Dinâmicas Regionais e Indicadores Sociais no Engajamento Político Brasileiro

Autora: Giovana Marques Trindade
Orientador: Guilherme Vieira Nunes Ludwig

Departamento de Estatística, IMECC, Unicamp – Campinas

Resumo

A partir do Estudo Eleitoral Brasileiro (ESEB) de 2014, realizado pelo CESOP, foi modelada a relação entre o nível de interesse político dos eleitores e variáveis socio-econômicas. Para a análise, são utilizadas técnicas de análise exploratória, regressão logística e análise espacial de resíduos que procura identificar a presença de características regionais, não coletadas, que podem trazer confundimentos espaciais nos resultados. O modelo logístico proposto bem como os resultados sobre a influência espacial nas respostas foram utilizados para predição a nível municipal do interesse em política. É constatado que embora seja detectado um efeito espacial, sua variância interfere na qualidade da previsão.

Palavras-chaves: Regressão logística, *deviance residuals*, variograma, krigagem.

1 Introdução

As análises realizadas em "Engajamento cívico e escolaridade superior: as eleições de 2014 e o comportamento político dos brasileiros" de Dias e Kerbauy [3] serviram de incentivo ao questionamento de que a participação e as opiniões políticas estão ligadas ao nível de instrução, bem como podem ter variabilidade relacionada à questões regionais em diferentes lugares no Brasil. Para o trabalho, foram utilizadas informações fornecidas pelo ESEB [5], um *survey* nacional pós-eleitoral de cunho acadêmico que coleta informações de diversos aspectos sociais e de participação política a fim de entender as dinâmicas eleitorais. O estudo de 2014 conta com 2506 entrevistas realizadas de forma estratificada por setor censitário pelo método de probabilidade proporcional ao tamanho [8], além de 630 feitas apenas no estado de São Paulo. A seleção dos entrevistados dentro de cada setor também foi feita estratificando-se por sexo, idade, grau de escolaridade e ramo de atividade cujas proporções foram retiradas dos dados mais recentes do PNAD (2012) e TSE (2014) [7, 10].

2 Materiais e Métodos

Modelos lineares generalizados (GLM) estendem os modelos de regressão de forma a incluir casos em que as variáveis respostas são discretas ou não normais [1]. Em particular, o modelo de regressão logística é um GLM utilizado para traçar uma relação probabilística entre uma variável resposta discreta binomial, ou multinomial no caso politômico, e variáveis explicativas.

Modelos lineares generalizados têm densidade de probabilidade expressa através da forma exponencial canônica, dada por $f(y|\theta,\phi) = \exp\{\phi[y\theta - b(\theta)] + c(y,\phi)\}$, em que b e c são funções conhecidas, θ representa uma função da esperança de Y, denominada função de ligação canônica,

e ϕ é o parâmetro de dispersão. Quando a variável resposta Y segue uma distribuição Binomial, a densidade na forma canônica é dada pela Equação (1).

$$Y \sim Binomial(\mathrm{n},\mu) \rightarrow f(\mathrm{y}|\mu) = \binom{\mathrm{n}}{\mathrm{y}} \mu_i^{\mathrm{y}} (1-\mu)^{\mathrm{n-y}}$$

$$f(\mathrm{y}|\mu) = \exp\left\{ \left[\mathrm{y} ln\left(\frac{\mu}{1-\mu}\right) + \mathrm{n} ln(1-\mu) \right] + ln\binom{\mathrm{n}}{\mathrm{y}} \right\} \tag{1}$$

em que $\phi = 1$, $c(\mathrm{y},\phi) = ln\binom{\mathrm{n}}{\mathrm{y}}$ e é necessária a utilização da função de ligação "logito" dada por $\theta = ln\left(\frac{\mu}{1-\mu}\right)$ para escrever no formato canônico, como na Equação (2).

$$f(\mathrm{y}|\mu) = exp\left\{ \left[\mathrm{y}\theta - \mathrm{n} ln(1+e^{\theta}) \right] + ln\binom{\mathrm{n}}{\mathrm{y}} \right\} \tag{2}$$

Através da ligação canônica da distribuição Binomial é possível traçar um modelo linear nos parâmetros, como propõem os modelos lineares generalizados. Através da função logito, o modelo é dado pela Equação (3).

$$ln\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} \tag{3}$$

de forma que

$$\mu_i = \frac{e^{\beta_0+\beta_1 X_{1i}+...+\beta_p X_{pi}}}{1+e^{\beta_0+\beta_1 X_{1i}+...+\beta_p X_{pi}}} \tag{4}$$

portanto através da estimação dos parâmetros β_j é possível inferir dado as informações X_j a probabilidade ou média μ_i de interesse da observação i, para $i = \{1,...,n\}$ e $j = \{1,...,p\}$.

Além do objetivo de traçar um modelo probabilístico de uma variável resposta procurando por variáveis que possam a explicar, pretende-se investigar a existência de evidências de que existe mais uma fonte de variabilidade, identificada através do espaço geográfico. A análise de presença de variabilidade espacial é feita através do estudo do variograma empírico dos resíduos do modelo, que para cada distância h possível entre pontos, calcula a média dos quadrados das diferenças entre resíduos separados por esta distância.

