



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
DOUTORADO EM ESTATÍSTICA

YURI ALVES DE ARAUJO

MODELOS DE REGRESSÃO ELÍPTICOS GEOGRAFICAMENTE
PONDERADOS

Recife

2021

YURI ALVES DE ARAUJO

**MODELOS DE REGRESSÃO ELÍPTICOS GEOGRAFICAMENTE
PONDERADOS**

Tese apresentada ao Curso de Doutorado em Estatística do Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de doutor em Estatística.
Área de Concentração: Estatística Aplicada

Orientador: Prof. Dr. Francisco José de Azevêdo Cysneiros

Coorientadora: Profa. Dra. Audrey Helen Mariz de Aquino
Cysneiros

Recife

2021

Catálogo na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

A663m Araújo, Yuri Alves de
 Modelos de regressão elípticos geograficamente ponderados / Yuri Alves de
 Araújo. – 2021.
 155 f.: il., fig., tab.

 Orientador: Francisco José de Azêvedo Cysneiros.
 Tese (Doutorado) – Universidade Federal de Pernambuco. CCEN,
 Estatística, Recife, 2021.
 Inclui referências.

 1. Estatística aplicada. 2. Modelos de regressão. I. Cysneiros, Francisco
 José de Azêvedo (orientador). II. Título.

310

CDD (23. ed.)

UFPE- CCEN 2021 - 118

YURI ALVES DE ARAÚJO

Modelos de Regressão Elípticos Geograficamente Ponderados

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Estatística.

Aprovada em: 31 de maio de 2021.

BANCA EXAMINADORA

Professor Dr. Francisco José de Azevedo Cysneiros
UFPE

Professora Dr^a Francielle de Lima Medina
UFPE

Professora Dr^a Fernanda De Bastiani
UFPE

Professor Dr. Gilberto A. Paula
USP

Professor Dr. Alan Ricardo da Silva
UNB

Dedico esse trabalho à minha família, por me apoiar em todo momento e por ser um porto seguro ao qual eu posso retornar sempre. Em memória ao meu pai Wanderley que não pode estar comigo neste momento, mas que sempre foi uma grande fonte de inspiração e força.

AGRADECIMENTOS

Gostaria de agradecer primeiramente aos meus pais Elianice e Wanderley que sempre estiveram ao meu lado me apoiando ao longo de toda a minha trajetória. Todo esse tempo eles acreditaram na minha capacidade e dedicação, isso só me fortaleceu ao longo dos anos. Graças a vocês consegui conquistar meus sonhos e chegar aonde estou hoje. Sou profundamente grato pelo apoio e amor incondicional que sempre me deram durante toda a minha vida.

Ao meu irmão Ygor e à toda minha família, a qual tenho o mais sincero respeito, carinho e admiração. Obrigado pela força de todos e por sempre me ajudarem a alcançar os meus sonhos. Vocês tem grande peso em minha vida.

Meu agradecimento mais profundo a Rommy e nossos seis filhos de quatro patas, Link, Lilith, Django, Lucy, Mary e Azazel. Todo esse tempo estiveram ao meu lado, até nos momentos mais difíceis, que não foram raros nestes últimos anos. Este período nos mostrou a verdade sobre nosso relacionamento, somos uma família. Cada gesto carinhoso, sorriso, mordida e miado durante essa fase da minha vida foram extremamente gratificante.

Aos meus orientadores Francisco José de Azevêdo Cysneiros e Audrey Helen Mariz de Aquino Cysneiros por sempre estarem presentes para indicar a direção correta que o trabalho deveria tomar. Deixo um agradecimento especial pela dedicação e disponibilidade de tempo, os quais foram fundamentais para a conclusão desta pesquisa. Por todo esse tempo sempre estiveram dispostos a ajudar nos momentos de dificuldade, suas orientações e conselhos pessoais foram fundamentais para superar os obstáculos que a vida me apresentou.

À Katia e toda sua família, pelo carinho incondicional e os cuidados nos momentos em que mais precisei. Sou extremamente grato por todo apoio e paciência que tiveram comigo ao longo desses anos.

Agradeço ao prof. Juvêncio Santos Nobre, por todo apoio e orientação ao longo da minha graduação, sempre me motivou a ingressar na pós-graduação e crescer profissionalmente. Meu muito obrigado!

Agradeço também a todos os professores do CCEN, e em especial aos Professores Alex Dias Ramos, Audrey Helen Mariz de Aquino Cysneiros, Francisco José de Azevêdo Cysneiros, Aldo William Medina Garay, Francielle de Lima Medina e Gauss Moutinho Cordeiro. Sem essa grande equipe de profissionais o meu esforço não teria se concretizado, vocês contribuíram imensamente para minha formação acadêmica.

Aos meus amigos de doutorado que compartilharam dos inúmeros desafios que enfrentamos, sempre prestativos e com um grande espírito colaborativo. Agradeço especialmente à Vinícius, César, Jucelino e Jodavid pelo companheirismo durante toda essa fase da minha vida e por todos os momentos divididos juntos. Espero que nossas amizades durem por muito tempo e que sempre tenhamos boas histórias para recordar. Foi muito bom poder contar com todos vocês!

A todos meus amigos de longa data que sempre estiveram ao meu lado, pela amizade incondicional e pelo apoio demonstrado ao longo de todo o período de tempo em que me dediquei a este trabalho. Um agradecimento especial à Pedro Janssen (PJ), Ramirez, Paulo Rogério (Chinês), Brendon e Ernand, por só quererem o meu bem e me valorizarem como pessoa e amigo. Obrigado pela amizade de todas as horas, eu sei sempre posso contar com vocês.

Aos servidores da Universidade Federal de Pernambuco (UFPE), principalmente a Valéria Bittencourt e Michelle Quemel, pela paciência e cuidado ao longo desses quatro anos de doutorado.

Quero agradecer também à UFPE por me oferecer ensino de qualidade e todo o suporte necessário para a conclusão dessa tese e ao apoio financeiro da FACEPE durante todo o desenvolvimento deste trabalho.

Por fim, gostaria de agradecer à todas as pessoas que contribuíram diretamente e indiretamente para meu sucesso e crescimento pessoal e profissional.

RESUMO

O uso dos modelos de regressão clássicos no contexto de dados espaciais podem levar a interpretações errôneas, pois este não leva em consideração as informações geoespaciais. Neste caso, a aplicação de modelos geograficamente ponderados para o estudo da não estacionariedade espacial dos parâmetros de regressão é uma alternativa bastante atrativa. No entanto, esses modelos costumam ser sensíveis na presença de valores discrepantes. Neste trabalho, desenvolvemos extensões da regressão geograficamente ponderada utilizando a família de distribuição elíptica para a suposição dos erros aleatórios. O uso desta família de distribuição permite ajustar um modelo menos sensível na presença de observações discrepantes, uma vez que abrange distribuições com caudas mais pesadas que a normal. Em específico, consideramos o uso das distribuições normal, t -Student, exponencial potência, Cauchy, logística do tipo I e II, t -Student generalizada, logística generalizada e normal contaminada. Métodos com base nos algoritmos score de Fisher e *backfitting* são propostos para estimar os parâmetros dos modelos. Além disso, descrevemos análises de resíduos e medidas diagnósticas com base na abordagem de influência global e local. Estudos de simulação sob diferentes cenários e aplicações práticas com dois conjuntos de dados reais são apresentados para avaliar o desempenho desses modelos na presença de observações discrepantes. A primeira aplicação refere-se ao estudo do desenvolvimento humano e da criminalidade nos municípios do Ceará, Brasil. A segunda aplicação aborda o desempenho educacional nos condados da Geórgia, EUA. Por fim, um pacote em R foi desenvolvido para disponibilizar a aplicação prática dos resultados teóricos obtidos nesta tese.

Palavras-chave: Análise espacial. Distribuições elípticas. Medidas de diagnóstico. Regressão geograficamente ponderada.

ABSTRACT

The use of the classic regression model in spatial data context may lead to misinterpretations because it do not consider geospatial information. In this instance, the application of geographically weighted models to study spatial nonstationarity of the regression parameters it is a very attractive alternative. However, these models are often sensitive in a presence of outliers. In this work, we develop extensions of the geographically weighted regression using the elliptical distribution family for the assumption of random errors. The use of this distribution family allows to fit a less sensitive model in the presence of outliers, since it covers distributions with heavier tails than normal. Specifically, we consider the use of the normal, t -Student, power exponential, Cauchy, types I and II logistics, generalized t -Student, generalized logistic, and contaminated normal distributions. Methods based on the Fisher score and backfitting algorithms are proposed to estimate the models parameters. In addition, we describe residuals analysis and diagnostic measures based on the global and local influence approach. Simulation studies under different scenarios and practical applications with two real datasets are presented to evaluate the performance of these models in the presence of outliers. The first application refers to the study of human development and crime in the municipalities of Ceará, Brazil. The second application addresses educational performance in the counties of Georgia, USA. A package in R was developed for the practical application of the theoretical results obtained in this thesis.

Keywords: Elliptical distribution. Diagnostic measures. Geographically weighted regression. Spatial analysis.

LISTA DE FIGURAS

Figura 1 – Mapa dos municípios do Ceará utilizados no ajuste dos modelos locais para os municípios “A” e “B”	30
Figura 2 – Função de densidade para algumas distribuições elípticas	34
Figura 3 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 da RGP (meio) e REGP com erros t -Student (direita) para o método CV.	54
Figura 4 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 da RGP (meio) e REGP com erros t -Student (direita) para o método CV.	54
Figura 5 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 da RGP (meio) e REGP com erros t -Student (direita) para o método AICc.	55
Figura 6 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 da RGP (meio) e REGP com erros t -Student (direita) para o método AICc.	55
Figura 7 – <i>Boxplots</i> da REQM para β_0 nos modelos geograficamente ponderados para os métodos CV (esquerda) e AICc (direita)	56
Figura 8 – <i>Boxplots</i> da REQM para β_1 nos modelos geograficamente ponderados para os métodos CV (esquerda) e AICc (direita)	56
Figura 9 – <i>Boxplots</i> da SQRp dos modelos geograficamente ponderados para os métodos CV (esquerda) e AICc (direita)	57
Figura 10 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 do RGP (meio) e REGP com erros t -Student (direita) para o método CV.	58
Figura 11 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 do RGP (meio) e REGP com erros t -Student (direita) para o método CV.	58
Figura 12 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 do RGP (meio) e REGP com erros t -Student (direita) para o método AICc.	59

Figura 13 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 do RGP (meio) e REGP com erros t -Student (direita) para o método AICc.	59
Figura 14 – <i>Boxplots</i> da REQM dos modelos geograficamente ponderados para os métodos CV (esquerda) e AICc (direita) na superfície de β_0	60
Figura 15 – <i>Boxplots</i> da REQM dos modelos geograficamente ponderados para os métodos CV (esquerda) e AICc (direita) na superfície de β_1	61
Figura 16 – <i>Boxplots</i> da SQRp dos modelos geograficamente ponderados para os métodos CV (esquerda) e AICc (direita)	61
Figura 17 – Histograma com densidade aproximada pela linha vermelha (esquerda) e <i>boxplot</i> (direita) da RPC	62
Figura 18 – Distribuição espacial (esquerda) e diagrama de Moran (direita) da RPC nos municípios cearenses	63
Figura 19 – Distribuição espacial (esquerda) e diagrama de Moran (direita) da TH nos municípios cearenses	64
Figura 20 – Distribuição espacial (esquerda) e diagrama de Moran (direita) da PCF nos municípios cearenses	64
Figura 21 – RPC versus TH (esquerda) e RPC versus PCF (direita) . . .	65
Figura 22 – Gráfico de quantis-quantis com envelope simulados dos resíduos padronizados para o modelo de regressão elíptico com erros normais (esquerda) e t -Student (direita)	66
Figura 23 – Diagramas de espalhamento de Moran dos resíduos padronizados para o modelo de regressão elíptico com erros normais (esquerda) e t -Student (direita)	66
Figura 24 – Gráfico de quantis-quantis com envelope simulados dos resíduos padronizados para o modelo de REGP com erros normais (esquerda) e t -Student (direita)	68
Figura 25 – Diagramas de espalhamento de Moran dos resíduos padronizados para o modelo de REGP com erros normais (esquerda) e t -Student (direita)	69

Figura 26 – Gráfico de alavancagem generalizada nos modelos de REGP com erros normais (esquerda) e t -Student (direita)	69
Figura 27 – Gráfico de influência local para perturbação na escala nos modelos REGP com erros normais (esquerda) e t -Student (direita)	70
Figura 28 – Gráfico de influência local para perturbação ponderação de casos nos modelos REGP com erros normais (esquerda) e t -Student (direita)	70
Figura 29 – Distribuição espacial das estimativas de β_0 considerando os dados completos para os modelos REGP com erros normais (esquerda) e t -Student (direita)	71
Figura 30 – Distribuição espacial das estimativas de β_0 considerando os dados após a exclusão de Fortaleza para os modelos REGP com erros normais (esquerda) e t -Student (direita)	71
Figura 31 – Distribuição espacial das estimativas de β_1 considerando os dados completos para os modelos REGP com erros normais (esquerda) e t -Student (direita)	72
Figura 32 – Distribuição espacial das estimativas de β_1 considerando os dados após a exclusão de Fortaleza para os modelos REGP com erros normais (esquerda) e t -Student (direita)	72
Figura 33 – Distribuição espacial das estimativas de β_2 considerando os dados completos para os modelos REGP com erros normais (esquerda) e t -Student (direita)	72
Figura 34 – Distribuição espacial das estimativas de β_2 considerando os dados após a exclusão de Fortaleza para os modelos REGP com erros normais (esquerda) e t -Student (direita)	73
Figura 35 – Distribuição espacial das estimativas significativas de β_0 (esquerda), β_1 (direita) e β_2 (abaixo) no modelo REGP com erros t -Student	74
Figura 36 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 da RGP (meio) e REGP com erros t -Student (direita).	91

Figura 37 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 da RGP (meio) e REGP com erros t -Student (direita).	91
Figura 38 – <i>Boxplots</i> dos valores ótimos para os parâmetros de suavização nos modelos de REGP e REGPM via validação cruzada .	92
Figura 39 – <i>Boxplots</i> da REQM (esquerda) e SQRp (direita) para os modelos de REGP e REGPM	92
Figura 40 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 da RGP (meio) e REGP com erros t -Student (direita).	94
Figura 41 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 da RGP (meio) e REGP com erros t -Student (direita).	94
Figura 42 – Superfície espacial verdadeira de β_2 (esquerda), superfícies espaciais médias das estimativas de β_2 da RGP (meio) e REGP com erros t -Student (direita).	95
Figura 43 – <i>Boxplots</i> dos valores ótimos para os parâmetros de suavização nos modelos de REGP e REGPM via validação cruzada .	95
Figura 44 – <i>Boxplots</i> da REQM (esquerda) e SQRp (direita) para os modelos de REGP e REGPM	96
Figura 45 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 da REGP (meio) e REGPM (direita).	98
Figura 46 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 da REGP (meio) e REGPM (direita).	98
Figura 47 – Superfície espacial verdadeira de β_2 (esquerda), superfícies espaciais médias das estimativas de β_2 da REGP (meio) e REGPM (direita).	98
Figura 48 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 da RGPM (meio) e REGPM (direita).	99

Figura 49 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 da RGPM (meio) e REGPM (direita).	99
Figura 50 – Superfície espacial verdadeira de β_2 (esquerda), superfícies espaciais médias das estimativas de β_2 da RGPM (meio) e REGPM (direita).	99
Figura 51 – <i>Boxplots</i> dos valores ótimos para os parâmetros de suavização nos modelos geograficamente ponderados com erros normais (esquerda) e <i>t</i> -Student (direita) via validação cruzada. .	100
Figura 52 – <i>Boxplots</i> da REQm para o modelo de REGPM em comparação com os modelos de RGPM (esquerda) e REGP com erros <i>t</i> -Student (direita).	101
Figura 53 – <i>Boxplots</i> da SQRp para o modelo de REGPM em comparação com os modelos de RGPM (esquerda) e REGP com erros <i>t</i> -Student (direita).	101
Figura 54 – Histograma com densidade aproximada pela linha vermelha (esquerda) e boxplot (direita) do <i>logit</i> da porcentagem da população com diploma de bacharel	103
Figura 55 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para o <i>logit</i> da porcentagem da população com diploma de bacharel nos condados da Geórgia	104
Figura 56 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para a população total em 1990 nos condados da Geórgia	104
Figura 57 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para a porcentagem da população rural nos condados da Geórgia	105
Figura 58 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para a porcentagem da população com 65 anos ou mais nos condados da Geórgia	105
Figura 59 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para a porcentagem da população nascida fora dos EUA nos condados da Geórgia	105

Figura 60 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para a porcentagem da população que vive abaixo da linha da pobreza nos condados da Geórgia	106
Figura 61 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para a porcentagem da população negra nos condados da Geórgia	106
Figura 62 – Gráficos de dispersão e valores para a correlação de Pearson das variáveis em estudo	107
Figura 63 – Gráfico de quantis-quantis com envelope simulado dos resíduos padronizados para o modelo de regressão elíptico com erros normais (esquerda) e <i>t</i> -Student (direita)	108
Figura 64 – Diagramas de espalhamento de Moran dos resíduos padronizados para o modelo de regressão elíptico com erros normais (esquerda) e <i>t</i> -Student (direita)	108
Figura 65 – Gráfico de quantis-quantis com envelope simulado dos resíduos padronizados para o modelo de regressão elíptico com distribuições normal (esquerda) e <i>t</i> -Student (direita)	110
Figura 66 – Diagramas de espalhamento de Moran dos resíduos padronizados para o modelo de regressão elíptico com distribuições normal (esquerda) e <i>t</i> -Student (direita)	110
Figura 67 – Resíduos padronizados versus valores ajustados para o modelo de REGPM com distribuições normal (esquerda) e <i>t</i> -Student (direita)	113
Figura 68 – Diagramas de espalhamento de Moran dos resíduos quantílicos para o modelo de REGPM com distribuições normal (esquerda) e <i>t</i> -Student (direita)	113
Figura 69 – Gráfico das medidas de alavancagem nos modelos de REGP com erros normais (esquerda) e <i>t</i> -Student (direita)	114
Figura 70 – Gráfico de influência local para perturbação na escala associado ao parâmetro β_1 nos modelos REGP com erros normais (esquerda) e <i>t</i> -Student (direita)	115

Figura 71 – Gráfico de influência local para perturbação ponderação de casos associado ao parâmetro β_1 nos modelos REGP com erros normais (esquerda) e t -Student (direita)	115
Figura 72 – Gráfico de influência local para perturbação na escala associado ao parâmetro β_2 nos modelos REGP com erros normais (esquerda) e t -Student (direita)	115
Figura 73 – Gráfico de influência local para perturbação ponderação de casos associado ao parâmetro β_2 nos modelos REGP com erros normais (esquerda) e t -Student (direita)	116
Figura 74 – Gráfico de influência local para perturbação na escala associado ao parâmetro β_3 nos modelos REGP com erros normais (esquerda) e t -Student (direita)	116
Figura 75 – Gráfico de influência local para perturbação ponderação de casos associado ao parâmetro β_3 nos modelos REGP com erros normais (esquerda) e t -Student (direita)	116
Figura 76 – Gráfico de influência local para perturbação na escala associado ao parâmetro β_4 nos modelos REGP com erros normais (esquerda) e t -Student (direita)	117
Figura 77 – Gráfico de influência local para perturbação ponderação de casos associado ao parâmetro β_4 nos modelos REGP com erros normais (esquerda) e t -Student (direita)	117
Figura 78 – Distribuição espacial das estimativas de β_0 considerando os dados completos para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	118
Figura 79 – Distribuição espacial das estimativas de β_0 considerando os dados após a exclusão de Bulloch para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	118
Figura 80 – Distribuição espacial das estimativas de β_0 considerando os dados após a exclusão de Clarke para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	118

Figura 81 – Distribuição espacial das estimativas de β_0 considerando os dados após a exclusão de Oconee para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	119
Figura 82 – Distribuição espacial das estimativas de β_1 considerando os dados completos para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	119
Figura 83 – Distribuição espacial das estimativas de β_1 considerando os dados após a exclusão de Bulloch para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	119
Figura 84 – Distribuição espacial das estimativas de β_1 considerando os dados após a exclusão de Clarke para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	120
Figura 85 – Distribuição espacial das estimativas de β_1 considerando os dados após a exclusão de Oconee para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	120
Figura 86 – Distribuição espacial das estimativas de β_2 considerando os dados completos para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	120
Figura 87 – Distribuição espacial das estimativas de β_2 considerando os dados após a exclusão de Bulloch para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	121
Figura 88 – Distribuição espacial das estimativas de β_2 considerando os dados após a exclusão de Clarke para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	121
Figura 89 – Distribuição espacial das estimativas de β_2 considerando os dados após a exclusão de Oconee para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	121
Figura 90 – Distribuição espacial das estimativas de β_3 considerando os dados completos para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	122

Figura 91 – Distribuição espacial das estimativas de β_3 considerando os dados após a exclusão de Bulloch para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	122
Figura 92 – Distribuição espacial das estimativas de β_3 considerando os dados após a exclusão de Clarke para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	122
Figura 93 – Distribuição espacial das estimativas de β_3 considerando os dados após a exclusão de Ocone para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	123
Figura 94 – Distribuição espacial das estimativas de β_4 considerando os dados completos para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	123
Figura 95 – Distribuição espacial das estimativas de β_4 considerando os dados após a exclusão de Bulloch para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	123
Figura 96 – Distribuição espacial das estimativas de β_4 considerando os dados após a exclusão de Clarke para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	124
Figura 97 – Distribuição espacial das estimativas de β_4 considerando os dados após a exclusão de Ocone para o modelo REGPM com erros normais (esquerda) e t -Student (direita)	124
Figura 98 – Distribuição espacial das estimativas significativas de β_0 (esquerda), β_1 (direita) e β_3 (abaixo) no modelo REGPM com erros t -Student	125
Figura 99 – Gráfico de quantis-quantis com envelope simulado dos resíduos padronizados para os modelos de regressão elíptico (esquerda) e REGP (direita) com erros t -Student	144
Figura 100 – Gráficos de influência local para perturbação de casos ponderados nos modelos de regressão elíptico (esquerda) e REGP (direita) com erros t -Student	146

LISTA DE TABELAS

Tabela 1 – Análise descritiva dos indicadores de desenvolvimento humano e criminais nos municípios cearenses	63
Tabela 2 – Análise descritiva das estimativas dos parâmetros e erros padrões assintóticos (entre parênteses) para os modelos de REGP com erros normais	67
Tabela 3 – Análise descritiva das estimativas dos parâmetros e erros padrões assintóticos (entre parênteses) para os modelos de REGP com erros <i>t</i> -Student	68
Tabela 4 – Análise descritiva dos indicadores educacionais e sociais nos condados da Geórgia	103
Tabela 5 – Valores- <i>p</i> para os testes de variabilidade espacial dos parâmetros da REGP	109
Tabela 6 – Análise descritiva das estimativas dos parâmetros e erros padrões assintóticos (entre parênteses) para os modelos de REGPM com erros normais	111
Tabela 7 – Análise descritiva das estimativas dos parâmetros e erros padrões assintóticos (entre parênteses) para os modelos de REGPM com erros <i>t</i> -Student	112
Tabela 8 – Valores dos parâmetros de suavização ótimos selecionadas para os modelos geograficamente ponderados	112

SUMÁRIO

1	INTRODUÇÃO E MOTIVAÇÃO	22
1.1	REVISÃO BIBLIOGRÁFICA	24
1.2	CONSIDERAÇÕES INICIAIS	27
1.2.1	Escala Espacial	27
1.2.2	Dependência, Autocorrelação e Não Estacionariedade Espacial	28
1.2.3	Regressão Geograficamente Ponderada	29
1.2.4	Índice de Moran e Teste de Autocorrelação Espacial	31
1.2.5	Teste de Não Estacionariedade Espacial	32
1.2.6	Família de Distribuições Elípticas	33
1.2.7	Otimização Por Seção Áurea	35
1.3	ORGANIZAÇÃO DA TESE	36
2	MODELO DE REGRESSÃO ELÍPTICO GEOGRAFICAMENTE	
	PONDERADA	38
2.1	ESPECIFICAÇÃO DO MODELO	39
2.2	PROCESSO DE ESTIMAÇÃO	40
2.2.1	Função de Ponderação Espacial	40
2.2.2	Função de Verossimilhança Local	41
2.2.3	Algoritmo de Estimação	43
2.2.4	Matriz de Projeção	43
2.2.5	Testes de Hipóteses	44
2.2.6	Seleção do Parâmetro de Suavização	45
2.3	MÉTODOS DE DIAGNÓSTICO	46
2.3.1	Análise de Resíduos	46
2.3.2	Distância de Cook Generalizada	47
2.3.3	Medidas de Alavancagem	48
2.3.4	Influência Local	49
2.3.4.1	Influência local na estimação	49
2.3.4.2	Influência local na predição	51
2.4	ESTUDO DE SIMULAÇÃO	52
2.4.1	Cenário 1 - Grade Regular	53
2.4.2	Cenário 2 - Grade Irregular	57

2.5	APLICAÇÃO	61
2.5.1	Análise Exploratória	62
2.5.2	Ajuste do Modelo	65
2.5.3	Análise de Sensibilidade	69
2.5.4	Análise Confirmatória	71
2.5.5	Interpretação do Modelo Ajustado	73
3	MODELOS DE REGRESSÃO ELÍPTICO GEOGRAFICA- MENTE PONDERADO MULTIESCALAR	75
3.1	ESPECIFICAÇÃO DO MODELO	76
3.2	PROCESSO DE ESTIMAÇÃO	77
3.2.1	Algortimo <i>backfitting</i>	78
3.2.2	Escolha dos Valores Iniciais	79
3.2.3	Critério de Convergência	80
3.2.4	Centralização das Variáveis Explicativas	80
3.2.5	Matriz de Projeção	81
3.2.6	Testes de Hipóteses	83
3.3	MÉTODOS DE DIAGNÓSTICO	84
3.3.1	Análise de Resíduos	85
3.3.2	Medidas de Alavancagem	86
3.3.3	Influência Local	86
3.3.3.1	Influência local na estimação	86
3.3.3.2	Influência local na predição	88
3.4	ESTUDO DE SIMULAÇÃO	89
3.4.1	Cenário 1 - Escalas Iguais	90
3.4.2	Cenário 2 - Escalas Diferentes	93
3.4.3	Cenário 3 - Escalas Diferentes na Presença de Observação Dis- crepante	96
3.5	APLICAÇÃO	102
3.5.1	Análise Exploratória	102
3.5.2	Ajuste do Modelo	107
3.5.3	Análise de Sensibilidade	114
3.5.4	Análise Confirmatória	117

3.5.5	Interpretação do Modelo Ajustado	125
4	IMPLEMENTAÇÃO COMPUTACIONAL	127
4.1	AJUSTE DO MODELO	127
4.2	VISUALIZAÇÃO DO AJUSTE	137
4.3	ANÁLISE DE DIAGNÓSTICO	141
5	CONCLUSÕES	148
5.1	RESULTADOS E CONTRIBUIÇÕES	149
5.2	TRABALHOS FUTUROS	150
	REFERENCIAS	152

1 INTRODUÇÃO E MOTIVAÇÃO

Modelos para análise de dados espaciais receberam um grande número de contribuições nas últimas décadas. De modo geral, a análise de dados espacial caracteriza-se pelo pressuposto básico de que observações geograficamente próximas devem apresentar características semelhantes. Nesse caso, a aplicação dos modelos de regressões usuais que não incorporam as tendências espaciais presentes nos dados, podem induzir interpretações completamente equivocadas. Isso ocorre pelo fato de que dados dessa natureza podem apresentar características particulares, como por exemplo, a autocorrelação espacial ou a não estacionariedade espacial. Naturalmente, essas características devem ser previamente verificadas e levadas em consideração no momento de escolher a técnica de modelagem adequada. Neste âmbito, autores como Fotheringham, Brunston e Charlton (2002) e Comber *et al.* (2020) sugerem a utilização da Regressão Geograficamente Ponderada (RGP) para a modelagem da não estacionariedade espacial presente nas relações entre a variável resposta e suas regressoras. Essa metodologia representa uma extensão natural e bem intuitiva dos modelos de regressão clássicos (com erros normais) ao permitir que seus parâmetros variem no espaço. Entretanto, essa classe de modelos assumem a normalidade dos erros, o que pode limitar sua aplicação a conjuntos de dados reais. De um modo geral, a suposição de normalidade para a distribuição dos erros aleatórios é comumente utilizada quando a característica de interesse é representada por uma variável contínua com suporte nos reais. Isto se dá pelo fato dessa suposição implicar em modelos com propriedades estatísticas desejáveis e um amplo desenvolvimento teórico. Consequentemente, isso induz a tentativa, em muitos casos errônea, de satisfazer essa suposição. Dentre as tentativas mais comuns podemos destacar as transformações de variáveis e manipulações dos dados. No entanto, isso pode levar a perda de informação ou dificultar bastante a interpretação do modelo utilizado. Além disso, é bem conhecido que a presença de observações com valores discrepantes é um problema muito comum para os modelos estatísticos clássicos. Isso se deve ao fato de que as estimativas dos modelos podem ser fortemente influenciadas por essas observações.

No contexto da estatística espacial, os métodos para solucionar este problema são tipicamente mais complicados porque diferentes tipos de observações podem ocorrer (Genton, 2001 e Fotheringham, Brunston e Charlton, 2002). Para os modelos de RGP, Fotheringham, Brunston e Charlton (2003) mostram que as estimativas do modelo tam-

bém podem ser altamente sensíveis na presença de valores discrepantes e propõem duas abordagens para resolver este problema. A primeira abordagem remove as observações com valores discrepantes e reajusta o modelo para comparar os resultados antes e depois da exclusão, deixando a cargo do pesquisador da área explicar a presença dessas observações. Nessa abordagem, há um alto custo computacional na detecção dos valores discrepantes, principalmente para conjuntos de dados com muitas observações. A ideia principal da segunda abordagem é reduzir o peso das observações com grandes valores residuais no processo de estimação. Nesse caso, aplicado-se um algoritmo robusto e automatizado que reajusta o modelo recursivamente até que o efeito dos valores discrepantes nas estimativas seja reduzido. Esta abordagem assume um método de detecção mais simples e desconsidera a interpretação subjetiva para a presença das observações discrepantes. No entanto, ambas abordagens são baseadas na remoção de valores discrepantes para o processo de estimação do modelo RGP. No contexto geral, um método alternativo robusto é considerar distribuições de caudas mais pesadas para os erros aleatórios. Autores como Westfall e Arias (2020) sugerem que esta abordagem é muito útil para reduzir a forte influência de valores discrepantes, uma vez que essas observações são mais prováveis de ocorrer em distribuições de caudas mais pesadas. Em particular, Fang, Kotz e Ng (1990) apresentam a família de distribuições elípticas, bem como a sua utilização no contexto da análise de regressão. Esta família contém distribuições com caudas leves e pesadas, como por exemplo a normal, *t*-Student e exponencial potência. Dessa forma, a utilização de distribuições pertencente a essa família para a suposição dos erros aleatórios em modelos de RGP é uma forma de contornar os problemas gerados pela presença de valores discrepantes nessa classe de modelos.

Adicionalmente, no contexto geral da análise de regressão é fundamental a verificação das suposições e da sensibilidade do modelo proposto. Para isso, são comumente desenvolvidas técnicas de diagnóstico, as quais podem ser consideradas sob diferentes enfoques. As técnicas de diagnóstico mais usuais tiveram início nos modelos lineares clássicos e seus conceitos vem sendo estendidos a diversas classes de modelos. No entanto, dependendo da complexidade do modelo essa tarefa pode ser bastante árdua. Na RGP há uma vasta discussão no que se refere a análise de resíduos, porém ainda não foi difundido estudos sobre medidas de influência para estes tipos de modelos. Visto que a detecção de observações possivelmente influentes representa um fator importante na modelagem de

dados, então é necessário também um estudo mais profundo neste tema. Essas técnicas são capazes de identificar observações que exercem influências desproporcionais no ajuste de um modelo.

Com base nesses fatos, o principal objetivo deste trabalho é propor uma classe de modelos geograficamente ponderados utilizando a família de distribuições elípticas para a componente aleatória. Além disso, apresentamos uma adaptação dessa classe de modelos considerando escalas geográficas variáveis, permitindo que cada relação de regressão possua sua própria variação espacial. Adicionalmente, procuramos desenvolver técnicas de diagnóstico para estas classes de modelos, como forma de avaliar a qualidade do ajuste e a sensibilidade do modelo ajustado. Por último, foi elaborada um conjunto de rotinas computacionais com a finalidade de facilitar a divulgação e aplicabilidade deste trabalho. Os resultados também foram utilizados para a produção de três artigos científicos, os quais encontram-se em processo de desenvolvimento e submissão. Alguns resultados obtidos nessa pesquisa também foram apresentados na 34^a edição do *Internacional Workshop on Statistical Modelling (IWSM)* e publicado no livro de resumo correspondente a essa edição do congresso.

1.1 REVISÃO BIBLIOGRÁFICA

A Regressão Geograficamente Ponderada (RGP) foi introduzido por Brunson, Fotheringham e Charlton (1996) e tem sido amplamente discutida na literatura por diversos autores. Essa classe de modelos assume que os parâmetros de regressão variam ao longo do espaço, possibilitando assim o estudo da não estacionariedade espacial presente nas relações de regressão (veja LeSage, 1999, Fotheringham, Brunson e Charlton, 2002 e Comber *et al.*, 2020, para mais detalhes). Em outras palavras, a RGP permite que a relação entre variável resposta e variáveis explicativas podem não ser equivalentes para regiões distintas no espaço, considerando nesse caso o ajuste de modelos locais associado a cada uma das regiões estudadas. Com base nessa premissa, diversos trabalhos teóricos foram desenvolvidos visando a adaptação dos modelos de RGP, bem como suas aplicações práticas em diferentes áreas de estudo. Por exemplo, Atkinson *et al.* (2003) desenvolveram a RGP para aplicar em dados com resposta binária, considerando o uso da distribuição logística para os erros aleatórios. A RGP logística multinomial foi proposta por Luo e Kanala (2008) aplicada a dados com variáveis do tipo categóricas. Para dados de contagem,

as distribuições Poisson e binomial negativa foram incorporadas na RGP por Nakaya *et al.* (2005) e Da Silva e Rodrigues (2014), respectivamente. Por fim, Da Silva e Lima (2017) propõem o modelo de RGP para dados espaciais no intervalo unitário ao assumir que os erros aleatórios seguem uma distribuição beta. Em todos os casos, a aplicação desses modelos em dados reais sugere que as relações de regressão são não-estacionárias no espaço e que os modelos usuais não acomodam bem as características espaciais presente nos dados.

Por outro lado, os modelos para análise de dados espaciais também apresentam problemas com a forte influência de valores discrepantes e as abordagens para lidar com isso tendem a ser mais complexas. Vários autores discutiram métodos para detectar observações com valores discrepantes em diferentes classes de modelos espaciais (ver Haslett *et al.*, 1991, Cerioli e Riani, 1999, Harris *et al.*, 2013, Jin *et al.*, 2015, Dai *et al.*, 2016 e De Bastiani *et al.*, 2018, por exemplo). O procedimento usual neste ponto é investigar as observações identificadas e considerar removê-las do banco de dados. No entanto, remover observação discrepantes que vêm puramente do processo de geração natural dos dados representa uma omissão de informações importantes e pode levar a interpretações errôneas. Desta forma, alguns autores propõem métodos alternativos de estimação robusta para modelos espaciais menos sensíveis a outliers (para mais detalhes, ver Genton, 1998, Genton, 2001, Militino, Palacios e Ugarte, 2003, Harris, Fotheringham e Juggins, 2010 e De Bastiani *et al.*, 2015). No âmbito da RGP, Fotheringham, Brunson e Charlton (2002) advertem que a influência de valores discrepantes também é um problema para esta classe de modelos e discutem possíveis soluções para reduzir a influência dessas observações. No entanto, Warsito *et al.* (2018) afirmam que ambas as abordagens têm um elemento de subjetividade, o que pode dificultar a interpretação dos resultados. Uma alternativa a essas abordagens é considerar uma distribuição de caudas mais pesadas para erros aleatórios, as quais são capazes de reduzir a forte influência dos valores discrepantes. Inúmeros trabalhos em modelos de regressão com distribuições de caudas pesadas podem ser encontrados (veja, por exemplo, Lange, Little e Taylor, 1989, Fang, Kotz e Ng, 1990, Cysneiros, Paula e Galea, 2007, Ibacache-Pulgar, Paula e Cysneiros, 2013, Vanegas, Rondon e Cysneiros, 2013 e De Bastiani *et al.*, 2015). Especificamente, Fang, Kotz e Ng (1990) propuseram uma classe de modelos de regressão com base na família de distribuições elípticas. Naturalmente, a extensão da RGP assumindo erros aleatórios elípticos representa uma alternativa atrativa

para a modelagem de dados espaciais com presença de valores discrepantes.

Outro aspecto fundamental na análise de dados espaciais é o conceito de escala. Em síntese, a escala espacial representa o nível em que os processos geográficos variam. No âmbito da RGP, Fotheringham, Yang e Kang (2017) destacam a importância desse conceito, expondo as limitações da RGP em relação a múltiplas escalas espaciais. De acordo com os autores supracitados, os modelos de RGP estão sob a suposição de que todos os parâmetros operam na mesma escala espacial. Entretanto, essa afirmação pode ser bastante irrealista dependendo da característica dos dados considerados no estudo científico. Por esse motivo, Fotheringham, Yang e Kang (2017) propõem uma extensão da RGP que funciona em diferentes escalas espaciais, a qual foi denominada por Regressão Geograficamente Ponderada Multiescalar (RGPM). Essa metodologia tem sido amplamente discutida nos últimos anos, obtendo diversos resultados teóricos e empíricos relevantes na área. Um estudo inferencial e um guia para a aplicação prática desses modelos podem ser encontrados em Yu *et al.* (2020) e Comber *et al.* (2020), respectivamente. Dessa forma, se ver necessário estender a RGP com erros aleatórios elípticos também para o caso de múltiplas escalas espaciais. Com isso, seria possível uma modelagem mais realista de diversos processos espaciais, sem preocupar-se com a influência de valores discrepantes.

Por fim, após a obtenção do ajuste de um modelo é necessário verificar as suposições assumidas na sua postulação, por meio de uma análise residual e um estudo da sensibilidade. Nesse contexto, Cox e Snell (1968) expressam uma definição geral para os resíduos, a qual pode ser usada como uma forma de avaliar as suposições associadas ao termo aleatório. Adicionalmente, para verificar a sensibilidade do modelo é possível destacar diversas propostas para medidas de diagnóstico. Para o objetivo de detectar observações que exercem pesos desproporcionais no processo de estimação do modelo, Wei, Hu e Fung (1998) propõem uma medida denominada de alavancagem generalizada. Sobre outro enfoque, Cook (1977) apresenta uma distância de Cook generalizada, a qual permite medir o efeito da exclusão de observações nas estimativas dos parâmetros. Por outro lado, uma técnica bastante inovadora, introduzida Cook (1986), permite avaliar a influência conjunta de observações sob pequenas perturbações tanto no modelo como nos dados. Essa técnica é conhecida como influência local e representa uma das medidas de diagnóstico mais sofisticadas até o presente momento. Levando em consideração esses estudos, autores como Galea, Paula e Cysneiros (2005) e Cysneiros, Paula e Galea (2007)

desenvolveram extensões para análise de resíduos e medidas de diagnóstico para modelos de regressão com erros aleatórios pertencente à família de distribuições elípticas. Os aspectos de robustez e métodos de diagnóstico para modelos de regressão com erros aleatórios elípticos univariados são detalhados em Villegas *et al.* (2013). Naturalmente, os estudos discutidos por esses autores podem ser estendidos também para modelos geograficamente ponderados, os quais são proposto nessa tese.

1.2 CONSIDERAÇÕES INICIAIS

Como mencionado anteriormente, o conteúdo abordado nesta tese refere-se a análise de dados espaciais sob determinada natureza, requerendo um conhecimento prévio dos fenômenos associados a esse tipo de dados. No entanto, esse conhecimento pode não ser tão trivial dependendo da área de domínio do leitor, sendo necessário nesse caso o esclarecimento de determinados fundamentos comumente associado a dados espaciais. Visto isso, a presente seção tem como objetivo introduzir alguns conceitos fundamentais neste âmbito, bem como ferramentas importantes para a análise de dados espaciais. Os conceitos discutidos aqui terão grande importância para a compreensão e definição dos modelos geograficamente ponderados apresentados nos próximos capítulos. Além disso, são abordadas brevemente determinadas metodologias para auxiliar a análise de dados espaciais através desses modelos, destacando-se a utilidade específica de cada uma das metodologias mencionadas.

1.2.1 Escala Espacial

De forma sucinta, a escala espacial é definida como a dimensão física de um objeto ou processo no espaço. Em outras palavras, entende-se por escala espacial o nível em que os fenômenos estudados variam no espaço. A importância da escala reside no fato de que dependendo dela se obterá descrições diferentes para o mesmo processo estudado. Desta forma, Todos nós temos uma ideia intuitiva do conceito de escala ao estar ciente das diferenças nas extensões dos elementos em função de sua área geográfica. Por exemplo, suponha que queremos estudar o comportamento da renda média de determinada população. Se medimos esta variável por país, notaremos que ela tem uma grande variação, possivelmente atribuída as diferenças entre as gestões de cada país. No entanto, se

medimos esta variável por dentro de um município específico, notaremos que terá um comportamento diferente, tendo possivelmente uma menor variação. Assim podemos notar que dependendo da escala espacial, os estudos realizados apresentaram diferentes características. Este conceito será particularmente importante na definição e aplicação dos modelos propostos nesta tese para a análise de dados espaciais.

1.2.2 Dependência, Autocorrelação e Não Estacionariedade Espacial

A dependência espacial ocorre quando o valor de uma variável em determinada unidade espacial é parcialmente função do valor da mesma variável em unidades vizinhas. Neste contexto, a essência da autocorrelação espacial é a mesma da dependência espacial, ou seja, observações retiradas de locais diferentes não são independentes umas das outras. De forma intuitiva, teremos que observações em locais mais próximos uns dos outros serão mais semelhantes (autocorrelação espacial positiva) ou menos semelhantes (autocorrelação espacial negativa) do que observações em locais mais distantes. Portanto, um elemento chave para medir e compreender a autocorrelação espacial é a descrição da relação entre o grau de similaridade entre as observações e a distância que as separa. A descrição da dependência espacial pode então ser usada para definir uma função de ponderação espacial que relaciona a similaridade das observações com a distância entre elas. Com base nesse conceito, a autocorrelação espacial irá então representar uma medida da magnitude e direção dessa relação.

Em contrapartida, um processo estacionário é um fenômeno homogêneo que se desenvolve no espaço. Nesse caso, diz-se que um processo $Z(x)$ é estritamente estacionário quando uma distribuição de dimensão finita permanece invariante sob qualquer translação arbitrária dos pontos por meio de d (distância de separação entre as observações). Além disso, um processo $Z(x)$ é fracamente estacionário se as seguintes condições forem satisfeitas:

- $Z(x)$ tem uma esperança finita e constante para cada ponto em seu domínio, o que implica que a esperança de os incrementos são zero

$$\mathbb{E}[Z(x+d) - Z(x)] = 0$$

- Para qualquer vetor h , a variância do incremento é definida e é uma função única da distância h

$$\text{Var}[Z(x+d) - Z(x)] = \mathbb{E}[Z(x+d) - Z(x)]^2 = 2\gamma(d)$$

Nesse sentido o conceito de não estacionariedade espacial dos parâmetros de um modelo de regressão refere-se à variação das relações entre as variáveis no espaço (uma discussão mais ampla a respeito desse conceitos pode ser encontrado em Fotheringham (2009)). Vale ressaltar que esses conceitos espaciais tem uma grande importância na análise de regressão, pois se não forem levados em consideração podem gerar resultados equivocados. A implicação metodológica nesse caso é que os coeficientes de regressão serão ineficientes para mostrar a magnitude da relação entre as variáveis e os testes de significância estatística sobre eles serão questionáveis devido a inflação nos erros padrões.

1.2.3 Regressão Geograficamente Ponderada

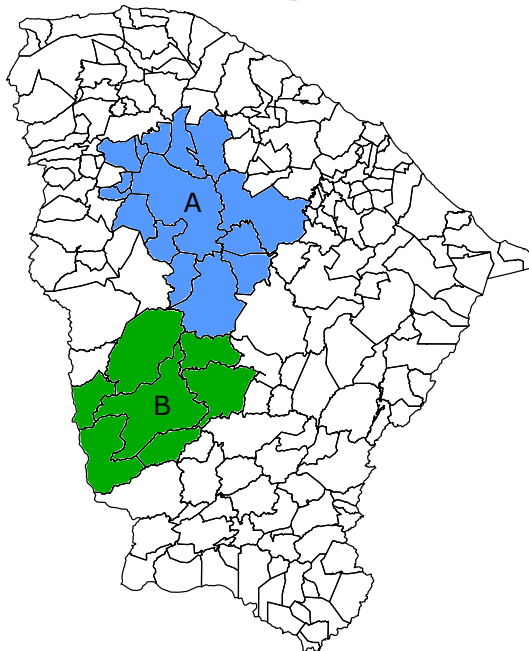
Como já foi mencionado anteriormente, os modelos de regressão geograficamente ponderados estão baseados em conceitos associados a dependência espacial e, principalmente, pela suposição de não estacionariedade espacial dos seus parâmetros de regressão. Esse modelos representam uma adaptação direta do conceito de regressão local para um contexto puramente geográfico, assumindo nesse caso que observações geograficamente próximas devem ter características semelhantes. Embora esse conceito seja intuitivo ainda não está tão claro como funciona a regressão geograficamente ponderada na análise de dados espaciais. Com isso em vista, consideramos a seguir um breve exemplo ilustrativo para uma melhor compreensão da ideia fundamental por trás dessa classe de modelos.

Assuma inicialmente que estamos estudando a relação entre as variáveis Y e X no estado do Ceará, de tal forma que Y representa a variável resposta e X uma variável explicativa. Nesse caso, ambas as variáveis foram medidas para cada um dos 184 municípios desse estado. Um modelo de regressão usual assume que a relação entre Y e X é igual para todos município, desconsiderando qualquer informação geográfica relevante. Esse modelo então levará em consideração a informação de todos os municípios conjuntamente para construir uma única medida que representa uma relação de regressão única para todo o estado. No entanto, a forma como as variáveis Y e X estão relacionadas pode ser diferente para determinadas regiões do Ceará e nesse caso uma única medida não será o suficiente para descrever sua variação no espaço.

Por outro lado, um modelo de regressão geograficamente ponderado irá descrever a relação entre as variáveis de forma separada para cada município, destacando variações locais dessa relação. Para isso um modelo local para cada município será ajustado de

maneira separada considerando apenas uma vizinhança próxima. A estrutura funcional desse modelo será essencialmente a mesma da regressão usual, mas a direção e a magnitude da relação entre Y e X poderão variar entre os locais. Em outras palavras, o ajuste do modelo local para o município “A” será realizado de forma equivalente a regressão usual, mas considerando apenas as observações dos municípios mais próximos (como é ilustrado na Figura 1 pela região de cor azul). O mesmo processo será realizado no ajuste do modelo local para o município “B” (utilizando a região destacada na Figura 1 pela cor verde), obtendo um conjunto de parâmetros diferentes devido a redução do espaço amostral total em ambos os casos. Esse processo é repetido para cada um dos municípios pertencentes ao estado do Ceará, obtendo então uma variação espacial para cada parâmetro de regressão. O resultado do ajuste dos modelos locais permitem que os parâmetros de regressão sejam interpretados como uma superfície espacial no estado do Ceará, as quais podem ressaltar comportamentos específicos para cada região do mapa. Naturalmente, a vizinhança utilizada no ajuste de cada modelo local é definida com base em alguma medida de proximidade, a qual será descrita formalmente no seguinte capítulo. Com base nesse princípio, as definições formais e as metodologias necessárias para a aplicação da regressão geograficamente ponderada serão apresentadas no decorrer desta tese.

Figura 1 – Mapa dos municípios do Ceará utilizados no ajuste dos modelos locais para os municípios “A” e “B”



Fonte: O autor (2021).

1.2.4 Índice de Moran e Teste de Autocorrelação Espacial

O Índice de Moran é uma medida de autocorrelação espacial que é mensurado através das localizações e dos valores observados da variável estudada. De maneira sucinta, o cálculo desse índice é obtido através da seguinte equação (Bivand e Wong, 2018):

$$I = \frac{\sum_{j=1}^n \sum_{l=1}^n w_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\sum_{j=1}^n \sum_{l=1}^n w_{ij} \sum_{j=1}^n (X_i - \bar{X})^2}.$$

em que X_i é o valor que a variável $\mathbf{X}$ assume no ponto i , $\bar{X}$ é a média amostral da variável $\mathbf{X}$, n é o número de observações e w_{ij} representa um peso espacial associado a uma medida de distância d_{ij} entre os pontos i e j . Nesse caso, diferentes métodos podem ser usados para calcular os pesos espaciais. Um dos métodos mais comuns é utilizar o conceito de vizinhança, ou seja, as observações que são consideradas vizinhas (sob algum critério preestabelecido) terão peso um enquanto as demais terão peso igual a zero. Note também que, o numerador dessa equação é um termo que representa uma medida de covariância enquanto que o denominador indica uma medida de variância. Isso faz com que esse índice se comporte como um coeficiente de correlação e geralmente varia entre os valores -1 e 1 , embora em teoria seja possível exceder ambos os limites. Nesse caso, se os pontos localizados a uma certa distância mostrarem valores semelhantes então uma correlação positiva é obtida. Por outro lado, se os pontos mostrarem valores muito diferentes, então será obtida uma correlação negativa. E caso não haja autocorrelação entre as observações então o índice calculado deverá ser próximo de zero.

Uma outra abordagem é realizar uma análise gráfica da dispersão entre os valores de $(\mathbf{x} - \bar{\mathbf{x}})$ e $\mathbf{W}(\mathbf{x} - \bar{\mathbf{x}})$, em que $\mathbf{W}$ é a matriz de pesos que contém os valores de w_{ij} . Este gráfico é conhecido como diagrama de espalhamento de Moran e também pode ser utilizado para avaliar a presença de autocorrelação espacial. Nesse caso, se a dispersão dos valores plotados nesse gráfico for distribuída aleatoriamente entre os quadrantes do plano cartesiano, então assumimos que não há presença de autocorrelação espacial. Por outro lado, se houver uma tendência na concentração desses valores no primeiro e terceiro quadrante há razões para suspeitar que haja uma autocorrelação espacial positiva na variável estudada. De forma similar, se a concentração dos valores estiverem no segundo e quarto quadrante, então há uma autocorrelação espacial negativa.

Adicionalmente, o índice de Moran também pode ser usado para realizar o teste de significância para a autocorrelação espacial. Nesse caso é testado, sob hipótese nula, se valor de I obtido para a variável estudada é estatisticamente igual a zero. Em Bivand e Wong, 2018 há uma discussão acerca da distribuição de probabilidade do índice de Moran sob essa hipótese nula e uma estatística de teste apropriada é apresentada. Como base nesse resultado é possível se a configuração espacial da uma variável ocorre de forma aleatória ou não. Nesta tese, esse teste é particularmente importante para verificar a presença de autocorrelação espacial nas variáveis estudadas e nos resíduos para os modelos de regressão ajustados. Vale ressaltar também que, há inúmeros outros indicadores de associação espaciais globais e locais, os quais também podem ser amplamente consideradas para a identificação de autocorrelação espacial presente nos dados. Para mais detalhes sobre medidas de autocorrelação espacial e das suas respectivas implementações computacionais, consulte Bivand e Wong, 2018, por exemplo.

1.2.5 Teste de Não Estacionariedade Espacial

Como mencionado anteriormente, os modelos de regressão geograficamente ponderados tem como base o pressuposto de que os parâmetros de regressão não são estacionários no espaço. Entretanto, uma dúvida válida é se determinado conjunto de estimativas para os parâmetros locais exibem uma variação espacial significativa. Nesse contexto, a variabilidade das estimativas locais pode ser utilizada como forma de examinar a plausibilidade da suposição de não estacionariedade espacial dos parâmetros de regressão. Para isto, assumamos inicialmente que V_k representa a variância das estimativas espaciais associada ao k -ésimo parâmetro de regressão. Valores altos dessa medida iriam indicar a não estacionariedade do seu respectivo parâmetro de regressão, embora seja provável que mesmo os parâmetros que não variem no espaço apresentem alguma variação nas estimativas locais. O ponto crucial nesse caso seria definir para que valores de V_k podemos considerar uma variação espacial significativa.

Uma forma de testar a significância da variabilidade espacial é determinar a distribuição amostral de V_k sob a hipótese nula que o parâmetro é estacionário no espaço. Nesse caso, uma abordagem aleatorizada via método de Monte Carlo seria uma alternativa plausível para obter essa distribuição de forma aproximada. Em outras palavras, sabe-se que sob hipótese nula qualquer permutação de pares (u_i, v_i) entre os pontos de dados

é igualmente provável. Por esse fato, o valor observado de V_k pode ser comparado aos valores obtidos a partir do reagrupamento aleatório dos dados no espaço e da repetição do ajuste do modelo geograficamente ponderado aos dados reagrupados, de tal forma que serve de base para o cálculo da significância do teste.

Com base nesse princípio, o conjunto de valores das medidas de variabilidade pode ser classificado do menor ao maior e a proporção de valores que excedem os dados observados representa a probabilidade de observar tal variação nas estimativas dos parâmetros locais de um processo estacionário. Dessa forma, quanto menor for essa probabilidade, mais confiança se tem em afirmar que o processo gerador das estimativas dos parâmetros locais é não estacionário. Em outras palavras, rejeitar a hipótese nula implica que as estimativas não são estacionárias no espaço, uma vez que quanto menor for nível de significância obtido menos provável que a variabilidade espacial observada seja proveniente de um fator puramente aleatório. Esse resultado pode então indicar quais relações de regressão são ou não estacionárias no espaço através da variabilidade espacial das estimativas locais. Mais detalhes sobre a utilização desses testes de hipóteses podem ser encontrados em Fotheringham, Brunsdon e Charlton, 2002, bem como proposta alternativas a esse teste.

1.2.6 Família de Distribuições Elípticas

A família de distribuições elípticas engloba inúmeras distribuições de probabilidades comumente utilizadas para análise de dados contínuos que assumem valores nos números reais. Esta família inclui as distribuições normal, t -Student, logística dos tipos I e II, exponencial de potência, entre outras. As distribuições pertencentes a essa família compartilham propriedades importantes proveniente da distribuição normal, no entanto podem apresentar caudas mais leves ou mais pesadas que as caudas da normal. Este fato possibilita uma maior flexibilidade para a análise de valores extremos. Nesse âmbito, a aplicação desta família de distribuição para modelos de regressão são amplamente discutidas na literatura e há inúmeros estudos que evidenciam sua utilidade na modelagem de dados com presença de observações discrepantes.

