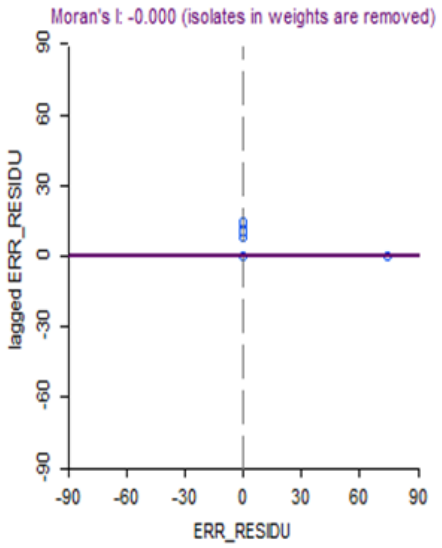
Tabela 3: Coeficiente padronizado.

	Regressão Simples	***Spatial Lag***	***Spatial Error***
Constante	13.40737457	1.436812793	11.95131
Renda_2010	25.77	19.64930361	23.01047
qtnov_pfed	27.70	10.29041682	14.04434
GINI	-16.7515136	-6.93420603	-12.4137
T_ANALF18M	-3.09586369	2.423619395	-4.1887
LAMBDA			62.06594768

Fonte: Elaboração própria dos autores, a partir do software R.

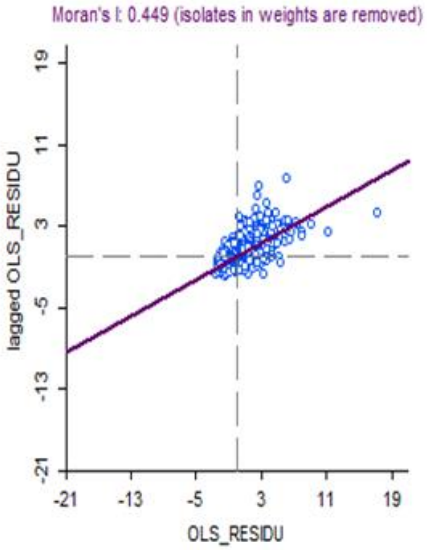
A partir da tabela acima, verifica-se que as variáveis, quando sujeitas a alguma técnica de incorporação de variável espacial perdem um pouco a sua importância. No modelo *Spatial Error*, por exemplo, a variável LAMBDA (que representa a média dos erros dos vizinhos), responde por boa parte do poder preditivo do modelo. Por outro lado, conforme evidenciado na (Tabela 3), há um ganho de eficiência e do poder preditivo nos modelos espaciais.

Gráfico 9: Moran's I – Erro Residual



Fonte: Elaboração própria dos autores, a partir do software GeoDa.