Definimos um processo espacial como *fracamente estacionário* e *isotrópico* se $\mathbb{E}(r(s)) = m$ é constante para todo s, e

$$\gamma(h) = \frac{1}{2}\mathrm{Var}\left(r(s) - r(s+h)\right) = \tau^2 + \sigma^2 \rho(\|h\|/\phi) \tag{5}$$

é função apenas da distância $\|h\|$ entre quaisquer duas localizações arbitrárias; onde $r(s)$ é o resíduo espacialmente indexado pela coordenada s, e $(r(s_i), r(s_i+h))$ com $i = \{1,...,n\}$ correspondem aos pares de resíduos separados pela distância $\|h\|$. A Equação (5) define a função variograma segundo [2]. Os parâmetros dos modelos de variograma são o "efeito pepita" τ^2, que representa $\gamma(0)$, ou seja variância do erro puro dos resíduos; o "patamar" $\tau^2 + \sigma^2$, que representa o valor máximo de $\gamma(h)$, a variância total dos erros (erro puro mais efeito aleatório espacial), e da "dependência" ϕ, que representa o decaimento da correlação entre resíduos em função da sua distância $\|h\|$.

O variograma empírico $\hat{\gamma}$ é construido com um conjunto finito de localizações espaciais. Ele será avaliado apenas em um conjunto finito de distâncias $0 < h_1 < h_2 < \ldots < h_p < M$, onde $M = \max_{ij} \|s_i - s_j\|$. Assuma, sem perda de generalidade, que $h_1, \ldots, h_p$ são equidistantes, e defina n^*

como $n^*(h_k) = \{i, j : h_k - \delta \leq \|x_i - x_j\| < h_k + \delta\}$ para $\delta = (h_k - h_{k-1})/2$ e $\cup_{k=1}^{p} n^*(h_k) = [0, M]$, tal que $|n^*(h_k)| > 0$ para $k = 1, 2, \ldots, p$. Então temos

$$\hat{\gamma}(h_k) = \frac{1}{2|n^*(h_k)|} \sum_{(i,j) \in n^*(h_k)} (r_i - r_j)^2.$$

Ao variograma empírico pode ser ajustado um modelo de variograma, como o exponencial, esférico, Gaussiano, entre outros [2]. O ajuste do variograma pode ser utilizado para mapear as regiões de maior ou menor correlação espacial através de um processo chamado "krigagem" ou em inglês, "*kriging*", que suaviza a predição espacial ao longo de um território com base nos parâmetros ajustados da curva. Esta suavização pode ser utilizada para mensurar o grau de influência que as características espacialmente estruturadas nos dados, que não são explicadas pelas covariáveis, têm sobre a variável resposta do modelo.

3 Resultados

Inicialmente foi realizada uma preparação das informações a serem utilizadas baseada na organização dos dados. A seguir estão dispostas as variáveis do ESEB utilizadas para o estudo e a descrição ou exemplos das respostas possíveis no questionário para cada uma delas.

1. Estado 26 estados e 1 distrito federal.

2. Região 5 regiões

3. Município 194 municípios

4. Idade resposta espontânea

5. Faixa de idade 7 faixas diferentes

6. Sexo feminino ou masculino

7. Nível de escolaridade 10 níveis

8. Quanto o entrevistado acredita que o governo deveria gastar com educação "Muito mais do que gasta atualmente", "Mais do que gasta atualmente", etc. 6 níveis diferentes

9. Se o entrevistado votou no primeiro turno 8 categorias diferentes, formadas da resposta à pergunta e da justificativa

10. Se o entrevistado votou no segundo turno 8 categorias diferentes, formadas da resposta à pergunta e da justificativa

11. Quanto o entrevistado acha que seu voto influencia no que acontece no Brasil 5 níveis

12. Quão interessado em política o entrevistado se autodeclara 4 níveis

13. Quão satisfeito o entrevistado está com a qualidade de ensino 10 níveis

14. Quão satisfeito o entrevistado está quanto à oportunidade de acesso ao ensino superior 10 níveis

15. Se o entrevistado teria votado caso o voto não fosse obrigatório Sim, Não ou Talvez

16. Se o entrevistado considera que ele e sua família mudaram de classe social nos últimos 8 anos Sim ou Não

17. A que classe o entrevistado e sua família pertenciam há 8 anos 7 categorias

18. A que classe o entrevistado e sua família pertencem atualmente 7 categorias

19. Renda domiciliar Resposta espontânea

20. Faixa de renda domiciliar 10 categorias

21. Número de pessoas que residem na casa do entrevistado Resposta espontânea

As variáveis 16, 17 e 18, que dizem respeito à classe social do entrevistado e sua família foram unidas numa única variável que corresponde à classe atual. As pessoas que responderam não ter mudado de classe social nos últimos 8 anos tinham como informação faltante a resposta da variável 18. Portanto para facilitar na análise exploratória, nestes casos esta variável foi preenchida com a resposta da variável 17.

As variáveis 19, 20 e 21, foram consolidadas numa única variável cujo valor é a divisão da renda dita na resposta espontânea pelo número de pessoas que moram com o entrevistado. Nos casos em que o entrevistado respondeu não saber a renda domiciliar, ou simplesmente deixou de responder, a nova variável assumiu o valor do "teto" (valor máximo da faixa de renda à qual a renda domiciliar do entrevistado pertence, informação contida na variável 20) dividido pelo número de pessoas que moram na casa. Desta forma a renda de cada pessoa da casa é menor ou igual ao valor desta nova variável, independente de haver pessoas desempregadas ou sem fonte de renda (como crianças) na residência.

Desta forma, ao estudar a distribuição da variável renda, nota-se, como apresentado na Figura 1, que ela não é normal, possui a extremidade esquerda elevada e uma cauda longa. Por conta disto a análise de algumas medidas podem ser interferidas negativamente pelo viés representado pelo pico da distribuição e pelos *outliers*. Sendo assim a aplicação da transformação logarítmica reduz o viés da distribuição e a normaliza. A nova variável dada pelo logarítmo da renda será representada no texto por "log-renda".