No contexto geral, dizemos então que uma variável aleatória Y pertence a uma distribuição elíptica univariada (ou uma densidade elíptica), denotada por $Y \sim S(\mu, \phi)$, se

sua função de densidade assumir a seguinte forma:

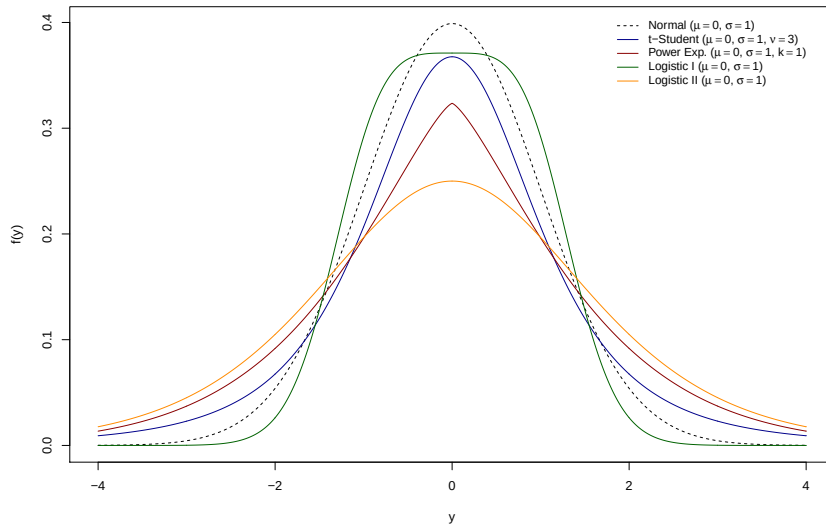
$$f(Y; \mu, \phi) = \frac{1}{\sqrt{\phi}} g \left\{ \frac{(y - \mu)^2}{\phi} \right\}, \quad y \in \mathbb{R},$$

em que $\mu \in \mathbb{R}$ é o parâmetro de localização, $\phi > 0$ é o parâmetro de escala e $g(\cdot)$ é chamada de função geradora de densidade, com $g(u) > 0$, para $u > 0$ e $\int_0^\infty u^{-1/2} g(u) du = 1$. A partir dessa definição podemos destacar algumas propriedades importantes e particularmente úteis para a construção dos modelos de regressão. Neste contexto, se $Y \sim S(\mu, \phi)$, então a função característica de Y é expressa por $\zeta_Y(t) = e^{it\mu} \Psi(t^2 \phi)$, $t \in \mathbb{R}$ para alguma função Ψ com $h(u) \in \mathbb{R}$ para todo $u > 0$. Consequentemente, a média e a variância de Y , quando existentes, são:

$$\mathbb{E}(Y) = \mu \quad \text{e} \quad \text{Var}(Y) = \xi \phi,$$

onde $\xi > 0$ é uma constante que não depende dos parâmetros μ e ϕ . Adicionalmente, assumindo que $Z = aY + b$ e $Y \sim S(\mu, \phi)$, então $Z \sim S(a\mu + b, a^2 \phi)$. Estas são propriedades particularmente importantes para a formulação dos modelos de regressão com erros aleatórios elípticos. Como forma de ilustrar o comportamento de algumas distribuições pertencente a essa família, a Figura 2 exibe as suas respectivas funções de densidades, revelando diferenças interessantes na forma das caudas de cada uma dessas distribuições. Para a obtenção das funções de densidades os valores dos parâmetros dessas distribuições foram pré-especificados, os quais são exibidos na legenda do gráfico supracitado.

Figura 2 – Função de densidade para algumas distribuições elípticas



Fonte: O autor (2021).

Note que, cada distribuição apresenta uma densidade distinta para a ocorrência de valores extremos, possibilitando uma modelagem menos sensível a presença de valores discrepantes. Por este fato, o uso dessa família de distribuição permite a modelagem de dados com a presença de observações discrepantes, reduzindo a forte influência que essas observações exercem na estimativa de modelos que assumem a normalidade dos erros aleatórios. Vale ressaltar que, além dos parâmetros de localização e escala algumas dessas distribuições podem apresentar também parâmetros de forma, os quais serão considerados fixos no processo de estimação dos modelos propostos nesta tese. Por exemplo, a distribuição *t*-Student é especificada através do parâmetro ν , o qual representa os graus de liberdade (g.l.) dessa distribuição. Nesse caso, um método específico para a seleção do valor ideal para ν no ajuste de modelos com erros elípticos pode ser considerado. O mesmo ocorre para o parâmetro de forma associado a distribuição exponencial potência (mais detalhes acerca da família de distribuições elípticas e modelos de regressão com erros elípticos podem ser encontrados em Kelker, 1970, Rao, 1990, Fang, Kotz e Ng, 1990 e Cysneiros e Paula, 2005).

1.2.7 Otimização Por Seção Áurea

Na aplicação de métodos estatísticos é bem comum nos depararmos em situações que precisamos obter o mínimo (ou o máximo) de uma função em termos de uma ou mais variáveis e a aplicação de algoritmos computacionais podem ser utilizados para este propósito. Neste contexto, algoritmos de minimização têm sido amplamente utilizados para resolver problemas de otimização em diversas áreas da ciência. Dentre os esses algoritmos, a otimização por seção áurea se destaca por sua simplicidade teórica e, em muitos casos, pelo baixo custo computacional exigido. De forma sucinta, a ideia por trás desse algoritmo é reduzir o intervalo total de uma determinada função de forma iterativa, com base na sequência de Fibonacci, até que o comprimento do intervalo seja suficientemente pequeno e convirja para o ponto mínimo dessa função. Essa metodologia é particularmente útil para a otimização de um parâmetro importante em modelos de regressão geograficamente ponderado, o qual será abordado com mais detalhes nos próximos capítulos.

Para a aplicação desse algoritmo, assuma inicialmente que desejamos obter o mínimo da função $f(x)$, a qual tem seu valor mínimo dentro de um intervalo $[a, b]$ preestabelecido. Nesse caso, o comprimento total desse intervalo é $b - a$ e desejamos

reduzir-lo ao retirar uma parcela desse intervalo que o mínimo não esteja. Com base nesse princípio, o próximo passo é selecionar dois pontos x_1 e x_2 dentro do intervalo total através dos seguintes operações:

$$x_1 = b - (b - a)r$$

$$x_2 = a + (b - a)r$$

em que $r = 0,618$ representa a razão áurea. Feito isso, deve-se identificar qual desses dois pontos retornam o menor valor para a função $f(x)$ e um novo intervalo deve ser estabelecido. Em outras palavras, se $f(x_1) > f(x_2)$, então a função decresce na direção de x_2 e consequentemente $(x_1, b]$ será o novo intervalo de busca contendo o mínimo dessa função. Por outro lado, se $f(x_1) < f(x_2)$, então a função decresce na direção de x_1 e o novo intervalo de busca nesse caso será $[a, x_2)$, o qual deverá conter o mínimo dessa função. Logo, o processo de redução do intervalo deve ser repetido até que o comprimento do intervalo seja suficientemente pequeno, obtendo assim o valor que minimiza a função $f(x)$ (mais detalhes acerca desse método de otimização podem ser encontrado em Greig, 1980 e Press *et al.*, 2007).

1.3 ORGANIZAÇÃO DA TESE

Nesta tese são discutidas propostas de modelos alternativos a RGP menos sensível à presença de valores discrepantes. No capítulo 2 propomos uma extensão da RGP considerando a família de distribuições elípticas para os erros aleatórios, a qual denominamos de Regressão Elíptica Geograficamente Ponderado (REGP). Apresentamos um processo de estimação do tipo score Fisher para esta classe de modelos baseando-se no conceito de verossimilhança local. Também discutimos a análise de resíduos e as medidas de diagnóstico baseadas na influência global e local para os modelos de REGP. Com o objetivo de avaliar o desempenho dos critérios para a escolha do parâmetro de suavização ótimo e a sensibilidade do modelo com relação a valores discrepantes, propomos estudos de simulação sob diferentes cenários. Por último, aplicamos os métodos desenvolvido a um exemplo prático baseado em dados reais referente a um estudo de desenvolvimento humano e criminalidade nos municípios do Ceará, Brasil. Os resultados obtidos são interpretados e algumas considerações finais são discutidas.

No capítulo 3 estendemos a REGP ao considerar múltiplas escalas geográficas,

permitindo que cada relação de regressão tenha sua própria variação espacial. Inicialmente reformulamos o modelo REGP assumindo um parâmetro de suavização distinto associado a cada variável explicativa, introduzindo assim a Regressão Elíptica Geograficamente Ponderada Multiescalar (REGPM). Com base na especificação do modelo, reescrevemos a REGP e REGPM na forma de um Modelos Aditivo Generalizado (MAG) e obtemos o processo de estimação dos parâmetros através de um algoritmo do tipo *backfitting*. Discutimos aspectos importantes acerca do método computacional deste algoritmo. Além disso, derivamos resultados inferenciais com base no processo de estimação. Apresentamos também a análise de resíduos e uma medida de alavancagem. Avaliamos empiricamente o modelo proposto através de diversos cenários simulados. Concluimos o capítulo com uma aplicação prática utilizando uma base de dados reais referente ao desempenho educacional nos condados da Geórgia, EUA. Algumas considerações finais acerca desse exemplo também são expostas.

Com a finalidade de facilitar a aplicação da teoria discutida nessa tese foi desenvolvido um pacote computacional no *software* R. Nesse contexto, no capítulo 4 é apresentado detalhadamente a sequência lógica da utilização das rotinas contidas no pacote, assim como uma breve orientação de como devem ser utilizadas. Essas rotinas contemplam funções destinadas ao processo de estimação, a descrição do ajuste, a análise de resíduos e as medidas de diagnóstico obtidos no decorrer desta tese. Para um melhor entendimento, alguns exemplos práticos são exibidos e brevemente discutidos.

Por fim, no capítulo 5 apresentamos uma síntese dos resultados alcançados ao longo dessa tese. Destacamos algumas conclusões importantes, discutimos nossas contribuições com base em pesquisas anteriores e expomos futuras linhas de pesquisa no âmbito deste trabalho.

2 MODELO DE REGRESSÃO ELÍPTICO GEOGRAFICAMENTE PONDERADA

Modelos espaciais são desenvolvidos para acomodar diferentes características geográficas intrínsecas ao processo de geração dos dados. Para modelar relações que variam espacialmente, a aplicação da RGP é uma alternativa muito atrativa. Entretanto, a RGP é construída sob a suposição de normalidade dos erros, podendo ser sensível à presença de valores discrepantes nos dados. Nesse caso, a utilização de distribuições com caudas mais pesadas presentes na família de distribuições elípticas ou métodos robustos podem reduzir a influência de valores discrepantes na estimação dos parâmetros do modelo. Como foi mencionado anteriormente, as distribuições pertencentes a essa família preservam propriedades importantes da distribuição normal, no entanto apresentam diferenças particulares na forma e suas caudas. Esse fato indica densidades distintas para a ocorrência de valores extremos, possibilitando uma modelagem menos sensível a presença de valores discrepantes dependendo da distribuição utilizada. Vale ressaltar também que a normal é um caso particular dessa família de distribuições, implicando que os resultados obtidos neste trabalho também se estende para o caso que os erros são normalmente distribuídos.

Neste contexto, o presente capítulo tem como objetivo propor a REGP, na qual representa uma generalização da RGP ao estender a suposição de normalidade através da família de distribuições elípticas. Com base nos resultados encontradas na literatura acerca dos modelos de regressão com erros elípticos, adaptou-se a RGP com finalidade de incluir as principais propriedades das distribuições elípticas em modelos geograficamente ponderados. Esta abordagem representa uma alternativa para modelar a não estacionaridade espacial para relações condicionais entre a variável resposta e as variáveis explicativas em base de dados georreferenciados com a presença de valores discrepantes. Vale ressaltar que a RGP representa um caso particularmente importante desta classe de modelos, quando assumimos que os erros são normalmente distribuídos. Dessa forma, a seguir é apresentada a definição dos modelos de REGP, ressaltando alguns conceitos preliminares quando necessário. Em torno disso, obtemos um processo de estimação com base em um método score de Fisher. Métodos de diagnóstico para detectar observações influentes e validar o modelo proposto também são propostos neste capítulo. Por fim, apresentamos os principais resultados obtidos em estudos de simulações e na aplicação prática a uma base de dados reais.

2.1 ESPECIFICAÇÃO DO MODELO

No contexto geral da modelagem estatística, um modelo de regressão linear global pode ser descrito pela seguinte equação:

$$y_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ik} + \epsilon_i \quad (2.1)$$

em que $i = 1, \dots, n$, β_k é o parâmetro associado à k -ésima variável explicativa, x_{ik} é o valor da i -ésima observação para a k -ésima variável explicativa e ϵ_i é o erro aleatório associado a i -ésima observação. Com base na equação (2.1), diferentes classes de modelos podem ser definidas por meio das suposições assumidas em relação aos erros aleatórios. No caso dos modelos de regressão elípticos, assumimos que os erros aleatórios são mutuamente independentes e pertencem à família de distribuições elípticas univariadas, a qual foi definida com detalhes no capítulo anterior.

Por outro lado, um modelo geograficamente ponderado estende a estrutura de regressão tradicional apresentada em (2.1) permitindo uma variação local dos parâmetros de regressão. A partir dessa definição torna-se possível estudar a não estacionariedade espacial na relação entre a variável respostas e suas respectivas variáveis explicativas, uma vez que cada local no espaço tem um valor diferente para os parâmetros de regressão. Com base neste princípio propomos então um modelo geograficamente ponderado com erros aleatórios pertencentes à família de distribuições elípticas univariadas, generalizando a classe dos modelos de RGP proposta inicialmente por Brunson, Fotheringham e Charlton (1996) e amplamente estudado na literatura nos últimos anos (veja Brunson, Fotheringham e Charlton (1996), LeSage, 1999, Fotheringham, Brunson e Charlton, 2002, Fotheringham, Brunson e Charlton (2003) e Comber *et al.*, 2020, para mais detalhes acerca dessa classe de modelos). Em outras palavras, a REGP é expressa através da equação dada a seguir:

$$y_i = \beta_0(h_i, v_i) + \sum_{k=1}^p \beta_k(h_i, v_i) x_{ik} + \epsilon_i \quad (2.2)$$

em que (h_i, v_i) indica as coordenadas geográficas previamente especificadas para o i -ésimo local, $\beta_k(h_i, v_i)$ é o parâmetro da k -ésima variável explicativa para o i -ésimo local, x_{ik} o valor da k -ésima variável explicativa no i -ésimo local e $\epsilon_i \sim S(0, \phi(h_i, v_i))$ representa o erros mutuamente independentes. Note que, o modelo formulado por meio da equação (2.2) é equivalente a definição de n modelos locais, ou seja, temos um conjunto de parâmetros distintos para cada local no espaço. Por este fato esse modelo pode ter mais parâmetros

desconhecidos do que observações, tornando-o não identificável. Para contornar esse problema é necessário considerar um processo de estimação alternativo, o qual será descrito com detalhes na próxima seção.

2.2 PROCESSO DE ESTIMAÇÃO

De forma semelhante aos modelos de RGP, temos que um modelo de REGP é potencialmente não identificável. Isso ocorre devido ao fato de que o modelo pode ter mais parâmetros do que dados observados, como foi mencionado na seção anterior. Neste contexto, Fotheringham, Brunsdon e Charlton (2002) sugerem que uma alternativa razoável para resolver esse problema é utilizar o conceito de verossimilhança local, introduzido por Tibshirani e Hastie (1987). Para isso, considere que queremos obter a estimativa dos parâmetros em um ponto arbitrário no espaço (h_i, v_i) , que não é necessariamente um ponto observado. Essa estimativa pode ser aproximada de forma local por um modelo de regressão elíptico clássico, mas considerando apenas uma vizinhança próxima ao i -ésimo local e uma função de ponderação espacial. Com base nesse princípio, a seguir é apresentado o processo de estimação proposto nesta tese para os parâmetros do modelo de REGP, bem como algumas definições fundamentais para a obtenção do algoritmo utilizado.

2.2.1 Função de Ponderação Espacial

A função de ponderação espacial tem como objetivo ponderar cada ponto de dados por sua distância ao ponto de regressão. Nesse âmbito, Fotheringham, Brunsdon e Charlton (2002) expõem um enfoque para definir esta função usando o conceito de funções *kernel*. Em síntese, a função de ponderação espacial $w_{ij} = S(d_{ij}, b)$ é baseada em alguma medida de distância d_{ij} entre as regiões i e j dentro de uma janela b , para $i = 1, \dots, m$ e $j = 1, \dots, n$. Neste trabalho as medidas de d_{ij} são obtidas através da distância de Minkowski, a qual é descrita por:

$$d_{ij} = \left(|h_i - h_j|^\tau + |v_i - v_j|^\tau \right)^{1/\tau}$$

em que τ representa o valor da potência associada a essa medida de distância. Vale ressaltar que, se τ assumir o valor 2 obtemos a distância Euclidiana como caso particular dessa medida. Além disso, para o caso que as coordenadas geográficas são definidas por meio da latitude e longitude no globo terrestre assumimos a medida de distância do grande

círculo, ou seja,

$$d_{ij} = 12742 \arcsen \left(\sqrt{\sin^2 \left(\frac{h_i - h_j}{2} \right) + \cos(h_i) \cos(h_j) \sin^2 \left(\frac{v_i - v_j}{2} \right)} \right)$$

No que diz respeito a função de ponderação, Brunson, Fotheringham e Charlton, 1998 afirmam que sua escolha tem pouco impacto nas estimativas dos parâmetros de um modelo geograficamente ponderado, no entanto seleção do valor ótimo para o parâmetro b representa um ponto fundamental no processo de estimação. Com base nesse fato e sem perda de generalidade, consideramos as funções de ponderações geográficas com base nos *kernels* gaussiano e biquadrado, definidos respectivamente por:

$$w_{ij} = \exp \left\{ -\frac{1}{2} \left(\frac{d_{ij}}{b} \right)^2 \right\} \quad \text{e} \quad w_{ij} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{b} \right)^2 \right]^2, & \text{se } d_{ij} < b \\ 0, & \text{c.c.} \end{cases},$$

em que $b > 0$ representa um parâmetro de suavização. Vale ressaltar que outras medidas de distâncias também podem ser consideradas, como por exemplo a distância de Minkowski. Além disso, as funções *kernel* geralmente assumem que o valor de b é fixo para todos os locais. No entanto, Fotheringham, Brunson e Charlton (2002) observam que isso pode causar uma estimativa com grande dispersão em modelos geograficamente ponderados, uma vez que os dados podem estar mais esparsos em uma região e mais densos em outra. Uma alternativa é utilizar parâmetros de suavização distintos para cada um dos locais, adaptando-se assim as diferenças de densidade entre as regiões de estudo. Uma função que utiliza esse tipo de abordagem é denominada por *kernel* adaptativo e consiste em considerar um parâmetro de suavização distinto para cada um dos locais estudados. Nesse caso, o valor do parâmetro de suavização b_i associada ao i -ésimo local é obtida a partir do conceito de vizinho mais próximo, de tal forma que todos os locais devem contemplar a mesma quantidade de observações no processo de estimação. Nesse caso a função *kernel* irá se adaptar as diferenças de densidades dos locais de acordo com o comportamento dos dados. Vale ressaltar também que outras medidas de distância e funções *kernels* podem ser utilizadas para a obtenção das ponderações geográficas (para mais detalhes, veja Fotheringham, Brunson e Charlton, 2002).

2.2.2 Função de Verossimilhança Local

Assuma inicialmente a hipótese de que a superfície dos parâmetros é aproximadamente plana na vizinhança de qualquer local estudado. Dessa forma, considerando

o i -ésimo local, definimos a função de verossimilhança local para o vetor de parâmetros $\boldsymbol{\theta}(i) = (\beta_0(i), \dots, \beta_p(i), \phi(i))^\top$ como sendo:

$$WL(\boldsymbol{\theta}(i)|D) = \sum_{j=1}^n \left\{ -\frac{1}{2} \log(\phi(i)) + \log\{g(u_j)\} \right\} w_{ij}, \quad (2.3)$$

em que (i) representa o ponto espacial (h_i, v_i) , D representa a base de dados espaciais, $u_j = (y_j - \mu_j)^2 / \phi(i)$ e w_{ij} é a função de ponderação espacial em termos da distância entre os pontos espaciais (i) e (j) . Com base na definição de verossimilhança local em (2.3), denota-se o vetor $\mathbf{U}(\boldsymbol{\beta}(i), \phi(i))$ como sendo a função escore para o i -ésimo local, o qual pode ser expresso por:

$$\mathbf{U}(\boldsymbol{\beta}(i), \phi(i)) = \begin{pmatrix} \frac{\partial WL(\boldsymbol{\theta}(i)|D)}{\partial \boldsymbol{\beta}(i)} \\ \frac{\partial WL(\boldsymbol{\theta}(i)|D)}{\partial \phi(i)} \end{pmatrix} = \begin{pmatrix} \phi(i)^{-1} \mathbf{X}^\top \mathbf{D}(v) \mathbf{W}_i (\mathbf{y} - \boldsymbol{\mu}) \\ (2\phi(i))^{-1} \{ \phi(i)^{-1} Q_v(\boldsymbol{\beta}(i)) - \text{tr}(\mathbf{W}_i) \} \end{pmatrix},$$

em que $\mathbf{y} = (y_1, \dots, y_n)$ representa a variável resposta, $\mathbf{X}$ é a matriz modelo, $\boldsymbol{\beta}(i) = (\beta_0(i), \dots, \beta_p(i))^\top$ é o vetor dos parâmetros de regressão, $Q_v(\boldsymbol{\beta}(i)) = \{\mathbf{y} - \boldsymbol{\mu}\}^\top \mathbf{D}(v) \mathbf{W}_i \{\mathbf{y} - \boldsymbol{\mu}\}$, $\mathbf{D}(v) = \text{diag}(v_1^*, \dots, v_n^*)$, $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^\top$, $\mathbf{W}_i = \text{diag}(w_{i1}, \dots, w_{in})$, $v_i^* = -2W_g(u_i) = -2g'(u_i)/g(u_i)$ e $\text{tr}(\cdot)$ é o operador traço. Ademais, a matriz de informação de Fisher observada para o i -ésimo local, denotada por $\ddot{\mathbf{L}}(\boldsymbol{\beta}(i), \phi(i))$, é expressa por:

$$\ddot{\mathbf{L}}(\boldsymbol{\beta}(i), \phi(i)) = \begin{pmatrix} \frac{\partial^2 WL(\boldsymbol{\theta}(i)|D)}{\partial \boldsymbol{\beta}(i) \partial \boldsymbol{\beta}(i)} & \frac{\partial^2 WL(\boldsymbol{\theta}(i)|D)}{\partial \boldsymbol{\beta}(i) \partial \phi(i)} \\ \frac{\partial^2 WL(\boldsymbol{\theta}(i)|D)}{\partial \phi(i) \partial \boldsymbol{\beta}(i)} & \frac{\partial^2 WL(\boldsymbol{\theta}(i)|D)}{\partial \phi(i) \partial \phi(i)} \end{pmatrix} = \begin{pmatrix} \ddot{\mathbf{L}}_{11} & \ddot{\mathbf{L}}_{12} \\ \ddot{\mathbf{L}}_{21} & \ddot{\mathbf{L}}_{22} \end{pmatrix},$$

em que $\ddot{\mathbf{L}}_{11} = -\phi(i)^{-1} \mathbf{X}^\top \mathbf{D}(a) \mathbf{W}_i \mathbf{X}$, $\ddot{\mathbf{L}}_{12} = \ddot{\mathbf{L}}_{21}^\top = 2\phi(i)^{-2} \mathbf{X}^\top \mathbf{W}_i \mathbf{b}$, $\ddot{\mathbf{L}}_{21} = 2\phi(i)^{-2} \mathbf{b}^\top \mathbf{W}_i \mathbf{X}$, $\ddot{\mathbf{L}}_{22} = \phi(i)^{-2} \{ \text{tr}(\mathbf{W}_i)/2 + \mathbf{u}^\top \mathbf{D}(c) \mathbf{W}_i \mathbf{u} - \mathbf{e}^\top \mathbf{D}(v) \mathbf{W}_i \mathbf{e} \}$, $\mathbf{u} = (u_1, \dots, u_n)^\top$, $\mathbf{e} = (e_1, \dots, e_n)^\top$, $\mathbf{b} = (b_1, \dots, b_n)^\top$, $\mathbf{D}(a) = \text{diag}(a_1, \dots, a_n)$, $\mathbf{D}(c) = \text{diag}(c_1, \dots, c_n)$, $e_i = y_i - \mu_i$, $c_i = W'_g(u_i)$, $a_i = -2\{W_g(u_i) + 2u_i W'_g(u_i)\}$ e $b_i = \{W_g(u_i) + 2u_i W'_g(u_i)\}(y_i - \mu_i)$. Adicionalmente, a matriz de informação de Fisher esperada para o i -ésimo local é dada por:

$$\mathbf{K}(\boldsymbol{\beta}(i), \phi(i)) = -\mathbb{E}(\ddot{\mathbf{L}}(\boldsymbol{\beta}(i), \phi(i))) = \begin{pmatrix} \frac{4d_g}{\phi(i)} \mathbf{X}^\top \mathbf{W}_i \mathbf{X} & \mathbf{0}_{p \times 1} \\ \mathbf{0}_{1 \times p} & \frac{\text{tr}(\mathbf{W}_i)}{4\phi(i)^2} (4f_g - 1) \end{pmatrix},$$

em que $d_g = \mathbb{E}[W_g^2(U^2)U^2]$, $f_g = \mathbb{E}[W_g^2(U^2)U^4]$ e $U \sim S(0, 1)$. Com isso, obtidas essas quantidades é possível especificar um algoritmo de estimação local para os parâmetros do modelo, o qual é descrito com detalhes a seguir.

2.2.3 Algoritmo de Estimação

Note que não é possível obter uma forma fechada para os estimadores de máxima verossimilhança local, sendo necessário o desenvolvimento de algum algoritmo iterativo para a maximização da função definida em (2.3). Nesse caso, o método escore de Fisher pode ser facilmente aplicado para obter as estimativas de máxima verossimilhança local de $\beta(i)$ e $\phi(i)$. Com base nesse método, desenvolvemos um algoritmo iterativo para os parâmetros do modelo de REGP para o i -ésimo local, o qual é expresso por:

$$\beta(i)^{(m+1)} = (\mathbf{X}^\top \mathbf{D}(v^{(m)}) \mathbf{W}_i \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}(v^{(m)}) \mathbf{W}_i \mathbf{y} \quad (2.4)$$

e

$$\phi(i)^{(m+1)} = \frac{1}{\text{tr}(\mathbf{W}_i)} \{\mathbf{y} - \mathbf{X} \beta(i)^{(m+1)}\}^\top \mathbf{D}(v^{(m+1)}) \mathbf{W}_i \{\mathbf{y} - \mathbf{X} \beta(i)^{(m+1)}\}, \quad (2.5)$$

com $m = 0, 1, 2, \dots$, em que m representa o número atual da iteração do algoritmo de estimação. Os valores iniciais de $\beta(i)$ e $\phi(i)$ podem ser obtidos a partir do modelo de regressão elíptico descrito em (2.1) considerando a mesma distribuição dos erros aleatórios assumida para o ajuste da REGP.

2.2.4 Matriz de Projeção

Considerando a convergência do processo de estimação para os parâmetros do modelo de REGP é possível obter uma matriz de projeção $\mathbf{S}$, de tal forma que $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$. Essa quantidade é fundamental na obtenção de resultados posteriores importantes para validação do modelo ajustado. No entanto, esse conceito é válido apenas quando os locais estudados apresentarem as mesmas coordenadas geográficas que os pontos de dados observados. Em outras palavras, se os dados foram coletados em pontos no espaço que não representam diretamente as regiões utilizadas para o ajuste do modelos locais, então essa abordagem não será válida. Neste contexto, considere previamente os resultados obtidos para a matriz de projeção da RGP em Fotheringham, Brunsdon e Charlton (2002). Com base nesse procedimento, obtemos que a i -ésima linha da matriz $\mathbf{S}$ na REGP é dada por

$\mathbf{x}_i(\mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_i \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_i$, ou seja,

$$\mathbf{S} = \begin{pmatrix} \mathbf{x}_1(\mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_1 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_1 \\ \mathbf{x}_2(\mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_2 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_2 \\ \vdots \\ \mathbf{x}_m(\mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_m \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_m \end{pmatrix}.$$

Nesse caso, o número efetivo de parâmetros (NEP) do modelo de REGP é expresso por $\text{NEP} = 2\nu_1 - \nu_2$, com $\nu_1 = \text{tr}(\mathbf{S})$ e $\nu_2 = \text{tr}(\mathbf{S}^\top \mathbf{S})$. Essa medida é uma forma de representar a quantidade de parâmetros obtidos no ajuste do modelo geograficamente ponderado, levando em consideração as variações espaciais dos parâmetros de regressão. Por outro lado, essa medida não leva em consideração a quantidade efetiva dos parâmetros de dispersão locais. Esse fato pode gerar estatísticas subestimadas, pois ainda não há uma proposta para medir adequadamente o número dos parâmetros de dispersão efetivos. Uma alternativa é considerar a estimativa do modelo REGP com o parâmetro de dispersão ϕ fixo, ou seja, assumir que ϕ não varia no espaço. Este método é apresentado como uma solução parcial para o cálculo do número efetivo de parâmetros para modelos de regressão geograficamente ponderado (ver Da Silva e Rodrigues (2014) e Da Silva e Lima (2017) para maiores detalhes). De forma similar, propomos um modelo de REGP com dispersão global, assumindo apenas um parâmetro ϕ_g fixo para todas os locais. Dessa forma, ϕ_g é estimado diretamente do modelo de regressão elíptico definido em (2.1) sob a mesma distribuição dos erros aleatórios e mantido fixo no processo de estimação dos demais parâmetros do modelo REGP. Consequentemente, o número efetivo de parâmetros nesse caso é especificado por $\text{NEP} = 2\nu_1 - \nu_2 + 1$.

2.2.5 Testes de Hipóteses

Sob as condições usuais de regularidade (Cox e Hinkley, 1974, Cap. 9) temos que os estimadores de máxima verossimilhança obtidos por meio das equações (2.4) e (2.5) são consistentes e assintoticamente normais. Isto é, considerando o i -ésimo modelo local, temos que:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}(i) - \boldsymbol{\beta}(i)) \xrightarrow{\mathcal{D}} \mathcal{N}_p(\mathbf{0}, \mathbf{J}_{\boldsymbol{\beta}(i)\boldsymbol{\beta}(i)}^{-1})$$

e

$$\sqrt{n}(\hat{\phi}(i) - \phi(i)) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{J}_{\phi(i)\phi(i)}^{-1})$$

em que $\mathbf{J}_{\beta(i)\beta(i)} = \lim_{n \rightarrow \infty} (4d_g/n\phi(i)) \mathbf{X}^\top \mathbf{W}_i \mathbf{X}$ e $\mathbf{J}_{\phi(i)\phi(i)} = \lim_{n \rightarrow \infty} (\text{tr}(\mathbf{W}_i)/4n\phi(i)^2)(4f_g - 1)$. Consequentemente, esse resultado é utilizado para a implementação dos testes de hipóteses para a significância dos parâmetros locais. No entanto, autores como Byrne, Charlton e Fotheringham (2009) e Da Silva e Fotheringham (2016) ressaltam que os testes de hipóteses para os parâmetros locais de modelos geograficamente ponderados sofrem com a ocorrência incorreta de falsos positivos (erro do tipo I) induzidos por múltiplas hipóteses. Neste contexto, entende-se falso positivo por rejeitar a hipótese nula quando esta é verdadeira. Visto isso, os autores supracitados propõem diferentes correções para o nível descritivo nos testes para manter uma taxa de erro ξ adequada. Essa correção evita a ocorrência de falsos positivos nos testes de significância dos parâmetros locais. Com base nos resultados apresentados em Da Silva e Fotheringham (2016), desenvolvemos a correção do nível descritivo na REGP por meio da seguinte forma:

$$\alpha = \frac{\xi}{\frac{\text{NEP}}{p}},$$

em que α representa o nível descritivo ajustado com base em uma taxa de erro ξ , NEP representa o número efetivo de parâmetros definido anteriormente e p a quantidade de variáveis explicativas incluindo o intercepto.

2.2.6 Seleção do Parâmetro de Suavização

Como mencionado anteriormente, o valor do parâmetro de suavização b tem grande impacto nas estimativas de um modelo geograficamente ponderado. Dessa forma, para a seleção do valor ótimo de b , propomos a adaptação dos métodos usualmente considerados para essa classe de modelos. No contexto da RGP, Fotheringham, Brunson e Charlton (2002) propõem os métodos baseados na validação cruzada (CV) e no critério da informação de Akaike corrigido (AICc). Esses métodos são extensivamente considerados e adaptados para as demais classes de modelos geograficamente ponderados (veja Luo e Kanala (2008), Nakaya *et al.* (2005), Da Silva e Rodrigues (2014) e Da Silva e Lima (2017) a título de exemplo). Especificamente, o método da validação cruzada consiste em minimizar o erro de predição em função de b , o qual pode ser expresso, no contexto da REGP, por:

$$\text{CV}(b) = \sum_{i=1}^n \left(y_i - \hat{y}_{\neq i}^*(b) \right)^2,$$

em que $\hat{y}_{\neq i}^*$ é o valor ajustado para o i -ésimo local, excluindo o ponto i no processo de estimação. Em contrapartida, o critério da informação Akaike corrigido procura minimizar a seguinte medida em função de b :

$$AIC_c = 2\nu_1 - 2\ell + 2\frac{\nu_1(\nu_1 + 1)}{n - \nu_1 - 1}$$

em que $\ell = \sum_{i=1}^n \left\{ -\log(\hat{\phi}(i))/2 + \log\{g(\hat{u}_i)\} \right\}$ e ν_1 é uma medida associada ao número efetivo de parâmetros do modelo de REGP. Para a minimização de um desses critérios em função de b foi considerada a técnica da seção áurea, descrita no capítulo anterior.

2.3 MÉTODOS DE DIAGNÓSTICO

A análise residual e as medidas de diagnóstico são amplamente utilizadas para a validação dos modelo de regressão. Com o intuito de verificar a qualidade do ajuste e as suposições do modelo, desenvolvemos a extensão de algumas medidas comumente utilizadas na regressão clássica para o caso da REGP. Dessa forma, propomos avaliar o impacto que a i -ésima observação exerce no i -ésimo modelo local. No entanto, esta abordagem novamente será válida apenas quando os locais estudados apresentarem as mesmas coordenadas geográficas que os pontos de dados observados.

2.3.1 Análise de Resíduos

No contexto geral, a análise de resíduos é comumente considerada quando se tem o interesse em avaliar a adequabilidade dos modelos de regressão. Os resíduos permitem estudar as observações cujo comportamento não é descrito completamente pelo modelo ajustado, medindo a discrepância entre os valores observados da variável resposta e seus correspondentes valores ajustados. Ademais, a análise de resíduos pode ser utilizada para verificar suposições de homocedasticidade, existência de pontos discrepantes e adequação das suposições assumidas para os erros aleatórios. Na literatura, há uma ampla discussão e inúmeras propostas de resíduos para modelos com erros elípticos. Em particular, estendemos o cálculo dos resíduos padronizados para a classe dos modelos de REGP (para mais detalhes da construção desses resíduos veja Galea, Paula e Cysneiros, 2005 e Cysneiros e Paula, 2008, a título de exemplo). Nesse caso, o i -ésimo elemento dos

resíduos padronizados é definido pela seguinte expressão:

$$r_{ti} = \frac{y_i - \hat{\mu}_i}{(\xi\hat{\phi}(i))^{1/2}[1 - (4d_g\xi)^{-1}\hat{h}_{ii}]^{1/2}},$$

em que $\hat{h}_{ii}$ é o i -ésimo elemento diagonal da matriz $\mathbf{H}_i = \mathbf{X}(\mathbf{X}^\top \mathbf{W}_i \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}_i$.

Em contrapartida, os resíduos componente desvio têm sido extensivamente estudados em modelos de regressão devido às suas propriedades estatísticas. Essa medida é baseada na estatística da razão de verossimilhanças generalizada e sua utilização é considerada para as diferentes classes de modelos (veja, por exemplo, Pregibon *et al.*, 1981, Pierce e Schafer, 1986, Davison e Gigli, 1989, McCullagh e Nelder, 1989 e Cysneiros, Paula e Galea, 2007). Com base nesses estudos, obtemos que o i -ésimo elemento dos resíduos componente desvio para o modelo de REGP é expresso por:

$$r_{di} = \text{sinal}(r_i) \sqrt{w_{ij}} \left\{ 2 \log\{g(0)\} - 2 \log[g(r_i^2)] \right\}^{1/2},$$

em que $r_i = (y_i - \hat{\mu}_i) / \sqrt{\hat{\phi}(i)}$. Adicionalmente, Dunn e Smyth (1996) apresentam uma definição geral de resíduos com boas propriedades estatísticas, a qual é denominada por resíduos quantílicos. Para obtenção dessa medida para o caso dos modelos de REGP, considere inicialmente $F(y, \mu, \phi)$ como sendo a função de distribuição acumulada da variável aleatória y . Dessa forma, sabe-se que se F é uma função contínua, então $F(y, \mu, \phi)$ é uniformemente distribuída no intervalo unitário. Com base nesse resultado, temos que o i -ésimo resíduo quantílico pode ser expresso por:

$$r_{qi} = \Phi^{-1}\{F(y_i, \hat{\mu}_i, \hat{\phi}(i))\},$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada da normal padrão. Nesse caso, devido às propriedades dessa função, sabe-se que r_{qi} segue uma distribuição normal padrão. Por esse fato, os resíduos quantílicos permitem verificar, de forma simples, as suposições do modelo relacionadas aos erros aleatórios (para mais detalhes veja Dunn e Smyth, 1996).

2.3.2 Distância de Cook Generalizada

A distância de Cook generalizada foi introduzida por Cook (1977) e é usada para medir o efeito da exclusão de observações na estimativa dos parâmetros do modelo. No contexto da REGP, esta medida pode ser representada para o i -ésimo local pela seguinte equação:

$$DG(\hat{\boldsymbol{\theta}}_{\{i\}}(i)) = (\hat{\boldsymbol{\theta}}_{\{i\}}(i) - \hat{\boldsymbol{\theta}}(i))^\top \mathbf{C}(\hat{\boldsymbol{\theta}}_{\{i\}}(i) - \hat{\boldsymbol{\theta}}(i)),$$

em que $\hat{\boldsymbol{\theta}}_{\{i\}}(i)$ é o estimador de máxima verossimilhança para $\boldsymbol{\theta}(i)$ excluindo a i -ésima observação e $\mathbf{C}$ é uma matriz definida positiva e assintoticamente equivalente à informação de Fisher esperada para $\boldsymbol{\theta}(i)$. No entanto, o cálculo de $\hat{\boldsymbol{\theta}}_{\{i\}}(i)$ pode ter alto custo computacional e uma aproximação dessa estimativa é geralmente considerada. Uma alternativa introduzida por Pregibon *et al.* (1981) para calcular $\hat{\boldsymbol{\theta}}_{\{i\}}(i)$, denominada por aproximação de um passo, considera uma expansão em série de Taylor até segunda ordem da função de verossimilhança do modelo sem a i -ésima observação. A utilização dessa aproximação para os modelos de REGP, considerando o i -ésimo local, implica no seguinte resultado:

$$\hat{\boldsymbol{\theta}}_{\{i\}}(i) = \begin{pmatrix} \hat{\boldsymbol{\beta}}(i) - \hat{\phi}^{-1/2} V(r_i) r_i (\mathbf{X}^\top \mathbf{W}_i \mathbf{X})^{-1} x_i w_{ii} / [4d_g(1 - \hat{h}_{ii})] \\ \hat{\phi}(i) - 2\hat{\phi}(V(r_i) r_i^2 - 1) w_{ii} / [\text{tr}(\mathbf{W}_i \boldsymbol{\Delta}_i)(4f_g - 1)] \end{pmatrix}$$

em que $V(r_i) = -2W_g(r_i)$, $\boldsymbol{\Delta}_i = \text{diag}(\delta_1, \dots, \delta_n)$, $\delta_i = 1$ e $\delta_j = 0$ para todo $j \neq i$. No caso em que $\mathbf{C}$ é a matriz de informações Fisher esperada e usando a aproximação de $\hat{\boldsymbol{\theta}}_{\{i\}}(i)$, então a distância de Cook generalizada para a i -ésima observação pode ser expressa da seguinte maneira:

$$DG(\hat{\boldsymbol{\theta}}_{\{i\}}(i)) = \frac{V^2(\hat{r}_i) \hat{r}_i^2 w_{ij} \hat{h}_{ii}}{4d_g(1 - \hat{h}_{ii})^2} + \frac{\text{tr}(\mathbf{W}_i)(V(\hat{r}_i) \hat{u}_i^2 - 1)^2 w_{ij}^2}{\text{tr}(\mathbf{W}_i \boldsymbol{\Delta}_i)(4f_g - 1)},$$

Deste modo, sugerimos que valores altos para $DG(\hat{\boldsymbol{\theta}}_{\{i\}}(i))$ podem revelar possíveis observações influentes no modelo ajustado.

2.3.3 Medidas de Alavancagem

De forma resumida, medidas de alavancagem buscam medir a variação dos valores ajustados quando seus valores observados são afetados por pequenas variações. Como uma proposta inicial, é possível detectar pontos de alavanca pelo estudo da diagonal principal de $\mathbf{S}$, uma vez que a interpretação dessa matriz de projeção é preservada para modelos de regressão geograficamente ponderados (Fotheringham, Brunsdon e Charlton (2002)). Dessa forma, seja s_{ii} o i -ésimo elemento diagonal de $\mathbf{S}$, então valores altos de s_{ii} indicam que a i -ésima observação é um possível ponto de alavanca.

Por outro lado, Wei, Hu e Fung (1998) propõem uma medida mais geral para detectar pontos de alavanca, a qual é denominada por alavancagem generalizada. Essa medida é obtida através da derivada dos valores preditos em termos da variável resposta, a qual é dada por:

$$GL(\boldsymbol{\theta}) = \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{y}} = \left\{ D_{\boldsymbol{\theta}}(-\ddot{L}_{\boldsymbol{\theta}\boldsymbol{\theta}})^{-1} \ddot{L}_{\mathbf{y}\boldsymbol{\theta}} \right\} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}},$$

em que $\mathbf{D}_\theta = \partial \boldsymbol{\mu} / \partial \boldsymbol{\theta}^\top$ e $\ddot{\mathbf{L}}_{\theta\mathbf{y}} = \partial^2 L(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \mathbf{y}$. Nos modelos de REGP, calculamos a medida da alavancagem generalizada para o i -ésimo modelo local, a qual pode ser expressa através da seguinte forma matricial:

$$GL(\hat{\boldsymbol{\theta}}(i)) = \mathbf{X}(\mathbf{X}^\top \mathbf{D}(\hat{a}) \mathbf{W}_i \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_i + \frac{4}{E\phi(\hat{i})^3} \mathbf{S}_i \mathbf{D}(\hat{v})^{-1} \hat{\mathbf{b}} \hat{\mathbf{b}}^\top (\mathbf{I}_n - \mathbf{S}_i),$$

em que $\mathbf{S}_i = \mathbf{X}\{\mathbf{X}^\top \mathbf{D}(\hat{a}) \mathbf{W}_i \mathbf{X}\}^{-1} \mathbf{X}^\top \mathbf{D}(\hat{a})$. Nesse caso, valores grandes para o i -ésimo elemento diagonal de $GL(\hat{\boldsymbol{\theta}}(i))$ indica que a i -ésima observação é possivelmente um ponto de alavanca. As observações destacadas devem ser investigada, avaliando o seu impacto na estimação do modelo.

2.3.4 Influência Local

Essa metodologia, proposta por Cook (1986), tem como objetivo verificar possíveis distorções no ajuste do modelo sob a presença de pequenas perturbações. No contexto da REGP, propomos medidas de influência local com base em duas funções objetivos. No primeiro caso avaliamos a influência de observações na estimação dos parâmetros através da distância entre as verossimilhanças. Em seguida, estudamos a influência de observações na predição do modelo com base em outra medida de distância. Em cada função objetivo consideramos diferentes tipos de perturbações, as quais serão descritas com mais detalhes a seguir.

2.3.4.1 Influência local na estimação

Assuma inicialmente um modelo de REGP, cuja sua função de verossimilhança local $WL(\boldsymbol{\theta}(i))$ é fornecida em (2.3). Nesse caso, $WL(\boldsymbol{\theta}(i)|\boldsymbol{\omega})$ representa a função de verossimilhança local incluindo o vetor de perturbação $\boldsymbol{\omega}$. Suponha também que haja algum vetor $\boldsymbol{\omega}_0$ de não perturbação, satisfazendo $WL(\boldsymbol{\theta}|\boldsymbol{\omega}_0) = WL(\boldsymbol{\theta})$. Assim, a distância entre as verossimilhanças locais é dada por:

$$LD(\boldsymbol{\omega}) = 2\{WL(\hat{\boldsymbol{\theta}}(i)) - WL(\hat{\boldsymbol{\theta}}(i)|\boldsymbol{\omega})\} \geq 0,$$

em que $\hat{\boldsymbol{\theta}}(i)$ e $\hat{\boldsymbol{\theta}}(i)|_{\boldsymbol{\omega}}$ representam as estimativas de máxima verossimilhança local do modelo com e sem perturbação, respectivamente. No contexto da REGP, o cálculo da influência local também pode ser realizado com base na curvatura normal. Nesse caso, a medida da

curvatura normal para δ na direção $\mathbf{l}$, $\|\mathbf{l}\| = 1$, é dada por:

$$C_{(i)\mathbf{l}}(\boldsymbol{\theta}) = 2 \left| \mathbf{l}^\top \Delta_{(i)}^\top (\ddot{\mathbf{L}}_{\boldsymbol{\theta}\boldsymbol{\theta}})^{-1} \Delta_{(i)} \mathbf{l} \right|.$$

em que $\Delta_{(i)} = \partial^2 WL(\boldsymbol{\theta}, \boldsymbol{\omega}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\omega}^\top$ é uma matriz de perturbações para o i -ésimo local. Como sugerido por Cook (1986), pode-se obter a maior curvatura $C_{(i)\mathbf{l}_{max}}$, na direção $\mathbf{l}_{max}$, que representa o autovetor associado ao maior autovalor de $\Delta_{(i)}^\top (\ddot{\mathbf{L}}_{\boldsymbol{\theta}\boldsymbol{\theta}})^{-1} \Delta_{(i)}$. O gráfico de índice para $C_{(i)\mathbf{l}_{max}}$ pode revelar as observações mais influentes em $\boldsymbol{\theta}$. Outra abordagem para análise de diagnóstico foi proposta por Lesaffre e Verbeke (1998), que é baseado no gráfico de índice de $C_{(i)j} = C_{(i)l_j}$, em que l_j é um vetor formado por zeros com um na j -ésima posição para cada $j = 1, 2, \dots, n$. Nesse caso, consideramos como observações atípicas aquelas que são superiores a $2\bar{C}_{(i)}$ com $\bar{C} = \sum_{j=1}^n C_{(i)j} / n$.

Visto isso, de acordo com o esquema de perturbação utilizado, a matriz Δ é estimada por $\hat{\Delta} = \Delta|_{(\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}, \boldsymbol{\omega}=\boldsymbol{\omega}_0)}$. Dessa forma, foram considerados dois esquemas de perturbação na estimativa dos parâmetros, os quais são apresentados a seguir.

- Perturbação na escala – Considere o esquema de perturbação no parâmetro de dispersão, para o i -ésimo modelo local, na forma de $\phi(i)/\omega_i$. Com isso, temos que $\boldsymbol{\omega}_0 = (1, \dots, 1)$ é o vetor de não perturbação e a função de verossimilhança local assume a seguinte forma:

$$WL(\boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{j=1}^n \left\{ -\frac{1}{2} \log(\phi(i)/\omega_j) + \log\{g(\omega_j u_j)\} \right\} w_{ij}.$$

Com base nesse esquema, a matriz de perturbação para o i -ésimo modelo local é expressa sob a forma $\Delta_{(i)} = (\Delta_{\boldsymbol{\beta}(i)}^\top, \Delta_{\phi(i)}^\top)^\top$, com elementos dado por:

$$\Delta_{\boldsymbol{\beta}(i)} = -\frac{2}{\phi(i)} \mathbf{X}^\top \mathbf{W}_i \mathbf{D}(b) \quad \text{e} \quad \Delta_{\phi(i)} = -\frac{1}{\phi(i)^2} \mathbf{W}_i \mathbf{D}(b).$$

- Perturbação ponderação de casos – Nesse caso, seja o vetor de não perturbação $\boldsymbol{\omega}_0 = (1, \dots, 1)$ e o vetor de perturbação $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)$. Com isso, a função de verossimilhança local é perturbada por meio de uma ponderação sob seguinte forma:

$$WL(\boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{j=1}^n \omega_j \left\{ -\frac{1}{2} \log(\phi(i)) + \log\{g(u_j)\} \right\} w_{ij}.$$

Portanto, assumindo esse esquema a matriz de perturbação para o i -ésimo modelo local é expressa por $\Delta_{(i)} = (\Delta_{\boldsymbol{\beta}(i)}^\top, \Delta_{\phi(i)}^\top)^\top$, com elementos dado por:

$$\Delta_{\boldsymbol{\beta}(i)} = \frac{1}{\phi(i)} \mathbf{D}(v) \mathbf{W}_i \mathbf{D}(e) \mathbf{X} \quad \text{e} \quad \Delta_{\phi(i)} = \frac{1}{2\phi(i)} \mathbf{W}_i \mathbf{V}(m).$$

2.3.4.2 Influência local na predição

Outra maneira de estudar a influência das observações no modelo é através do seu impacto na predição. Nesse caso, considere inicialmente o vetor $\mathbf{q}$ com dimensão $p \times 1$ de valores das variáveis explicativas, para o qual não temos necessariamente uma resposta observada. Com isso, a função de predição em $\mathbf{q}$ é expressa por:

$$\hat{\boldsymbol{\mu}}(\mathbf{q}) = \sum_{k=1}^p q_j \hat{\beta}(i)_k.$$

De maneira análoga, a predição em $\mathbf{q}$ com base no modelo perturbado é expresso pela seguinte equação:

$$\hat{\boldsymbol{\mu}}(\mathbf{q}, \boldsymbol{\omega}) = \sum_{k=1}^p q_j \hat{\beta}(i)_{k\boldsymbol{\omega}}.$$

Dessa modo, assumindo a função objetivo $f(\mathbf{q}, \boldsymbol{\omega}) = \{\hat{\boldsymbol{\mu}}(\mathbf{q}) - \hat{\boldsymbol{\mu}}(\mathbf{q}, \boldsymbol{\omega})\}^2$, então a curvatura normal na direção unitária $\mathbf{l}$ assume a forma:

$$\mathbf{C}_{(i)\mathbf{l}} = |\mathbf{l}^\top \ddot{f} \mathbf{l}|,$$

em que $\ddot{f} = \partial^2 f(\mathbf{q}, \boldsymbol{\omega}) / \partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^\top$ é avaliada em $\boldsymbol{\omega}$ e $\hat{\beta}_{(i)}$. Logo, calculando-se essa derivada obtemos o seguinte resultado:

$$\ddot{f} = \frac{\partial^2}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^\top} f = -2 \Delta_{(i)}^\top \ddot{\mathbf{L}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \mathbf{q} \mathbf{q}^\top \ddot{\mathbf{L}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \Delta_{(i)}.$$

Novamente, pode-se utilizar a maior curvatura, na qual tem a direção $\mathbf{l}_{max}$, como forma de avaliar a influência das observações na predição do modelo. Os dois esquemas de perturbação a seguir são considerados para a avaliar o impacto das observações na predição do modelo.

- Perturbação aditiva na resposta – Considere a perturbação da variável resposta na forma $y_{i\boldsymbol{\omega}} = y_i + \omega_i s_i$, em que s_i é a estimativa do desvio padrão de y_i . Com base nisso, temos que $\boldsymbol{\omega}_0 = (0, \dots, 0)$ é o vetor de não perturbação e a função de verossimilhança local é expressa por:

$$WL(\boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{j=1}^n \left\{ -\frac{1}{2} \log(\phi(i)) + \log\{g(u_{j\boldsymbol{\omega}})\} \right\} w_{ij}.$$

Nesse caso, a matriz de perturbação para o i -ésimo local é expressos pela seguinte forma matricial:

$$\Delta_{\beta(i)} = \frac{1}{\phi(i)} \mathbf{X}^\top \mathbf{W}_i \mathbf{D}(a).$$

- Perturbação aditiva nos preditores – Considere a perturbação da k -ésima variável explicativa contínua na forma $x_{ik\omega} = x_{ik} + \omega_i s_k$, com s_k representando um fator de escala. Novamente sabe-se que $\omega_0 = (0, \dots, 0)$ é o vetor de não perturbação e a função de verossimilhança local é dada por:

$$WL(\theta, \omega) = \sum_{j=1}^n \left\{ -\frac{1}{2} \log(\phi(i)) + \log\{g(u_{j\omega})\} \right\} w_{ij}.$$

Logo, os elementos da matriz de perturbação para o i -ésimo modelo local nesse caso é expressos por:

$$\Delta_{\beta(i)} = s_k(\mathbf{F}\mathbf{D}(g)\mathbf{D}(e) - \phi(i)^{-1}\beta(i)_k\mathbf{X}^\top\mathbf{D}(a))\mathbf{W}_i,$$

em que $\mathbf{F}$ é uma matriz $p \times n$ de zeros com o valor 1 na k -ésima linha.

2.4 ESTUDO DE SIMULAÇÃO

Os estudos de simulação propostos nessa seção têm a finalidade de verificar os desempenhos dos modelos de REGP. De forma geral, o propósito desse estudo de simulação é responder as seguintes dúvidas:

- Os modelos de regressão geograficamente ponderados conseguem reproduzir as superfícies locais dos parâmetros reais de maneira satisfatória?
- De que forma a presença de observações com valores discrepantes pode afetar a superfícies estimada dos parâmetros em modelos de regressão geograficamente ponderados?
- A utilização de distribuições com caudas mais pesadas para os erros aleatórios gera estimativas espaciais menos sensíveis na presença de observações com valores discrepantes?
- Os resultados obtidos são válidos para estruturas espaciais com base em grades regulares e irregulares?

Visto isso, desenvolvemos estudos de simulações com propriedades conhecidas e uma superfície espacial bem definida para os parâmetros de regressão. Como forma de comparar desempenho dos modelos geograficamente ponderados consideramos o uso de duas medidas específicas. Neste âmbito, para mensurar a capacidade dos modelos RGP e REGP em replicar a superfície dos parâmetros conhecidos, utilizamos a raiz do erro

quadrático médio (REQM), a qual é dada por:

$$\text{REQM}_j = \sqrt{\frac{1}{n} \sum_{i=1}^n \left(\beta_j(h_i, v_i) - \hat{\beta}_j(h_i, v_i) \right)^2}$$

em que $\hat{\beta}_j(h_i, v_i)$ é a estimativa do j -ésimo coeficiente para o i -ésimo local. Por outro lado, a qualidade do ajuste para ambos os modelos é medida por meio da soma dos quadrados residuais ponderados (SQRp), definida através da seguinte forma:

$$\text{SQRp} = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{v_i^{*-1}}$$

em que $\hat{y}_i$ é a estimativa da variável resposta para o i -ésimo local e $v_i^* = -2g'(u_i)/g(u_i)$. Dessa forma, nas seções a seguir são apresentados os resultados obtidos com base nessas medidas para dois cenários simulados distintos.