Gráfico 10: Moran's I – OLS Residual

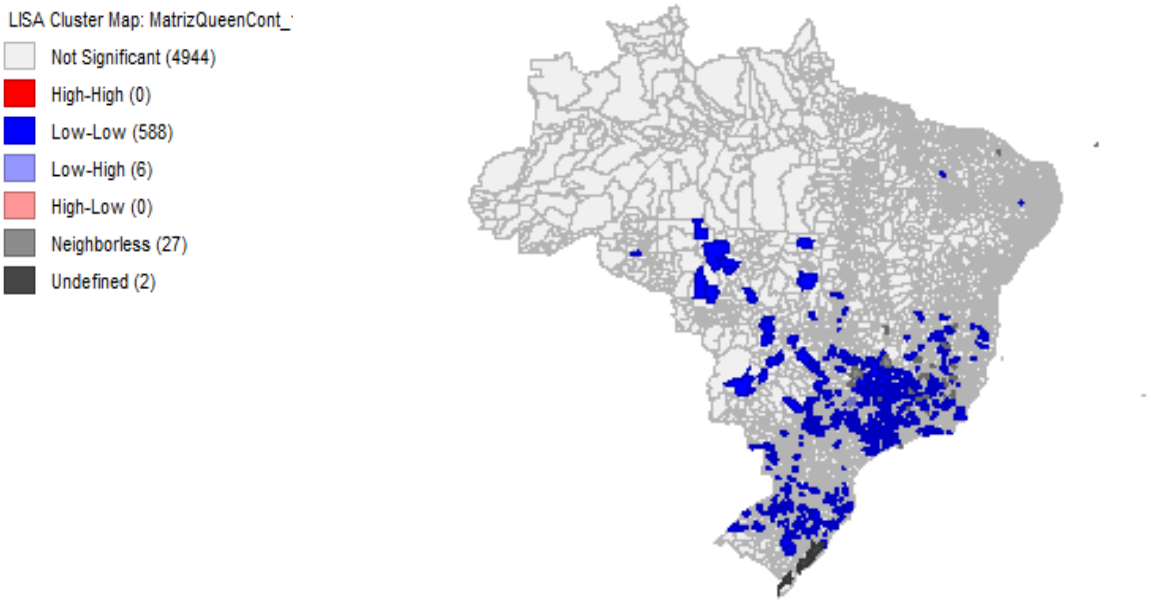


Fonte: Elaboração própria dos autores, a partir do software GeoDa.

Os Gráficos 9 e 10 acima, evidenciam o ganho do modelo *Spatial Error* na diminuição da autocorrelação espacial dos resíduos. A comparação entre *Índice de Moran* dos resíduos da regressão simples e os resíduos do modelo *Spatial Error* deixam claro essa diferença. Saiu-se de um patamar de 0,449 para um próximo a zero e negativo. O que significa que o modelo não está errando mais em

uma região do que em outra. O erro está espalhado de uma forma mais geral no mapa. Essa constatação fica mais clara de visualizar no Lisa Cluster abaixo (Figura 15):

Figura 15: Mapa Lisa Cluster – Resíduos do Modelo Spatial Error.



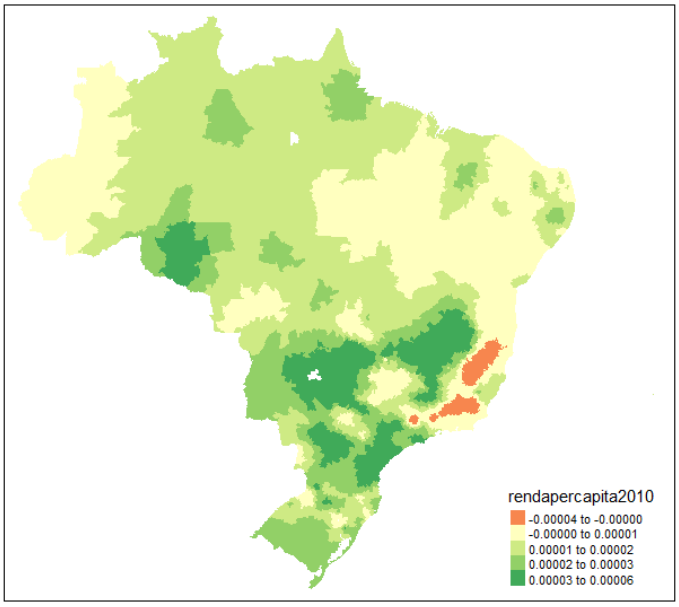
Fonte: IBGE – Elaboração própria dos autores, a partir do software GeoDa.

Analisando os bons resultados, surge a pergunta: é necessário rodar o modelo GWR? Infelizmente não foi possível manter a hipótese de homocedasticidade nos modelos globais executados. Por essa razão e por evidências de que existem relações regionais no fenômeno estudado, prosseguimos com a execução do modelo local, o GWR.