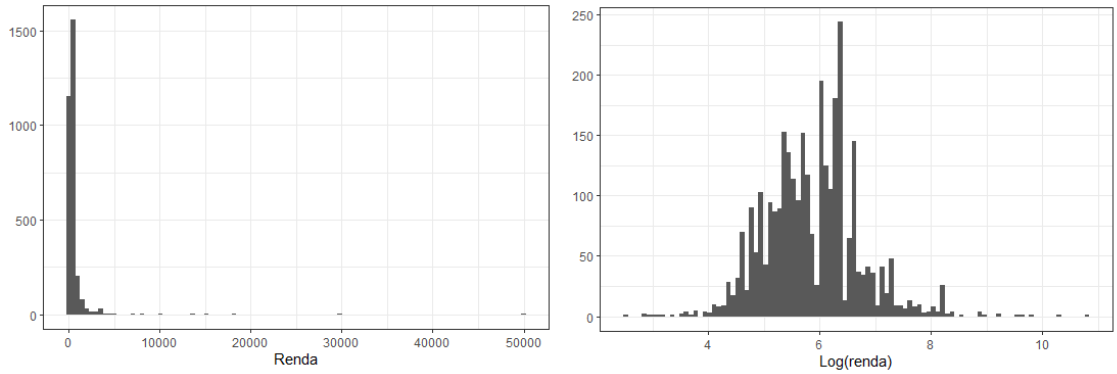

Figura 1: Distribuição da renda e do log-renda

Na Figura 2 um gráfico apresenta a relação entre o valor log-renda e os níveis de classe social.

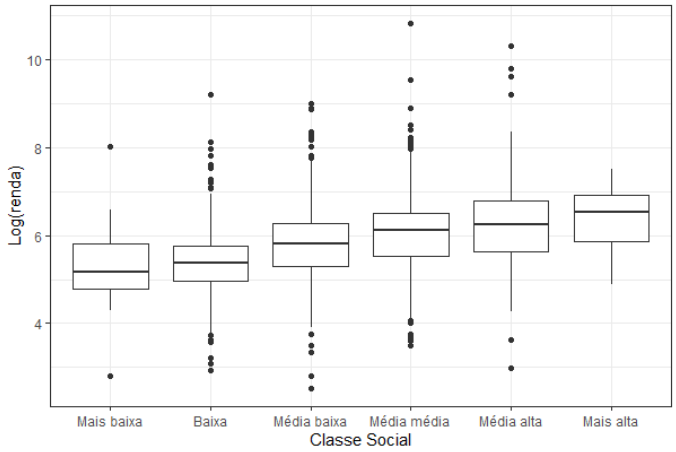

Figura 2: Boxplot da transformação logarítmica da renda por classe social

É possível notar pelos *outliers* dos *boxplots* que existe uma correção positiva entre o log-renda e os níveis de classe social entretanto não determinística devido ao fato de que ambas informações foram dadas arbitrariamente pelo entrevistado e que podem carregar incoerências motivadas pela incerteza que se tem sobre elas.

Outra variável manipulada, mas com propósito de que fosse ajustada aos requisitos da regressão logística foi a de número 12, que representa o interesse político autodeclarado. Ela possuia 4 níveis diferentes, que foram adaptados à dois níveis: pouco interessado e interessado, conforme a Tabela 1.

Tabela 1: Adaptação das respostas do interesse político

Resposta inicial	Resposta final
Nada interessado	Pouco interessado
Pouco interessado	Pouco interessado
Interessado	Interessado
Muito interessado	Interessado
Não sabe/ Não respondeu	Não sabe / Não respondeu

O modelo escolhido não conta com as respostas "Não sabe / Não respondeu" tanto das variáveis explicativas quando do interesse em política autodeclarado. 2905 observações divididas entre 2043 pouco interessados e 862 interessados o compõem.

A construção do modelo de regressão logística foi feita testando a significância das variáveis explicativas 7, 8, 11, 13, 14 e log-renda, e comparando diversos modelos cada um com combinações diferentes do uso destas variáveis. E após alguns testes o modelo escolhido utiliza a escolaridade, a oportunidade de acesso ao ensino superior, o quanto o entrevistado acha que seu voto influencia no que acontece no Brasil e o log-renda para explicar o grau de interesse político.

Foram realizados testes de significância assintóticos como sugeridos em [11], em que o p-valor assintótico é obtido usando como referência que a soma dos quadrados obtida pela inclusão da covariável no modelo dividida pela soma dos quadrados dos erros segue uma distribuição χ^2_{df}. Todas as variáveis têm relação significativa com a resposta, através dos resultados da Tabela 2.

Na tabela 3 são apresentadas as estimativas dos coeficientes multiplicadores das covariáveis. Estes coeficientes correspondem aos valores de $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{12}$ que são estimativas de máxima verossimilhança de $\beta = (\beta_0, \beta_1, ..., \beta_{12})$ e a expressão do modelo de regressão é dado como na Equação (3). Na codificação do problema por casela ou nível de referência, os coeficientes das variáveis indicadoras

Tabela 2: Tabela ANOVA sequencial do modelo com p-valor assintótico

Variável	df	Deviance	Resid. Df	Resid. Dev	P($>$ Chi)
Escolaridade	9	61.644	2895	3471.2	< 0.05
Oportunidade de acesso ao ensino superior	1	15.884	2894	3455.3	< 0.05
Influência do voto	1	16.613	2893	3438.7	< 0.05
log-renda	1	43.618	2892	3395.1	< 0.05

representam o incremento no intercepto quando o nível de escolaridade da observação é representado por esta variável. Neste caso as variáveis X_1 a X_9 são do tipo "indicadora", que conjuntamente representam o nível de escolaridade do entrevistado, tendo como nível de referência o "Analfabeto" (representado por β_0), as variáveis X_{10} e X_{11} são números inteiros que representam *scores* do quanto o entrevistado considera que teve oportunidade de acesso ao ensino superior e o quanto ele considera que seu voto influencia no que acontece no Brasil, respectivamente, e X_{12} é a variável contínua log-renda.