2.4.1 Cenário 1 - Grade Regular

Para avaliar o desempenho dos modelos de REGP em dados com observações discrepantes, desenvolvemos inicialmente um estudo de simulação com base em uma grade regular de dimensões 25×25 para a estrutura espacial. Diante disso, simulamos 100 replicações com base em um modelo geograficamente ponderado, assumindo o seguinte processo para a geração dos dados:

$$y_i = \beta_0(h_i, v_i) + \beta_1(h_i, v_i)x_i + \epsilon_i \quad (2.6)$$

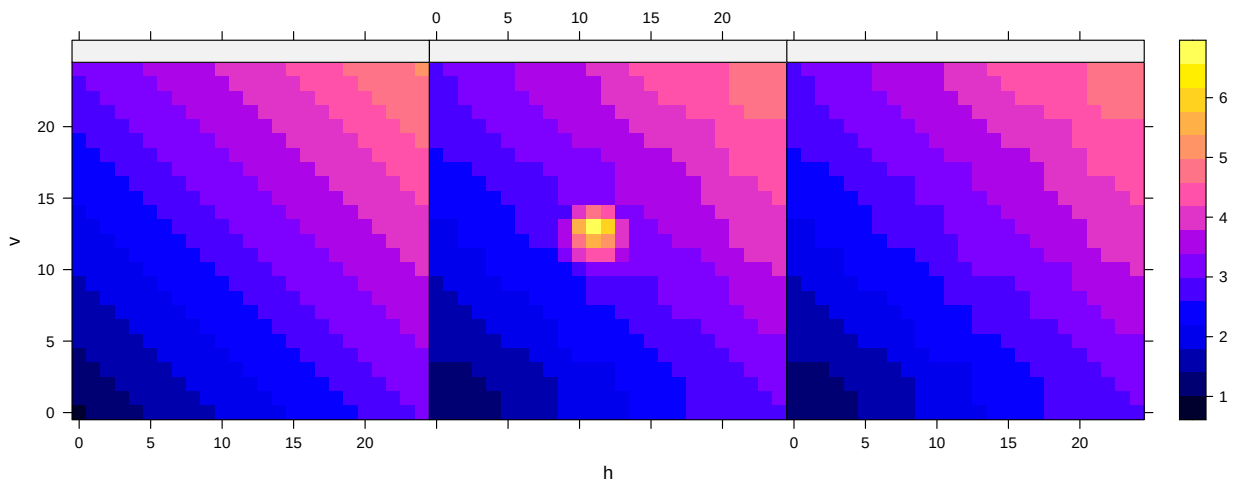
em que $x = (x_1, \dots, x_n)$ e $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ foram gerados a partir de uma distribuição normal padrão. Nesse caso, as coordenadas (h_i, v_i) representam o ponto central da região (definida por um polígono retangular de área um) associada à i -ésima observação, em que $h_i = 0, 1, \dots, 25$ e $v_i = 0, 1, \dots, 25$. Para os coeficientes de regressão, projetamos duas superfícies de parâmetros diferentes na mesma escala de variação com base nas seguintes regras:

$$\beta_0(h_i, v_i) = 1 + \frac{1}{12}(h_i + v_i^+) \quad \text{e} \quad \beta_1(h_i, v_i) = 1 + \frac{1}{12}(h_i + v_i^-)$$

em que v_i^+ e v_i^- representam i -ésimo valor em ordem crescente e decrescente de $v = (v_1, \dots, v_n)$, respectivamente. Para introduzir uma observação discrepante foi selecionado o local referente às coordenadas (11,12) e substituímos o valor da variável resposta pelo seu valor máximo mais três vezes o desvio padrão. Feito isso, ajustou-se os modelos

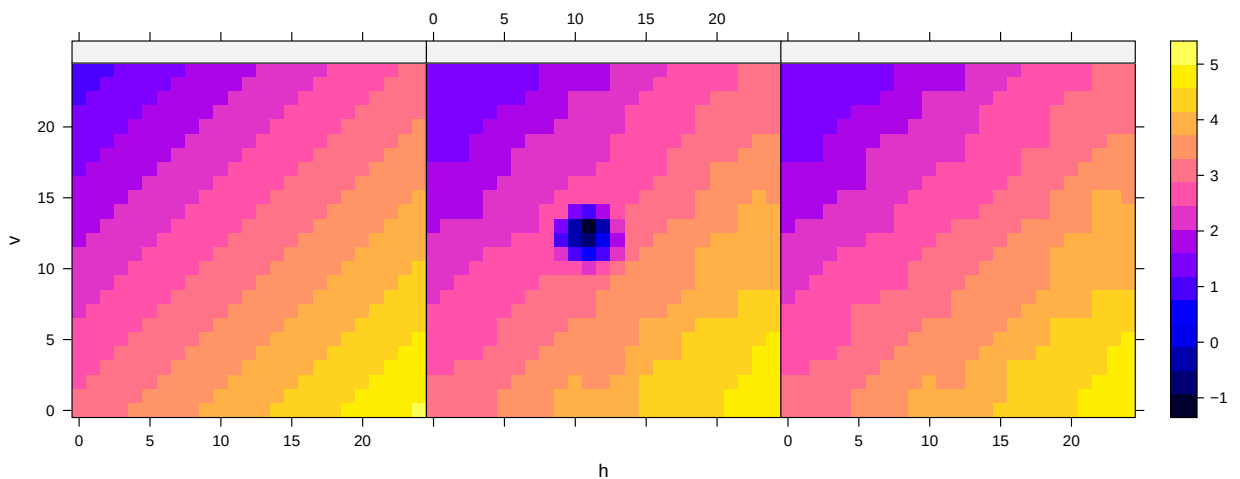
de RGP e REGP assumindo a distribuição t -Student com 3 graus de liberdades (g.l.) para os erros aleatórios. No processo de estimação considerou-se o uso da função kernel gaussiano adaptativo e a distância Euclidiana para as ponderações espaciais. A otimização do parâmetro de suavização foi realizada através dos métodos CV e AICc. Baseado nesses ajustes, as Figuras 3-6 mostram as superfícies espaciais para as médias das estimativas dos parâmetros de regressão, considerando ambos os métodos para a seleção do parâmetro de suavização ótimo. Nesse caso, as médias das estimativas obtidas nos modelos geograficamente ponderados são comparadas com seus respectivos valores reais.

Figura 3 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 da RGP (meio) e REGP com erros t -Student (direita) para o método CV.



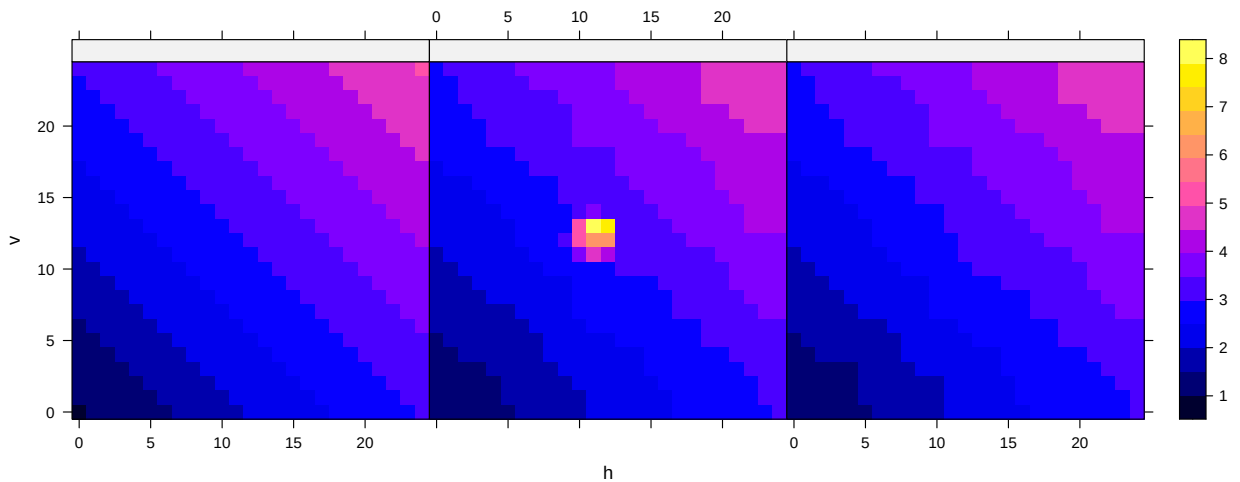
Fonte: O autor (2021).

Figura 4 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 da RGP (meio) e REGP com erros t -Student (direita) para o método CV.



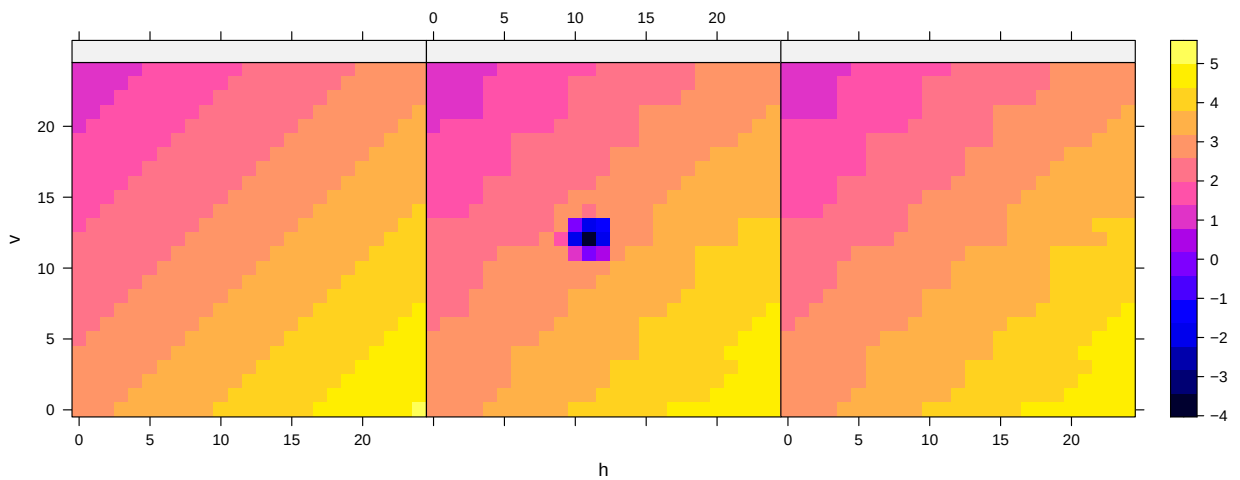
Fonte: O autor (2021).

Figura 5 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 da RGP (meio) e REGP com erros t -Student (direita) para o método AICc.



Fonte: O autor (2021).

Figura 6 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 da RGP (meio) e REGP com erros t -Student (direita) para o método AICc.

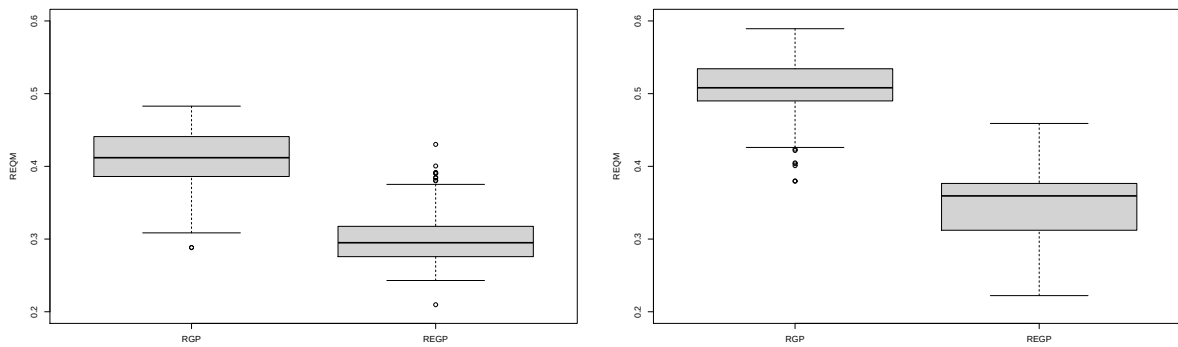


Fonte: O autor (2021).

Note que há uma distorção dos parâmetros estimados pela RGP apenas nas regiões próximas à observação discrepante. No entanto, esse comportamento não é observado nas estimativas obtidas através do ajuste da REGP com erros t -Student. Observou-se também que o número de vizinhos considerados para os valores ótimos obtidos dos parâmetros de suavização em ambos métodos de seleção foram pequenos em relação ao número de observações, apresentando uma baixa variabilidade entre as simulações. Este fato indica que ambos modelos selecionam valores que representam bem a variação espacial real dos parâmetros de regressão. Além disso, a Figura 7 e 8

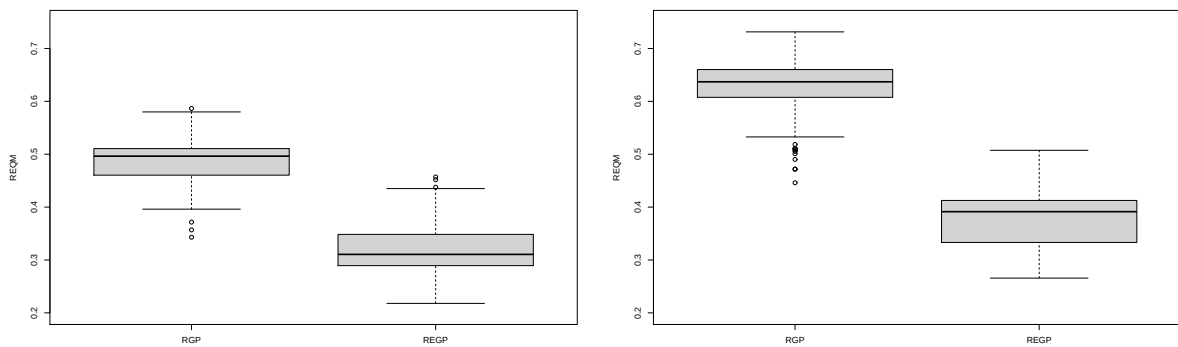
mostram os *boxplots* da REQM para ambos modelos com base nas estimativas de β_0 e β_1 , respectivamente. Note que, em geral, o modelo de REGP com erros t -Student mostra uma maior precisão para replicar as superfícies espaciais reais de ambos parâmetros. Além disso, conforme mostrado na Figura 9, o modelo REGP com erros t -Student possui também as melhores estatísticas para os valores da SQRp. Todos esses resultados indicam que a distribuição t -Student permite o ajuste de um modelo geograficamente ponderado menos sensível à presença de observações discrepantes.

Figura 7 – *Boxplots* da REQM para β_0 nos modelos geograficamente ponderados para os métodos CV (esquerda) e AICc (direita)



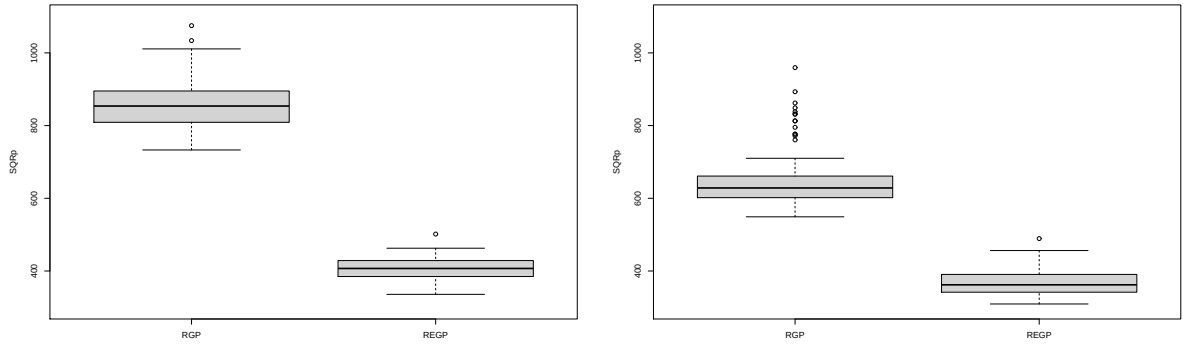
Fonte: O autor (2021).

Figura 8 – *Boxplots* da REQM para β_1 nos modelos geograficamente ponderados para os métodos CV (esquerda) e AICc (direita)



Fonte: O autor (2021).

Figura 9 – *Boxplots* da SQRp dos modelos geograficamente ponderados para os métodos CV (esquerda) e AICc (direita)



Fonte: O autor (2021).

2.4.2 Cenário 2 - Grade Irregular

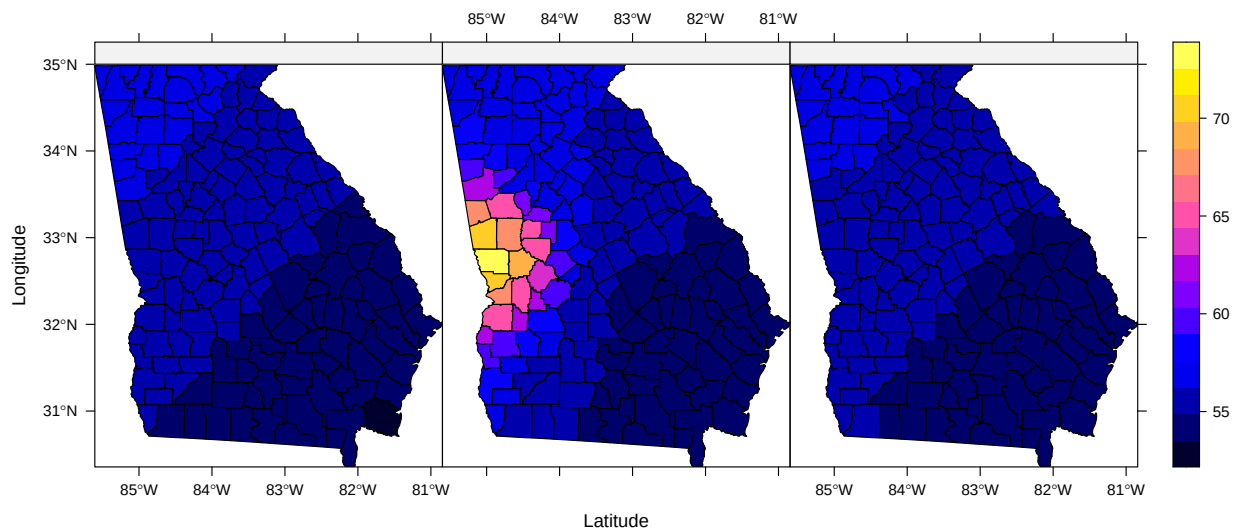
O objetivo desse cenário é avaliar se há diferenças nos resultados obtidos anteriormente ao utilizar uma estrutura espacial com base em uma grade irregular. Como forma de tornar o processo de simulação mais realista, usamos um esquema espacial baseado no mapa dos 159 condados da Geórgia, EUA. Com base nesse esquema foi realizado um estudo com 1000 replicações através do mesmo processo de geração definido em (2.6) utilizado para o primeiro cenário simulado. No entanto, para esse estudo as coordenadas (h_i, v_i) representam a latitude e longitude dos centroides de cada um dos condados, onde $\mathbf{h} = (h_1, \dots, h_n)$ e $\mathbf{v} = (v_1, \dots, v_n)$. As informações geográficas dos condados da Geórgia foram retirados de um estudo comumente utilizada para a aplicação de modelos geograficamente ponderados, o qual será descrito com mais detalhes no seguinte capítulo. Para os coeficientes de regressão, projetamos novamente duas superfícies de parâmetros diferentes na mesma escala de variação por meio das equações dadas a seguir:

$$\beta_0(h_i, v_i^+) = 1 + \frac{1}{r}(h_i + v_i^+) \quad \text{e} \quad \beta_1(h_i, v_i^-) = 1 + \frac{1}{r}(h_i + v_i^-)$$

em que $r = [\max(\mathbf{v}) - \min(\mathbf{v}) - \max(\mathbf{u}) - \min(\mathbf{u})]/4$, e v_i^+ e v_i^- representam o i -ésimo valor em ordem crescente e decrescente de $\mathbf{v} = (v_1, \dots, v_n)$, respectivamente. Para introduzir uma observação discrepante foi escolhida arbitrariamente o condado de Harris com coordenadas $(-83.6882, 33.3159)$ para o centroide dessa região. Com base nessas coordenadas, observa-se que esse condado esta situado na região centro-oeste do estado da Georgia. Em cada um das réplicas, substituímos a variável resposta pelo seu valor máximo com um acréscimo de três vezes o seu desvio padrão. Em seguida, ajustou-se os modelos de RGP e REGP supondo a distribuição t -Student com 3 graus de liberdade para os erros aleatórios. As

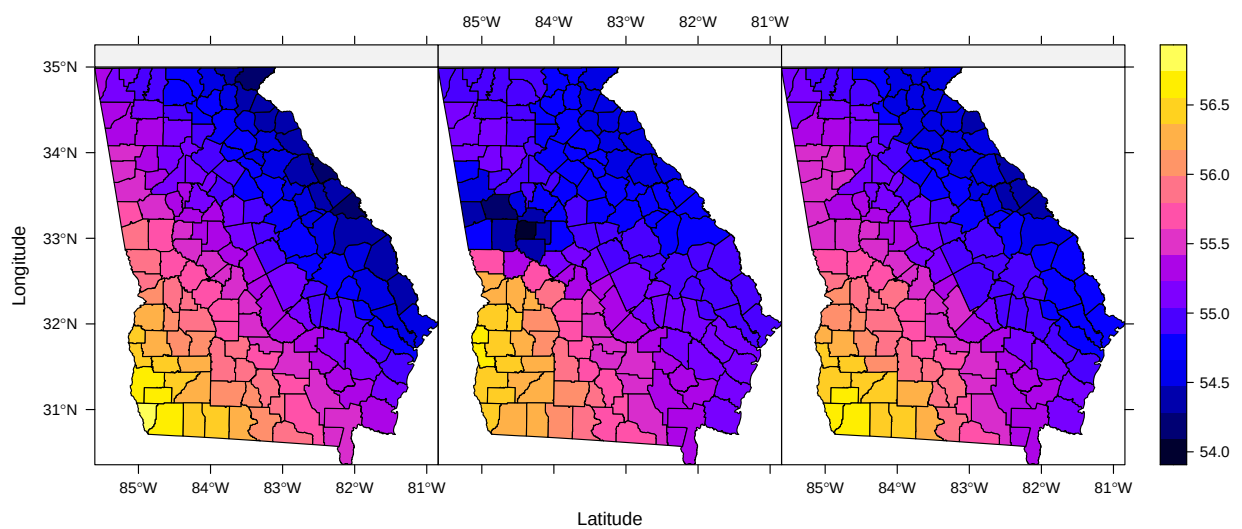
ponderações espaciais foram calculadas com base em um *kernel* gaussiano adaptativo e na medida de distância do grande círculo, uma vez que as coordenadas geográficas são representadas por meio da latitude e longitude. Com base nessas condições, as Figuras 10-13 mostram as superfícies espaciais para as médias das estimativas dos parâmetros de regressão, considerando ambos os métodos para a otimização do parâmetro de suavização.

Figura 10 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 do RGP (meio) e REGP com erros *t*-Student (direita) para o método CV.



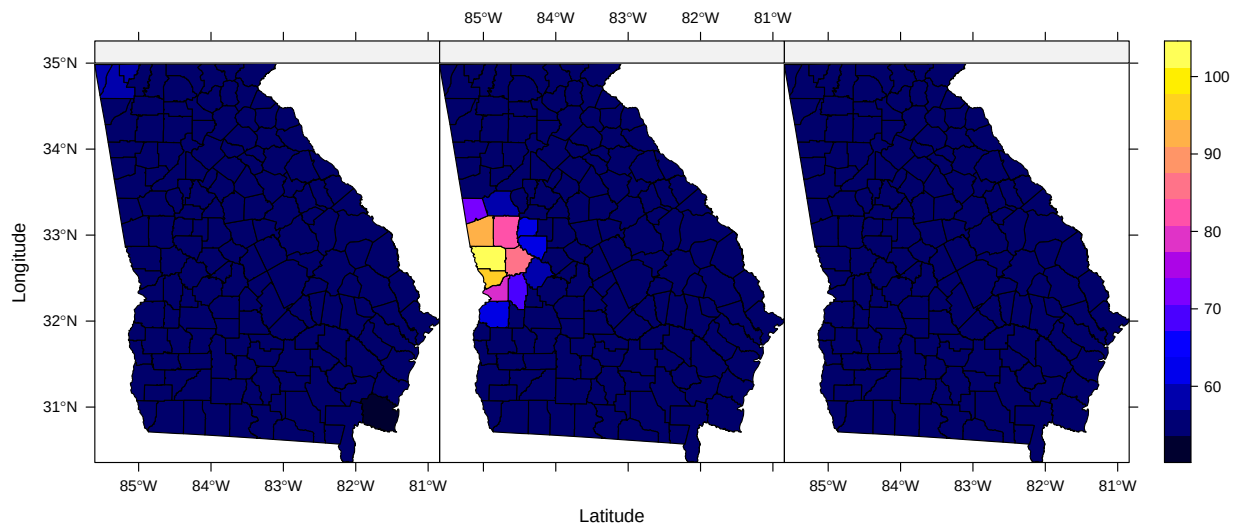
Fonte: O autor (2021).

Figura 11 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 do RGP (meio) e REGP com erros *t*-Student (direita) para o método CV.



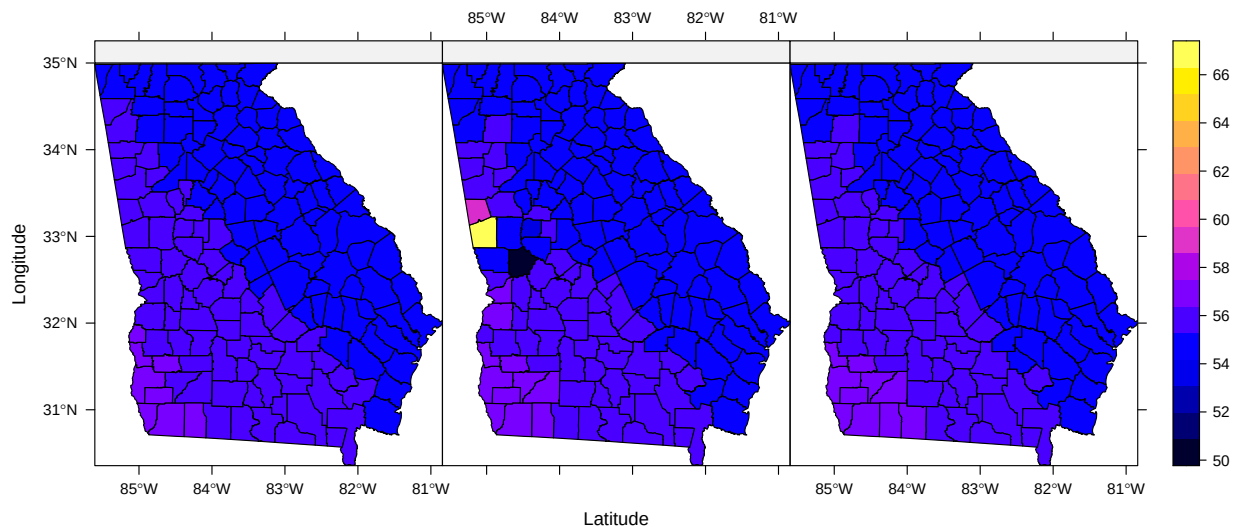
Fonte: O autor (2021).

Figura 12 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 do RGP (meio) e REGP com erros t -Student (direita) para o método AICc.



Fonte: O autor (2021).

Figura 13 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 do RGP (meio) e REGP com erros t -Student (direita) para o método AICc.



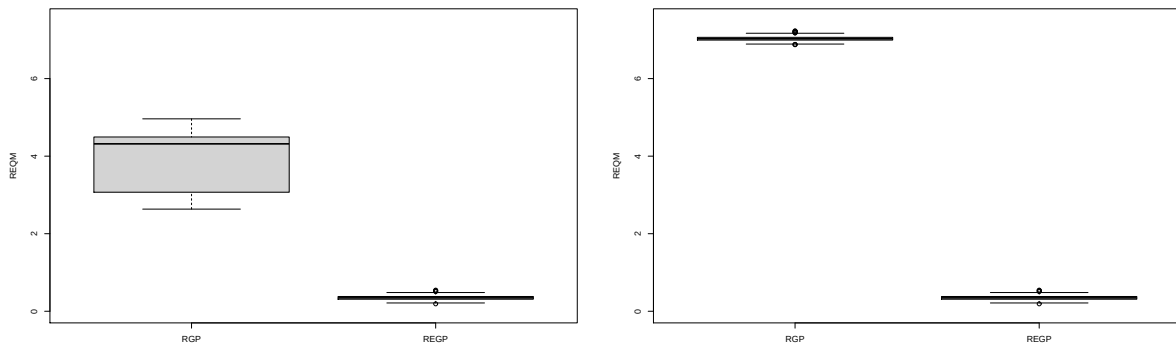
Fonte: O autor (2021).

Novamente, observa-se que há uma distorção da superfície espacial em ambos parâmetros de regressão estimados pela RGP nas regiões próximas ao condado de Harris, ou seja, os valores estimados dos parâmetros de regressão para alguns condados da região centro-oeste do estado da Georgia são bem diferentes dos seus respectivos valores reais. Essa distorção é ainda mais clara na distribuição espacial do intercepto, apresentando valores muito mais altos que seus respectivos parâmetros locais reais. Ressaltando novamente que o ajuste da RGP é sensível à presença de observações discrepantes, de forma semelhante aos resultados obtidos no cenário anterior. Entretanto, esse comportamento não é observado

nas estimativas obtidas através do ajuste da REGP com erros t -Student. Nesse caso, o uso do modelo REGP em uma grade espacial irregular obteve também ajustes com menor sensibilidade a observação discrepante, apresentando uma superfície média estimada bem próxima dos valores reais. Observou-se também que o modelo RGP apresenta uma grande variabilidade da quantidade de vizinhos associados aos valores dos parâmetros de suavização ótimos obtidos através do método de validação cruzada.

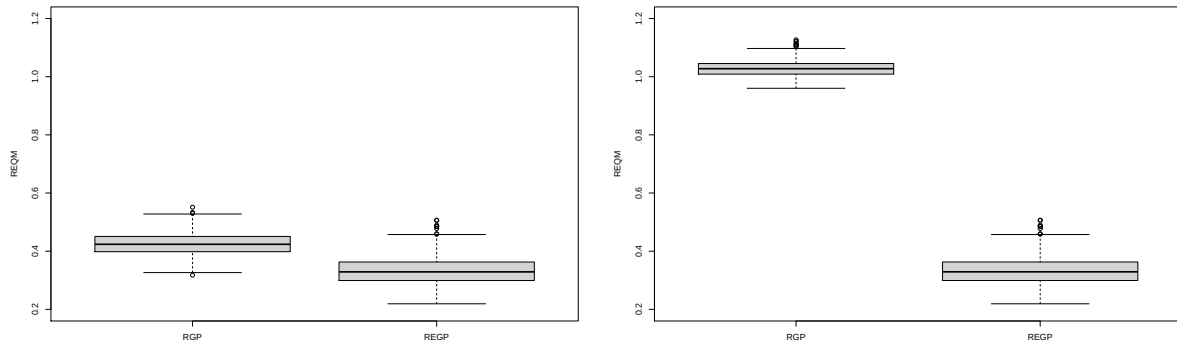
Por outro lado, o modelo REGP com erros t -Student obtém um número de vizinhos bem menores em relação ao número de observações para os valores ótimos e pouca variabilidade entre as simulações, permitindo representar melhor a variação geográfica dos parâmetros. Já para o caso do método via AICc, ambos modelos selecionam um número pequeno de vizinhos para os valores ótimos, obtendo pouca variabilidade entre as simulações. Neste caso, podemos observar que a seleção do parâmetro de suavização tem consequências claras na superfície dos parâmetros de regressão estimados e o método CV tem maior dificuldade em selecionar um parâmetro de suavização ideal em dados com presença de observações discrepantes. Por fim, as Figuras 14 e 15 exibem os *boxplots* da REQM para as estimativas de β_0 e β_1 , respectivamente. Novamente o modelo de REGP com erros t -Student mostra uma maior precisão para replicar as superfícies espaciais reais de ambos parâmetros. Por fim, conforme mostrado na Figura 16, o modelo REGP com erros t -Student possui as melhores estatísticas para os valores da SQRp. Esses resultados são semelhantes ao cenário anterior, mostrando que a distribuição t -Student permite o ajuste de um modelo geograficamente ponderado menos sensível à presença de observações discrepantes também para casos com estruturas espaciais irregulares.

Figura 14 – *Boxplots* da REQM dos modelos geograficamente ponderados para os métodos CV (esquerda) e AICc (direita) na superfície de β_0



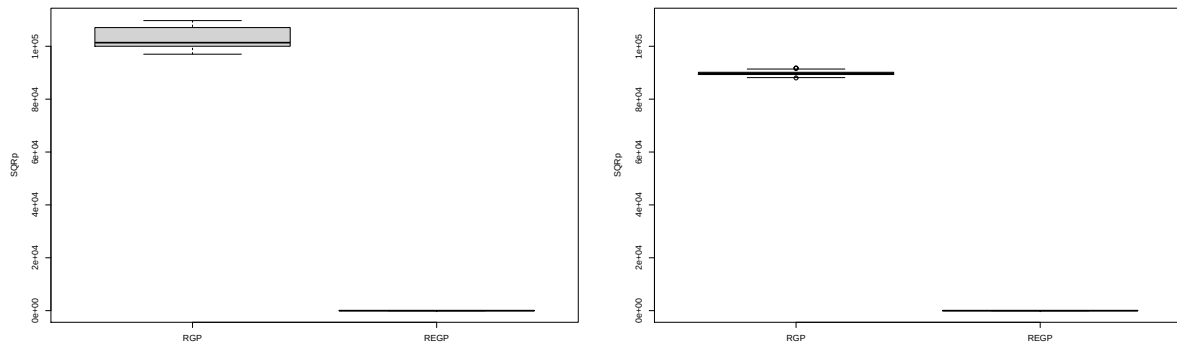
Fonte: O autor (2021).

Figura 15 – *Boxplots* da REQM dos modelos geograficamente ponderados para os métodos CV (esquerda) e AICc (direita) na superfície de β_1



Fonte: O autor (2021).

Figura 16 – *Boxplots* da SQRp dos modelos geograficamente ponderados para os métodos CV (esquerda) e AICc (direita)



Fonte: O autor (2021).

2.5 APLICAÇÃO

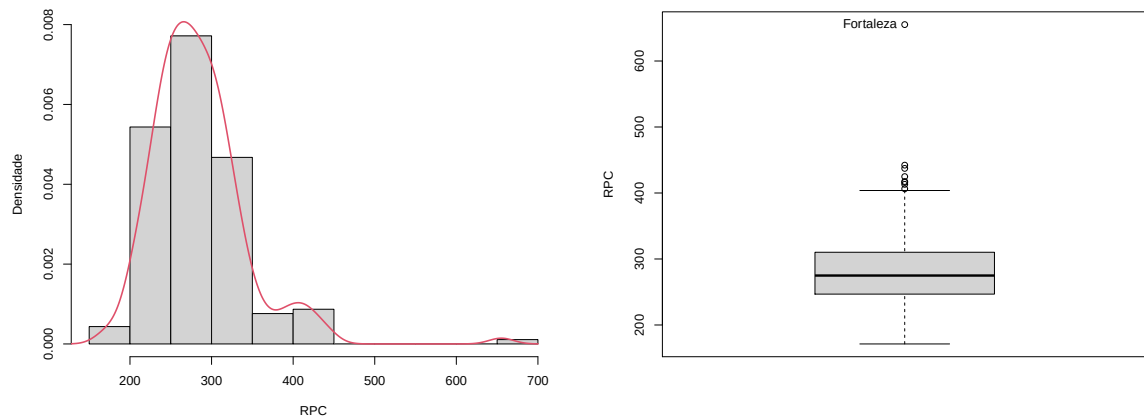
Para ilustrar os métodos propostos nesse capítulo analisamos uma base de dados referente ao estudo do desenvolvimento humano e criminalidade no ano de 2010 para os 184 municípios do Ceará, Brasil. Este estudo considera três variáveis: renda per capita média do 4º quinto mais pobre (RPC), taxa de homicídios por 100.000 habitantes (TH) e percentual de crianças que vivem em domicílios onde nenhum dos moradores tem ensino fundamental completo (PCF). Vale ressaltar que, as observações da variável RPC representa a média da renda per capita mensal de uma parcela da população pobre associada a cada município durante o ano de 2010. Nesse caso, a renda per capita é obtida pela soma das renda pelo total de habitantes no respectivo município. As informações contidas na base de dados foi extraída do estudo censitário de 2010 realizado no Brasil, disponível através do “Atlas do Desenvolvimento Humano” e do “Atlas da Violência”. Esses dados foram produzidos pelo Instituto de Pesquisa Econômica Aplicada, Fórum

Brasileiro de Segurança Pública e Fundação João Pinheiro. As informações geoespaciais foram obtidas por meio do pacote **geobr** presente no *software* R. Esse pacote facilita o acesso de dados espaciais fornecido pelo Instituto Brasileiro de Geografia e Estatística, o qual fornece diversas informações geográficas para diferentes regiões do Brasil. O objetivo desta aplicação é avaliar a relação entre renda de uma parcela específica da população pobre com os indicadores educacionais e criminais no estado do Ceará.

2.5.1 Análise Exploratória

Inicialmente consideramos uma breve análise exploratória e um estudo de correlação das variáveis apresentadas anteriormente. Dessa forma, o histograma e o *boxplot* da RPC são exibidos na Figura 17. Note que essa variável apresenta um comportamento aproximadamente simétrico, porém há presença de algumas observações com valores discrepantes. Nesse caso, destaca-se o município de Fortaleza por apresentar um valor extremamente discrepante da renda per capita do 4º quinto mais pobre em comparação com os demais municípios.

Figura 17 – Histograma com densidade aproximada pela linha vermelha (esquerda) e *boxplot* (direita) da RPC



Fonte: O autor (2021).

Na Tabela 1 são apresentadas algumas estatísticas descritivas para as variáveis RPC, TH e PCF. Observe que, em todas variáveis temos uma leve assimetria e há uma pequena variabilidade dos valores da RPC e do PCF. Além disso, destaca-se um alto valor para a curtose da RCP, indicando que uma distribuição com caudas mais pesadas pode descrever melhor o comportamento dessa variável.

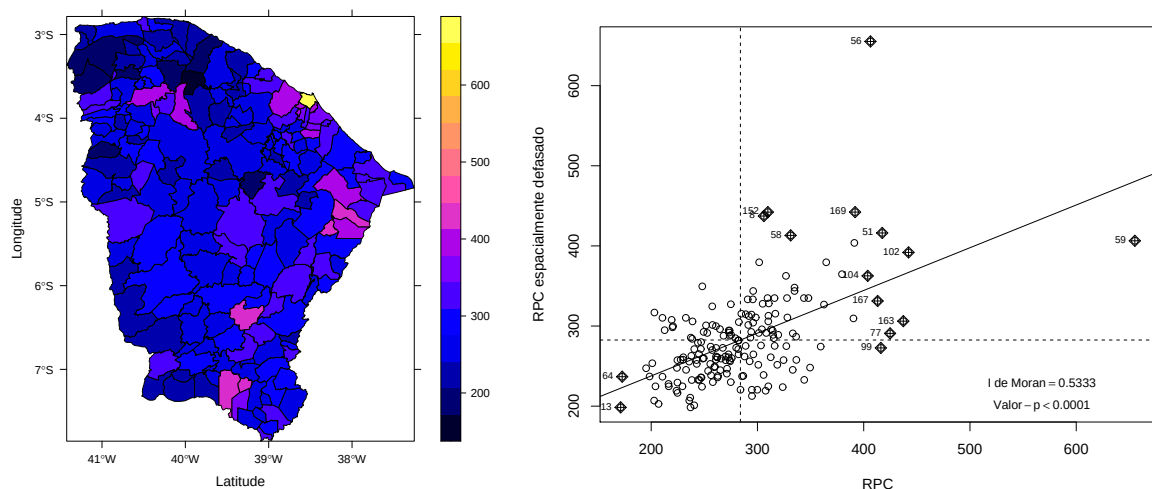
Tabela 1 – Análise descritiva dos indicadores de desenvolvimento humano e criminais nos municípios cearenses

	Minimo	Mediana	Média	Máximo	CV	Assimétria	Curtose
RPC	171,27	274,92	283,97	655,16	0,20	1,85	11,18
TH	0,00	17,41	20,73	67,28	0,71	0,80	3,26
PCF	21,24	41,28	41,06	62,13	0,18	-0,07	2,82

Fonte: O autor (2021).

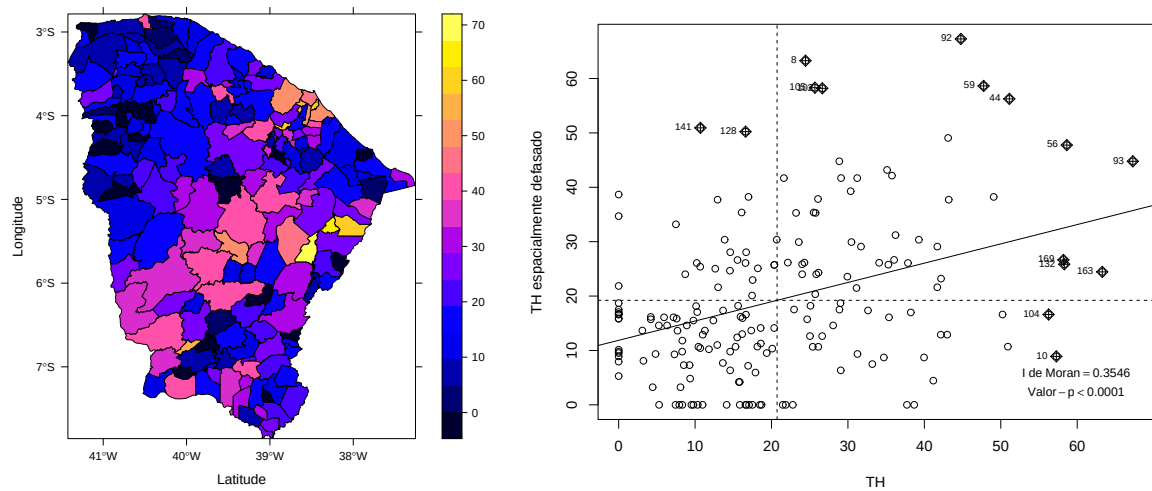
Como forma de avaliar possíveis dependências espaciais presente nas variáveis estudadas, as Figuras 18-20 mostram os comportamentos espaciais das variáveis RPC, TH e PCF nos municípios do Ceará. Além disso, são destacados os índices de Moran e os valores- p para o teste de autocorrelação espacial nos gráficos apresentados. Note que a RCP possui os maiores valores nos municípios ao nordeste, os quais fazem parte da região metropolitana do estado. Esse mesmo comportamento é observado na TH, no entanto o PCF tem os menores valores na região nordeste. Isso dá uma ideia de como as variáveis estudadas estão relacionadas entre si. Por outro lado, essas relações não são equivalentes em todos os municípios, indicando uma heterogeneidade espacial na relação dessas variáveis.

Figura 18 – Distribuição espacial (esquerda) e diagrama de Moran (direita) da RPC nos municípios cearenses



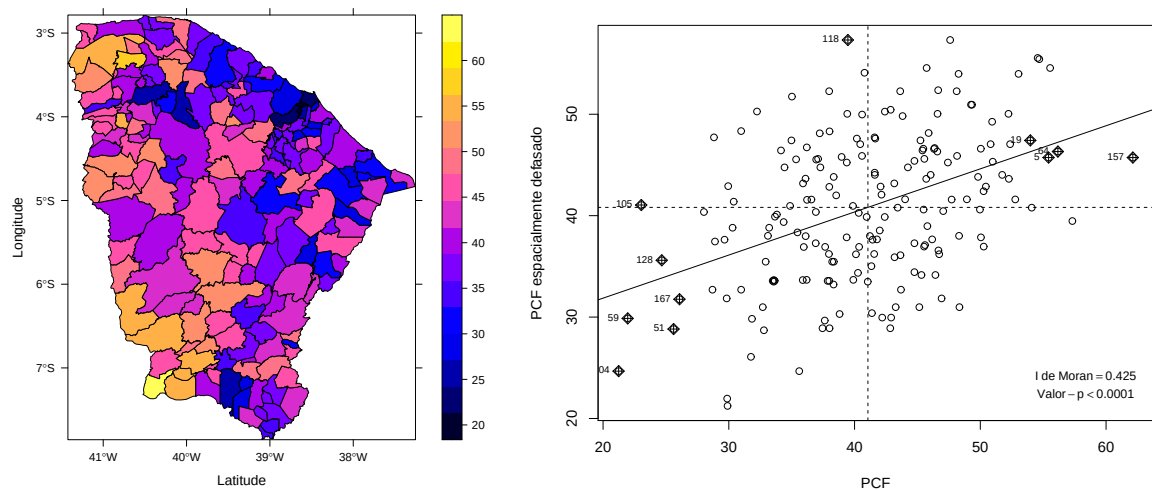
Fonte: O autor (2021).

Figura 19 – Distribuição espacial (esquerda) e diagrama de Moran (direita) da TH nos municípios cearenses



Fonte: O autor (2021).

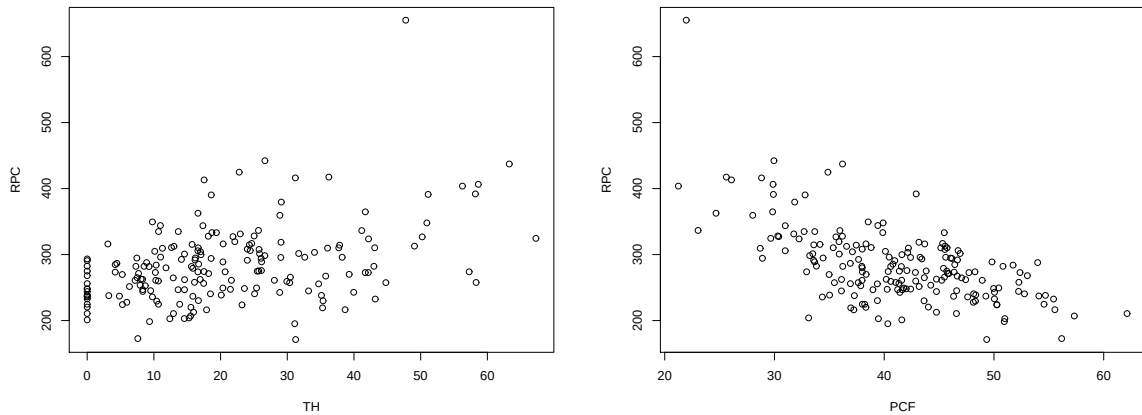
Figura 20 – Distribuição espacial (esquerda) e diagrama de Moran (direita) da PCF nos municípios cearenses



Fonte: O autor (2021).

Adicionalmente, os diagramas de espalhamento de Moran mostram tendências no primeiro e terceiro quadrantes presentes em todas as variáveis estudadas, as quais apresenta correlações espaciais significativas e relativamente altas. Por fim, a Figura 21 são exibidos os gráficos de dispersão entre a RPC e as demais variáveis explicativas, destacando relações lineares entre essas variáveis. Note também que os gráficos de dispersão ressaltam uma relação diretamente proporcional entre a RCP e a TH, indicando que a TH de alguma forma aumenta a renda desta parcela da população pobre. Por outro lado, o PCF tem um efeito contrário na RPC, apresentando uma relação inversamente proporcional.

Figura 21 – RPC versus TH (esquerda) e RPC versus PCF (direita)



Fonte: O autor (2021).

2.5.2 Ajuste do Modelo

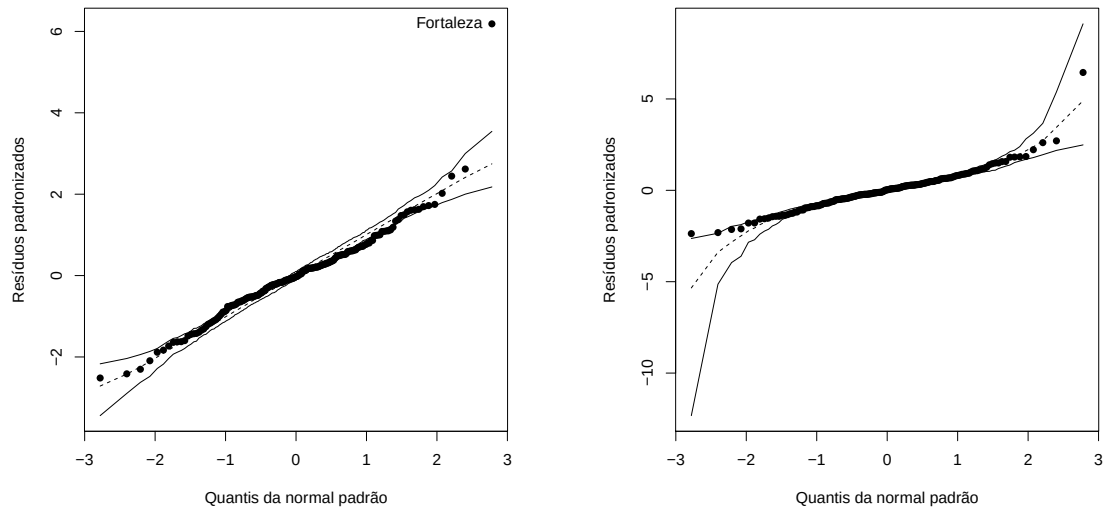
Visto que identificamos relações lineares entre as variáveis na análise exploratória, então propomos inicialmente um modelo de regressão elíptico global através da seguinte equação linear:

$$RPC_i = \beta_0 + \beta_1 TH_i + \beta_2 PCF_i + \epsilon_i,$$

em que $\epsilon_i \sim S(0, \phi)$ representa os erros aleatórios. O ajuste desse modelo ao considerar erros normalmente distribuídos constatou a presença de um valor extremamente discrepante nos resíduos (veja Figura 22), o qual apresentou uma forte influência na estimação dos parâmetros. Além disso, nota-se vários pontos fora das bandas com 95% de confianças no gráfico de quantis-quantis dos resíduos padronizados para o modelo com erros normais. Por esse fato, propomos a utilização de uma distribuição com caudas mais pesadas para os erros aleatórios. De acordo com o valor do AICc, foi selecionado o modelo ajustado por meio da distribuição t -Student com 4 graus de liberdade. Nessa seleção foram considerados apenas uma grade de valores inteiros (de 3 até 30) para os graus de liberdade. Considerou-se também o ajuste dos modelos elípticos com distribuições Cauchy, exponencial potência e logísticas do tipo I e II, no entanto o modelo com erros t -Student apresentou os melhores valores de AICc. Com isso, constatamos que o uso dessa distribuição para os erros aleatório foi capaz de reduzir significativamente a influência do valor discrepante destacado anteriormente no modelo com erros normais. Contudo, ambos os modelos apresentaram uma forte autocorrelação espacial nos resíduos, com podemos observar nos diagramas de

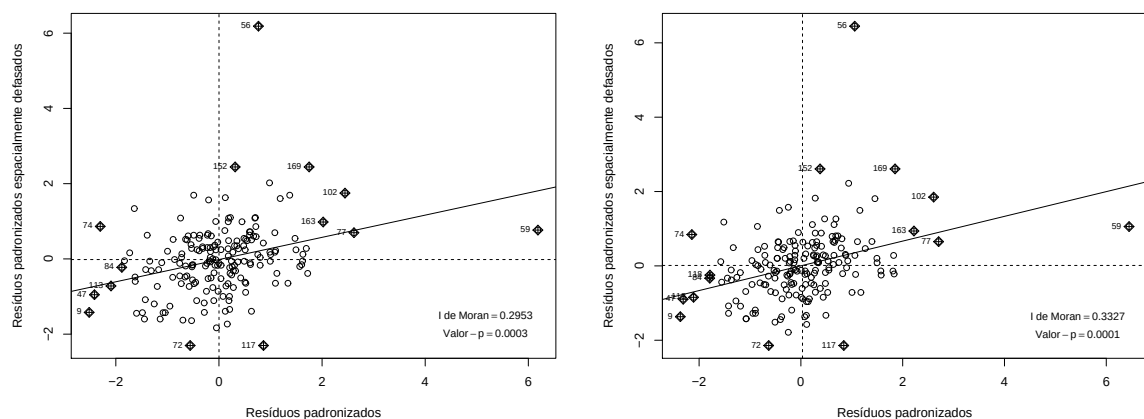
espalhamento de Moran (veja Figura 23). Em ambos os casos obtemos altos valores para o índice de Moran, destacando-se autocorrelações espaciais significativas.

Figura 22 – Gráfico de quantis-quantis com envelope simulados dos resíduos padronizados para o modelo de regressão elíptico com erros normais (esquerda) e *t*-Student (direita)



Fonte: O autor (2021).

Figura 23 – Diagramas de espalhamento de Moran dos resíduos padronizados para o modelo de regressão elíptico com erros normais (esquerda) e *t*-Student (direita)



Fonte: O autor (2021).

Por este fato, o uso da REGP pode ser uma melhor alternativa para medir a relação entre a RPC e as variáveis explicativas, pois permite a variação espacial dos

parâmetros de regressão. Dessa forma, propomos o ajuste de modelos REGP através da seguinte equação:

$$\text{RPC}_i = \beta_0 + \beta_1(h_i, v_i)\text{TH}_i + \beta_2(h_i, v_i)\text{PCF}_i + \epsilon_i,$$

em que $\epsilon_i \sim S(0, \phi(h_i, v_i))$ representa os erros aleatórios. Neste caso, foi comparado os ajustes dos modelos ajustados supondo erros normais e t -Student. Utilizamos um *kernel* adaptativo biquadrado como função de ponderação espacial, o critério CV para seleção do parâmetro de suavização ótimo e o AICc selecionou 3 g.l. para o modelo com erros t -Student. Considerou-se novamente uma grade de valores inteiros (de 3 até 30) para a seleção dos graus de liberdade e mesmas distribuições mencionadas anteriormente para os erros aleatórios dos modelos REGP. Nesse caso o modelo REGP com erros t -Student apresentou os melhores valores para o AICc.

Com base nesses ajustes, as Tabelas 2 e 3 apresentam uma análise descritiva dos parâmetros estimados e dos erros padrões assintóticos para os modelos de REGP com erros normais e t -Student, respectivamente. Observe que, em ambos modelos, os parâmetros apresentam uma alta variação em torno das medidas obtidas pelo ajuste dos modelos globais. Vale ressaltar que, o modelo de REGP com distribuição t -Student apresenta melhores estatísticas para os erros padrões. Além disso, de acordo com os critérios de informação de Akaike (AIC), Akaike corrigido (AICc) e Bayesiano (BIC), a REGP com distribuição t -Student é selecionada com o melhor ajuste.

Tabela 2 – Análise descritiva das estimativas dos parâmetros e erros padrões assintóticos (entre parênteses) para os modelos de REGP com erros normais

	Minimo	Mediana	Média	Máximo	Global
$\hat{\beta}_0$	279,95 (38,56)	438,57 (58,81)	435,85 (61,13)	540,56 (118,81)	435,19 (19,70)
$\hat{\beta}_1$	-1,08 (0,42)	0,94 (0,67)	0,93 (0,74)	1,94 (1,30)	1,12 (0,22)
$\hat{\beta}_2$	-7,25 (0,82)	-4,15 (1,36)	-4,24 (1,40)	-0,69 (2,65)	-4,25 (0,44)
$\hat{\phi}$	548,69 (203,38)	1244,96 (383,73)	1479,37 (481,39)	3307,26 (1176,37)	1854,17 (193,31)
AIC = 1860,20		AICc = 1869,63		BIC = 1945,87	

Fonte: O autor (2021).