A Abordagem do modelo GWR é um pouco diferente daquela que vinha sendo adotada com os modelos Globais. Enquanto que nas regressões utilizando técnicas globais os parâmetros valem para o mapa inteiro, nos modelos locais, como o GWR, os parâmetros vão mudando ao longo do Mapa. Desse modo, as análises são diferentes. Vamos verificar que, em determinadas regiões o parâmetro renda tem mais influência, em outras a variável quantidade de deputados estaduais irá influenciar mais, ou seja, há uma questão regional a ser explorada.

Dessa forma, consegue-se visualizar o poder de explicação de cada variável ao longo do Mapa. Na (Figura 16), que apresenta o Mapa com o desempenho da variável de renda per capita (rendapercapita2010), verifica-se que nas regiões sudeste e sul a variável possui mais influência nas outras regiões do Brasil. Por outro lado, é também na região sudeste que se nota a menor influência, mais especificamente na região serrana do Rio de Janeiro e no leste do estado de Minas Gerais.

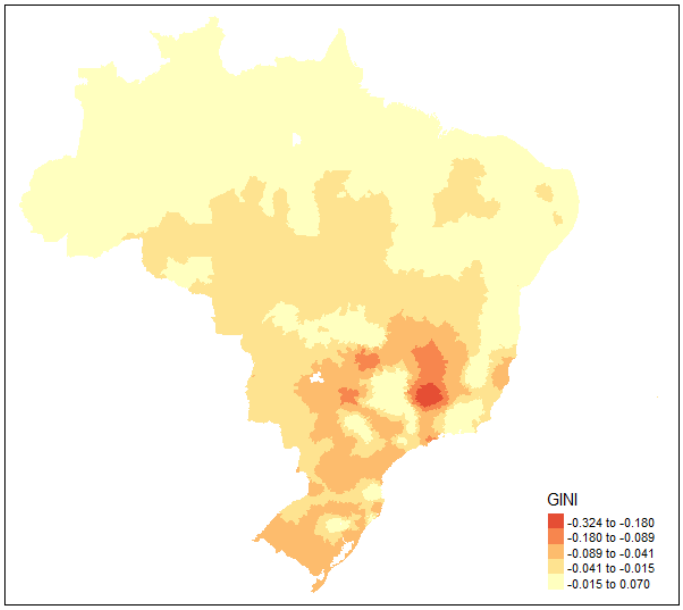
Figura 16: Mapa GWR, renda per capita.



Fonte: IBGE Elaboração própria dos autores, a partir do software R.

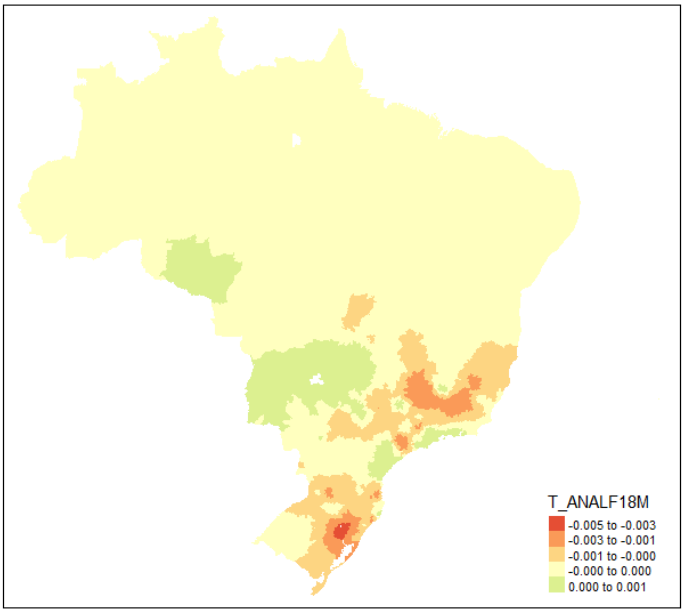
No tocante ao coeficiente de GINI (GINI) e à taxa de analfabetismo (T_ANALF_18M), as variáveis conseguem explicar melhor no Sul e Sudeste. Por outro lado, o poder de influência das variáveis diminui quando se avança em direção as regiões Norte e Nordeste, conforme verificado nas figuras (Figura 17) e (Figura 18) abaixo.