Tabela 3: Coeficientes do modelo

Variável	Estimativa
Nível de referência (Analfabeto)	-3.872
Até 3ª série do Ensino Fundamental	-0.002
Ensino Fundamental 1 completo	-0.016
Até 7ª série do Ensino Fundamental	-0.183
Ensino Fundamental 2 completo	-0.291
Ensino Médio incompleto	0.054
Ensino Médio completo	0.004
Ensino Universitário incompleto	0.102
Ensino Universitário completo	0.529
Graduação ou mais	0.787
Oportunidade de acesso ao ensino superior	0.056
Influência do voto	0.162
log-renda	0.345

A maioria dos valores preditos do modelo giram em torno de 20% à 40%, e isto pode ser justificado pelo fato de que a maioria das pessoas se autodeclarou nada ou pouco interessadas. Um histograma das probabilidades preditas é apresentado na Figura 3.

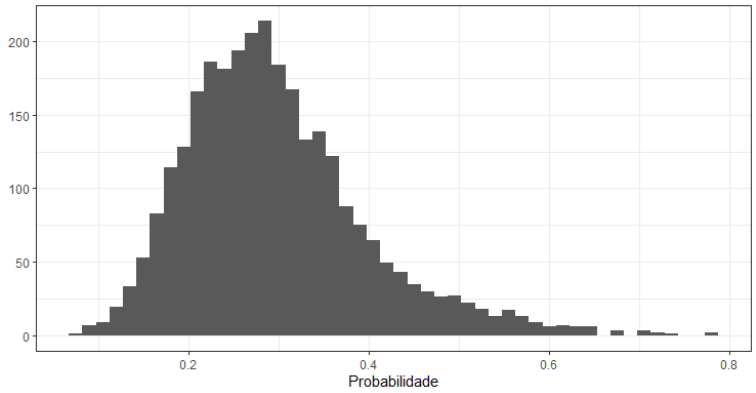

Figura 3: Distribuição dos valores preditos do interesse político autodeclarado

Na Figura 4 é apresentado a dispersão espacial em São Paulo das médias por município dos resíduos do modelo. Esta visualização é importante para análise de presença de aleatoriedade na dispersão dos resíduos.

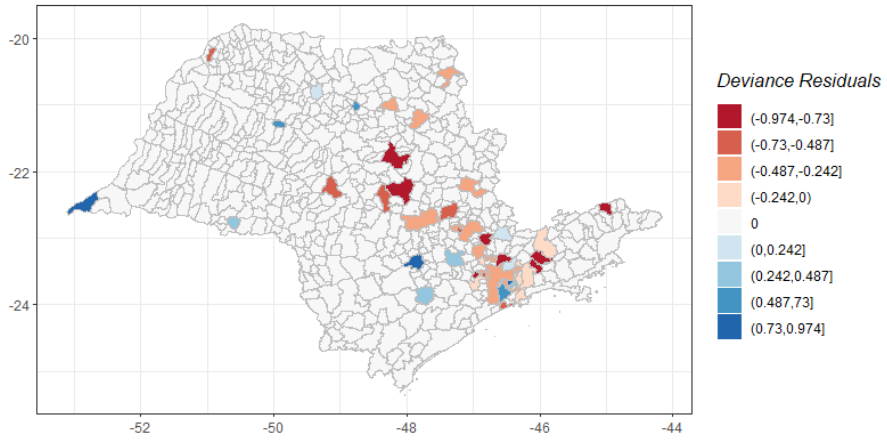

Figura 4: Dispersão espacial dos resíduos do modelo

A análise de resíduos para GLM não é tão intuitiva quanto para o modelo de regressão normal, por exemplo. E a proposta de análise espacial previamente apresentada em Materiais e Métodos foi feita sob os *deviance residuals* do modelo, que são uma função da razão de verossimilhança entre os valores reais e os estimados, dada pela equação (6) em que sgn(x) é uma função que retorna o sinal (positivo ou negativo) de x, y_i são os valores reais da variável respostas (zero ou um) e $\hat{y}_i$ são os valores ajustados

$$d_i = sgn(y_i - \hat{y}_i)\sqrt{2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(1-y_i)\log\left(\frac{1-y_i}{1-\hat{y}_i}\right)}. \tag{6}$$

quando o valor em y_i é 0, o limite do termo $2y_i \log\left(\frac{y_i}{\hat{y}_i}\right)$ existe e é 0. O mesmo acontece com o termo $2(1-y_i)\log\left(\frac{1-y_i}{1-\hat{y}_i}\right)$ quando o valor em y_i é 1.

Ao variograma empírico dos *deviance residuals* foi ajustado um modelo de variograma teórico exponencial, por mínimos quadrados, com uma curva dada pela Equação (7).

$$\hat{\gamma}(h) = 0.857 + 1.556\left(1 - \exp\left(\frac{-3h}{4.661}\right)\right), \tag{7}$$

em que 0.857 é o "efeito pepita", 0.857 + 1.556 é o "patamar" e 4.661 é a "dependência", indicando um raio de alta dependência local de aproximadamente 466km [4, 9]. A Figura 5 apresenta o variograma empírico feito com uso da distância euclidiana, que não se mostrou interferir prejudicialmente nos resultados de interesse, e sua curva de ajuste. A unidade de medida no eixo das distâncias corresponde à 100 quilômetros do território real. E a Figura 6, o "kriging" dos resíduos.