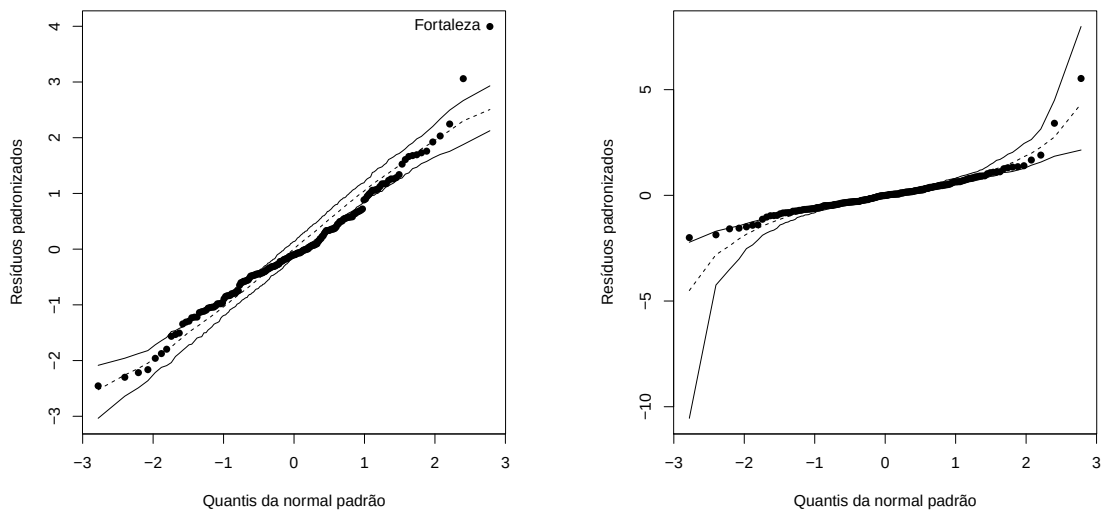
Tabela 3 – Análise descritiva das estimativas dos parâmetros e erros padrões assintóticos (entre parênteses) para os modelos de REGP com erros t -Student

	Minimo	Mediana	Média	Máximo	Global
$\hat{\beta}_0$	246,59 (37,33)	417,14 (49,92)	409,16 (53,39)	586,64 (119,31)	413,56 (16,66)
$\hat{\beta}_1$	-1,13 (0,32)	0,94 (0,60)	0,95 (0,66)	2,24 (1,36)	0,98 (0,19)
$\hat{\beta}_2$	-7,24 (0,76)	-3,71 (1,13)	-3,59 (1,22)	-0,11 (2,89)	-3,68 (0,37)
$\hat{\phi}$	184,68 (110,04)	666,03 (292,47)	663,47 (323,44)	2445,34 (1417,96)	946,484 (130,538)
AIC = 1826,02		AICc = 1839,38		BIC = 1926,76	

Fonte: O autor (2021).

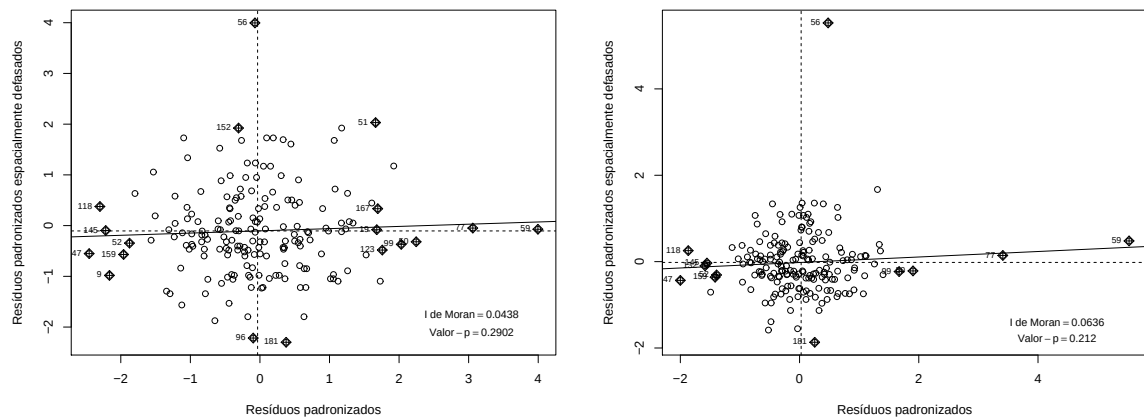
Adicionalmente, nas Figuras 24 e 25 é apresentado a análise de resíduos para os modelos de REGP ajustados. Nota-se novamente a presença de um valor altamente discrepante nos resíduos do modelo com erros normais, porém o uso da distribuição t -Student para os erros aleatórios demonstra acomodar bem a observação destacada. Além disso, os diagramas de espalhamento de Moran revelam correlações espaciais não significativas, indicando que ambos os modelos de REGP contornaram o problema da dependência espacial ressaltada nos resíduos dos modelos elípticos globais. Isso se dá pelo fato de que ao assumir a não estacionariedade espacial dos parâmetros de regressão obtemos modelos ajustados sem a presença autocorrelação espacial nos resíduos.

Figura 24 – Gráfico de quantis-quantis com envelope simulados dos resíduos padronizados para o modelo de REGP com erros normais (esquerda) e t -Student (direita)



Fonte: O autor (2021).

Figura 25 – Diagramas de espalhamento de Moran dos resíduos padronizados para o modelo de REGP com erros normais (esquerda) e *t*-Student (direita)

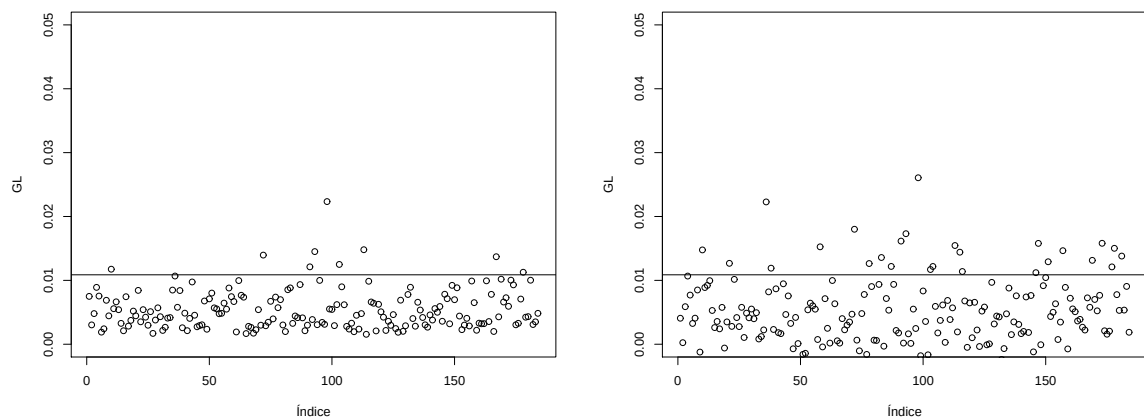


Fonte: O autor (2021).

2.5.3 Análise de Sensibilidade

Para avaliar a qualidade do ajuste e sensibilidade dos modelos de REGP, foram calculadas algumas medidas de diagnóstico no modelos ajustados. Inicialmente, na Figura 26 são exibidos os gráficos com a medida da alavancagem generalizada para os modelos de REGP ajustados. Em ambos ajustes não é destacado nenhuma observação como um possível ponto de alavanca, uma vez que todas observações apresentam valores bem próximos pra essa medida.

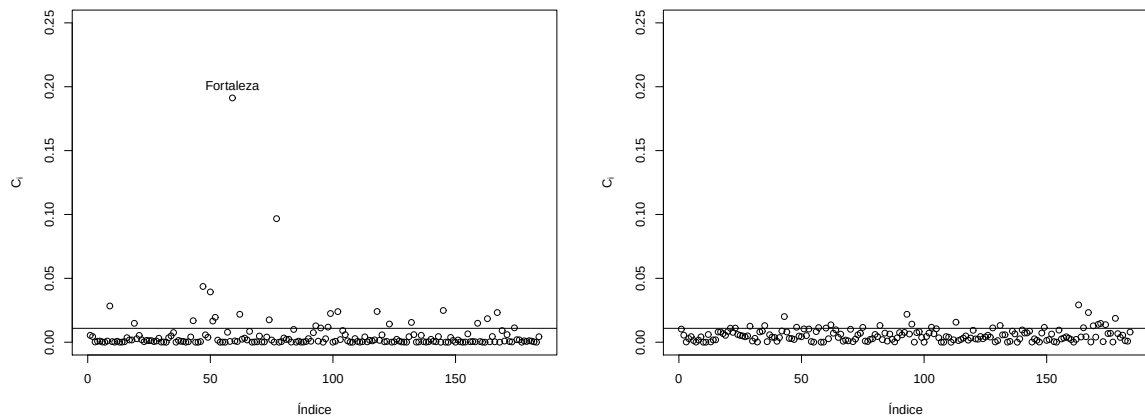
Figura 26 – Gráfico de alavancagem generalizada nos modelos de REGP com erros normais (esquerda) e *t*-Student (direita)



Fonte: O autor (2021).

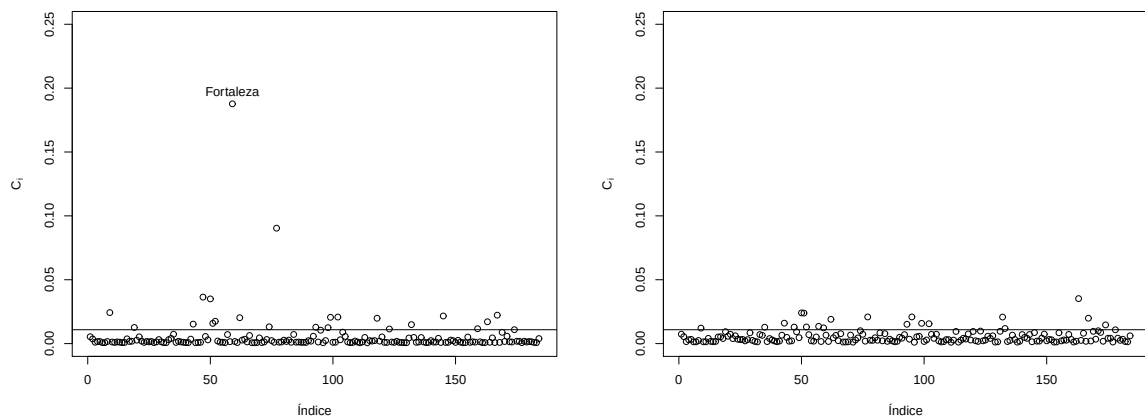
Por outro lado, nas Figuras 27 e 28 são apresentados os gráficos de influência local sob os esquemas de perturbações na escala e ponderação de casos, respectivamente. Note que, o município de Fortaleza é destacado como possível observação influente no modelo de REGP com erros normais em ambos esquemas de perturbação. Por outro lado, no modelo REGP com erros t -Student nenhuma observação é destacada como possivelmente influente. Esse fato indica que o modelo REGP com erros t -Student produz estimativas menos sensível à presença de valores discrepantes.

Figura 27 – Gráfico de influência local para perturbação na escala nos modelos REGP com erros normais (esquerda) e t -Student (direita)



Fonte: O autor (2021).

Figura 28 – Gráfico de influência local para perturbação ponderação de casos nos modelos REGP com erros normais (esquerda) e t -Student (direita)

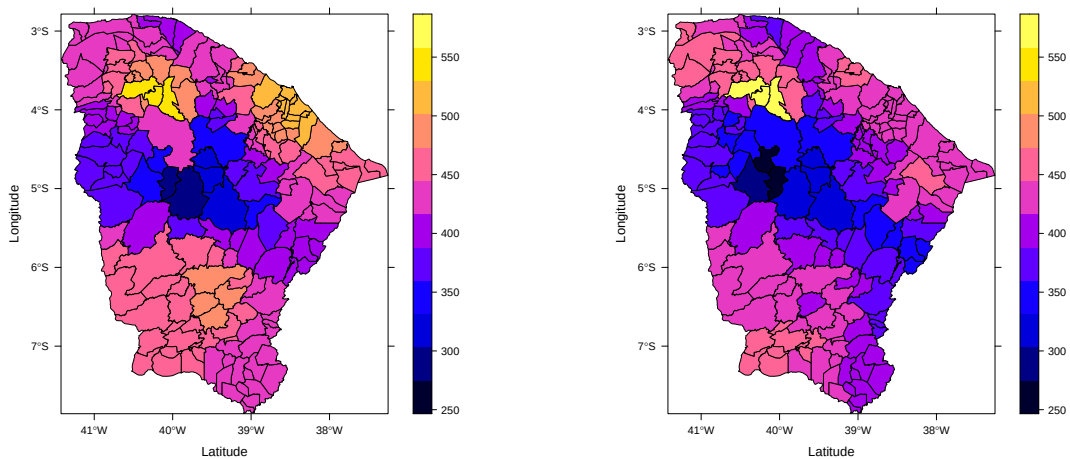


Fonte: O autor (2021).

2.5.4 Análise Confirmatória

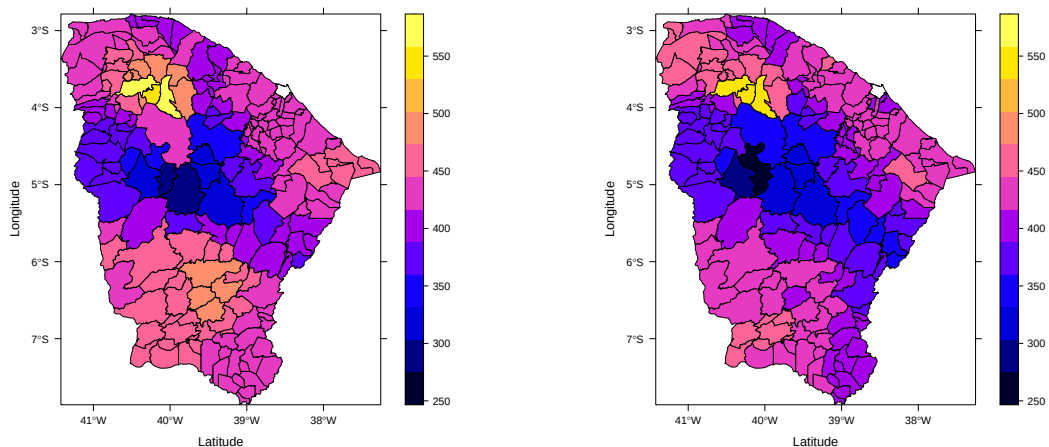
Como destacado anteriormente, os métodos de diagnósticos para a REGP apontaram uma forte influência da observação referente ao município de Fortaleza no modelo com erros normais. Nesse caso, para verificar se Fortaleza é de fato uma observação influente, considerou o reajuste dos modelos de REGP excluindo essa observação da base de dados. Os resultados ao reajustar os modelos foram comparados diretamente com os obtidos anteriormente para os dados completos. Dessa forma, as Figuras 29-34 mostram a comparação entre as distribuições espaciais das estimativas dos parâmetros de regressão para os modelos de REGP ao excluir o município de Fortaleza.

Figura 29 – Distribuição espacial das estimativas de β_0 considerando os dados completos para os modelos REGP com erros normais (esquerda) e t -Student (direita)



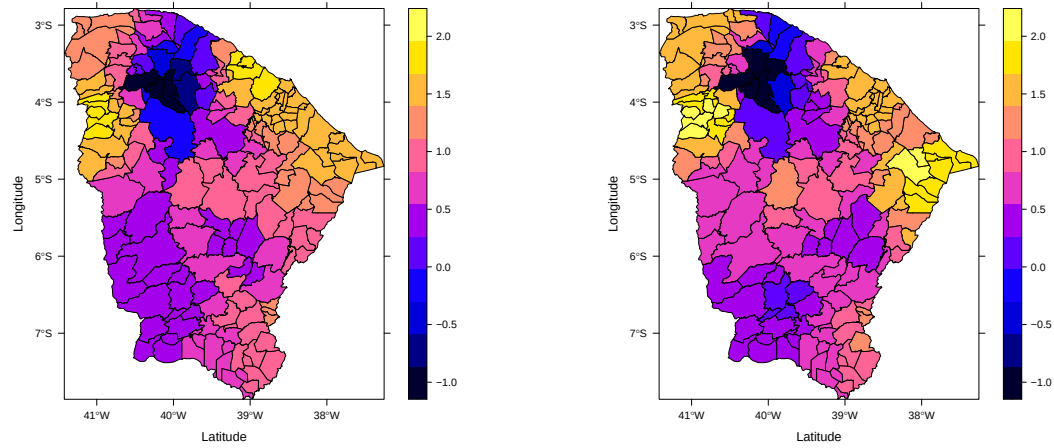
Fonte: O autor (2021).

Figura 30 – Distribuição espacial das estimativas de β_0 considerando os dados após a exclusão de Fortaleza para os modelos REGP com erros normais (esquerda) e t -Student (direita)



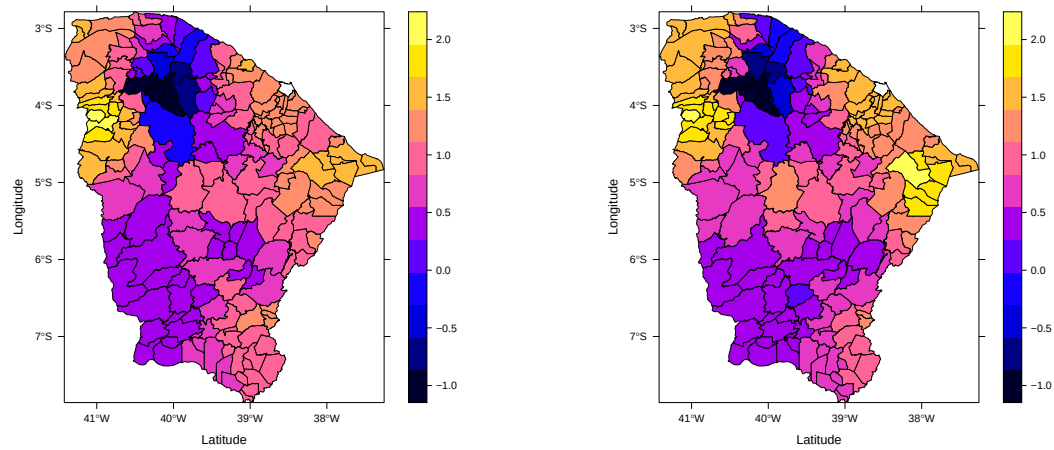
Fonte: O autor (2021).

Figura 31 – Distribuição espacial das estimativas de β_1 considerando os dados completos para os modelos REGP com erros normais (esquerda) e t -Student (direita)



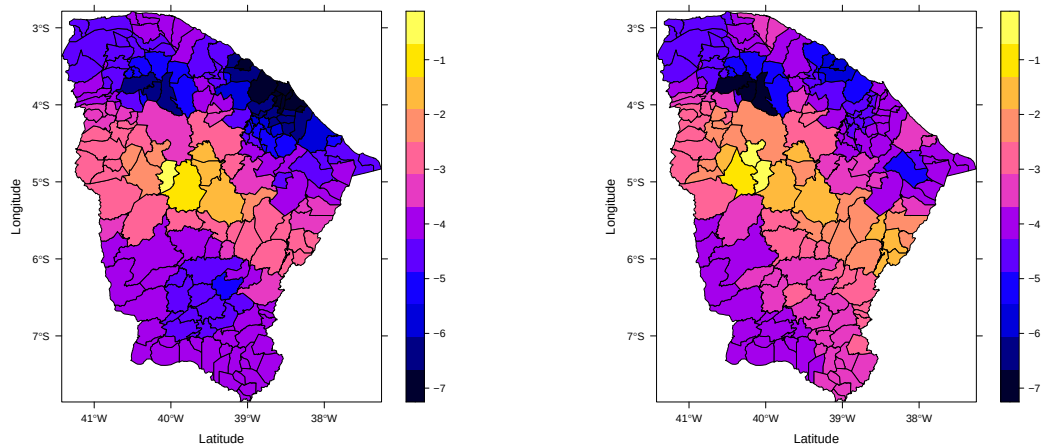
Fonte: O autor (2021).

Figura 32 – Distribuição espacial das estimativas de β_1 considerando os dados após a exclusão de Fortaleza para os modelos REGP com erros normais (esquerda) e t -Student (direita)



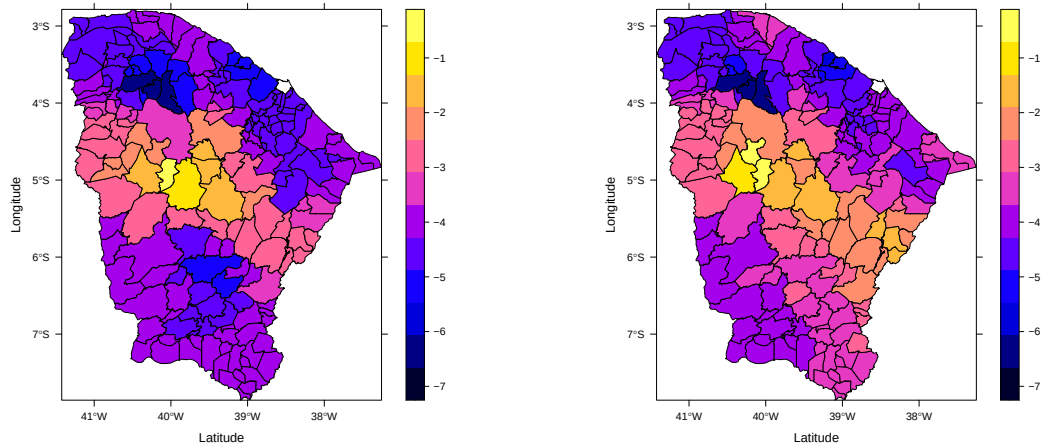
Fonte: O autor (2021).

Figura 33 – Distribuição espacial das estimativas de β_2 considerando os dados completos para os modelos REGP com erros normais (esquerda) e t -Student (direita)



Fonte: O autor (2021).

Figura 34 – Distribuição espacial das estimativas de β_2 considerando os dados após a exclusão de Fortaleza para os modelos REGP com erros normais (esquerda) e t -Student (direita)



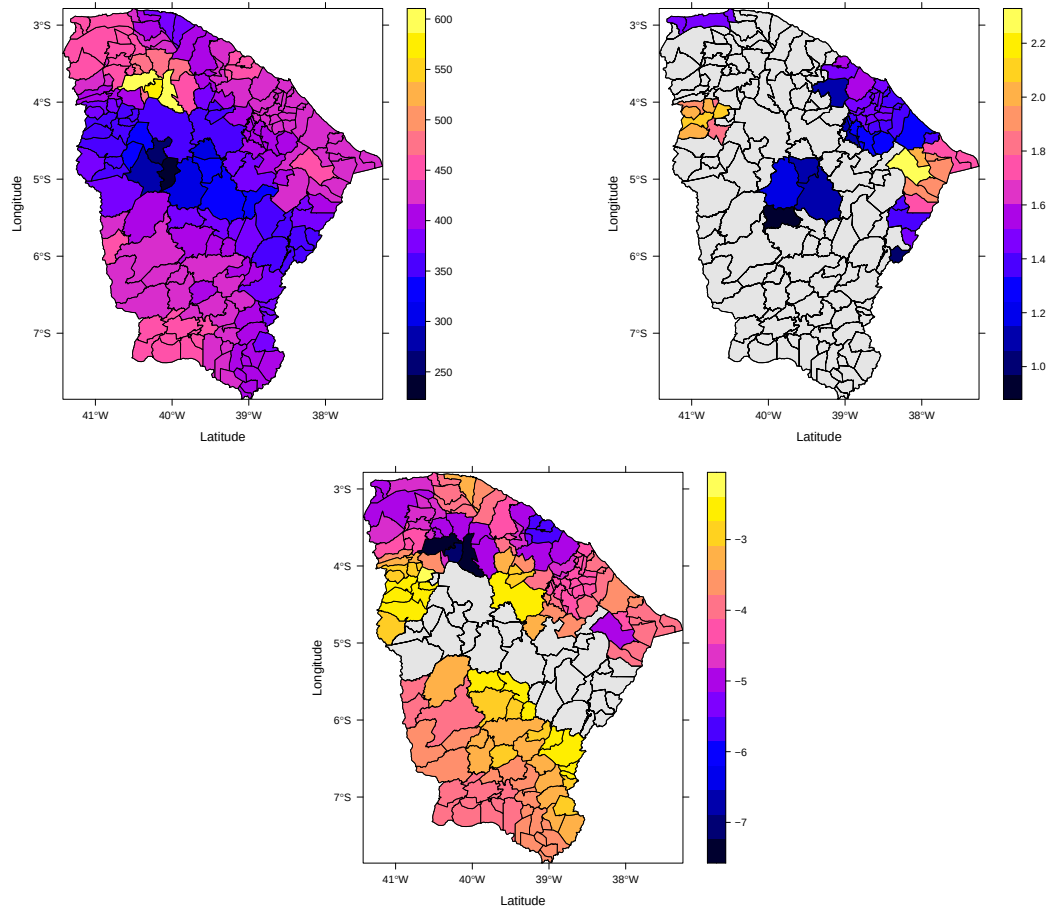
Fonte: O autor (2021).

Observe que, as distribuições espaciais das estimativas obtidas no modelo de REGP com erros normais ao desconsiderar Fortaleza mostram mudanças extremas nos resultados em torno desse município. Em outras palavras, a retirada dessa observação muda a interpretação da variação local dos parâmetros. Isso ocorre devido ao valor da RPC deste município ser muito maior que os demais, representando uma observação extremamente discrepante na variável resposta. Por outro lado, a distribuição espacial das estimativas no modelo de REGP com erros t -Student não apresenta alterações substanciais ao retirar o município de Fortaleza, conservando as mesmas interpretações dos resultados. Isso evidencia que o modelo REGP com erros t -Student mostrou-se menos sensível a presença de valores discrepantes.

2.5.5 Interpretação do Modelo Ajustado

Devido aos resultados obtidos anteriormente, o modelo REGP com erros t -Student obteve os melhores valores para os critérios de informação e menor sensibilidade às observações influentes. Dessa forma, optou-se pela utilização da REGP com erros t -Student para descrever o comportamento da RPC condicionado a TH e ao PCF. Dessa forma, as distribuições espaciais das estimativas significativas (assumindo o nível descritivo corrigido com uma taxa de erro fixada em 5%) dos parâmetros de regressão desse modelo são apresentadas na Figura 35, destacando as estimativas não significativas na escala de cores pela cor cinza.

Figura 35 – Distribuição espacial das estimativas significativas de β_0 (esquerda), β_1 (direita) e β_2 (abaixo) no modelo REGP com erros t -Student



Fonte: O autor (2021).

Observa-se que, à medida que TH aumenta há também um aumento nos valores da RPC. Essa relação é vista em todo o estado do Ceará, no entanto o seu efeito ($\hat{\beta}_1$) é significativo apenas em alguns municípios nas regiões central, noroeste e nordeste. Além disso, os gráficos indicam que o decréscimo do PCF implica no aumento dos valores da RPC, porém essa relação é significativa apenas nos municípios da região norte e sul. Nota-se também que o efeito de ambas variáveis explicativas tem magnitudes diferentes para cada município devido à variação espacial dos coeficientes de regressão.

3 MODELOS DE REGRESSÃO ELÍPTICO GEOGRAFICAMENTE PONDERADO MULTIESCALAR

Como discutido no capítulo anterior, a REGP se mostrou uma ótima ferramenta para a modelagem de dados espaciais sob a suposição de não estacionariedade espacial nas relações condicionais entre a variável resposta e as variáveis explicativas. Em geral, as classes de modelos geograficamente ponderados dependem diretamente de uma função de ponderação geográfica baseada na distância entre as regiões estudadas dentro de uma janela previamente estabelecida. Dessa forma, o processo para a seleção do parâmetro de suavização ótimo representa um fator de grande importância na estimação do modelo, pois o valor do parâmetro de suavização determinará o tamanho efetivo da amostra para a estimação dos parâmetros locais (Fotheringham, Brunsdon e Charlton, 2002). Em síntese, o tamanho da janela determina a taxa de decaimento dos pesos geográficos em torno de um determinado local. Visto isso, valores pequenos para o parâmetro de suavização faz com que os pesos diminuam rapidamente conforme a distância aumenta, possibilitando assim que os coeficientes de regressão mudem rapidamente no espaço. Isso possibilita destacar anomalias locais e produz um modelo com viés pequeno. No sentido contrário, valores maiores para o parâmetro de suavização geram pesos que diminuem lentamente conforme a distância aumenta, fazendo com que os coeficientes de regressão mudem suavemente e tenham uma variância menor. À vista disso, é esperado que seleção do parâmetro de suavização tenha um impacto direto no viés e na variância dos parâmetros locais.

Como mencionado no primeiro capítulo, as escalas espaciais podem variar de acordo com o processo estudado, ou seja, diferentes processos não necessariamente terão as mesmas escalas de variação. Nesses casos, poderíamos distinguir entre micro e macro processos, processo locais e globais, entre outros. A título de exemplo, se tivermos o interesse em estudar os preços dos imóveis notaríamos facilmente que estes variam de forma diferente se os estudarmos por município, estado ou país. Neste contexto, se entendemos como processo a relação entre duas variáveis, o conceito de janela seria essencialmente uma questão de escala e o tamanho da janela indicaria em qual escala espacial variam os processos. Isto implica que as classes de modelos geograficamente ponderados usualmente assumem que todos os processos em estudo operam na mesma escala espacial, determinada por meio de um único parâmetro de suavização no ajuste do modelo. No entanto, essa suposição pode ser irrealista e gerar interpretações completamente equivocadas.

Em contrapartida, Fotheringham, Yang e Kang (2017) apresenta a classe dos modelos de RGPM, a qual permite que seus parâmetros de regressão variem em diferentes escalas espaciais. A ideia básica é supor que cada variável explicativa esteja associada a um parâmetro de suavização diferente, permitindo que cada relação de regressão tenha sua própria variação espacial. No entanto, essa classe de modelos ainda está suscetível à influência de possíveis observações discrepantes. Com isto em vista, propomos a introdução da REGPM como uma alternativa a RGPM na aplicação de dados com presença de observações discrepantes. Naturalmente, os modelos de REGPM são essencialmente uma REGP que funcionam em várias escalas espaciais, no entanto alguns conceitos adicionais devem ser levados em consideração. Nesse capítulo iremos abordar a construção e estimação dos modelos de REGPM, bem como os conceitos e teorias necessárias para tal propósito. Além disso, são discutidos métodos de diagnósticos para a REGPM com base na abordagem de influência local, estudos de simulações e uma aplicação prática a dados reais.

3.1 ESPECIFICAÇÃO DO MODELO

Como foi apresentado anteriormente, os modelos de REGP são especificados por meio da equação (2.2) e representam uma extensão dos modelos elípticos assumindo a variação local dos seus parâmetros. Em contrapartida, os modelos de REGPM permitem que essas variações locais tenham diferentes escalas espaciais para cada parâmetro de regressão, ou seja,

$$y_i = \beta_{b_0}(h_i, v_i) + \sum_{k=1}^p x_{ik} \beta_{b_k}(h_i, v_i) + \epsilon_i \quad (3.1)$$

em que (h_i, v_i) indica as coordenadas espaciais do i -ésimo local, $\beta_{b_k}(h_i, v_i)$ é o parâmetro associado à k -ésima variável explicativa para o i -ésimo local, b_k indica o parâmetro de suavização associado à k -ésima variável explicativa, x_{ik} o valor da k -ésima variável explicativa no i -ésimo local e $\epsilon_i \sim S(0, \phi(h_i, v_i))$ são erros aleatórios mutuamente independentes. De acordo com a equação (3.1), cada coeficiente de regressão está relacionado a um parâmetro de suavização diferente, a qual deverá ser selecionada com base em algum dos métodos definidos no capítulo anterior. Os parâmetros de suavização selecionados devem então ser capazes de refletir a escala espacial associada ao seu respectivo coeficiente de regressão. Com base nessa definição, se assumimos que todos os parâmetros de suavização são iguais, então teríamos um modelo de REGP como caso particular dos modelos REGPM.

Note que, pela complexidade da especificação do modelo de REGPM, há a necessidade de adaptar o processo de estimação proposto no capítulo anterior. Isso se dá pelo fato de que, na REGPM cada coeficiente de regressão tem seu próprio parâmetro de suavização, implicando em uma matriz de ponderação diferente associado a cada um desses coeficientes. Uma forma de contornar esse problema é reescrever a equação do modelo em (3.1) sob uma forma mais conveniente. Em outras palavras, na forma como os modelos de REGP e REGPM são definidos é possível representá-los sob a mesma estrutura de um modelo aditivo generalizado (MAG). Essa classe de modelos assume que a variável resposta é expressa por uma soma de variáveis preditoras suavizadas. Nesse caso, cada superfície espacial dos parâmetros pode ser associada a um termo aditivo não paramétrico, de tal forma que:

$$\mathbf{Y} = \sum_{k=0}^p \mathbf{f}_k + \boldsymbol{\epsilon} \quad (3.2)$$

em que $\mathbf{f}_k = (\beta_{b_k}(h_1, v_1)x_{1k}, \dots, \beta_{b_k}(h_n, v_n)x_{nk})^\top$ é o k -ésimo termo aditivo do modelo e $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^\top$ representa os erros aleatórios mutuamente independentes. Em particular, a variável preditora $\mathbf{x}_0 = (x_{i0}, \dots, x_{in})^\top$, com $x_{i0} = 1$ para todo $i = 1, \dots, n$, irá representar o intercepto do modelo. Dessa forma, cada termo aditivo $\mathbf{f}_k$ pode ser interpretado como uma função da k -ésima variável preditora suavizada por meio do parâmetro b_k , uma vez que o parâmetro de suavização está associada com a suavidade da função de predição. Naturalmente, os valores ótimos selecionados para os parâmetros de suavização podem indicar que um modelo mais simples forneça um ajuste adequado. No caso que todos os valores dos parâmetros de suavização sejam próximos e relativamente grandes, então a aplicação de um modelo elíptico global poderia fornecer um bom ajuste. Por outro lado, se os valores dos parâmetros de suavização forem próximas e relativamente pequenos, então um modelo REGP seria suficiente para representar o comportamento dos dados.

3.2 PROCESSO DE ESTIMAÇÃO

Para a obtenção do processo de estimação dos modelos de REGPM será necessário o uso de um algoritmo iterativo do tipo *backfitting*, proposto inicialmente por Breiman e Friedman (1985) e adaptado para a classe dos modelos geograficamente ponderados com múltiplas escalas em Fotheringham, Yang e Kang (2017). De forma resumida, propomos obter as estimativas do modelo de REGPM maximizando a verossimilhança local para

cada termo aditivo por meio do ajuste de múltiplas REGP. Esse procedimento tem como base a definição da REGPM sob a forma de um MAG, a qual foi apresentada na seção anterior. Dito isso, a partir dos resultados obtidos para a estimação dos parâmetros da REGP, a seguir é descrito de forma detalha o algoritmo de estimação para os modelos de REGPM.

3.2.1 Algoritmo *backfitting*

Esse algoritmo consiste em estimar, de forma recursiva e iterativa, as funções de suavização $\mathbf{f}_k$ com base nos demais termos aditivos estimados. Cada função $\mathbf{f}_k$ é estimada através de uma REGP, obtendo um parâmetro de suavização ótimo e uma superfície de valores estimados para o k -ésimo parâmetro de regressão. Dessa forma, o processo de estimação para os parâmetros do modelo de REGPM é descrito pelos seguintes passos.

1. Inicialize os valores dos termos aditivos ajustados $(\hat{\mathbf{f}}_{0,\dots,p})$ com base em algum critério de decisão.
2. Através dos valores iniciais definidos no passo anterior, um conjunto de estimativas para $\mathbf{y}$ é obtido e consequentemente os resíduos ordinais são expressos por:

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \sum_{k=0}^p \hat{\mathbf{f}}_k$$

3. Obtenha um valor ótimo para o parâmetro de suavização associado ao intercepto e ajuste um modelo de REGP regressando $(\hat{\mathbf{f}}_0 + \hat{\boldsymbol{\epsilon}})$ em $\mathbf{X}_0$. Dessa forma, sob convergência do algoritmo de estimação do modelo de REGP, a função de suavização atualizada é dado pela seguinte forma:

$$\hat{\mathbf{f}}_0^* = \text{diag}(\mathbf{X}_0)(\mathbf{X}_0^\top \mathbf{D}(\hat{v}^{(m)}) \mathbf{W}_{i0} \mathbf{X}_0)^{-1} \mathbf{X}_0^\top \mathbf{D}(\hat{v}^{(m)}) \mathbf{W}_{i0} (\hat{\mathbf{f}}_0 + \hat{\boldsymbol{\epsilon}})$$

em que $\mathbf{W}_{i0}$ representa a matriz de ponderações espaciais para o intercepto no i -ésimo local. Com isso, os valores dos resíduos ordinais são novamente atualizados a partir das novas estimativas obtidas para $\mathbf{f}_0^*$ e baseando-se na expressão do passo anterior.

4. Repita o passo anterior para cada variável explicativa, regressando $(\hat{\mathbf{f}}_k + \hat{\boldsymbol{\epsilon}})$ em $\mathbf{X}_k$ por meio de um modelo REGP. Com isso, a função de suavização associada a k -ésima variável atualizada é dada pela seguinte expressão:

$$\hat{\mathbf{f}}_k^* = \text{diag}(\mathbf{X}_k)(\mathbf{X}_k^\top \mathbf{D}(\hat{v}^{(m)}) \mathbf{W}_{ik} \mathbf{X}_k)^{-1} \mathbf{X}_k^\top \mathbf{D}(\hat{v}^{(m)}) \mathbf{W}_{ik} (\hat{\mathbf{f}}_k + \hat{\boldsymbol{\epsilon}})$$

em que $\mathbf{W}_{ik}$ representa a matriz de ponderações espaciais para o k -ésimo parâmetro de regressão no i -ésimo local. Dessa forma, atualize novamente os valores dos resíduos a partir das novas estimativas obtidas para $\mathbf{f}_k$ sob a expressão dada no segundo passo.

5. Esse processo deverá ser repetido até atingir algum critério de convergência.

Nesse processo de estimação assume-se que as estimativas dos parâmetros de dispersão são obtidas do ajuste prévio de uma REGP e são mantidas fixas durante a execução do algoritmo. Com isso, ao final do processo iterativo obtemos uma superfície local estimada para cada coeficiente de regressão, os quais terão escalas espaciais individuais.

3.2.2 Escolha dos Valores Iniciais

Como ressalta Fotheringham, Yang e Kang (2017) os valores iniciais das estimativas locais é um fator importante para o algoritmo de estimação em modelos geograficamente ponderados com múltiplas escalas. O autor supracitado argumenta que a escolha dos valores iniciais não deve influenciar a seleção dos parâmetros de suavização ótimas e superfície espacial das estimativas, mas pode afetar diretamente o número de iterações necessárias para obter convergência. Embora a escolha dos valores utilizados para inicializar o algoritmo seja uma decisão particular do usuário algumas sugestões podem ser levadas em consideração. Neste contexto, as seguintes opções podem ser utilizadas para garantir que não haja um custo computacional alto no processo de estimação.

1. Todas as estimativas devem ser inicializadas com o valor zero.
2. Os valores iniciais são retirados do ajuste de um modelo elíptico global.
3. Os valores iniciais são retirados do ajuste da REGP.

Naturalmente, essas opções sugeridas estão sujeitas a tempos de processamentos diferentes dependendo da base de dados em estudo. Em casos que a relação de regressão seja aproximadamente estacionária no espaço, um valor inicial para este parâmetro com base no ajuste de um modelo global deve estar próximo da estimativa final. Por outro lado, as estimativas dos parâmetros que são claramente não estacionários no espaço devem ter convergência mais rápida usando os valores iniciais a partir do ajuste da REGP. Em outras palavras, não há como prever qual opção terá o melhor desempenho computacional. Com isso em vista, optamos por utilizar os valores iniciais através das estimativas obtidas com base no ajuste um modelo de REGP, embora isso não deve ser considerado como uma

regra definitiva em trabalhos futuros.

3.2.3 Critério de Convergência

Outro fator importante que deve ser levado em consideração é o critério de convergência usado para a aplicação do algoritmo de estimação. No contexto dos modelos geograficamente ponderados com múltiplas escalas, Fotheringham, Yang e Kang (2017) sugere o uso de dois métodos baseadas na diferença relativa entre iterações sucessivas. Esse conceito é denominado por pontuação de mudança e é amplamente utilizado como critério de convergência. Nesse contexto, a primeira opção pontua mudanças entre as somas dos quadrados residuais ($RSS = \sum_{i=1}^n \hat{\epsilon}^2$) e é definida pela seguinte expressão:

$$SOC_{RSS} = \frac{|RSS_{new} - RSS_{old}|}{RSS_{new}}$$

em que RSS_{new} e RSS_{old} representam a soma dos quadrados residuais para a iteração atual e anterior, respectivamente. Como alternativa, o segundo critério pontua as mudanças entre os termos aditivos estimados $\hat{\mathbf{f}}_{ik}$, sendo expressa por:

$$SOC_f = \sqrt{\frac{\sum_{k=0}^p \frac{\sum_{i=1}^n (\hat{\mathbf{f}}_{ik}^{new} - \hat{\mathbf{f}}_{ik}^{old})^2}{n}}{\sum_{i=1}^n (\sum_{k=1}^p \hat{\mathbf{f}}_{ik}^{new})}}$$

em que $\hat{\mathbf{f}}_{ik}^{new}$ e $\hat{\mathbf{f}}_{ik}^{old}$ representam o k -ésimo termo aditivo associado ao i -ésimo local para a iteração atual e anterior, respectivamente. Novamente, vale ressaltar que os critérios de paradas apresentados são apenas sugestões e a escolha do critério deve ser uma decisão particular do usuário. Nesse trabalho consideramos a utilização do critério de parada baseado na pontuação pelo somas dos quadrados residuais, porém um critério alternativo pode ser facilmente implementado.

3.2.4 Centralização das Variáveis Explicativas

Outro fator importante para o algoritmo é a centralização de todas as variáveis explicativas, facilitando uma interpretação a priori dos parâmetros de suavização individualmente e reduzindo consideravelmente o tempo de processamento do algoritmo. A padronização consiste basicamente em centralizar as variáveis preditoras em zero subtraindo sua média de cada observação. Este procedimento permite que os parâmetros de suavização selecionados indiquem diretamente a escala espacial das relações condicionais

entre a variável resposta e as variáveis explicativas. Naturalmente, não considerar as padronizações implica na mesma interpretação, porém o custo computacional será bem maior. Esse procedimento é particularmente útil quando a utilização dos modelos de REGPM tiver um objetivo explanatório para a escolha do modelo adequado, uma vez que os parâmetros de suavização selecionados indicam quais parâmetros apresentam variações espaciais.

3.2.5 Matriz de Projeção

Assim como na REGP, a obtenção da matriz de projeção é crucial para derivar alguns resultados inferenciais nos modelos de REGPM. Com essa matriz é possível mensurar o número de parâmetros efetivos e, conseqüentemente, calcular medidas associadas à qualidade de ajuste do modelo. De acordo com os resultados obtidos no capítulo anterior, sabe-se que:

$$\mathbf{S} = \begin{pmatrix} \mathbf{x}_1(\mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_1 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_1 \\ \vdots \\ \mathbf{x}_n(\mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_n \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_n \end{pmatrix}.$$

Com base nos resultados apresentados em Yu *et al.* (2020), essa expressão matricial pode ser definida através de uma forma alternativa. Para isso é necessário reescrever o modelo definido na equação (2.2) sob a forma de um MAG, expresso por meio da equação (3.2) ao definir um modelo REGPM. Com isso, torna-se possível obter as estimativas das funções aditivas em termos dos coeficientes locais estimados. Nesse caso, a matriz $\mathbf{S}$ pode ser decomposta em matrizes de projeções aditivas, que por sua vez são associadas a cada variável explicativa usada na postulação do modelo. Essas matrizes $\mathbf{R}_k$ projetam $\mathbf{y}$ em cada um dos termos aditivos ajustados, tal que:

$$\hat{\mathbf{f}}_k = \mathbf{R}_k \mathbf{y}.$$

Com isso, a matriz de projeção $\mathbf{S}$ pode ser obtida através da soma das matrizes $\mathbf{R}_k$. Em outras palavras, os valores ajustados para a variável resposta pode ser expresso pela seguinte forma:

$$\hat{\mathbf{y}} = \mathbf{S} \mathbf{y} = \sum_{K=0}^p \mathbf{R}_k \mathbf{y}.$$

No entanto, para obter as matrizes $\mathbf{R}_k$ precisamos expressar $\hat{\mathbf{f}}_k$ em função dos coeficientes locais estimados. Logo, utilizando os resultados obtidos anteriormente para o ajuste do modelo de REGPM, temos que:

$$\hat{\mathbf{f}}_k = \begin{pmatrix} x_{1k}\hat{\beta}(1)_k \\ \vdots \\ x_{nk}\hat{\beta}(n)_k \end{pmatrix}.$$

Assumindo que $\mathbf{e}_k$ é um vetor $1 \times p$ de zeros com um na k -ésima posição, então sabe-se que $\hat{\beta}(i)_k = \mathbf{e}_k \hat{\beta}(i)$. Assim, substituindo a expressão do vetor de coeficientes estimado para o i -ésimo local, sob convergência do processo iterativo dado em (2.4), obtemos:

$$\hat{\beta}(i)_k = \mathbf{e}_k (\mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_i \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_i \mathbf{y}.$$

Consequentemente, o k -ésimo termo aditivo estimado pode ser expresso diretamente pela seguinte estrutura matricial:

$$\hat{\mathbf{f}}_k = \begin{pmatrix} x_{1k} \mathbf{e}_k (\mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_1 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_1 \\ \vdots \\ x_{nk} \mathbf{e}_k (\mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_n \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_n \end{pmatrix} \mathbf{y}.$$

Como isso, a matriz de projeção associada ao k -ésimo termo aditivo do modelo REGP é descrita pela seguinte forma matricial:

$$\mathbf{R}_k = \begin{pmatrix} x_{1k} \mathbf{e}_k (\mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_1 \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_1 \\ \vdots \\ x_{nk} \mathbf{e}_k (\mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_n \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D}(\hat{v}) \mathbf{W}_n \end{pmatrix} \quad (3.3)$$

Esse resultado possibilita o cálculo da matriz de projeção para o modelo de REGPM, realizado em conjunto com o processo de estimação definido anteriormente. Para isso, assuma a inicialização do algoritmo *backfitting* usando as estimativas dos parâmetros locais na REGP assumindo uma única escala espacial para todas variáveis explicativas. Neste ponto do algoritmo calcule as matrizes de projeção aditiva para cada variável explicativa, por meio da expressão dada em (3.3) e do ajuste da REGP. Feito isso, durante a estimação do k -ésimo termo aditivo calcule a matriz de projeção parcial $\mathbf{A}_k$ associada a k -ésima variável explicativa, de tal forma que:

$$\hat{\mathbf{f}}_k^* = \mathbf{A}_k (\hat{\mathbf{f}}_k + \hat{\epsilon})$$

em que $\hat{\mathbf{f}}_k^*$ é a estimativa de $\mathbf{f}_k$ atualizada e $\hat{\mathbf{f}}_k$ representa o termo aditivo ajustado na iteração anterior. No entanto, através do processo de estimação e da definição de $\mathbf{R}_k$, obtemos o seguinte resultado:

$$\hat{\mathbf{f}}_k + \hat{\epsilon} = \hat{\mathbf{f}}_k + \mathbf{y} - \sum_{k=1}^p \hat{\mathbf{f}}_k = \mathbf{R}_k \mathbf{y} + \mathbf{y} - \mathbf{S} \mathbf{y}$$

Com isso, as estimativas para a k -ésima função de suavização $\hat{\mathbf{f}}_k^*$ pode ser reescrita na seguinte forma:

$$\hat{\mathbf{f}}_k^* = \mathbf{A}_k (\mathbf{R}_k \mathbf{y} + \mathbf{y} - \mathbf{S} \mathbf{y}) = \mathbf{R}_k^* \mathbf{y}$$

em que $\mathbf{R}_k^* = \mathbf{A}_k \mathbf{R}_k + \mathbf{A}_k - \mathbf{A}_k \mathbf{S}$ representa a matriz de projeção do k -ésimo termo aditivo atualizada. Com isso, a matriz de projeção aditiva do modelo REGPM $\mathbf{S}^*$ pode ser expressa pela seguinte forma iterativa:

$$\mathbf{S}^{(m+1)} = \mathbf{S}^{(m)} - \mathbf{R}_k^{(m)} + \mathbf{R}_k^{(m+1)}$$

Logo, o valor dessa matriz deve então ser atualizado sempre que a estimativa de algum termo aditivo é calculada durante o algoritmo de estimação. Ao final desse processo, sob a convergência do algoritmo de estimação, obtemos a matriz de projeção $\mathbf{S}^*$ para o modelo de REGPM. Com base nesse resultado, obtemos que o valor $\text{NEP}^* = 2\nu_1^* - \nu_2^*$ representa o número efetivo de parâmetros nos modelos de REGPM, com $\nu_1^* = \text{tr}(\mathbf{S}^*)$ e $\nu_2^* = \text{tr}(\mathbf{S}^{*\top} \mathbf{S}^*)$. Vale ressaltar que essa medida não leva em consideração a quantidade efetiva dos parâmetros de dispersão locais e a estimação do ϕ fixo também pode ser usado como uma solução parcial para esse problema. Com isso, de forma semelhante ao capítulo anterior, é possível obter o ajuste do modelo de REGPM assumindo um parâmetro de dispersão global ϕ_g no processo de estimação. Nesse caso específico, a estimativa de ϕ_g é obtida no início do algoritmo de estimação através do ajuste de um modelo de regressão elíptico definido em (2.1) sob a mesma distribuição dos erros aleatórios. Como consequência, o número efetivo de parâmetros nesse caso é especificado por $\text{NEP}^* = 2\nu_1^* - \nu_2^* + 1$.

3.2.6 Testes de Hipóteses

Sob as condições usuais de regularidade (Cox e Hinkley, 1974, Cap. 9) temos que os estimadores de máxima verossimilhança obtidos através do algoritmo de estimação

backfitting são consistentes e assintoticamente normais. Dito isso, considerando o i -ésimo modelo local associado a k -ésima variável explicativa, sabe-se que:

$$\sqrt{n}(\hat{\beta}_{ik} - \beta_{ik}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_k^{-1})$$

em que $J = \lim_{n \rightarrow \infty} \frac{4dg}{n\phi(i)} \mathbf{X}_k^\top \mathbf{W}_{ik} \mathbf{X}_k$. Naturalmente, esse resultado pode ser utilizado como forma de obter os testes de hipótese para a significância de cada um dos parâmetros locais. Entretanto, vale ressaltar que a aplicação dos testes de significância para parâmetros locais em modelos geograficamente ponderados sofrem com a ocorrência incorreta de falsos positivos (erro do tipo I) induzidos por múltiplas hipóteses. A solução desse problema foi discutida no capítulo anterior e uma correção para o nível descritivo α para a REGPM foi considerada. No entanto, Yu *et al.* (2020) argumenta que nos modelos geograficamente ponderados com múltiplas escalas espaciais o ajuste do valor de α deve ser realizado separadamente para cada variável explicativa. Isso se deve pelo fato de que cada superfície de parâmetros tem sua própria escala espacial. Nesse caso, foi discutido anteriormente que as R_k representam uma matriz de projeção específica para cada variável explicativa. Essas matrizes podem ser usadas para derivar o número de parâmetros efetivos associado a cada uma das variáveis explicativas, de tal forma que:

$$\text{NEP}_k = \text{tr}(\mathbf{R}_k)$$

em que NEP_k representa o número efetivo de parâmetros associado à k -ésima variável explicativa. Com base nesse resultado, o nível descritivo ajustado α_k para os parâmetros locais associados à k -ésima variável explicativa com uma taxa de erro ξ é dado pela seguinte expressão:

$$\alpha_k = \frac{\xi}{\text{NEP}_k}$$

3.3 MÉTODOS DE DIAGNÓSTICO

Como foi ressaltado anteriormente, as medidas de diagnóstico são fundamentais para a validação do modelo. Dessa forma, adaptamos algumas dessas medidas discutidas na REGP para o caso de múltiplas escalas espaciais. De forma geral, propomos avaliar o impacto que a i -ésima observação realiza no modelo ajustado para o i -ésimo local. No entanto, como a REGPM apresenta um processo de estimação mais complexo baseado em múltiplos ajustes, então há uma dificuldade adicional para definição e obtenção das

medidas de diagnóstico. Por este fato, em alguns casos é mensurado o impacto da i -ésima observação para cada parâmetro de forma individual. Além disso, vale ressaltar que essa abordagem novamente será válida apenas quando os locais em estudo apresentarem as mesmas coordenadas geográficas que os pontos de dados observados.

3.3.1 Análise de Resíduos

Como discutido no capítulo anterior, os resíduos são amplamente utilizados para verificar as suposições do modelo proposto. Além disso, a análise de resíduos é particularmente importante para a identificação de observações discrepantes no ajuste do modelo. Para a REGPM consideramos o uso de duas formas simples de resíduos, visto que a complexidade do modelo dificulta a obtenção de propostas mais refinadas. Nesse contexto, inicialmente propomos uma medida baseada na padronização dos resíduos ordinais pela variância de y , a qual é denominada por resíduos de Pearson. Com base no modelo ajustado, o i -ésimo elemento para os resíduos de Pearson é expresso pela seguinte equação:

$$r_{pi} = \frac{y_i - \hat{\mu}_i}{(\xi \hat{\phi}(i))^{1/2}},$$

Além disso, como mencionado no capítulo anterior, os resíduos quantílicos podem ser uma boa alternativa para essa medida de discrepância, devido suas propriedades estatísticas. Esses resíduos também são amplamente utilizada em classes de modelos não-paramétricos e semi-paramétricos, como é o caso dos MAG, por exemplo. Para obtenção desses resíduos para o caso dos modelos de REGPM, considere novamente que $F(y, \mu, \phi)$ representa a função de distribuição acumulada da variável aleatória y , sendo essa contínua e uniformemente distribuída no intervalo unitário. Dito isso, o i -ésimo resíduo quantílico pode ser expresso por:

$$r_{qi} = \Phi^{-1}\{F(y_i, \hat{\mu}_i, \hat{\phi}(i))\},$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada da normal padrão. Nesse caso, pelo mesmo motivo exposto no capítulo anterior, sabe-se que r_{qi} também segue uma distribuição normal padrão. Devido a esse fato, os resíduos quantílicos permitem verificar facilmente as suposições do modelo relacionadas aos erros aleatórios.

3.3.2 Medidas de Alavancagem

Assim como discutimos no capítulo anterior, uma forma relevante para a detecção dos pontos de alavanca é através do estudo da matriz que projeta as observações de y em seus valores ajustados. No caso dos modelos de REGPM, essa matriz é representada pela quantidade $\mathbf{S}^*$, obtida anteriormente através do algoritmo de estimação. Resumidamente, os valores da diagonal principal dessa matriz podem indicar observações que exercem um peso desproporcional na estimação do modelo. Dessa forma, seja s_{ii}^* o i -ésimo elemento diagonal de $\mathbf{S}^*$, então valores altos de s_{ii}^* indicam que a i -ésima observação é um possível ponto de alavanca.

3.3.3 Influência Local

De forma semelhante ao capítulo anterior, adaptamos as medidas e influência local propostas inicialmente por Cook (1986) com base em duas funções objetivos sob alguns esquemas de perturbação. No entanto, essas medidas são obtidas individualmente para cada variável explicativa, devendo ser interpretadas de forma separadas. Dito isto, a obtenção das medidas de influência local para a k -ésima variável explicativa nos modelos de REGPM é descrita com detalhes a seguir.