Figura 17: Mapa GWR, Índice de GINI.



Fonte: Atlas Brasil – Elaboração própria dos autores, a partir do software R.

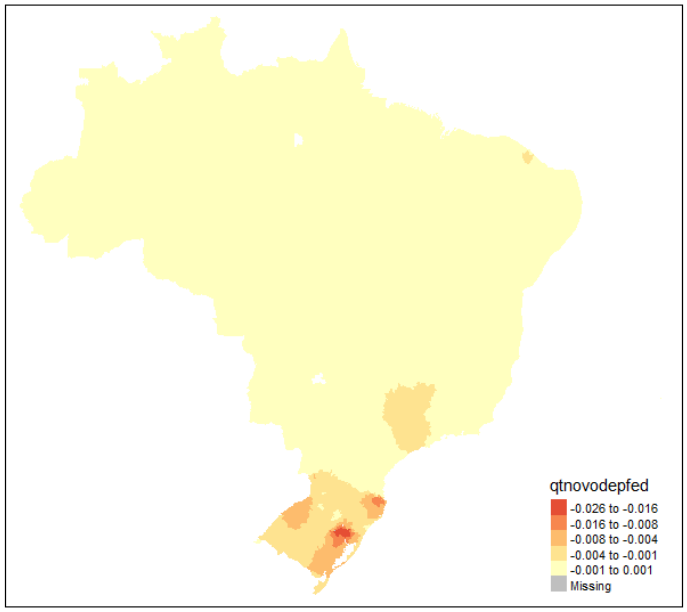
Figura 18: Mapa GWR, taxa de analfabetismo.



Fonte: Atlas Brasil – Elaboração própria dos autores, a partir do software R.

Em relação ao total de Deputados Federais candidatos pelo partido NOVO, praticamente só conseguiu ter influência no Sul do país, conforme ilustrado na (Figura 19).

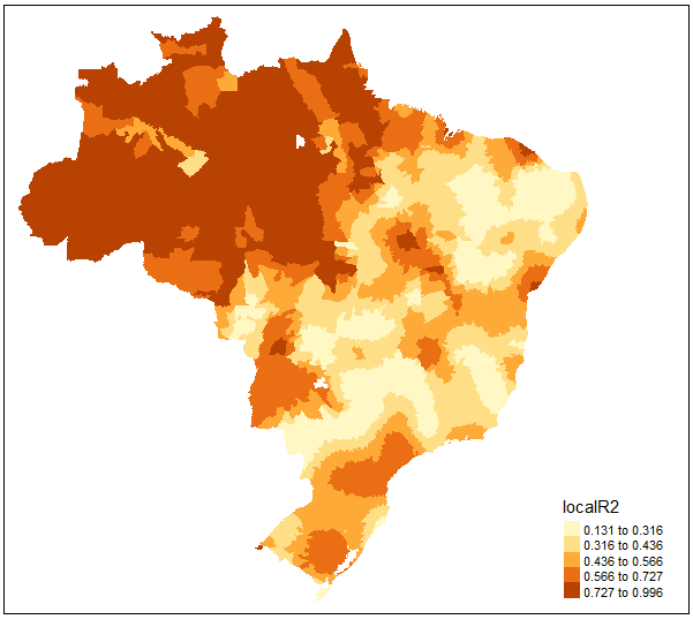
Figura 19: Mapa GWR, total Deputados Federais candidatos pelo partido NOVO.



Fonte: TSE – Elaboração própria dos autores, a partir do software R.