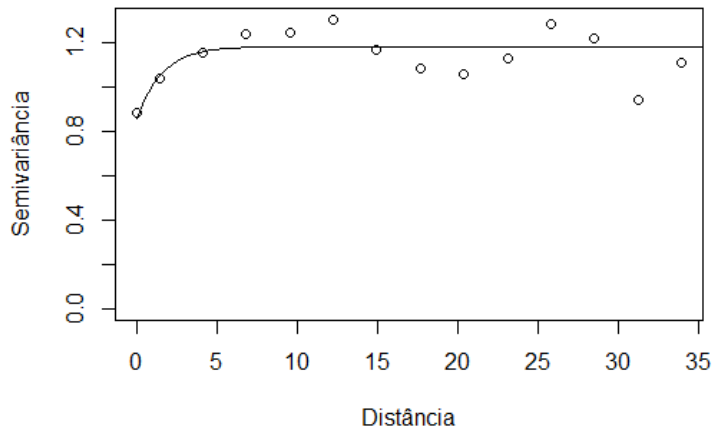

Figura 5: Variograma dos resíduos do modelo logístico utilizando todos os municípios. A distância é aproximadamente em centenas de quilômetros (projeção cartográfica do pacote `sp` [4, 9]).

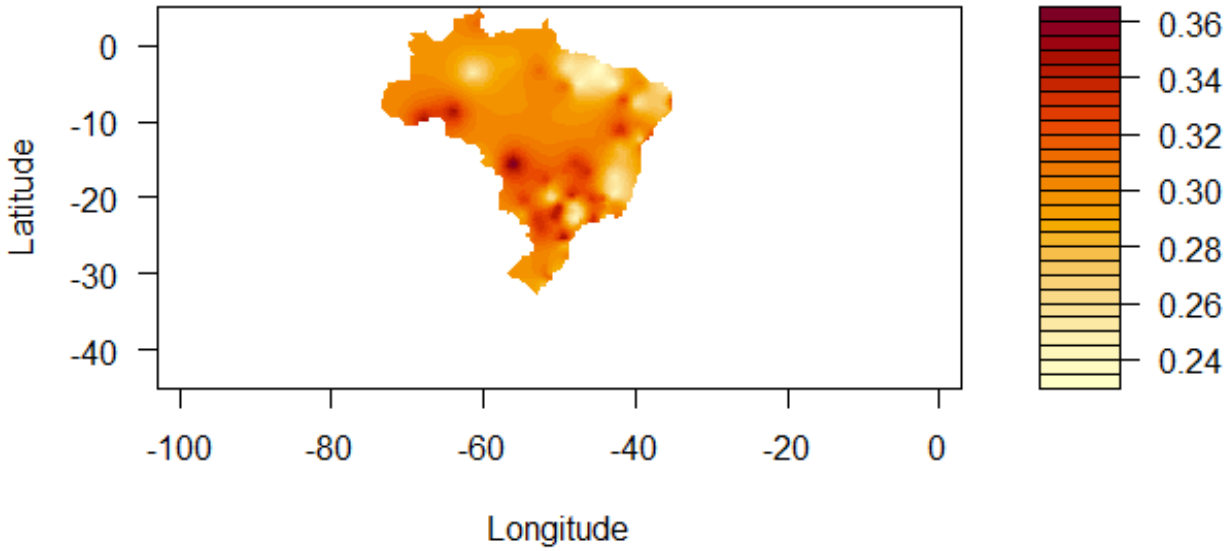

Figura 6: Predição espacial dos resíduos no território nacional

Dados obtidos nos sites do Tribunal Superior Eleitoral [10] e do Instituto Brasileiro de Geografia Estatística [7] com informações sobre escolaridade de eleitores e renda per capita municipal, foram consolidados para serem utilizados para predizer o nível de interesse em política municipal, bem como os resíduos suavizados pelo *kriging* da Figura 6. Um novo modelo feito sob os dados do ESEB com uso restrito às variáveis "escolaridade" e "log-renda", sem perda de suas significâncias quando comparado ao modelo já apresentado, foi ajustado a fim de que a predição fosse a nível municipal.

Isto foi feito levando em consideração que as variáveis "Oportunidade de acesso ao ensino superior" e "Quanto o entrevistado acredita que seu voto influencia no que acontece no Brasil" são pessoais e de difícil estimação, e que elas estejam aleatoriamente distribuídas ao longo da população de um município. Quanto aos níveis de escolaridade, algumas alterações foram feitas afim de que fosse possível uma adequação aos níveis presentes nos dados do TSE. A categoria "Graduação ou mais" que representava anteriormente 1.5% das observações utilizadas no modelo passou a ser representada pela categoria "Ensino Universitário Completo", de forma que juntas, representassem 9% das observações. O novo modelo, inclui como variável os valores preditos do resíduo, com intuito de que eles representem fontes regionais de variabilidade. Nas Tabelas 5 e 4 são apresentados os novos coeficientes dos níveis de escolaridade, do log-renda e dos resíduos, e os testes de significância assintóticos.

Tabela 4: Tabela ANOVA sequencial do modelo a nível municipal com p-valor assintótico

Variável	Df	Deviance	Resid. Df	Resid. Dev	P($>$ Chi)
Escolaridade	8	59.52	2896	3473.3	< 0.05
log-renda	1	41.417	2895	3395.1	< 0.05
Resíduos espaciais	1	12.177	2894	3395.1	< 0.05

Tabela 5: Coeficientes do modelo a nível municipal

Variável	Estimativa
Nível de referência (Analfabeto)	-2.824
Até 3ª série do Ensino Fundamental	-0.031
Ensino Fundamental 1 completo	-0.005
Até 7ª série do Ensino Fundamental	-0.201
Ensino Fundamental 2 completo	-0.338
Ensino Médio incompleto	0.017
Ensino Médio completo	-0.016
Ensino Universitário incompleto	0.063
Ensino Universitário completo	0.542
log-renda	0.301
Resíduos espaciais	0.514

Para previsão, algumas adaptações foram necessárias para resolver problemas de dados faltantes ou completamento de dados, como por exemplo:

Renda per capita As cidades cuja renda per capita consta como menor que 500 reais tiveram esta informação preenchida por 500 reais, pois nestes casos isto era uma informação faltante.