3.3.3.1 Influência local na estimação

No processo de estimação da REGPM cada variável explicativa tem sua própria função de verossimilhança local, devido aos múltiplos ajustes dos modelos de REGP. Dessa forma, assuma que $WL_k(\boldsymbol{\theta})$ representa a verossimilhança local associada à k -ésima variável explicativa obtida através da equação (2.3) regressando $(\hat{\mathbf{f}}_k + \hat{\boldsymbol{\epsilon}})$ em $\mathbf{X}_k$. Suponha agora que $WL_k(\boldsymbol{\theta}|\boldsymbol{\omega})$ representa a função de verossimilhança local associada à k -ésima variável explicativa incluindo o vetor de perturbação $\boldsymbol{\omega}$. Assuma também que há um vetor $\boldsymbol{\omega}_0$ de não perturbação, de tal forma que $WL_k(\boldsymbol{\theta}|\boldsymbol{\omega}_0) = WL_k(\boldsymbol{\theta})$. Com isso, a distância entre as verossimilhanças locais é expressa por:

$$LD(\boldsymbol{\omega})_k = 2\{WL_k(\hat{\boldsymbol{\theta}}) - WL_k(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})\} \geq 0,$$

em que $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ representam as estimativas de máxima verossimilhança local para a k -ésima variável explicativa do modelo com e sem perturbação, respectivamente. Nesse caso, o

cálculo da influência local pode ser realizado por meio da curvatura normal para $\boldsymbol{\delta}$ na direção $\mathbf{l}$, $\|\mathbf{l}\| = 1$, a qual é dada pela seguinte expressão:

$$\mathbf{C}_{(i)k}(\boldsymbol{\theta}) = 2 \left| \mathbf{l}^\top \boldsymbol{\Delta}_{(i)k}^\top (\ddot{\mathbf{L}}_{\boldsymbol{\theta}\boldsymbol{\theta}k})^{-1} \boldsymbol{\Delta}_{(i)k} \mathbf{l} \right|.$$

em que $\ddot{\mathbf{L}}_{\boldsymbol{\theta}\boldsymbol{\theta}k}$ é a matriz informação de Fisher observada do modelo ajustado para a k -ésima variável explicativa e $\boldsymbol{\Delta}_{(i)k} = \partial^2 W L_k(\boldsymbol{\theta}, \boldsymbol{w}) / \partial \boldsymbol{\theta} \partial \boldsymbol{w}^\top$ é uma matriz de perturbações para a k -ésima variável explicativa no i -ésimo local. Nesse caso, as mesmas abordagens sugeridas no capítulo anterior e descritas em Cook (1986) e Lesaffre e Verbeke (1998) podem ser consideradas. Com isso, os esquemas de perturbação considerados para estimativa dos parâmetros são descritos a seguir.

- Perturbação na escala – Considere o esquema de perturbação no parâmetro de dispersão, para o modelo associado à k -ésima variável explicativa no i -ésimo local, na forma de $\phi(i)/\omega_i$. Feito isso, temos que $\boldsymbol{\omega}_0 = (1, \dots, 1)$ é o vetor de não perturbação e a função de verossimilhança local é descrita pela forma:

$$W L_k(\boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{j=1}^n \left\{ -\frac{1}{2} \log(\phi(i)/\omega_j) + \log\{g(\omega_j u_j)\} \right\} w_{ijk}.$$

Portanto, os elementos da matriz de perturbação associada à k -ésima variável explicativa no i -ésimo local são expressos sob a forma $\boldsymbol{\Delta}_{(i)k} = (\boldsymbol{\Delta}_{\boldsymbol{\beta}(i)k}^\top, \boldsymbol{\Delta}_{\phi(i)k}^\top)^\top$, com elementos dados por:

$$\boldsymbol{\Delta}_{\boldsymbol{\beta}(i)k} = -\frac{2}{\phi(i)} \mathbf{X}_k^\top \mathbf{W}_{ik} \mathbf{D}(b) \quad \text{e} \quad \boldsymbol{\Delta}_{\phi(i)k} = -\frac{1}{\phi(i)^2} \mathbf{W}_{ik} \mathbf{D}(b).$$

- Perturbação ponderação de casos – Para este caso assuma inicialmente o vetor de não perturbação $\boldsymbol{\omega}_0 = (1, \dots, 1)$ e o vetor de perturbação $\boldsymbol{\omega} = (\omega_1, \dots, \omega_n)$. Dessa forma, a função de verossimilhança local associada à k -ésima variável explicativa é ponderada da seguinte forma:

$$W L_k(\boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{j=1}^n \omega_j \left\{ -\frac{1}{2} \log(\phi(i)) + \log\{g(u_j)\} \right\} w_{ijk}.$$

Portanto, assumindo esse esquema a matriz de perturbação para a k -ésima variável explicativa no i -ésimo local é expressa por $\boldsymbol{\Delta}_{(i)k} = (\boldsymbol{\Delta}_{\boldsymbol{\beta}(i)k}^\top, \boldsymbol{\Delta}_{\phi(i)k}^\top)^\top$, com elementos dado por:

$$\boldsymbol{\Delta}_{\boldsymbol{\beta}(i)k} = \frac{1}{\phi(i)} \mathbf{D}(v) \mathbf{W}_{ik} \mathbf{D}(e) \mathbf{X}_k \quad \text{e} \quad \boldsymbol{\Delta}_{\phi(i)k} = \frac{1}{2\phi(i)} \mathbf{W}_{ik} \mathbf{V}(m).$$

3.3.3.2 Influência local na predição

De forma análoga ao capítulo anterior, considere a função de predição em $\mathbf{q}$ para o modelo perturbado, a qual pode ser expressa por:

$$\hat{\boldsymbol{\mu}}(\mathbf{q}, \boldsymbol{\omega}) = \sum_{k=1}^p q_j \hat{\beta}(i)_{k\boldsymbol{\omega}}.$$

Naturalmente, assumimos que existe um vetor de não perturbação $\boldsymbol{\omega}_0$ de tal forma que $\hat{\boldsymbol{\mu}}(\mathbf{q}) = \hat{\boldsymbol{\mu}}(\mathbf{q}, \boldsymbol{\omega})$ representa a função de predição em $\mathbf{q}$ sem perturbação. Com isso, definimos a função objetivo por meio da seguinte equação:

$$f(\mathbf{q}, \boldsymbol{\omega}) = \{\hat{\boldsymbol{\mu}}(\mathbf{q}) - \hat{\boldsymbol{\mu}}(\mathbf{q}, \boldsymbol{\omega})\}^2. \quad (3.4)$$

Visto isto, temos o interesse em investigar a curvatura normal na superfície formada por $\boldsymbol{\omega}$ e $f(\mathbf{q}, \boldsymbol{\omega})$ em torno de $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. Desse modo, a curvatura normal na direção unitária $\mathbf{l}$ associada à k -ésima variável explicativa é expressa pela forma:

$$C_{(i)k}(\boldsymbol{\theta}) = |\mathbf{l}^\top \ddot{f}_k \mathbf{l}|,$$

em que $\ddot{f}_k = \partial^2 f(\mathbf{q}, \boldsymbol{\omega}) / \partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^\top$ é avaliada em $\boldsymbol{\omega}$ e $\hat{\beta}_{(i)k}$. Portanto, essa medida associada à k -ésima variável explicativa é expressa por:

$$\ddot{f}_k = \frac{\partial^2}{\partial \boldsymbol{\omega} \partial \boldsymbol{\omega}^\top} f = -2 \Delta_{(i)k}^\top \ddot{L}_{\boldsymbol{\theta}\boldsymbol{\theta}k}^{-1} \mathbf{q} \mathbf{q}^\top \ddot{L}_{\boldsymbol{\theta}\boldsymbol{\theta}k}^{-1} \Delta_{(i)k}.$$

Nesse caso, uma sugestão é utilizar a maior curvatura, na qual tem a direção $\mathbf{l}_{max}$, como forma de avaliar a influência das observações na predição do modelo. Os dois esquemas de perturbação a seguir são considerados para a avaliar o impacto das observações na predição do modelo.

- Perturbação aditiva na resposta – Considere a perturbação da variável resposta na forma $y_{i\boldsymbol{\omega}} = y_i + \omega_i s_i$, em que s_i é a estimativa do desvio padrão de y_i . Nesse caso, temos que $\boldsymbol{\omega}_0 = (0, \dots, 0)$ é o vetor de não perturbação e a função de verossimilhança local assume a seguinte forma:

$$WL_k(\boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{j=1}^n \left\{ -\frac{1}{2} \log(\phi(i)) + \log\{g(u_{j\boldsymbol{\omega}})\} \right\} w_{ijk}.$$

Nesse caso, a matriz de perturbação associada à k -ésima variável explicativa para o i -ésimo local é expressa pela seguinte forma matricial:

$$\Delta_{\beta(i)k} = \frac{1}{\phi(i)} \mathbf{X}_k^\top \mathbf{W}_{ik} \mathbf{D}(a).$$

- Perturbação aditiva nos preditores – Considere a perturbação da k -ésima variável explicativa contínua na forma $x_{ik\omega} = x_{ik} + \omega_i s_k$, com s_k representando um fator de escala. Dessa forma, temos que $\omega_0 = (0, \dots, 0)$ é o vetor de não perturbação e a função de verossimilhança local para a k -ésima variável explicativa é expressa por:

$$WL_k(\theta, \omega) = \sum_{j=1}^n \left\{ -\frac{1}{2} \log(\phi(i)) + \log\{g(u_{j\omega})\} \right\} w_{ijk}.$$

Logo, a matriz de perturbação associada à k -ésima variável explicativa para o i -ésimo local para esse esquema de perturbação é expressa pela seguinte forma matricial:

$$\Delta_{\beta(i)_k} = s_k(\mathbf{F}\mathbf{D}(g)\mathbf{D}(e) - \phi(i)^{-1}\beta(i)_k\mathbf{X}_k^\top\mathbf{D}(a))\mathbf{W}_{ik},$$

em que $\mathbf{F}$ é uma matriz de zeros $1 \times n$ com o valor 1 na k -ésima linha.

3.4 ESTUDO DE SIMULAÇÃO

Com o objetivo de avaliar o desempenho dos modelos de REGPM sob distintas situações, foram desenvolvidos diferentes cenários simulados com base em estudos empíricos anteriores para modelos de RGPM (ver, por exemplo, Fotheringham, Yang e Kang, 2017 e Yu *et al.*, 2020). De forma geral, o propósito desse estudo de simulação é responder às seguintes dúvidas:

- A REGPM produz estimativas confiáveis para as escalas espaciais através dos parâmetros de suavização independentes para cada variável explicativa?
- A REGPM fornece estimativas semelhantes à REGP quando as superfícies locais de todos parâmetros variam na mesma escala espacial?
- A REGPM fornece estimativas mais precisas para as superfícies locais dos parâmetros sob diferentes escalas espaciais em relação a REGP?
- A REGPM gera estimativas espaciais menos sensíveis na presença de valores discrepantes em relação a RGPM ao assumir distribuições de caudas mais pesadas para os erros aleatórios?
- A REGPM produz estimativas mais precisas da variável resposta na presença de valores discrepantes sob diferentes escalas espaciais em comparação com os modelos de REGP e RGPM ao assumir distribuições de caudas mais pesadas para os erros aleatórios?

Dessa forma, examinamos o comportamento desses modelos aplicando-os a cenários simulados com propriedades conhecidas e diferentes níveis de heterogeneidade

especial entre as superfícies locais dos parâmetros. Para isso, foi produzido uma grade regular de dimensões 25×25 e três cenários distintos para a geração dos dados. Para comparar o desempenho entre os modelos ajustados utilizamos as mesmas medidas consideradas nas simulações do capítulo anterior. Dito isto, a seguir é definido os três diferentes cenários de simulação, bem como os resultados obtidos através dos respectivos estudos.

3.4.1 Cenário 1 - Escalas Iguais

O primeiro cenário tem como finalidade avaliar o desempenho do modelo REGPM quando os coeficientes de regressão variam na mesma escala espacial. Com base nesse contexto, simulamos 100 replicações supondo um modelo geograficamente ponderado na forma:

$$y_i = \beta_0(h_i, v_i) + \beta_1(h_i, v_i)x_i + \epsilon_i$$

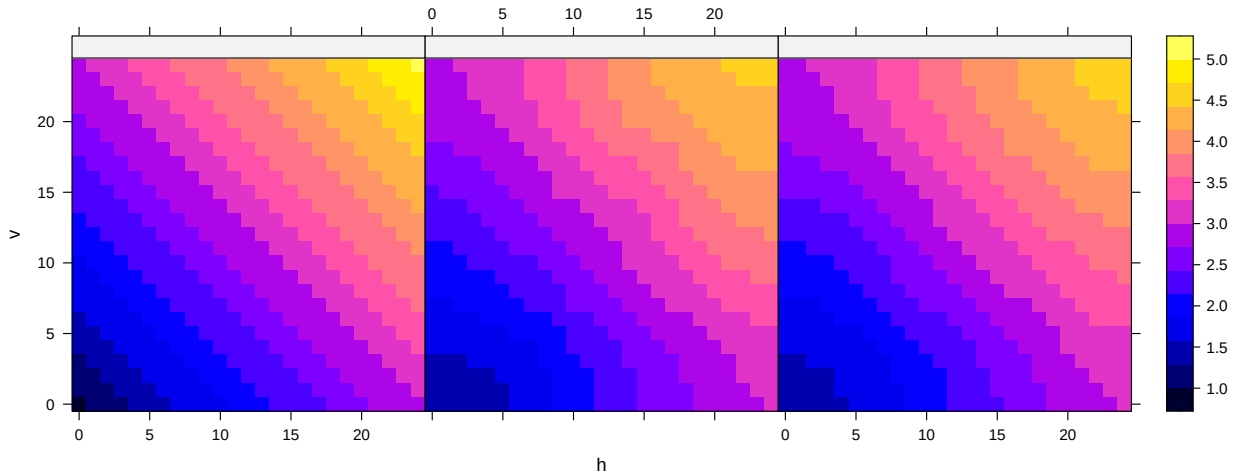
em que $x = (x_1, \dots, x_n)$ foram gerados a partir de uma distribuição normal padrão e $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ de uma distribuição t -Student com 3 graus de liberdade. As coordenadas (h_i, v_i) representam o ponto central da região (definida por um polígono retangular de área um) associada à i -ésima observação, em que $h_i = 0, 1, \dots, 25$ e $v_i = 0, 1, \dots, 25$. Para os coeficientes de regressão, projetamos duas superfícies de parâmetros diferentes na mesma escala de variação com base nas seguintes regras:

$$\beta_0(h_i, v_i) = 1 + \frac{1}{12}(h_i + v_i^+) \quad \text{e} \quad \beta_1(h_i, v_i) = 1 + \frac{1}{12}(h_i + v_i^-),$$

em que v_i^+ e v_i^- representam i -ésimo valor em ordem crescente e decrescente de $v = (v_1, \dots, v_n)$, respectivamente. Dessa forma, ajustamos os modelos de REGP e REGPM para cada uma das simulações, assumindo uma distribuição t -Student com 3 graus de liberdade para ambos modelos. No processo de estimação considerou-se o uso da função *kernel* biquadrado e a distância Euclidiana para o calculo das ponderações espaciais. A seleção do parâmetro de suavização foi realizado apenas por meio do método da validação cruzada. Visto isso, nas Figuras 36 e 37 são apresentadas as distribuições espaciais dos parâmetros reais e as respectivas médias das estimativas obtidas através dos modelos ajustados. Note que a superfície espacial das estimativas em ambos modelos são bem próximas dos valores reais, indicando que os resultados são semelhantes quando as escalas espaciais dos parâmetros de regressão são as mesmas. Com base nesse estudo de simulação,

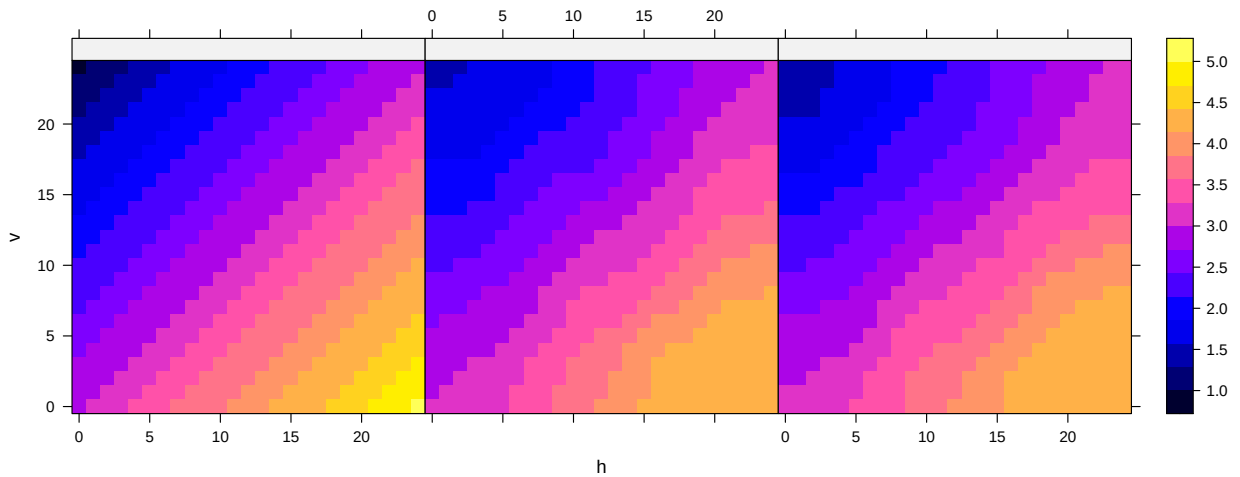
constatou-se que a REGPM fornece estimativas semelhantes à REGP quando as superfícies locais de todos parâmetros variam na mesma escala espacial.

Figura 36 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 da RGP (meio) e REGP com erros t -Student (direita).



Fonte: O autor (2021).

Figura 37 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 da RGP (meio) e REGP com erros t -Student (direita).

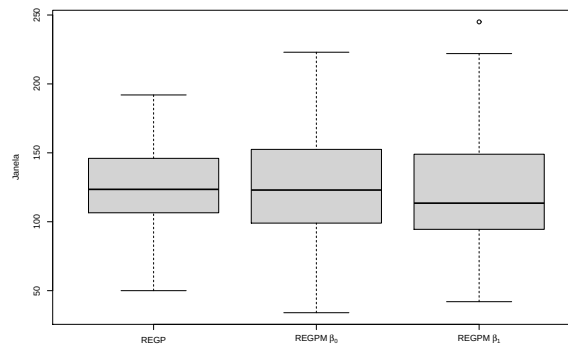


Fonte: O autor (2021).

Além disso, a Figura 38 mostra os *boxplots* dos valores ótimos selecionados para os parâmetros de suavização dos modelos geograficamente ponderados através do método da validação cruzada. Nesse caso, como as duas superfícies de parâmetro têm as mesmas escalas espaciais, a quantidade de vizinhos considerados para os valores ótimos dos parâmetros de suavização da REGP e REGPM são bem próximos. Em outras palavras, a

REGPM produz estimativas confiáveis para as escalas espaciais através dos parâmetros de suavização independentes para cada variável explicativa. Este fato indica mais uma vez que a REGPM apresenta estimativas confiáveis para as escalas espaciais quando as relações de regressão têm o mesmo grau de heterogeneidade espacial.

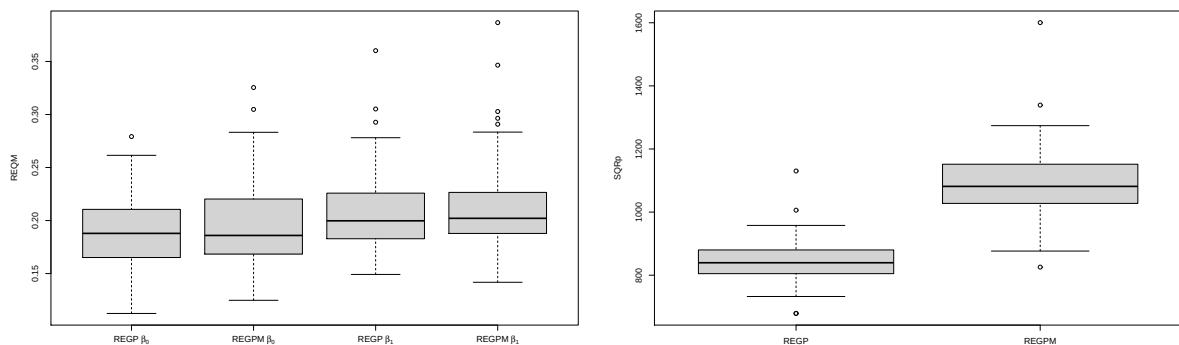
Figura 38 – *Boxplots* dos valores ótimos para os parâmetros de suavização nos modelos de REGP e REGPM via validação cruzada



Fonte: O autor (2021).

Adicionalmente, a Figura 39 mostra os *boxplots* dos valores das medidas REQM e SQRp para os coeficientes estimados dos modelos de REGP e REGPM com erros aleatórios *t*-Student. Observe que, os modelos ajustados apresentam valores semelhantes de REQM, reforçando a premissa de que ambos modelos conseguem replicar de maneira semelhante a superfície real dos parâmetros quando estão na mesma escala espacial. O mesmo ocorre para os valores de SQRp, no entanto o modelo REGP apresenta resultados ligeiramente melhores na predição da variável resposta.

Figura 39 – *Boxplots* da REQM (esquerda) e SQRp (direita) para os modelos de REGP e REGPM



Fonte: O autor (2021).

3.4.2 Cenário 2 - Escalas Diferentes

Este cenário simulado tem o objetivo de verificar o desempenho do modelo REGPM quando os coeficientes de regressão apresentam escalas espaciais distintas. De forma semelhante ao cenário anterior, simulamos 100 replicações supondo um modelo geograficamente ponderado através da seguinte formula:

$$y_i = \beta_0(h_i, v_i) + \beta_1(h_i, v_i)x_{1i} + \beta_2(h_i, v_i)x_{2i} + \epsilon_i$$

em que $x = (x_1, \dots, x_n)$ foram gerados a partir de uma distribuição normal padrão e $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ de uma distribuição t -Student com 3 graus de liberdade. As coordenadas (h_i, v_i) representam o ponto central da região (definida por um polígono retangular de área um) associada à i -ésima observação, em que $h_i = 0, 1, \dots, 25$ e $v_i = 0, 1, \dots, 25$. Para os coeficientes de regressão, projetamos três superfícies de parâmetros diferentes com escalas de variações espaciais distinta, baseado nas seguintes regras:

$$\beta_0(h_i, v_i) = 3,$$

$$\beta_1(h_i, v_i) = 1 + \frac{1}{12}(h_i + v_i^-)$$

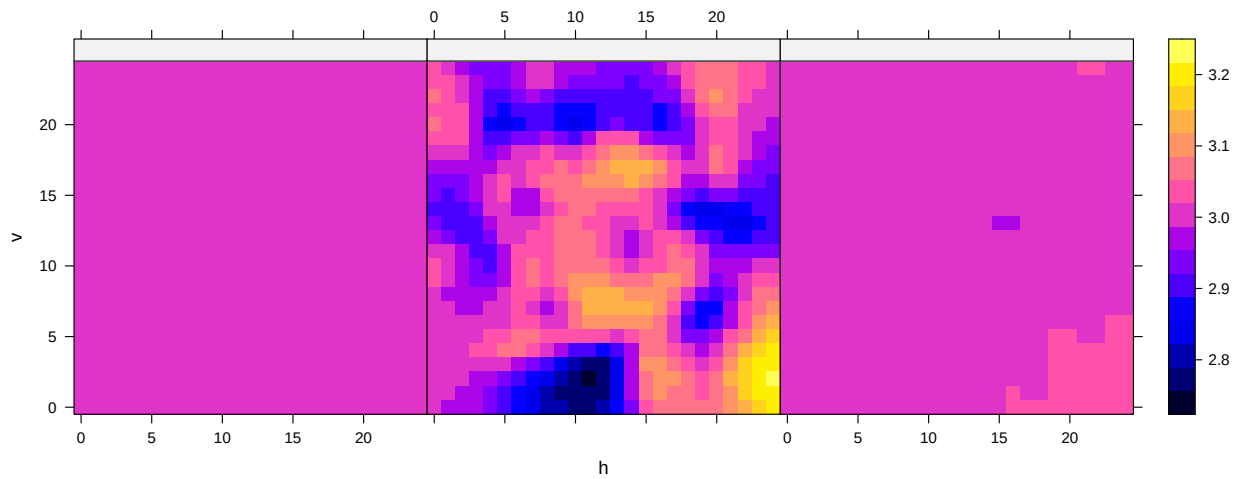
e

$$\beta_2(h_i, v_i) = 1 + \frac{1}{324} \left(36 - \left(6 - \frac{h_i}{2} \right)^2 \right) \left(36 - \left(6 - \frac{v_i}{2} \right)^2 \right).$$

De acordo com Fotheringham, Yang e Kang, 2017 os coeficientes β_0 , β_1 e β_2 apresentam uma escala de variação baixa, média e alta, respectivamente. Dessa forma, ajustamos os modelos de REGP e REGPM para cada uma das simulações, considerando uma distribuição t -Student com 3 graus de liberdade. Para o processo de estimação foi utilizado uma função *kernel* biquadrado e a distância Euclidiana para a construção das ponderações espaciais. A seleção do parâmetro de suavização ótimo é realizada novamente por meio do método da validação cruzada. Dessa forma, as Figuras 40-42 exibem as superfícies espaciais dos parâmetros reais e as respectivas médias das estimativas obtidas pelo ajuste dos modelos geograficamente ponderados. Observe que o modelo de REGP tem dificuldade em replicar a real superfície dos parâmetros quando as escalas espaciais são diferentes. Isso é claramente observado na distribuição espacial de β_0 , uma vez que

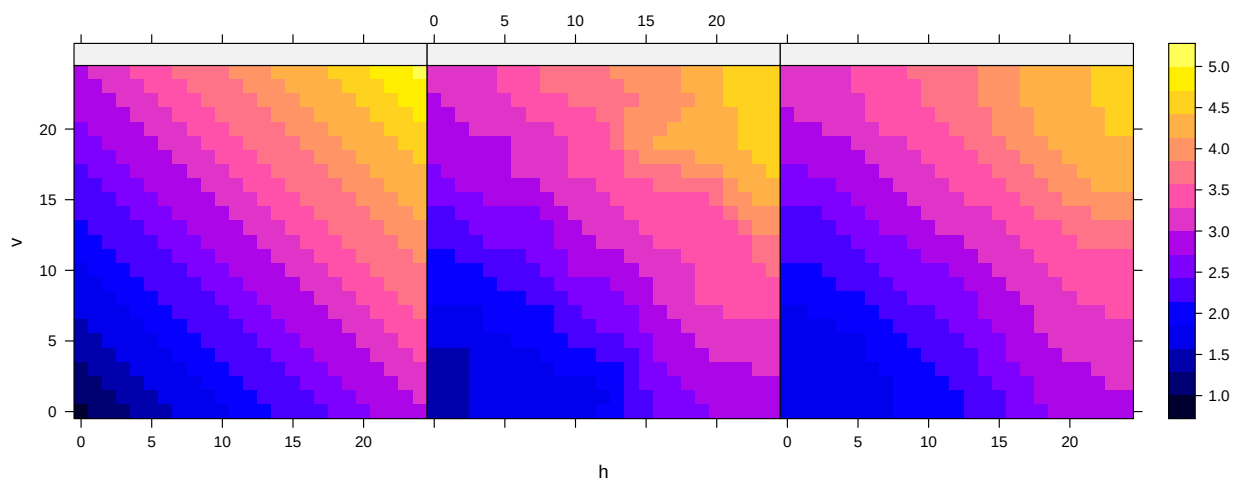
as suas estimativas tem uma variação espacial razoável quando na verdade deveria ser constante no espaço. Nota-se também uma leve distorção nas estimativas do parâmetro β_1 , entretanto as estimativas de β_2 são bem próximas dos valores reais. Isso ocorre pelo fato de que o parâmetro de suavização único selecionado na REGP consegue representar bem a escala espacial de β_2 , mas o mesmo não ocorre para os demais parâmetros de regressão. Nitidamente isso não é um problema para a REGPM, pois os parâmetro de suavização individuais para cada parâmetro conseguem representar bem todas as escalas espaciais.

Figura 40 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 da RGP (meio) e REGP com erros t -Student (direita).



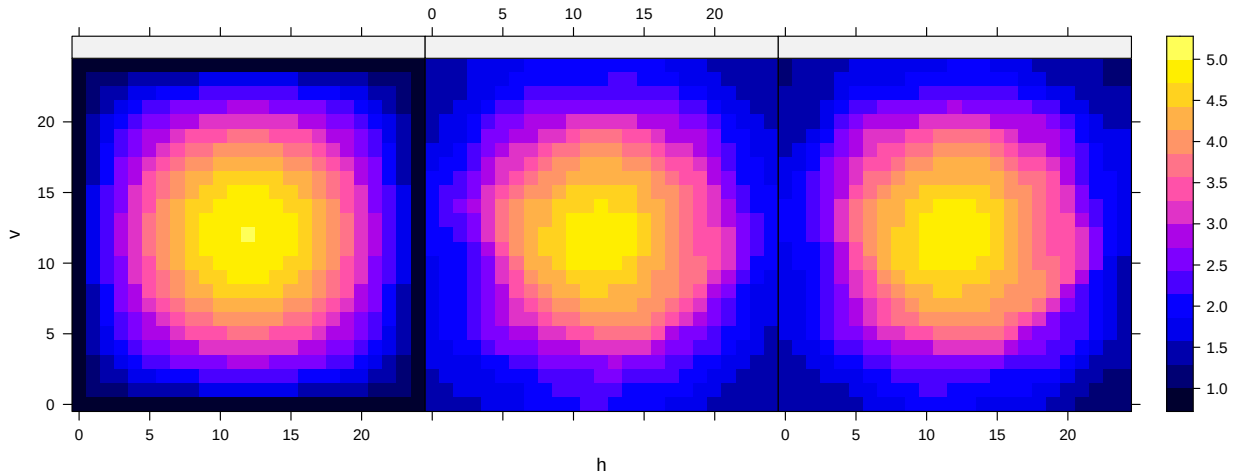
Fonte: O autor (2021).

Figura 41 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 da RGP (meio) e REGP com erros t -Student (direita).



Fonte: O autor (2021).

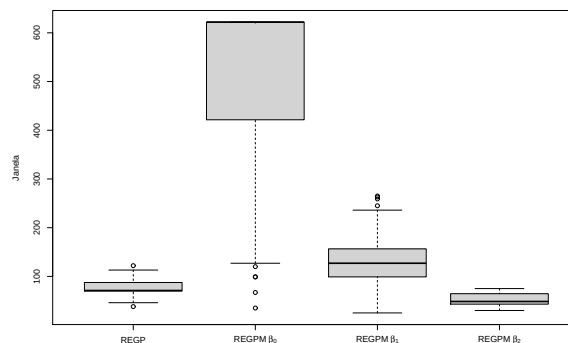
Figura 42 – Superfície espacial verdadeira de β_2 (esquerda), superfícies espaciais médias das estimativas de β_2 da RGP (meio) e REGP com erros t -Student (direita).



Fonte: O autor (2021).

Além disso, a Figura 43 exibe os *boxplots* da quantidade de vizinhos obtidos na otimização dos parâmetros de suavização por meio do método da validação cruzada. Como as superfícies de parâmetros tem graus de heterogeneidade espacial diferentes, então os parâmetros de suavização selecionadas nos modelos REGPM conseguem captar bem a variação das escalas espaciais. Nesse caso, os modelos selecionam valores altos para os parâmetros de suavização associados a β_0 , medianos para β_1 e pequenos para β_2 , refletindo bem a heterogeneidade espacial real dos parâmetros de regressão. Por outro lado, a REGP assume que todos os parâmetros variam na mesma escala espacial através de um único parâmetro de suavização, impossibilitando a distinção das escalas entre os parâmetros de regressão. Por esse fato, as estimativas obtidas por meio da REGP não reproduzem bem a superfície espacial real dos parâmetros de regressão.

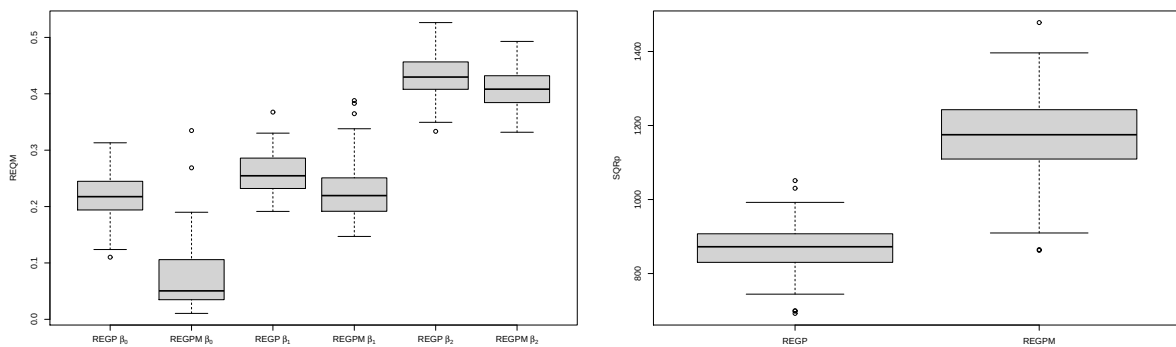
Figura 43 – *Boxplots* dos valores ótimos para os parâmetros de suavização nos modelos de REGP e REGPM via validação cruzada



Fonte: O autor (2021).

Na Figura 44 são exibidos os *boxplots* dos valores da REQM e SQRp para a dos modelos de REGP e REGPM. Observe que, os valores da REQM indicam que a REGPM consegue replicar melhor a superfície real dos parâmetros, uma vez que levam em consideração a escala espacial variável entre os parâmetros. Isso é ressaltado principalmente em relação ao parâmetro β_0 , pois o mesmo deveria ser fixo no espaço e o modelo REGP induz uma superfície espacial das estimativas de forma equivocada. Adicionalmente, ambos os modelos conseguem replicar bem a variável resposta com precisão, no entanto o modelo REGP tem valores de SQRp relativamente melhores. Isso deve ocorrer pelo fato de que as estimativas do parâmetro β_0 na REGP apresentam uma variação espacial inadequada, capturando efeitos residuais do modelo no processo de estimação. Em contrapartida esse fenômeno não acontece na REGPM, pois as estimativas de β_0 são essencialmente constates no espaço. Portanto, esses resultados mostram que o modelo de REGPM fornece estimativas mais precisas para as superfícies locais dos parâmetros quando as escalas são distintas.

Figura 44 – *Boxplots* da REQM (esquerda) e SQRp (direita) para os modelos de REGP e REGPM



Fonte: O autor (2021).

3.4.3 Cenário 3 - Escalas Diferentes na Presença de Observação Discrepante

O último cenário tem como objetivo avaliar o desempenho da REGPM na presença de observações discrepantes, assumindo que os coeficientes de regressão apresentam escalas espaciais distintas. Novamente foram simulados 100 replicações supondo um modelo

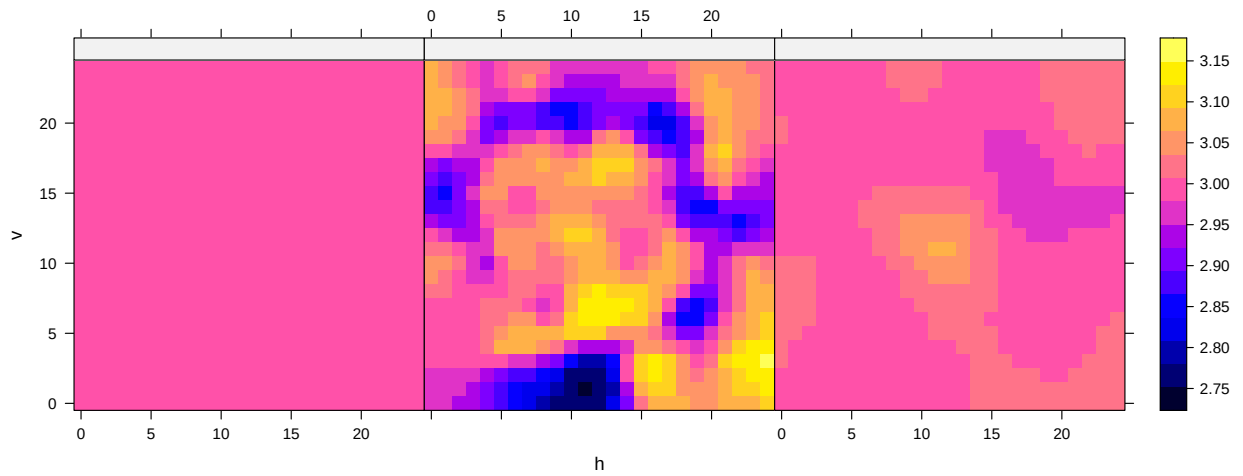
geograficamente ponderado na forma:

$$y_i = \beta_0(h_i, v_i) + \beta_1(h_i, v_i)x_{1i} + \beta_2(h_i, v_i)x_{2i} + \epsilon_i$$

em que $x = (x_1, \dots, x_n)$ e $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ foram gerados a partir de uma distribuição normal padrão. As coordenadas (h_i, v_i) representam o ponto central da região (definida por um polígono retangular de área um) associada à i -ésima observação, em que $h_i = 0, 1, \dots, 25$ e $v_i = 0, 1, \dots, 25$. Nesse cenário, projetamos três superfícies de parâmetros com escalas de variações distintas, com base nas mesmas regras de geração apresentadas no cenário anterior. Isto significa novamente que os coeficientes β_0 , β_1 e β_2 apresentam uma escala de variância baixa, média e alta, respectivamente. Para cada uma das simulações, selecionamos o local referente às coordenadas (11,12) e substituímos o valor da variável resposta pelo seu valor máximo mais três vezes o desvio padrão. Com isso, ajustou-se inicialmente os modelos de RGP e REGP assumindo erros aleatórios t -Student com 3 graus de liberdade. Em seguida foram ajustados também os modelos de RGPM e REGPM supondo erros aleatórios t -Student com 3 graus de liberdade. Para todos os casos, assumimos mais uma vez a função *kernel* biquadrado e a distância Euclidiana para as ponderações espaciais. O método da validação cruzada foi utilizado para a otimização do parâmetro de suavização. Esse procedimento foi repetido para cada uma das 100 replicas, comparando os resultados obtidos entre todos os modelos ajustados.

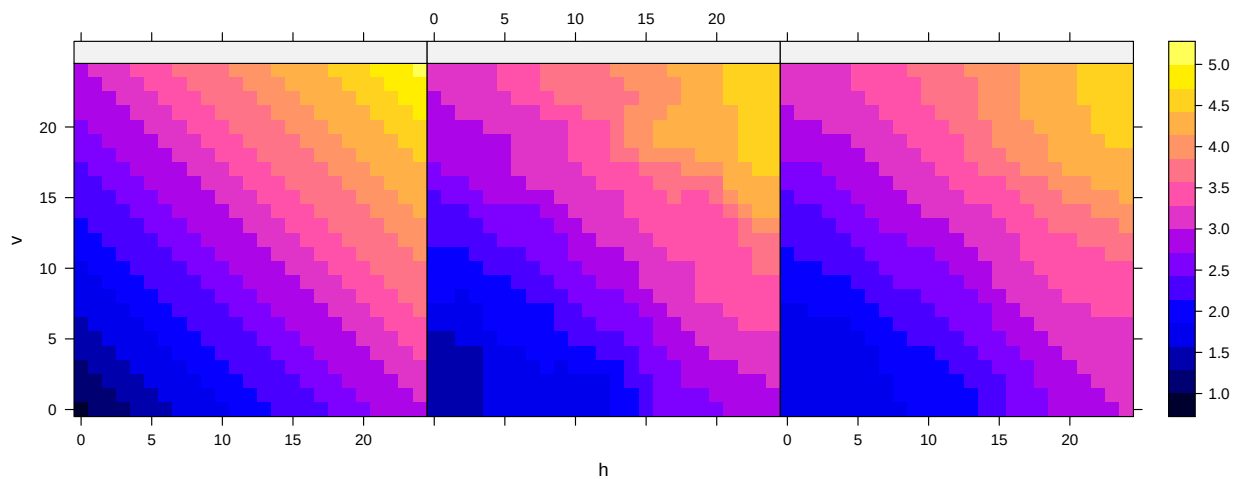
Com base nesse cenário simulado, as Figuras 45-50 exibem as distribuições espaciais para as médias das estimativas dos parâmetros de regressão dos modelos ajustados para todas as simulações. Os gráficos das Figuras 45-47 comparam os valores reais dos parâmetros com suas respectivas estimativas obtidas no ajuste dos modelos de REGP e REGPM com erros t -Student. Observe que, o modelo REGPM apresentou as superfícies espaciais mais próximas dos valores reais para todos coeficientes, obtendo resultados similares ao segundo cenário de simulação. Por outro lado, os gráficos contidos nas Figuras 48-50 fazem a mesma comparação entre os modelos de RGPM e REGPM com erros aleatórios t -Student. Nesse caso é possível notar que ambos representam bem as escalas espaciais dos parâmetros, entretanto observa-se uma distorção clara na superfície dos valores estimados pelos modelos com erros normais em torno da observação discrepante. Isto indica que o modelo RGPM embora apresente superfícies melhores que o modelo de REGP com erros t -Student, ainda é claramente sensível a presença de valores discrepantes.

Figura 45 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 da REGP (meio) e REGPM (direita).



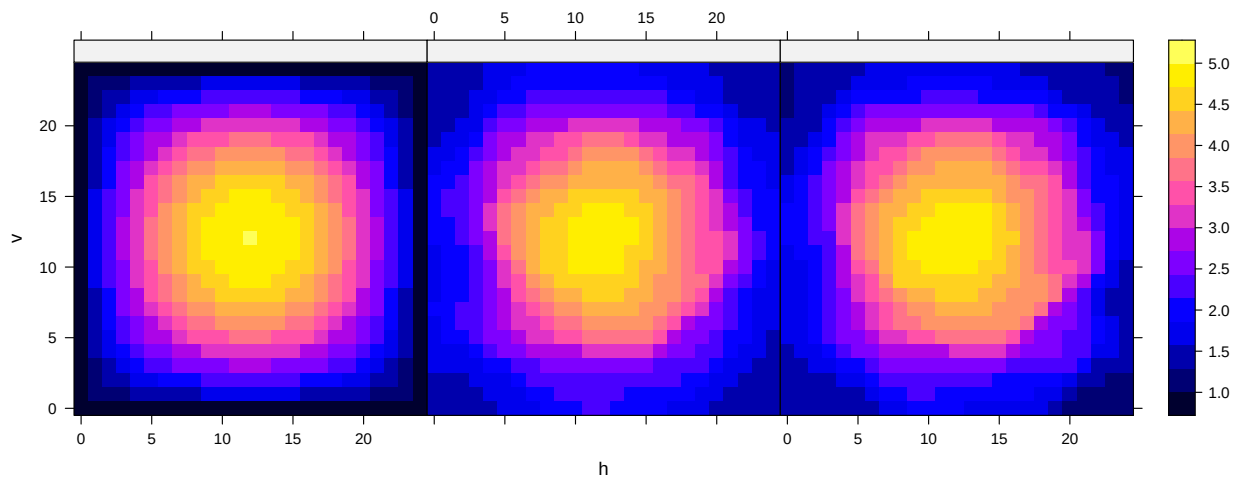
Fonte: O autor (2021).

Figura 46 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 da REGP (meio) e REGPM (direita).



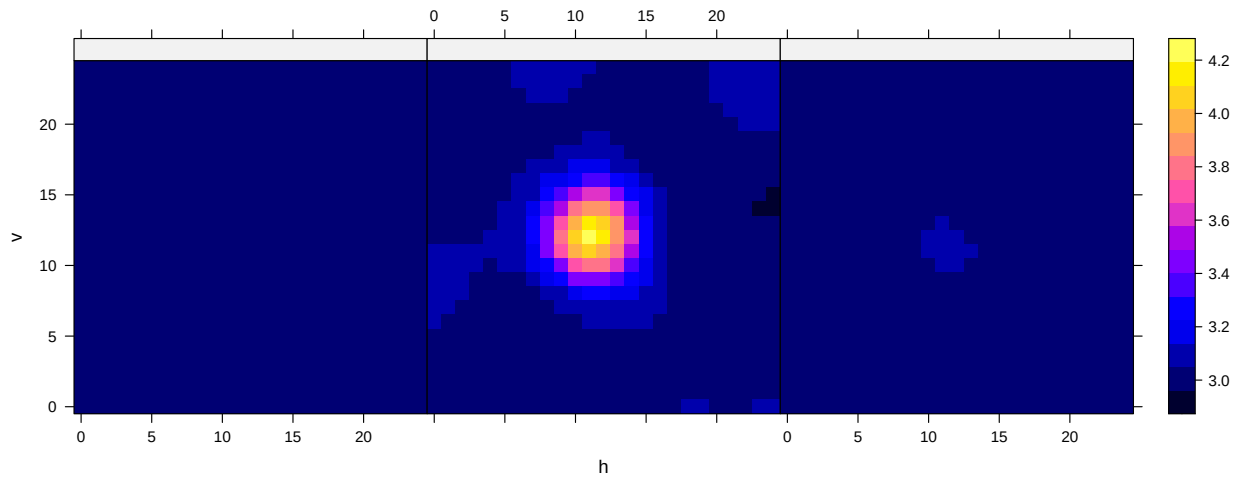
Fonte: O autor (2021).

Figura 47 – Superfície espacial verdadeira de β_2 (esquerda), superfícies espaciais médias das estimativas de β_2 da REGP (meio) e REGPM (direita).



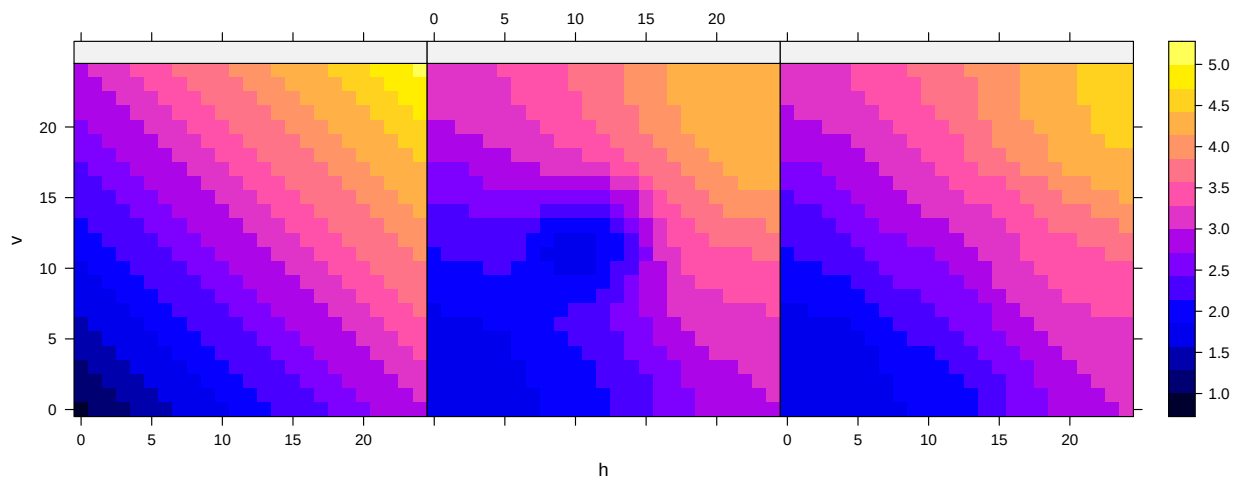
Fonte: O autor (2021).

Figura 48 – Superfície espacial verdadeira de β_0 (esquerda), superfícies espaciais médias das estimativas de β_0 da RGPM (meio) e REGPM (direita).



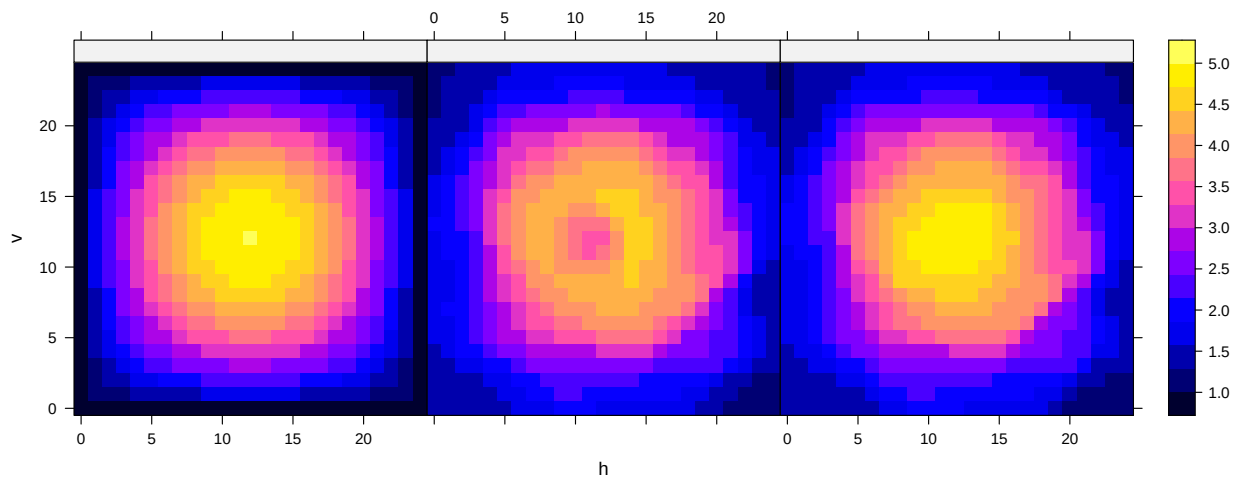
Fonte: O autor (2021).

Figura 49 – Superfície espacial verdadeira de β_1 (esquerda), superfícies espaciais médias das estimativas de β_1 da RGPM (meio) e REGPM (direita).



Fonte: O autor (2021).

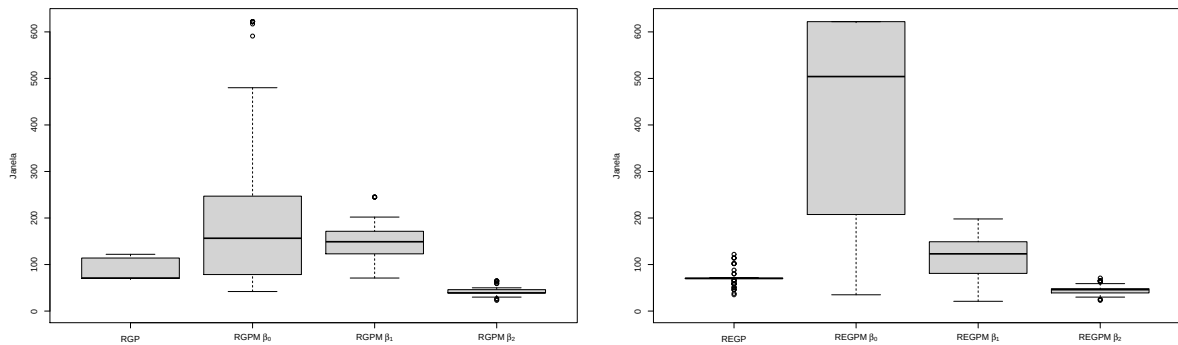
Figura 50 – Superfície espacial verdadeira de β_2 (esquerda), superfícies espaciais médias das estimativas de β_2 da RGPM (meio) e REGPM (direita).



Fonte: O autor (2021).

Adicionalmente, a Figura 51 mostram os *boxplots* da quantidade de vizinhos considerados para os valores ótimos dos parâmetros de suavização selecionados obtidas pelo método da validação cruzada. De forma semelhante ao cenário anterior, os parâmetros de suavização selecionadas nos modelos de REGPM conseguem captar bem a variação das escalas espaciais para cada parâmetro. Observa-se que, em ambos modelos de REGPM selecionam valores altos para β_0 , medianos para β_1 e pequenos para β_2 , representando bem a real escala espacial dos parâmetros. No entanto, a REGPM com erros normais tende a selecionar valores bem menores para o número de vizinhos para o parâmetro de suavização associado a β_0 em comparação a REGPM com erros *t*-Student. Isso é um reflexo da sensibilidade desse modelo em relação a observações discrepantes, afetando não apenas as estimativas dos parâmetros como também a seleção da janela ótima.

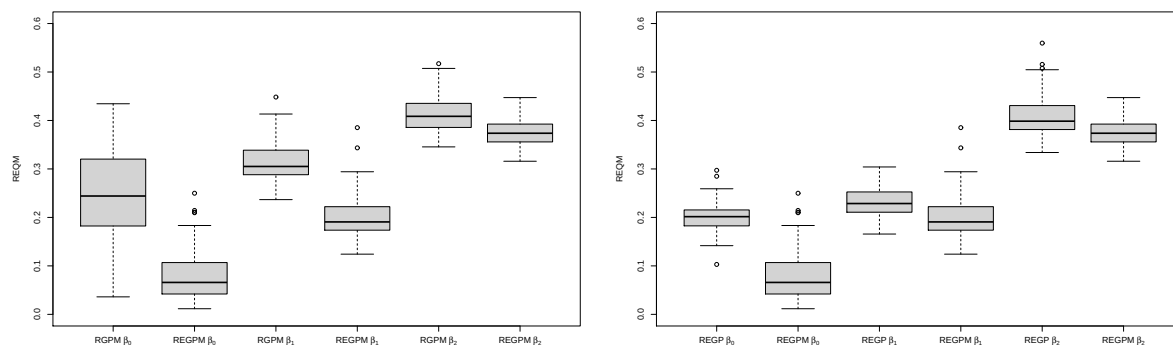
Figura 51 – *Boxplots* dos valores ótimos para os parâmetros de suavização nos modelos geograficamente ponderados com erros normais (esquerda) e *t*-Student (direita) via validação cruzada.



Fonte: O autor (2021).

Além disso, a Figura 52 exibe os *boxplots* dos valores da REQM obtidos no ajuste dos modelos geograficamente ponderados. Observe que o modelo de REGPM com erros *t*-Student mostram maior precisão para replicar as superfícies de parâmetros conhecidas em comparação com os demais ajustes. Embora o modelo de RGP com erros *t*-Student não apresente estimativas sensíveis na presença de valores discrepantes, um único parâmetro de suavização não é suficiente para representar bem as variações espaciais dos parâmetros. Por outro lado, a estimação do modelo de RGP sofre forte influência da presença do valor discrepante, ainda que esse modelo represente bem a escala espacial dos parâmetros.