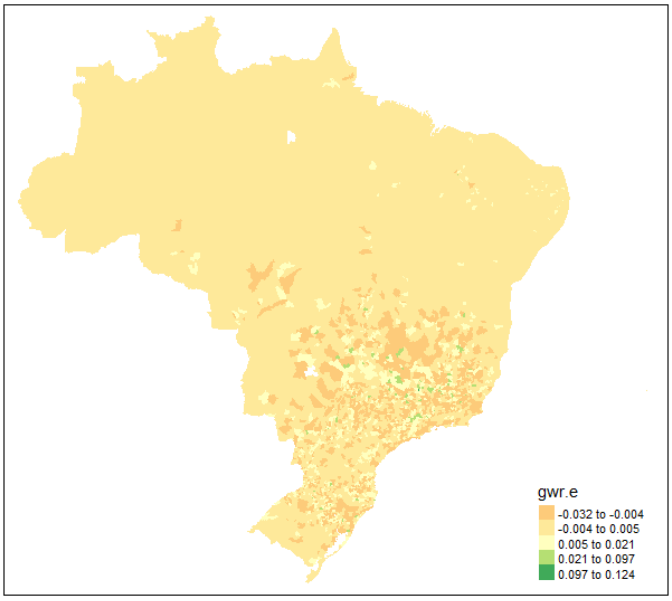
Finalmente, analisando o (R^2) ao longo do Mapa (Figura 20) verifica-se que o modelo tem um melhor desempenho na região Norte, com um valor de (R^2) variando entre 0,727 e 0,996.

Figura 20: Mapa R^2.



Fonte: Elaboração própria dos autores, a partir do software R.

Figura 21: Mapa GWR, residual Error.



Fonte: Elaboração própria dos autores, a partir do software R.

No tocante a metodologia técnica utilizada no GWR, Albuquerque (2017), destaca que as duas principais funções de ponderação encontradas na literatura são a função Gaussiana (Normal ou, em inglês, *Gaussian*) e a função Biquadrática (em inglês *Bisquare*), sendo que para cada uma delas existem dois tipos de expressões para cada uma das funções Gaussiana e Biquadrática, que se diferenciam por meio da escolha do parâmetro b (bandwidth) a ser utilizado (se fixo ou variável). Esse parâmetro é a faixa de distância ou número de vizinhos usado para cada regressor local.

Sendo assim, foram testados os modelos para as duas funções, sendo verificado o parâmetro (b) fixo e o variável para a função Gaussiana e o fixo para a função Biquadrática, que obteve pior desempenho, conforme pode ser verificado na (Tabela 4) abaixo:

Tabela 4: Comparativo de funções, GWR Gaussiano Fixo, GWR Gaussiano Variável e GWR Biquadrática Fixo.

Função	**AICC**	**AIC**	**Soma dos Quadrados Residual**	**R^2 Quasi-global**
GWR Gaussiano Fixo	-38662,5	-39197,9	0,2622115	0,7583213
GWR Gaussiano Variável	-39337,7	-39921,4	0,2288075	0,7891096
GWR Biquadrático Fixo	-36558,3	-36583,6	0,4531078	0,5823733

Fonte: Elaboração própria dos autores, a partir do software R.

É possível perceber que, para uma regressão geograficamente ponderada, a que ofereceu melhor resposta foi a que utilizou a função Gaussiana de parâmetro variável, tanto na comparação pelo critério de Akaike (AIC), quanto na comparação pelo coeficiente de determinação (R^2).

COMPARAÇÃO DOS MODELOS

Tendo-se obtido os resultados para todos os modelos, a (Tabela 5) apresenta a comparação entre seus AICs para se verificar qual o melhor modelo para explicar a variável dependente, que é percentual de votos no partido NOVO para presidente por município.

Tabela 5: Comparativo AIC, Regressão Simples, Regressão Spatial Error, Regressão Spatial Leg e Regressão GWR.

	Regressão Simples	***Regressão Spatial Error***	***Regressão Spatial Lag***	***Regressão Gwr (Gaussiana variável)***
Akaike Info Criterion	-36108.2	-38690.9	-38403.3	-39921,4

Fonte: Elaboração própria dos autores, a partir do software R.

Verifica-se que o modelo GWR, foi o que apresentou o melhor resultado segundo o critério de comparação definido.