Escolaridade Foi feita uma ponderação pelo número de observações nos coeficientes dos níveis que representam o Ensino Fundamental incompleto, pois nos dados do TSE não existia distinção entre Ensino Fundamental 1 incompleto, Ensino Fundamental 1 completo e Ensino Fundamental 2 incompleto, como há nos dados do ESEB. A ponderação dos coeficientes foi feita da seguinte forma: $\hat{\beta}_{novo} = \frac{189*\hat{\beta}_1+381*\hat{\beta}_2+330*\hat{\beta}_3}{189+381+330}$ em que $\hat{\beta}_1, \hat{\beta}_2$ e $\hat{\beta}_3$ são respectivamente os coeficientes dos níveis "Até 3ª série do Ensino Fundamental", "Ensino Fundamental 1 completo" e "Até 7ª série do Ensino Fundamental", e os números que os multiplicam representam o número de observações utilizadas no modelo que pertencem a estas categorias. Sendo assim o coeficiente $\hat{\beta}_{novo}$ representa o coeficiente do nível "Ensino Fundamental incompleto".

Desta forma, dois conjuntos de valores preditos foram calculados com uso da Equação 4 a partir do modelo de nível municipal, um incluindo o efeito espacial e outro não, a fim de promover a possibilidade de comparação. A Figura 7 apresenta o território nacional colorido segundo os dois conjuntos de valores preditos. O mapa à esquerda apresenta a dispersão dos valores preditos considerando nula a variabilidade espacial identificada, e o da direita a incrementa no cálculo. Na comparação é preciso atentar-se que existe diferença nas escalas de cores dos mapas. Percebe-se que o incremento da informação que gera a variabilidade espacial suaviza os valores preditos, o que não é um resultado positivo para o estudo. Entretanto a Figura 8 apresenta a dispersão territorial do desvio padrão dos valores estimados do efeito espacial, e é possível notar que ainda nos pontos em que o desvio é mínimo, ele é alto quando comparado aos valores estimados apresentados na figura 6.

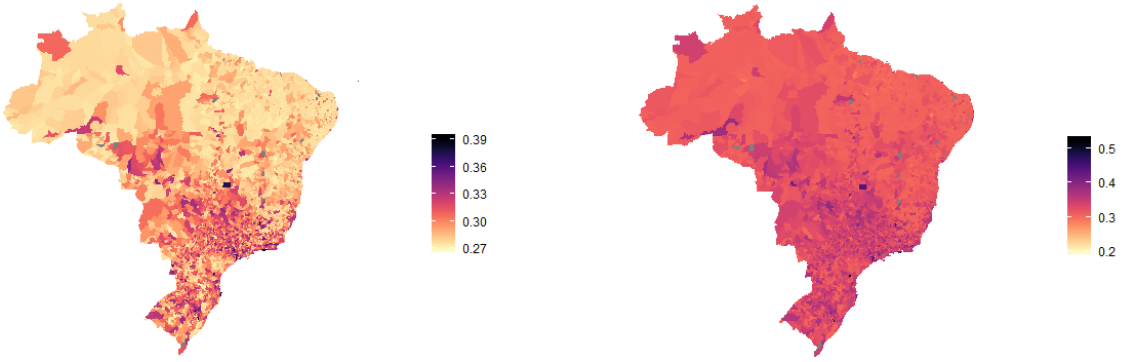

Figura 7: Dispersão territorial dos valores preditos

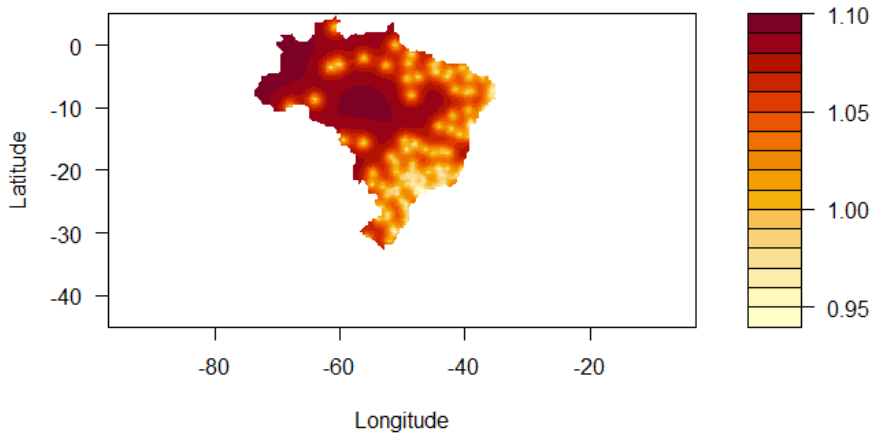

Figura 8: Dispersão territorial do desvio padrão do efeito especial

Outro resultado interessante pode ser obtido em uma região menor com maior número de dados, o que justifica o Estado de São Paulo ter sido analisado anteriormente na Figura 4. Sendo assim, também foi feito sob os resultados do Estado de São Paulo a mesma análise de variograma e krigagem feitas sob o conjunto inteiro de dados.