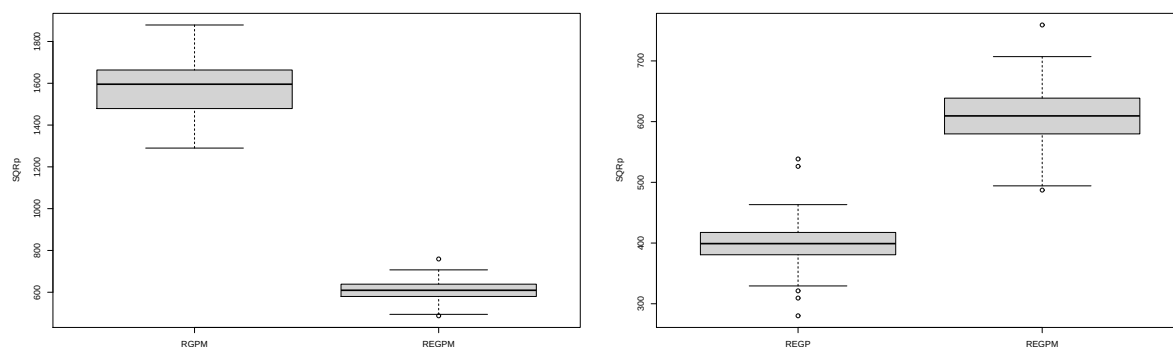
Figura 52 – *Boxplots* da REQM para o modelo de REGPM em comparação com os modelos de RGPM (esquerda) e REGP com erros *t*-Student (direita).



Fonte: O autor (2021).

Por fim, a Figura 53 exibe os *boxplots* dos valores da SQRp obtidos no ajuste dos modelos geograficamente ponderados. Nesse caso o modelo de RGPM tem as piores medidas de SQRp, indicando que a presença do valor discrepante afeta também a precisão em replicar os valores da variável resposta. Adicionalmente, os modelos com erros *t*-Student obtêm melhores valores de SQRp, no entanto o modelo REGP tem resultados relativamente melhores nesse ponto. Isso deve ocorrer pelo mesmo fato destacado no segundo cenário, visto que parâmetro β_0 na REGP apresenta uma variação espacial inadequada e captura efeitos residuais do modelo. Esses resultados mostram que a REGPM com erros *t*-Student fornece estimativas mais precisas para as superfícies locais dos parâmetros quando há valores discrepantes e diferenças de escalas espaciais entre os parâmetros.

Figura 53 – *Boxplots* da SQRp para o modelo de REGPM em comparação com os modelos de RGPM (esquerda) e REGP com erros *t*-Student (direita).



Fonte: O autor (2021).

3.5 APLICAÇÃO

Para ilustrar os métodos desenvolvidos neste capítulo consideramos a base de dados sobre o desempenho educacional nos condados da Geórgia, EUA. Esses dados são extensivamente usados na literatura em estudos de modelos geograficamente ponderados e está disponível no software R (R Core Team, 2019) sob o nome `georgia` através do pacote `spgwr`. As informações da base de dados foram extraídas do censo de 1990 nos EUA e contém um total de 159 observações. O objetivo desta aplicação é avaliar como a população com diploma de bacharel está relacionado com determinados indicadores sociais. Para isto estabelecemos um modelo em que a porcentagem da população com diploma de bacharel ($PctBach$) é regredida em termos das seguintes variáveis explicativas: população total em 1990 em cem milhares ($TotPop90$), porcentagem da população definida como rural ($PctRural$), porcentagem da população com 65 anos ou mais ($PctEld$), porcentagem da população nascida fora dos EUA ($PctFB$), porcentagem da população que vive abaixo da linha da pobreza ($PctPov$) e porcentagem da população negra ($PctBlack$). Um aspecto interessante deste base de dados é a presença de valores discrepantes, destacadas no ajuste de modelos normais em análises anteriores. Com isso, o ajuste de modelos supondo distribuições com caudas mais pesadas para os erros aleatórios pode apresentar resultados melhores para descrever o comportamento do percentual da população com diploma de bacharel. No entanto, como a variável resposta representa uma porcentagem, então foi considerado a transformação *logit* descrita da seguinte forma:

$$L.PctBach = \log\left(\frac{PctBach}{100 - PctBach}\right) \quad (3.5)$$

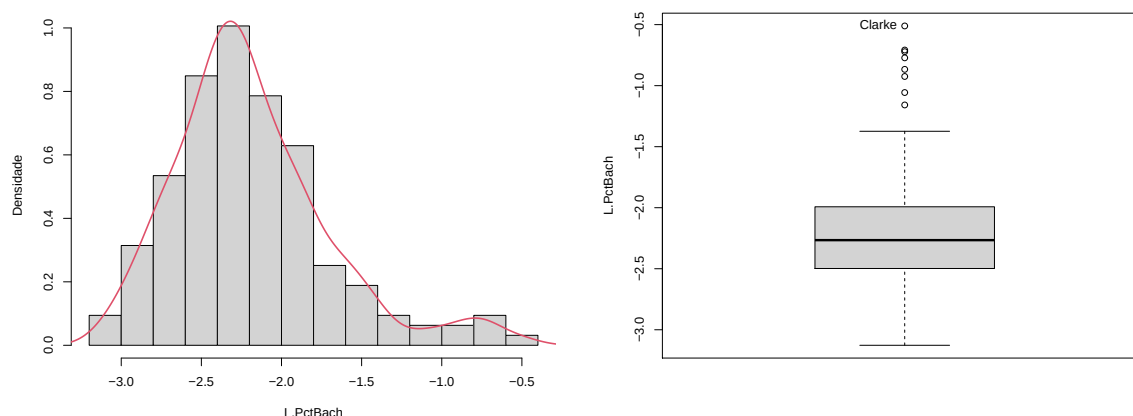
Dessa forma, os demais resultados dessa aplicação são obtidos com base na análise dessa variável transformada, uma vez que os modelos propostos nessa tese são destinados à variáveis respostas contínuas com suporte nos reais. Vale ressaltar que essa transformação não impossibilita a interpretação dos resultados em termos da porcentagem da população com diploma de bacharel.

3.5.1 Análise Exploratória

Consideramos inicialmente uma breve análise exploratória com o intuito de observar os comportamentos individuais e conjuntos das variáveis utilizadas na aplicação. Como forma de verificar o comportamento da variável resposta, a Figura 54 mostra o

histograma e o *boxplot* do *logit* da porcentagem da população com diploma de bacharel. Observar-se a presença de alguns valores discrepantes na distribuição da variável resposta e uma leve assimetria a direita. Nesse caso, o condado de Clarke é destacada por apresentar o maior valor da variável resposta, sendo esse um valor bem discrepante em relação aos demais condados.

Figura 54 – Histograma com densidade aproximada pela linha vermelha (esquerda) e boxplot (direita) do *logit* da porcentagem da população com diploma de bacharel



Fonte: O autor (2021).

Adicionalmente, na Tabela 4 são apresentados alguns valores descritivos para todas as variáveis presentes na base de dados. De modo geral, as variáveis explicativas apresentam comportamentos bem distintos com diferentes medidas descritivas. Além disso, a variável resposta apresenta valores moderados tanto para o coeficiente de variação quanto para a curtose.

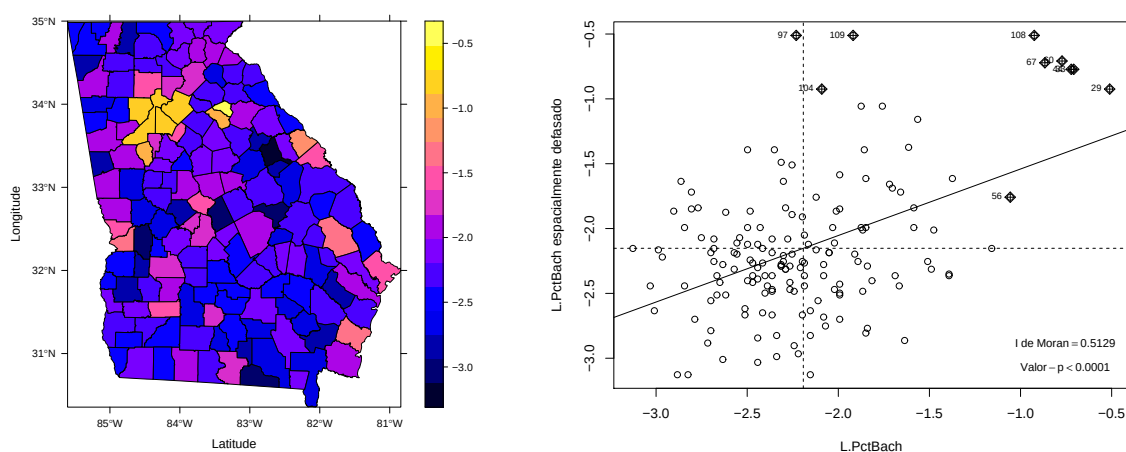
Tabela 4 – Análise descritiva dos indicadores educacionais e sociais nos condados da Geórgia

	Minimo	Mediana	Média	Máximo	C.V.	Assimetria	Curtose
L.PctBach	-3,13	-2,27	-2,19	-0,51	-0,22	0,98	4,32
TotPop90	0,02	0,17	0,41	6,49	2,05	5,00	30,98
PctRural	2,50	72,30	70,18	100,00	0,39	-0,71	2,84
PctEld	1,46	12,07	11,74	22,96	0,26	-0,18	4,09
PctFB	0,04	0,72	1,13	6,74	1,09	2,37	8,99
PctPov	2,60	18,60	19,34	35,90	0,38	0,08	2,31
PctBlack	0,00	27,64	27,39	79,64	0,63	0,24	2,42

Fonte: O autor (2021).

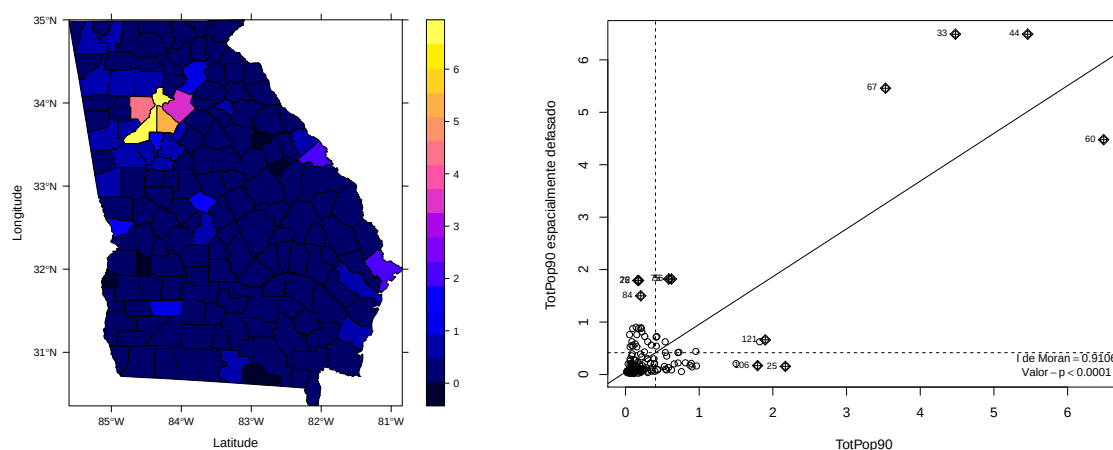
Adicionalmente, para avaliar possíveis tendências espaciais presente nas variáveis estudadas, as Figuras 55-61 mostram as distribuições espaciais e os diagramas de espalhamento de Moran dos indicadores educacionais e sociais nos condados da Geórgia. Destaca-se também os índices de Moran e o valores- p para os testes de autocorrelação espacial nos gráficos apresentados. Observe que as variáveis L.PctBach, TotPop90 e PctFB têm os valores mais altos em alguns condados no norte da Geórgia, mas as variáveis PctRural e PctPov têm os valores mais baixos na mesma região. No entanto, as relações observadas não são equivalentes em todas as regiões da Geórgia, indicando assim que há uma possível heterogeneidade espacial na relação entre a variável respostas e as respectivas variáveis explicativas.

Figura 55 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para o *logit* da porcentagem da população com diploma de bacharel nos condados da Geórgia



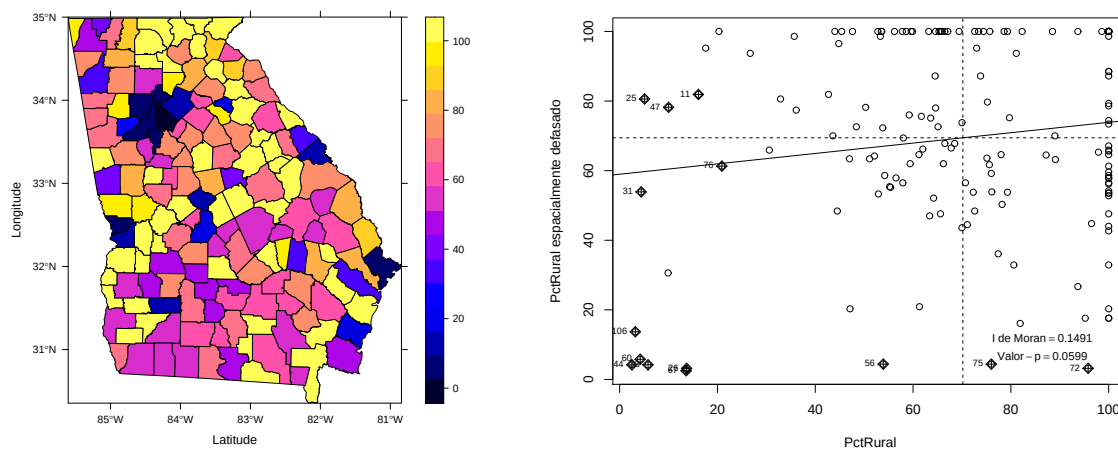
Fonte: O autor (2021).

Figura 56 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para a população total em 1990 nos condados da Geórgia



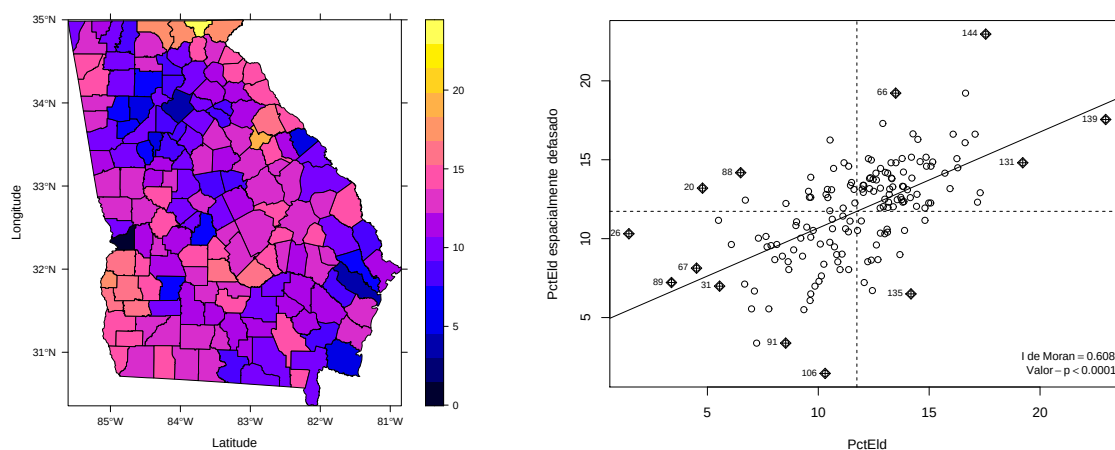
Fonte: O autor (2021).

Figura 57 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para a porcentagem da população rural nos condados da Geórgia



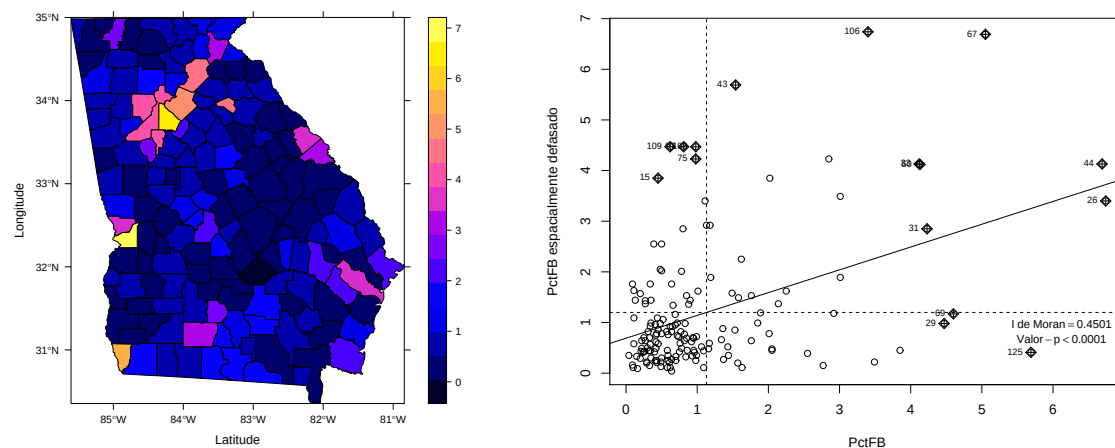
Fonte: O autor (2021).

Figura 58 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para a porcentagem da população com 65 anos ou mais nos condados da Geórgia



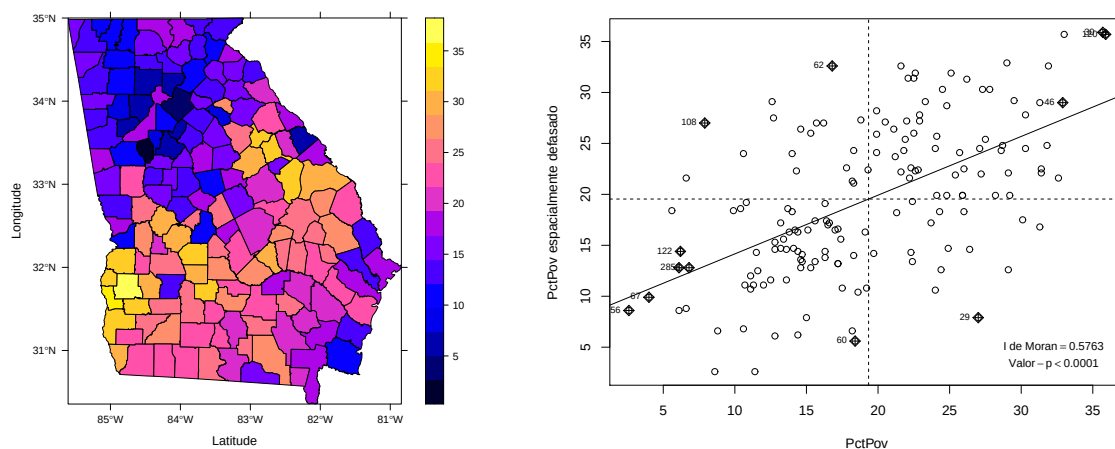
Fonte: O autor (2021).

Figura 59 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para a porcentagem da população nascida fora dos EUA nos condados da Geórgia



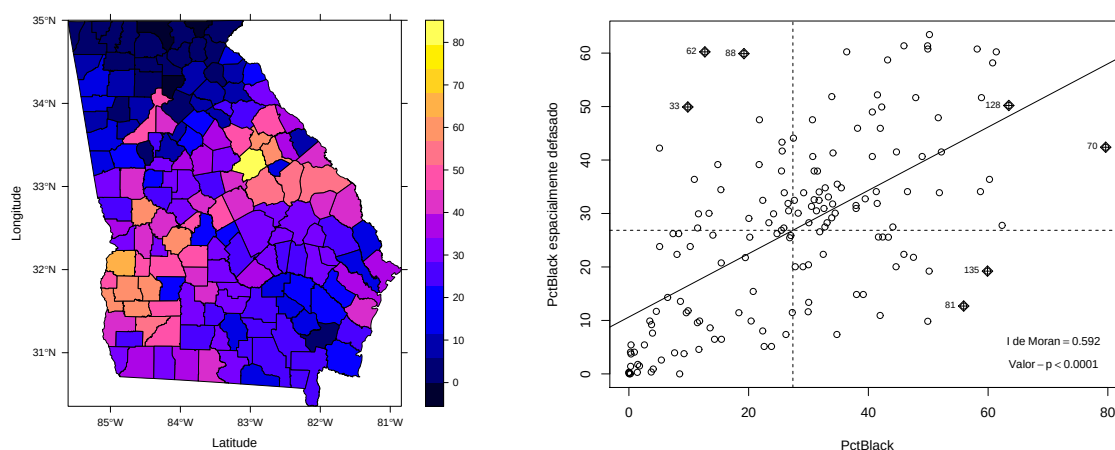
Fonte: O autor (2021).

Figura 60 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para a porcentagem da população que vive abaixo da linha da pobreza nos condados da Geórgia



Fonte: O autor (2021).

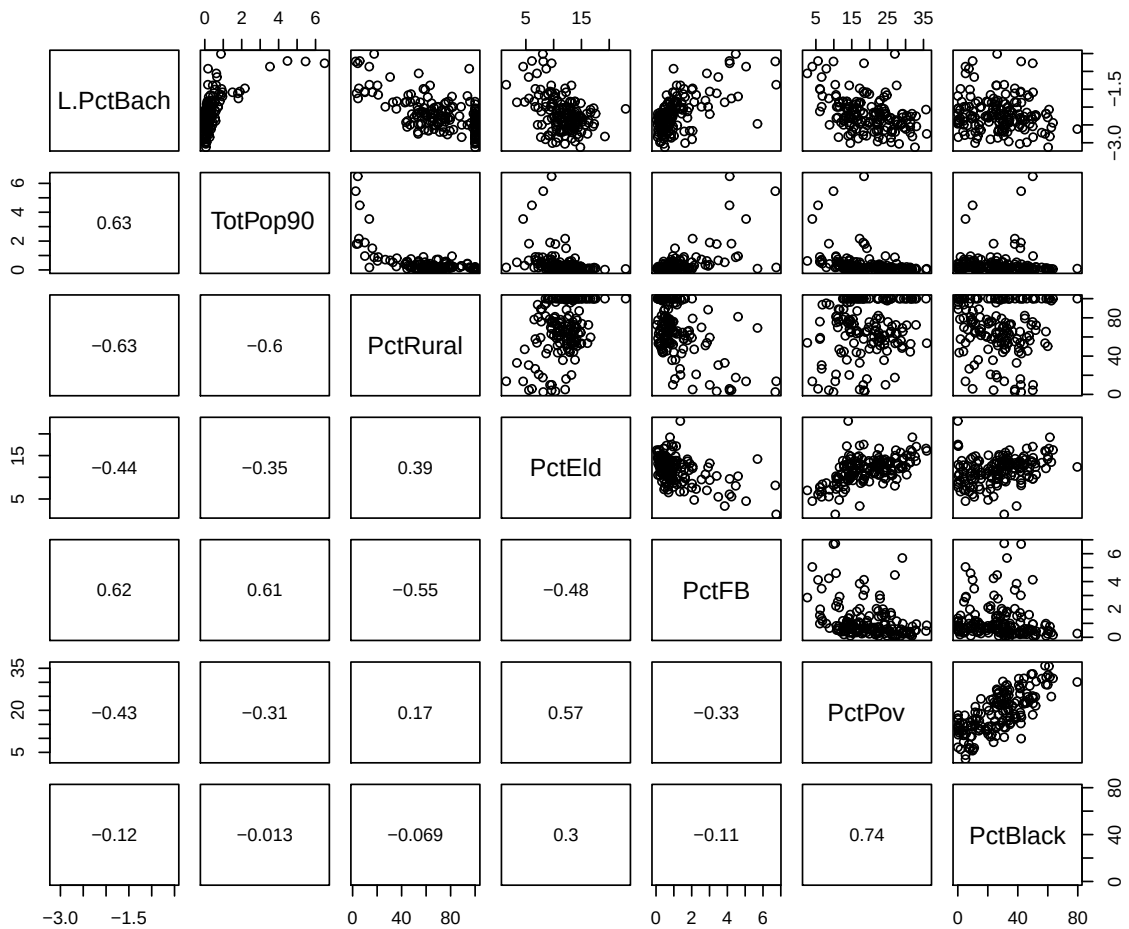
Figura 61 – Distribuição espacial (esquerda) e diagrama de Moran (direita) para a porcentagem da população negra nos condados da Geórgia



Fonte: O autor (2021).

Vale ressaltar que os resultados dos testes mostram autocorrelação espacial significativa e altos valores do índice de Moran para todas as variáveis, exceto para a porcentagem da população definida como rural. Ademais, como forma de investigar melhor possíveis relações entre as variáveis estudadas, a Figura 62 mostra os gráficos de dispersão no painel superior e os respectivos valores para a correlação de Pearson no painel inferior. Os gráficos destacam possíveis relações lineares com correlação moderada entre o *logit* da porcentagem da população com diploma de bacharel e todas as variáveis explicativas, exceto para a porcentagem da população negra. Nota-se também uma forte relação linear entre a porcentagem da população que vive abaixo da linha da pobreza e a porcentagem da população negra, podendo gerar um problema de multicolinearidade.

Figura 62 – Gráficos de dispersão e valores para a correlação de Pearson das variáveis em estudo



Fonte: O autor (2021).

3.5.2 Ajuste do Modelo

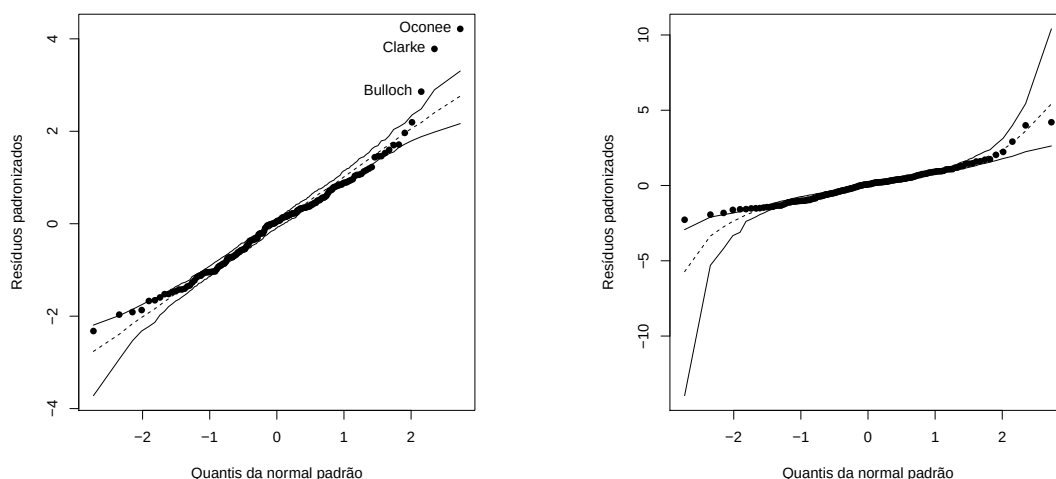
Uma análise prévia dos dados indicou que a porcentagem da população com 65 anos ou mais e a porcentagem da população negra não apresentaram uma contribuição significativa para explicar o *logit* da porcentagem da população com diploma de bacharel. Com isso, propomos inicialmente o ajuste do modelo de regressão elíptico na forma:

$$L.PctBach_i = \beta_0 + \beta_1 TotPop90_i + \beta_2 PctRural_i + \beta_3 PctFB_i + \beta_4 PctPov_i + \epsilon_i,$$

em que $\epsilon_i \sim S(0, \phi)$ representa os erros aleatórios. Primeiramente foi propomos o uso da distribuição normal para os erros aleatórios, no entanto o ajuste desse modelo constatou a presença de valores discrepantes que exerciam forte influência na estimação dos parâmetros (ver Figura 63). Por este motivo, assumimos então o uso da distribuição *t*-Student (com 3 g.l. selecionado via AIC) para os erros aleatórios, a qual resultou em um ajuste menos

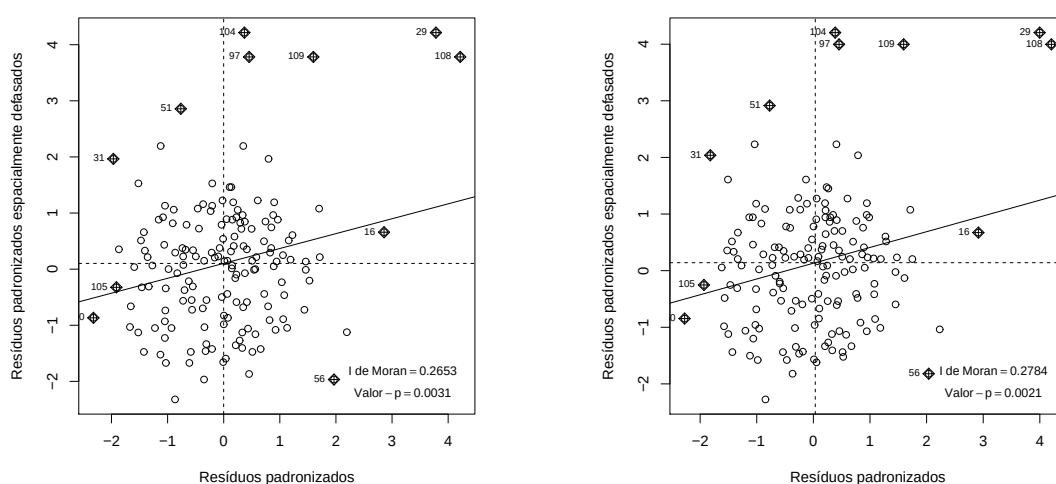
sensível a observações discrepantes. Essa seleção foi realizada de maneira equivalente ao capítulo anterior, considerando apenas uma grade de valores inteiros (de 3 até 30) para os graus de liberdade.

Figura 63 – Gráfico de quantis-quantis com envelope simulado dos resíduos padronizados para o modelo de regressão elíptico com erros normais (esquerda) e *t*-Student (direita)



Fonte: O autor (2021).

Figura 64 – Diagramas de espalhamento de Moran dos resíduos padronizados para o modelo de regressão elíptico com erros normais (esquerda) e *t*-Student (direita)



Fonte: O autor (2021).

Vale ressaltar que, considerou-se também o ajuste dos modelos elípticos com distribuições Cauchy, exponencial potência e logísticas do tipo I e II, no entanto o modelo

com erros t -Student apresentou os melhores valores de AICc. Observou-se que todos os modelos ajustados apresentaram uma forte e significativa autocorrelação espacial nos resíduos padronizados, como podemos observar na Figura 64 através dos diagramas de espalhamento de Moran. Nesse caso, o uso da REGP pode ser uma melhor alternativa para medir a relação entre o *logit* da porcentagem da população com diploma de bacharel e as variáveis explicativas, pois permite a variação espacial dos parâmetros de regressão. Desse modo, foi ajustado modelos de REGP através da seguinte forma funcional:

$$\begin{aligned} \text{L.PctBach}_i = & \beta(i)_0 + \beta(i)_1 \text{TotPop90}_i + \beta(i)_2 \text{PctRural}_i + \beta(i)_3 \text{PctFB}_i \\ & + \beta(i)_4 \text{PctPov}_i + \epsilon_i, \end{aligned}$$

em que $\epsilon_i \sim S(0, \phi(i))$ representa os erros aleatórios. Novamente, consideramos o uso das distribuições normal e t -Student com 3 g.l. para os erros aleatórios. No processo de estimação foi utilizado um *kernel* biquadrado adaptativo para as funções de ponderação, o critério da validação cruzada para seleção do parâmetro de suavização ótimo e o AICc para escolha dos graus de liberdade do modelo t -Student. No processo de seleção desse modelo, considerou-se novamente uma grade de valores inteiros (de 3 até 30) para a seleção dos graus de liberdade e mesmas distribuições mencionadas anteriormente para os erros aleatórios. Feito isso, aplicamos os testes de não estacionariedade espacial aleatorizados para os parâmetros de ambos modelos de REGP ajustados. Na Tabela 5 são exibidos os valores- p desses testes para cada parâmetro de regressão, destacando que apenas as estimativas de β_1 e β_3 aparentam não ser estacionárias no espaço.

Tabela 5 – Valores- p para os testes de variabilidade espacial dos parâmetros da REGP

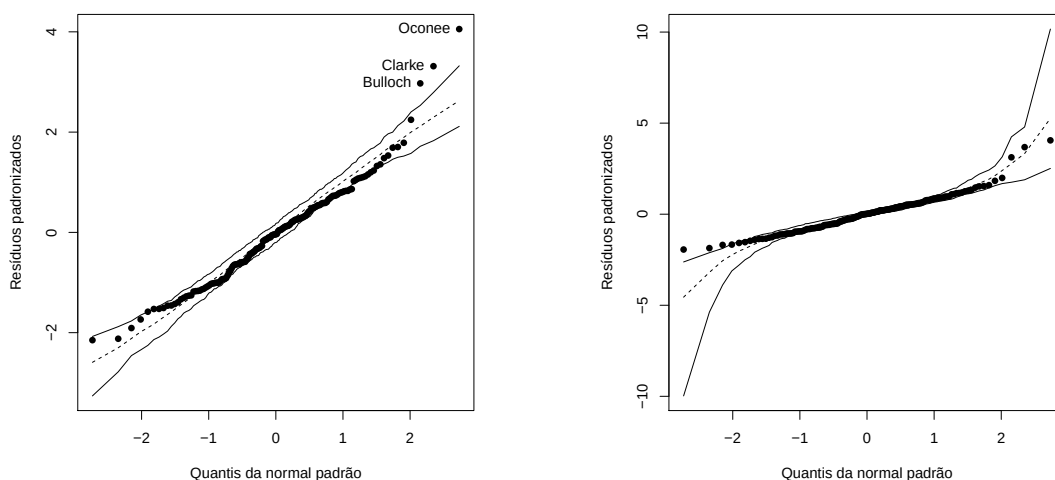
	REGP com erros normais	REGP com erros t -Student
$\hat{\beta}_0$	0,0720	0,4620
$\hat{\beta}_1$	0,0110	0,0020
$\hat{\beta}_2$	0,0500	0,2610
$\hat{\beta}_3$	0,0000	0,0550
$\hat{\beta}_4$	0,7970	0,7530

Fonte: O autor (2021).

Além disso, constatou-se mais uma vez que o ajuste do modelo de REGP com erros normais apresenta observações discrepantes destacadas na análise de resíduos (ver Figura 65). Por outro lado, o ajuste do modelo de REGP com erros t -Student obteve um

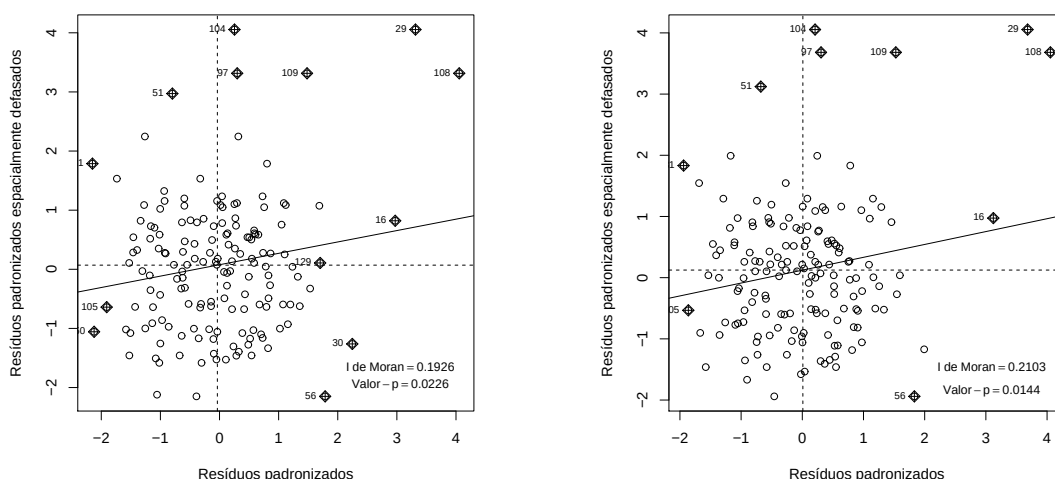
ajuste menos sensível à presença de observações discrepantes. Nota-se também que ambos modelos ainda exibem resíduos com forte dependência espacial como podemos observar na Figura 66, destacando tendências nos primeiros e terceiros quadrantes semelhante.

Figura 65 – Gráfico de quantis-quantis com envelope simulado dos resíduos padronizados para o modelo de regressão elíptico com distribuições normal (esquerda) e t -Student (direita)



Fonte: O autor (2021).

Figura 66 – Diagramas de espalhamento de Moran dos resíduos padronizados para o modelo de regressão elíptico com distribuições normal (esquerda) e t -Student (direita)



Fonte: O autor (2021).

Observa-se nesse caso que, apesar de haver uma pequena redução desses valores em comparação com os modelos de regressão globais, a autocorrelação espacial de ambos os resíduos ainda é considerada significativa. Além disso, como foi ressaltado que apenas

os parâmetros β_1 e β_3 apresentam variação no espaço e a análise de resíduos indicou presença de autocorrelação espacial, então há indícios que a REGP não apresenta um ajuste adequado. Por este fato, uma alternativa é considera a aplicação da REGPM, visto que cada parâmetro pode ter sua própria escala de variação espacial. Nesse contexto, foi proposto o ajuste dos modelos na seguinte forma:

$$\begin{aligned} \text{L.PctBach}_i = & \beta(i)_{b_0} + \beta(i)_{b_1} \text{TotPop90}_i + \beta(i)_{b_2} \text{PctRural}_i + \beta(i)_{b_3} \text{PctFB}_i \\ & + \beta(i)_{b_4} \text{PctPov}_i + \epsilon_i, \end{aligned}$$

em que $\epsilon_i \sim S(0, \phi(i))$ representa os erros aleatórios. Nesse caso, assumimos as distribuições normal e t -Student com 3 g.l. para os erros aleatórios. Mais uma vez assumimos a grade de valores inteiros (de 3 até 30) para a seleção dos graus de liberdade e mesmas distribuições mencionadas nos ajustes anteriores, obtendo como resultado que o modelo REGPM com erros t -Student apresentou o melhor valor para o AICc. Além disso, foi utilizado um *kernel* biquadrado adaptativo e a distância Euclidiana para as funções de ponderação e o critério da validação cruzada para seleção do parâmetro de suavização ótima. As Tabelas 6 e 7 apresentam uma análise descritiva dos parâmetros estimados e erros padrões assintóticos (entre parênteses) para os modelos de REGPM ajustados.

Tabela 6 – Análise descritiva das estimativas dos parâmetros e erros padrões assintóticos (entre parênteses) para os modelos de REGPM com erros normais

	Minimo	Mediana	Média	Máximo	Global
$\hat{\beta}_0$	-1,721 (0,164)	-1,505 (0,189)	-1,501 (0,189)	-1,312 (0,207)	-1,624 (0,135)
$\hat{\beta}_1$	0,073 (0,049)	0,077 (0,050)	0,109 (0,060)	0,213 (0,092)	0,123 (0,041)
$\hat{\beta}_2$	-0,007 (0,001)	-0,007 (0,001)	-0,007 (0,001)	-0,007 (0,001)	-0,006 (0,001)
$\hat{\beta}_3$	0,042 (0,031)	0,105 (0,036)	0,097 (0,035)	0,141 (0,038)	0,093 (0,027)
$\hat{\beta}_4$	-0,019 (0,005)	-0,018 (0,005)	-0,018 (0,005)	-0,018 (0,005)	-0,015 (0,004)
$\hat{\phi}$	0,089 (0,013)	0,092 (0,014)	0,092 (0,014)	0,096 (0,016)	0,096 (0,011)
AIC = 1320,6790		AICc = 1488,4680		BIC = 1382,2580	

Fonte: O autor (2021).

Tabela 7 – Análise descritiva das estimativas dos parâmetros e erros padrões assintóticos (entre parênteses) para os modelos de REGPM com erros t -Student

	Minimo	Mediana	Média	Máximo	Global
$\hat{\beta}_0$	-1,648 (0,212)	-1,525 (0,232)	-1,535 (0,231)	-1,452 (0,247)	-1,618 (0,127)
$\hat{\beta}_1$	0,108 (0,059)	0,114 (0,064)	0,151 (0,086)	0,310 (0,148)	0,134 (0,039)
$\hat{\beta}_2$	-0,006 (0,002)	-0,006 (0,002)	-0,006 (0,002)	-0,006 (0,002)	-0,006 (0,001)
$\hat{\beta}_3$	0,044 (0,039)	0,077 (0,044)	0,074 (0,044)	0,097 (0,048)	0,078 (0,025)
$\hat{\beta}_4$	-0,019 (0,006)	-0,019 (0,006)	-0,019 (0,006)	-0,019 (0,007)	-0,016 (0,003)
$\hat{\phi}$	0,040 (0,013)	0,049 (0,014)	0,048 (0,014)	0,056 (0,017)	0,068 (0,009)
AIC = 969,7798	AICc = 1133,3960			BIC = 1007,4290	

Fonte: O autor (2021).

Observe que, apenas as estimativas dos parâmetros β_0 , β_1 e β_3 apresentam variações em torno das medidas globais em ambos modelos. Note também que, de acordo com os critérios de informação Akaike (AIC), Akaike corrigido (AICc) e Bayesian (BIC), a REGPM com distribuição t -Student é selecionado como o melhor ajuste. Adicionalmente, a Tabela 8 compara os parâmetros de suavização selecionadas para cada parâmetro de regressão no ajuste dos modelos geograficamente ponderados multiescalares. Nota-se que, os parâmetros de suavização selecionadas para os modelos REGPM indicam que as estimativas dos parâmetros β_2 e β_4 são constantes no espaço, enquanto que as estimativas dos parâmetros β_0 , β_1 e β_3 destacam não estacionariedade espacial. Isso se dá pelo fato de que um parâmetro de suavização muito próximo da quantidade de observações resulta em parâmetros de regressão com pouca variabilidade espacial.

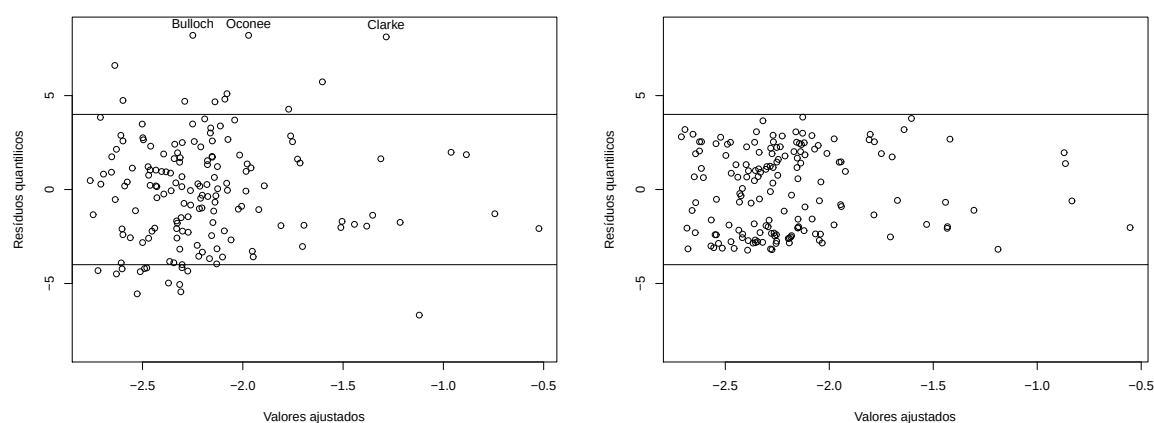
Tabela 8 – Valores dos parâmetros de suavização ótimos selecionadas para os modelos geograficamente ponderados

	REGPM com erros normais	REGPM com erros t -Student
$\hat{\beta}_0$	31	67
$\hat{\beta}_1$	124	109
$\hat{\beta}_2$	157	157
$\hat{\beta}_3$	118	142
$\hat{\beta}_4$	157	157

Fonte: O autor (2021).

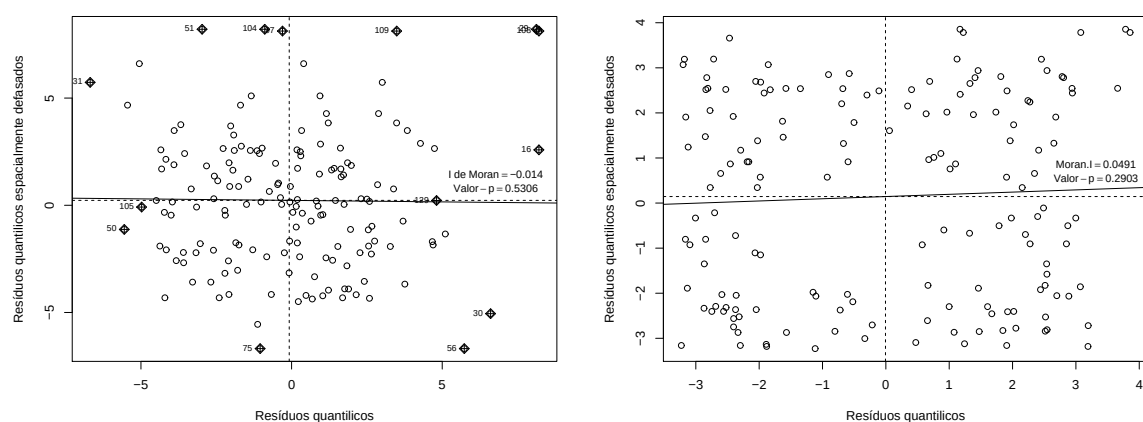
Na Figura 67 são apresentados os gráficos dos resíduos quantílicos versus os valores ajustados para ambos modelos de REGPM, destacando os condados de Bulloch, Clarke e Oconee como observações discrepantes no modelo com erros normais. Além disso, na Figura 68 são exibidos os diagramas de espalhamento de Moran dos resíduos quantílicos para os modelos de REGPM, bem como os respectivos índices de Moran e valores-p para os testes de autocorrelação espacial. Os resultados obtidos indicam que ambos os modelos de REGPM não apresentam autocorrelação espacial nos resíduos.

Figura 67 – Resíduos padronizados versus valores ajustados para o modelo de REGPM com distribuições normal (esquerda) e *t*-Student (direita)



Fonte: O autor (2021).

Figura 68 – Diagramas de espalhamento de Moran dos resíduos quantílicos para o modelo de REGPM com distribuições normal (esquerda) e *t*-Student (direita)

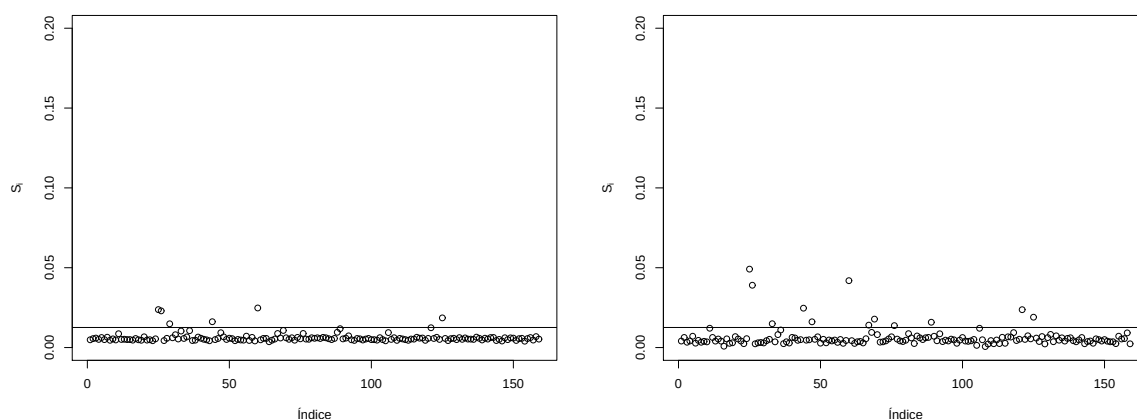


Fonte: O autor (2021).

3.5.3 Análise de Sensibilidade

Para avaliar a qualidade do ajuste e a sensibilidade dos modelos de REGPM, foram aplicadas algumas das medidas de diagnóstico propostas no presente capítulo. Na Figura 69 são exibidos os gráficos das medidas da alavancagem baseadas nas matrizes de projeção para os modelos de REGPM ajustados. Em ambos os casos nenhuma observação é destaca como um possível ponto de alavanca, uma vez que todas observações apresentam valores bem próximos pra essa medida.

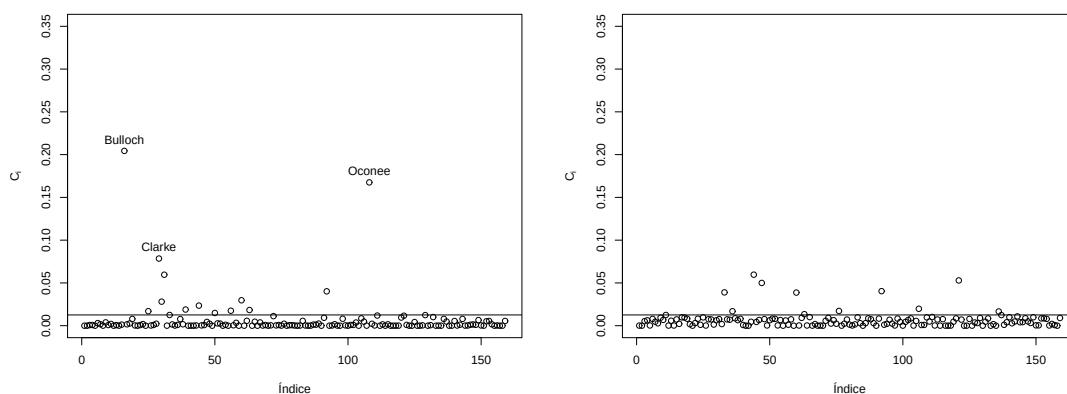
Figura 69 – Gráfico das medidas de alavancagem nos modelos de REGP com erros normais (esquerda) e t -Student (direita)



Fonte: O autor (2021).

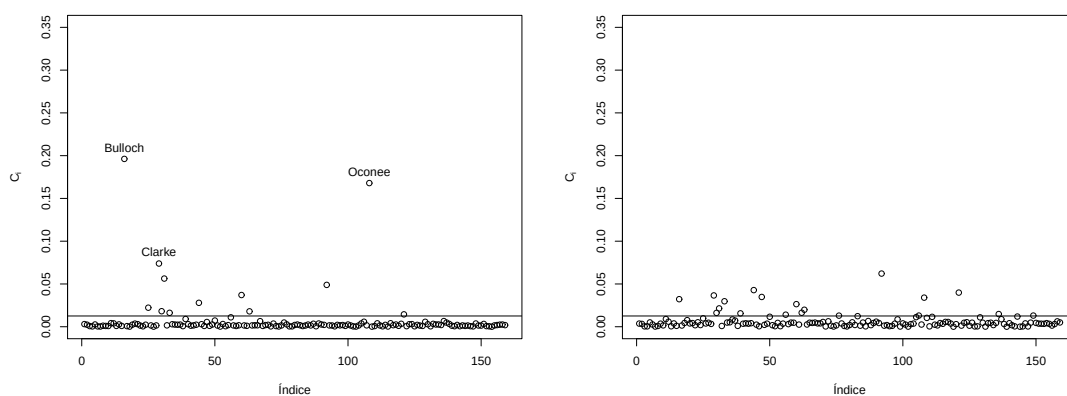
Adicionalmente, nas Figuras 70-77 são apresentados os gráficos das medidas de influência local sob as perturbações na escala e ponderação de casos associado a estimação de cada parâmetro dos modelos REGPM. Note que, os condados de Bulloch, Clarke e Oconee são destacados como possíveis observações influentes na estimação dos parâmetros no modelo com erros normais em ambos esquemas de perturbação. Esses condados estão situados na região nordeste da Geórgia, os quais apresentam altos valores para a porcentagem da população com diploma de bacharel. Por outro lado, no modelo REGPM com erros t -Student nenhuma observação é destacada como possivelmente influente.

Figura 70 – Gráfico de influência local para perturbação na escala associado ao parâmetro β_1 nos modelos REGP com erros normais (esquerda) e t -Student (direita)



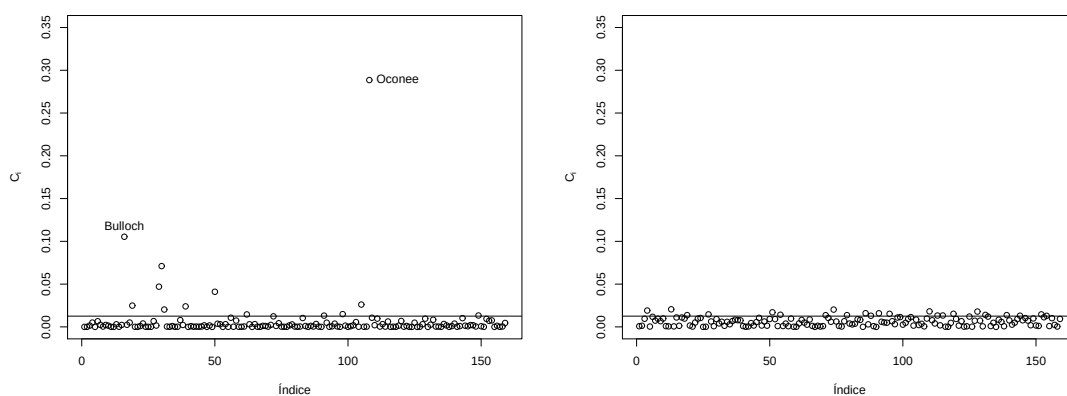
Fonte: O autor (2021).

Figura 71 – Gráfico de influência local para perturbação ponderação de casos associado ao parâmetro β_1 nos modelos REGP com erros normais (esquerda) e t -Student (direita)



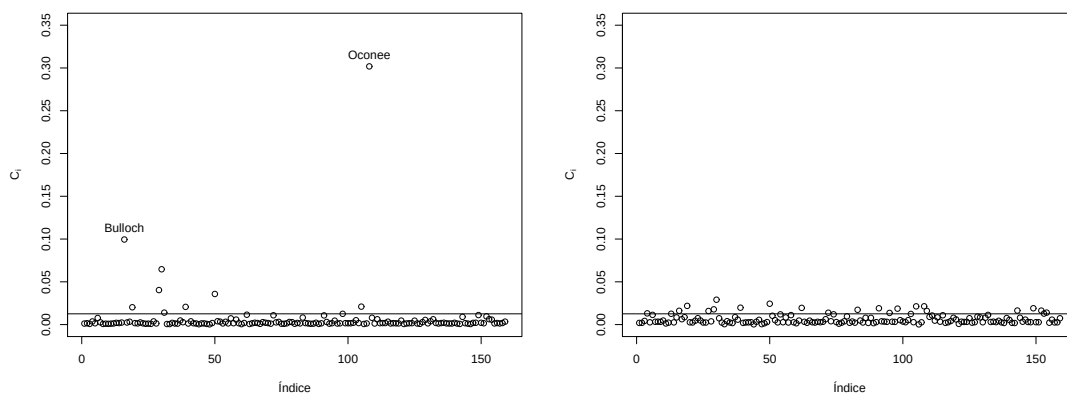
Fonte: O autor (2021).

Figura 72 – Gráfico de influência local para perturbação na escala associado ao parâmetro β_2 nos modelos REGP com erros normais (esquerda) e t -Student (direita)



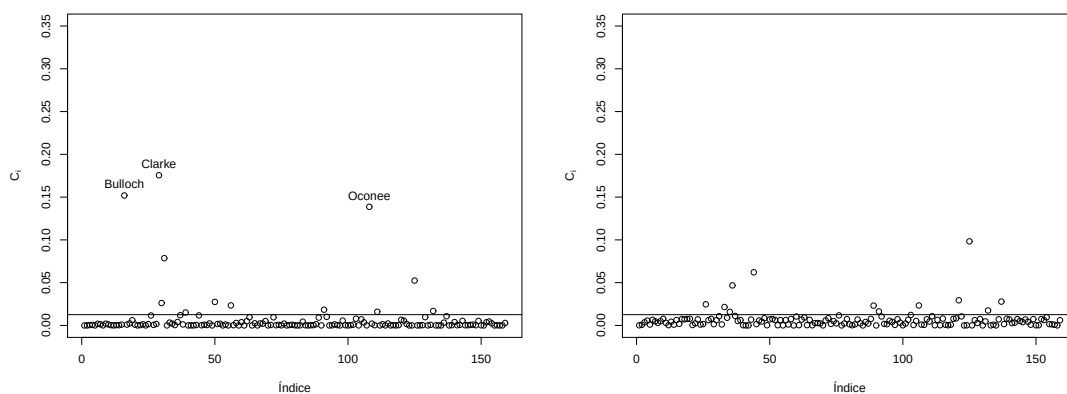
Fonte: O autor (2021).