Os parâmetros estimados pelos modelos OLS e GWR foram comparados na (Tabela 6). A variação dos coeficientes, de negativo no mínimo a positivo no máximo, mostra que, de acordo com a localidade, há variação significativa dos coeficientes. Considerando a mediana, apenas a variável "qtnovodepfed", que representa a quantidade de candidatos a deputado federal em um estado, apresentou sinal diferente do valor de coeficiente obtido para OLS.

Tabela 6: Distribuição dos coeficientes do total de Deputados Federais candidatos pelo partido NOVO, dos modelos OLS e GWR.

Variáveis	OLS	GWR Gaussiano Variável				
		Mínimo	1º quartil	Mediana	3º quartil	Máximo
Intercepto	1,66E-02	-2,54E-02	3,49E-03	1,23E-02	4,31E-02	5,13E-01
rendapercapita2010	2,11E-05	-4,22E-05	9,72E-06	1,43E-05	2,32E-05	5,76E-05
qtnovodepfed	2,73E-04	-2,60E-02	-4,92E-04	-7,48E-05	1,36E-04	1,45E-03
T_ANALF18M	-6,79E-05	-5,34E-03	-4,76E-04	-8,48E-05	1,42E-05	1,31E-03
GINI	-3,71E-02	-3,24E-01	-4,00E-02	-1,68E-02	-4,40E-03	6,95E-02

Fonte: Elaboração própria dos autores, a partir do software R.

A razão para os valores aparentemente baixos para os coeficientes é que a variável dependente também tem valor baixo, já que se trata da porcentagem de votos do partido NOVO por município, podendo ser, por isso, no máximo 1, além do fato de que o partido novo obteve apenas 2,5% dos votos totais (2,5E-02).

Realizando a interpretação dos coeficientes pelo resultado da mediana, verifica-se, primeiro, que o valor do intercepto de 1,23E-02 representa o valor esperado para a variável dependente quando todas as variáveis explicativas são 0.

Já a renda per capita, aponta que um aumento de R$ 1,00 na renda de 2010 de um município tem o efeito de aumentar o percentual de voto no novo em 1,43E-05. As outras variáveis (qtnovodepfed, T_ANALF18M e GINI), considerando a mediana, têm efeito negativo na variável dependente, sendo que a redução de uma unidade para cada uma delas, tem efeito de reduzir o percentual de votos no NOVO em -7,48E-05, -8,48E-05 e -1,68E-02 respectivamente.

CONCLUSÃO E CONSIDERAÇÕES FINAIS

O objetivo principal deste trabalho, foi verificar uma serie de dados socioeconômicos para examinar a influência destas no aspecto eleitoral brasileiro, afim de identificar características e as principais variáveis demográficas que definiram, definem ou que melhor definirão os eleitores ao candidato à presidência pelo partido NOVO a partir dos municípios com foco principal, na influência geográfica e definição da identidade do eleitor. O tema, determinantes do voto, vem tomando cada vez mais

espaço nos estudos eleitorais. Entender e mapear o mecanismo de escolha do eleitor não somente ajuda os partidos políticos na tomada de decisão, mas também, ajuda a compreender a evolução da democracia no Brasil.

Ao início do processo investigativo e seleção de variáveis, utilizando a verificação de correlação para determinantes de votos, levou-se em consideração na seleção, a correlação entre variáveis explicativas, etapa cuidadosamente imprescindível para evitar multicolinearidade. Desta maneira, foi possível chegar nas variáveis renda per capita, total de Deputados Federais candidatos pelo partido NOVO, índice de GINI e taxa de analfabetismo da população de 18 anos ou mais de idade como preditoras e determinantes na escolha do candidato pelo NOVO.