Para analisar os *deviance residuals* desta parcela das informações surgiu a dúvida se bastaria filtra-los do modelo original ou modelar novamente a partir da mesma função utilizada anteriormente

mas agora apenas sob as 1056 observações do Estado e analisar seus resíduos. Os dois conjuntos de resíduos foram obtidos e comparados através do gráfico da Figura 9. Concluí-se que a variabilidade é pequena e a diferença entre eles não impactaria os resultados do procedimento metodológico. Sendo assim, escolheu-se utilizar os resíduos de São Paulo do modelo original.

Figura 9: Dispersão dos dois conjuntos de resíduos

O variograma dos *deviance residuals* de São Paulo foi ajustado por uma curva gaussiana dada pela Equação 8.

$$\hat{\gamma}(h) = 0.521 + 3.475\left(1 - \exp\left(\frac{-3h^2}{6.015^2}\right)\right) \tag{8}$$

Em que 0.521 é o "efeito pepita", 0.521 + 3.475 é o "patamar" e 6.015 é a "dependência", indicando um raio de alta dependência local de aproximadamente 600km [4,9]. Note que a dependência usando apenas os dados de São Paulo parece ser maior, mas o tamanho amostral efetivo é menor e consequentemente a incerteza sobre a dependência também é maior. A Figura 10 apresenta o variograma dos resíduos de São Paulo e sua curva de ajuste. E a Figura 11, o mapa do efeito espacial predito.

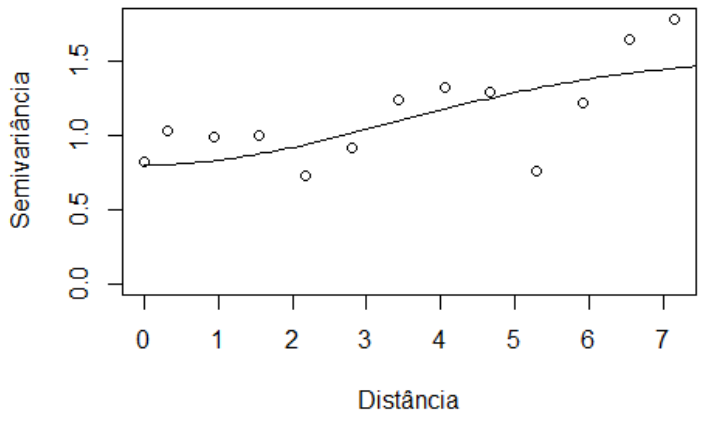

Figura 10: Variograma dos resíduos de São Paulo do modelo logístico. A distância é aproximadamente em centenas de quilômetros (projeção cartográfica do pacote `sp` [4,9]).

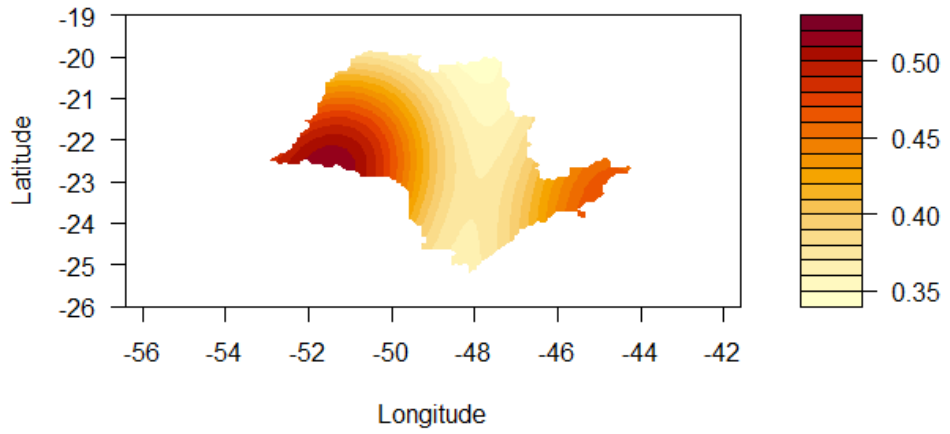

Figura 11: Predição espacial dos resíduos no território do Estado de São Paulo

O cálculo dos valores preditos incluindo e não incluindo o efeito espacial foi feito da mesma forma como para o conjunto completo de dados. A Figura 12 apresenta os mapas de valores preditos sem inclusão do efeito espacial, à esquerda, e com, à direita. E a Figura 13, a dispersão dos desvios padrões. É possível notar a mesma característica identificada na comparação feita na Figura 7, há uma alta suavização que pode ser explicada pelos altos desvios padrões identificados na Figura 13.