Figura 73 – Gráfico de influência local para perturbação ponderação de casos associado ao parâmetro β_2 nos modelos REGP com erros normais (esquerda) e t -Student (direita)



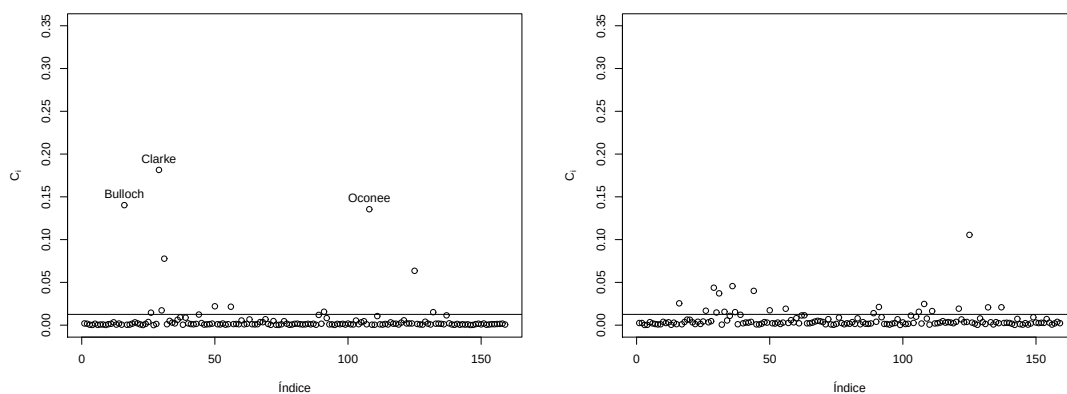
Fonte: O autor (2021).

Figura 74 – Gráfico de influência local para perturbação na escala associado ao parâmetro β_3 nos modelos REGP com erros normais (esquerda) e t -Student (direita)



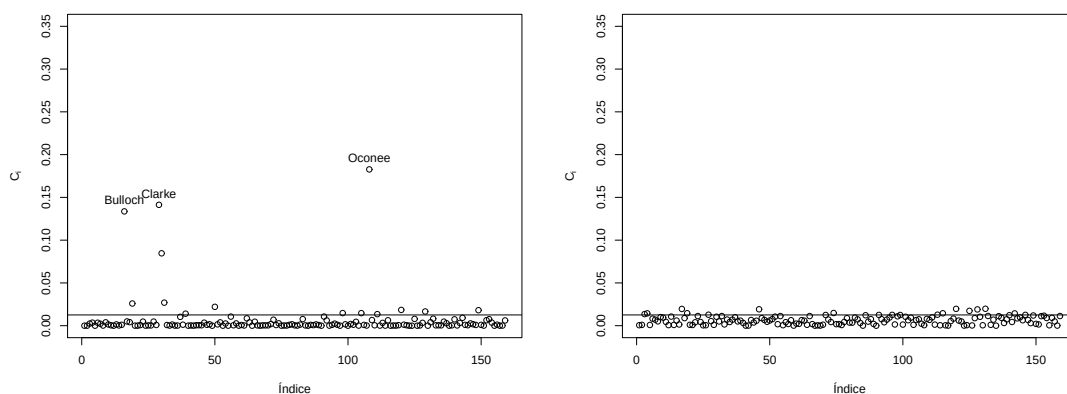
Fonte: O autor (2021).

Figura 75 – Gráfico de influência local para perturbação ponderação de casos associado ao parâmetro β_3 nos modelos REGP com erros normais (esquerda) e t -Student (direita)



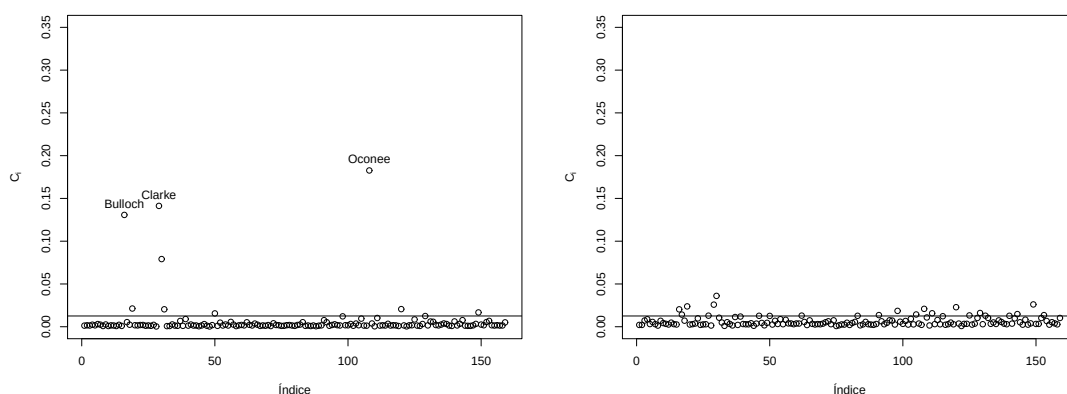
Fonte: O autor (2021).

Figura 76 – Gráfico de influência local para perturbação na escala associado ao parâmetro β_4 nos modelos REGP com erros normais (esquerda) e t -Student (direita)



Fonte: O autor (2021).

Figura 77 – Gráfico de influência local para perturbação ponderação de casos associado ao parâmetro β_4 nos modelos REGP com erros normais (esquerda) e t -Student (direita)

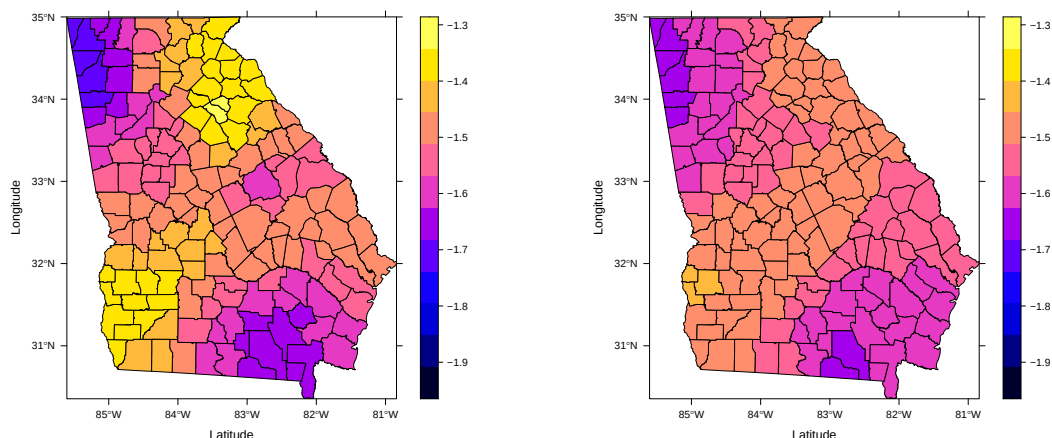


Fonte: O autor (2021).

3.5.4 Análise Confirmatória

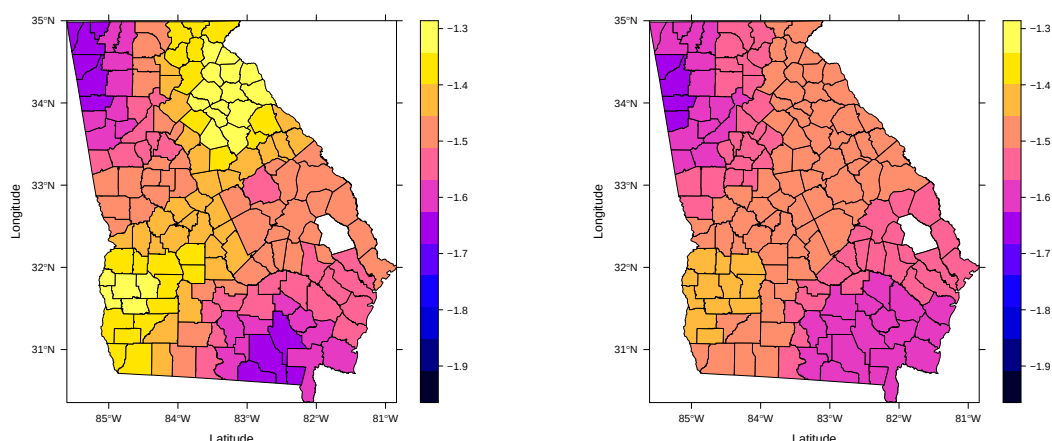
Como foi destacado na análise de sensibilidade, os métodos de diagnósticos para a REGPM apontaram uma forte influência da observação referente aos condados de Bulloch, Clarke e Oconee no modelo com erros normais. Nesse sentido, para verificar se estes condados são de fato observações influentes, considerou o reajuste dos modelos de REGPM excluindo as observações destacadas da base de dados. Os resultados ao reajustar os modelos foram comparados diretamente com os obtidos anteriormente com os dados completos. Visto isso, as Figuras 78-97 mostram a comparação entre as distribuições espaciais das estimativas dos parâmetros de regressão para os modelos de REGPM excluindo-se, individualmente, os condado de Bulloch, Clarke e Ocoone.

Figura 78 – Distribuição espacial das estimativas de β_0 considerando os dados completos para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



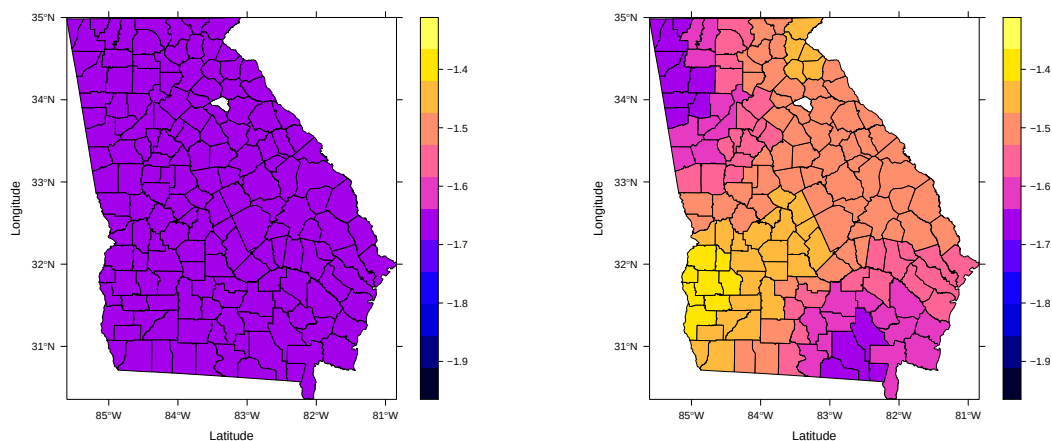
Fonte: O autor (2021).

Figura 79 – Distribuição espacial das estimativas de β_0 considerando os dados após a exclusão de Bulloch para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



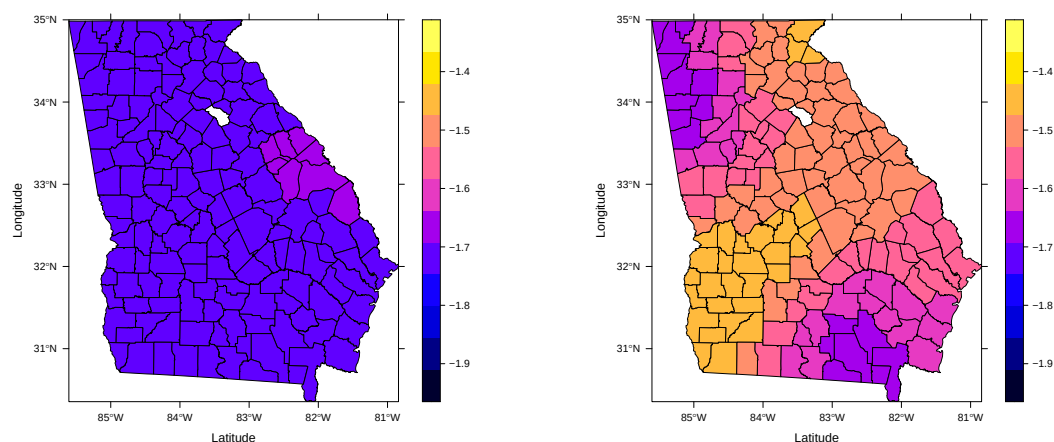
Fonte: O autor (2021).

Figura 80 – Distribuição espacial das estimativas de β_0 considerando os dados após a exclusão de Clarke para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



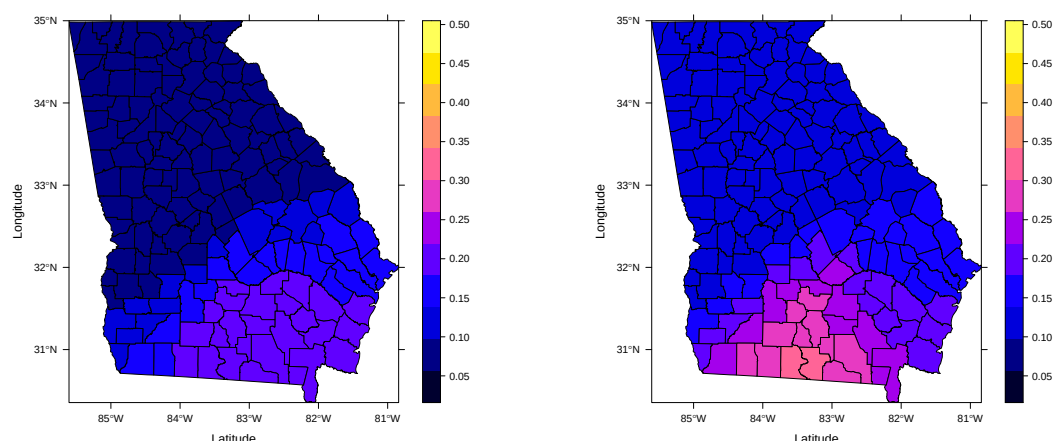
Fonte: O autor (2021).

Figura 81 – Distribuição espacial das estimativas de β_0 considerando os dados após a exclusão de Oconee para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



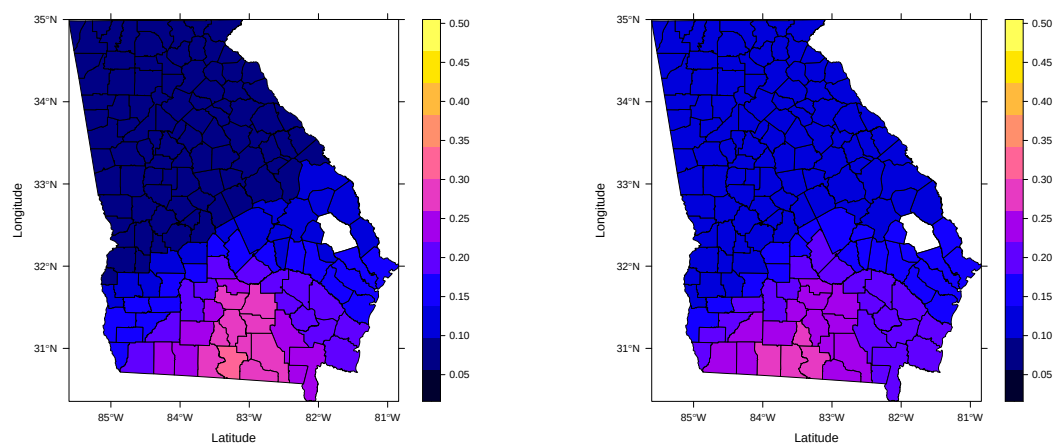
Fonte: O autor (2021).

Figura 82 – Distribuição espacial das estimativas de β_1 considerando os dados completos para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



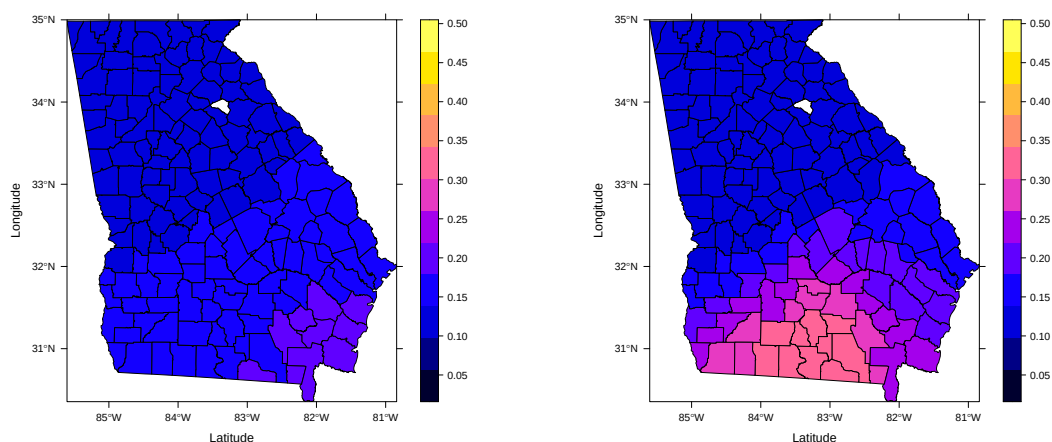
Fonte: O autor (2021).

Figura 83 – Distribuição espacial das estimativas de β_1 considerando os dados após a exclusão de Bulloch para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



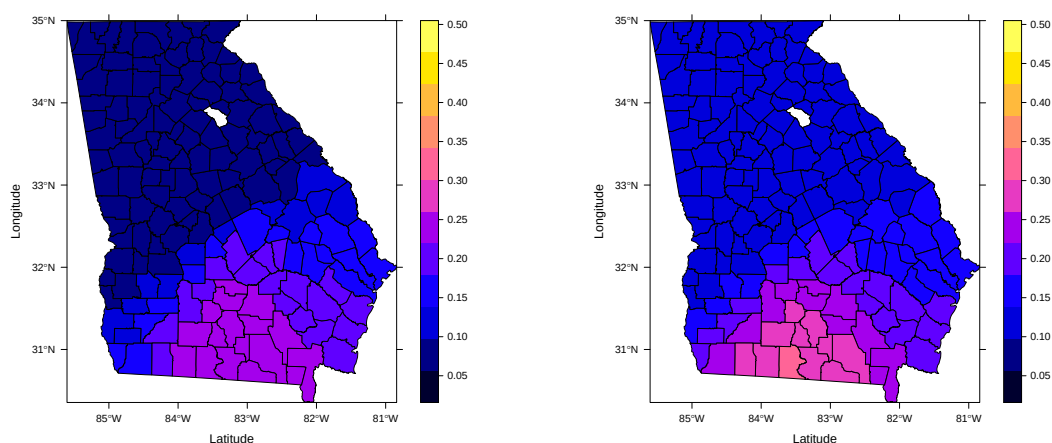
Fonte: O autor (2021).

Figura 84 – Distribuição espacial das estimativas de β_1 considerando os dados após a exclusão de Clarke para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



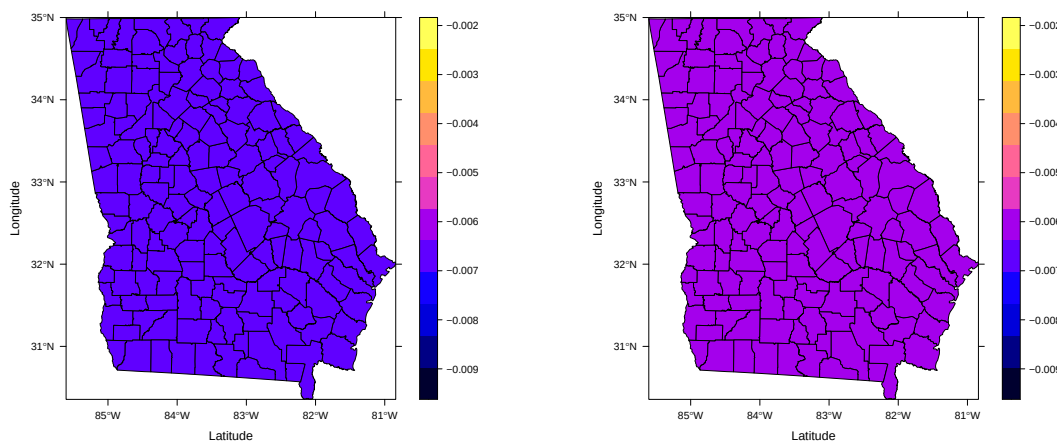
Fonte: O autor (2021).

Figura 85 – Distribuição espacial das estimativas de β_1 considerando os dados após a exclusão de Oconee para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



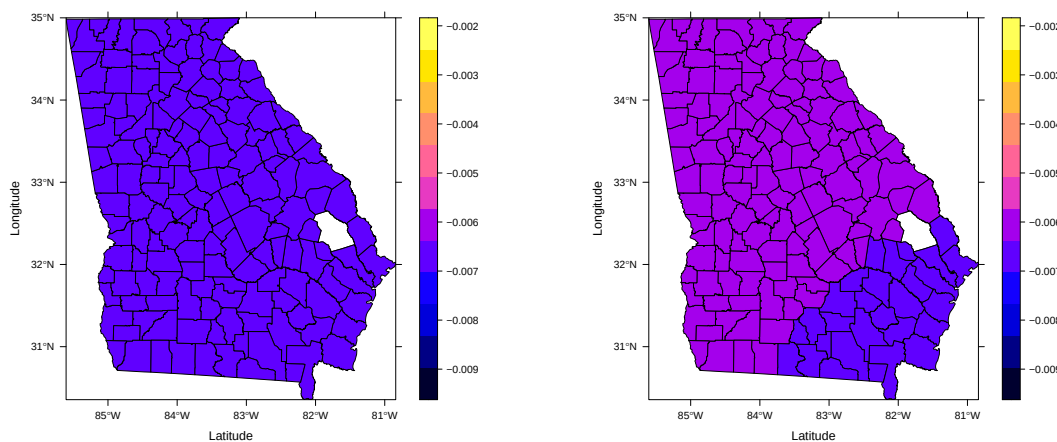
Fonte: O autor (2021).

Figura 86 – Distribuição espacial das estimativas de β_2 considerando os dados completos para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



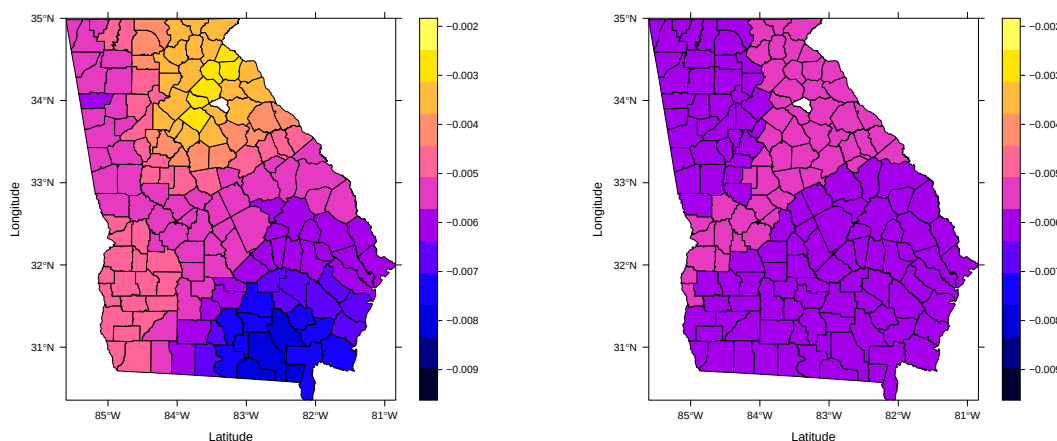
Fonte: O autor (2021).

Figura 87 – Distribuição espacial das estimativas de β_2 considerando os dados após a exclusão de Bulloch para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



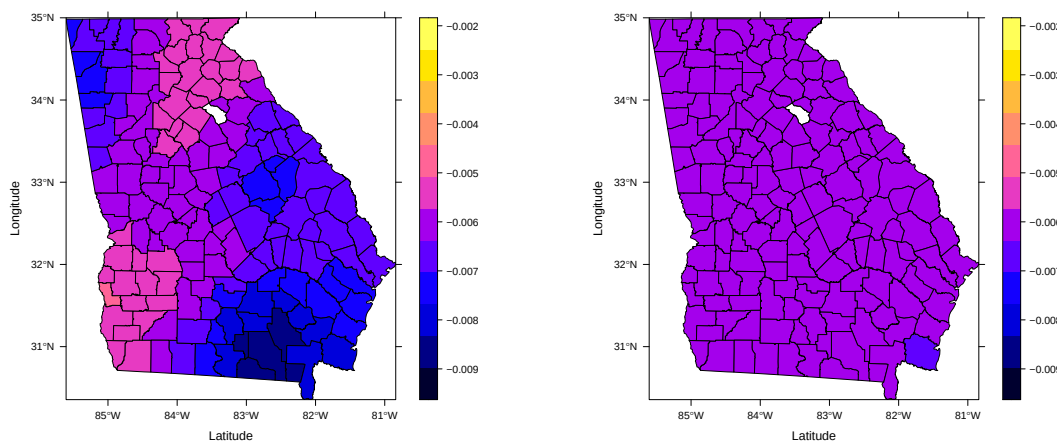
Fonte: O autor (2021).

Figura 88 – Distribuição espacial das estimativas de β_2 considerando os dados após a exclusão de Clarke para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



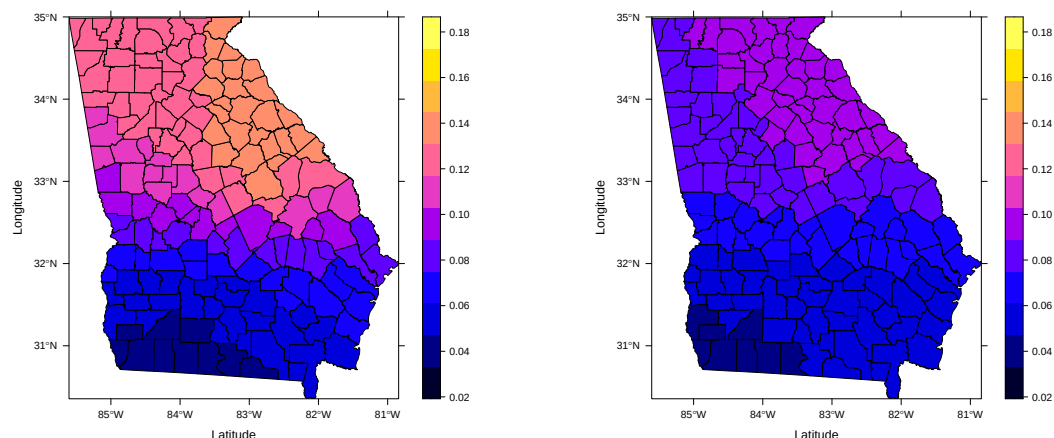
Fonte: O autor (2021).

Figura 89 – Distribuição espacial das estimativas de β_2 considerando os dados após a exclusão de Oconee para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



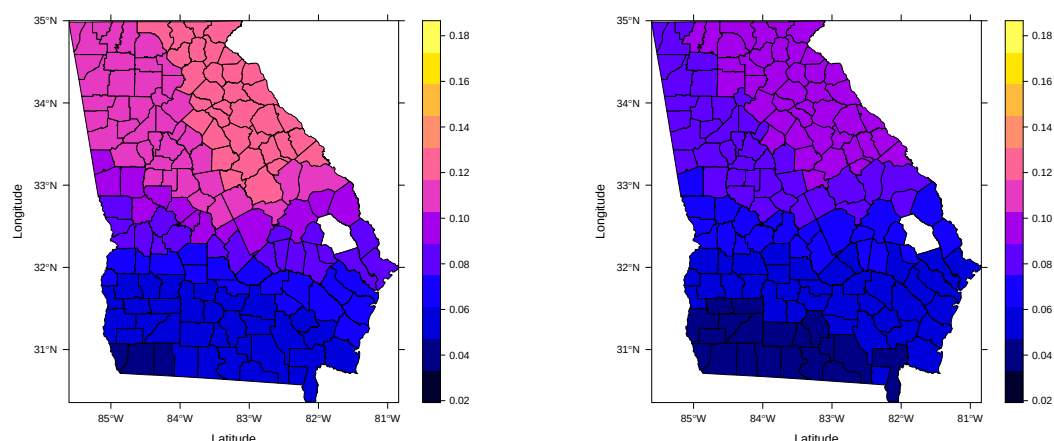
Fonte: O autor (2021).

Figura 90 – Distribuição espacial das estimativas de β_3 considerando os dados completos para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



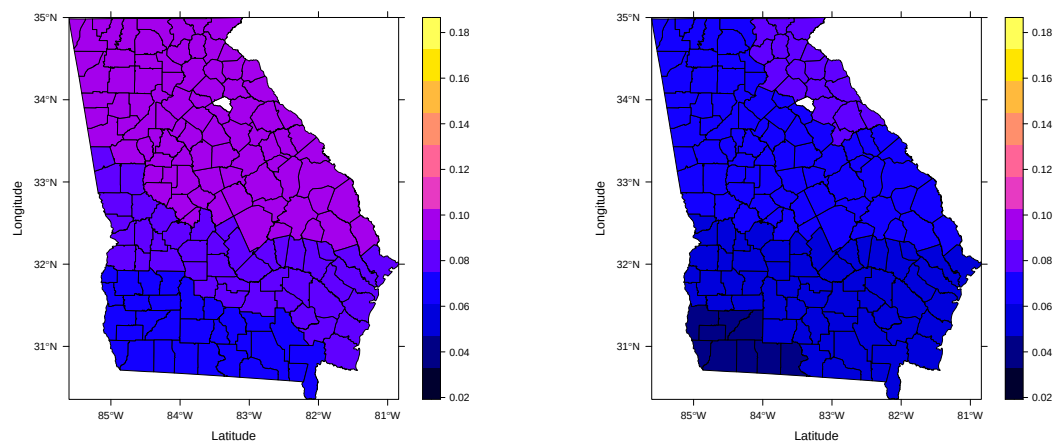
Fonte: O autor (2021).

Figura 91 – Distribuição espacial das estimativas de β_3 considerando os dados após a exclusão de Bulloch para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



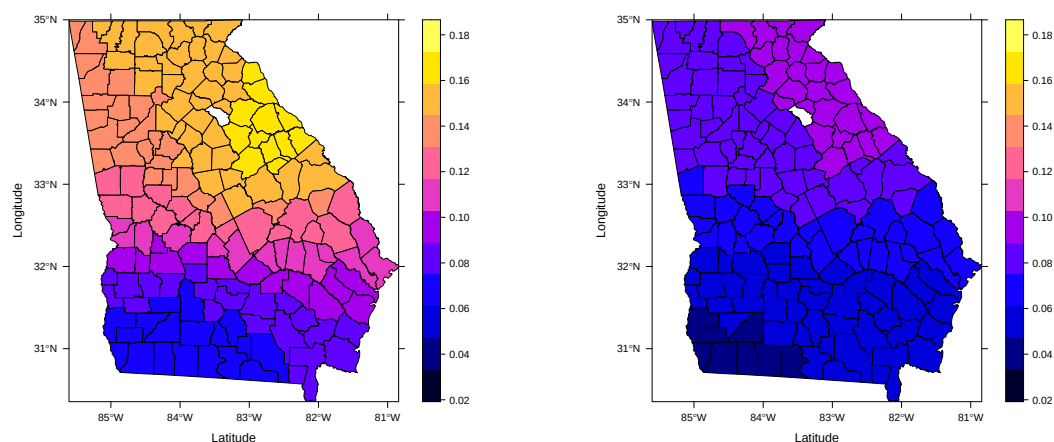
Fonte: O autor (2021).

Figura 92 – Distribuição espacial das estimativas de β_3 considerando os dados após a exclusão de Clarke para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



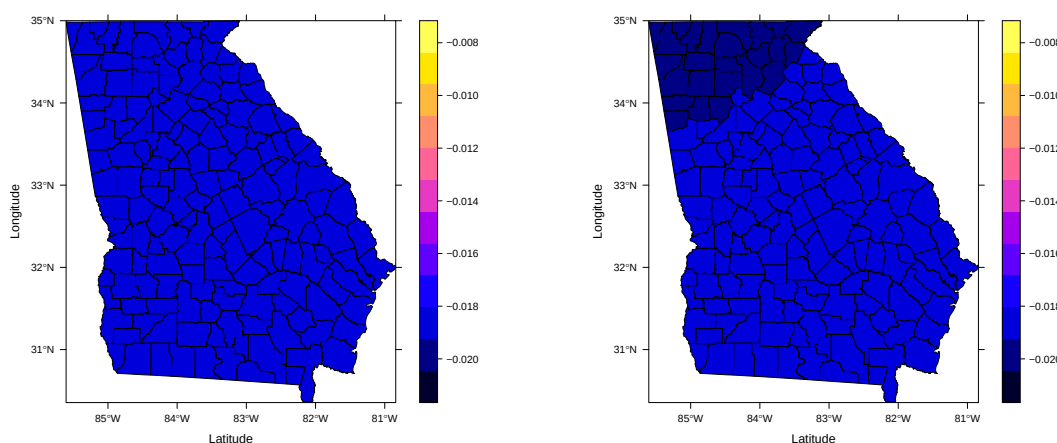
Fonte: O autor (2021).

Figura 93 – Distribuição espacial das estimativas de β_3 considerando os dados após a exclusão de Oconee para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



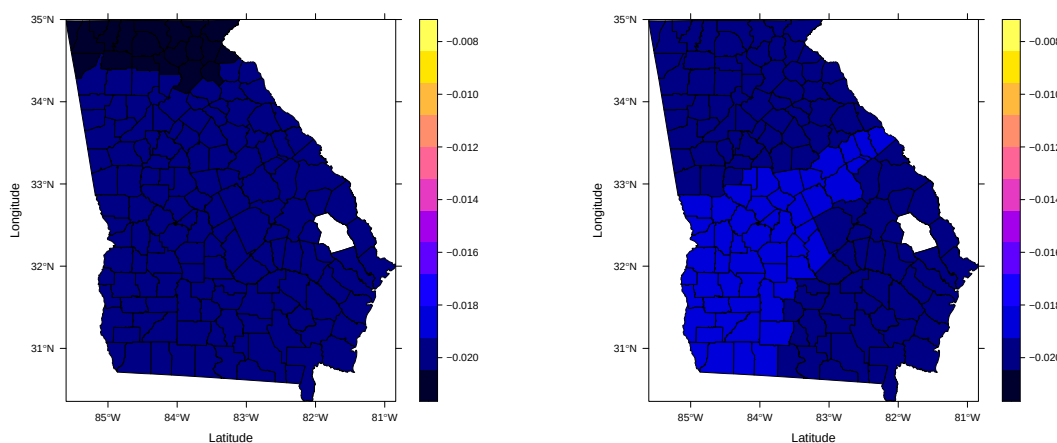
Fonte: O autor (2021).

Figura 94 – Distribuição espacial das estimativas de β_4 considerando os dados completos para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



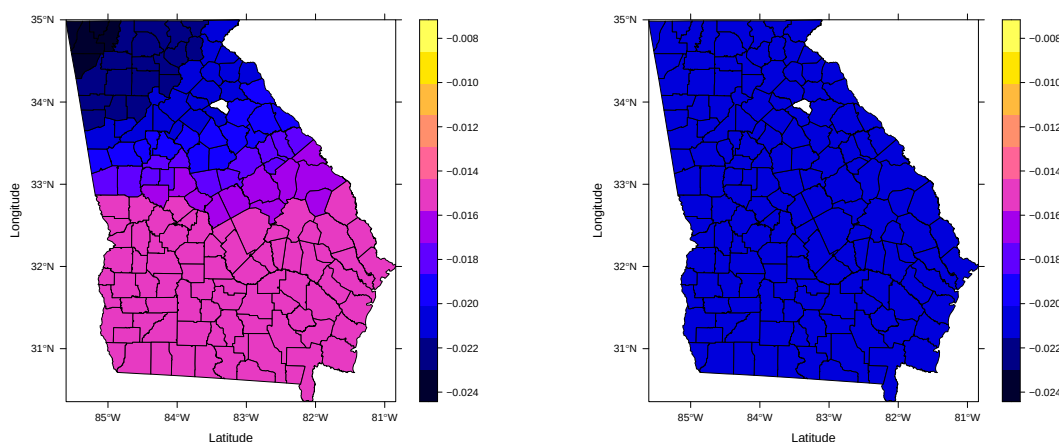
Fonte: O autor (2021).

Figura 95 – Distribuição espacial das estimativas de β_4 considerando os dados após a exclusão de Bulloch para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



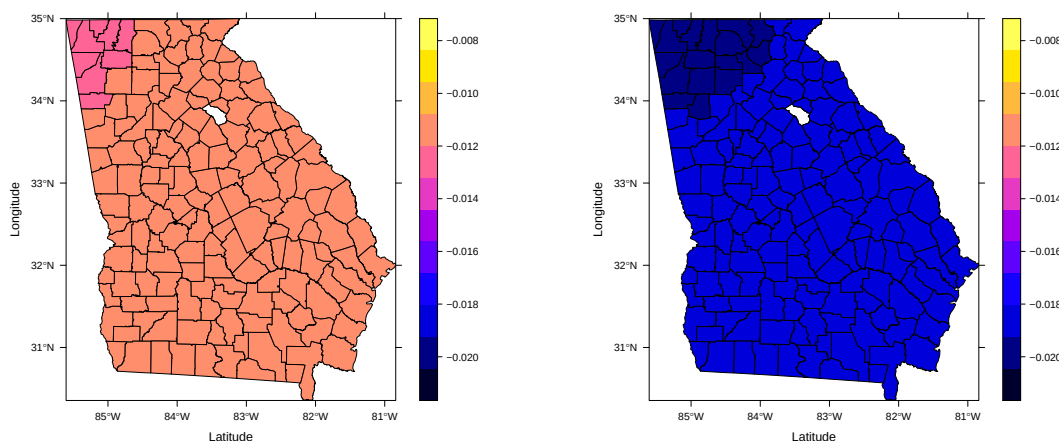
Fonte: O autor (2021).

Figura 96 – Distribuição espacial das estimativas de β_4 considerando os dados após a exclusão de Clarke para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



Fonte: O autor (2021).

Figura 97 – Distribuição espacial das estimativas de β_4 considerando os dados após a exclusão de Oconee para o modelo REGPM com erros normais (esquerda) e t -Student (direita)



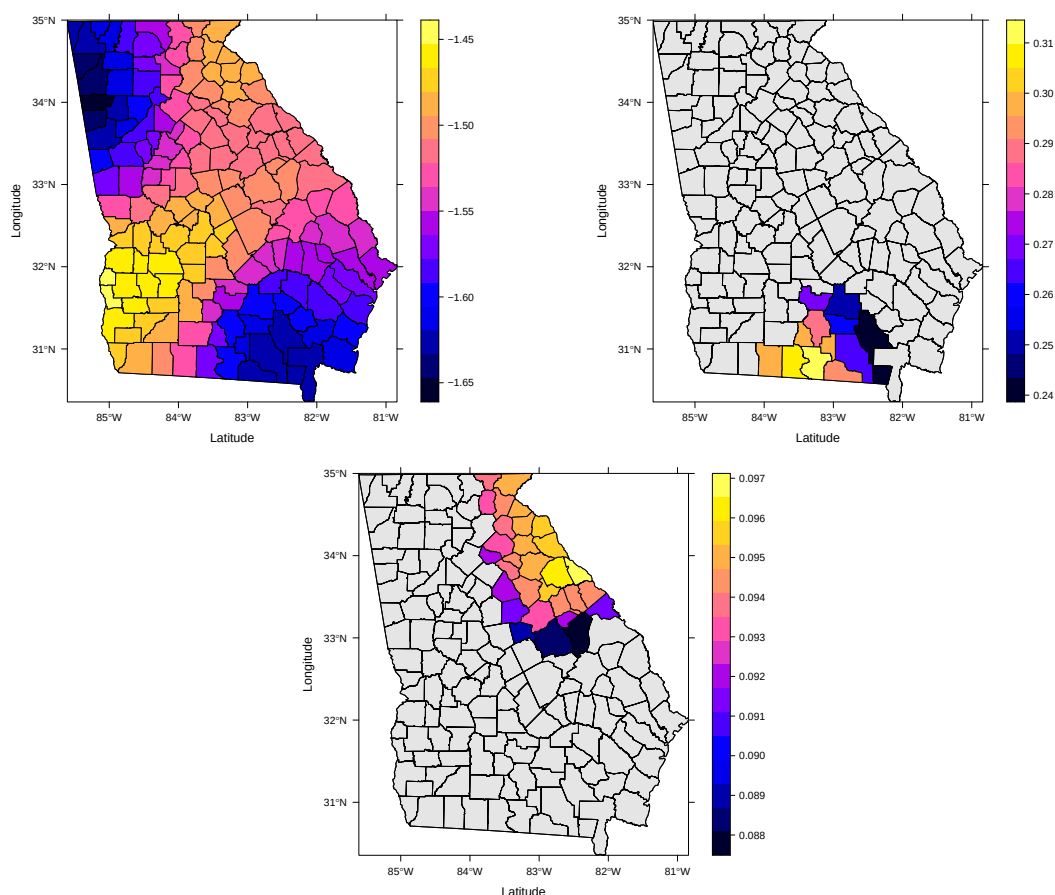
Fonte: O autor (2021).

Observe que, as distribuições espaciais das estimativas obtidas no modelo de REGPM com erros normais ao desconsiderar os condados de Bulloch, Clarke e Ocoone mostram mudanças extremas nas estimativas dos parâmetros. Isto é, a retirada dessas observações muda completamente a interpretação da variação local dos parâmetros. Por outro lado, a distribuição espacial das estimativas no modelo de REGPM com erros t -Student não apresenta alterações substanciais ao retirar os condados mencionados anteriormente, conservando as mesmas interpretações dos resultados obtidos com os dados completos. Esse fato indica que o modelo REGPM com erros t -Student produz estimativas menos sensível à presença de valores discrepantes.

3.5.5 Interpretação do Modelo Ajustado

Conforme apresentado nos resultados anteriores, o modelo REGPM com erros t -Student mostrou melhores valores para os critérios de informação e menor sensibilidade às observações discrepantes. O ajuste desse modelo indica que os parâmetros β_2 e β_4 são significativos (ao nível de 95% de confiança) e praticamente constantes no espaço, uma vez que o parâmetro de suavização selecionado é muito próxima ao valor total de observações. Os valores descritivos das estimativas desses parâmetros podem ser consultados na Tabela 7, ressaltando que todos os locais apresentam valores bem próximo da sua média. Em contrapartida, os parâmetros β_0 , β_1 e β_3 apresentam comportamentos não estacionários no espaço, os quais são destacados pelo ajuste do modelo REGPM com erros t -Student. Visto isso, as distribuições espaciais das estimativas significativas (assumindo o nível descritivo corrigido com uma taxa de erro fixada em 5%) desses parâmetros são apresentadas na Figura 98, destacando as estimativas não significativas na escala de cores pela cor cinza.

Figura 98 – Distribuição espacial das estimativas significativas de β_0 (esquerda), β_1 (direita) e β_3 (abaixo) no modelo REGPM com erros t -Student



Fonte: O autor (2021).

Note inicialmente que o modelo foi ajustado para a variável PctBach transformada, entretanto devido as propriedades da transformação *logit* é possível interpretar os resultados obtidos em termos da porcentagem de população com bacharel. Em outras palavras, o aumento do valor da variável L.PctBach também implicará no aumento da porcentagem de população com bacharel, no entanto isso ocorre em escalas diferentes. Com base nesse fato podemos destacar que, em média o acréscimo de cem mil unidade na população total, quando as demais variáveis se mantêm fixas, implica em mais pontos percentuais da população com diploma de bacharel. No entanto, esse acréscimo é significativo apenas em poucos condados na região sul da Georgia. Nota-se também que o acréscimo no percentual da população definida como rural, fixando as demais variáveis, diminui o percentual da população com diploma de bacharel. Nota-se também que o acréscimo da porcentagem de estrangeiros aumenta a porcentagem de diplomas de bacharelado, mais esse comportamento é significativo apenas em alguns condados na região nordeste da Georgia. Por fim, o acréscimo de um ponto percentual na população que vive abaixo da linha da pobreza, quando as demais variáveis são fixadas, implica na diminuição dos pontos percentuais na população com diploma de bacharel. Todos esses resultados indicam que o desenvolvimento educacional para o ensino superior é explicado por diferentes indicadores, apresentando relações particulares de acordo com a região considerada.

4 IMPLEMENTAÇÃO COMPUTACIONAL

Com a finalidade de facilitar a aplicação das metodologias desenvolvidas nesta pesquisa em novas áreas, foi elaborado um conjunto de rotinas computacionais através da linguagem R (R Core Team, 2019). A escolha desse *software* computacional se dar pelo fato do seu código fonte ser aberto, podendo ser facilmente manipulado pelos seus usuários. Essas rotinas foram implementadas por meio do pacote denominado **gwer**, o qual está disponível no CRAN do *software* R. O pacote **gwer** contém um diversas funções bases exclusivas e bancos de dados utilizados como exemplo práticos em algumas dessas funções. É possível agrupar todas estas funções em três categorias principais, as quais são especificadas a seguir.

1. Ajuste do modelo: funções para a obtenção dos modelos de regressão elíptico, REGP e REGPM ajustados a uma base de dados fornecida pelo usuário.
2. Visualização do ajuste: funções utilizadas para resumir os resultados principais obtidos pelo ajuste dos modelos de regressão elíptico, REGP e REGPM.
3. Análise de diagnóstico: funções para calcular e apresentar graficamente as medidas de diagnóstico nos modelos de regressão elíptico, REGP e REGPM.

A seguir são apresentadas as funções que estão contidas em cada uma dessas categorias, explicando os seus parâmetros de entrada e saída. Além disso, é discutido como essas funções devem ser utilizadas no *software* R, assim como alguns exemplos para suas manipulações. Vale ressaltar que esse pacote computacional está em constante aperfeiçoamento e algumas estruturas explicadas nesse capítulo podem sofrer futuras alterações.

4.1 AJUSTE DO MODELO

As funções discutidas nessa seção são a base para aplicação dos algoritmos de estimação abordados nessa pesquisa. Em particular, a função **family.elliptical** e suas sub-rotinas são fundamentais para especificar e exibir as definições acerca da componente aleatória do modelo. Para extrair as informações da componente aleatória de um objeto previamente ajustado usa-se diretamente a função **family**, a qual possui a seguinte sintaxe:

```
family(object, ...)
```

Esta função deve receber um objeto que contenha a saída de uma função utilizada para o ajuste do modelo especificado pelo usuário e retornará uma lista de informações da distribuição utilizada nesse ajuste. Este objeto pode ser da classe `elliptical`, `gwer` ou `gwer.multiscale`, as quais serão discutidas posteriormente. Por outro lado, a especificação da componente aleatória é feita através das sub-rotinas da função `family.elliptical`. Para isso é preciso decidir qual distribuição de probabilidade será utilizada para o ajuste do modelo e fornecer os valores dos parâmetros fixos necessários. Nesse sentido, a seguir é apresentado as sub-rotinas dessa função para cada uma das distribuições implementadas. Uma dessas funções deverá então ser fornecida como argumento de entrada para a função de ajuste através do parâmetro adequado.

```
Normal()
Cauchy()
LogisI()
LogisII()
Student(df = stop("no df argument"))
Powerexp(k = stop("no k argument"))
Glogis(parma = stop("no alpha=alpha(m) or m argument"))
Gstudent(parm = stop("no s or r argument"))
Cnormal(parmt = stop("no epsi or sigma argument"))
```

Com isso, para realizar o ajuste do modelo de regressão linear global com componente aleatória pertencente à família de distribuições elípticas utiliza-se a função `elliptical`, a qual retorna um objeto da classe `elliptical`. Este objeto armazena inúmeros resultados obtidos através do ajuste de um modelo de regressão elíptico a base de dados fornecida pelo usuário. De forma sucinta, essa função definida sob a seguinte sintaxe:

```
elliptical(formula = formula(data), family = Normal, data = sys.parent(),
dispersion = NULL, weights, subset, na.action = "na.fail", method =
"elliptical.fit", control = glm.control(epsilon = 1e-04, maxit = 100,
trace = F), model = F, x = F, y = T, contrasts = NULL, offset, ...)
```

Em contrapartida, a função `gwer` refere-se ao ajuste de um modelo de REGP para um dado parâmetro de suavização fornecido pelo usuário. A partir desta função é

possível especificar diferentes características associadas a função de ponderação espacial e a medida de distância, as quais deverão ser utilizadas no processo de estimação do modelo. Esta função retorna um objeto da classe `gwer`, contendo as saídas obtidas pelo ajuste do modelo de REGP. A sintaxe base dessa função é dada pelo seguinte código computacional:

```
gwer(formula, data, regression.points, bandwidth, kernel = "bisquare",
p = 2, theta = 0, adapt = NULL, hatmatrix = FALSE, family = Normal,
longlat = NULL, dMat, weights, dispersion = NULL, subset, na.action =
"na.fail", method = "gwer.fit", control = glm.control(epsilon = 1e-04,
maxit = 100, trace = F), model = FALSE, x = FALSE, y = TRUE, contrasts
= NULL, offset, spdisp = TRUE, parplot = FALSE,...)
```

Visto que a função `gwer` funciona apenas para um parâmetro de suavização preestabelecido e fornecido pelo usuário. Por este motivo foi desenvolvido uma função para a otimização desse parâmetro de suavização. Para isto, foi considerada a técnica da seção áurea para a minimização de alguma função objetivo especificada (para maiores detalhes da implementação computacional dessa técnica consulte Press *et al.*, 2007). Nessa pesquisa discutimos os métodos de seleção do parâmetro de suavização baseados na minimização dos critérios de CV, AIC e SV, os quais foram apresentados no segundo capítulo. É necessário também atribuir as características assumida para as ponderações espaciais, as quais devem está em concordância com o uso posterior da função `gwer`. Com isso, ao final da execução dessa função é retornado o valor ótimo do parâmetro de suavização usado nas ponderações espaciais, de acordo com a medida de distância especificada e o critério atribuído pelo usuário. De modo geral, a função para seleção do parâmetro de suavização ótimo, denominada por `bw.gwer`, é utilizada a partir da sintaxe dada a seguir:

```
bw.gwer(formula, data, approach="CV", kernel="bisquare", adaptive = F,
spdisp = 'local', dispersion, p=2, theta=0, longlat=F, dMat, family =
Normal())
```

Por fim, o ajuste da REGPM é obtido através da função `gwer.multiscale`, a qual foi implementada com base no método de estimação descrito no capítulo anterior. De forma semelhante a função `gwer`, o usuário deve fornecer as informações necessárias para a obtenção das ponderações espaciais usadas no algoritmo de estimação. As metodologias discutidas anteriormente para reduzir o custo computacional também estão implementadas

nessa função, as quais devem ser escolhidas no momento da sua utilização. Nesse caso, os resultados obtidos através da aplicação dessa função são retornados através de um objeto da classe `gwer.multiscale` ao final da sua execução. Dito isso, a seguir é apresentado a sintaxe base para o uso dessa função:

```
gwer.multiscale(formula, data, coords, kernel = "bisquare", approach =
"CV", adaptive = FALSE, criterion = "dCVR", fit.points, family = Normal,
threshold = 0.00001, dMats, p.vals, theta.vals, longlat = NULL, bws0,
bw.seled = rep(F, length(bws0)), bws.thresholds = rep(0.1, length(dMats)),
bws.reOpts = 5, spdisp = 'local', verbose = F, weights, dispersion = NULL,
na.action = "na.fail", hatmatrix = T, control = glm.control(epsilon =
1e-04, maxit = 100, trace = F), model = FALSE, x = FALSE, y = TRUE,
contrasts = NULL, parplot = FALSE, max.iterations = 2000, subset, offset,
predictor.centered = rep(T, length(bws0) - 1), nlower = 10, ...)
```

Note que, embora essas três funções sejam caracterizadas por diferentes argumentos de entrada, alguns desses têm as mesmas definições e finalidades. Visto isto, todos os argumentos utilizados para a execução dessas funções são listados a seguir de maneira conjunto como forma de evitar informações repetidas. Para cada um desses argumentos é considerado uma breve explicação de como estes devem ser utilizados e suas respectivas finalidades.

- **formula:** Especifica de forma simbólica a componente sistemática do modelo, descrita como uma função linear dos parâmetros de regressão. Isto é feito fornecendo a relação de regressão como entrada para esse argumento, sob a forma padrão usualmente utilizada nas funções do *software* R (veja, por exemplo, a função `lm`).
- **family:** Especifica de forma simbólica a componente aleatória do modelo, a qual deve pertencer à família de distribuições elípticas. Para isto, é necessário o uso das sub-rotinas associadas a função *"elliptical.family"*. A entrada desse argumento deve ser uma função que descreve a distribuição de probabilidade assumida para os erros, assim como os parâmetros fixos indispensáveis para sua especificação.
- **data:** Recebe o objeto contendo a base de dados sob uma formatação conveniente. No caso da função `elliptical` este argumento deve conter um objeto estar no formato `data.frame`, `list` ou `environment`. Por outro lado, as funções `gwer` e `gwer.multiscale` precisa receber como argumento um objeto no formato

`SpatialPointsDataFrame` ou `SpatialPolygonsDataFrame`, os quais podem ser definidos a partir do pacote `sp`.

- **bandwidth**: Argumento que recebe o valor do parâmetro de suavização usado na construção das ponderações espaciais. Esse valor pode ser fornecido pelo usuário com base em algum conhecimento a priori acerca dos dados ou por meio de um processo de otimização (veja `bw.gwer`).
- **approach**: Argumento para definir o método utilizado no processo de otimização do parâmetro de suavização. Os métodos implementados são realizados pela minimização das funções descritas no capítulo 2 através da seção áurea.
- **kernel**: Define a função de ponderação geográfica utilizada no ajuste dos modelos de REGP e REGPM. Nas funções para o ajuste desses modelos há algumas opções implementadas para esse argumento, como por exemplo os *kernel* gaussiano e biquadrado expressos no segundo capítulo.
- **p** e **theta**: Utilizados para o cálculo das medidas distância espaciais entre os locais da base de dados para o processo de estimação da REGP. O argumento **p** representa o valor da potência na distância de Minkowski. O valor padrão desse argumento é 2, indicando uma distância euclidiana. O valor de **theta** representa o ângulo em radianos para rotacionar o sistema de coordenadas, caso seja conveniente. O valor padrão desse argumento é 0, ou seja, não considera nenhuma rotação no sistema de coordenadas.
- **adapt**: É um argumento lógico que define o tipo de parâmetro de suavização utilizado no ajuste dos modelos de REGP e REGPM. Caso receba o valor `FALSE` a função assume que as ponderações espaciais são obtidas através de um parâmetro de suavização fixo para todos os locais. Por outro lado, se esse argumento for `TRUE` a função calcula as ponderações espaciais com base em parâmetros de suavização que se adaptam as distribuições espaciais de cada local.
- **hatmatrix**: É um argumento lógico utilizado para indicar se as matrizes de projeção deverão ser calculadas e retornadas ao usuário (`TRUE`) ou não (`FALSE`) nos modelos de REGP e REGPM. No caso que o usuário queira obter medidas de diagnóstico com base no ajuste desses modelos é necessário que este argumento receba o valor `TRUE`. Vale ressaltar que esse argumento só poderá receber o valor `TRUE` se locais estudados apresentarem as mesmas coordenadas geográficas que os pontos de dados observados.