Mesmo mantendo o cuidado na seleção das variáveis, testou-se o VIF afim de verificar, ainda, a severidade da multicolinearidade. O resultado obtido seguiu conforme o esperado, nenhuma variável apresentou valor acima do máximo permitido pelo teste, indicando baixa correlação entre as variáveis explicativas. Em segundo, para verificar a existência de autocorrelação espacial, com o objetivo de considerar o espaço geográfico no modelo preditivo, foi analisado o *I'Moran* local da variável dependente, em que foi obtido uma autocorrelação positiva e significativa de 0,747, o que constatou fortes evidências para existência de um padrão espacial dos votos destinados ao partido NOVO, indicando também, a consideração desse parâmetro no modelo.

O modelo de regressão linear múltipla mostrou ser significante para explicar a variável dependente, permitindo-se realizar a análise espacial para incorporação e melhora do poder de explicação do modelo. Em síntese, todos os testes para diagnostico de dependência espacial apresentaram resultados significativos, sendo necessário a execução dos modelos Spatial Error e Spatial Lag para verificar qual dos dois modelos seria o melhor. Sendo assim, a comparação do desempenho dos modelos pelo AIC indicou o GWR como melhor modelo espacial, mais especificamente com a função Gaussiana

de parâmetro variável (-39.921,4), sendo também o melhor modelo pelo coeficiente de determinação R^2 (0,79).

Desse modo, ao levarmos em consideração o espaço geográfico, obtivemos uma melhora significativa no modelo preditivo e determinante para agremiação de votos para presidente do partido NOVO. No aspecto analítico e político, sempre há espaço para incorporação de variáveis e parâmetros ainda não identificados e não explicados pelo modelo, que possam futuramente contribuir ainda mais para a tomada de decisões e ações a serem seguidas pelo partido.

Apesar do ganho preditivo na consideração do espaço geográfico, o avanço na tecnologia e na ciência de dados estão permitindo cada vez mais a aproximação e identificação de eleitores e potenciais eleitores de um determinado candidato político, não apenas isso, estão permitindo a segmentação de eleitores por características individuais, possibilitando o engajamento de marketing indutivo pessoa a pessoa. A exemplo, um dos modelos de psicometria como o *OCEAN (Openness, Conscientiousness, Extraversion, Agreeableness and Neuraticism),* foi incorporado aos dados socioeconômicos, geográficos e dados coletados a partir de redes sociais e aplicado pela *Cambridge Analytica* nas eleições presidências americanas de 2016, sendo considerado por muitos especialistas, marco determinante no resultado final concebendo vitória ao então candidato republicando, Donald Trump (Kaiser, 2019).

BIBLIOGRAFIA

ALBUQUERQUE, Pedro Henrique Melo; MEDINA, Fabio Augusto Scalet; SILVA, Alan Ricardo da. Regressão Logística Geograficamente Ponderada Aplicada a Modelos de Credit Scoring. Rev. contab. finanç., São Paulo, v. 28, n. 73, p. 93-112, Apr. 2017. Available from <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1519-70772017000100093&lng=en&nrm=iso>. access on 30 June 2020.

ANSELIN, L., *Exploring Spatial Data with GeoDa™: A Workkbook, University of Illinois*, 2005.

ANSELIN, L., SRIDHARAN, S., & GHOLSTON, S. (2007). Using Exploratory Spatial Data Analysis to Leverage Social Indicator Databases: The Discovery of Interesting Patterns. Social Indicators Research, 82, 287-309.

BIVAND, ROGER e PIRAS, GIANFRANCO, *Comparing Implementations of Estimation Methods for Spatial Econometrics*, Journal of Statistical Software; Vol 63 (2015).

BRAGA, Maria do Socorro Sousa; PIMENTEL JR, Jairo. Os partidos políticos brasileiros realmente não importam?. Opinião Pública, v. 17, n. 2, p. 271-303, 2011.

BURNHAM, Kenneth P.; ANDERSON, David R. Multimodel inference: understanding AIC and BIC in model selection. Sociological methods & research, v. 33, n. 2, p. 261-304, 2004.

CAMPBELL, A., *The American Voter*, New York, 1960.

CARREIRAO, Y. D. S. (2002). Identificação ideológica e voto para presidente. Opinião Pública, 8(1), 54-79.