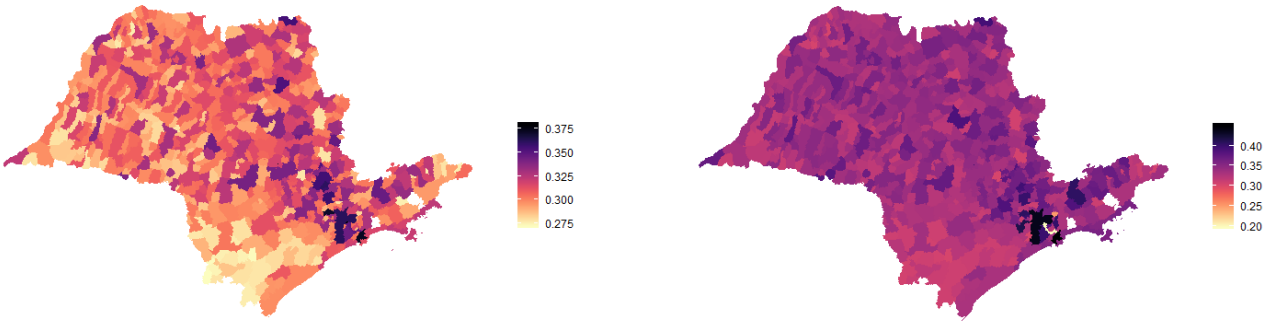

Figura 12: Dispersão territorial dos valores preditos

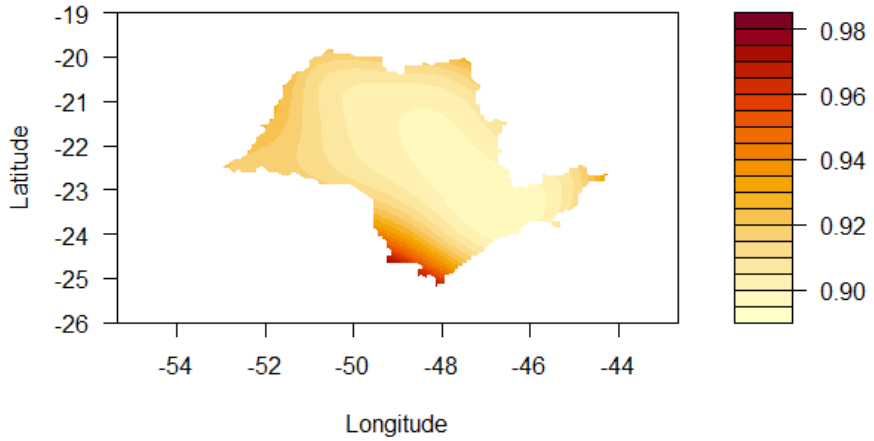

Figura 13: Dispersão territorial do desvio padrão do efeito especial

4 Discussão

Os resultados do modelo de regressão logística apontam a existência de uma influência positiva da escolaridade e da renda no interesse político autodeclarado, de forma que quanto maiores o nível de instrução e a renda, maior a probabilidade de que o indivíduo afirme seu interesse em política. A significância da covariável escolaridade é dada através dos maiores níveis de instrução, ou seja, existe uma discrepância maior de probabilidade de afirmação de interesse em política entre pessoas que concluíram o ensino universitário e as demais. Além disto as variáveis "oportunidade de acesso ao ensino superior" e "qual a influência do voto no que acontece no Brasil" também têm uma relação positiva com a variável resposta.

A metodologia proposta pode ser mais facilmente aplicada a modelos de regressão normais pelo fato de que neles os resíduos são expressos linearmente através da curva de ajuste, e após identificado um efeito espacial, os resíduos podem ser decompostos aditivamente em dois componentes, um que represente a variabilidade espacial agora conhecida, e outro que represente o erro puro restante. Desta forma, ao passo que se reconhece a existência de uma fonte de variabilidade, identificada e representada no espaço geográfico, diminui-se a ignorância quanto aos resultados que podem ser obtidos através do modelo. No caso logístico, os resíduos são função da razão de verossimilhança entre os valores reais e os ajustados, e não compõem a curva de regressão. Desta forma, a alternativa tomada foi a de incrementar a suavização dos resíduos ao longo da malha territorial como uma nova variável explicativa do modelo, e dada sua alta significância, utilizar os coeficientes estimados para a previsão de resultados. Entretanto este procedimento, bem como a alta variância do efeito espacial, podem ter influenciado para uma alta suavização de valores preditos.

Outra constatação importante é a de que incluir o efeito espacial obtido dos resíduos de um modelo linear não altera significativamente os coeficientes de regressão. Isto evita o problema de *confundimento espacial* que acontece quando uma covariável significativa representa a mesma fonte de variabilidade identificada espacialmente, que existiria em uma modelagem simultânea dos efeitos [6]. O exame de confundimentos (espaciais ou multicolinearidade) e investigação de relações causais é de interesse para pesquisa futura.

Referências

[1] A. Agresti. *An Introduction to Categorical Data Analysis*. John Wiley & Sons, 2018.

[2] R. S. Bivand, E. J. Pebesma e V. Gómez-Rubio. *Applied Spatial Data Analysis with R*. Springer, 2008.

[3] A. L. V. Dias e M. T. M. Kerbauy. "Engajamento cívico e escolaridade superior: as eleições de 2014 e o comportamento político dos brasileiros." *Revista de Sociologia e Política*, pp. 149–181, 2015.

[4] Federal Communications Commission. "Reference points and distance computations" *Code of Federal Regulations (Annual Edition). Title 47: Telecommunication*. 73 (208), 2016. URL: `https://www.govinfo.gov/content/pkg/CFR-2016-title47-vol4/pdf/CFR-2016-title47-vol4-sec73-208.pdf`.

[5] R. Meneghello *et al.* Estudo Eleitoral Brasileiro. URL: `https://www.cesop.unicamp.br/por/eseb`

[6] J. S. Hodges e B. J. Reich. "Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love." *The American Statistician*, pp. 325–334, 2010.

[7] Instituto Brasileiro de Geografia Estatística. *Pesquisa Nacional por Amostra de Domicílios.* URL: `https://www.ibge.gov.br/estatisticas/sociais/educacao/9127-pesquisa-nacional-por-amostra-de-domicilios.html?edicao=18338&t=publicacoes`

[8] S. L. Lohr. *Sampling: Design and Analysis*. Nelson Education, 2009.

[9] E. J. Pebesma e R. S. Bivand. "Classes and methods for spatial data in R." *R News*, 5(2), 9–13. URL: `https://cran.r-project.org/web/packages/sp/index.html`

[10] Tribunal Superior Eleitoral. URL: `http://www.tse.jus.br/`.

[11] W. N. Venables e B. D. Ripley. *Modern Applied Statistics with S-PLUS*. Springer, 2013.