- **longlat**: É um argumento lógico para especificar o tipo das coordenadas geográficas utilizadas para calcular as medidas de distância. Use o valor **TRUE** se as coordenadas fornecidas na base de dados representarem longitude e latitude e **FALSE** caso contrário. Caso esse argumento seja **TRUE** será considerado a distância do grande círculo para o cálculo das ponderações geográficas ao invés da distância de Minkowski.
- **dMat**: Argumento que recebe uma matriz de distância pré-especificada utilizada no processo e estimação da REGPM. Essa matriz pode ser obtida por meio de algum conhecimento a priori do usuário ou calculada por meio de alguma função específica (veja a função **gw.dist** do pacote **GWmodel**, por exemplo).
- **dMats**: Argumento que recebe uma lista de matrizes de distância usadas para o processo de estimação dos parâmetros na REGPM. Note que, esse argumento permite estabelecer uma medida de distância distinta associada a cada variável preditora no algoritmo de estimação. Isto é possível no ajuste de um modelo de REGPM uma vez que cada parâmetro de regressão local é estimado por meio de ponderações espaciais individuais.
- **p.vals** e **theta.vals**: Utilizados para o cálculo das medidas distância espaciais entre os locais da base de dados para o processo de estimação da REGPM. O argumento **p.vals** recebe um vetor (de comprimento igual ao número de preditores) de números positivos usados para definir a potência na distância de Minkowski para cada preditor. Por outro lado, o argumento **theta.vals** deve ser um vetor (de comprimento igual ao número de preditores) com os ângulos em radianos para girar o sistema de coordenadas para cada preditor.
- **bws0**: Argumento utilizado para estabelecer os valores iniciais dos parâmetros de suavização no processo de estimação da REGPM. Nesse caso, o usuário deve fornecer um vetor (de comprimento igual ao número de preditores) com os valores iniciais dos parâmetros de suavização associados a cada preditor utilizado no ajuste da REGM.
- **bw.seled**: Argumento utilizado para especificar quais parâmetros de suavização deverão passar por um processo de otimização ou deverão ser fixas no processo de estimação. O usuário deve fornecer um vetor lógico (de comprimento igual ao número de preditores) de tal forma que, os elementos com valor **TRUE** indica que seu respectivo parâmetro de suavização deverá ser mantido fixo na execução do algoritmo. Os valores dos parâmetros de suavização são retirados do argumento

`bws0`, descrito no item anterior.

- **`bws.thresholds`**: Define os valores limites para o qual o algoritmo assume que o parâmetro de suavização associado a um parâmetro específico convergiu ou não. Nesse caso o usuário deve fornecer um vetor (de comprimento igual ao número de preditores) de valores reais positivos que define a distância limite associada a cada preditor. Esses valores representam a distância máxima entre os valores selecionados para os parâmetro de suavização em cada iteração com seu respectivo valor no passo anterior.
- **`bws.re0pts`**: Define o número de passos necessários para obter convergência na otimização dos parâmetros de suavização do modelo de REGPM. Nesse caso o usuário deve fornecer um vetor (de comprimento igual ao número de preditores) com valores inteiros positivos associados a cada preditor. Logo, durante a otimização de cada parâmetro de suavização é necessário que o valor ótimo obtido seja igual ao passo anterior de acordo com o número de vezes fornecida por este argumento. Isto serve para evitar uma escolha que não seja ideal devido à ilusão de convergência.
- **`dispersion`**: Argumento usado para fornecer, de forma opcional, um valor fixo positivo para o parâmetro de dispersão. Se o usuário optar por atribuir um valor a esse argumento, então o algoritmo assume que o parâmetro de dispersão é fixo em todo processo de estimação.
- **`weights`**: Argumento utilizado para introduzir uma ponderação a priori no processo de estimação do modelo. Caso o usuário tenha interesse, deverá ser fornecido um vetor de pesos (de comprimento igual ao número de dados observados) a esse argumento, o qual irá executar uma reponderação na expressão dos estimadores para os parâmetros de regressão.
- **`subset`**: Argumento usado para especificar, de forma opcional, um vetor a partição das observações que serão usadas no processo de estimação do modelo. Nesse caso, o usuário deverá fornecer um vetor com valores lógicos indicando quais observações deveram ser consideradas no processo de estimação.
- **`na.action`**: Recebe, de forma opcional, uma função que indica o que deve ocorrer com as observações que contenham valores **NA** na base de dados.
- **`method`**: Define o método de otimização usado para a estimação dos parâmetros. O padrão é considerar os métodos escore de Fisher descritos nesta pesquisa, no entanto

também é possível não fazer o ajuste e retornar apenas a matriz modelo.

- **criterion**: Argumento para definir a condição estabelecida para convergência do algoritmo *backfitting* no processo de estimação da REGPM. O valor padrão desse argumento é `dCVR` representando o critério SOC_{RSS} descrito no capítulo anterior.
- **control**: Representa uma lista de argumentos utilizados para controlar o processo de estimação do modelo, especificada por meio da função `glm.control`.
- **model**: É um valor lógico que indica se o quadro do modelo deve ser incluído como argumento de saída da função (`TRUE`) ou não (`FALSE`).
- **y e x**: Representam valores lógico que indicam se o vetor de resposta e a matriz modelo usados no processo de estimação devem ser inclusos como argumentos de saída da função (`TRUE`) ou não (`FALSE`).
- **contrasts**: É uma lista de argumentos para definir e utilizar contrastes no modelo (Para mais detalhes veja `contrasts.arg` presente na função `model.matrix.default`).
- **offset**: Argumento que pode ser usado para especificar um componente a priori para ser incluída no preditor linear durante o ajuste do modelo (de forma equivalente a função `glm`). Para mais detalhes veja a função `model.offset` presente no pacote base do *software* R.
- **spdisp**: Argumento que define a variação espacial do parâmetro de dispersão nos modelos de REGP e REGPM.
- **parplot**: Argumento lógico para estabelecer se a função irá fornecer os boxplots dos parâmetros locais como saída gráfica (`TRUE`) ou não (`FALSE`).
- **verbose**: Argumento lógico utilizado para visualizar (`TRUE`) ou não (`FALSE`) o progresso no processo de otimização do parâmetro de suavização.
- **max.iterations**: Define o número máximo de iterações para o algoritmo *backfitting* no processo de estimação do modelo REGPM.
- **predictor.centered**: Argumento que recebe um vetor lógico de comprimento igual ao número de variáveis preditoras (excluindo-se o preditor relacionado ao intercepto), indicando qual variável será centralizada na média. Se o k -ésimo elemento desse vetor for `TRUE`, então o k -ésimo preditor será centralizado em função de sua média. Este procedimento é útil para reduzir o tempo computacional de processamento da função `gwer.multiscale`, como foi discutido no capítulo anterior.
- **nlower**: Estabelece o valor mínimo de vizinhos mais próximos para a seleção do

parâmetro de suavização adaptativo no cálculo das ponderações espaciais. O usuário deve fornecer um valor inteiro positivo para este argumento.

- **df**: É o argumento usado para definir os graus de liberdade da distribuição *t*-Student.
- **s e r**: São os argumentos usados para definir os parâmetros de forma da distribuição *t*-Student generalizada.
- **alpha**: É o argumento usado para definir o parâmetro de forma das distribuições normal contaminada e logística.
- **mp**: É o argumento usado para definir o parâmetro de forma da distribuição logística generalizada.
- **epsi e sigmap**: São os argumentos usados para definir os parâmetros de dispersão da distribuição normal contaminada.
- **k**: É o argumento usado para definir o parâmetro de forma da distribuição exponencial potência.
- **...**: Argumentos adicionais que podem ser usados na função e que não foram descritos diretamente pela sintaxe da função.

Uma vez explicado cada argumento necessário para a utilização das rotinas destinadas ao ajuste desses modelos, considerou-se uma breve demonstração das suas funcionalidades. Assim, como forma de exemplificar a utilização das funções discutidas nesta seção consideramos uma breve aplicação em uma base de dados reais. Dessa forma, considere o exemplo prático apresentado no capítulo anterior, referente aos dados censitários do estado da Georgia, EUA. Esses dados podem ser facilmente obtidos através do pacote `spgwr`, usando o comando `data(georgia, package="spgwr")` carregando a base de dados no objeto `gSRDF`. Assuma então os mesmos modelos propostos nessa análise, os quais podem ser ajustados por meio dos seguintes comandos:

```
data(georgia, package="spgwr")
formula <- PctBach ~ TotPop90 + PctRural + PctFB + PctPov

ajuste.gt <- elliptical(formula, family = Student(df = 3), data = gSRDF,
method = "elliptical.fit")

regp.b <- bw.gwer(formula, data = gSRDF, family = Student(df = 3),
approach = 'cv', adapt = TRUE, kernel = 'bisquare')
```

```
ajuste.regp <- gwer(formula, data = gSRDF, family = Student(df=3),
  bandwidth = regp.b, adapt = TRUE, hatmatrix = TRUE,
  spdisp = TRUE, method = "gwer.fit")

ajuste.regpm <- gwer.multiscale(formula, data = gSRDF, family = Student(3),
  approach = 'cv', kernel = 'bisquare',
  bws0 = rep(regp.b, 5), hatmatrix = TRUE,
  adaptive = TRUE, predictor.centered =
  rep(T, 5-1))
```

Nesse caso, o objeto `formula` representa a expressão simbólica para a componente sistemática usada como argumento em todas as funções de ajuste utilizadas. Em todos os casos é especificado a distribuição *t*-Student com 3 g.l. para os erros aleatórios do modelo. Os resultados obtidos no ajuste da regressão elíptico foram armazenados em `ajuste.gt`, sendo possível extrair diversas informações desse objeto. Para o ajuste da REGP o parâmetro de suavização foi selecionado através da minimização da validação cruzada, assumindo um kernel adaptativo biquadrado com distâncias calculadas através das coordenadas presente nos dados `gSRDF`. O valor do parâmetro de suavização ótimo é armazenado em `regp.b` e utilizado como argumento dos ajustes das regressões geograficamente ponderadas. Os resultados do ajuste da REGP são armazenados no objeto `ajuste.regp`, os quais podem ser visualizado posteriormente. No caso do ajuste da REGPM são fornecidos os mesmos argumentos em comum com a REGP. Todos os parâmetros de suavização são inicializados pelo valor armazenado em `regp.b` e otimizadas através do algoritmo *backfitting*. Além disso, todas as variáveis explicativas foram centralizadas em zero com o objetivo de obter um ajuste explanatório de forma mais rápida. Em ambos ajustes dos modelos geograficamente ponderados considerou-se o cálculo da matriz de projeção, sendo fundamental para a obtenção das medidas de diagnósticos posteriormente. Por fim, observe que as funções implementadas nesse pacote fornecem a possibilidade de diferentes ajustes para a mesma base de dados, os quais dependem da especificação dos argumentos passados a essas funções.

4.2 VISUALIZAÇÃO DO AJUSTE

Note que as funções `elíptical`, `gwer` e `gwer.multiscale` fornecem como resultado algumas informações obtidas através dos seus respectivos ajustes a base de dados considerada. Esses resultados já podem ser utilizados para avaliar a qualidade do ajuste de maneira básica e interpretar o modelo ajustado. Essas informações são retornadas por estas funções como uma extensa lista de objetos, os quais são detalhados no suporte da própria função no *software* R. No entanto, extrair as informações mais úteis pode ser um trabalho árduo, principalmente se o usuário ainda não estiver familiarizado com essas rotinas. Nesse caso, o uso da função `print` é uma opção inicial para destacar os principais resultados obtidos por meio do ajuste de algum modelo. A título de exemplo, considere o ajuste da REGP obtido na seção anterior e armazenado no objeto `ajuste.regp`. Dessa forma, a aplicação da função `print` nesse objeto retorna a seguinte saída:

```
> print(ajuste.regp)
```

Call:

```
gwer(formula = formula, data = gSRDF, bandwidth = regp.b, adapt = TRUE,
      hatmatrix = TRUE, family = Student(df = 3), method = "gwer.fit",
      spdisp = TRUE)
```

Kernel function: bisquare

Adaptive quantile: TRUE (about 159 of 159 data points)

	Min.	1st Qu.	Median	3rd Qu.	Max.
X.Intercept.	9.0514e+00	1.2053e+01	1.3784e+01	1.4848e+01	1.5815e+01
TotPop90	1.8808e-05	2.3733e-05	2.5075e-05	3.0303e-05	8.8092e-05
PctRural	-6.5529e-02	-4.9288e-02	-4.3031e-02	-3.1917e-02	-1.7378e-02
PctFB	5.9448e-02	3.6304e-01	1.1484e+00	1.4094e+00	1.6793e+00
PctPov	-1.7613e-01	-1.3693e-01	-1.1674e-01	-9.4462e-02	-1.2957e-02
dispersion	2.6128e+00	2.9013e+00	3.4941e+00	4.3019e+00	4.6016e+00
Global					
	13.4523				
	0.0000				
	-0.0419				
	0.9452				

-0.1127

4.0235

Number of data points: 159

Effective number of parameters (residual: $2\text{traceS} + \text{traceS}'\text{S}$): 17.62782

Effective degrees of freedom (residual: $n - 2\text{traceS} - \text{traceS}'\text{S}$): 141.3722

Effective number of parameters (model: traceS): 14.49974

Effective degrees of freedom (model: $n - \text{traceS}$): 144.5003

AIC: 770.6953 AICc: 773.8276 BIC: 815.1936

Note que, o início da saída apresenta os argumentos utilizados na execução da função e algumas informações que caracterizam o modelo proposto pelo usuário. Além disso, essa função retorna algumas medidas descritivas para os parâmetros locais do modelo, assim como seus respectivos valores globais (obtidos através do ajuste da regressão global). Por fim, algumas medidas associadas a qualidade do ajuste são fornecidas ao usuário. As saídas dos objetos `ajuste.gt` e `ajuste.regpm` ao aplicar a função `print` também devem ser semelhantes, embora ressalte as características particulares de cada modelo. No entanto, as informações fornecidas desta forma são até certo ponto limitadas e formas alternativas podem ser consideradas. Nesse sentido, as funções `summary.elliptical` e `summary.gwer` também fornecem informações adicionais sobre o ajuste dos modelos discutidos nessa pesquisa. Em ambos os casos, utiliza-se na prática apenas a seguinte sintaxe:

```
summary(object, ...)
```

De forma semelhante a função `print`, o usuário deve fornecer o objeto de ajuste da classe `elliptical` ou `gwer` para o argumento `object`. Nesse caso, a função `summary` calcula medidas adicionais com base no ajuste obtido previamente e as exibe de forma mais completa que a função `print`. Considerando o exemplo anterior, é possível obter a seguinte saída:

```
summary(ajuste.regp)
```

```
Call: gwer(formula = formula, data = gSRDF, bandwidth = regp.b, adapt = TRUE,
           hatmatrix = TRUE, family = Student(df = 3), method = "gwer.fit",
           spdisp = TRUE)
```

```
Summary of pearson Residuals:
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
	-1.94010	-0.52249	0.03074	0.11956	0.44307	5.99298

Summary of GWER coefficient estimates:

	Min.	1st Qu.	Median	Mean
(Intercept)	9.05141e+00	1.20531e+01	1.37839e+01	1.33507e+01
TotPop90	1.88084e-05	2.37325e-05	2.50748e-05	3.16463e-05
PctRural	-6.55287e-02	-4.92876e-02	-4.30309e-02	-4.14227e-02
PctFB	5.94483e-02	3.63042e-01	1.14836e+00	9.55848e-01
PctPov	-1.76129e-01	-1.36934e-01	-1.16742e-01	-1.12379e-01
dispersion	2.61280e+00	2.90132e+00	3.49412e+00	3.58699e+00
	3rd Qu.	Max.	Global	
	1.48484e+01	1.58153e+01	1.34523e+01	
	3.03028e-05	8.80924e-05	2.70321e-05	
	-3.19173e-02	-1.73775e-02	-4.19231e-02	
	1.40944e+00	1.67929e+00	9.45183e-01	
	-9.44619e-02	-1.29569e-02	-1.12723e-01	
	4.30193e+00	4.60162e+00	4.02349e+00	

Summary of GWER standard errors estimates:

	Min.	1st Qu.	Median	Mean
(Intercept).se	1.75474e+00	1.87499e+00	1.93243e+00	1.97712e+00
TotPop90.se	4.36797e-06	4.85552e-06	6.45504e-06	8.09742e-06
PctRural.se	1.51860e-02	1.62668e-02	1.71086e-02	1.70833e-02
PctFB.se	2.93572e-01	3.80740e-01	4.02823e-01	3.93586e-01
PctPov.se	4.58327e-02	5.15164e-02	5.50399e-02	5.51334e-02
dispersion.se	7.71345e-01	8.47744e-01	1.01676e+00	1.00972e+00
	3rd Qu.	Max.	Global	
	2.054699472	2.395774348	1.068586849	
	0.000010659	0.000016137	0.000003278	
	0.017776511	0.019813031	0.009457989	
	0.420609426	0.487084387	0.212349343	


```
0.058255458 0.071056616 0.028931405
1.156791011 1.363067722 0.638166504
```

```
Number of data points: 159
```

```
Global dispersion parameter for Student: 2.88277 (2.84574)
```

```
Effective number of parameters (residual: 2traceS + traceS'S): 17.6278
```

```
Effective degrees of freedom (residual: n - 2traceS - traceS'S): 141.372
```

```
Effective number of parameters (model: traceS): 14.4997
```

```
Effective degrees of freedom (model: n - traceS): 144.5
```

```
-2*Log-Likelihood: 741.696 AIC: 770.695 AICc: 773.828 BIC: 815.194
```

Observa-se que além das mesmas informações exibidas por meio da função `print` há também resultados referentes aos erros padrões e os resíduos. Adicionalmente, a função `anova.elliptical` pode ser utilizada para obter a análise de variância do modelo de regressão elíptico. Essa função se aplica apenas a objetos da classe `elliptical` e pode ser usada de forma simples pela seguinte sintaxe:

```
anova.elliptical(object, dispersion = NULL, ...)
```

Nesse caso, o objeto de ajuste deve ser fornecido através do argumento `object`, assim como na função `summary`. O argumento `dispersion` é usado para especificar um valor fixo para o coeficiente de dispersão. Caso não seja fornecido nenhum valor a função retira essa informação do próprio objeto de ajuste. Um exemplo prático da utilização dessa função, é mostrado a seguir aplicando a função `anova` ao objeto `ajuste.gt`.

```
> anova(ajuste.gt)
```

```
Analysis of Deviance Table
```

```
Error distribution: Student
```

```
Response:PctBach
```

```
Terms added sequentially (first to last)
```

	Df	Deviance	Resid. Df	-2*LL	Pr(>Chi)
NULL			158	934.29	
TotPop90	1	111.969	157	822.32	< 2.2e-16 ***

```
PctRural  1    13.606          156 808.72 0.0002255 ***
PctFB     1    15.904          155 792.81 6.663e-05 ***
PctPov    1    11.183          154 781.63 0.0008253 ***
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nesse caso, é retornado como saída uma tabela referente a análise de variância do modelo ajustado. Vale ressaltar que, as funções `summary` e `anova` também gera uma lista de saídas, as quais podem ser armazenadas em um objeto e utilizadas posteriormente. Naturalmente, é possível extrair ainda mais resultados através do manuseio da lista de saídas de cada uma das funções presentes no pacote. Para mais detalhes consulte a interface de ajuda do pacote por meio do `help` da respectiva função utilizada.

4.3 ANÁLISE DE DIAGNÓSTICO

Uma vez realizado o ajuste de um modelo é natural verificar suas suposições e sensibilidade às observações influentes. Neste contexto, foram implementadas algumas rotinas para a aplicação dos métodos de diagnóstico proposto na presente pesquisa. Essas rotinas foram divididas em diferentes funções presentes no pacote `gwer`. Inicialmente, as funções `residuals.elliptical` e `residuals.gwer` calculam diferentes tipos de resíduos para um modelo previamente ajustado. Nesse caso, a função `residuals.elliptical` obtém as medidas de resíduos do modelo de regressão elíptico, enquanto que a função `residuals.gwer` faz o mesmo procedimento para a REGP. No entanto, a sintaxe de ambas funções são equivalentes e descrita da seguinte forma:

```
residuals(object, type = c("stand", "ordinal", "response",
"pearson", "desvio"), ...)
```

A diferença está na classe do objeto de entrada, podendo ser `elliptical` ou `gwer` de acordo com o ajuste realizado pelo usuário. Em ambos os casos deve ser especificado qual tipo de resíduos a função deve retornar por meio do argumento `type`, embora esta ação não seja exigida. No caso em que o usuário não forneça o tipo de resíduo desejado o programa irá retornar os resíduos padronizados. Além disso, por meio desse pacote é possível obter o gráfico de quantis-quantis com envelopes simulados para os resíduos padronizados dos modelos de regressão elíptica e REGP. Nesse caso, as funções

`elliptical.envelope` e `gwer.envelope` podem ser executadas para obter esses gráficos para os objetos de classe `elliptical` e `gwer`, respectivamente. A sintaxe base dessas duas funções é descrita da seguinte forma:

```
elliptical.envelope <- function (object, B = 100, arg, xlab = NULL,
ylab = NULL, ident = NULL, ident.labels = NULL, ...)
```

```
gwer.envelope <- function (object, B = 100, arg, xlab = NULL,
ylab = NULL, ident = NULL, ident.labels = NULL, ...)
```

Assim sendo, o argumento `B` deve receber um valor inteiro positivo referente ao número de réplicas de Monte Carlo consideradas na obtenção dos envelopes simulados. Adicionalmente, é preciso que o usuário forneça o valor dos parâmetros fixos da distribuição os erros assumidos no ajuste do modelo. Esse valor é repassado, quando necessário, ao parâmetro `arg` no formato de um vetor numérico com os valores dos parâmetros. Os argumentos `xlab` e `ylab` são utilizados para rotular os eixos gráficos, de maneira semelhante a função `plot` nos pacotes base do *software* R. Caso o usuário queira destacar alguma observação no gráfico isso deve ser informado por meio dos argumentos `ident` e `ident.labels`. Neste caso, `ident` deve receber o número de observações que o usuário deseja destacar e `ident.labels` deve ser um vetor (de comprimento igual ao número de observações) contendo os respectivos rótulos das observações. Como forma de ilustrar o uso das funções de resíduos, a seguir é apresentado a aplicação dessa função para as classes `elliptical` e `gwer` referente aos modelos ajustados anteriormente. Com isso, é obtido como saída os vetores de resíduos e os gráficos de quantis-quantis para os respectivos resíduos (ver Figura 99).

```
> res.gt <- residuals(ajuste.gt, type = "stand") ; round(res.gt, 3)
[1] -0.251 -0.461 -0.596 0.790 0.344 -0.508 -0.450 -0.688 -0.623 -0.724
[11] 0.388 0.130 -0.584 -0.220 0.685 2.871 0.472 -0.725 1.217 0.139
[21] 0.067 0.089 -1.071 -0.411 -0.146 0.511 -0.973 1.199 6.262 1.587
[31] -1.905 -0.460 1.329 0.197 -0.647 2.158 -1.150 0.610 -0.920 0.296
[41] -0.239 -0.009 0.335 -0.189 -0.239 0.597 0.942 -0.810 0.335 -1.309
[51] -0.587 -0.462 0.085 0.438 -0.145 3.067 0.020 1.320 0.201 -0.340
[61] 0.123 -0.730 1.724 -0.351 -0.625 0.161 0.845 0.017 -0.113 0.127
```

```

[71] -0.789  1.208 -0.142 -0.569 -0.404  0.023  0.057 -0.286  0.676 -0.356
[81] -0.110 -0.227 -0.681  0.621  0.109 -0.548  0.443  1.111 -0.525  0.144
[91] -1.006  1.035  0.449  0.402  0.423  0.407  0.253 -0.590 -0.484  0.252
[101] -0.335  0.811  0.576  0.337 -1.194 -0.781 -0.292  5.346  1.305 -0.777
[111]  1.421  0.175 -0.765  0.144 -1.107  0.137  0.218  0.233  0.721 -0.639
[121] -0.483  1.295  0.188  0.030 -1.545 -0.158  0.753  0.573  1.671  0.018
[131] -0.248 -1.034  0.146  0.270  0.178  0.888  0.495  0.188  0.651 -0.622
[141]  0.403 -0.414 -0.579  0.505 -0.378 -1.075 -0.479  0.015 -0.585  0.257
[151] -0.008 -0.396  0.664  1.240 -0.473  0.373  0.376  0.207 -0.667

> elliptical.envelope(ajuste.gt, B = 100, arg = 3, xlab =
+ "Quantis da normal padrão", ylab =
+ "Resíduos padronizados")

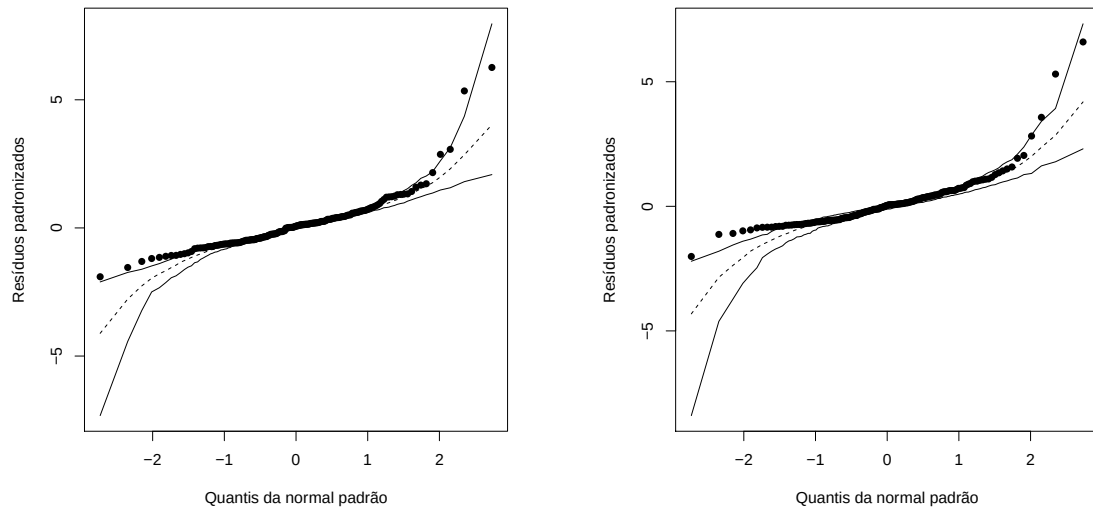
> res.regp <- residuals(ajuste.regp, type = "stand") ; round(res.regp, 3)
[1] -0.186 -0.162 -0.736  0.632  0.063 -0.601 -0.536 -0.556 -0.679 -0.464
[11]  0.355  0.029 -0.419  0.075  0.846  3.572  0.630 -0.802  1.099  0.459
[21]  0.107  0.197 -0.623 -0.156 -0.257  0.318 -0.690  1.163  6.597  1.500
[31] -2.012 -0.551  1.345  0.352 -0.367  2.042 -0.836  0.633 -1.127  0.175
[41] -0.235 -0.098  0.490 -0.458 -0.280  0.576  0.674 -0.562  0.124 -0.761
[51] -0.659 -0.679  0.068  0.720 -0.185  2.821  0.344  1.087  0.111 -0.120
[61]  0.096 -0.985  1.927 -0.234 -0.621  0.139  0.552 -0.450 -0.752  0.268
[71] -0.606  1.046 -0.301 -0.572 -0.449 -0.206  0.057 -0.381  0.587 -0.305
[81] -0.034 -0.318 -0.867  0.586  0.095 -0.455  0.453  1.018 -0.122  0.071
[91] -0.610  0.608  0.409  0.386  0.722  0.299  0.104 -0.730 -0.488  0.091
[101] -0.490  0.744  1.010  0.162 -1.088 -0.803 -0.368  5.308  1.273 -0.708
[111]  1.422  0.183 -0.758  0.064 -0.946  0.031 -0.107  0.144  0.488 -0.723
[121] -0.565  1.064  0.078  0.063 -0.372 -0.122  0.637  0.428  1.579 -0.109
[131] -0.208 -0.838  0.002  0.290 -0.011  0.751  0.888  0.139  0.394 -0.843
[141]  0.457 -0.545 -0.730  0.374 -0.404 -0.621 -0.561 -0.051 -0.574  0.232
[151] -0.047 -0.533  0.921  0.985 -0.650  0.413  0.288  0.220 -0.811

> gwer.envelope(ajuste.regp, B = 100, arg = 3, xlab =
+ "Quantis da normal padrão", ylab =

```

```
+ "Resíduos padronizados")
```

Figura 99 – Gráfico de quantis-quantis com envelope simulado dos resíduos padronizados para os modelos de regressão elíptico (esquerda) e REGP (direita) com erros t -Student



Fonte: O autor (2021).

Os resultados fornecidos por essas funções têm grande relevância para verificar as suposições do modelo em relação aos erros aleatórios. No entanto, essas medidas não são suficientes para verificar a qualidade do ajuste e a sensibilidade do modelo no contexto geral. Nesse sentido, as funções `elliptical.diag`, `gwer.diag` e `gwer.multiscale.diag` podem ser utilizadas para calcular todas as medidas de diagnóstico propostas nesta pesquisa de maneira bem simples. Naturalmente, a sintaxe dessas funções é bem semelhante e descritas da seguinte forma:

```
elliptical.diag(object, ...)
```

```
gwer.diag(object, ...)
```

```
gwer.multiscale.diag(object, ...)
```

Cada uma dessas funções refere-se ao respectivo ajuste de uma das classes de modelos implementadas nesse pacote. Para isso o usuário deve fornecer como argumento de entrada um objeto da respectiva classe do modelo ajustado. Em todos os casos uma lista de matrizes é retornada, a qual contém todas as medidas de diagnósticos do

ajuste fornecido. Essa saída pode ser armazenada em um objeto de interesse e utilizada posteriormente da forma que for mais conveniente para o usuário. Naturalmente, a análise gráfica dessas medidas pode facilmente ser realizada pelo usuário da rotina, embora exija um conhecimento prévio das rotinas utilizadas. Nesse contexto, foi desenvolvido um conjunto de função para facilitar ainda mais a construção dos gráficos referente a análise de diagnóstico. Caso o usuário do pacote tenha interesse, as funções `elliptical.diag.plots`, `gwer.diag.plots` e `gwer.multiscale.diag.plots` podem ser utilizadas para visualizar graficamente essas medidas de diagnósticos para os respectivos modelos ajustados. Estas funções também funcionam de maneira semelhantes, as quais são descritas pela seguinte sintaxe:

```
elliptical.diag.plots(object, ellipticaldiag = NULL, which, subset = NULL,
                      iden = FALSE, labels = NULL, ret = FALSE, ...)
```

```
gwer.diag.plots(object, gwerdiag = NULL, which, subset = NULL, iden = F,
                 labels = NULL, ret = F,...)
```

```
gwer.multiscale.diag.plots(object, gwerdiag = NULL, which, subset = NULL,
                            iden = F, labels = NULL, ret = F,...)
```

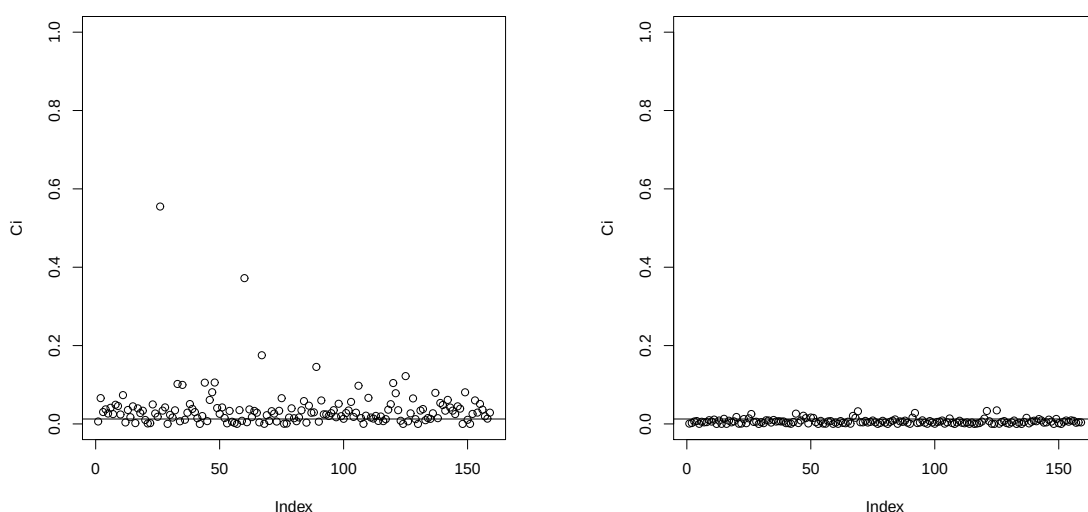
Novamente, cada funções refere-se ao respectivo ajuste de uma das classes de modelos implementadas previamente através das rotinas desse pacote. Essas funções retornaram um menu no console do *software* R e o usuário deverá selecionar o gráfico de diagnóstico desejado. É possível fornecer as medidas de diagnósticos obtidas anteriormente para os argumentos `ellipticaldiag` e `gwerdiag`, evitando que as mesmas medidas sejam calculadas novamente. Isso afeta apenas o tempo computacional para a execução da função, uma vez que as medidas são obtidas através do mesmo cálculo. Para que essas medidas sejam retornadas como saída dessas funções basta fornecer o valor lógico `TRUE` para o argumento `ret`. Caso o usuário tenha o interesse em destacar um único gráfico em específico é preciso fornecer ao argumento `which` o número correspondente no painel de saída. Esta ação irá fazer com que a saída gráfica da função não passe por um menu no console. Além disso, os argumentos `labels` e `iden` podem ser utilizados para fornecer, respectivamente, os rótulos e a quantidade de observações destacadas nos gráficos. Nesse caso, deve ser fornecido um vetor de rótulos para `labels` e um valor inteiro positivo para

iden. Para uma melhor compreensão, a seguir é apresentado um exemplo do uso dessas funções com base nos modelos ajustados anteriormente por meio dos seguintes comandos:

```
> diag.gt <- elliptical.diag(ajuste.gt)
> elliptical.diag.plots(ajuste.gt, ellipticaldiag = diag.gt,
which=11)

> diag.regp <- gwer.diag(ajuste.regp)
> gwer.diag.plots(ajuste.regp, gwerdiag = diag.regp, which=11)
```

Figura 100 – Gráficos de influência local para perturbação de casos ponderados nos modelos de regressão elíptico (esquerda) e REGP (direita) com erros t -Student



Fonte: O autor (2021).

A partir desses comandos, calcula-se as medidas de diagnósticos dos ajustes armazenados nos objetos `ajuste.gt` e `ajuste.regp`. Essas medidas são então utilizadas nas funções `elliptical.diag.plots` e `gwer.diag.plots` para realizar a análise gráfica apropriada. No exemplo apresentado é retornado os gráficos de influência local para perturbação de casos ponderados (ver Figura 100). Naturalmente, o usuário das rotinas pode obter os gráficos na forma que for mais conveniente. Por fim, a última função desse pacote refere-se à aplicação dos testes para testar a hipótese de variabilidade espacial dos parâmetros de regressão. O conceito teórico referente a esse teste de hipótese para os modelos geograficamente ponderados pode ser consultado em Fotheringham, Brunson e

Charlton (2002). No contexto computacional, esse teste pode ser aplicado ao ajuste de um modelo REGP por meio da seguinte sintaxe:

```
gwer.montecarlo(formula, family = Normal, data = list(),
nsims = 99, kernel = "bisquare", adaptive = F,
bw, p = 2, theta = 0, dispersion = NULL,
longlat = F, dMat, control = glm.control(
epsilon = 1e-04, maxit = 100, trace = F))
```

Para isto, o usuário deve fornecer o mesmo conjunto de entradas utilizadas para ajustar um modelo de REGP. Adicionalmente, o argumento `nsims` refere-se ao número de réplicas de Monte Carlo usado no cálculo da estatística de testes. Essa função retorna como saída uma tabela com os valores-p para os parâmetros de cada um dos preditores. Uma aplicação simples dessa função com base nos ajustes feitos nesse capítulo é considerada a seguir. Os resultados ressaltam a maioria dos parâmetros locais não apresentam variabilidade espacial.

```
> gwer.montecarlo(formula, data = gSRDF, family = Student(df=3),
+                 kernel='bisquare', adaptive = TRUE, bw = regp.b)
```

Tests based on the Monte Carlo significance test

p-value	
(Intercept)	0.36
TotPop90	0.00
PctRural	0.32
PctFB	0.06
PctPov	0.65

5 CONCLUSÕES

Na presente tese foram abordadas metodologias para análise de dados espaciais assumindo a não estacionariedade espacial entre as relações da variável resposta com suas variáveis regressoras. Nesse contexto, a regressão geograficamente ponderada mostrou-se uma alternativa viável que permite a modelagem de dados espaciais, uma vez que seus parâmetros podem variar localmente no espaço. Essa classe de modelos apresenta estimativas locais que podem ser mapeadas e interpretadas com facilidade, no entanto há pouca flexibilidade em relação à suposição dos erros aleatórios. Por esse fato, as classes propostas nessa tese são alternativas a esse tipo de modelagem supondo, nesse caso, que os erros pertençam à família de distribuições elípticas. Contudo, esta classe de modelos geograficamente ponderados com erros elípticos, proposta no presente trabalho, assume que todas as relações de regressão variam na mesma escala geográfica. Como discutido nessa tese, essa suposição pode ser bastante irrealista e levar a interpretações errôneas. Neste contexto, propomos também uma extensão dessa classe de modelos que funciona em várias escalas. Essa extensão flexibiliza ainda mais o modelo de REGP, permitindo que cada relação de regressão tenha sua própria escala espacial.

Adicionalmente, foram discutidos resultados empíricos com base em diversos cenários de simulação, como uma forma de avaliar o desempenho das classes de modelos propostas nesta tese. Estes estudos de simulação indicaram que o uso de distribuições pertencentes a família elíptica na REGP gera estimativas mais robusta na presença de valores discrepantes em comparação com a RGP clássica. Aplicações em base de dados reais também foram consideradas, os quais apresentaram resultados satisfatórios para a detecção de valores discrepantes que apresentam fortes influências nas estimativas dos parâmetros. A influência dessas observações na modelagem dos dados foi contornada através da utilização de distribuições com caudas mais pesadas que a normal. Como isso, os modelos propostos nesta tese possibilitam uma modelagem da não estacionariedade espacial na relação de regressão, apresentando uma estimação robusta na presença de valores discrepantes. Nesse contexto, a seguir são discutidos alguns resultados específicos obtidos através da utilização desses modelos, bem como possíveis melhoramentos para a aplicação em trabalhos futuros.

5.1 RESULTADOS E CONTRIBUIÇÕES

Como foi discutido anteriormente, a RGP pode ser particularmente sensível a presença de valores discrepantes, sendo um problema ocasional nesses modelos de regressão (Fotheringham, Brunston e Charlton, 2002). As possíveis soluções apresentadas em Fotheringham, Brunston e Charlton (2002) tem suas limitações e apresentam elementos de subjetividade, como foi ressaltado em Warsito *et al.* (2018). Nesse sentido, os modelos propostos nesta tese, demonstraram acomodar bem os valores discrepantes ao assumir distribuições com caudas pesadas para os erros aleatórios. Além disso, foram apresentadas propostas inovadoras de métodos para o diagnóstico dos modelos de REGP, até então não explorado na literatura para a RGP. Esses métodos se mostraram eficazes para detectar observações influentes e avaliar as suposições do modelo.

Por outro lado, vimos que a escala é um conceito de suma importância na análise de dados espaciais, a qual é diretamente associada à definição do parâmetro de suavização em modelos de regressão geograficamente ponderados. No entanto, assim como na RGP, assumimos que as relações de regressão nos modelos de REGP variam na mesma escala geográfica. Com a finalidade de flexibilizar essa limitação na RGP, Fotheringham, Yang e Kang (2017) introduziu uma classe de modelos geograficamente ponderados que funciona em diferentes escalas geográficas. Posteriormente, Yu *et al.* (2020) obteve resultados inferenciais importantes nessa extensão da RGP, permitindo uma análise mais completa nessa classe de modelos. Esta extensão da RGP, como sugere Comber *et al.* (2020), possibilita uma análise mais realista dos dados, uma vez que cada relação de regressão possui sua própria escala espacial. Com base nesses estudos, desenvolvemos a REGPM como uma proposta de REGP que funciona em diferentes escalas geográficas, ao assumir que cada relação de regressão possui um parâmetro de suavização distinto.

Dessa forma, como principais contribuições obtidas ao longo do desenvolvimento teórico e prático dessa tese, destaca-se:

- A introdução da REGP, estendendo a RGP para o caso em que a componente aleatória pertence à família de distribuições elípticas. Esses modelos permitem a modelagem de dados espaciais usando o conceito de ponderações geográficas, porém apresentando menor sensibilidade a presença de observações discrepantes.
- A proposta da REGP considerando diferentes escalas espaciais para cada relação de regressão, a qual foi intitulada de REGPM. Através das escalas espaciais variáveis é

possível descrever de forma mais precisa a não estacionariedade espacial presente na relação de regressão em um modelo geograficamente ponderado com erros aleatórios pertencente à família de distribuições elípticas.

- A obtenção das medidas de diagnósticos para a classe dos modelos de REGP, tais como resíduos, alavancagem generalizada, distância de Cook generalizada e influência local. Algumas dessas medidas também foram estendidas para os modelos de REGPM.
- A implementação computacional do pacote **gwer** no *software* R para a aplicação da REGP e REGPM. Esse pacote conta com diversas funções destinadas tanto ao ajuste do modelo quanto a análise de diagnóstico.
- Uma aplicação prática da REGP com dados reais referente a um estudo sobre o desenvolvimento humano e à criminalidade dos municípios cearenses no ano de 2016, mostrando a importância da utilização desses modelos em dados com presença de valores discrepantes.
- Uma aplicação prática da REGPM com dados reais referente ao estudo do desempenho educacional nos condados da Geórgia, EUA. Essa análise ressalta a importância da utilização de diferentes escalas espaciais na interpretação dos resultados.
- A apresentação dos primeiros resultados obtidos em um congresso internacional, 34th IWSM, em julho de 2019. Esses resultados foram publicados no livro de resumo do correspondente congresso.
- A produção de um artigo científico intitulado “*Geographically weighted elliptical regression model: Estimation, residuals and diagnostic analysis*”, o qual encontra-se em processo de submissão à revista internacional *Computers & Geosciences*.
- O desenvolvimento de materiais teóricos e práticos destinados a produções de outros dois artigos científicos referente a REGPM e o pacote computacional.

5.2 TRABALHOS FUTUROS

Como trabalhos a serem desenvolvidos futuramente a partir do conteúdo discutido nessa tese, destaca-se as seguintes possibilidades:

- Introduzir modelos de regressão espaço-temporais com erros aleatórios elípticos utilizando funções de ponderação espaço-temporal ou o conceito de regressão em painel. Estes tipos de modelos levariam em conta a interação direta entre as

dinâmicas espacial e temporal do modelo.

- Propor modelos de regressão espaço-temporais com erros aleatórios elípticos, considerando a inclusão de dinâmicas temporais através de um processo autoregressivos de médias móveis (ARMA). Esta seria uma proposta mais complexa e flexível para a introdução da dinâmica temporal na REGP.
- Considerar técnicas bayesianas para a formulação da REGP e demais extensões desses modelos, fornecendo um processo de estimação alternativo aos métodos clássicos.
- Propor modelos de RGP com o enfoque em famílias distribuições distintas, aumentando assim a flexibilidade da componente aleatória do modelo.
- Implementar novas rotinas e melhorar o desempenho computacional do pacote **gwer** no *software* R, facilitando ainda mais a aplicação dessas novas classes de modelos à problemas reais.

Finalmente, espera-se que o desenvolvimento desse estudo possa acrescentar novas possibilidades na modelagem da não estacionariedade espacial dos parâmetros em modelos de regressão, e que os resultados teóricos e práticos obtidos possam motivar novas produções científicas nessa área de pesquisa.

REFERENCIAS

- ATKINSON, P. M.; GERMAN, S. E.; SEAR, D. A.; CLARK, M. J. Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression. **Geographical Analysis**, Wiley Online Library, v. 35, n. 1, p. 58–82, 2003.
- BIVAND, R. S.; WONG, D. W. Comparing implementations of global and local indicators of spatial association. **TEST**, Springer, v. 27, n. 3, p. 716–748, 2018.
- BREIMAN, L.; FRIEDMAN, J. H. Estimating optimal transformations for multiple regression and correlation. **Journal of the American statistical Association**, Taylor & Francis Group, v. 80, n. 391, p. 580–598, 1985.
- BRUNSDON, C.; FOTHERINGHAM, A. S.; CHARLTON, M. E. Geographically weighted regression: a method for exploring spatial nonstationarity. **Geographical Analysis**, Wiley Online Library, v. 28, n. 4, p. 281–298, 1996.
- BRUNSDON, C.; FOTHERINGHAM, S.; CHARLTON, M. Geographically weighted regression. **Journal of the Royal Statistical Society: Series D (The Statistician)**, Wiley Online Library, v. 47, n. 3, p. 431–443, 1998.
- BYRNE, G.; CHARLTON, M.; FOTHERINGHAM, S. Multiple dependent hypothesis tests in geographically weighted regression. In: UNIVERSITY OF NEW SOUTH WALES. **Proceedings of the 10th International Conference on GeoComputation**. [S.l.], 2009.
- CERIOLI, A.; RIANI, M. The ordering of spatial data and the detection of multiple outliers. **Journal of computational and graphical statistics**, Taylor & Francis, v. 8, n. 2, p. 239–258, 1999.
- COMBER, A.; BRUNSDON, C.; CHARLTON, M.; DONG, G.; HARRIS, R.; LU, B.; LÜ, Y.; MURAKAMI, D.; NAKAYA, T.; WANG, Y. *et al.* The gwr route map: a guide to the informed application of geographically weighted regression. **arXiv preprint arXiv:2004.06070**, 2020.
- COOK, R. D. Detection of influential observation in linear regression. **Technometrics**, Taylor & Francis Group, v. 19, n. 1, p. 15–18, 1977.
- COOK, R. D. Assessment of local influence. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 133–169, 1986.
- COX, D. R.; HINKLEY, D. V. **Theoretical statistics**. [S.l.]: London: Chapman and Hall, 1974.
- COX, D. R.; SNELL, E. J. A general definition of residuals. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 248–275, 1968.
- CYSNEIROS, F. J. A.; PAULA, G. A. Restricted methods in symmetrical linear regression models. **Computational Statistics & Data Analysis**, Elsevier, v. 49, p. 689–708, 2005.
- CYSNEIROS, F. J. A.; PAULA, G. A. Residuals and their statistical properties in symmetrical nonlinear models. **Statistics & Probability Letters**, Elsevier, v. 78, n. 12, p. 3269–3273, 2008.

- CYSNEIROS, F. J. A.; PAULA, G. A.; GALEA, M. Heteroscedastic symmetrical linear models. **Statistics & Probability Letters**, Elsevier, v. 77, n. 11, p. 1084–1090, 2007.
- Da Silva, A. R.; FOTHERINGHAM, A. S. The multiple testing issue in geographically weighted regression. **Geographical Analysis**, Wiley Online Library, v. 48, n. 3, p. 233–247, 2016.
- Da Silva, A. R.; LIMA, A. d. O. Geographically weighted beta regression. **Spatial Statistics**, Elsevier, v. 21, p. 279–303, 2017.
- Da Silva, A. R.; RODRIGUES, T. C. V. Geographically weighted negative binomial regression—incorporating overdispersion. **Statistics and Computing**, Springer, v. 24, n. 5, p. 769–783, 2014.
- DAI, X.; JIN, L.; SHI, A.; SHI, L. Outlier detection and accommodation in general spatial models. **Statistical Methods & Applications**, Springer, v. 25, n. 3, p. 453–475, 2016.
- DAVISON, A.; GIGLI, A. Deviance residuals and normal scores plots. **Biometrika**, Oxford University Press, v. 76, n. 2, p. 211–221, 1989.
- De Bastiani, F.; CYSNEIROS, A. H. M. A.; URIBE-OPAZO, M. A.; GALEA, M. Influence diagnostics in elliptical spatial linear models. **TEST**, Springer, v. 24, p. 332–340, 2015.
- De Bastiani, F.; URIBE-OPAZO, M.; GALEA, M.; CYSNEIROS, A. Case-deletion diagnostics for spatial linear mixed models. **Spatial statistics**, Elsevier, v. 28, p. 284–303, 2018.
- DUNN, P. K.; SMYTH, G. K. Randomized quantile residuals. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 5, n. 3, p. 236–244, 1996.
- FANG, K.; KOTZ, S.; NG, K. **Symmetric Multivariate and Related Distributions**. [S.l.]: London: Chapman and Hall, 1990.
- FOTHERINGHAM, A.; BRUNSDON, C.; CHARLTON, M. Geographically weighted regression: The analysis of spatially varying relationships. **John Wiley & Sons**, v. 13, 01 2002.
- FOTHERINGHAM, A. S. “the problem of spatial autocorrelation” and local spatial statistics. **Geographical analysis**, Wiley Online Library, v. 41, n. 4, p. 398–403, 2009.
- FOTHERINGHAM, A. S.; BRUNSDON, C.; CHARLTON, M. **Geographically weighted regression**. [S.l.]: John Wiley & Sons, Limited West Atrium, 2003.
- FOTHERINGHAM, A. S.; YANG, W.; KANG, W. Multiscale geographically weighted regression (mgwr). **Annals of the American Association of Geographers**, Taylor & Francis, v. 107, n. 6, p. 1247–1265, 2017.
- GALEA, M.; PAULA, G. A.; CYSNEIROS, F. J. A. On diagnostics in symmetrical nonlinear models. **Statistics & Probability Letters**, Elsevier, v. 73, n. 4, p. 459–467, 2005.
- GENTON, M. G. Spatial breakdown point of variogram estimators. **Mathematical Geology**, Springer, v. 30, n. 7, p. 853–871, 1998.

GENTON, M. G. Robustness problems in the analysis of spatial data. In: **Spatial statistics: Methodological aspects and applications**. [S.l.]: Springer, 2001. p. 21–37.

GREIG, D. M. **Optimisation**. [S.l.]: Longman Publishing Group, 1980.

HARRIS, P.; BRUNSDON, C.; CHARLTON, M.; JUGGINS, S.; CLARKE, A. Multivariate spatial outlier detection using robust geographically weighted methods. **Mathematical Geosciences**, Springer, v. 46, n. 1, p. 1–31, 2013.

HARRIS, P.; FOTHERINGHAM, A. S.; JUGGINS, S. Robust geographically weighted regression: a technique for quantifying spatial relationships between freshwater acidification critical loads and catchment attributes. **Annals of the Association of American Geographers**, Taylor & Francis, v. 100, n. 2, p. 286–306, 2010.

HASLETT, J.; BRADLEY, R.; CRAIG, P.; UNWIN, A.; WILLS, G. Dynamic graphics for exploring spatial data with application to locating global and local anomalies. **The American Statistician**, Taylor & Francis Group, v. 45, n. 3, p. 234–242, 1991.

IBACACHE-PULGAR, G.; PAULA, G. A.; CYSNEIROS, F. J. A. Semiparametric additive models under symmetric distributions. **TEST**, Springer, v. 22, n. 1, p. 103–121, 2013.

JIN, L.; DAI, X.; SHI, A.; SHI, L. Detection of outliers in mixed regressive-spatial autoregressive models. **Communications in statistics-theory and methods**, Taylor & Francis, v. 45, n. 17, p. 5179–5192, 2015.

KELKER, D. Distribution theory of spherical distributions and a location-scale parameter generalization. **Sankhyā: The Indian Journal of Statistics, Series A**, JSTOR, p. 419–430, 1970.

LANGE, K. L.; LITTLE, R. J.; TAYLOR, J. M. Robust statistical modeling using the t distribution. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 84, n. 408, p. 881–896, 1989.

LESAFFRE, E.; VERBEKE, G. Local influence in linear mixed models. **Biometrics**, JSTOR, p. 570–582, 1998.

LESAGE, J. P. **The theory and practice of spatial econometrics**. [S.l.]: University of Toledo. Toledo, Ohio, 1999. v. 28.

LUO, J.; KANALA, N. K. Modeling urban growth with geographically weighted multinomial logistic regression. In: INTERNATIONAL SOCIETY FOR OPTICS AND PHOTONICS. **Geoinformatics 2008 and Joint Conference on GIS and Built Environment: The Built Environment and Its Dynamics**. [S.l.], 2008. v. 7144, p. 71440M.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**. [S.l.]: CRC press, 1989. v. 37.

MILITINO, A. F.; PALACIOS, M. B.; UGARTE, M. D. Robust trend parameters in a multivariate spatial linear model. **Test**, Springer, v. 12, n. 2, p. 445–457, 2003.

NAKAYA, T.; FOTHERINGHAM, A. S.; BRUNSDON, C.; CHARLTON, M. Geographically weighted poisson regression for disease association mapping. **Statistics in medicine**, Wiley Online Library, v. 24, n. 17, p. 2695–2717, 2005.

PIERCE, D. A.; SCHAFER, D. W. Residuals in generalized linear models. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 81, n. 396, p. 977–986, 1986.

PREGIBON, D. *et al.* Logistic regression diagnostics. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 9, n. 4, p. 705–724, 1981.

PRESS, W. H.; WILLIAM, H.; TEUKOLSKY, S. A.; VETTERLING, W. T.; SAUL, A.; FLANNERY, B. P. **Numerical recipes 3rd edition: The art of scientific computing**. [S.l.]: Cambridge university press, 2007.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2019. Disponível em: <<https://www.R-project.org/>>.

RAO, B. P. Remarks on univariate elliptical distributions. **Statistics & Probability Letters**, Elsevier, v. 10, n. 4, p. 307–314, 1990.

TIBSHIRANI, R.; HASTIE, T. Local likelihood estimation. **Journal of the American Statistical Association**, Taylor & Francis Group, v. 82, n. 398, p. 559–567, 1987.

VANEGAS, L. H.; RONDON, L. M.; CYSNEIROS, F. J. A. Assessing robustness of inference in symmetrical nonlinear regression models. **Communications in Statistics-Theory and Methods**, Taylor & Francis, v. 42, n. 9, p. 1692–1711, 2013.

VILLEGAS, C.; PAULA, G. A.; CYSNEIROS, F. J. A.; GALEA, M. Influence diagnostics in generalized symmetric linear models. **Computational Statistics & Data Analysis**, Elsevier, v. 59, p. 161–170, 2013.

WARSITO, B.; YASIN, H.; ISPRIYANTI, D.; HOYYI, A. Robust geographically weighted regression of modeling the air polluter standard index (apsi). In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2018. v. 1025, n. 1, p. 012096.

WEI, B.-C.; HU, Y.-Q.; FUNG, W.-K. Generalized leverage and its applications. **Scandinavian Journal of Statistics**, Wiley Online Library, v. 25, n. 1, p. 25–37, 1998.

WESTFALL, P. H.; ARIAS, A. L. **Understanding Regression Analysis: A Conditional Distribution Approach**. [S.l.]: CRC Press, 2020.

YU, H.; FOTHERINGHAM, A. S.; LI, Z.; OSHAN, T.; KANG, W.; WOLF, L. J. Inference in multiscale geographically weighted regression. **Geographical Analysis**, Wiley Online Library, v. 52, n. 1, p. 87–106, 2020.