CARVALHO, Alexandre Xavier Ywata; ALBUQUERQUE, Pedro Henrique Melo. Tópicos em econometria espacial para dados cross-section. Texto para Discussão, 2010.

DE CARREIRÃO SOUZA, Yan; BARBETTA, Pedro Alberto. A eleição presidencial de 2002: a decisão do voto na região da grande São Paulo. Revista Brasileira de Ciências Sociais, v. 19, n. 56, p. 75-93, 2004.

EMILIANO, P. C., *Fundamentos e Aplicações dos Critérios de Informação: Akaike e Bayesiano*, P.41, 2009.

FIELD, A. Discovering statistics using SPSS. 3 ed. London: SAGE publications Ltd; 2009. p. 822.

FRANCISCO, Eduardo de Rezende. Indicadores de renda baseados em consumo de energia elétrica: abordagens domiciliar e regional na perspectiva da estatística espacial. 2010. Tese de Doutorado.

GUJARATI, D.N. Econometria: princípios, teoria e aplicações práticas. Tradução de Cristina Yamagami; Revisão técnica de Salvatore Benito Virgillito. São Paulo: Saraiva Educação, 2019.

KAISER, B., *Targeted: My Inside Story of Cambrige Analytica and How Trump, Brexit and Facebook Broke Democracy*, 2019.

KEY, V.O., *Southen Politics in State and Nation*, New York, 1949.

MONTGOMERY, D. C., PACK, E. A., VINING, G. G. Introduction to linear regression analysis. Johon, Wiley and Sons, Inc., New York, 612p, 2006.

O'LOUGHLIN, J. (2003). Spatial analysis in political geography. A companion to political geography, (Feb), 30-46.

PEIXOTO, Vitor; RENNÓ, Lucio. Mobilidade social ascendente e voto: as eleições presidenciais de 2010 no Brasil. Opinião Pública, v. 17, n. 2, p. 304-332, 2011.

PEREIRA, Carlos; RENNÓ, Lúcio. O que é que o reeleito tem? O retorno: o esboço de uma teoria da reeleição no Brasil. Brazilian Journal of Political Economy, v. 27, n. 4, p. 664-683, 2007.

RENNÓ, Lucio R. Escândalos e voto: as eleições presidenciais brasileiras de 2006. Opinião Pública, v. 13, n. 2, p. 260-282, 2007.

RENNÓ, Lúcio; CABELLO, Andrea. As bases do lulismo: a volta do personalismo, realinhamento ideológico ou não alinhamento?. Revista Brasileira de Ciências Sociais, v. 25, n. 74, p. 39-60, 2010.

RIBEIRO, Ednaldo; CARREIRÃO, Yan; BORBA, Julian. Sentimentos partidários e atitudes políticas entre os brasileiros. Opinião Pública, v. 17, n. 2, p. 333-368, 2011.

RIBEIRO, Ednaldo; CARREIRÃO, Yan; BORBA, Julian. Party feelings and antipetismo: constraints and covariates. Opinião Pública, v. 22, n. 3, p. 603-637, 2016.

SIEGFRIED, A., T*ableau Politique de la France de I'Ouest sous la IIIe République*. Paris, Imprimerie National, 1995 (Republicação).

SINGER, A. (1999). Esquerda e direita no eleitorado brasileiro: a identificação ideológica nas disputas presidenciais de 1989 e 1994. Edusp.

SPECK, Bruno Wilhelm; BALBACHEVSKY, Elizabeth. Identificação partidária e voto. As diferenças entre petistas e peessedebistas. Opinião Pública, v. 22, n. 3, p. 569-602, 2016.

TERRON, S. L. (2012). Geografia eleitoral em foco. Em Debate: Periódico de Opinião Pública e Conjuntura Política: ano 4, n. 2 (maio 2012).

TOBLER W., (1970) "A computer movie simulating urban growth in the Detroit region". Economic Geography, 46(Supplement): 234–240.